MH3511 Data Analysis with Computer

Group Project

Diabetes Health Indicators Analysis

GitHub Repository
https://github.com/Oganesson0221/Diabetes_Health_Indicators_Analysis/

Name	Contribution	Matriculation Number
Low Jo Yi, Nicole	Summary Statistics, Proportional Testing, Formatting	U2321370D
Tian Yumeng	Summary Statistics, Single Variable Hypothesis Testing	U2340561G
Lu ShanShan	Data description, finding associations, Wilcoxon Rank Sum Test, Ordinal Logistic Regression	U2320618J
Mehta Rishika	Data description, Summary Statistics, research questions, Statistical analysis, Machine Learning	U2323133H
Zhao Qixian	Hypothesis questions, feature engineering, code consolidation	U2321752L

Abstract:

This study analyzes the Diabetes 012 Health Indicators dataset from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) to identify key predictors of diabetes status (no diabetes, prediabetes, diabetes) among U.S. adults. Using exploratory data analysis and statistical modeling, we examine associations between various physiological, lifestyle, and socioeconomic factors and diabetes prevalence.

Table of Contents

1. Introduction

The **"Diabetes 012 Health Indicators"** dataset (Figure 1) available on Kaggle is sourced from the **Behavioral Risk Factor Surveillance System (BRFSS) 2015**, a large-scale health survey conducted in the United States. The CSV titled "diabetes_012_health_indicators_BRFSS2015.csv" comprises **253,680 observations** and **22 variables**, providing a comprehensive overview of various health indicators related to diabetes.

2. Project Objectives and Research Questions

Project Objectives

1. Conduct comprehensive exploratory data analysis to identify statistically significant relationships between health indicators and diabetes status, including correlation analysis, distribution comparisons, and outlier detection.
2. Apply multivariate statistical methods to create prediction models for diabetes risk, evaluating each predictor variable's statistical significance and effect size.
3. Perform comparative statistical analysis across demographic subgroups to identify significant differences in diabetes risk factors and prevalence.

Research Questions

1. What categorical health indicators (e.g., HighBP, HighChol, DiffWalk) are most strongly associated with diabetes status, and how do their associations differ across diabetic, prediabetic, and non-diabetic individuals?
2. Is there a statistically significant association between having high blood pressure and being diagnosed with diabetes?
3. Does the age distribution vary significantly among individuals with different diabetes statuses, indicating age as a contributing factor to diabetes?
4. Is there a significant relationship between the number of physically unhealthy days and diabetes status, suggesting that diabetes affects physical well-being?
5. Does mean Body Mass Index (BMI) significantly differ among individuals with varying diabetes statuses, indicating a link between BMI and diabetes?
6. Is the proportion of non-diabetic individuals among those with high blood pressure significantly different from the commonly cited figure of 55%?
7. Do individuals with diabetes experience a significantly different distribution of mentally unhealthy days compared to non-diabetic individuals, implying a mental health impact of diabetes?

3. Data Description

The target variable, **Diabetes_012**, categorizes individuals into three groups:

- **0** – No diabetes (or only during pregnancy)
- **1** – Prediabetes
- **2** – Diabetes

The dataset includes key health metrics such as **Body Mass Index (BMI)**, **blood pressure levels (HighBP)**, **cholesterol status (HighChol, CholCheck)**, **lifestyle factors (smoking, physical activity, alcohol consumption, fruit/vegetable intake)**, and **pre-existing health conditions (stroke, heart disease, difficulty walking)**. Additionally, it records **demographic information**, including **age, sex, education level, and income bracket**.

1. **Diabetes_012** - Diabetes status, with values 0 (No diabetes), 1 (Prediabetes), or 2 (Diagnosed diabetes), on nominal scale
2. **HighBP –** Blood Pressure, with values 0 (No high blood pressure) or 1 (High blood pressure), nominal scale
3. **HighChol –** Cholesterol, with values 0 (No high cholesterol) or 1 (High cholesterol), on nominal scale
4. **CholCheck** - Cholesterol check in the past five years, with values 0 (No check) or 1 (Check done), on nominal scale
5. **BMI**: Body Mass Index, numeric (double), **min. 12** and **max. 98,** on ratio scale
6. **Smoker**: Smoked at least 100 cigarettes in their lifetime, with values **0** (No) or **1**(Yes), on nominal scale
7. **Stroke**: Ever had a stroke, with values **0** (No) or **1** (Yes), on nominal scale
8. **HeartDiseaseorAttack**: History of coronary heart disease or myocardial infarction, with values **0** (No) or **1** (Yes), on nominal scale
9. **PhysActivity**: Conduct physical activity (excluding job-related activity) in the past 30 days, with values **0** (No) or **1** (Yes), on nominal scale
10. **Fruits**: Consumes fruit at least once per day, with values **0** (No) and **1** (Yes), on nominal scale
11. **Veggies**: Consumed vegetables once or more per day, with values **0** (No) or **1**(Yes), which is on a nominal scale
12. **HvyAlcoholConsump**: Participant is a heavy drinker (having more than 14 drinks per week for adult men and having more than 7 drinks per week for adult women), with values **0** (No) or **1**(Yes), which is on a nominal scale
13. **AnyHealtcare**: Participant has health care coverage such as health insurance, with values **0** (have) or **1**(do not have), which is on a nominal scale

14. **NoDocbcCost**: participant could not visit a doctor due to cost in the past 12 months, with values **0** (No) or **1**(Yes), which is on a nominal scale
15. **GenHlth**: Participant's rating of their general health, with values **1**(excellent), **2**(very good), **3**(good), **4**(fair), or **5**(poor), which is on an ordinal scale
16. **MentHlth**: Number of days when participants are in bad mental health (due to stress, depression, problems with emotions, etc) in the past 30 days, takes a value **from 0 to 30**, which is on a ratio scale
17. **PhysHlth**: Number of days when participants are in bad physical health (due to physical illness and injury etc) in the past 30 days, takes a value **from 0 to 30**, which is on a ratio scale
18. **DiffWalk**: Participants have serious difficulty walking or climbing stairs, with values **0** (No) or **1**(Yes), which is on a nominal scale
19. **Sex**: **1** (Male), **0** (Female) which is on a nominal scale
20. **Age**: Participant's age category, which is on an ordinal scale, takes one of the following values:

Value	Age
1	18-24
2	25-29
3	30-34
4	35-39
5	40-44
6	45-49
7	50-54
8	55-59
9	60-64
10	65-69
11	70-74
12	75-79
13	80 or above

21. **Education**: Education level, which is on an ordinal scale, takes one of the following values:

Value	Education
1	Never attended school or only attended kindergarten
2	Grades 1 through 8
3	Grades 9 through 11
4	Grade 12 or GED
5	College 1 year to 3 years
6	College 4 years or more

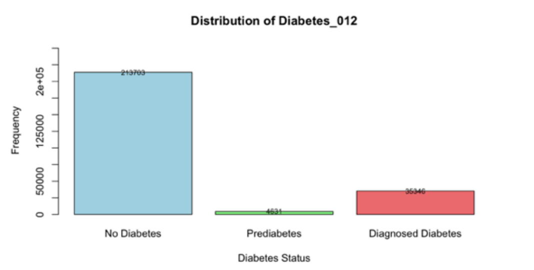
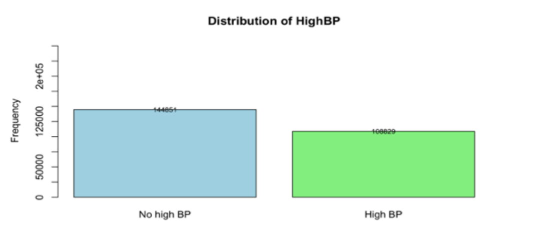
22. **Income**: Income level, which is on an ordinal scale, takes one of the following values:

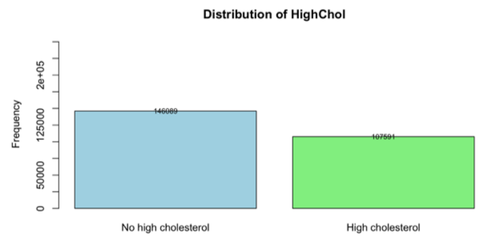
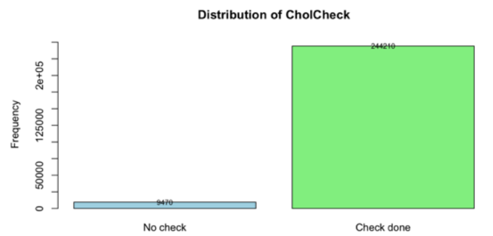
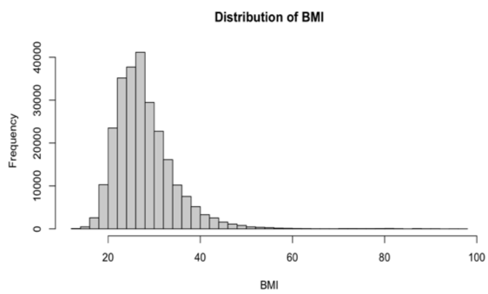
Value	Income
1	Less than $10,000
2	$10,000 to less than $15,000
3	$15,000 to less than $20,000
4	$20,000 to less than $25,000
5	$25,000 to less than $35,000
6	$35,000 to less than $50,000
7	$50,000 to less than $75,000
8	$80000 or more

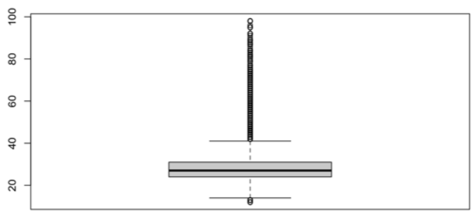
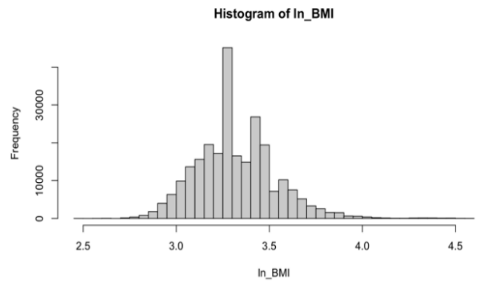
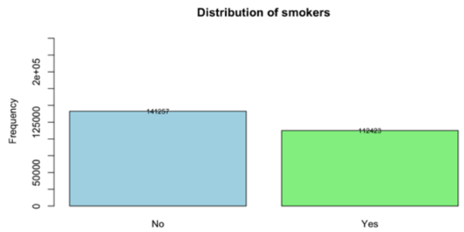
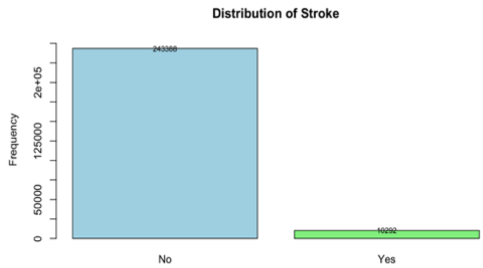
4. Data Analysis

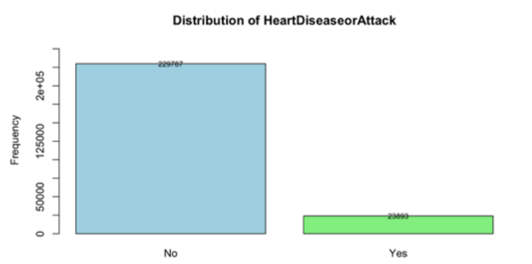
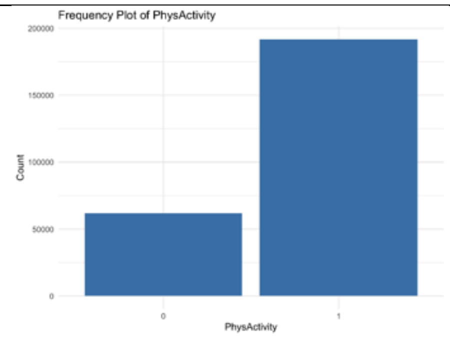
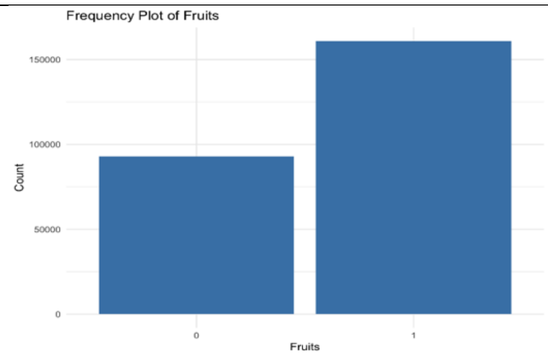
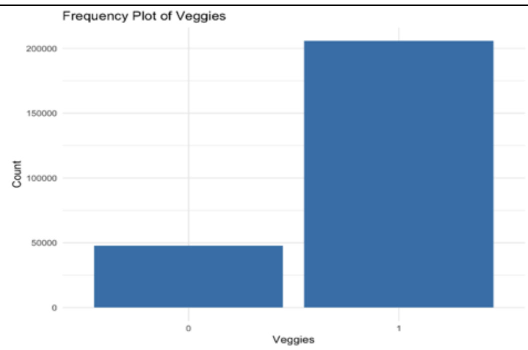
In this section, we shall look into the data in more detail. Each variable is investigated individually to look for possible outliers and/or to perform a transformation to avoid highly skewed data. We will also be investigating the general relationship between different variables and Diabetes_012, our main variable of interest.

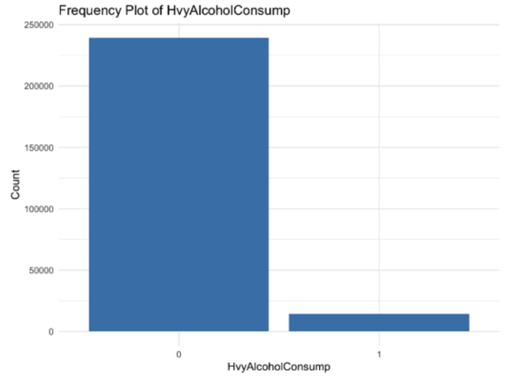
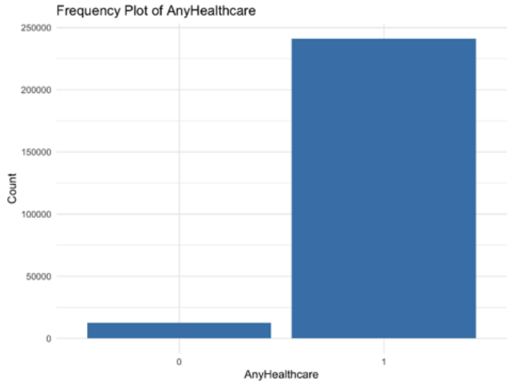
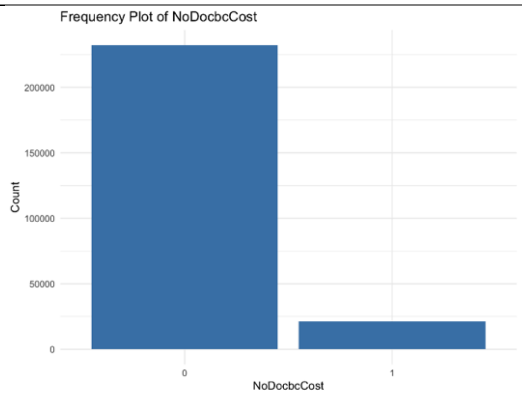
4.1 Summary Statistics

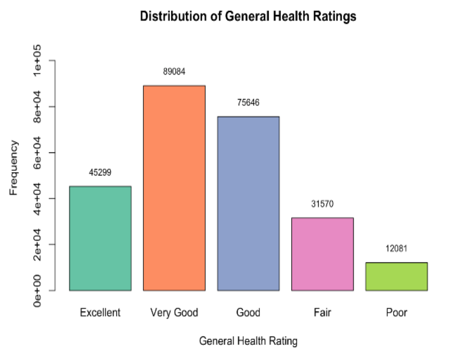
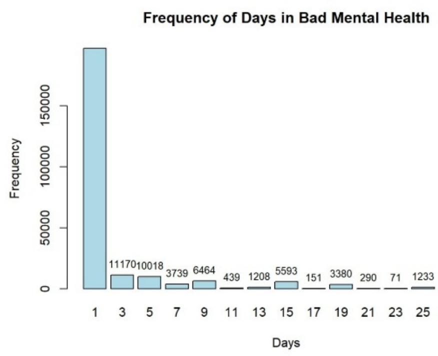
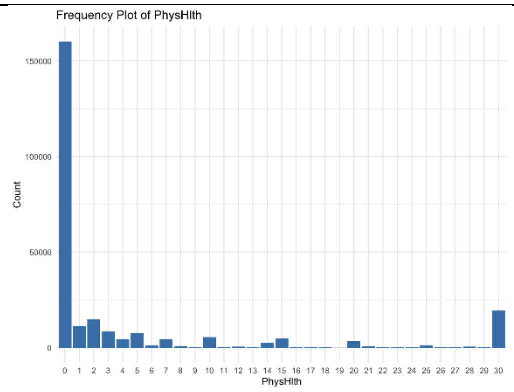
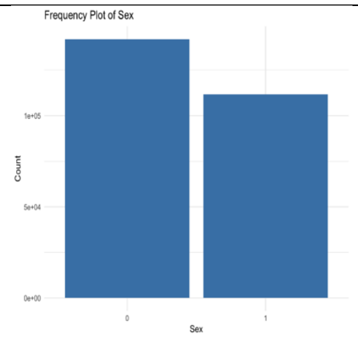
Diabetes_012		*Diabetes_012* consists of 3 categories: 0 for “No diabetes” (213,703), 1 for “Prediabetes” (4,631) and 2 for “Diagnosed Diabetes”(35,346).
HighBP		*HighBP is a binary factor which* consists of 2 categories: 0 for “No high blood pressure” (144,851) and 1 for “High blood pressure”(108,829). Based on the chart, there are more people who do not have high blood pressure than those who do.

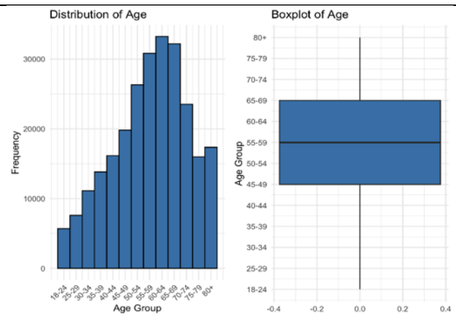
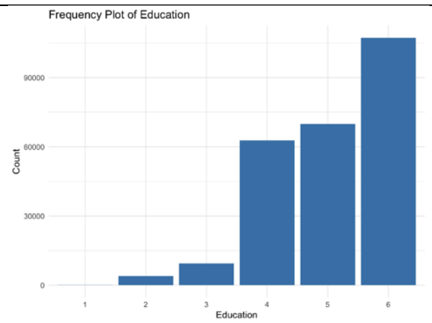
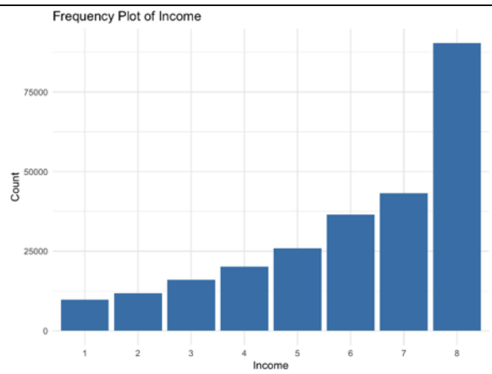
HighChol		*HighChol* consists of 2 categories: 0 for "No high cholesterol" (146,089) and 1 for "High cholesterol" (107,591). Based on the chart, there are more people who do not have high cholesterol as compared to those who do.
CholCheck		*CholCheck* consists of 2 categories: 0 for "No check" (9,470) and 1 for "Check" (244,210). Based on the chart, there are more people who have done their cholesterol check in the past 5 years than those who have not.
BMI		Based on the histogram, the distribution of *BMI* seems to be right skewed with a longer right tail. This means that there is a concentration of lower values, while a few extreme values are spread out to the right, which may be considered as outliers. It is a numeric variable with a mean of 28.38, ranging from 12 to 98, and a median of 27.

<table>
<tr><td></td><td>
</td><td>As seen from the boxplot of BMI, there are many outliers on the right side of the plot and the right whisker is longer than the left one, which explains why there is a longer right tail on the distribution of BMI.</td></tr>
<tr><td></td><td>
</td><td>After applying log transformation on the histogram of BMI, the distribution of ln_BMI now appears to be more evenly distributed with a bell-shaped curve and symmetrical distribution.</td></tr>
<tr><td>Smoker</td><td>
</td><td>It is a binary factor with 141,257 non-smokers (0) and 112,423 smokers (1). Based on the chart, there are more people who have never smoked more than 100 cigarettes in their life than those who have.</td></tr>
<tr><td>Stroke</td><td>
</td><td>It is a binary factor where 243,388 indicate no stroke history (0) and 10,292 a history of stroke (1). Based on the chart, there are more people who have never had a stroke before than those who have.</td></tr>
</table>

HeartDiseaseorAttack		It is a binary factor with 229,787 indicating no heart disease or attack (0) and 23,893 a history of heart disease or attack (1). Based on the chart, there are more people who do not have any history of coronary heart disease or myocardial infarction than those who do.
PhysActivity		Based on the chart, more individuals engage in physical activity (191,920) than no physical activity (61,760).
Fruits		It is a binary factor where 92,782 indicate no fruit consumption (0) and 160,898 indicate fruit consumption (1). Based on the chart, more consumption of fruit is observed.
Veggies		A binary factor with 47,839 indicating no vegetable consumption (0) and 205,841 vegetable consumption (1). Higher vegetables consumption is observed among people.

HvyAlcoholConsump		It is a binary factor where 239,424 indicate no heavy alcohol consumption (0) and 14,256 heavy alcohol consumption (1). Based on the chart, a greater population has lower alcohol consumption.
AnyHealthcare		It is a binary factor with 12,417 indicating no healthcare access (0) and 241,263 access to healthcare (1). The large number of individuals are unable to afford healthcare underscores the economic strain.
NoDocbcCost		It is a binary factor where 232,326 observations indicate no cost barrier to seeing a doctor (0) and 21,354 indicate a cost barrier (1). Based on the chart a majority of people could afford to visit the doctor in the past 12 months.

GenHlth		The majority of respondents reported "Very Good" (89,084) and "Good" (75,646) health, while the smallest group reported "Poor" health (12,081). The general health ratings distribution shows that most participants rate their health as "Very Good" or "Good," while fewer report "Fair" or "Poor" health.
MenHlth		Most participants report 0 bad mental health days, but a small subset experiences frequent poor mental health.
PhysHlth		A numeric variable representing physical health days, with a mean of 4.242 and a median of 0, suggesting most respondents reported 0 days of poor physical health.
Sex		A binary factor where 141,974 observations are female (0) and 111,706 are male (1). Based on the chart, more females participated in the survey.

Age		A numeric variable with a mean of 8.032 (likely representing age groups), ranging from 1 to 13, and a median of 8.
Education		A categorical factor with 6 levels, where the majority of respondents have higher education levels (107,325 in level 6).
Income		A categorical factor with 8 levels, where the majority of respondents fall in the higher income brackets (90,385 in level 8).

4.2 Associations between variables and Diabetes_012

4.2.1 Categorical Variables (Nominal Scale)

Cramer's V test is used to find associations between Diabetes_012 and other nominal data. A higher Cramer's V value indicates the variable has a stronger association with Diabetes_012. From the table below, the top 3 variables that have a stronger association with Diabetes_012 are HighBP, DiffWalk and HighChol.

	Category	CramersV_Value
1	HighBP	0.27219111
2	HighChol	0.21067124
3	CholCheck	0.06802124
4	Smoker	0.06311427
5	Stroke	0.10722761
6	HeartDiseaseorAttack	0.18028072
7	PhysActivity	0.12221836
8	Fruits	0.04232050
9	Veggies	0.05935909
10	HvyAlcoholConsump	0.05789607
11	AnyHealthcare	0.01650162
12	NoDocbcCost	0.03951385
13	DiffWalk	0.22442454
14	Sex	0.03144593

Figure 1. Result of Cramer's V Test

4.2.2 Categorical Variables (Ordinal Scale)

Kruskal-Wallis Test is used to find the association between ordinal variables and Diabetes_012. If the p-value of the test is less than 0.05, we will reject the null hypothesis, which is the distribution of the ordinal variable is the same across all 3 groups of Diabetes_012 (no diabetes, prediabetic, diabetic). From the table below, the p-value of the Kruskal-Wallis Test is less than 0.05 for all 4 ordinal variables shown in the table. Hence, we will reject the null hypothesis for all 4 nominal variables and conclude for each ordinal variable, their distribution are not the same across the 3 groups. Hence, there may be associations between the 4 ordinal variables and Diabetes_012.

OrdinalVar	Kruskal_statistic	PValue
GenHlth	22480.925	0
Age	8811.763	0
Education	4083.037	0
Income	7558.899	0

Figure 2. Result of Kruskal-Wallis Test for categorical variables

4.2.3 Numerical Variables (Ratio scale)

Since MentHlth and PhysHlth do not follow a normal distribution, we will use the Kruskal-Wallis Test to find their association with Diabetes_012 by comparing their median value across the 3 groups of Diabetes_012 (no diabetes, prediabetic and diabetic). The p-value for the test for both variables is less than 0.05, hence, we can reject the null hypothesis and conclude that there is not enough evidence to say the median values for MentHlth or PhysHlth are the same across the 3 groups of Diabetes_012. Hence, there may be some association between MentHlth and Diabetes_012 and PhysHlth and Diabetes_012.

	NumVar	K_statistic	P_value
Kruskal-Wallis chi-squared	MentHlth	528.9106	1.407759e-115
Kruskal-Wallis chi-squared1	PhysHlth	6661.8780	0.000000e+00

Figure 3. Result of Kruskal-Wallis Test for numerical variables

Since logBMI follows a normal distribution, as seen from the qqplot below. We can use one-way ANOVA to check if logBMI is associated with Diabetes_012 by comparing the mean value of logBMI across the 3 groups of Diabetes_012.

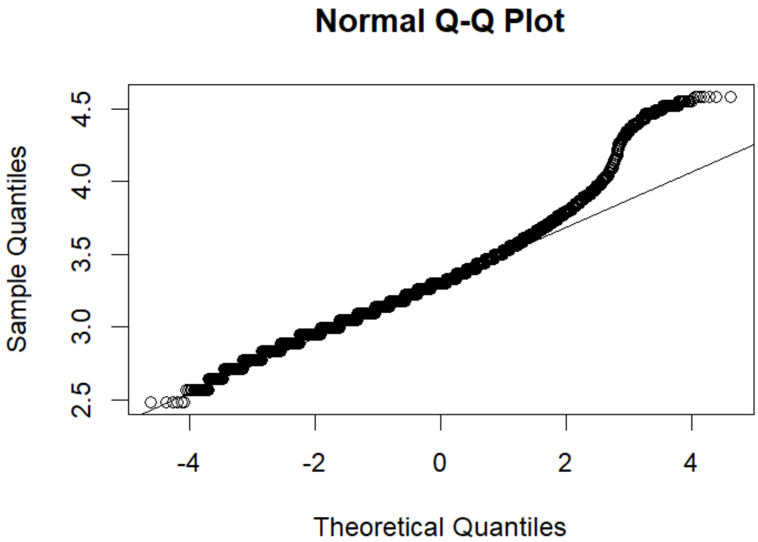


Figure 4. qqplot for logBMI

Since the p-value is less than 0.05, we can reject null hypothesis and conclude that the means of logBMI differs significantly between the 3 groups of Diabetes_012. Hence, there is strong association between logBMI and Diabetes_012.

```
                 Df Sum Sq Mean Sq F value Pr(>F)    
Diabetes_012      2    613  306.69    7394 <2e-16 ***
Residuals    253677  10522    0.04                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5. Results of one-way ANOVA test of logBMI vs Diabetes_012

In conclusion, some variables that show significant association with Diabetes_012 and are worth exploring further are HighBP, HighChol, DiffWalk, GenHlth, Age, Education, Income, logMentHlth, logPhysHlth and logBMI.

5. Statistical Analysis

5.1 Hypothesis testing (Single Variable)

5.1.1 Chi-Square Test

Do diabetic patients have high blood pressure?

- H0: There is no significant association between high blood pressure status and diabetes status.
- H1: There is a significant association between high blood pressure status and diabetes status.

```
	Pearson's Chi-squared test

data:  table_highbp
X-squared = 18795, df = 2, p-value < 2.2e-16
```

Figure 6. Results of Chi-square Test of HighBP vs Diabetes_012

We conducted a Chi-Square Test of Independence to investigate the relationship between high blood pressure status and diabetes status. The test produced a Chi-square statistic of 18,795 with 2 degrees of freedom, and a p-value less than 2.2e-16. This extremely small p-value indicates strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that there is a statistically significant association between diabetes status and high blood pressure. In other words, individuals with diabetes are more likely to have high blood pressure compared to those without diabetes.

5.1.2 Kruskal-Wallis Rank Sum Test

Is age a factor for getting diabetes?

- H0: The distribution of diabetes status is the same across all age categories.
- H1: The distribution of diabetes status differs significantly across age categories.

```
	Kruskal-Wallis rank sum test

data:  Age by as.factor(Diabetes_012)
Kruskal-Wallis chi-squared = 8811.8, df = 2, p-value < 2.2e-16
```

Figure 7. Results of Kruskal-Wallis Rank Sum Test of Age vs Diabetes_012

We conducted a Kruskal-Wallis's test to assess whether age influences diabetes status. The test showed a significant difference in age distribution across the three diabetes groups ($\chi^2 = 8811.8$, df = 2, $p < 2.2e-16$). As the p-value is below 0.05, we reject the null hypothesis, concluding that the evidence suggests that age is significantly associated with diabetes status—older individuals are more likely to have diabetes. Although the proportion of diabetes cases does not continue to rise in the 80+ age group according to the bar chart, the overall trend indicates that the likelihood of having diabetes increases with age. This slight deviation may be due to a smaller sample size in the 80+ group. Therefore, we still conclude that older individuals are generally more likely to have diabetes.

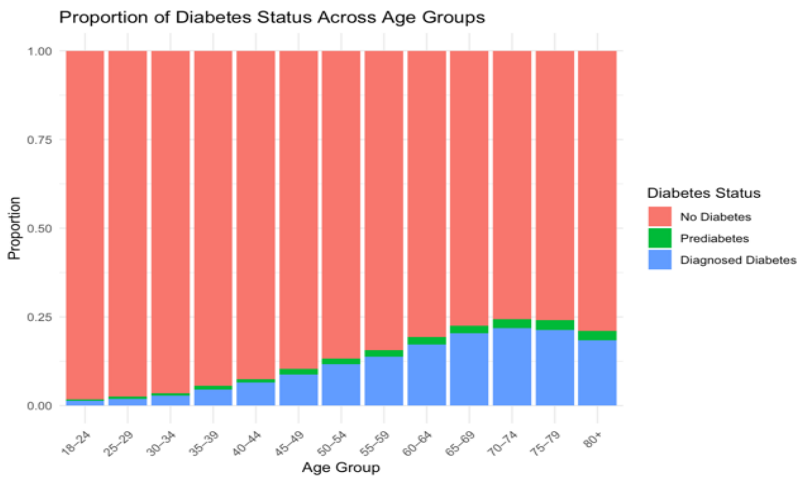


Figure 8. Bar Chart of Age vs Diabetes_012

Do people with diabetes have poorer physical health?

- H0: There is no association between physically unhealthy days (PhysHlth) and diabetes status (Diabetes_012).
- H1: There is an association between physically unhealthy days (PhysHlth) and diabetes status (Diabetes_012).

```
> kruskal.test(PhysHlth ~ factor(Diabetes_012), data = x)

        Kruskal-Wallis rank sum test

data:  PhysHlth by factor(Diabetes_012)
Kruskal-Wallis chi-squared = 6661.9, df = 2, p-value < 2.2e-16
```

Figure 9. Results of Kruskal-Wallis Rank Sum Test of PhysHlth vs Diabetes_012

We first perform the Kruskal-Wallis's test to examine the relationship between the number of physically unhealthy days and diabetes status. The test result showed a statistically significant difference in the number of physically unhealthy days across the three diabetes status groups (**χ^2 = 6661.9, df = 2, p < 2.2e-16**), **suggesting an association** between physical health and diabetes status. Therefore, we **reject the null hypothesis** and conclude that there is not enough evidence to say that there is no relationship between physical health and diabetes status. The association may be due to the impact of diabetes on individuals' physical health, potentially leading to reduced activity levels.

We then grouped the number of physically unhealthy days into five categories and calculated the proportion of each diabetes status within those groups to make a stacked bar chart to show how the proportion of diabetes, prediabetes, and no diabetes changes as the number of unhealthy days increases.

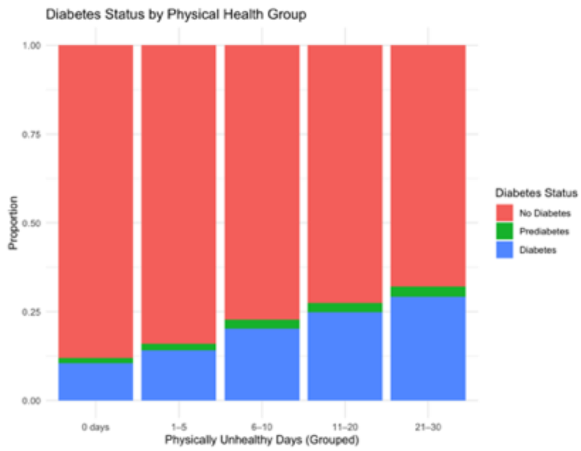


Figure 10. Bar Chart of PhysHlth vs Diabetes_012

5.1.3 One-Way ANOVA Test

Does diabetes affect one's BMI?

- H0: There is no significant difference in mean BMI across the three diabetes status groups.
- H1: At least one diabetes status group has a significantly different mean BMI compared to the others.

```
                            Df Sum Sq Mean Sq F value Pr(>F)    
as.factor(Diabetes_012)      2    613  306.69    7394 <2e-16 ***
Residuals               253677  10522    0.04                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> EtaSq(fit)
                            eta.sq eta.sq.part
as.factor(Diabetes_012) 0.05508495  0.05508495
```

Figure 11. Results of One-way ANOVA Test of logBMI vs Diabetes_012

We conducted a one-way ANOVA test to examine if logBMI differs by diabetes status. The results showed a significant effect, with F statistics = 7394**,** p < 2e-16, providing strong evidence to reject the null hypothesis. The eta squared value was 0.0551, which suggests that approximately 5.51% of the variance in logBMI can be explained by diabetes status, indicating a small to moderate effect size. Thus, diabetes status has a statistically significant but modest impact on logBMI.

5.1.4 Proportional Test

What proportion of individuals with high blood pressure do not have diabetes?
It is generally said that around 55% of people who have high blood pressure are non-diabetic (De Feo et al., 2021) . Therefore, to test whether the observed proportion matches the expected 55%, we performed a proportional test with confidence level 95% (i.e alpha = 0.05).

- H0: p = 0.55
- H1: p $\neq$ 0.55

In this case, x is the number of people who have high blood pressure but no diabetes while n is the number of people who do not have diabetes. After performing proportion test, we find out that the p-value is 2.2e-16. Since the p-value is smaller than value of alpha, we reject the null hypothesis and conclude that the true proportion of non-diabetic individuals among high blood pressure patients is not 55%.

```
> prop.test(x,n, p0, conf.level = 0.95)

        1-sample proportions test with continuity correction

data:  x out of n, null probability p0
X-squared = 27624, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.55
95 percent confidence interval:
 0.3690836 0.3731848
sample estimates:
        p
0.3711319
```

Figure 12. Results of Proportional Test of HighBP vs Diabetes_012

5.1.5 Wilcoxon Rank Sum Test

Do people with diabetes have the same mental health compared to people without diabetes?

- H0: The distribution of the number of days with bad mental health is the same between the two groups
- H1: The distribution of the number of days with bad mental health is different
- Between the two groups

Since our question focuses only on people with or without diabetes, we will first remove prediabetic individuals. From our visualization using a bar plot, the distribution for the number of days in bad mental health does not follow any specific distribution, and the mean value is very close to zero. Hence, we decided to use the Wilcoxon Rank Sum Test to answer our question.

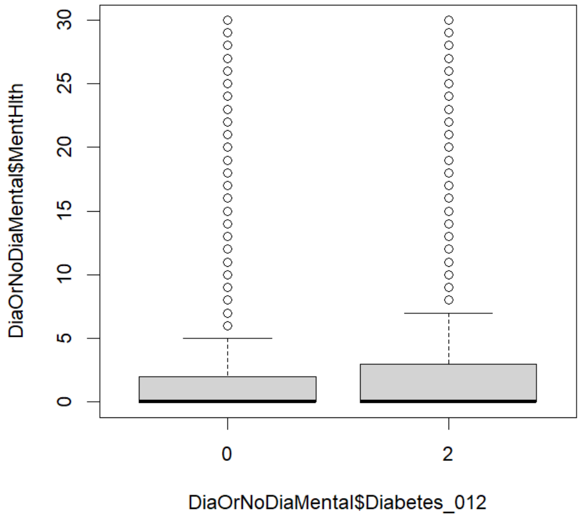


Figure 13. Bar plot of MentHlth vs Diabetes_012

Since the p-value is less than 0.05, we can reject the null hypothesis and conclude that the distribution for the number of days with bad mental health between diabetic and non-diabetic individuals is not the same. This means that diabetic individuals do not have the same mental health as non-diabetic individuals. This may be due to diabetic individuals facing more stressful days as their daily life may be inconvenienced due to them having diabetes; for example, they must take medication constantly or they have to go for checkups regularly.

```
> result <- wilcox.test(DiaOrNoDiaMental$MentHlth~DiaOrNoDiaMental$Diabetes_012,
+                        data = DiaOrNoDiaMental, exact = FALSE)
> result

        Wilcoxon rank sum test with continuity correction

data:  DiaOrNoDiaMental$MentHlth by DiaOrNoDiaMental$Diabetes_012
W = 3564682390, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Figure 14. Results of Wilcoxon Rank Sum Test of MentHlth vs Diabetes_012

6. Appendix

6.1 Hypothesis testing (Multi Variable)

6.1.1 Ordinal Logistic Regression (Proportional Odds Model)

Do physiological factors have a combined effect on diabetes?

- H0: The interactions between the physiological risks (HighBP, HighChol and BMI) have no significant impact on diabetes severity.
- H1: The interactions between the physiological risks (HighBP, HighChol and BMI) have a significant impact on diabetes severity.

Since our target variable, Diabetes_012, has an inherent order (0 = No diabetes, 1 = Prediabetes, 2 = Diabetes), the Ordinal Logistic Regression will account for the ordinal nature. We will be using three variables in building the model, HighBP, HighChol and BMI. BMI will be split into 4 levels (1 – underweight, 2 – healthy, 3 – overweight and 4 – obese). We will then convert each level as a factor.

```
> #convert BMI into categorical
> diabetesData$BMIlevel <- cut(diabetesData$BMI,
+                              breaks = c(-Inf, 18.5, 24.9, 29.9, Inf),
+                              labels = c(1, 2, 3, 4))
> # Convert to factor for categorical analysis
> diabetesData$BMIlevel <- as.factor(diabetesData$BMIlevel)
> physioFactors<-data.frame(bmiLevel = diabetesData$BMIlevel, bpLevel = as.factor(diabetesData$HighBP),
+                           cholLevel = as.factor(diabetesData$HighChol), diabeteStatus = as.factor(diabetesData$Diabetes_012))
> str(physioFactors)
'data.frame':   253680 obs. of  4 variables:
 $ bmiLevel     : Factor w/ 4 levels "1","2","3","4": 4 3 3 3 2 3 4 3 4 2 ...
 $ bpLevel      : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 1 ...
 $ cholLevel    : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 1 2 2 1 ...
 $ diabeteStatus: Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 3 1 ...
```

Appendix 1. Splitting of BMI into levels and selecting physiological variables

Then, we will start building the model using the polr() function in R and check if using the model is appropriate by using the brant(). The omnibus probability is less than 0.05, indicating that the effect of high Bp/cholesterol on diabetes severity is not consistent across all outcome levels, meaning that the impact might differ between ‘no diabetes’, ‘prediabetes’ and ‘diabetes’. Hence, using Ordinal Logistic Regression result may be biased for the two variables.

```
> model <- polr(factor(diabeteStatus, ordered = TRUE) ~ bpLevel + cholLevel + bmiLevel,
+               data = physioFactors, Hess = TRUE)
> #check if its appropriate to use the model
> brant(model)
--------------------------------------------
Test for        X2      df      probability
--------------------------------------------
Omnibus         85.44   5       0
bpLevel1        79.04   1       0
cholLevel1      5.43    1       0.02
bmiLevel2       0.07    1       0.79
bmiLevel3       0.02    1       0.89
bmiLevel4       0       1       0.94
--------------------------------------------
```

Appendix 2. Building and checking of Ordinal Logistic Regression Model

```
Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 219006.5 on 507341 degrees of freedom

Log-likelihood: -109503.2 on 507341 degrees of freedom

Number of Fisher scoring iterations: 6

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1'

Exponentiated coefficients:
                  bpLevel1:1                  bpLevel1:2                cholLevel1:1                cholLevel1:2
                   0.3065137                   0.2910582                   0.3098362                   0.3124963
                   bmiLevel2                   bmiLevel3                   bmiLevel4          bpLevel1:cholLevel1
                   0.9820018                   0.5459730                   0.2414677                   1.2359628
         bpLevel1:bmiLevel2          bpLevel1:bmiLevel3          bpLevel1:bmiLevel4         cholLevel1:bmiLevel2
                   0.7836831                   0.8534446                   1.0148347                   1.1765157
       cholLevel1:bmiLevel3        cholLevel1:bmiLevel4 bpLevel1:cholLevel1:bmiLevel2 bpLevel1:cholLevel1:bmiLevel3
                   1.2468377                   1.4015292                   1.0477740                   1.0540522
bpLevel1:cholLevel1:bmiLevel4
                   0.8522229
```

Appendix 3. Logistic Model Summary and Coefficients

Since the Omnibus probability is less than 0.05, it is inappropriate to use the model, so we switched to the Partial Proportional Odds Model to allow some variables (bpLevel and cholLevel) to violate some assumptions while others stay constant.

```
> model_ppo <- vglm(diabetesStatus ~ bpLevel * cholLevel * bmiLevel,
+                   family = cumulative(parallel = FALSE ~ bpLevel + cholLevel),
+                   data = physioFactors)

> summary(model_ppo)

Call:
vglm(formula = diabetesStatus ~ bpLevel * cholLevel * bmiLevel,
    family = cumulative(parallel = FALSE ~ bpLevel + cholLevel),
    data = physioFactors)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept):1                  3.65452    0.15116  24.177  < 2e-16 ***
(Intercept):2                  3.84924    0.15125  25.450  < 2e-16 ***
bpLevel1:1                    -1.18249    0.23358  -5.062 4.14e-07 ***
bpLevel1:2                    -1.23423    0.23364  -5.283 1.27e-07 ***
cholLevel1:1                  -1.17171    0.23388  -5.010 5.44e-07 ***
cholLevel1:2                  -1.16316    0.23392  -4.972 6.61e-07 ***
bmiLevel2                     -0.01816    0.15461  -0.117    0.906
bmiLevel3                     -0.60519    0.15321  -3.950 7.81e-05 ***
bmiLevel4                     -1.42102    0.15265  -9.309  < 2e-16 ***
bpLevel1:cholLevel1            0.21185    0.31752   0.667    0.505
bpLevel1:bmiLevel2            -0.24375    0.23853  -1.022    0.307
bpLevel1:bmiLevel3            -0.15847    0.23597  -0.672    0.502
bpLevel1:bmiLevel4             0.01473    0.23511   0.063    0.950
cholLevel1:bmiLevel2           0.16256    0.23894   0.680    0.496
cholLevel1:bmiLevel3           0.22061    0.23644   0.933    0.351
cholLevel1:bmiLevel4           0.33756    0.23587   1.431    0.152
bpLevel1:cholLevel1:bmiLevel2  0.04667    0.32421   0.144    0.886
bpLevel1:cholLevel1:bmiLevel3  0.05264    0.32053   0.164    0.870
bpLevel1:cholLevel1:bmiLevel4 -0.15991    0.31961  -0.500    0.617
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 4. Partial Proportional Odds Model

Then we use anova() to check if the combined interactions among the 3 variables affect diabetes severity.

```
> anova(model_ppo, type='III')
Analysis of Deviance Table (Type III tests: each term added last)

Model: 'cumulative', 'VGAMordinal', 'VGAMcategorical'

Links: 'logitlink'

Response: diabetesStatus

                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
bpLevel                    2   116.38    507343     219123 < 2.2e-16 ***
cholLevel                  2    25.90    507343     219032 2.376e-06 ***
bmiLevel                   3  1538.08    507344     220545 < 2.2e-16 ***
bpLevel:cholLevel          1     0.44    507342     219007 0.5066527    
bpLevel:bmiLevel           3    29.39    507344     219036 1.854e-06 ***
cholLevel:bmiLevel         3    12.87    507344     219019 0.0049169 ** 
bpLevel:cholLevel:bmiLevel  3   16.70    507344     219023 0.0008131 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 5. ANOVA Results for Interaction Effects

Since the p-value is less than 0.05, we reject the null hypothesis and conclude that interactions between HighBP, HighChol, and BMI have a significant impact on diabetes severity.

6.1.2 Feature Engineering: Nested Multivariable Model Comparison

In this section, we are going to find out how the Socioeconomic Status (SES) is affecting diabetes. From our feature selection 4.2.2, we have identified two core features: Income and Education, with high Kruskal Statistics of 7558 and 4083, respectively, and a p-value of 0. In this section, we are investigating their combined effects in predicting diabetes by building up a core model using ordinal logistic regression that contains Income and Education.

```
> print(summary(model_core))
Call:
polr(formula = Diabetes_012 ~ Income + Education, data = data, 
    Hess = TRUE)

Coefficients:
                Value Std. Error  t value
Income.L    -0.994840    0.01911 -52.0506
Income.Q    -0.290902    0.01751 -16.6097
Income.C    -0.022437    0.01714  -1.3092
Income^4    -0.134144    0.01714  -7.8243
Income^5     0.027141    0.01689   1.6069
Income^6    -0.060633    0.01652  -3.6713
Income^7     0.006494    0.01631   0.3981
Education.L -0.587868    0.10268  -5.7254
Education.Q -0.146538    0.09348  -1.5675
Education.C  0.105492    0.06634   1.5902
Education^4 -0.118006    0.03888  -3.0351
Education^5 -0.016676    0.02205  -0.7561

Intercepts:
                               Value   Std. Error t value
No Diabetes|Prediabetes         1.2474  0.0292    42.6928
Prediabetes|Diagnosed Diabetes  1.3969  0.0293    47.7344

Residual Deviance: 241196.47 
AIC: 241224.47 
```

Appendix 6. Core model metrics summary

The model output displays several polynomial contrasts. For instance, the linear contrast for *Income* (Income.L) has a coefficient of -0.994840 with a standard error of 0.01911, resulting in a t-value of -52.05. Similarly, for *Education*, the linear contrast (Education.L) is -0.587868 (SE = 0.10268, t = -5.73). These statistics indicate that **higher income and education levels are strongly associated with lower odds of being in a**

higher diabetes category. The overall model fit is summarized by a residual deviance of 241196.47 and an AIC of 241224.47, which serve as benchmarks for model comparison.

We further explore two more predictors, AnyHealthcare and NoDocbcCost—which are intended to capture aspects of healthcare access and SES, but perform poorly in 4.2.1, with CramersV_Value of only 0.0165 and 0.0395 respectively. We want to investigate whether by extending to these two predictors in addition to core model, we can achieve a better extended model.

```
> print(summary(model_extended))
Call:
polr(formula = Diabetes_012 ~ Income + Education + AnyHealthcare +
    NoDocbcCost, data = data, Hess = TRUE)

Coefficients:
                  Value Std. Error  t value
Income.L      -1.021281    0.01944 -52.5276
Income.Q      -0.294605    0.01756 -16.7784
Income.C      -0.025494    0.01718  -1.4844
Income^4      -0.124833    0.01718  -7.2658
Income^5       0.019853    0.01692   1.1731
Income^6      -0.056592    0.01654  -3.4206
Income^7       0.007449    0.01634   0.4558
Education.L   -0.648884    0.10315  -6.2904
Education.Q   -0.126040    0.09389  -1.3425
Education.C    0.101609    0.06662   1.5252
Education^4   -0.117936    0.03904  -3.0208
Education^5   -0.019541    0.02213  -0.8828
AnyHealthcare  0.605883    0.02798  21.6546
NoDocbcCost    0.103823    0.01893   5.4832

Intercepts:
                              Value   Std. Error t value
No Diabetes|Prediabetes         1.8090  0.0394    45.9257
Prediabetes|Diagnosed Diabetes  1.9589  0.0394    49.6745

Residual Deviance: 240670.86
AIC: 240702.86
```

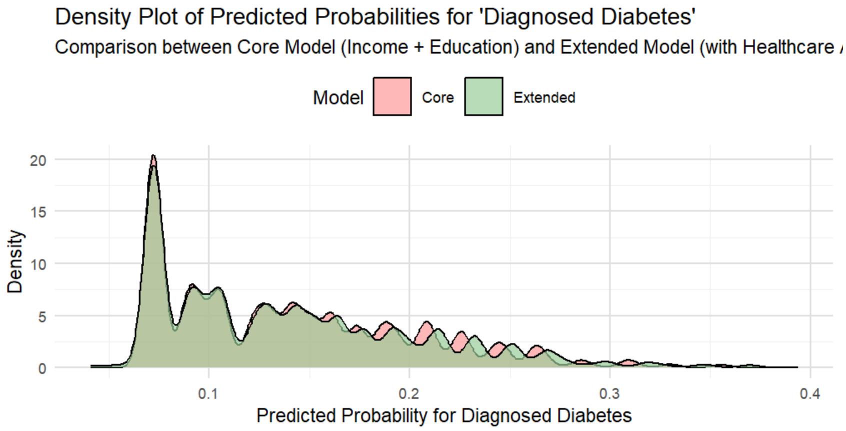
Appendix 7. Extended model (with 2 more parameters) metrics summary

When AnyHealthcare and NoDocbcCost are added, the coefficients for *Income* and *Education* remain relatively consistent, but intercepts in this extended model shift to 1.8090 and 1.9589, respectively. The extended model achieves a residual deviance of 240670.86 and an AIC of 240702.86, indicating a better fit relative to the core model.

```
> print(lr_test)
Likelihood ratio tests of ordinal regression models

Response: Diabetes_012
                                          Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
1                            Income + Education    253666   241196.5
2 Income + Education + AnyHealthcare + NoDocbcCost    253664   240670.9 1 vs 2     2 525.6157       0
```

Appendix 8.Likelihood Ratio Test for core model vs extended model

Model comparison is performed using the AIC and a likelihood ratio test, on a global scale fit. A drop of approximately 521.6 AIC points suggests that the extended model provides a better balance between model complexity and goodness of fit. The likelihood ratio test further reinforces this finding with an LR statistic of 525.62 (with 2 degrees of freedom) and a p-value effectively 0. This significant test statistic confirms that the addition of *AnyHealthcare* and *NoDocbcCost* leads to a statistically significant improvement in model performance.

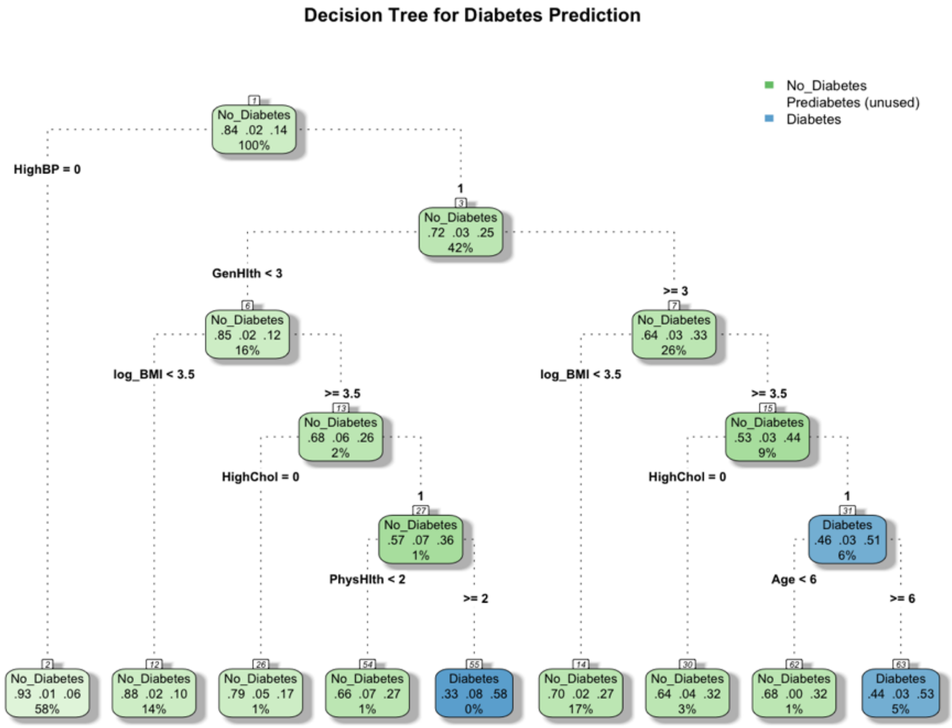


Appendix 9. Density Plot of Predicted Probabilities for Diagnosed Diabetes

The extended model shifts probability mass across categories, generally assigning higher risk to "Diagnosed Diabetes" and lowering the chance of "No Diabetes." This adjustment reflects the model's sensitivity to additional risk factors, providing a more nuanced risk assessment. The extended model's improved global fit indicates it captures real-world diabetes risk patterns more effectively, enhancing prediction accuracy and identifying more high-risk individuals than the core model with greater precision.

6.2 Machine Learning

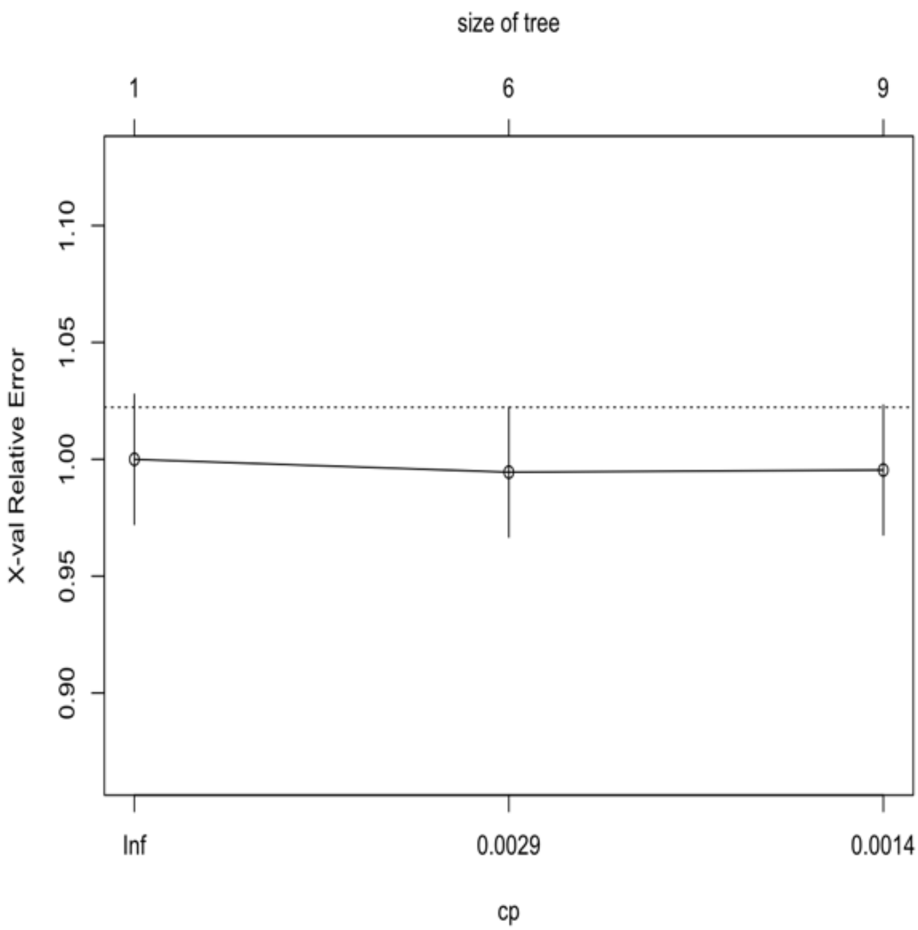
We perform a decision tree analysis to predict diabetes status based on various health indicators. We start by transforming variables like BMI and converting categorical variables to factors. A sample of 10,000 data points is used for faster processing. The data is split into training and test sets, and a decision tree model is built using the rpart function with multiple predictors. The tree is visualized, pruned to prevent overfitting, and then used to make predictions on the test set. A confusion matrix is created to evaluate the model's performance, and feature importance is analyzed and visualized to identify the most influential predictors for diabetes status.



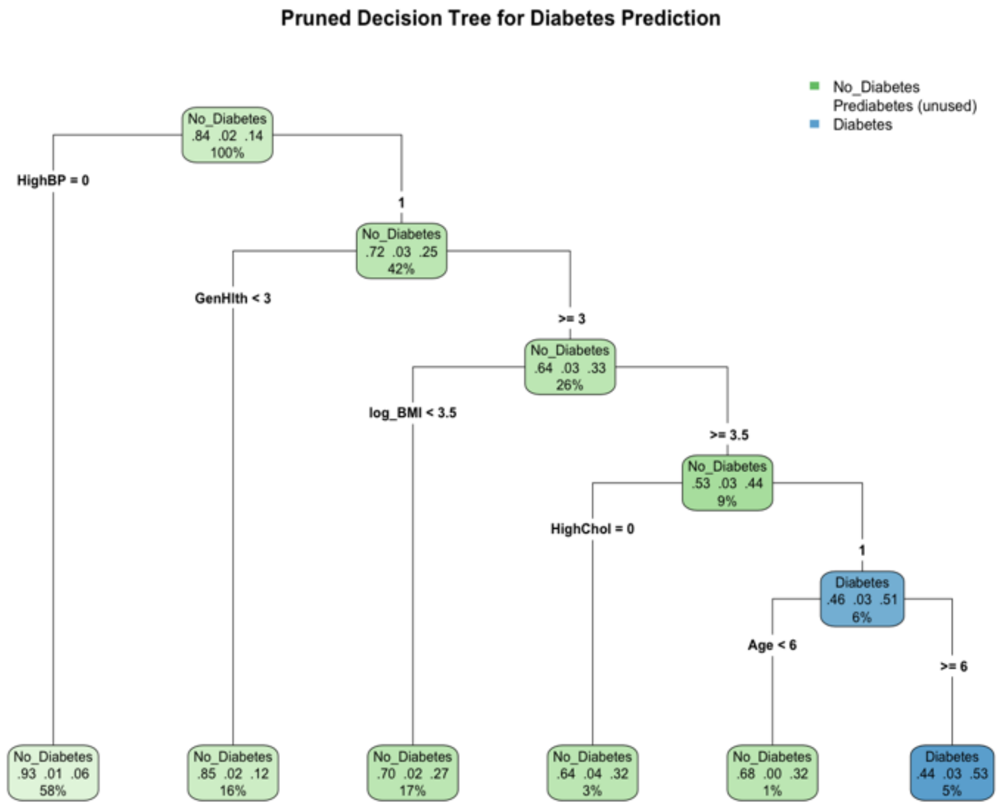
Appendix 10. Decision Tree for Diabetes Prediction

```
> # Print the complexity parameter table
> printcp(tree_model)

Classification tree:
rpart(formula = Diabetes_012 ~ HighBP + HighChol + log_BMI +
    Smoker + Stroke + HeartDiseaseorAttack + PhysActivity + Fruits +
    Veggies + HvyAlcoholConsump + AnyHealthcare + NoDocbcCost +
    GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Age + Education +
    Income, data = train_data, method = "class", control = rpart.control(minsplit = 20,
    minbucket = 10, cp = 0.001, maxdepth = 5))

Variables actually used in tree construction:
[1] Age      GenHlth  HighBP   HighChol log_BMI  PhysHlth

Root node error: 1089/7002 = 0.15553

n= 7002

         CP nsplit rel error  xerror     xstd
1 0.0045914      0   1.00000 1.00000 0.027847
2 0.0018365      5   0.96878 0.99449 0.027784
3 0.0010000      8   0.96327 0.99541 0.027795
```

Appendix 11.Decision Tree Complexity Parameter (CP) Metrics



Appendix 12. Cross-Validation Relative Error by Tree Size



Appendix 13. Pruned Decision Tree for Diabetes Prediction

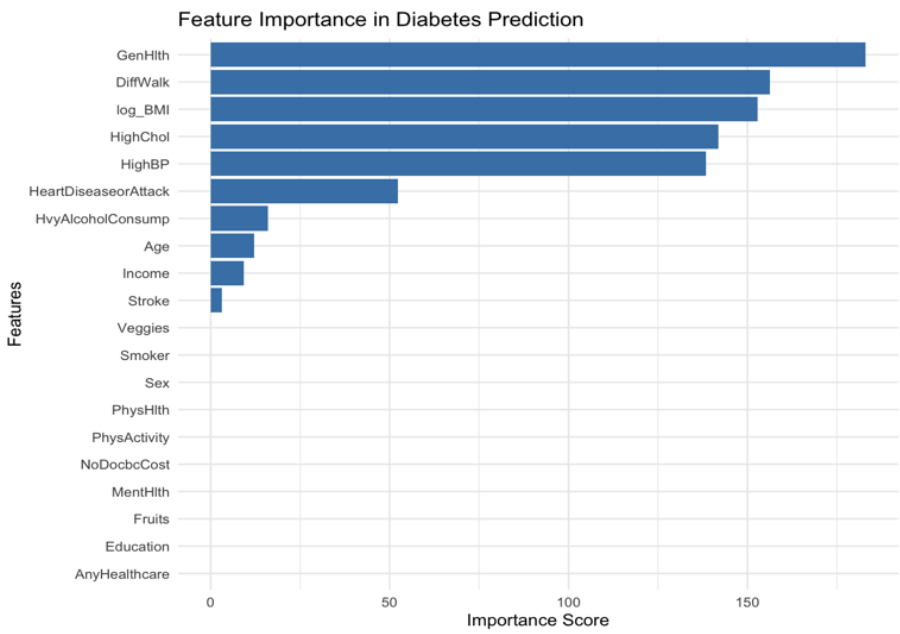
```
> # Create confusion matrix
> confusion_matrix <- confusionMatrix(predictions, test_data$Diabetes_012)
> print(confusion_matrix)
Confusion Matrix and Statistics

             Reference
Prediction    No_Diabetes Prediabetes Diabetes
  No_Diabetes        2456          42      342
  Prediabetes           0           0        0
  Diabetes             77           5       76

Overall Statistics

               Accuracy : 0.8446
                 95% CI : (0.8311, 0.8574)
    No Information Rate : 0.8449
    P-Value [Acc > NIR] : 0.5325

                  Kappa : 0.1916

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: No_Diabetes Class: Prediabetes Class: Diabetes
Sensitivity                      0.9696            0.00000         0.18182
Specificity                      0.1742            1.00000         0.96822
Pos Pred Value                   0.8648                NaN         0.48101
Neg Pred Value                   0.5127            0.98432         0.87958
Prevalence                       0.8449            0.01568         0.13943
Detection Rate                   0.8192            0.00000         0.02535
Detection Prevalence             0.9473            0.00000         0.05270
Balanced Accuracy                0.5719            0.50000         0.57502
>
```

Appendix 14. Confusion Matrix and Classification Metrics

An accuracy of 84.46% was observed after pruning the tree to ensure that the model does not overfit.

```
> # Feature importance
> importance <- varImp(pruned_tree)
> importance <- importance[order(-importance$Overall), , drop = FALSE]
> print(importance)
                       Overall
GenHlth             183.052884
DiffWalk            156.306071
log_BMI             152.888340
HighChol            141.814679
HighBP              138.490844
HeartDiseaseorAttack  52.391301
HvyAlcoholConsump    16.048299
Age                  12.071868
Income                9.212019
Stroke                3.186151
Smoker                0.000000
PhysActivity          0.000000
Fruits                0.000000
Veggies               0.000000
AnyHealthcare         0.000000
NoDocbcCost           0.000000
MentHlth              0.000000
PhysHlth              0.000000
Sex                   0.000000
Education             0.000000
```

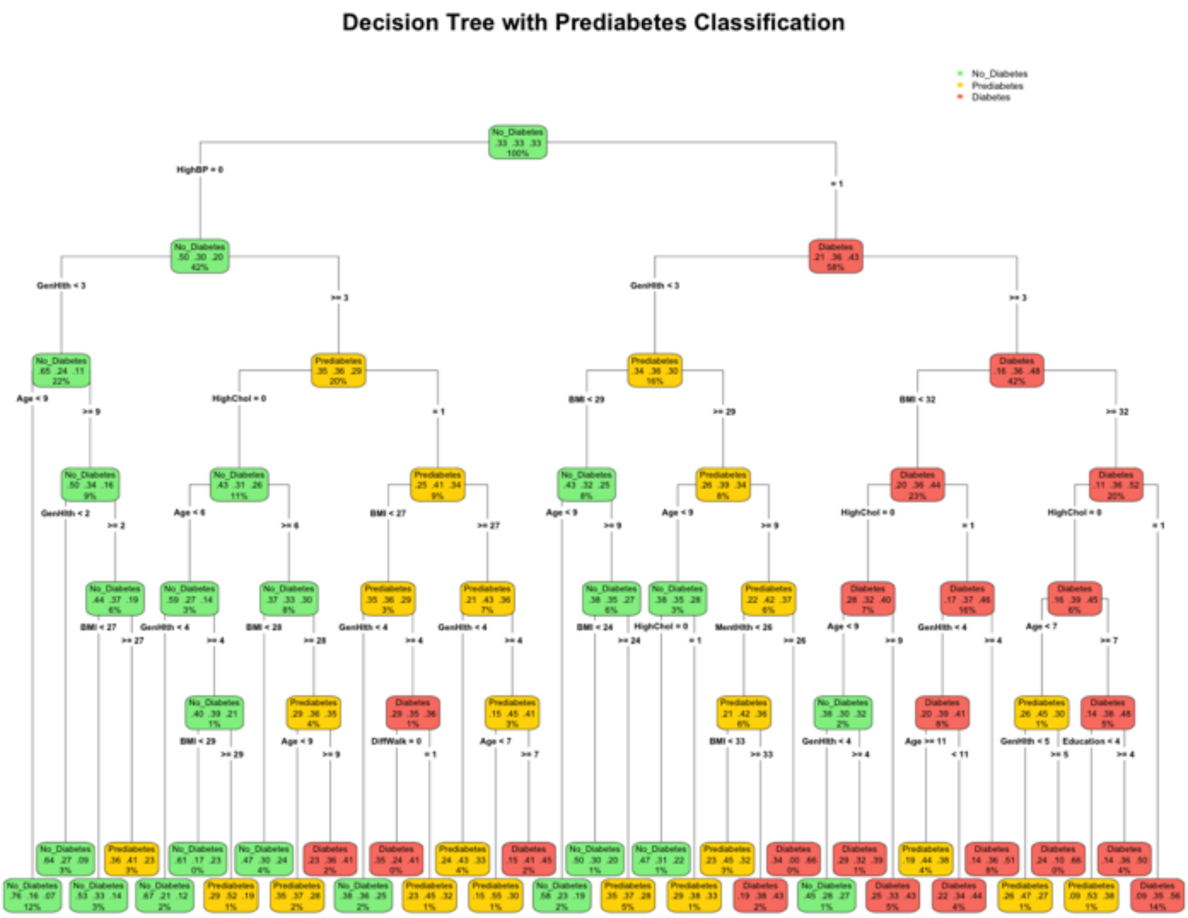
Appendix 15. Feature Importance Metrics for Diabetes Prediction



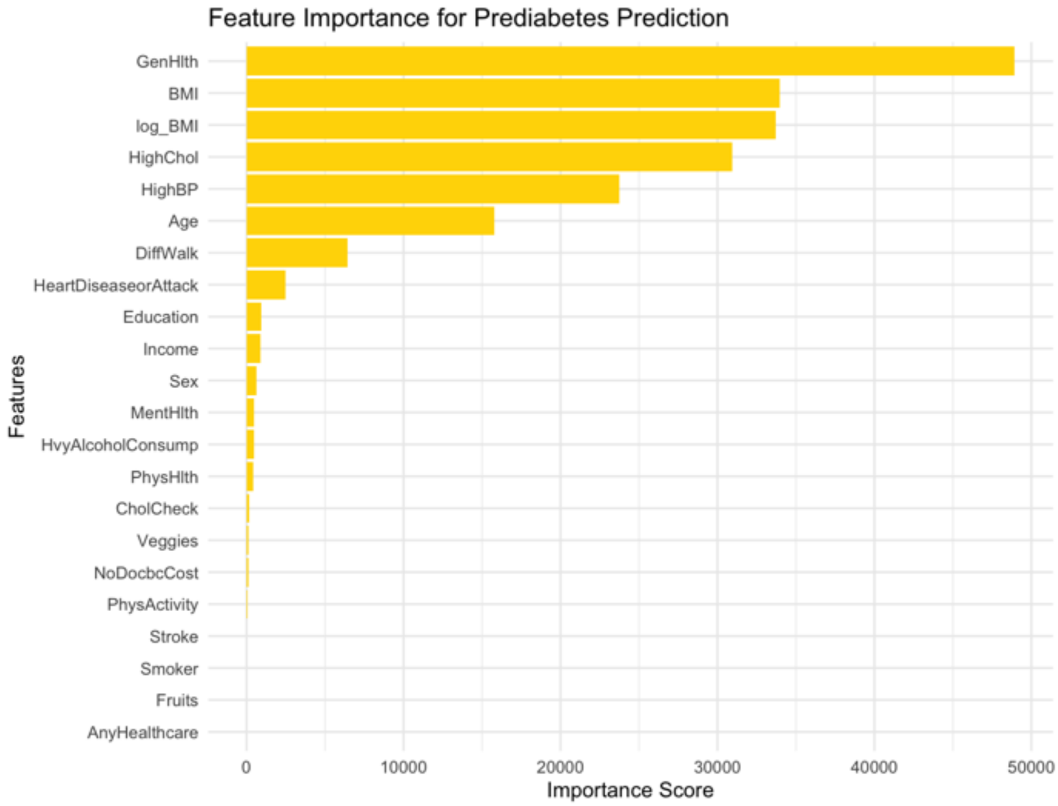
Appendix 16. Feature Importance for Diabetes Prediction

The key features of importance for prediction of diabetes were found to be GenHlth, DiffWalk, BMI, HighChol, and HighBP.

Additionally, we also performed a decision tree analysis for predicting diabetes status, with a specific focus on handling **prediabetes cases**. After loading and preparing the dataset, it addresses class imbalance by upsampling prediabetes cases in the training set. The decision tree is built using various predictors, with special handling for prediabetes, such as penalizing misclassifications of prediabetes more heavily. The model is visualized with prediabetes highlighted, and its performance is evaluated through a confusion matrix. Feature importance is calculated and visualized to identify key predictors for prediabetes.



Appendix 17. Decision Tree with Prediabetes Classification



Appendix 18. Feature Importance for Prediabetes Prediction

The key features of importance to predict pre-diabetes (or likelihood of people of becoming a prey to diabetes) are GenHlth, BMI, HighChol, HighBP, and Age.

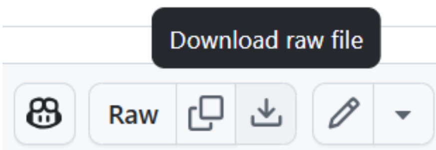
6.3 R Code

Full code can be found here:

https://github.com/Oganesson0221/Diabetes_Health_Indicators_Analysis/blob/main/Project.Rmd

Viewable HTML file can be found here where you use the navigation bar to direct to parts that may interest you, please click into this link, click “Download raw file”, and view in browser:

https://github.com/Oganesson0221/Diabetes_Health_Indicators_Analysis/blob/main/Project.html



Preview:

Introduction

Project Objectives and Research Questions

Install and Load Required Packages

Data Loading and Initial Exploration

Data Cleaning and Preprocessing

Exploratory Data Analysis (EDA)

Associations Between Variables and Diabetes_012

Statistical Analysis (Hypothesis Testing)

Multi-Variable Analysis

Machine Learning

Conclusion

Code

Diabetes Health Indicators Analysis

Zhao Qixian | Mehta Rishika | Tian Yumeng | Low Jo Yi, Nicole | Lu ShanShan

2025-04-13

Introduction

This document analyzes the diabetes health indicators from the BRFSS2015 dataset. We explore the data's structure, conduct exploratory data analysis (EDA), test associations through various statistical methods, build predictive models using ordinal logistic regression and decision trees, and compare model performance.

Project Objectives and Research Questions

Objectives

- Conduct comprehensive EDA to identify statistically significant relationships between health indicators and diabetes status.
- Apply both statistical tests and machine learning models to predict and explain diabetes risk.
- Compare the influence of demographic, lifestyle, and physiological variables on diabetes.

Research Questions

1. Which health indicators (e.g., blood pressure, cholesterol) are strongly associated with diabetes status?
2. How do factors such as age, BMI, and physical activity differ across diabetes groups?
3. Can ordinal regression and decision tree models accurately classify diabetes status?
4. What improvements are observed when including additional socioeconomic indicators in the prediction models?

7. References

Teboul, A. (2021, 8 November). *Diabetes Health Indicators Dataset*. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

De Feo, M., Del Pinto, R., Pagliacci, S., Grassi, D., & Ferri, C. (2021, 9 April). Real-World Hypertension Prevalence, awareness, treatment, and control in adult diabetic individuals: An Italian nationwide epidemiological survey. *High Blood Pressure & Cardiovascular Prevention*, *28*(3), 301–307. https://doi.org/10.1007/s40292-021-00449-7