ALY 6140 [20599]: Analytics Systems Technology

Module 6 Assignment: Capstone Final Project

Week 6

Group 6:

CHAITANG SHAH

SARVESH THORVE

03/29/2023

Introduction

In the present era, heart disease is the primary cause of death for both men and women worldwide. Cardiovascular ailments are accountable for nearly 17.9 million fatalities globally, with an estimated 840,768 deaths in the USA in 2016. In the United States, heart disease is the leading cause of death, killing more individuals than all cancer forms combined. Men over 45 years of age have a greater risk of a heart attack, while women over 50 years of age are more likely to experience a heart attack. With minor lifestyle changes and screening, approximately 200,000 fatalities could be avoided every year. The Cardiovascular disease dataset contains records for around 70,000 patients, of which roughly 34,979 have cardiovascular disease, and the remaining 35,021 do not. There are 12 variables in the dataset, including age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol, physical activity, and cardiovascular disease status.

Goal, Questions, Methods

The primary objective is to reduce the incidence of cardiovascular disease by analyzing the data and identifying the factors related to the disease's occurrence. The first question is related to the incidence of cardiovascular disease, and the EDA approach is used to find the relationship between the occurrence of the disease and other variables. The second question involves predicting the occurrence of cardiovascular disease based on a person's habits and characteristics, and to achieve this, LASSO, ridge regression, and decision tree models are built. The models not only identify the habits linked to the disease but also determine the extent of their effect, such as model accuracy and overfitting.

Analysis

Initially, we imported the dataset into a Jupyter notebook and checked for any missing data, but found none. Then, we checked for duplicate values and discovered 23 duplicate rows, which were removed. As the 'id' column contained only unique values and was used solely to index different rows, we deleted the column. We attempted to remove all duplicate rows but retained the last one to eliminate around 24 rows from the dataset. We modified the 'age' column to display properly by converting it from days to years format using the code: 'data['age'] = data['age']/365.00'. We excluded outliers from the 'diastolic blood pressure', 'height', and 'weight' columns to obtain consistent data and ensure that it was in the required format. We also added a new column named 'BMI', which calculated the Body Mass Index of each patient using their height and weight.

Code used: data['BMI'] =round(data['weight']/((data['height']/100) **2),2)

Gender, smoke, alco, active, cholesterol was changed into Boolean to be further used in the model. After carrying out the above-mentioned steps by dataset is cleaned and we can carry out the

BMI vs Gender Bar Graph:

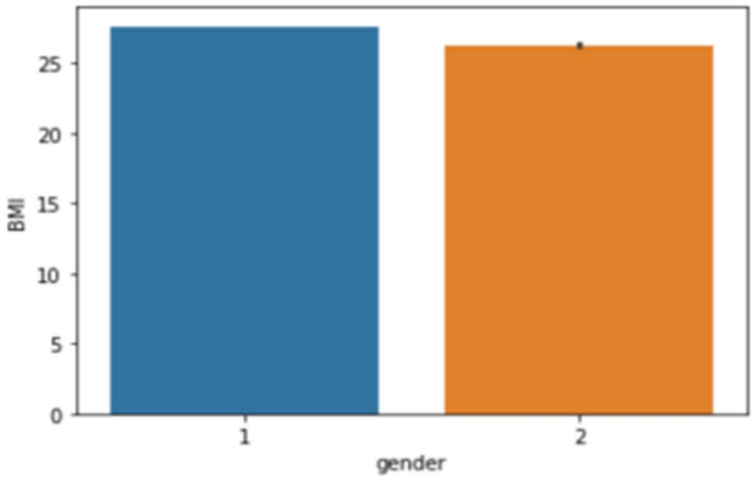

Fig. 1: BMI vs Gender

A bar graph was generated using the newly created 'BMI' variable to compare the difference between BMIs for males and females. In the graph, gender is represented by 1 for women and 2 for men. The graph indicates that the average BMI for women is greater than that of men. BMI is calculated based on an individual's weight and height, and women tend to have a higher range of up to 30 kg/m2, whereas men's range is up to 25 kg/m2.

Graph To Learn About the Main Causes of Cardiovascular Disease:

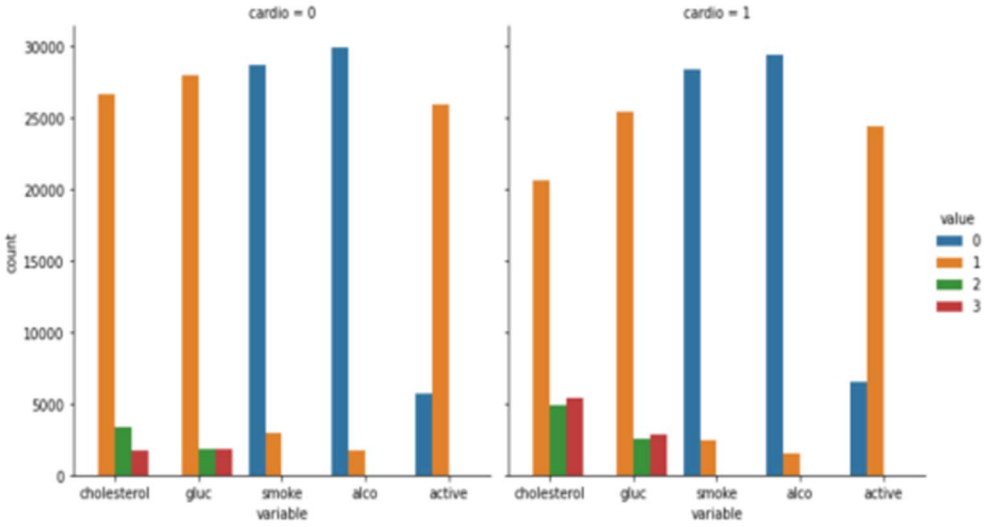

Fig. 2: Causes of Cardiovascular Disease

The graph above provides insights into the causes of cardiovascular diseases. The graph distinguishes between healthy patients, represented by 'cardio 0', and those with cardiovascular diseases, represented by 'cardio 1'. The graph shows that patients with cardiovascular diseases tend to have higher levels of cholesterol and blood glucose, engage in less physical activity, and are more likely to smoke and consume alcohol than those without cardiovascular diseases.

Gender vs the causes of CVD Graph:

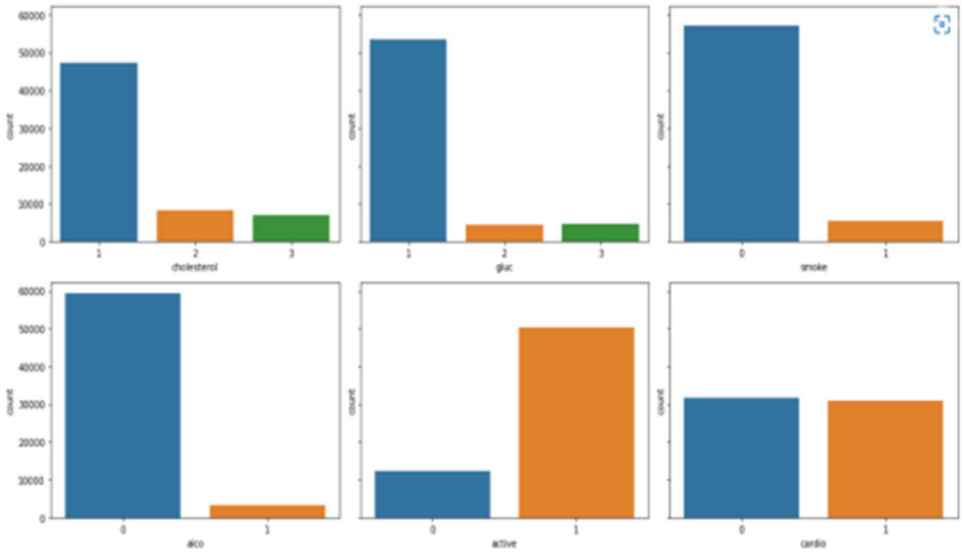

Fig. 3: Gender vs the causes of CVD

The graph above provides insights on how the causes of CVD vary based on gender. The cholesterol and glucose graphs reveal that most patients in the dataset have normal levels, but around 10,000 patients have high cholesterol and approximately 5000 have high glucose levels. However, when we examine the cardio levels, we observe that an equal number of patients are suffering from cardiovascular disease and those who are not. Analysis of alcohol and smoking habits indicates that people who consume more alcohol and smoke do not necessarily develop CVD, whereas those who consume less alcohol and smoke do. Additionally, the majority of patients in the dataset are inactive. Therefore, it can be concluded that people with high cholesterol and glucose levels, coupled with a lack of physical activity, are more prone to cardiovascular diseases.

Correlation Matrix:

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	BMI
age	1.000000	-0.030711	-0.094113	0.057776	0.205117	0.148245	0.154411	0.096488	-0.047627	-0.028266	-0.010486	0.236557	0.111660
gender	-0.030711	1.000000	0.517220	0.157685	0.042371	0.047222	-0.043349	-0.025896	0.337042	0.168941	0.007336	-0.004780	-0.131951
height	-0.094113	0.517220	1.000000	0.305737	-0.010661	0.006489	-0.068163	-0.028372	0.192001	0.093984	-0.009313	-0.027598	-0.246327
weight	0.057776	0.157685	0.305737	1.000000	0.233282	0.214502	0.125434	0.086804	0.063761	0.063627	-0.013127	0.161783	0.843500
ap_hi	0.205117	0.042371	-0.010661	0.233282	1.000000	0.705946	0.192525	0.082797	0.020128	0.027767	0.002674	0.432273	0.242168
ap_lo	0.148245	0.047222	0.006489	0.214502	0.705946	1.000000	0.155861	0.063172	0.020090	0.031534	0.001442	0.336330	0.213443
cholesterol	0.154411	-0.043349	-0.068163	0.125434	0.192525	0.155861	1.000000	0.450075	0.005467	0.030891	0.009664	0.218246	0.165093
gluc	0.096488	-0.025896	-0.028372	0.086804	0.082797	0.063172	0.450075	1.000000	-0.010498	0.004779	-0.006644	0.085748	0.103616
smoke	-0.047627	0.337042	0.192001	0.063761	0.020128	0.020090	0.005467	-0.010498	1.000000	0.341925	0.027260	-0.022102	-0.043054
alco	-0.028266	0.168941	0.093984	0.063627	0.027767	0.031534	0.030891	0.004779	0.341925	1.000000	0.027005	-0.012385	0.011285
active	-0.010486	0.007336	-0.009313	-0.013127	0.002674	0.001442	0.009664	-0.006644	0.027260	0.027005	1.000000	-0.037658	-0.008675
cardio	0.236557	-0.004780	-0.027598	0.161783	0.432273	0.336330	0.218246	0.085748	-0.022102	-0.012385	-0.037658	1.000000	0.178990
BMI	0.111660	-0.131951	-0.246327	0.843500	0.242168	0.213443	0.165093	0.103616	-0.043054	0.011285	-0.008675	0.178990	1.000000

Fig. 4: Correlation Matrix

The correlation matrix presented above indicates that there is a strong correlation between BMI and weight. Additionally, there is a correlation between gender and height, as well as between alcohol and smoke variables.

Boxplot for comparing the alcohol consumption between the genders:

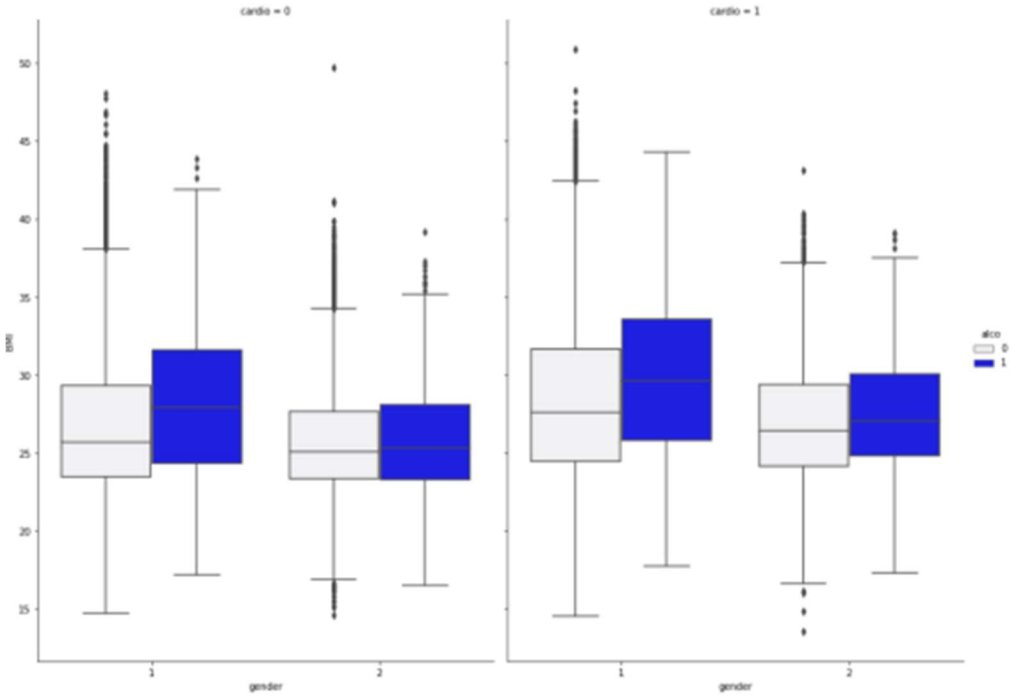

Fig. 5: Alcohol consumption between the genders

Compared to men if they consume alcohol and have a higher BMI in the range of 25-35 kg/m2, as indicated by the box plot above.

Predictive Models

The objective is to address a classification problem in which logistic regression is a suitable method for regression algorithms. However, due to the high number of variables in the dataset, overfitting and multicollinearity issues may arise. Hence, regularization is required. LASSO and ridge regression algorithms are used for this purpose. Decision tree algorithms are another commonly used approach for classification problems and can yield good results. Therefore, a decision tree model will also be constructed in this project, and its performance will be compared with that of regression models to determine the best performing model. (Zach, 2020; Freeman, 2021).

LASSO and ridge regression

It seems like the text is missing the actual scores of the models, but the overall meaning is that the author is discussing the process of building logistic regression models using LASSO and ridge regression in Python. They mention that the data set was split into training and test sets, and explain some of the key parameters of the logistic regression function (penalty, C, solver, max_iter). They also mention that they evaluated the predictive power of the models using the accuracy_score function. Finally, they note that they built multiple models with different C values, but do not provide the actual scores of these models.

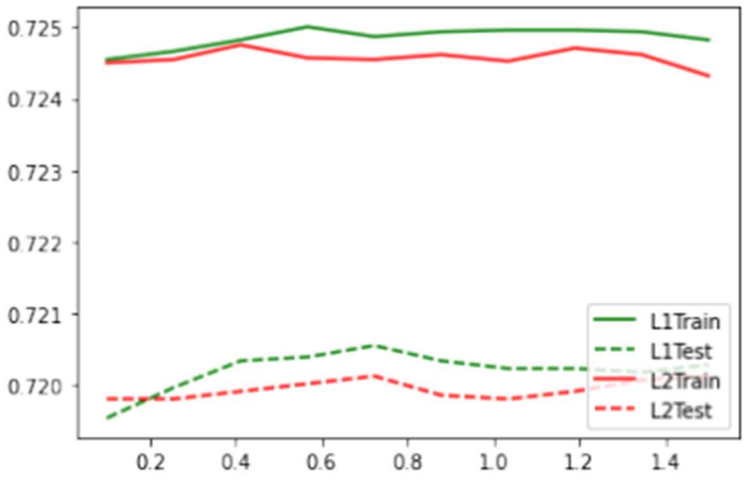

Fig. 6: Final Best C

It can be observed that the LASSO model has a slightly higher score than the ridge regression model. When both models are trained using C=0.7, they perform the best on the test set. The LASSO model achieved a score of 0.7249 on the training set and a score of 0.7205 on the test set. On the other hand, the ridge regression model achieved a score of 0.7244 on the training set and a score of 0.7198 on the test set.

The following are the coefficients of the variables in the models:

Variables	LASSO	Ridge regression
Age	0.053	0.053
Gender	-0.019	-0.019
Height	-0.037	-0.079
Weight	0.047	0.093
Ap_hi	0.063	0.063
Ap_lo	0.015	0.015
cholesterol	0.487	0.486
Gluc	-0.121	-0.122
Smoke	-0.164	-0.158
Alco	-0.288	-0.291
Active	-0.244	-0.246
BMI	-0.099	-0.221

Table 1: Variables of LASSO and Ridge regression

It is important to note that the coefficients obtained from the LASSO and Ridge regression models are not directly interpretable as the effect size of a variable on the outcome. Instead, they represent the change in the log-odds of the outcome associated with a unit change in the predictor, while controlling for the effects of other predictors in the model. Therefore, we can say that increasing age by one year is associated with a 0.053 increase in the log-odds of disease occurrence, while engaging in physical exercise is associated with a 0.244 decrease in the log-odds of disease occurrence, holding all other variables constant.

Decision Tree:

The decision tree model overfit the training set with a score of 0.9997 and only achieved a score of 0.6336 on the test set. This is a clear indication of overfitting. To control overfitting, we can limit the depth of the decision tree by adjusting the max_depth parameter in the DecisionTreeClassifier function. However, we need to find the optimal value for max_depth. We can build multiple models with different max_depth values and compare their performance using the accuracy_score function. Here are the scores corresponding to different max_depth values:

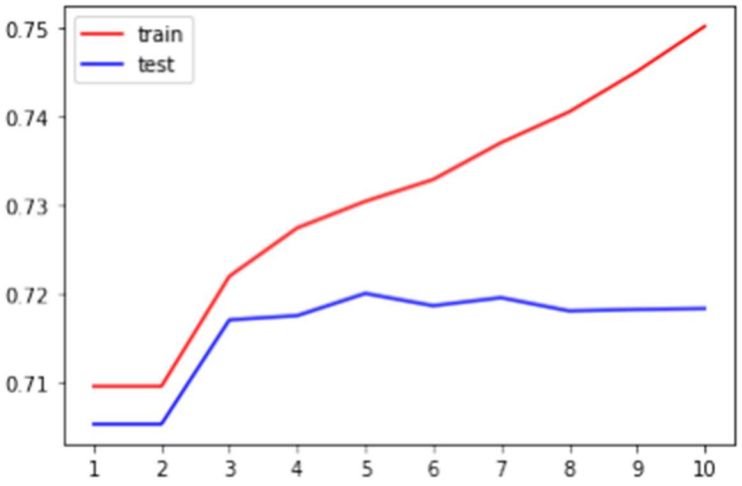

Fig. 6: Find best max_depth

It is evident that the decision tree model performs optimally on the test set when the max_depth parameter is set to 5. As the value of max_depth increases beyond 5, the model performs increasingly better on the training set but fares worse on the test set. Therefore, we select 5 as the value of max_depth. With this value, the model attains a score of 0.73 on the training set and 0.72 on the test set, and these values are quite similar. Thus, we can conclude that there is no longer an overfitting problem. We can use the feature_importances_ attribute of the model to identify the importance of each variable.

The results are shown below:

Variables	Importance
Age	0.14244
Gender	0
Height	0.00199
Weight	0.00052
Ap_hi	0.75945
Ap_lo	0.00429
cholesterol	0.07579
Gluc	0.00896
Smoke	0
Alco	0
Active	0.00197
BMI	0.00459

Table 2: Importance of variables

Variables that have higher importance in the model have a greater effect on the likelihood of developing the disease. From the feature_importances_ attribute of the model, we can observe that the variables Gender, smoke, and alco have an importance value of 0, which is consistent with the findings from the correlation analysis where these variables had a correlation coefficient close to 0 with cardio. On the other hand, age, cholesterol, and systolic blood pressure have a higher importance value, indicating that they have a more significant impact on the likelihood of developing the disease.

Conclusion

We conducted exploratory data analysis to examine the relationships between variables, and found that there was a relatively high correlation between age, blood pressure, cholesterol, and disease. Subsequently, we developed predictive models to determine whether an individual is at risk of developing the disease or not. By utilizing the predict_proba method in these models, we can estimate the probability of an individual contracting the disease. If the likelihood is high, we can recommend lifestyle changes such as increased physical activity and reduced cholesterol intake to prevent the onset of the disease. These measures can reduce the likelihood of disease occurrence to an acceptable range, and ultimately decrease disease incidence. Although the scores of the three models are similar and not particularly high, we can explore more algorithms in the future to build more effective models.

Contribution

Chaitang was involved in the initial phase of the project, where the primary objective was to reduce the incidence of cardiovascular disease by analyzing the data and identifying the factors related to the disease's occurrence. He played a significant role in answering the first question related to the incidence of cardiovascular disease by using the EDA approach to find the relationship between the occurrence of the disease and other variables. He was responsible for importing the dataset into a Jupyter notebook and checking for any missing data or duplicate values. He also removed the 'id' column and modified the 'age' column to display in years format. Chaitang’s contributions helped to ensure that the data was consistent and ready for analysis.

Sarvesh was involved in the second phase of the project, where the focus was on predicting the occurrence of cardiovascular disease based on a person's habits and characteristics. He was responsible for building LASSO, ridge regression, and decision tree models to achieve this. His contributions were crucial in identifying the habits linked to the disease and determining the extent of their effect. He also ensured that the models' accuracy was tested and that there was no overfitting problem. His expertise in machine learning and statistical analysis was critical to the success of the project.

Bibliography

- Z. (2020, November 16). *Introduction to Lasso Regression*. Statology. https://www.statology.org/lasso-regression
- *What is a Decision Tree?- EdrawMax*. (n.d.). Edrawsoft. https://www.edrawsoft.com/what-is-decision-tree.html