COMP 562 Final Project Report

May 5, 2022

Zhuo Yao Tan (PID: 730550548), Xin Yue Zhang (PID: 730549984), Lingshan Zhao (PID: 730550553), Tobias Østmo Hermansen (PID: 730549377)

1. Introduction

Our group has decided to analyse a 2020 annual CDC survey data of 400,000 adults on their personal key indicators of heart disease. The dataset is extracted from Kaggle [1], which has 18 variables reduced from nearly 300 variables and is suitable for applying machine learning algorithms. The purpose of our project is to determine if a patient is likely to get heart disease based on key indicators. With heart disease as one of the main causes of death for people of most races in the United States of America [2], our group believes that it is of paramount importance to ascertain the likelihood of people getting heart diseases in their lifetime. The source code for data cleaning, data preprocessing as well as model training is uploaded on our GitHub repository at this link: https://github.com/xyzhangg/COMP562-Project.

2. Data Cleaning

First, we converted all the string data into factors. This includes questions on gender, race, age category, general health, as well as all questions with "Yes/No" as answers. The responses that can be measured on a scale are given numerical values. For example, "No" and "Yes" are represented as 0 and 1 respectively, while responses for questions such as that for general health ranges from "Poor" to "Excellent" and are therefore represented as values from 0 to 4. As for questions with responses that cannot be measured on a scale, such as age category and race, extra dummy variables are created to represent each of the different answers to those questions. As a result, age category is represented by 13 variables while race is represented by 6 variables. Lastly, we split the data into training and test data with a ratio of $80:20$, and set the variable "Heart Disease" as the response variable.

3. Data Balancing

It is crucial to note that the classes are rather imbalanced. The number of persons without heart disease is an estimated 11 times of those with heart disease. In order to address this imbalance, we decided to use Synthetic Minority Oversampling Technique (SMOTE) [3]. SMOTE selects a sample randomly from the minority class and determines its k-nearest neighbours, then selects one of the neighbours to calculate the Euclidean distance to it from the sampled point. The distance vector is then multiplied by a random number between 0 to 1 and added to the sampled point to create a new minority class observation. Using this method, we were able to synthesise data entries to balance out the number of positive and negative heart disease cases.

4. Data Modelling

4.1 Linear Regression

We trained three different linear regression models: Lasso, Ridge, and Elastic Net. These models were first trained using the initial unbalanced data set and the results are shown below.

Lasso	Ridge	Elastic Net
Actual Predicted 0 1 0 58047 4937 1 444 531	Actual Predicted 0 1 0 58098 4997 1 393 471	Actual Predicted 0 1 0 58084 4983 1 407 485

Model	Accuracy	Sensitivity	Specificity
Lasso	91.5868	9.711046	99.24091
Ridge	91.57273	8.613753	99.3281
Elastic Net	91.57273	8.869788	99.30417

Although the accuracy results are rather high, we cannot conclude that the models are indeed accurate since predicting "No" all the time would produce a similar level of accuracy due to the imbalanced data. Thus, we balanced the data using SMOTE and trained the three models again, with the following results.

Lasso	Ridge	Elastic Net
Actual Predicted 0 1 0 44243 1278 1 14248 4190	Actual Predicted 0 1 0 44472 1297 1 14019 4171	Actual Predicted 0 1 0 44233 1270 1 14258 4198

4.2 Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis model assumes that classes follow a Gaussian distribution, hence we need to do some additional data preprocessing before generating the model. All of the binary variables were discarded and we chose the variables for BMI, PhysicalHealth, MentalHealth, GenHealth, and SleepTime to train the model.

The results of the model are shown here:

Actual		
Predicted	0	1
0	37836	1621
1	20655	3847

4.3 Quadratic Discriminant Analysis (QDA)

The Quadratic Discriminant Analysis model requires the same assumption as the LDA model, hence we chose the same variables to train the model.

The results of the model are shown here:

Actual		
Predicted	0	1
0	47076	2777
1	11415	2691

4.4 K-Nearest Neighbours (kNN)

The K-Nearest-Neighbours model was trained with a value of K = 9 and the results are shown here:

Actual		
Predicted	0	1
0	53419	4255
1	5072	1213

5. Results and Conclusion

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Lasso	75.72507	76.62765	75.6407
Ridge	76.05341	76.28018	76.03221
Elastic Net	75.72195	76.77396	75.6236
LDA	65.17144	70.35479	64.68687
QDA	77.81079	49.21361	80.48418
kNN	85.41722	22.18361	91.32858

From the table above, the Ridge Regression model has the highest percentages for all three metrics (Accuracy, Sensitivity, Specificity) that were utilized to analyze the models. This thus serves as evidence that out of the six models used, the Ridge Regression model proves to be the most applicable and effective model for determining how likely a person will get heart diseases in their lifetime.

It is important to note that from the perspective of healthcare, having a higher false positive rate is significantly better than having a higher false negative rate. Therefore in the context of healthcare, sensitivity is a better performance indicator than specificity. It is much safer to misdiagnose a person who does not actually have heart disease with the condition, as opposed to misdiagnosing a person who actually has heart disease as without the condition. It is more costly to misclassify a potential heart disease patient as negative than to misclassify a healthy patient as positive.

On further analysis, we note that the K-Nearest Neighbors model has the highest accuracy and specificity amongst the six models. However, it also has by far the lowest sensitivity and is significantly lower than all the other models. As mentioned earlier that sensitivity is the best performance indicator in this context, this thus proves that the K-Nearest Neighbors model is not very functional.

6. References

1. Pytlak, K. (2022, February 16). Personal key indicators of heart disease. Kaggle. Retrieved May 3, 2022, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

2. Centers for Disease Control and Prevention. (2019, December 9). Know your risk for heart disease. Centers for Disease Control and Prevention. Retrieved May 3, 2022, from https://www.cdc.gov/heartdisease/risk_factors.htm

3. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2011, June 9). Smote: Synthetic minority over-sampling technique. arXiv.org. Retrieved May 2, 2022, from https://arxiv.org/abs/1106.1813