Diabetes Health Indicator Analytics

Wanqing Li 20146670 | 18wl19@queensu.ca
Yulu Chen 20150856 | 18yc124@queensu.ca
Lu Chen 20164422 | 18lc44@queensu.ca
Yingqi Xu 20158700 | 18yx84@queensu.ca
Sixuan Zhang 20204414 | 19sz61@queensu.ca
Yiying Chen 20198951 | 19yc56@queensu.ca

Instructor: Dr. Brian Ling
April 2023

Contributions:
Wanqing Li: organizations, data visualizations, build clustering model (discarded), presentation, introduction, discussion and conclusion
Sixuan Zhang: data visualizations, build decision tree model, presentation
Yingqi Xu: data visualizations, build random forest model, presentation
Yiying Chen: data visualizations, build KNN model, presentation
Yulu Chen: data visualizations, build Naïve Bayes model, presentation
Lu chen: data visualizations, build logistic regression model, presentation

1.0 Introduction

According to IDF Diabetes statistics [1], there are five thirty seven million adults all over the world living with diabetes which means 1 in 10 adults will experience diabetes. Diabetes is responsible for 6.7 million deaths in 2021 which means 1 Death every 5 seconds. Diabetic patients more likely to be hospitalized due to serious risks such as Blindness, Heart Disease, Loss of arm or legs, kidney failure and reduce lifetime etc. Meanwhile, Diabetes also causes hundreds of billion US dollars in health expenditure around the world. That's why it's important for us to understand how to prevent and manage diabetes effectively. By analyzing risk factors. and use. We hope that we can help people to know their diabate possibility as an early screening tool. Make people live healthier. And reduce the overall burden of this disease on our society. To achieve our goal, we chose a dataset called Diabete Health Indicator [2] from Kaggle, prior works on Kaggle are mostly using the Neural Networks to train prediction models, and we only use statistical methods in this report to try to compare with them.

2.0 Dataset visualization

Strategies such as weight loss, healthy eating, exercise, and medical treatments can help mitigate the harm of the disease. Although most diabetes are hard to cure, there are various approaches that can help alleviate its negative effects in patients, such as adopting healthy habits like weight loss, healthy eating, and regular exercise, as well as seeking medical treatment. Early diagnosis is crucial in promoting lifestyle changes and more effective treatment, highlighting the significance of predictive models for diabetes risk as essential tools for public and health officials [2].

To better visualize the dataset, we divide the dataset detail into 3 parts in 2.1, 2.2 and 2.3.

2.1 shows personal information, 2.2 shows health status, and 2.3 shows living habits and eating habits. And we summarized a graph in ***figure 1*** to show each group, the each group title represented as a light orange box, the green box means categorical attribute, yellow box means numeric attribute, blue box means binary attribute.

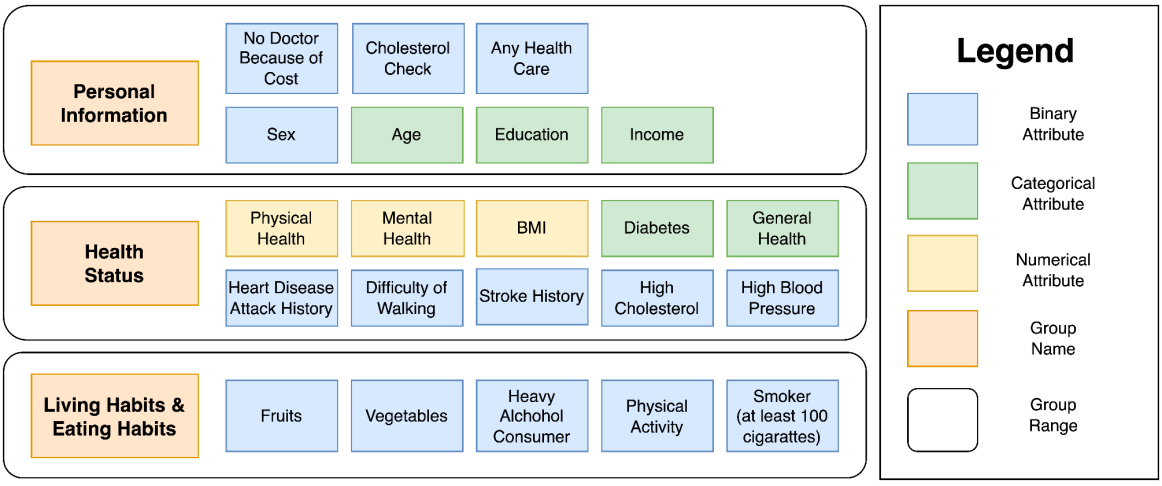

Figure 1. Dataset Conclusion [draw.io: link]

2.1 Personal Information

We can see from the ***figure 2***, it shows the relationship between age groups and diabetes. From this figure it indicates that the median age of individuals with diabetes is higher than the median age of individuals without diabetes and individuals with prediabetes. This suggests that age may be a risk factor for developing diabetes. ***Figure 3*** is a violin plot that displays the distribution of body mass index (BMI) for females and males. It suggests that there is a small difference in the distribution of BMI values between males and females, with females having a wider distribution of values and slightly higher median BMI value. ***Figure 4*** bar plot that displays the count of records in the dataset for each age group. It suggests that there is a trend of increasing the number of records with increasing age, with a peak around the age group of 45-49. ***Figure 5*** histogram shows the distribution of BMI values in the dataset. It is a unimodal distribution with a peak around 25-30 (health BMI range is 18.5 - 24.9), indicating that the majority of the individuals in the dataset have a BMI in the overweight or obese range.

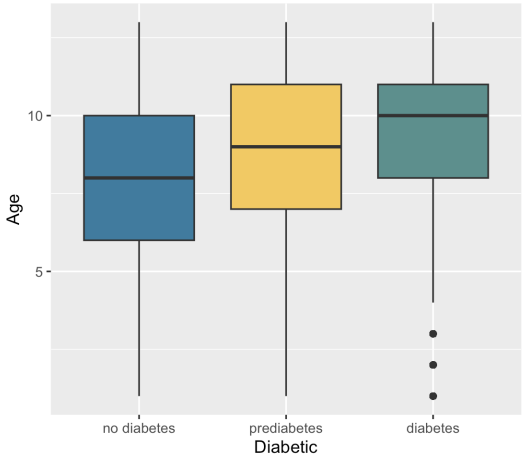

Figure 2. Age vs Diabete

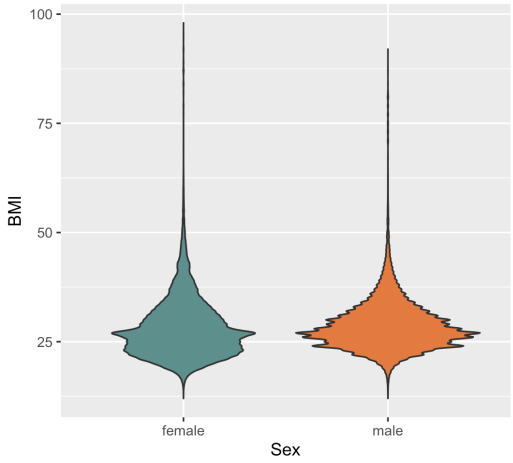

Figure 3. Sex vs BMI

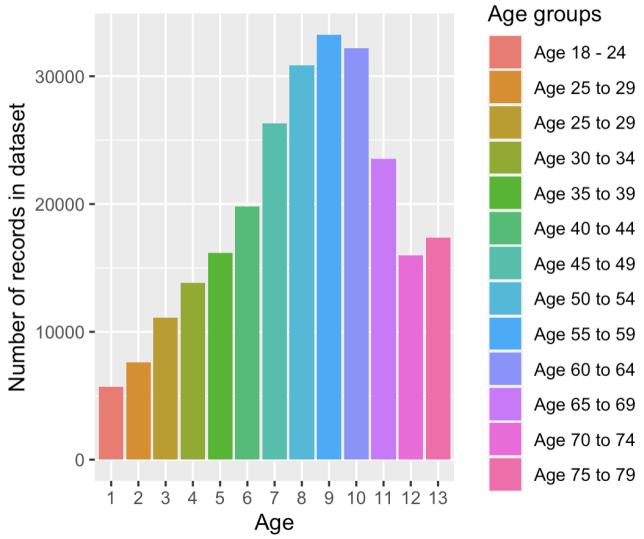

Figure 4. Age distribution

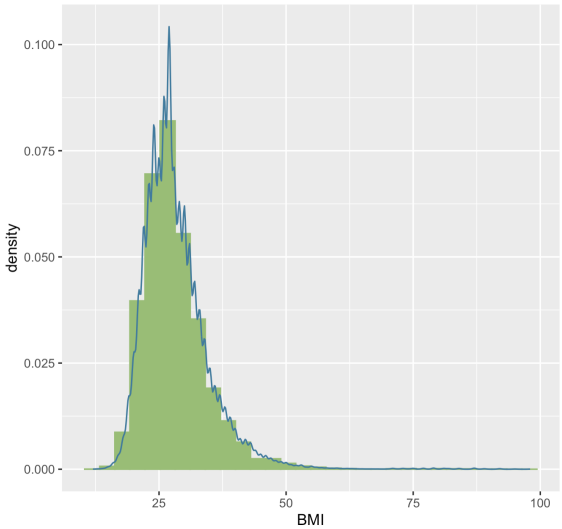

Figure 5. BMI histogram

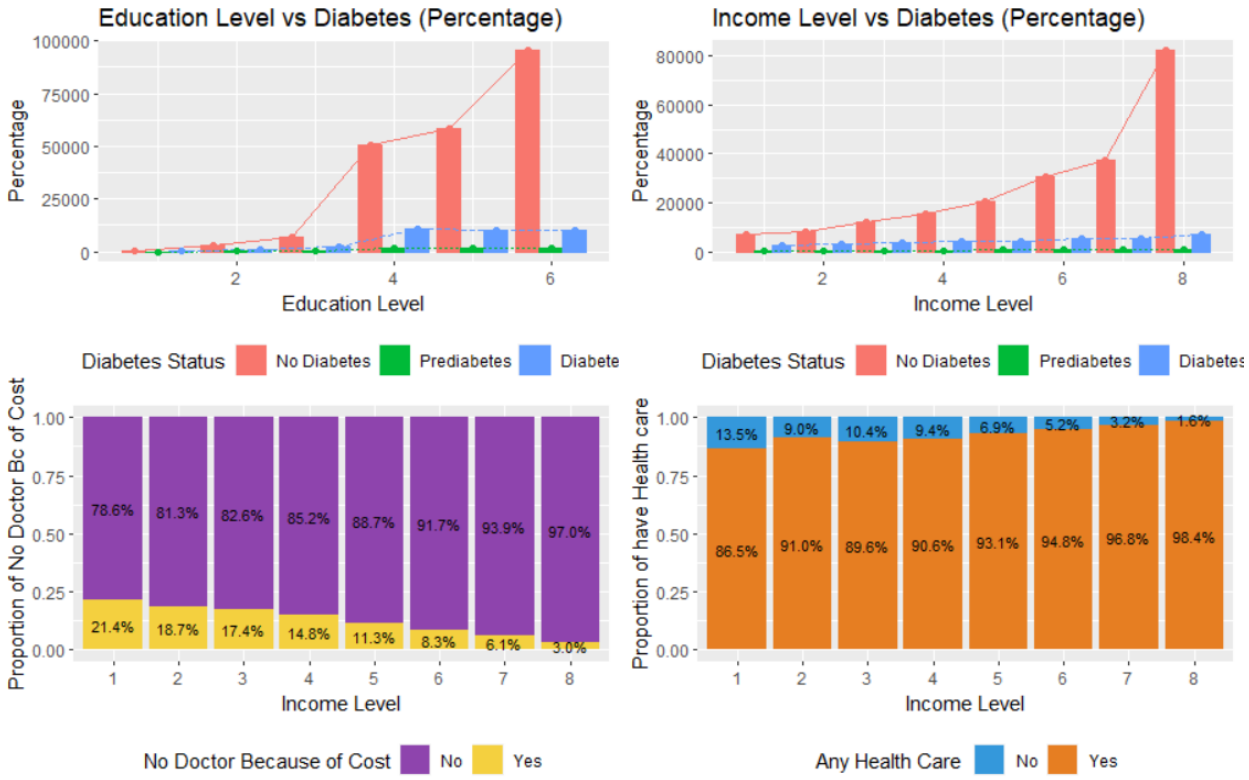

Figure 6. Education level and Income level

The ***figure 6*** displays trends in diabetes status and health care access across different education and income levels. Generally, from the upper left corner and upper right corner of ***figure 6***, as education and income levels increase, the percentage of individuals with no diabetes also increases, while the percentage of individuals with diabetes decreases. Prediabetes percentages remain relatively low across all education and income levels.

In terms of health care access, the bottom left corner of ***figure 6*** shows the percentage of individuals who did not visit a doctor because of cost decreases as income level increases, indicating that higher-income individuals are less likely to face financial barriers when seeking medical care. On the other hand, the bottom right corner of ***figure 6*** shows the percentage of individuals with any health care coverage increases with higher income levels, suggesting that higher-income individuals have better access to health care services. Overall, these trends highlight the associations between socioeconomic factors, such as education and income, and health outcomes, as well as access to health care services.

2.2 Individual Health Status

The present report provides a box plot that shows the average BMI of individuals with varying diabetes statuses in our study. The values "0", "1", and "2" represent the absence of diabetes, pre-diabetes, and diabetes, respectively. From the ***figure 7*** we can know the diabetes label

distribution is not very balanced, the “no diabetes” occupied a very large amount of records in the dataset while “prediabetes” and “diabetes” has less representation in the dataset.

It is noteworthy that all three groups in ***figure 8*** have an average BMI greater than 25, indicating that being overweight is prevalent among individuals with varying diabetes statuses in our study. However, our analysis reveals a positive correlation between BMI and diabetes status, with higher BMI levels being associated with an increased likelihood of developing diabetes. Thus, it is reasonable to conclude that higher BMI is a significant predictor of diabetes risk in our sample. These findings have important implications for public health initiatives aimed at preventing and managing diabetes, particularly among individuals who are overweight or obese. Further research may be necessary to validate these findings and identify other factors that may influence diabetes risk in this population.

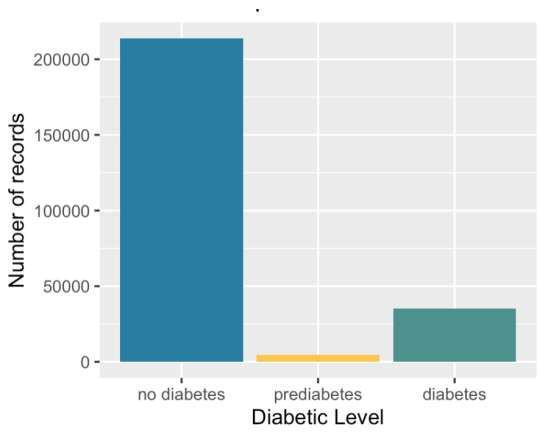

Figure 7. Diabetes Label Distribution

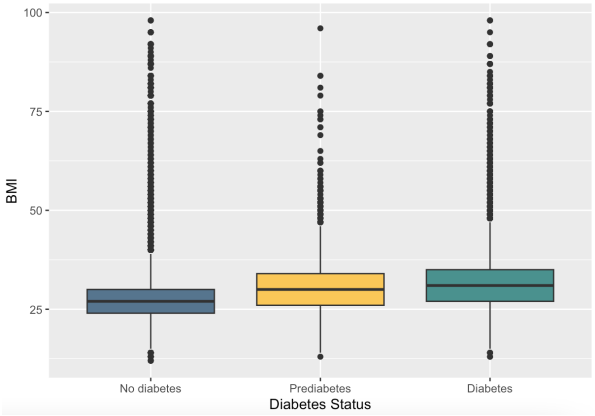

Figure 8. Diabetes status vs BMI

The current report presents a bar plot that illustrates the health status of individuals in our study. Based on the ***figure 9***, it is evident that a smaller proportion of the participants reported difficulty walking, heart disease, or stroke. However, approximately 50% of the participants in our study reported having high blood pressure and high cholesterol. The bar graph in ***figure 10*** shows the relationship between different individual health statuses and diabetes status. We observed that all health statuses contribute to a similar proportion of diabetes, and among them, individuals with difficulty walking and heart disease or attack have a higher risk of developing diabetes. This finding indicates that there may be some sample bias present in our study, as individuals with high blood pressure and high cholesterol are known to be more likely to develop diabetes and constitute a significant proportion of our sample. Therefore, this suggests caution is necessary when generalizing the results of this study to the broader population. Further research may be required to account for potential sample bias and obtain a more representative sample.

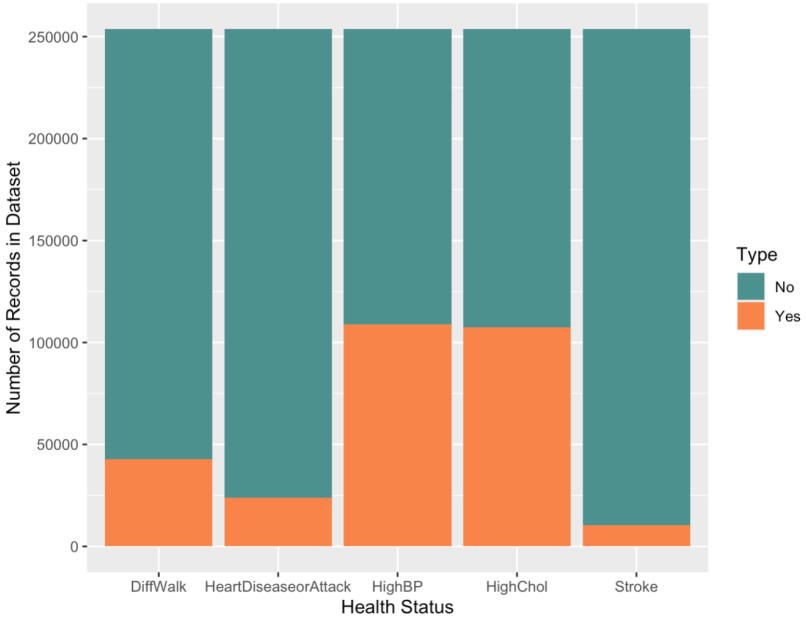

Figure 9. Individual health status

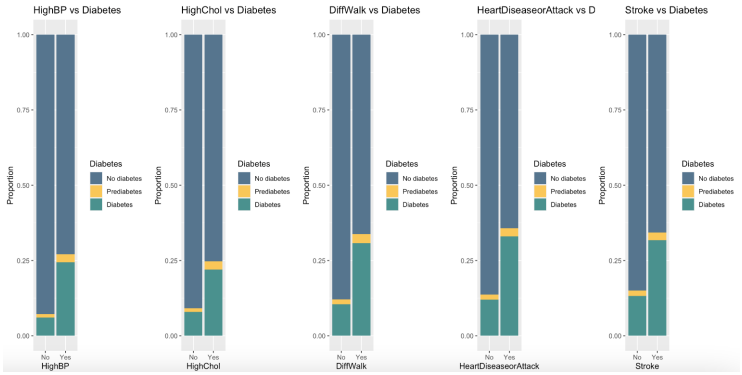

Figure 10. Health status vs Diabetes status

Figure 11 displays the proportions of participants with different diabetes types and their average general health scores. People without diabetes constitute the majority among all participants, and they predominantly report being in "very good" health. In contrast, those with prediabetes or diabetes form a minority and mainly report being in "good" health. The chart primarily highlights two data trends. First, individuals without diabetes demonstrate a healthier general trend compared to those with prediabetes or diabetes. Second, the disparity in participant numbers between those without diabetes and those with diabetes may indicate potential bias in further studies, as there are significantly more participants without diabetes.

Regarding health conditions in more detail, this bar plot compares different diabetes types and participants' perceptions of their health, including mental and physical health, over the past 30 days. ***Figure 12*** reveals that individuals generally feel worse as they progress from diabetes type 0 to type 2, both in terms of mental and physical health. The difference is more pronounced for physical health. People with diabetes may feel unwell physically for an average of 8 days over the past 30 days, while those without diabetes experience only 4 days of feeling unwell during

the same period. The difference in mental health is relatively smaller. According to the graph, people with prediabetes and diabetes have nearly the same number of days with mental health issues over the past 30 days.

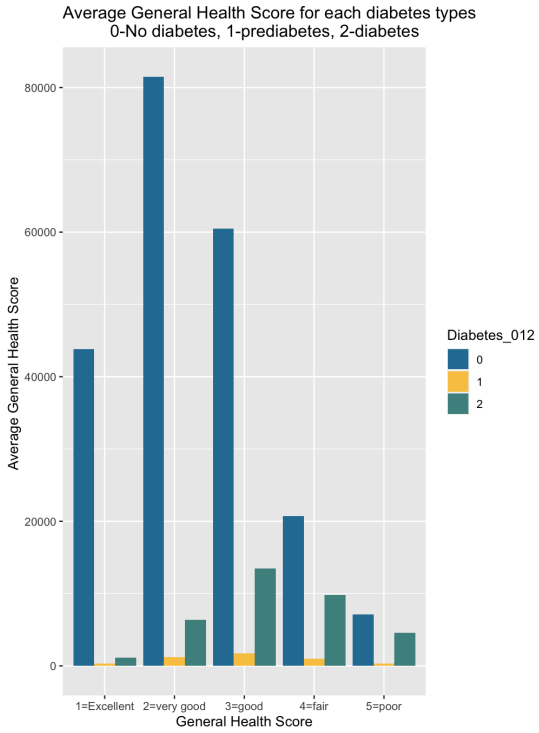

Figure 11: Average General Health Score v.s. General Health Score for diabetes types 0 (No diabetes), 1(prediabetes) , 2 (diabetes)

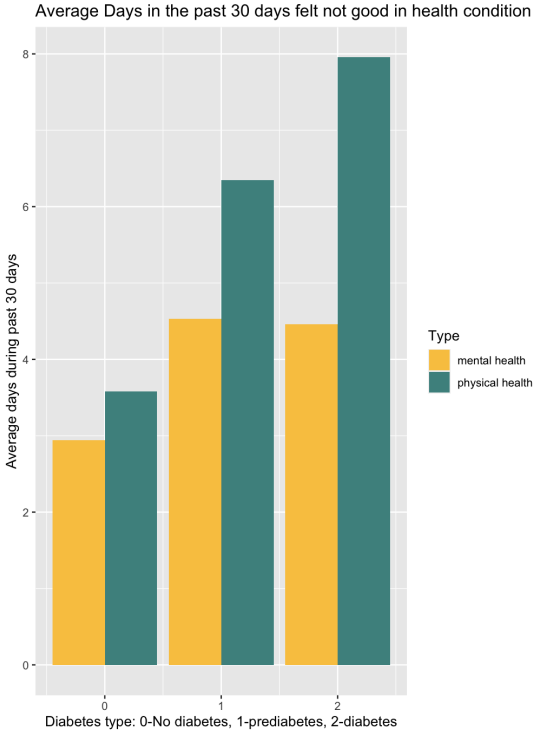

Figure 12. Average days during the past 30 days v.s. Diabetes type for mental health and physical health

2.3 Living habits and eating habits

Figure 13 features a set of four charts that compare different lifestyles with regard to their association with diabetes. The charts utilize the labels "yes" and "no," with "yes" denoting individuals who are vegetarian, heavy drinkers, smokers, and physically active, and "no" indicating those who are not. The graphs reveal that vegetarians have the lowest prevalence of diabetes and prediabetes, accounting for only 15%. Conversely, heavy drinkers are more likely to be diabetes-free than non-drinkers, with nearly 90% of them not having diabetes. Smokers have a higher incidence of diabetes, at approximately 20%, compared to non-smokers. Individuals who engage in regular physical activity have a lower risk of developing diabetes, with nearly 88% of them being diabetes-free. Conversely, those who do not engage in physical activity have a 75% chance of being diabetes-free. Therefore, we can get the result from above that veggie, smoke, and physical activity do have some positive effects in diabetes.

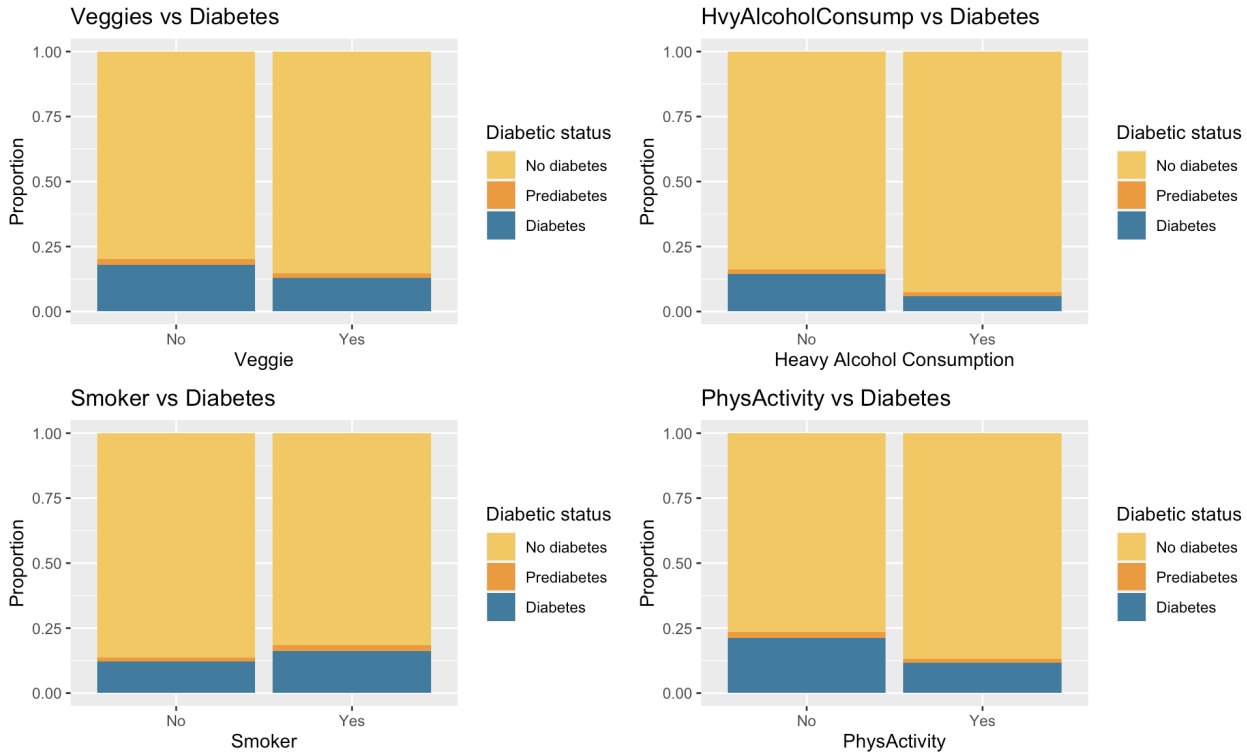

Figure 13. Lifestyles (Veggies, Heavy Alcohol Consumption, Smoker, Physical Activity) vs Diabetes

3.0 Methodology

In the beginning of methodologies, we did preprocessing steps like train test splitting to 7:3, and remove all the missing values. Each method subsection will be divided into model explanation, preprocessing (if it has extra steps), and result analysis.

3.1 Naïve Bayes

Model explanation: The Naïve Bayes classifier is a supervised nonlinear classification algorithm that calculates the conditional probability of the events based on Bayes Theorem, and it assumes the predictors are independent [3].

The formula of Bayes Theorem is given by $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$.

The Naïve Bayes classifier is suitable for the present dataset as it can handle both numerical and categorical data, including high-dimensional datasets [4]. The Naïve Bayes classifier uses different methods to estimate the class conditional distributions. In this dataset, we utilize kernel density estimation (KDE) to estimate the class conditional densities for numerical variables. Besides, the Categorical distribution applies the Bernoulli distribution for binary variables and applies the Multinomial distribution for categorical variables with more than two outcomes.

Result: Naïve Bayes classifier has an accuracy of 0.84 on this dataset, indicating that 84% of testing samples are correctly classified. However, the accuracy that predicts ‘no diabetes' is 0.81,

this reveals that accuracy is not useful in indicating the performance of prediabetic and diabetic classes. As shown in ***table 1***, the sensitivity of no diabetes is 0.96, which means that 96% of samples with diabetes are correctly identified. However, the sensitivity of prediabetes and diabetes groups are 0 and 0.25 respectively, this indicates that the probability of correctly distinguishing prediabetic and diabetic cases is low.

Table 1. Sensitivity and Specificity of Naïve Bayes.

	No diabetes	prediabetes	diabetes
Sensitivity	0.96	0	0.25
Specificity	0.24	1	0.95

3.2 Decision tree

Model explanation: In decision tree methods, the predictor space is divided into simple regions to predict outcomes for new observations. The mean (for continuous response) or mode (for categorical response) response value of the training observations in the region to which the new observation belongs is typically used. The splitting rules used to segment the predictor space can be summarized in a tree. These methods can be used for both regression and classification problems, with classification trees used for categorical outcomes and regression trees used for continuous outcomes [5]. Decision trees are versatile, as they can handle both categorical and numerical data, making them a suitable tool for a wide range of data types in a given dataset.

Preprocessing: Train test is splitted in two methods, which are training data 1:200 and training data 7:3.

Result: In ***figure 15***, when we split the data into a small training data, the tree will show more branches, when we split the data into larger training data, ***figure 14*** shows that the tree will show small branches. Classification accuracy in the larger training data is 0.84 and accuracy in the smaller training data is 0.76, which shows the data is effective in classifying data. For the CV plot, ***figure 16*** shows that there are no terminal nodes of the tree, which means it's hard to improve the prediction accuracy and it's a simple model.

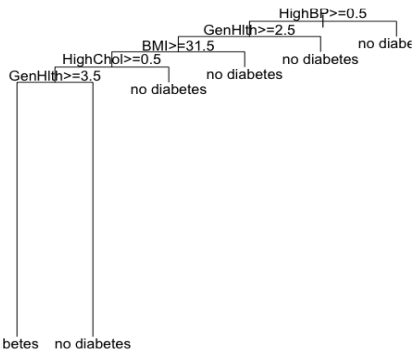

Figure 14. Large training data 7:3

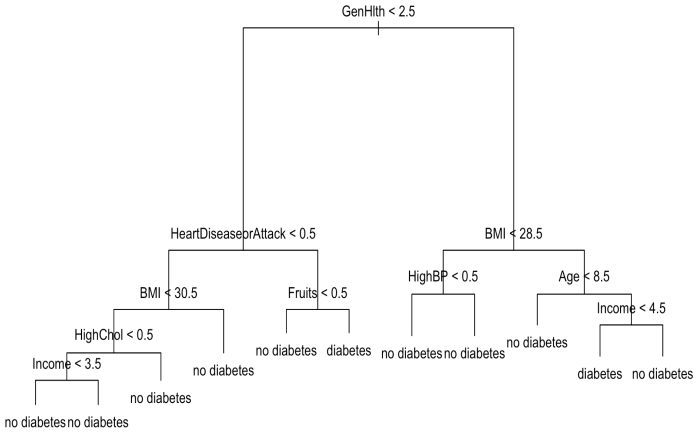

Figure 15. Small training data 1:200

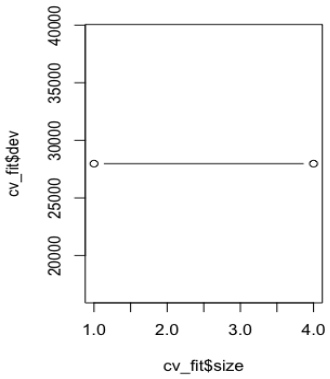

Figure 16. CV-plot

3.3 Random forest

Model explanation: Random Forest (RF) is a popular classifier that uses multiple decision trees on different subsets of data to improve predictive accuracy. Random forest is suitable for our problem as it can handle both categorical and numerical data, including age, BMI, gender, and other relevant variables. It captures complex interactions between different variables, such as the interaction between age, BMI, and gender, which provide insights into how different variables affect diabetes health status. Moreover, it is robust to outliers and noise, reducing the risk of overfitting through the technique called bootstrapping to create multiple decision trees [7]. Therefore, random forest is considered during the analysis as it can generate the importance of each feature and improve the accuracy of predictions compared to using a single decision tree.

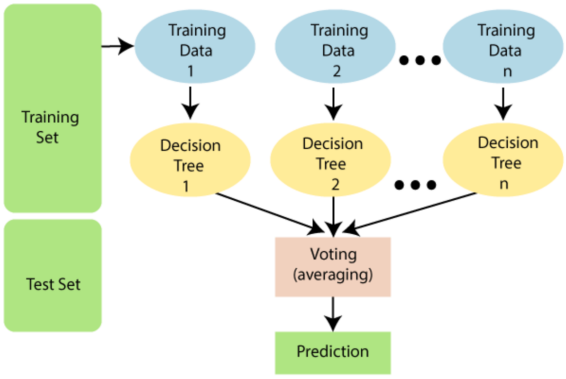

Figure 17. Algorism of RF [6]

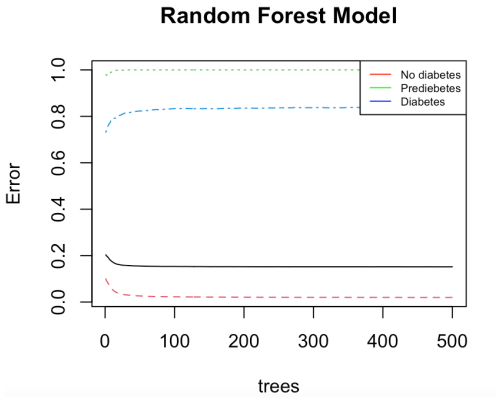

Figure 18. Plot of training RF

Result: Following training of the random forest model with default settings, we generated a confusion matrix and corresponding plot (***Figure 18 and Table 2)***. The model demonstrated strong performance in identifying instances with no diabetes, with a classification error rate of only 0.02. However, it showed poor performance in identifying pre-diabetes, with all instances misclassified and a classification error rate of 1.0. Furthermore, the model demonstrated inadequate performance in identifying diabetes, with a classification error rate of 0.84. The corresponding plot displays the classification error rate versus the number of trees for different diabetes-status groups. The black line in the plot represents the average error rate across all groups.

Table 2. RF Confusion matrix

Prediction \ Reference	No Diabetes	Prediabetes	Diabetes	Class.error
No Diabetes	167593	0	3370	0.02
Pre-diabetes	3411	0	294	1.0
Diabetes	23700	1	4576	0.84

Table 3. RF Confusion matrix of test set

Prediction \ Reference	No Diabetes	Pre-diabetes	Diabetes
No Diabetes	41945	859	5931
Pre-diabetes	3411	0	294
Diabetes	795	67	1138

The trained model on the test dataset in ***table 3*** had an accuracy of 0.85 (***Table 5***) , which indicates the proportion of correct predictions over all predictions. However, it was concluded that the model demonstrated good identification of no diabetes but poor identification of pre-diabetes and diabetes based on the specificity and sensitivity statistics in ***table 4***.

Additionally, the Kappa value of 0.19 in ***table 5*** suggested poor agreement between predicted and actual diabetes status, indicating a need for further tuning for better performance.

Table 4. Sensitivity and specificity of RF

	No Diabetes	Pre-diabetes	Diabetes
Sensitivity	0.98	0.00	0.16
Specificity	0.15	1.00	0.98

Table 5. Accuracy and kappa value of RF

Accuracy	Kappa
0.85	0.19

The variable importance plot in ***figure 19*** for this model showed that BMI was the most important variable, contributing more to reducing the Gini impurity index, indicating better splits in the decision tree. This finding could aid in feature selection and improving model performance by removing irrelevant variables. Conversely, the variable smoker had lower importance, indicating it was less useful in predicting the response variable.

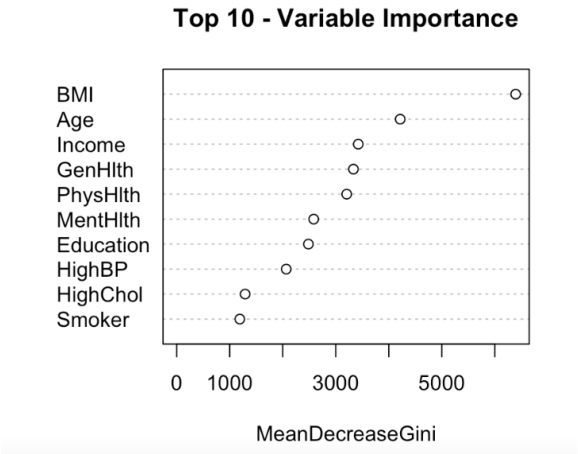

Figure 19. Variable importance plot of RF model

3.4 Logistic regression

Model explanation: Logistic regression is a popular method for binary classification, is utilized in this project to identify individuals with diabetes based on multiple factors. The fitted model reveals each factor's contribution to diabetes likelihood. Examining the model's coefficients allows us to discern the relative importance of these factors, informing targeted prevention strategies.

Preprocessing: Spearman and Pearson correlation coefficients were assessed to confirm the logistic regression model's validity, revealing minimal multicollinearity among binary, ordinal

(spearman) and numeric (pearson) features. The Spearman correlation result in ***figure 20*** and Pearson correlation result in ***figure 21*** shows little correlation between features, the highest correlation value is 0.4 (weak correlation). This allows for the confident inclusion of all features without requiring dropping, combining, or regularization.

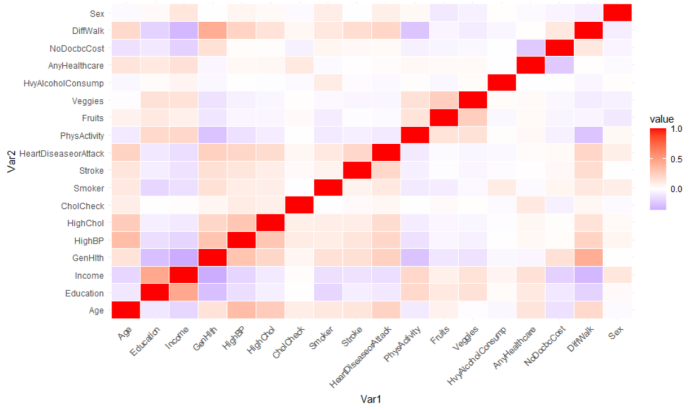

Figure 20. Spearman correlation for ordinal, binary data

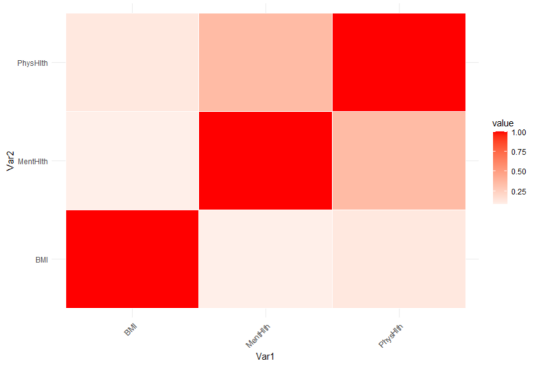

Figure 21. Pearson correlation for numerical values

Result: The performance of the logistic regression model in identifying individuals with diabetes is characterized by an imbalanced sensitivity and specificity, which is also evident in the confusion matrix in ***figure 23***. The model demonstrates a high sensitivity of 0.98, effectively identifying individuals without diabetes. However, its specificity is significantly low at 0.16, indicating a poor ability to correctly detect those with diabetes. This imbalance results in a less meaningful accuracy (0.86) and AUC score (0.83) in ***figure 22***, even they are considered high, while the Kappa statistic of 0.19 highlights the model's limited agreement beyond chance. In summary, the model faces challenges in accurately distinguishing between individuals with and without diabetes, necessitating improvements to achieve better balance and overall performance.

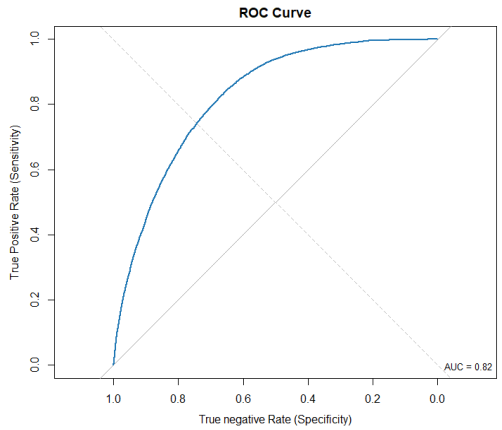

Figure 22. ROC curve of logistic regression

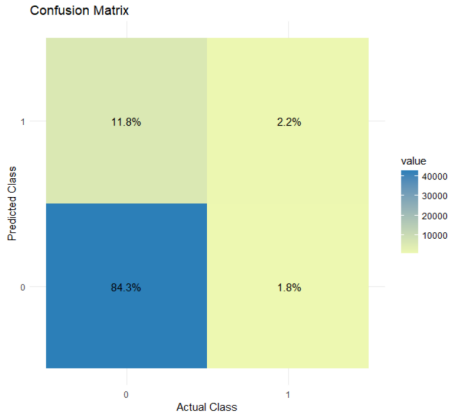

Figure 23. Confusion matrix of logistic regression performance

3.5 K Nearest Neighbor (KNN)

Model explanation: The K-Nearest Neighbor (KNN) model is used for classification. For each data point, we attempt to find a value, k, representing its nearest neighbors. In this model, we identify the k neighbors using Euclidean distance [8]. After finding the k nearest neighbors, we successfully assign responses to the testing data we possess. We then use the predicted data to compare with the actual testing data, determining accuracy by counting how many results we predicted correctly. A KNN model can accommodate various types of data, including categorical data, which constitutes the majority of our dataset. KNN is suitable for this dataset since it is designed for classification tasks. In our dataset, the objective is to predict diabetes types, meaning our response variable is categorical data. Additionally, we have numerous categorical variables, further supporting the suitability of KNN for this dataset.

Preprocessing: The dataset contains three types of responses, and we will first normalize all the data by min-max normalization to scale them between 0 and 1. By using the cross-validation method, I set the range of k from 15 to 50, and obtained the best k with highest accuracy for this dataset is 45.

Result: The accuracy from KNN modeling is about 84%, and kappa for 0.12 with a large portion of correct prediction coming from people without diabetes. There's also no prediction on prediabetes type at all, and this may be caused by the small portion of prediabetes type in original data.

Table 6. Table of Sensitivity and Specificity

	No Diabetes	Pre-diabetes	Diabetes
Sensitivity	0.99	0.00	0.10
Specificity	0.10	1.00	0.99

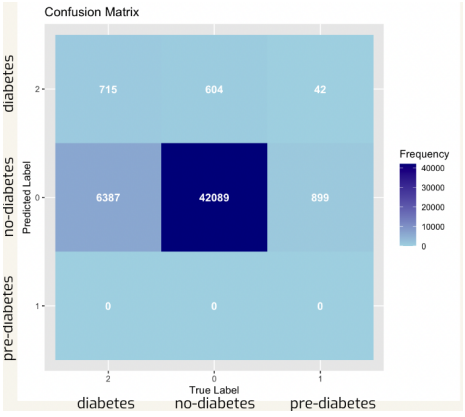

Figure 24. Confusion matrix of KNN model

4.0 Result Analysis and Discussion

Table 7. performance comparison (logistic regression is not included since it has only 2 classes)

Model (3 classes)	**Accuracy**	**Kappa**
Naive Bayes	84%	0.23
Decision Tree	84%	0.07
Random Forest	85%	0.19
KNN	84%	0.12

Result Analysis: From ***table 7***, we can tell the accuracy of these classification methods is similar with an average of approximately 84% to 85%. To obtain a more holistic measure of performance, we utilized Cohen's Kappa. And kappa indicates that Naïve Bayes performed the best, demonstrating a fair agreement. This may be attributed to the imbalanced distribution of classes in the dataset, as Naïve Bayes is known to perform well in such scenarios. Conversely, decision trees showed the worst performance which is 0.07 in Kappa, while other classification methods also demonstrated poor agreement (Kappa < 0.20) in Kappa. This could possibly be because these classification methods overfit on the dataset, leading to poor generalization on unseen data.

Limitations: According to above analysis, we conclude our limitation to dataset bias and model selection bias.

Because our dataset label distribution is extremely unbalanced, so it perform particularly poorly in identifying "pre-diabetes" and "diabetes" which shows in KNN's confusion matrix. This may because of the **dataset label distribution bias**, the number of "pre-diabetes" cases is considerably lower than other labels, while the number of "diabetes" cases is also much lower than "no diabetes".

Additionally, the **model selection biases** should be taken into account. For example, Naïve Bayes may not be an ideal choice for predicting correlated attribute dataset, as diabetes risk factors are interdependent, and this classifier performs poorly on unseen data. Random forest may introduce biases when dealing with imbalanced datasets. Decision trees are sensitive to large numbers of features, and it is hard to find a pattern in our dataset that contains over 20 different features.

Solution: To solve above biases, several potential solutions exist.

(1) Resampling, such as oversampling and undersampling. Oversampling labels with a smaller number of records (e.g. "prediabetes") to increase the representation of minority class in data,

and undersampling labels with majority class (e.g. “no diabetes”) to reduce its dominance in data.
(2) Adding more data points or noises to avoid overfitting on this specific dataset and reduce underfitting for minority classes.
(3) Using pre-train technique to utilize its knowledge from previous training experiences on imbalance dataset to achieve higher performance.
(4) Using feature selection, reduce the features has lower correlation between our label, only select features has higher correlation to train the model
(5) Utilizing ensemble methods like bagging, boosting, and stacking to allow models to find more patterns in data.

Lesson Learned: From this group project, we know that cooperation and time management are crucial, because we must communicate effectively, collaborate, and plan carefully to make sure of our efficiency. Additionally, it has provided us with opportunities to enhance our analytical abilities and recognize our limitations, which helps us to grow and develop our skills further.

5.0 References

[1] International Diabetes Federation. (2021). Resources. Diabetes Atlas. Retrieved April 20, 2023, from https://diabetesatlas.org/resources/?gclid=EAIaIQobChMIiavtuqO3_gIVBQtlCh1L0AgDEAAYASAAEgJkx_D_BwE

[2] Teboul, A. (2021). Diabetes health indicators dataset. Kaggle. Retrieved April 20, 2023, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[3] Zhang Z. (2016). Naïve Bayes classification in R. Annals of translational medicine, 4(12), 241. https://doi.org/10.21037/atm.2016.03.38

[4] *What is Naïve Bayes | IBM*. (n.d.). https://www.ibm.com/topics/naive-bayes#:~:text=The%20Na%C3%AFve%20Bayes%20classifier%20is,a%20given%20class%20or%20category.

[5] Ling, B. (n.d.). Decision Tree. Retrieved April 19, 2023, from https://onq.queensu.ca/d2l/le/content/763470/viewContent/4648485/View

[6] JavaTpoint. (n.d.). Machine Learning Random Forest Algorithm - Javatpoint. Www.javatpoint.com. https://www.javatpoint.com/machine-learning-random-forest-algorithm

[7] Introduction to Random Forest in Machine Learning. (n.d.). Engineering Education (EngEd) Program | Section. Retrieved April 20, 2023, from https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/#:~:text=Advantages%20of%20random%20forest%201%20It%20can%20perform

[8] Ling, B. (n.d.). K-nearest neighbors. Retrieved April 19, 2023, from https://brian-ling.github.io/k-nearest-neighbors.html

Acknowledgement: Our work is polished and grammar is fixed by ChatGPT.