ECO395M Final Project

Youngseok Yim (EID: yy9739)

4/24/2023

Abstract

This analysis aims to develop predictive models that identify risk factors for diabetes, and answer the research questions of what medical and lifestyle factors are most related to diabetes. The source of the data used for this analysis is the Behavioral Risk Factor Surveillance System (BRFSS) survey conducted by the Centers for Disease Control and Prevention (CDC) in 2017. The data were cleaned by dropping missing values and ambiguous answers, resulting in one dependent variable and 13 independent variables. Multiple logistic regression was used to analyze the relationship between the dependent and independent variables, and the logistic regression model achieved an accuracy of 85.98% on the test set. The results suggest that individuals with heart conditions, high cholesterol, and high blood pressure are at greater risk for diabetes. The study also found that lifestyle factors such as smoking, physical exercise are not strongly associated with diabetes.

1. Introduction and the Goal of the analysis

Diabetes is a prevalent chronic disease in the US that impacts the health of millions of individuals, including both type 1 and 2 diabetes. This disease is characterized by a loss of glucose regulation in the blood, leading to reduced quality of life and life expectancy. High blood sugar levels in people with diabetes can cause serious complications such as heart disease, vision loss, lower-limb amputation, and kidney disease. To improve early diagnosis and intervention, it is worth developing predictive models that identify risk factors for diabetes and pre-diabetes. Therefore, this analysis aims to answer the following research questions: What medical risk factors are most related to diabetes (including pre-diabetes)? What lifestyle factors, such as food intake, physical exercise, smoking/alcohol, are most related to diabetes (including pre-diabetes)?

2. Source of the dataset

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC) that collects state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The current survey format started in 2011, with more than 500,000 interviews conducted in the states, the District of Columbia, participating U.S. territories, and other geographic areas. Although the overall survey structure and items are the same, new questions are added, and some old questions are omitted/corrected every year. For this analysis, I selected the 2017 data, as they have more detailed questions regarding food intake behaviors of U.S. residents compared to other years.

3. Documentation of the dataset and data-cleaning

According to BRSFF, the surveyees are non-institutionalized adult population aged 18 years or older who reside in the United States in 2017. Respondents are identified through telephone-based methods. The BRFSS questionnaire consists of three components: a core component, optional modules, and state-added questions. The core component includes inquiries about current health perceptions, conditions, behaviors (such as health status, health care access, alcohol consumption, tobacco use, fruit and vegetable consumption, and HIV/AIDS risks), and demographic information. The optional BRFSS modules are sets of questions focused on specific topics such as pre-diabetes, diabetes, sugar-sweetened beverages, sun exposure, caregiving, shingles, and cancer survivorship.

For this analysis, we selected 14 questions and following answers from the core component and optional modules(Table 1). We cleaned the data by dropping missing values and ambiguous answers, such as ‘don’t know’, and came up with one dependent variable and 13 independent variables. While examining the questions, fruit and vegetable consumption data was entirely dropped due to its inconsistencies and extreme values. Since all of the answers in the original survey questions are designed for the respondents to answer with numerical or categorical numbers, we further renamed and reorganized each variable to fit the statistical model we were going to use. Additionally, by examining detailed statistics of each variable, we excluded potential invalid samples.

(Please refer to Appendix for a detailed data cleaning process conducted using Python and Pandas.)

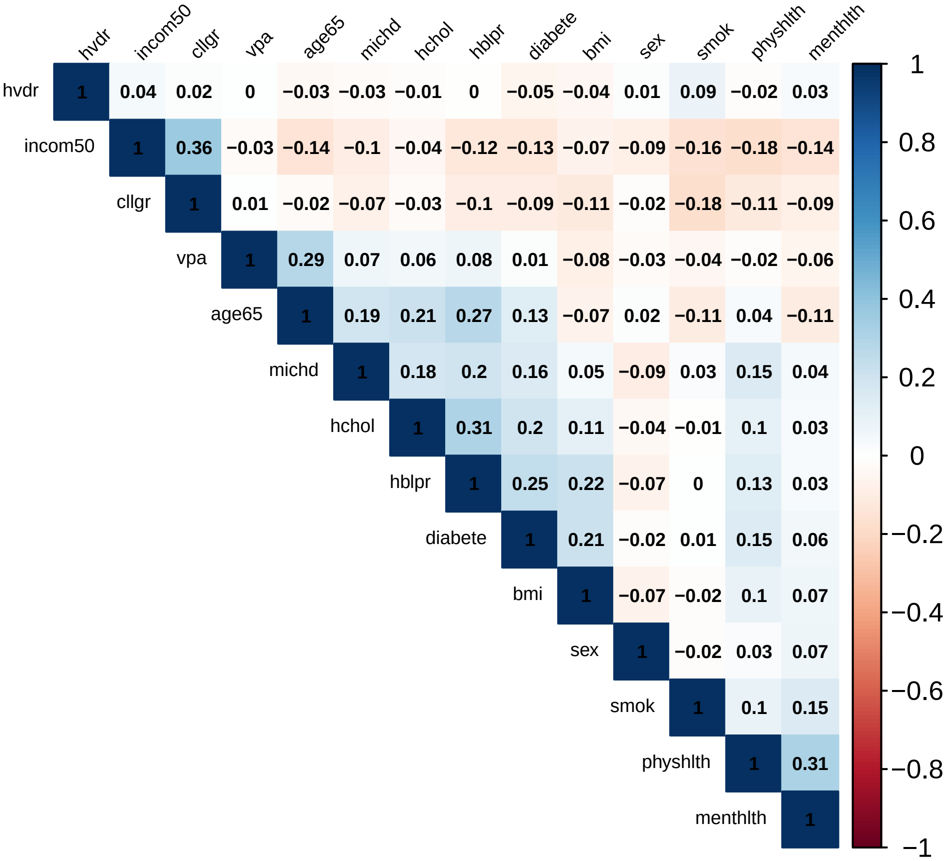
Table 1: Description of the variables

Variable Name	Description	Variable Type
diabete	diabete, pre- or borderline diabetes(no=0, yes=1)	categorical
bmi	body mass index	numerical
physhlth	number of days not feeling in good physical conditions in the past 30 days	numerical
menthlth	number of days not feeling in good mental conditions in the past 30 days	numerical
michd	respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	categorical
hchol	high cholesterol status (no=0, yes=1)	categorical
hblpr	high blood pressure status (no=0, yes=1)	categorical
vpa	vigourous physical activities (minutes/week)	numerical
hvdr	heavy drinker status (no=0, yes=1)	categorical
smok	smoking status (no=0, yes=1)	categorical
incom50	annual income status (<$50,000=0, >=$50,000=1)	categorical
cllgr	scollege/technical School Graduation Status (no=0, yes=1)	categorical
sex	gender (men=0, women=1)	categorical
age65	Age 65 or older (yes=0, no=1)	categorical

4. Methodology of the Analysis

4.1. Check multicollinearity: To check the presence of multicollinearity, I created a correlation coefficient heat map. I also calculated the variance inflation factor (VIF) for each independent variable. All VIF values were less than 2, indicating the absence of strong multicollinearity.

Figure 1: Correlation Heatmap of independent variables



	hvdr	incom50	cllgr	vpa	age65	michd	hchol	hblpr	diabete	bmi	sex	smok	physhlth	menthlth
hvdr	1	0.04	0.02	0	−0.03	−0.03	−0.01	0	−0.05	−0.04	0.01	0.09	−0.02	0.03
incom50		1	0.36	−0.03	−0.14	−0.1	−0.04	−0.12	−0.13	−0.07	−0.09	−0.16	−0.18	−0.14
cllgr			1	0.01	−0.02	−0.07	−0.03	−0.1	−0.09	−0.11	−0.02	−0.18	−0.11	−0.09
vpa				1	0.29	0.07	0.06	0.08	0.01	−0.08	−0.03	−0.04	−0.02	−0.06
age65					1	0.19	0.21	0.27	0.13	−0.07	0.02	−0.11	0.04	−0.11
michd						1	0.18	0.2	0.16	0.05	−0.09	0.03	0.15	0.04
hchol							1	0.31	0.2	0.11	−0.04	−0.01	0.1	0.03
hblpr								1	0.25	0.22	−0.07	0	0.13	0.03
diabete									1	0.21	−0.02	0.01	0.15	0.06
bmi										1	−0.07	−0.02	0.1	0.07
sex											1	−0.02	0.03	0.07
smok												1	0.1	0.15
physhlth													1	0.31
menthlth														1

Table 2: Variance Inflation Factor (VIF)

	bmi	physhlth	menthlth	michd	hchol	hblpr	vpa	hvdr	smok	incom50	cllgr	sex	age65
vif	1.1	1.17	1.16	1.11	1.15	1.26	1.1	1.02	1.09	1.24	1.19	1.04	1.29

As shown in Figure 1, since all of the correlation coefficients are < 0.5, there is no obvious evidence of strong multicollinearity. Also, in Table 2, variance inflation factor(VIF) for each variable is less than 2, which is a good indicator of the absence of strong multicollinearity.

4.2. Multiple linear regression and R-squared value:

```
## [1] 0.1296276
```

I performed multiple linear regression to analyze the relationship between the dependent and independent variables. The R-squared value of the model was 0.13, indicating that only 13% of the variance in the dependent variable could be explained by the independent variables. This suggests that the multiple OLS model does not fit the data well.

4.3. Multiple logistic regression: Instead I performed multiple logistic regression using all 13 selected independent variables to fit the categorical dependent variable. The logistic regression model is a powerful tool for modeling the relationship between a categorical dependent variable and one or more independent variables.

Table 3: Logistic Regression Results with original 13 independent variables

term	estimate	std.error	statistic	p.value	odds_ratio
bmi	-0.05467	0.00052	-104.14467	0.00000	0.94680
physhlth	0.01947	0.00078	25.01510	0.00000	1.01967
menthlth	-0.00136	0.00093	-1.46442	0.14308	0.99864
michd	0.44718	0.01994	22.42888	0.00000	1.56390
hchol	0.59216	0.01357	43.64929	0.00000	1.80789
hblpr	0.86655	0.01419	61.06278	0.00000	2.37869
vpa	-0.00045	0.00003	-17.94269	0.00000	0.99955
hvdr	-0.92087	0.03611	-25.49997	0.00000	0.39817
smok	-0.49487	0.02070	-23.91124	0.00000	0.60965
incom50	-0.74419	0.01386	-53.68149	0.00000	0.47512
cllgr	-0.50726	0.01409	-36.01307	0.00000	0.60214
sex	-0.43529	0.01260	-34.55787	0.00000	0.64708
age65	-0.01860	0.01436	-1.29522	0.19525	0.98157

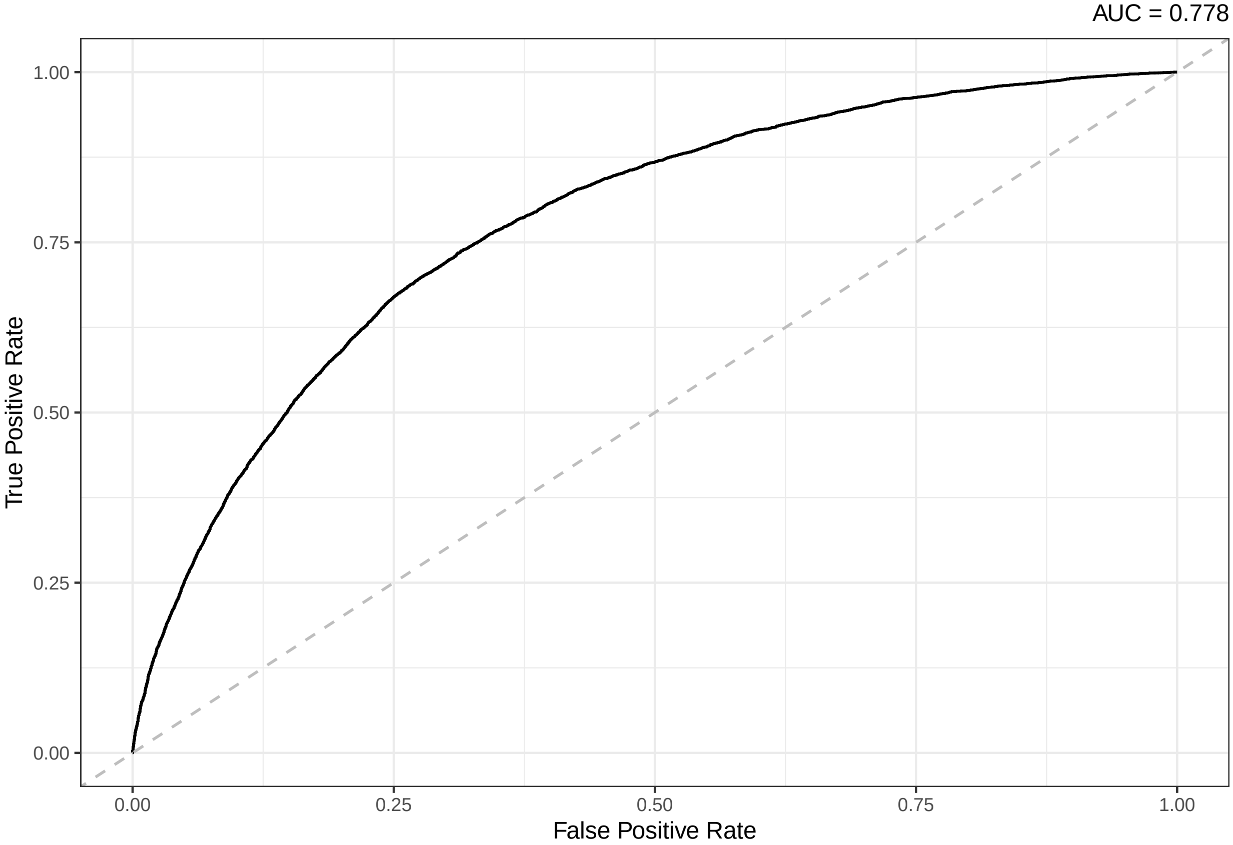
4.4. Model selection and validation using AUC value and ROC curve: After identifying that the independent variables menthlth and age65 had insignificant coefficients at the 5% level of significance, I removed them and conducted a multiple logistic regression with the 11 remaining variables. To validate the model, I split the data into training and testing groups and performed a validation test using machine learning. The test involved evaluating the model's performance based on the AUC value and ROC curve.

Table 4: Confusion matrix and accuracy

	0	1
0	44708	568
1	6837	716

Accuracy	0.8598

Figure 2: ROC Curve and AUC



The logistic regression model achieved an accuracy of 85.98% on the test set, which means that 85.98% of the predictions made by the model were correct. However, accuracy alone may not be the best measure of performance for a binary classification model. In this case, the model achieved an AUC of 0.778, which indicates that the model is able to distinguish between positive and negative cases with 77.8% accuracy. The AUC value ranges from 0.5 to 1.0, with a higher value indicating better performance. Therefore, our model's AUC of 0.778 can be considered good, but there is still room for improvement.

5. Result and conclusion

Table 5: Logistic Regression Results with 11 independent variables and odds ratio

term	estimate	std.error	statistic	p.value	odds_ratio
bmi	-0.05484	0.00052	-106.23152	0	0.94663
physhlth	0.01914	0.00074	25.88194	0	1.01933
michd	0.44342	0.01973	22.47433	0	1.55802
hchol	0.58961	0.01344	43.85747	0	1.80329
hblpr	0.86314	0.01389	62.15415	0	2.37060
vpa	-0.00046	0.00002	-19.00588	0	0.99954
hvdr	-0.92166	0.03610	-25.53295	0	0.39786
smok	-0.49554	0.02045	-24.22613	0	0.60924
incom50	-0.74134	0.01377	-53.83643	0	0.47648
cllgr	-0.50836	0.01406	-36.16683	0	0.60148
sex	-0.43858	0.01245	-35.22260	0	0.64495

According to the logistic regression model, there are certain medical risk factors and lifestyle factors that are closely associated with diabetes. The results show that individuals with heart conditions (michd) have a 1.56 times greater tendency to develop diabetes. Similarly, those with high cholesterol (hchol) have a 1.80 times greater tendency to develop diabetes. The strongest association was found with high blood pressure (hblpr), with individuals having a 2.37 times greater tendency to develop diabetes. These findings suggest that managing and treating these medical risk factors may help prevent or reduce the risk of developing diabetes.

On the other hand, the logistic regression model also revealed some interesting lifestyle factors that could play a significant role in preventing diabetes. Individuals with higher annual income of $50,000 or more (incom50) had a 0.48 times lower tendency to develop diabetes. Likewise, those with higher education of bachelor's degree or more (cllgr) had a 0.60 times lower tendency to develop diabetes. This suggests that increasing income and education level could be effective in preventing or reducing the risk of developing diabetes.

In conclusion, the logistic regression analysis suggests that medical risk factors such as high blood pressure and lifestyle factors such as annual income and education level are most closely associated with diabetes. Therefore, it is important to manage and treat medical conditions like high blood pressure and increase income and education level to prevent or reduce the risk of developing diabetes. By taking steps to manage and treat these risk factors, individuals may be able to improve their overall health and well-being and reduce their risk of developing diabetes.

Appendix: data cleaning using Python and Pandas(codes)

*The original data can be downloaded from https://www.cdc.gov/brfss/annual_data/annual_2017.html

Read in the dataset

```
df = pd.read_sas('../data/LLCP2017.XPT', encoding='utf-8')
```

Select columns for analysis after screening through the codebook

```
df_sel = df[['DIABETE3','_BMI5', 'PHYSHLTH','MENTHLTH', '_MICHD', '_RFCHOL1', '_RFHYPE5',
             'PA1VIGM_', '_RFDRHV5', '_RFSMOK3', '_INCOMG', '_EDUCAG', 'SEX', '_AGE65YR']]
```

Drop missing values

```
df_sel = df_sel.dropna()
df_sel.count()
```

DIABETE3

```
 # remove 7(don't know) and 9(refused)
 # change order and scale
  # 1 >> 1 (Yes diabetes)
  # 2 >> 1 (Yes Pre- or borderline diabetes)
  # 3 >> 0 (No)
  # 4 >> 1 (Yes Pre- or borderline diabetes)

df_sel['DIABETE3'].value_counts()
df_sel = df_sel[df_sel['DIABETE3'] != 7]
df_sel = df_sel[df_sel['DIABETE3'] != 9]
df_sel['DIABETE3'] = df_sel['DIABETE3'].astype(int)
df_sel['DIABETE3'].replace({1:1,2:1,3:0,4:1},inplace=True)
```

_BMI5

```
 # The original value is BMI * 100. Therefore, divide every value with 100
df_sel['_BMI5'] = df_sel['_BMI5'].div(100)
df_sel['_BMI5'].value_counts()
```

PHYSHLTH

```
 # change 88 to 0 (no bad physical health days)
 # remove 77(don't know) and 99(refused)
df_sel = df_sel[df_sel['PHYSHLTH'] != 77]
df_sel = df_sel[df_sel['PHYSHLTH'] != 99]
df_sel['PHYSHLTH'] = df_sel['PHYSHLTH'].replace({88:0})
df_sel['PHYSHLTH'].value_counts()
```

MENTHLTH

```
 # change 88 to 0 (no bad mental health days)
 # remove 77(don't know) and 99(refused)
df_sel = df_sel[df_sel['MENTHLTH'] != 77]
df_sel = df_sel[df_sel['MENTHLTH'] != 99]
df_sel['MENTHLTH'] = df_sel['MENTHLTH'].replace({88:0})
df_sel['MENTHLTH'].value_counts()
```

_MICHD

```
 # change order, scale
  # 1 >> 1 (Reported having MI or CHD)
  # 2 >> 0 (Did not report having MI or CHD)
df_sel['_MICHD'].replace({1:1,2:0},inplace=True)
df_sel['_MICHD'] = df_sel['_MICHD'].astype(int)
df_sel['_MICHD'].value_counts()
```

_RFCHOL1

```
 # change order, scale
  # 1 >> 0 (No)
  # 2 >> 1 (Yes)
df_sel['_RFCHOL1'].replace({1:0,2:1},inplace=True)
df_sel['_RFCHOL1'] = df_sel['_RFCHOL1'].astype(int)
df_sel['_RFCHOL1'].value_counts()
```

_RFHYPE5

```
 # removing 9(don't know)
 # change order, scale
  # 1 >> 0 (No)
  # 2 >> 1 (Yes)
df_sel=df_sel[df_sel['_RFHYPE5'] != 9]
df_sel['_RFHYPE5'].replace({1:0,2:1},inplace=True)
df_sel['_RFHYPE5'] = df_sel['_RFHYPE5'].astype(int)
df_sel['_RFHYPE5'].value_counts()
```

PA1VIGM_

```
df_sel['PA1VIGM_'] = df_sel['PA1VIGM_'].round(0)
df_sel['PA1VIGM_'].value_counts()
```

_RFDRHV5

```
  # 1 >> 0 (Not a heavy Drinker)
  # 2 >> 1 (Is Heavy Drinker)
  # Remove 9 >> (Don't know)
df_sel = df_sel[df_sel['_RFDRHV5'] != 9]
```

```
df_sel['_RFDRHV5'].replace({1:0,2:1},inplace=True)
df_sel['_RFDRHV5'] = df_sel['_RFDRHV5'].astype(int)
df_sel['_RFDRHV5'].value_counts()
```

_RFSMOK3

```
 # removing 9(don't know)
 # change order, scale
  # 1 >> 0 (No)
  # 2 >> 1 (Yes)
df_sel=df_sel[df_sel['_RFSMOK3'] != 9]
df_sel['_RFSMOK3'].replace({1:0,2:1},inplace=True)
df_sel['_RFSMOK3'] = df_sel['_RFSMOK3'].astype(int)
df_sel['_RFSMOK3'].value_counts()
```

_INCOMG

```
 # removing 9(don't know)
 # change order, scale
 # 1,2,3,4 ->> 0 (<$50,000)
 # 5 ->> 1 (>=$50,000)
df_sel=df_sel[df_sel['_INCOMG'] != 9]
df_sel['_INCOMG'].replace({1:0,2:0,3:0,4:0,5:1},inplace=True)
df_sel['_INCOMG'] = df_sel['_INCOMG'].astype(int)
df_sel['_INCOMG'].value_counts()
```

_EDUCAG - check,replace,remove

```
 # removing 9(don't know)
 # 1,2,3 ->> 0 (have not graduated college or technical school)
 # 4 ->> 1 (have graduated college or technical school)
df_sel=df_sel[df_sel['_EDUCAG'] != 9]
df_sel['_EDUCAG'].replace({1:0,2:0,3:0,4:1},inplace=True)
df_sel['_EDUCAG'] = df_sel['_EDUCAG'].astype(int)
df_sel['_EDUCAG'].value_counts()
```

SEX

```
  #  1 ->> 0 male
  #  2 ->> 1 female
  #  9 ->> refused (remove)
df_sel=df_sel[df_sel['SEX'] != 9]
df_sel['SEX'] = df_sel['SEX'].astype(int)
df_sel['SEX'].replace({1:0,2:1},inplace=True)
df_sel['SEX'].value_counts()
```

_AGE65YR

```
  #  1 ->> 0 18-64
  #  2 ->> 1 >=65
  #  3 ->> unknown (remove)
df_sel=df_sel[df_sel['_AGE65YR'] != 3]
df_sel['_AGE65YR'] = df_sel['_AGE65YR'].astype(int)
df_sel['_AGE65YR'].replace({1:0,2:1},inplace=True)
df_sel['_AGE65YR'].value_counts()
```

save to CSV

```
df_sel.to_csv('../data/2017_diabetes_cleaned_vld.csv', index = False)
```