REASEARCH PAPER ON HEART DISEASE

MIE 1413

APRIL 5, 2020

STUDENT NAME: YUCHAO WU (1000651984),
HUANGLEI LIN (1003207031)

Table of Contents

Research Paper on Heart Disease

Abstract: Through building a logistic regression model to predict heart disease using demographics, behavioural and medical condition features, we've learned that holding other features constant, people who had previous had a stroke is almost 2 (1.84) times higher than for people who hadn't. The possibility of getting diagnosed with heart disease for males is 62% higher than for females. Increases in age, number of cigarettes smoked per day and systolic blood pressure also increase the possibility of getting diagnosed with heart disease.

1.0 Introduction

In contemporary society, there is a considerable amount of people suffering from heart and cardiovascular disease (CVD) all around the world. This has been an ongoing issue for most countries during the past several decades. It seems that the cause of heart disease is due to numerous reasons. Many scholars suggested a number of variables and factors that could increase the risks of heart disease, such as age, overweight, medical history and family history. Rachel R (2011) [1]investigated the mechanisms underlying the sex difference in risk of coronary heart disease and concluded that whether the differences are biological or related to differences in smoking behaviour between men and women is unclear. Marianne U Jakobsen (2004)[2]'s study suggests that coronary heart disease risk relates to both the quantity and the quality of dietary fats. On the medical history side, Nicolas J Stapelberg (2012)[3] reviews heart rate variability in major depressive disorder and coronary heart disease.

The purpose of the study is straight forward - to explore some important factors (e.g., age, lifestyle and medical condition) that could increase the risks of heart disease and understand how they are correlated with heart disease. For example, we would like to study whether men or women are most susceptible to heart disease and would having a healthy lifestyle decrease the chance of having heart disease? Knowing this information could eventually make some contribution to heart disease prevention.

[1] Huxley, R. R., & Woodward, M. (2011, August 10). Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies

[2] Jakobsen, U., M., Kim, Schroll, Marianne, Heitmann, & L., B. (2004, July 15). Dietary Fat and Risk of Coronary Heart Disease: Possible Effect Modification by Gender and Age

[3] Stapelberg, N. J., Hamilton-Craig, I., Neumann, D. L., Shum, D. H. K., & McConnell, H. (n.d.). Mind and heart: Heart rate variability in major depressive disorder and coronary heart disease - a review and recommendations - Nicolas J Stapelberg, Ian Hamilton-Craig, David L Neumann, David HK Shum, Harry McConnell, 2012

2.0 Data Description

The source data for studying heart disease is publicly available on the Kaggle website and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The data set consists of 15 features (independent variables) and 1 dependent variable. The dataset overview can be found in Table 2-1.

Table 2-1 Dataset Overview

Feature Category	Feature	Description	Data Type
Demographics	Sex	male or female	Nominal
	Age	approximate age of the patient	Continuous
	Education	years of education	Continuous
Behavioural	Current Smoker	whether or not the patient is a current smoker	Nominal
	Cigs Per Day	the number of cigarettes that the person smoked on average in one day	Continuous
Medical(history)	BP Meds	whether or not the patient was on blood pressure medication	Nominal
	Prevalent Stroke	whether or not the patient had previously had a stroke	Nominal
	Prevalent Hyp	whether or not the patient was hypertensive	Nominal
	Diabetes	whether or not the patient had diabetes	Nominal
Medical(current)	Tot Chol	total cholesterol level	Continuous
	Sys BP	systolic blood pressure	Continuous
	Dia BP	diastolic blood pressure	Continuous
	BMI	Body Mass Index	Continuous
	Heart Rate	heart rate	Continuous
	Glucose	glucose level	Continuous
Predict variable (desired target)	10 year risk of coronary heart disease CHD	binary: “1”, means “Yes”, “0” means “No”	Nominal

There are mainly three categories of independent variables: demographic variables include sex, age and education; behavioral variables include smoking habit and medical condition variables

Table 2-2 Dataset Descriptive Statistics

Feature	null_ count	null_ percenage	non_null_ count	non_null_ percentage	distinct_ count	distinct_ percentage	max	min	mean	median	standard_ deviation
Sex	0	0	4240	1	2	0.0005	1	0	0.429245	0	0.495027
Age	0	0	4240	1	39	0.0092	70	32	49.580189	49	8.572942
Education	105	0.0248	4135	0.9752	4	0.0009	4	1	1.979444	2	1.019791
Current Smoker	0	0	4240	1	2	0.0005	1	0	0.494104	0	0.500024
Cigs Per Day	29	0.0068	4211	0.9932	33	0.0078	70	0	9.005937	0	11.922462
BP Meds	53	0.0125	4187	0.9875	2	0.0005	1	0	0.029615	0	0.169544
Prevalent Stroke	0	0	4240	1	2	0.0005	1	0	0.005896	0	0.076569
Prevalent Hyp	0	0	4240	1	2	0.0005	1	0	0.310613	0	0.462799
Diabetes	0	0	4240	1	2	0.0005	1	0	0.025708	0	0.15828
Tot Chol	50	0.0118	4190	0.9882	248	0.0585	696	107	236.699523	234	44.591284
Sys BP	0	0	4240	1	234	0.0552	295	83.5	132.354599	128	22.0333
Dia BP	0	0	4240	1	146	0.0344	142.5	48	82.897759	82	11.910394
BMI	19	0.0045	4221	0.9955	1364	0.3217	56.8	15.54	25.800801	25.4	4.07984
Heart Rate	1	0.0002	4239	0.9998	73	0.0172	143	44	75.878981	75	12.025348
Glucose	388	0.0915	3852	0.9085	143	0.0337	394	40	81.963655	78	23.954335
10 year risk of coronary heart disease CHD	0	0.00%	4240	100.00%	2	0.05%	1	0	0.151887	0	0.358953

include a couple of historical and current medical tests. Out of the 15 independent variables, 7 are nominal and 8 are continuous.

Table 2-2 has some descriptive statistics of the dataset. As is shown in the table, almost all features are of high quality. Education, Cigs Per Day, BP Meds, Tot Chol, BMI, Heart Rate, and Glucose have some missing values, however the percentages of missing are all lower than 10%. Although the problem is not so series, it will be dealt with before applying logistic regression to predict heart disease.

3.0 Methods & Results – Part 1

3.1 Chi-Square Test of Nominal Features

The Chi-Square test of independence is a test usually used to see if there is a relationship between two categorical variables. In our case, Chi-Square tests are performed to nominal features with the target variable to investigate whether one nominal feature is independent from the target. This will help us to better understand statistically the relationship between each feature and the target and enable feature selection of nominal variables based on the test results. Table 3-1 displays the results of Chi-Square tests on nominal features.

Table 3-1 Chi-Square Test Results of Nominal Features

Feature	Data Type	Chi-Square	P-Value
Sex	Nominal	32.6183349	0.0000
Current Smoker	Nominal	1.49720354	0.2211
BP Meds	Nominal	30.6459909	0.0000
Prevalent Stroke	Nominal	14.0336573	0.0002
Prevalent Hyp	Nominal	132.456286	0.0000
Diabetes	Nominal	38.4823381	0.0000

The **Null Hypothesis** of a Chi-Square test is that there is no relationship between the nominal feature and the target.

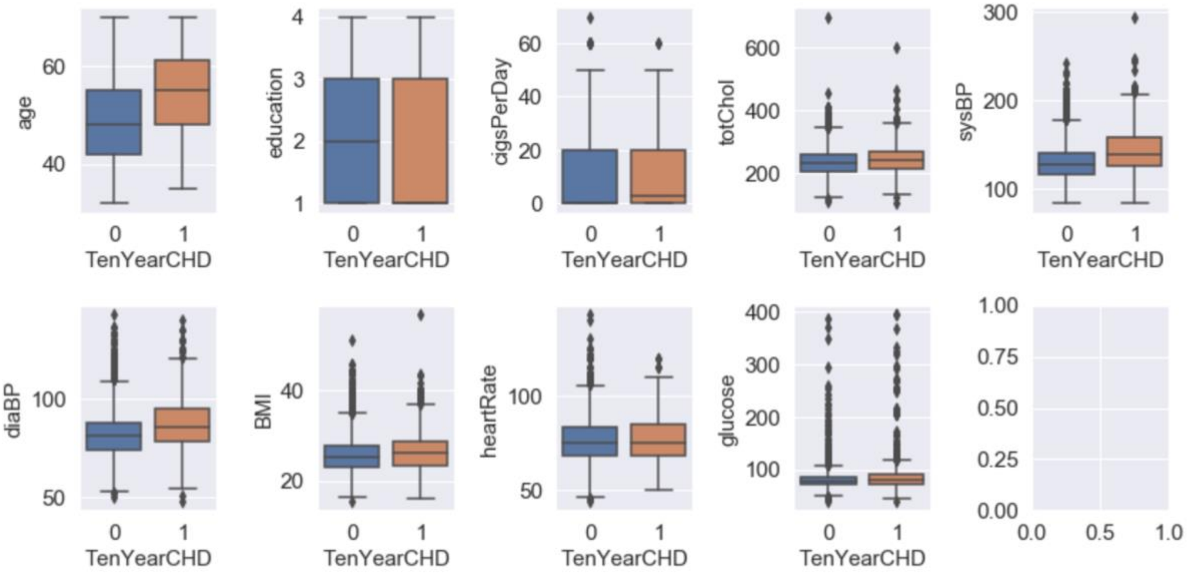
If the p-value of the test is lower than 0.05, we are 95% confident to reject the Null Hypothesis and conclude that there is a relationship between the feature and the target. From Table 3-1, we can conclude that out of the 6 nominal features, only Current Smoker doesn’t have any relationship with the target and thus will not be included in further analysis. Additional insights from Chi-Square tests are that Prevalent Hypertensive indicator, Diabetes indicator have strong relationships with the target, since their Chi-Square values are higher than other nominal features.

We would also be interested to learn more about which levels of the categorical feature are responsible for the relationship to the target. We will be able to answer this after constructing a logistic regression model.

3.2 T-test of Continuous Features

Graph 3-1 shows the box plot of each continuous feature grouped by the target variable: people who have 10-year risk of coronary heart disease (CHD) and people who don't. It's not difficult to notice that people who have 10-year risk of CHD are older, have more cigarettes per day, higher total cholesterol level, systolic blood pressure, diastolic blood pressure, Body Mass Index and glucose level compared to people who don't.

Graph 3-1 Box Plot of Continuous Variables by Target



Independent sample t-test is usually used to statistically measure whether the average (expected) value differs significantly across independent samples. Although we do see some trends in the box plots, t-tests are performed to see if one continuous feature's distribution is significantly different for people who have 10-year risk of CHD and people who don't. Table 3-2 shows the results from t-tests.

Table 3-2 T-test Results of Continuous Features

Feature	Data Type	T-Statistics	P Value
Age	Continuous	375.101126	0.0
Education	Continuous	109.919121	0.0
Cigs Per Day	Continuous	48.3349239	0.0
Tot Chol	Continuous	345.412168	0.0
Sys BP	Continuous	390.64847	0.0
Dia BP	Continuous	452.174409	0.0
BMI	Continuous	407.780945	0.0
Heart Rate	Continuous	409.867171	0.0
Glucose	Continuous	222.363567	0.0

The **Null Hypothesis** of t-test is that independent samples have identical average (expected) value. With a p-value lower than 0.05, we will be 95% confident to reject the null hypothesis and conclude that independent samples have different average (expected) value.

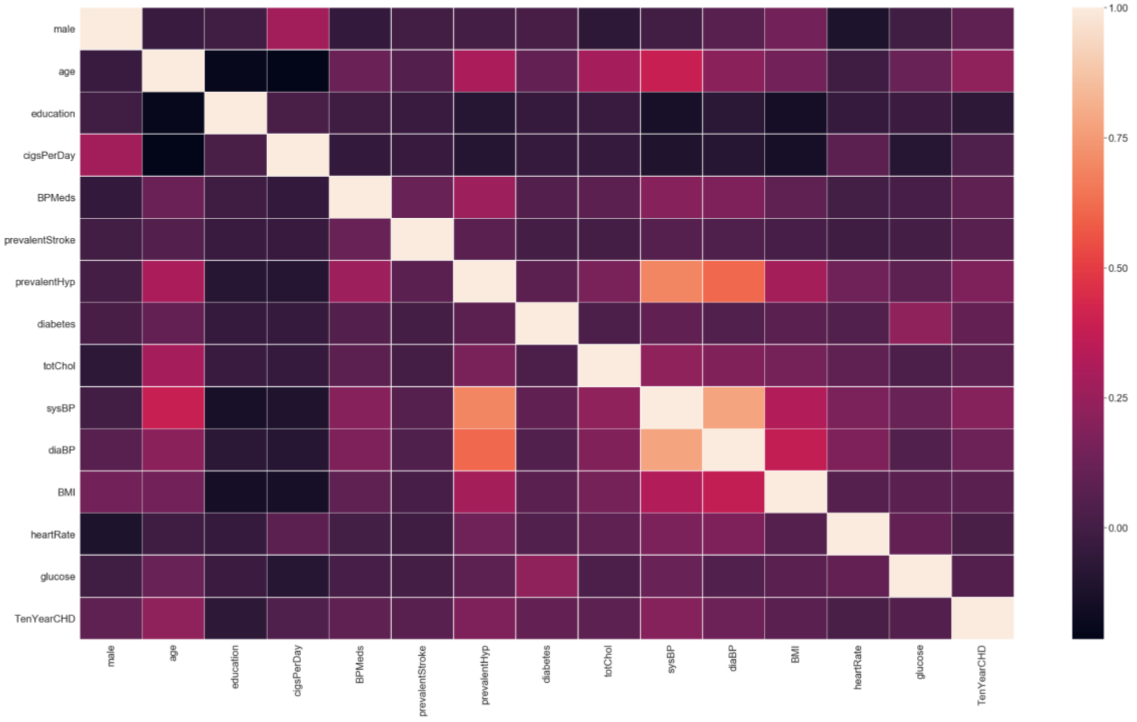
Therefore, from Table 3-2, we can conclude that distribution of all continuous features is different for people who have 10-year risk of CHD and people who don't, which could be an indication of the prediction power of our continuous variables. Out of all continuous features, Dia BP, Heart Rate and BMI have comparatively higher t-statistics. They might be more predictive of heart disease.

3.3 Correlation Matrix

Since multicollinearity reduces the precision of the estimate coefficients of logistic regression, making it hard to trust the p-values of independent features for determining which features are statistically significant, it needs to be tested before applying logistic regression.

Correlation matrix of the dataset is thus plotted and analyzed in hope of removing features that are highly correlated with one another. However, we didn't find any correlations between features higher than 0.8 or lower than -0.8. We are ready for building a regression model.

Graph 3-2 Dataset Correlation Matrix



4.0 Methods & Results – Part 2

4.1 Missing Values

As we've noticed before, Education, Cigs Per Day, BP Meds, Tot Chol, BMI, Heart Rate and Glucose have some missing values and they have to be imputed for the purpose of building a regression model. We felt that the nature of missing values in Education and Cigs Per Day is different from the other features. The reason for a person to not fill out Education could be that he/she doesn't have any education or he/she doesn't have the lowest level of education offered for selection. Similarly, a person didn't fill out Cigs Per Day probably because he/she doesn't smoke. Therefore, we decided to impute the missing values for Education and Cigs Per Day to 0 and impute the missing values for BP Meds, Tot Chol, BMI, Heart Rate and Glucose to be the median of the non-missing values for the respective feature.

4.2 Logistic Regression

Our goal of building the logistic regression model is that it should be a high-quality model that includes as few features as possible, which will make it easier for us to interpret model results. Thus firstly, we decided to use backward elimination to do some further feature selection. This process of course shouldn't compromise the predictive power of the model. Secondly, train and test split were performed and the train dataset was fed to a logistic regression model. Thirdly, trained model was applied on the test dataset for performance evaluation. Graph 4-1 illustrates the process flow of our methodology.

Graph 4-1 Process Flow

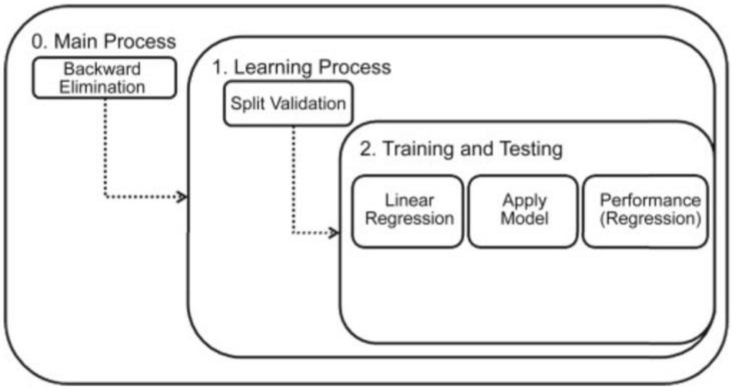
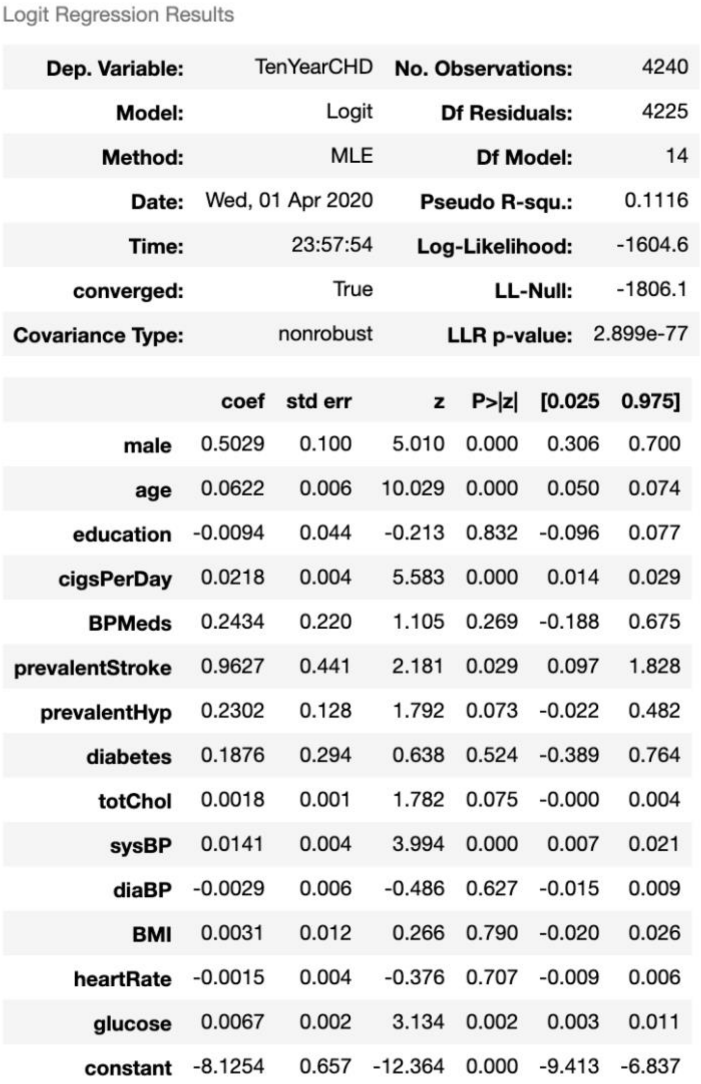


Table 4-1 contains the results summary of the initial logistic regression model built using all features from our previous analysis. It's not difficult to observe that some features have p-values higher than 0.05. Backward elimination of features is thus performed. Each time, the feature

Table 4-1 Logistic Regression Summary Table Before Backward Elimination



```
Logit Regression Results

       Dep. Variable:       TenYearCHD   No. Observations:        4240
              Model:            Logit       Df Residuals:        4225
             Method:              MLE           Df Model:          14
               Date: Wed, 01 Apr 2020     Pseudo R-squ.:      0.1116
               Time:         23:57:54     Log-Likelihood:     -1604.6
          converged:             True            LL-Null:     -1806.1
    Covariance Type:        nonrobust        LLR p-value:   2.899e-77

                   coef  std err        z  P>|z|  [0.025  0.975]
           male  0.5029    0.100    5.010  0.000   0.306   0.700
            age  0.0622    0.006   10.029  0.000   0.050   0.074
      education -0.0094    0.044   -0.213  0.832  -0.096   0.077
     cigsPerDay  0.0218    0.004    5.583  0.000   0.014   0.029
         BPMeds  0.2434    0.220    1.105  0.269  -0.188   0.675
 prevalentStroke 0.9627    0.441    2.181  0.029   0.097   1.828
   prevalentHyp  0.2302    0.128    1.792  0.073  -0.022   0.482
       diabetes  0.1876    0.294    0.638  0.524  -0.389   0.764
        totChol  0.0018    0.001    1.782  0.075  -0.000   0.004
          sysBP  0.0141    0.004    3.994  0.000   0.007   0.021
          diaBP -0.0029    0.006   -0.486  0.627  -0.015   0.009
            BMI  0.0031    0.012    0.266  0.790  -0.020   0.026
      heartRate -0.0015    0.004   -0.376  0.707  -0.009   0.006
        glucose  0.0067    0.002    3.134  0.002   0.003   0.011
       constant -8.1254    0.657  -12.364  0.000  -9.413  -6.837
```

with the highest p-value was removed from the modelling process. This process is done multiple times until all the features left in the modelling process are with p-values less than 0.05.

Table 4-2 Logistic Regression - Feature Odds Ratio and P-value

Features Selected after Backward Elimination	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	P-value
male	1.339436	1.960696	1.620564	0.000
age	1.054621	1.079399	1.066938	0.000
cigsPerDay	1.014207	1.029606	1.021877	0.000
prevalentStroke	1.208851	6.684146	2.842559	0.017
sysBP	1.013223	1.021203	1.017205	0.000

glucose	1.00441	1.010843	1.007622	0.000
constant	0.000098	0.000452	0.000211	0.000

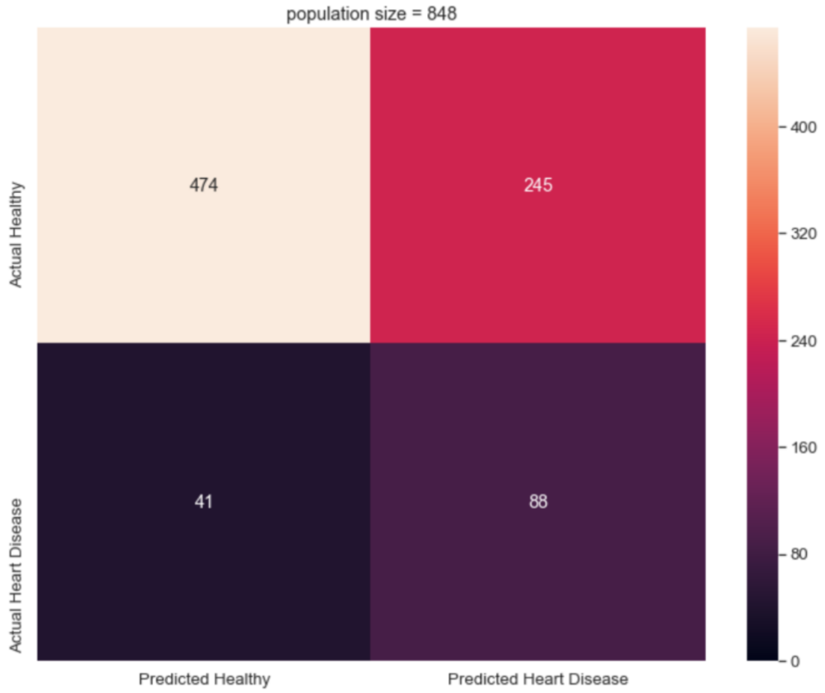
Table 4-2 shows the features that survived the backward elimination process. The following interpretations were made from Table 4-2 Odds Ratio and P-value statistics:

- Holding all other independent features constant, the possibility of getting heart disease for people who had previous had a stroke is almost 2 (1.84) times higher than for people who hadn’t.
- Holding all other independent features constant, the possibility of getting diagnosed with heart disease for males is 62% higher than for females.
- Holding all other independent features constant, with every one year of increase in age, there is 6.7% increase in the possibility of getting heart disease.
- Holding all other independent features constant, with one extra cigarette per day, there is a 2.2% increase in the possibility of getting diagnosed with heart disease.
- Holding all other independent features constant, with every one level of increase in Systolic Blood Pressure, there is a 1.7% increase in the possibility of getting heart disease.

Then train and test split was performed and 20% of our data was put aside as the test dataset. The train dataset is fed into a logistic regression model with the class_weight parameter set to ‘balanced’ since the class distribution of our target variable is imbalanced (only 15% of our population have heart disease while 85% are healthy).

The confusion matrix on the test dataset, as shown in graph 4-2, is one of the metrics we’ve used to evaluate the model performance. Graph 4-2 shows that the total population in the test data is 848, with 129 people actually will be diagnosed with heart disease. Out of the 129 people who actually will be diagnosed, we successfully predicted 88 of them, indicating that the recall of our model is 68%. However, we predicted extra 245 people to have heart disease in order to capture the 88 people who actually have heart disease, which indicates that the precision of our model is 26%. The f1 score of our model is thus 38% and the accuracy is 66%.

Graph 4-2 Confusion Matrix on Test Data



We did choose to sacrifice the accuracy of our model to be able to successfully capture more people who actually have heart disease. The accuracy of the model can go up to around 90%, but with many more **Type II** errors. This is not advisable as we do believe that in our specific case, a False Negative (ignoring the probability of heart disease when actually there is one) is more dangerous than a False Positive. Heart disease is a dangerous disease which needs to be treated as early as possible to avoid health damages. And for our wrong predictions (those 225 people in our test data), it would always be a good thing for people to monitor their health status and maybe make changes for healthier lifestyles.

5.0 Discussion & Conclusion

From the research, we've learned that one's gender, age, cigarettes smoked per day, prevalent stroke, systolic blood pressure and glucose level play significant roles in heart disease prediction. The conclusion is solid, since all these features are selected from statistical tests and backward elimination process and end up with p-values lower than 0.05 in the logistic regression model.

In general, people who had previously had a stroke are more susceptible to heart disease than people who hadn't. Men are more susceptible to heart disease than women. According to the data from the Public Health Agency of Canada's Canadian Chronic Disease Surveillance System (CCDSS), men are two times more likely to suffer a heart attack than women and tend to be newly diagnosed with heart disease about 10 years younger than women [4]. Increases in age, number of cigarettes smoked per day and systolic blood pressure also increase the possibility of getting diagnosed with heart disease. Based on the statistics from American Stroke Association, about two-thirds of CVD deaths occur in people age of 75 or even older and the leading (top 3) causes of death in older women and men (>65 years of age) were disease of heart (NO.1), cancer (NO.2) and chronic lower respiratory disease (NO.3) [5]. Another research from American Heart Association also indicates that nearly 20 percent of the deaths caused by CVD are due to cigarette smoking and the non-smokers who are regularly exposed to second-hand smoke have a 25 to 30 percent increased risk of coronary heart disease than those not exposed [6]. Furthermore, glucose level increase causes a negligible increase in the possibility of getting heart disease, which might be due to the presence of good glucose level in total glucose level.

For our specific case, ignoring the probability of getting heart disease when actually there is one is more dangerous than being predicted with heart disease when actually doesn't have it. Therefore, the model is fine tuned in order to reduce health risks and encourage people to monitor their health status.

What we've learned from the research process is that we could have done less statistical tests between each feature with the target variable. Because the detected relationships from statistically tests are not deterministic for the prediction power of independent variables. On the other hand, backward elimination process for feature selection of logistic regression model is probably the more solid way of selecting features.

References:

1. Huxley, R. R., & Woodward, M. (2011, August 10). Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies

2. Jakobsen, U., M., Kim, Schroll, Marianne, Heitmann, & L., B. (2004, July 15). Dietary Fat and Risk of Coronary Heart Disease: Possible Effect Modification by Gender and Age

3. Stapelberg, N. J., Hamilton-Craig, I., Neumann, D. L., Shum, D. H. K., & McConnell, H. (n.d.). Mind and heart: Heart rate variability in major depressive disorder and coronary heart disease - a review and recommendations - Nicolas J Stapelberg, Ian Hamilton-Craig, David L Neumann, David HK Shum, Harry McConnell, 2012

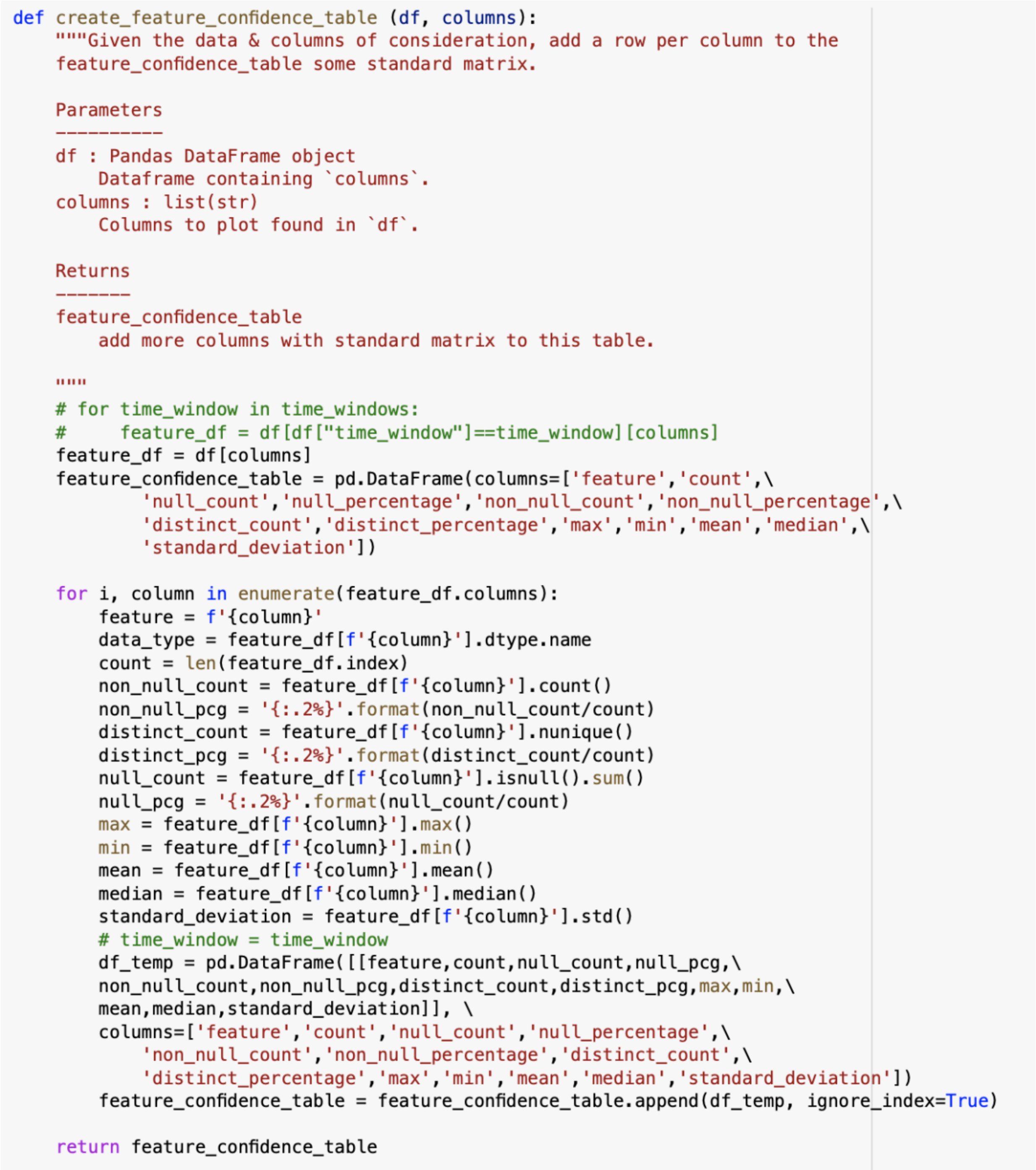
4. Heart Disease in Canada. (2017, 02 10). Retrieved from Government of Canada: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

5. Older Americans & Cardiovascular Diseases. (2016). Retrieved from American Stroke Association: https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_483970.pdf

6. How Smoking Affects Heart Health. (2020, 03 03). Retrieved from U.S. Food & Drug Administration: https://www.fda.gov/tobacco-products/health-information/how-smoking-affects-heart-health

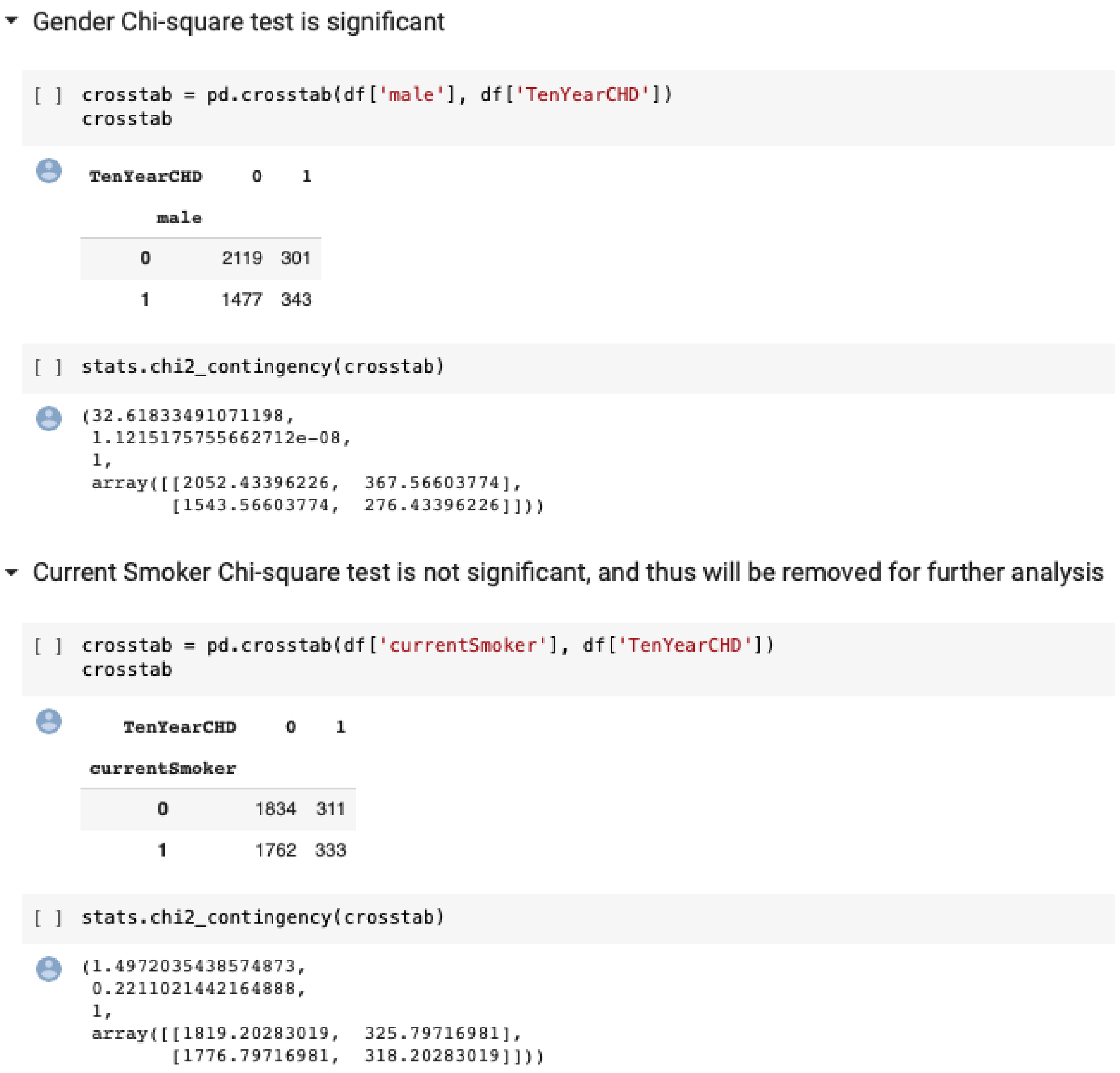
Appendices: Statistical Python Code

Appendix A: Feature Descriptive



```
def create_feature_confidence_table (df, columns):
    """Given the data & columns of consideration, add a row per column to the
    feature_confidence_table some standard matrix.

    Parameters
    ----------
    df : Pandas DataFrame object
        Dataframe containing `columns`.
    columns : list(str)
        Columns to plot found in `df`.

    Returns
    -------
    feature_confidence_table
        add more columns with standard matrix to this table.

    """
    # for time_window in time_windows:
    #     feature_df = df[df["time_window"]==time_window][columns]
    feature_df = df[columns]
    feature_confidence_table = pd.DataFrame(columns=['feature','count',\
            'null_count','null_percentage','non_null_count','non_null_percentage',\
            'distinct_count','distinct_percentage','max','min','mean','median',\
            'standard_deviation'])

    for i, column in enumerate(feature_df.columns):
        feature = f'{column}'
        data_type = feature_df[f'{column}'].dtype.name
        count = len(feature_df.index)
        non_null_count = feature_df[f'{column}'].count()
        non_null_pcg = '{:.2%}'.format(non_null_count/count)
        distinct_count = feature_df[f'{column}'].nunique()
        distinct_pcg = '{:.2%}'.format(distinct_count/count)
        null_count = feature_df[f'{column}'].isnull().sum()
        null_pcg = '{:.2%}'.format(null_count/count)
        max = feature_df[f'{column}'].max()
        min = feature_df[f'{column}'].min()
        mean = feature_df[f'{column}'].mean()
        median = feature_df[f'{column}'].median()
        standard_deviation = feature_df[f'{column}'].std()
        # time_window = time_window
        df_temp = pd.DataFrame([[feature,count,null_count,null_pcg,\
        non_null_count,non_null_pcg,distinct_count,distinct_pcg,max,min,\
        mean,median,standard_deviation]], \
        columns=['feature','count','null_count','null_percentage',\
            'non_null_count','non_null_percentage','distinct_count',\
            'distinct_percentage','max','min','mean','median','standard_deviation'])
        feature_confidence_table = feature_confidence_table.append(df_temp, ignore_index=True)

    return feature_confidence_table
```

Appendix B: Crosstab Analysis of Nominal Features to Target



▾ Gender Chi-square test is significant

```
[ ] crosstab = pd.crosstab(df['male'], df['TenYearCHD'])
    crosstab
```

TenYearCHD	0	1
male		
0	2119	301
1	1477	343

```
[ ] stats.chi2_contingency(crosstab)
```

```
(32.61833491071198,
 1.1215175755662712e-08,
 1,
 array([[2052.43396226,  367.56603774],
        [1543.56603774,  276.43396226]]))
```

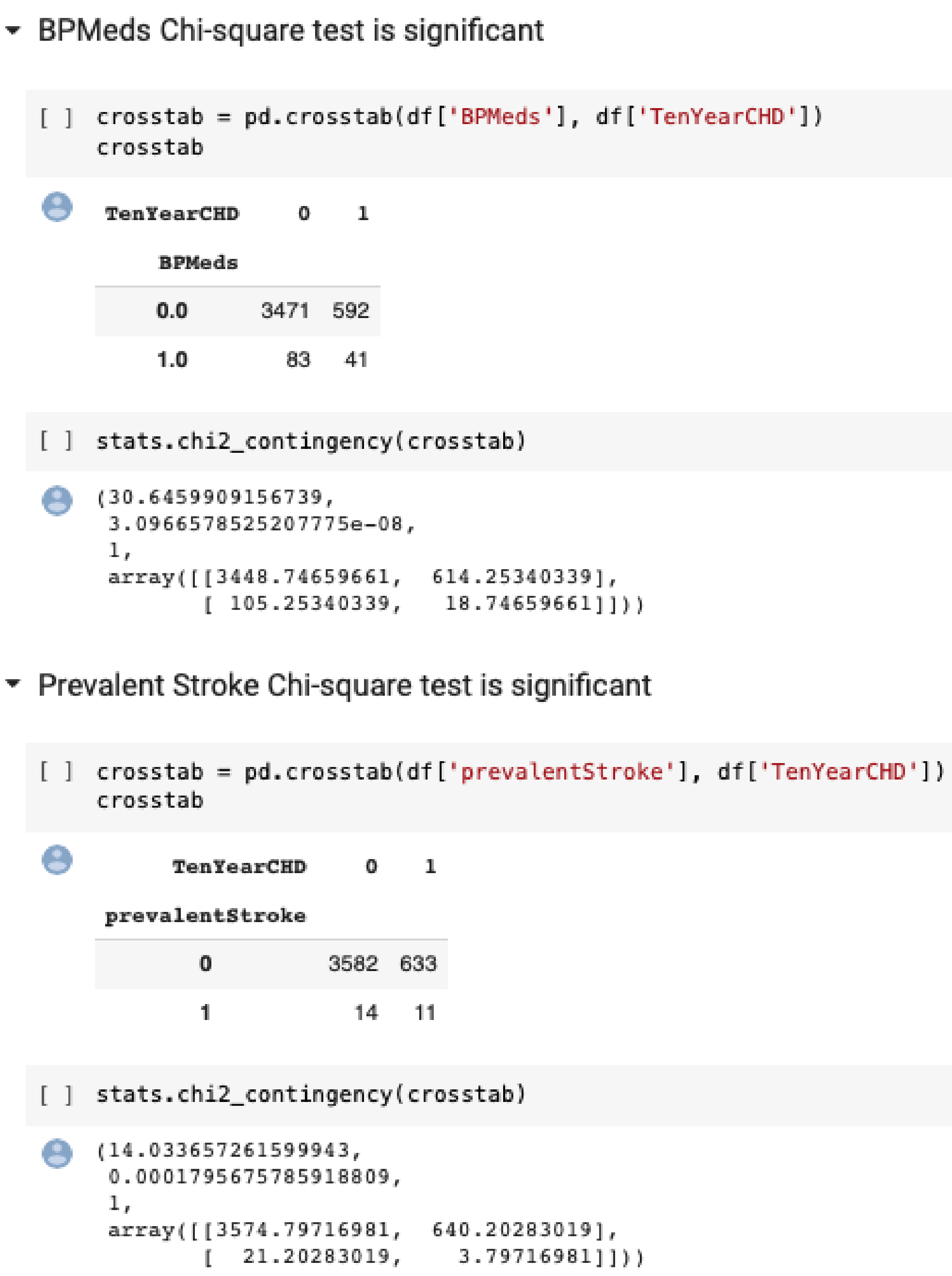
▾ Current Smoker Chi-square test is not significant, and thus will be removed for further analysis

```
[ ] crosstab = pd.crosstab(df['currentSmoker'], df['TenYearCHD'])
    crosstab
```

TenYearCHD	0	1
currentSmoker		
0	1834	311
1	1762	333

```
[ ] stats.chi2_contingency(crosstab)
```

```
(1.4972035438574873,
 0.2211021442164888,
 1,
 array([[1819.20283019,  325.79716981],
        [1776.79716981,  318.20283019]]))
```



BPMeds Chi-square test is significant

```
[ ] crosstab = pd.crosstab(df['BPMeds'], df['TenYearCHD'])
    crosstab
```

```
TenYearCHD      0     1
    BPMeds
       0.0   3471   592
       1.0     83    41
```

```
[ ] stats.chi2_contingency(crosstab)
```

```
(30.6459909156739,
 3.0966578525207775e-08,
 1,
 array([[3448.74659661,  614.25340339],
        [ 105.25340339,   18.74659661]]))
```

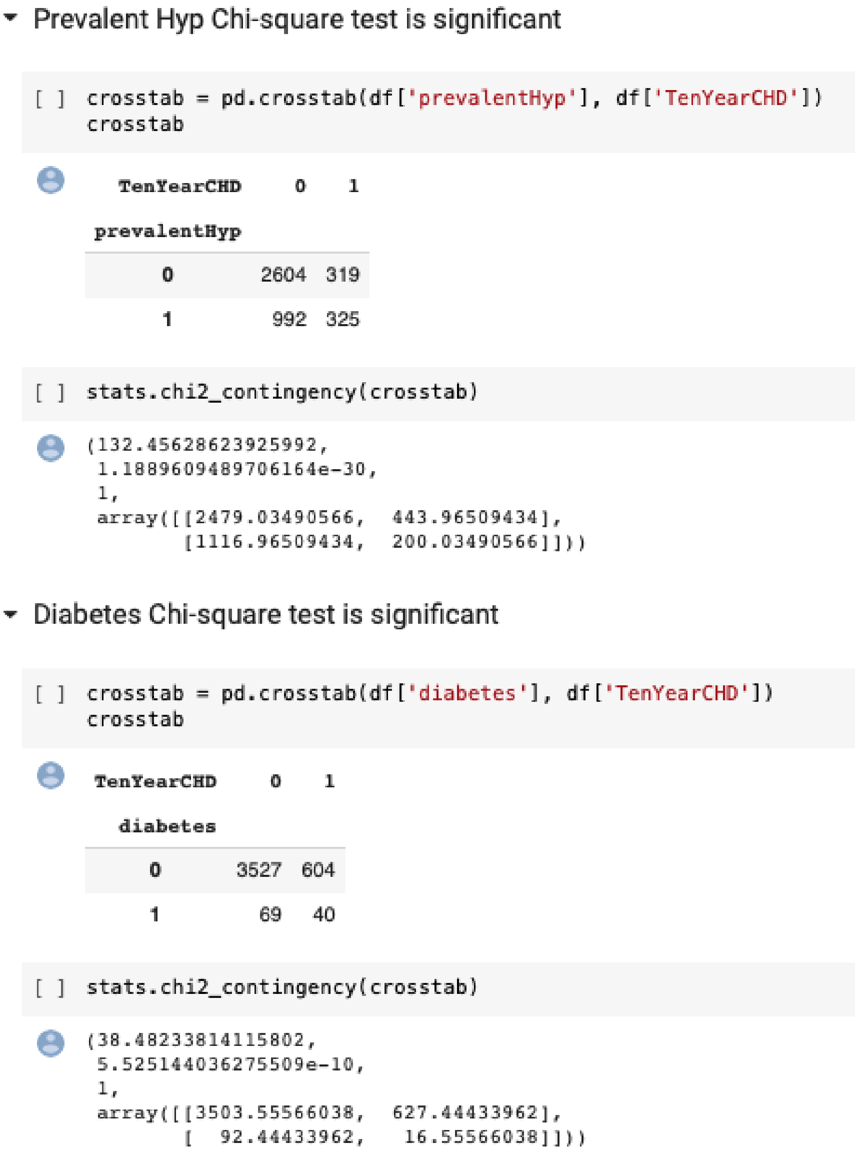
Prevalent Stroke Chi-square test is significant

```
[ ] crosstab = pd.crosstab(df['prevalentStroke'], df['TenYearCHD'])
    crosstab
```

```
     TenYearCHD      0     1
prevalentStroke
              0   3582   633
              1     14    11
```

```
[ ] stats.chi2_contingency(crosstab)
```

```
(14.033657261599943,
 0.00017956757859l8809,
 1,
 array([[3574.79716981,  640.20283019],
        [  21.20283019,    3.79716981]]))
```



▾ Prevalent Hyp Chi-square test is significant

```
[ ]  crosstab = pd.crosstab(df['prevalentHyp'], df['TenYearCHD'])
     crosstab
```

TenYearCHD	0	1
prevalentHyp		
0	2604	319
1	992	325

```
[ ]  stats.chi2_contingency(crosstab)
```

```
(132.45628623925992,
 1.1889609489706164e-30,
 1,
 array([[2479.03490566,  443.96509434],
        [1116.96509434,  200.03490566]]))
```

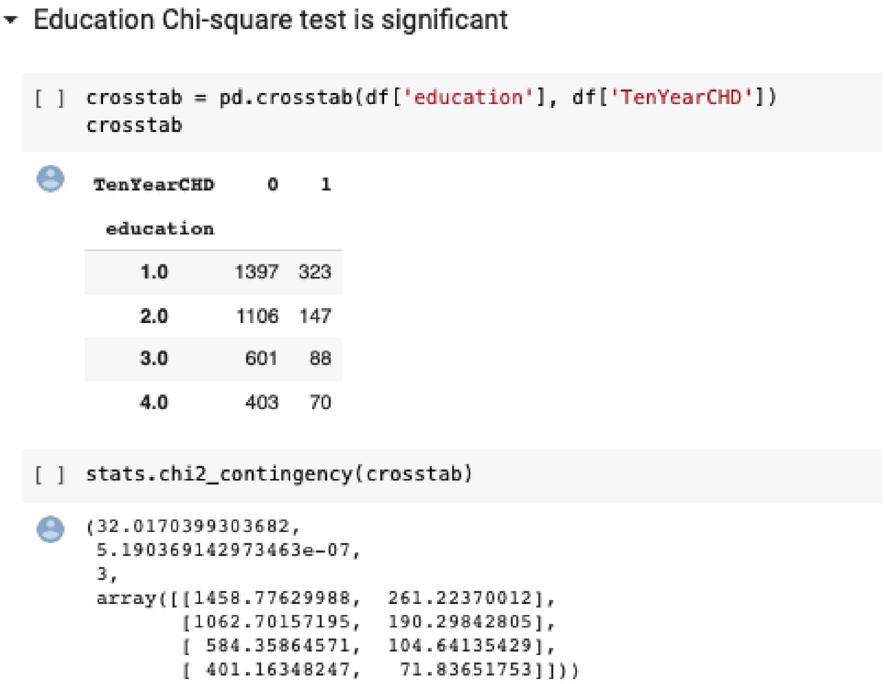
▾ Diabetes Chi-square test is significant

```
[ ]  crosstab = pd.crosstab(df['diabetes'], df['TenYearCHD'])
     crosstab
```

TenYearCHD	0	1
diabetes		
0	3527	604
1	69	40

```
[ ]  stats.chi2_contingency(crosstab)
```

```
(38.48233814115802,
 5.525144036275509e-10,
 1,
 array([[3503.55566038,  627.44433962],
        [  92.44433962,   16.55566038]]))
```



Education Chi-square test is significant

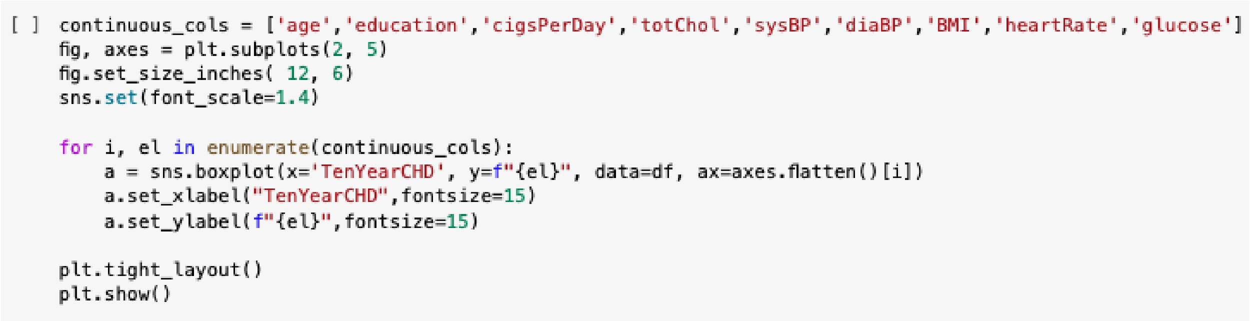
```
[ ] crosstab = pd.crosstab(df['education'], df['TenYearCHD'])
    crosstab
```

TenYearCHD	0	1
education		
1.0	1397	323
2.0	1106	147
3.0	601	88
4.0	403	70

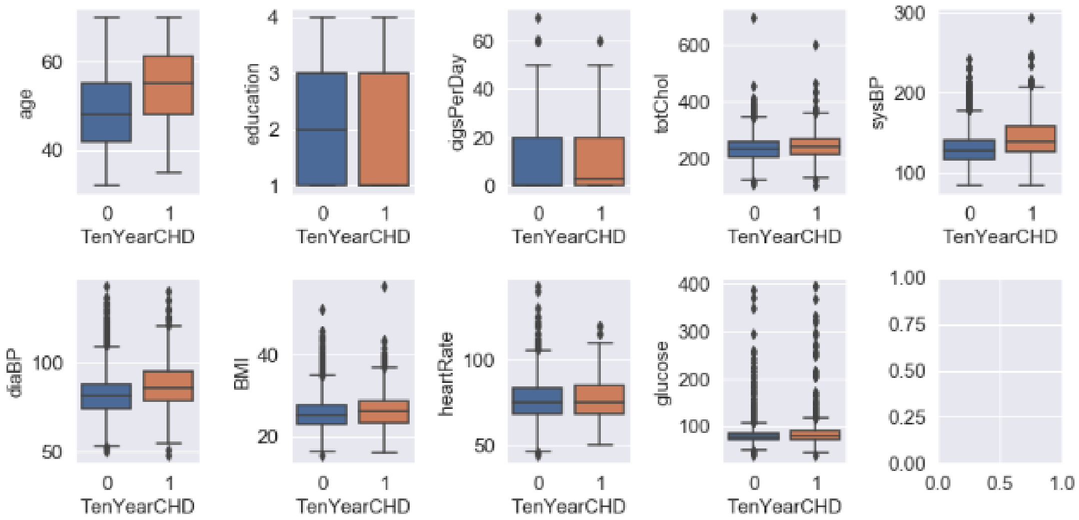
```
[ ] stats.chi2_contingency(crosstab)
```

```
(32.0170399303682,
 5.190369142973463e-07,
 3,
 array([[1458.77629988,  261.22370012],
        [1062.70157195,  190.29842805],
        [ 584.35864571,  104.64135429],
        [ 401.16348247,   71.83651753]]))
```

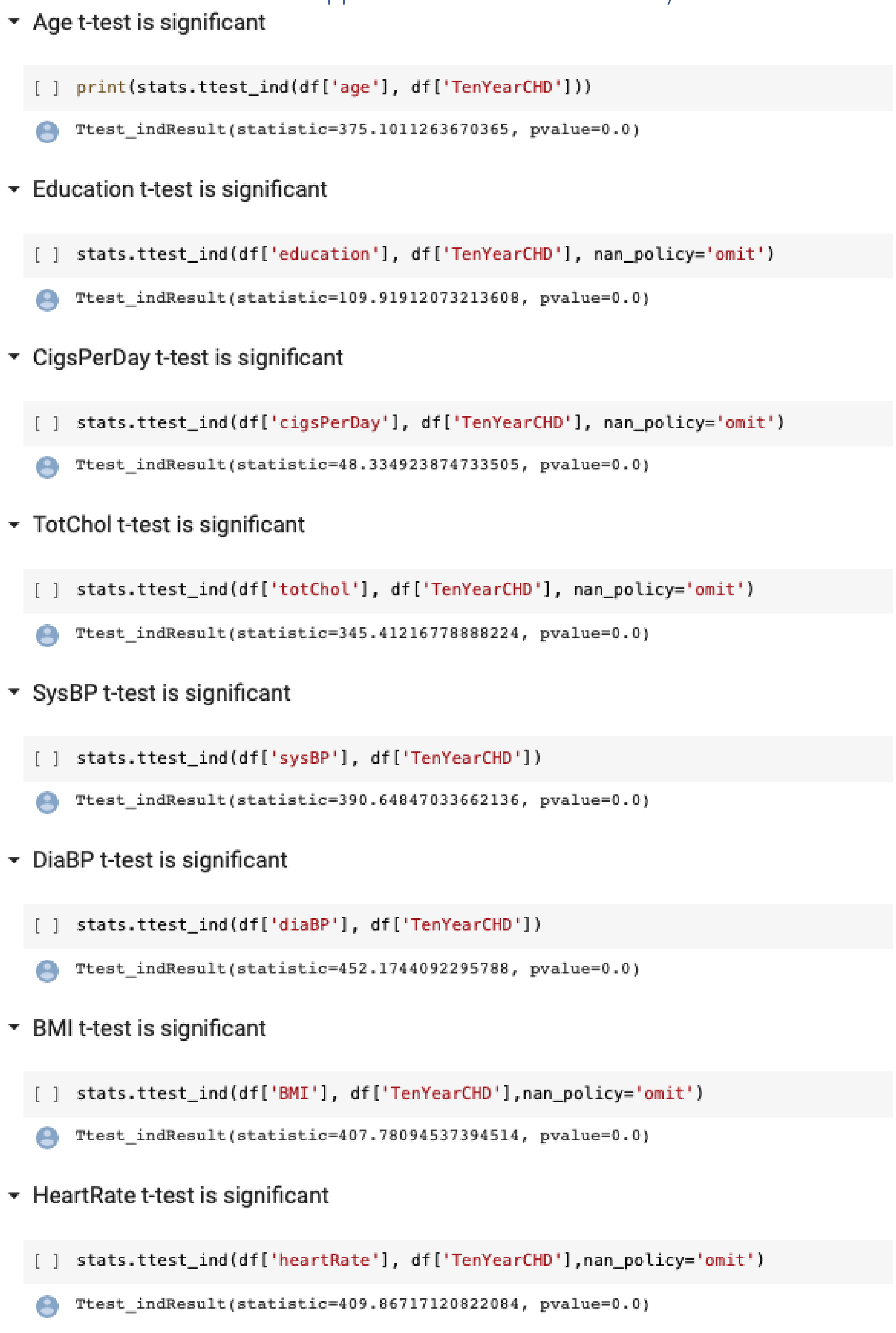
Appendix C: Box Plots of Continuous Features to Target



```
[ ] continuous_cols = ['age','education','cigsPerDay','totChol','sysBP','diaBP','BMI','heartRate','glucose']
    fig, axes = plt.subplots(2, 5)
    fig.set_size_inches( 12, 6)
    sns.set(font_scale=1.4)

    for i, el in enumerate(continuous_cols):
        a = sns.boxplot(x='TenYearCHD', y=f"{el}", data=df, ax=axes.flatten()[i])
        a.set_xlabel("TenYearCHD",fontsize=15)
        a.set_ylabel(f"{el}",fontsize=15)

    plt.tight_layout()
    plt.show()
```



Appendix D: T- test Statistical Python Code



Age t-test is significant

```
[ ] print(stats.ttest_ind(df['age'], df['TenYearCHD']))
```

```
Ttest_indResult(statistic=375.1011263670365, pvalue=0.0)
```

Education t-test is significant

```
[ ] stats.ttest_ind(df['education'], df['TenYearCHD'], nan_policy='omit')
```

```
Ttest_indResult(statistic=109.91912073213608, pvalue=0.0)
```

CigsPerDay t-test is significant

```
[ ] stats.ttest_ind(df['cigsPerDay'], df['TenYearCHD'], nan_policy='omit')
```

```
Ttest_indResult(statistic=48.334923874733505, pvalue=0.0)
```

TotChol t-test is significant

```
[ ] stats.ttest_ind(df['totChol'], df['TenYearCHD'], nan_policy='omit')
```

```
Ttest_indResult(statistic=345.41216778888224, pvalue=0.0)
```

SysBP t-test is significant

```
[ ] stats.ttest_ind(df['sysBP'], df['TenYearCHD'])
```

```
Ttest_indResult(statistic=390.64847033662136, pvalue=0.0)
```

DiaBP t-test is significant

```
[ ] stats.ttest_ind(df['diaBP'], df['TenYearCHD'])
```

```
Ttest_indResult(statistic=452.1744092295788, pvalue=0.0)
```

BMI t-test is significant

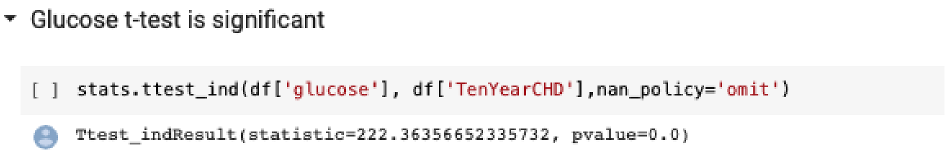
```
[ ] stats.ttest_ind(df['BMI'], df['TenYearCHD'],nan_policy='omit')
```

```
Ttest_indResult(statistic=407.78094537394514, pvalue=0.0)
```

HeartRate t-test is significant

```
[ ] stats.ttest_ind(df['heartRate'], df['TenYearCHD'],nan_policy='omit')
```

```
Ttest_indResult(statistic=409.86717120822084, pvalue=0.0)
```

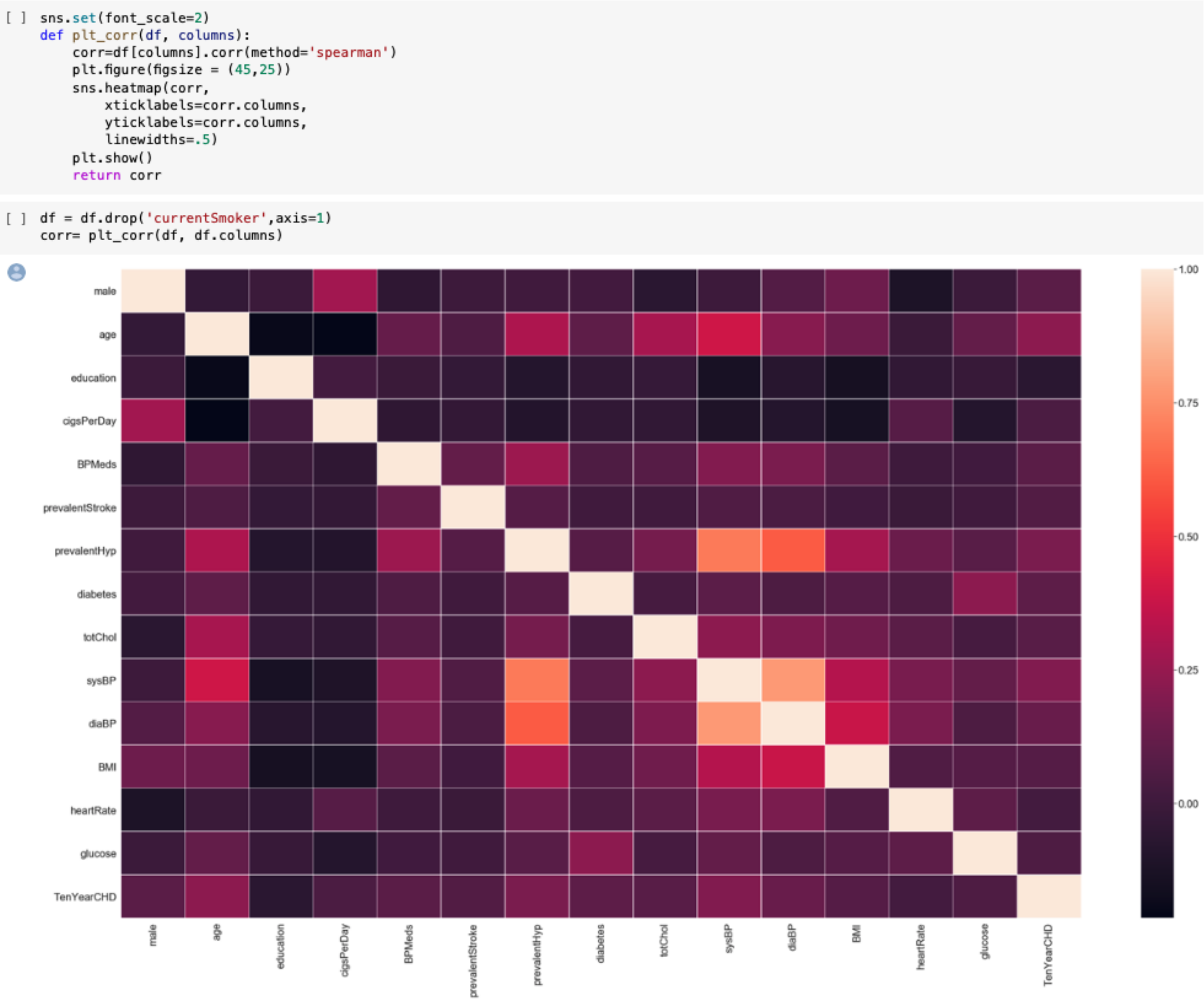


Glucose t-test is significant

```
[ ] stats.ttest_ind(df['glucose'], df['TenYearCHD'],nan_policy='omit')
```

```
Ttest_indResult(statistic=222.36356652335732, pvalue=0.0)
```

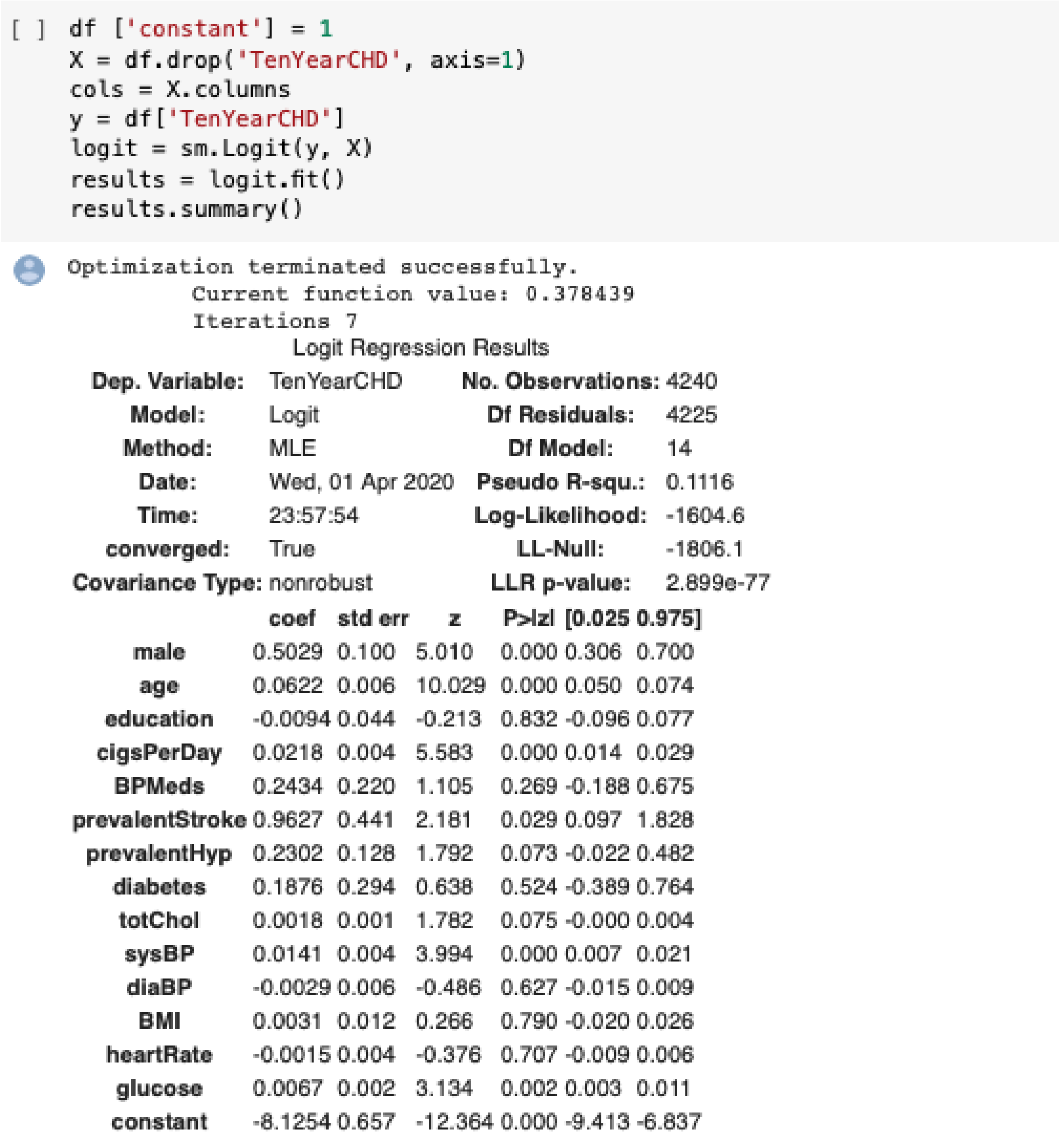
Appendix E: Correlation Analysis of Features Target



```
[ ] sns.set(font_scale=2)
    def plt_corr(df, columns):
        corr=df[columns].corr(method='spearman')
        plt.figure(figsize = (45,25))
        sns.heatmap(corr,
            xticklabels=corr.columns,
            yticklabels=corr.columns,
            linewidths=.5)
        plt.show()
        return corr
```

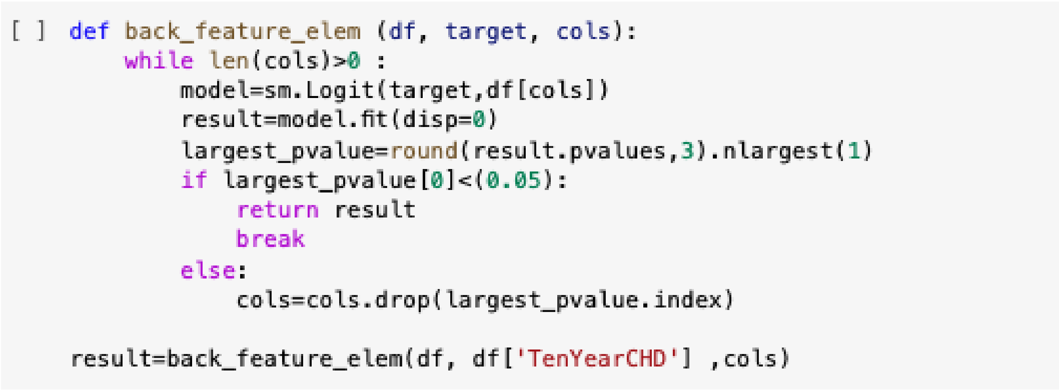
```
[ ] df = df.drop('currentSmoker',axis=1)
    corr= plt_corr(df, df.columns)
```

Appendix F: Backward Elimination

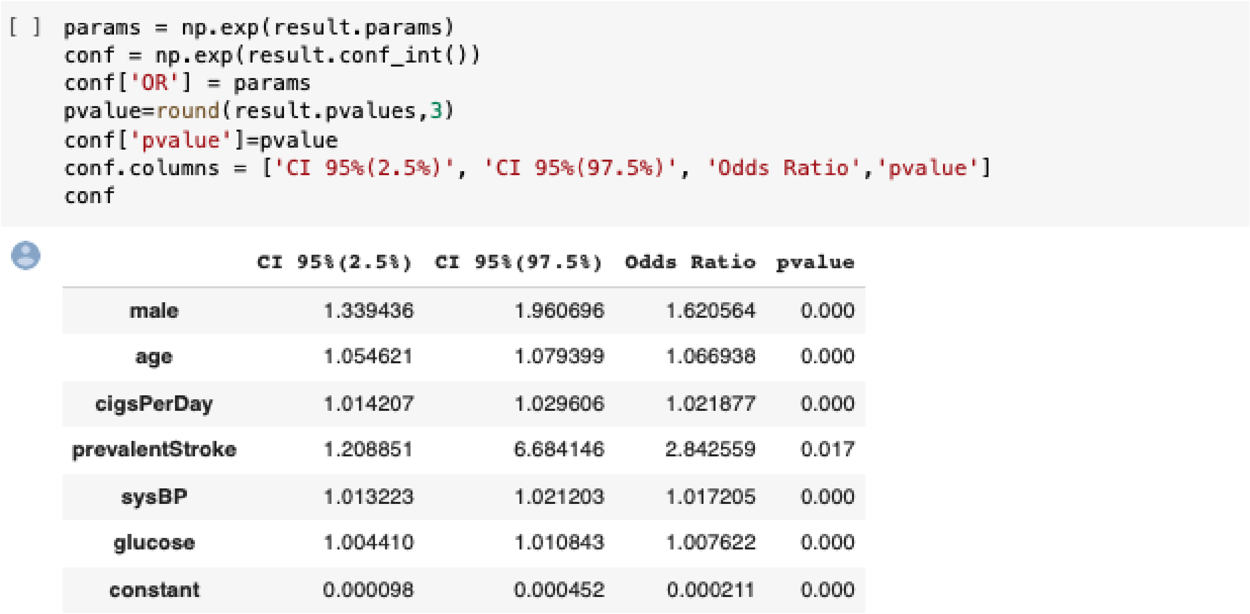


```
[ ] df ['constant'] = 1
    X = df.drop('TenYearCHD', axis=1)
    cols = X.columns
    y = df['TenYearCHD']
    logit = sm.Logit(y, X)
    results = logit.fit()
    results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.378439
         Iterations 7
                       Logit Regression Results
   Dep. Variable:     TenYearCHD        No. Observations: 4240
       Model:         Logit               Df Residuals:   4225
      Method:         MLE                   Df Model:     14
       Date:          Wed, 01 Apr 2020   Pseudo R-squ.:   0.1116
       Time:          23:57:54           Log-Likelihood:  -1604.6
     converged:       True                  LL-Null:      -1806.1
Covariance Type:      nonrobust          LLR p-value:     2.899e-77
                       coef    std err   z       P>|z|  [0.025  0.975]
       male          0.5029   0.100   5.010    0.000  0.306   0.700
        age          0.0622   0.006   10.029   0.000  0.050   0.074
     education      -0.0094   0.044   -0.213   0.832  -0.096  0.077
    cigsPerDay       0.0218   0.004   5.583    0.000  0.014   0.029
      BPMeds         0.2434   0.220   1.105    0.269  -0.188  0.675
 prevalentStroke     0.9627   0.441   2.181    0.029  0.097   1.828
   prevalentHyp      0.2302   0.128   1.792    0.073  -0.022  0.482
     diabetes        0.1876   0.294   0.638    0.524  -0.389  0.764
      totChol        0.0018   0.001   1.782    0.075  -0.000  0.004
       sysBP         0.0141   0.004   3.994    0.000  0.007   0.021
       diaBP        -0.0029   0.006   -0.486   0.627  -0.015  0.009
        BMI          0.0031   0.012   0.266    0.790  -0.020  0.026
     heartRate      -0.0015   0.004   -0.376   0.707  -0.009  0.006
      glucose        0.0067   0.002   3.134    0.002  0.003   0.011
     constant       -8.1254   0.657   -12.364  0.000  -9.413  -6.837
```

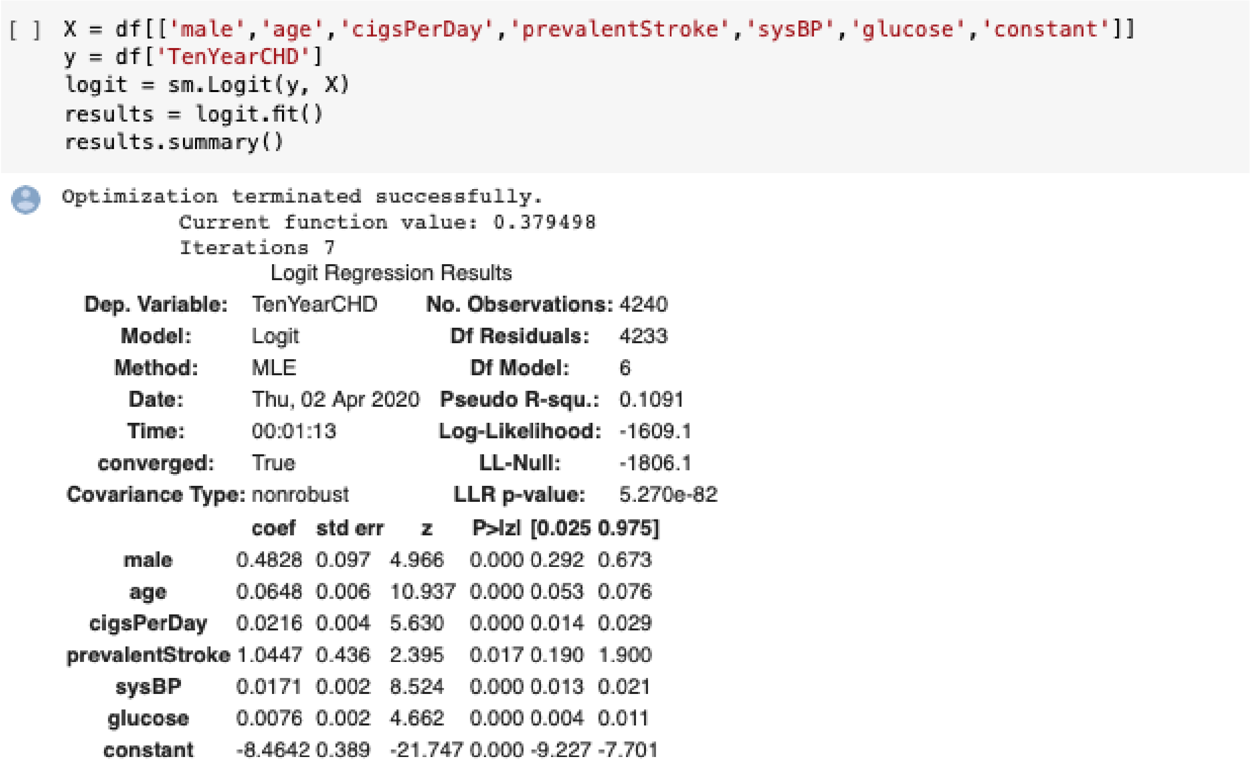


```
[ ] def back_feature_elem (df, target, cols):
        while len(cols)>0 :
            model=sm.Logit(target,df[cols])
            result=model.fit(disp=0)
            largest_pvalue=round(result.pvalues,3).nlargest(1)
            if largest_pvalue[0]<(0.05):
                return result
                break
            else:
                cols=cols.drop(largest_pvalue.index)

    result=back_feature_elem(df, df['TenYearCHD'] ,cols)
```



```
[ ] params = np.exp(result.params)
    conf = np.exp(result.conf_int())
    conf['OR'] = params
    pvalue=round(result.pvalues,3)
    conf['pvalue']=pvalue
    conf.columns = ['CI 95%(2.5%)', 'CI 95%(97.5%)', 'Odds Ratio','pvalue']
    conf
```

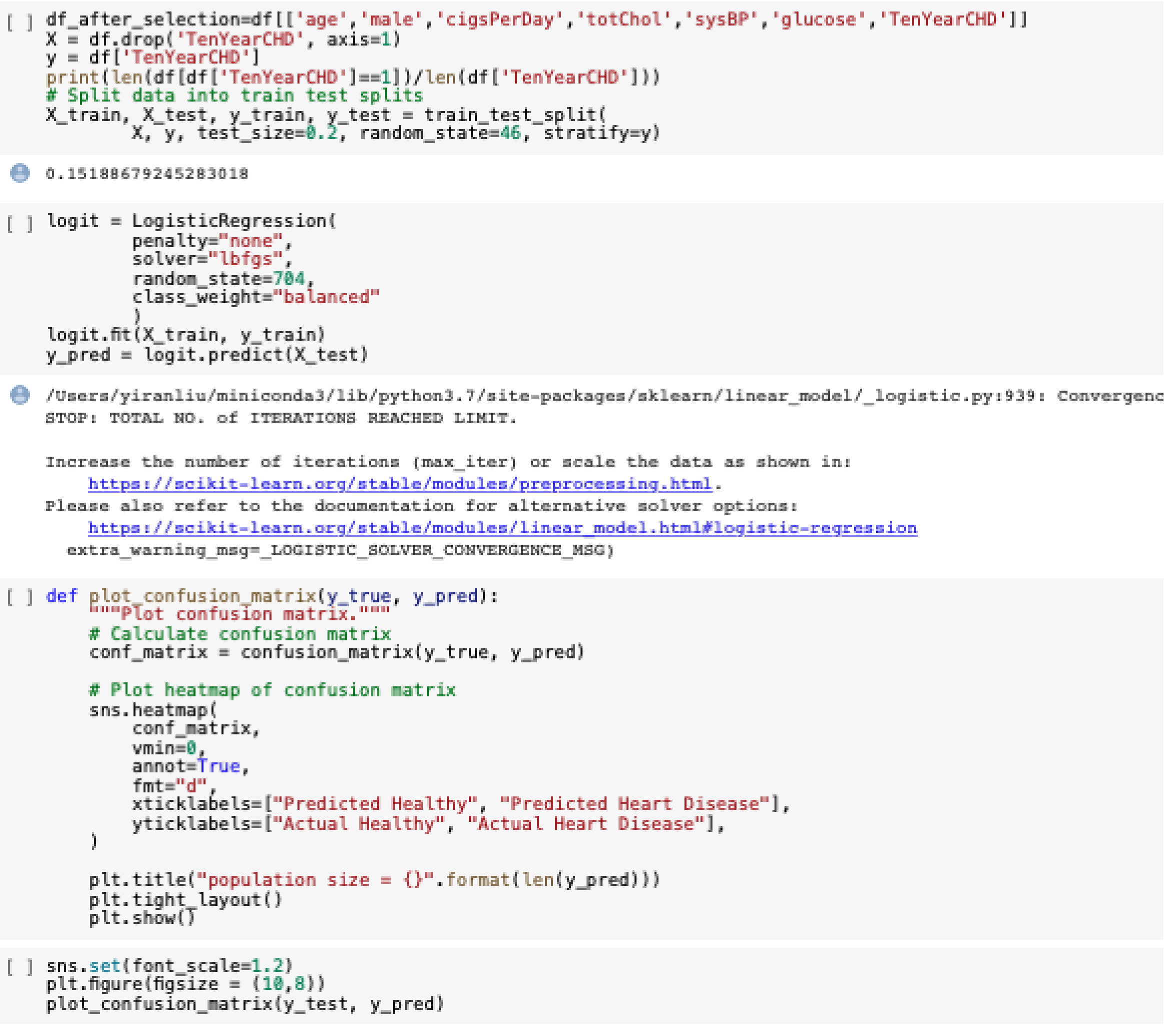
	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
male	1.339436	1.960696	1.620564	0.000
age	1.054621	1.079399	1.066938	0.000
cigsPerDay	1.014207	1.029606	1.021877	0.000
prevalentStroke	1.208851	6.684146	2.842559	0.017
sysBP	1.013223	1.021203	1.017205	0.000
glucose	1.004410	1.010843	1.007622	0.000
constant	0.000098	0.000452	0.000211	0.000



```
[ ] X = df[['male','age','cigsPerDay','prevalentStroke','sysBP','glucose','constant']]
    y = df['TenYearCHD']
    logit = sm.Logit(y, X)
    results = logit.fit()
    results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.379498
         Iterations 7
                      Logit Regression Results
   Dep. Variable:  TenYearCHD        No. Observations: 4240
          Model:   Logit                Df Residuals:  4233
         Method:   MLE                     Df Model:   6
           Date:   Thu, 02 Apr 2020   Pseudo R-squ.:   0.1091
           Time:   00:01:13          Log-Likelihood:   -1609.1
      converged:   True                     LL-Null:   -1806.1
Covariance Type:   nonrobust            LLR p-value:   5.270e-82
                  coef    std err   z       P>|z| [0.025  0.975]
     male         0.4828  0.097    4.966    0.000 0.292   0.673
      age         0.0648  0.006    10.937   0.000 0.053   0.076
  cigsPerDay      0.0216  0.004    5.630    0.000 0.014   0.029
prevalentStroke   1.0447  0.436    2.395    0.017 0.190   1.900
     sysBP        0.0171  0.002    8.524    0.000 0.013   0.021
    glucose       0.0076  0.002    4.662    0.000 0.004   0.011
    constant     -8.4642  0.389   -21.747   0.000 -9.227  -7.701
```

Appendix G: Logistic Regression



```
[ ] df_after_selection=df[['age','male','cigsPerDay','totChol','sysBP','glucose','TenYearCHD']]
    X = df.drop('TenYearCHD', axis=1)
    y = df['TenYearCHD']
    print(len(df[df['TenYearCHD']==1])/len(df['TenYearCHD']))
    # Split data into train test splits
    X_train, X_test, y_train, y_test = train_test_split(
            X, y, test_size=0.2, random_state=46, stratify=y)
```

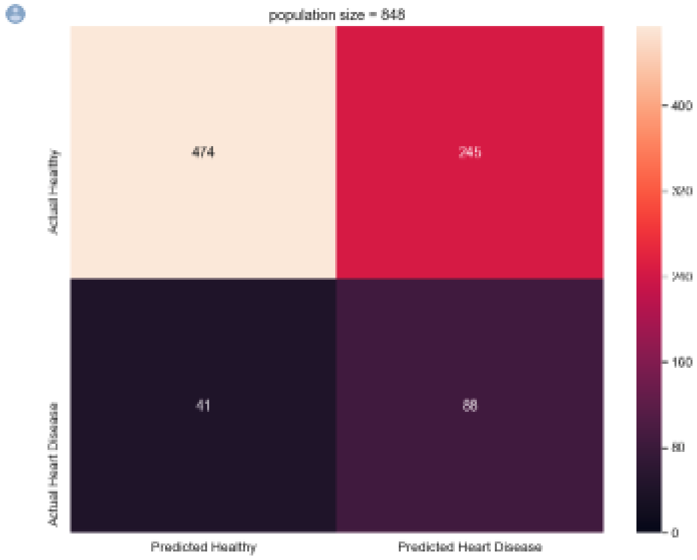
```
0.15188679245283018
```

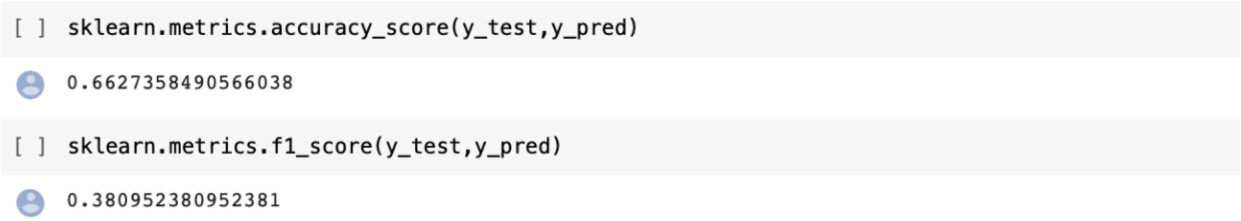
```
[ ] logit = LogisticRegression(
            penalty="none",
            solver="lbfgs",
            random_state=704,
            class_weight="balanced"
            )
    logit.fit(X_train, y_train)
    y_pred = logit.predict(X_test)
```

```
/Users/yiranliu/miniconda3/lib/python3.7/site-packages/sklearn/linear_model/_logistic.py:939: Convergenc
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html.
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
```

```
[ ] def plot_confusion_matrix(y_true, y_pred):
        """Plot confusion matrix."""
        # Calculate confusion matrix
        conf_matrix = confusion_matrix(y_true, y_pred)

        # Plot heatmap of confusion matrix
        sns.heatmap(
            conf_matrix,
            vmin=0,
            annot=True,
            fmt="d",
            xticklabels=["Predicted Healthy", "Predicted Heart Disease"],
            yticklabels=["Actual Healthy", "Actual Heart Disease"],
        )

        plt.title("population size = {}".format(len(y_pred)))
        plt.tight_layout()
        plt.show()
```

```
[ ] sns.set(font_scale=1.2)
    plt.figure(figsize = (10,8))
    plot_confusion_matrix(y_test, y_pred)
```





```
sklearn.metrics.accuracy_score(y_test,y_pred)
```

```
0.6627358490566038
```

```
sklearn.metrics.f1_score(y_test,y_pred)
```

```
0.380952380952381
```