A Data Mining Analysis of Heart Diseases Factors

Exploring factors that affect heart disease using the BRFSS dataset

Ravan SADIGLI
Department of Computer engineering

Akdeniz University
Antalya, Turkey

***Abstract*— A heart attack occurs when blood flow to the heart is blocked. Heart failure occurs when the heart cannot pump enough blood to meet your body's needs. The factors that affect heart disease and heart attack are roughly the same. Causes of heart disease or attack, high blood pressure, age, cholesterol, diabetes, smoking, etc. We want to find out if demographic, social, and behavioral factors also play a role. In this study, people with heart disease and heart attack, we present a survey of relevant behavioral risk factors. This study applies data mining techniques to the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset with more than 250,000 observations with 22 variables. The results show that some of these variables and some associations between heart disease may help prevent heart disease.**

Keywords—heart disease, heart attack, classification, ensemble, KNN, random forest, bagging, CVD

I. INTRODUCTION

The heart is one of the human body's most important organs. The general term used to cover malfunctions of the heart is Heart Disease, or sometimes Cardiovascular diseases. Though there are multiple forms of heart disease, this paper focuses on the two most common: Heart Attack and Heart Failure. Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. 85% of these deaths were due to heart attacks and strokes. More than three-quarters of CVD deaths occur in low- and middle-income countries. Untreated heart disease can result in emergency visits, hospitalization, and death. Therefore, focused efforts are needed to reduce the social and economic consequences of heart disease.

There are several initiatives around the world that aim to reduce the difficulties caused by CVD. One of the keys is to identify the population at risk to appropriately allocate healthcare resources and initiatives. Numerous research findings have linked various socio-demographic factors with a high prevalence of CVD. Important factors include high blood pressure, age, cholesterol, diabetes, smoking, physical activity, etc.

We used data from the 2015 Behavioral Risk Factor Surveillance System to examine the effect of various demographic, social and behavioral factors on CVD prevalence in a representative sample of U.S. adults nationwide. We applied various data mining techniques, including neural networks, knn, random forest classifier, ada boost classifier, and bagging classifier, in order to choose the best model that predicts heart disease prevalence.

II. ANALYSIS

A. Data Source

The application of data mining techniques to predict risk factors for certain diseases has been gaining popularity in the medical literature over the past few decades. This phenomenon is primarily due to the efficiency and simplicity of the techniques developed in the field of data mining.

In this study, the 2015 BRFSS dataset was used. BRFSS is a nationwide random dial telephone survey for US adults aged 18+. BFRSS data are weighted to adjust for the probability of participant selection and subsequently stratified to reflect the age and gender distribution of the general population. Interviews are conducted via landline and cell phones and include data for the 50 US states, the District of Columbia, Guam, and Puerto Rico.

B. Analysis

The goal of the analysis was to identify risk factors associated with heart disease and assess the effect of each independent variable on target variables. Specifically, neural networks, KNN, Ada boost, random forest, and bagging classifications are applied to the binary classification task of predicting heart disease based on survey responses to 21 of the questions. We compared five models in order to identify the model that predicted heart disease with higher accuracy. Out of the 253,680 survey responses collected from the 2015 BRFSS survey for use in this study, 23,893 had heart disease or had previously had a heart attack. The remainder of the survey respondents does not have heart disease and this large difference causes the imbalance class as illustrated Figure1.

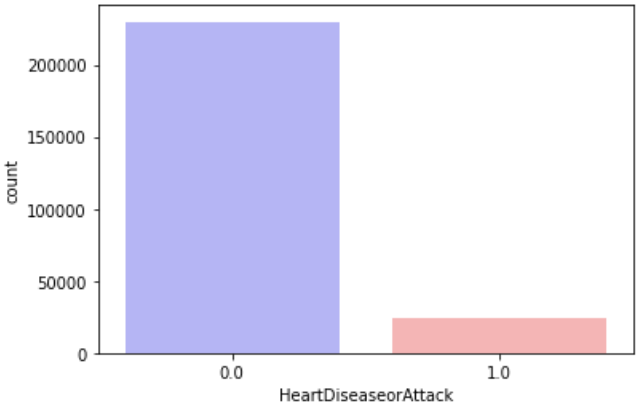

Figure 1: Number of people with and without the disease

Class imbalance causes poor performance of the model through training and testing. To avoid class imbalance, we followed three strategies to deal with this issue. For the first,

we split the dataset into two, those with and without heart disease have the same number of rows, thus eliminating the class imbalance problem. Second, to achieve high accuracy for 5 different models, we split the dataset into 60% and 40% without heart disease and with heart disease, respectively. For the third, we randomly split the dataset without known features.

C. *Analysis by gender*

From Figure 2, we can see that men with heart disease have a higher probability of having heart disease than women.

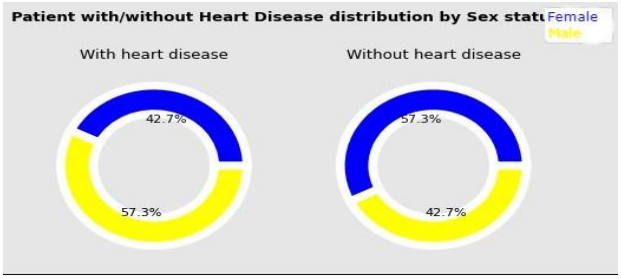

Figure 2: Pie chart for Gender

This has an important difference and needs to be taken into account. Women's bodies are pear-shaped, while men's bodies are usually apple-shaped. When women gain weight, it usually goes down to their hips and thighs. Also, men do not have estrogen protection. Estrogen can keep women's cholesterol levels in check and reduce a major heart disease risk factor.

D. *Analysis by smoker*

Smoking is the well-known factor that affects the heart badly. The question given to those surveyed said they had smoked at least 100 cigarettes in their lifetime. This may raise suspicion, but any cigarette smoked can weaken the heart, regardless of age group. From Figure 3,

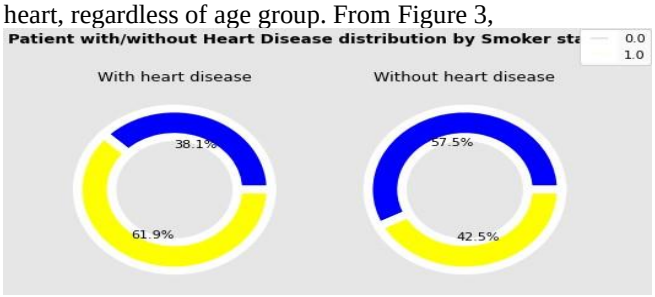

Figure 3: Pie chart for Smoking

we can easily see that 62% of people with heart disease is smoker. Smoking is empowering to narrow the coronary arteries, which causes a heart attack.

E. *Analysis by Alcohol Consumption*

Heavy alcohol consumption can cause high blood pressure, heart disease, or heart attack. Heavy alcohol consumption can also contribute to the disorder that affects the heart muscle. Alcohol can also contribute to obesity and the long list of health problems that can come with it.

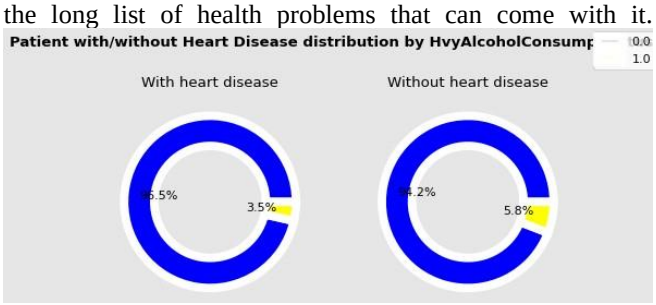

Figure 4: Pie chart for Alcohol

However, as shown in Figure 4, We can easily see that the entire population consumes less alcohol. In the question given to the participants, it was said that adult men drank more than 14 drinks per week and adult women more than 7 drinks per week. That's why heavy alcohol use is less than expected. From Figure 4, we can see that 96.5% of those with heart disease do not consume heavy alcohol in their normal lives. Only 3.5% drink heavily and have heart disease. We can say that heavy alcohol consumption does not affect heart disease much, if we observe who consumes heavy alcohol and heart disease, it contains very little of the data, as illustrated in figure 4. In addition, we should not forget that the problem in the dataset is only that alcohol is predominantly consumed. However, if we compare smoke and alcohol consumption from the pie chart, we can see that smoking affects heart disease more.

F. *Analysis by Stroke*

As in Figure 5, those with stroke and heart disease included 16.5% of the data set.

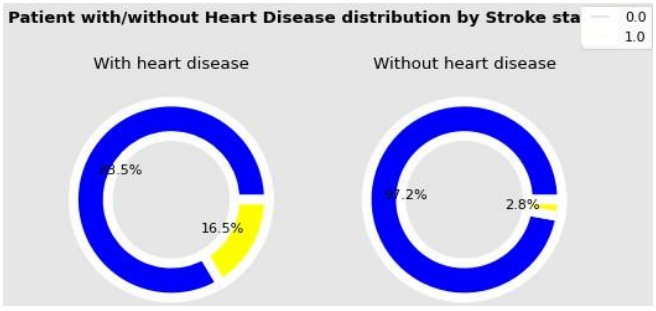

Figure 5: Pie chart for Stroke

This seems like a small part, but we can see that there are also fewer people who have had a stroke from the dataset. But these observations mean that those who have had a stroke are much more likely to getting heart disease.

G. *Analysis by Age*

As shown in Figure 6, we can see the age distribution in heart disease. The age distribution is divided into fourteen different age categories as follows: "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-older.

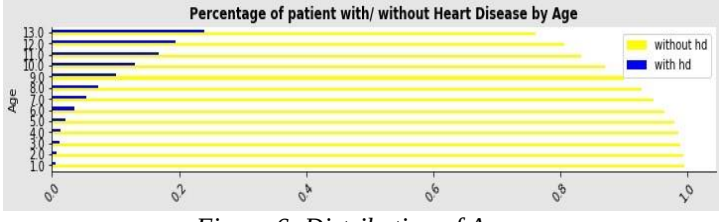

Figure 6: Distribution of Age

According to this pie chart, group 13, or age 65 or older, is the most at risk. As it is understood from the dataset, there is a direct ratio between age distribution and heart disease. As age increases, the probability of getting heart disease increases.

H. Analysis by Income

As shown in Figure 7, we can see the income distribution for heart disease. Contrary to the age distribution, there is an

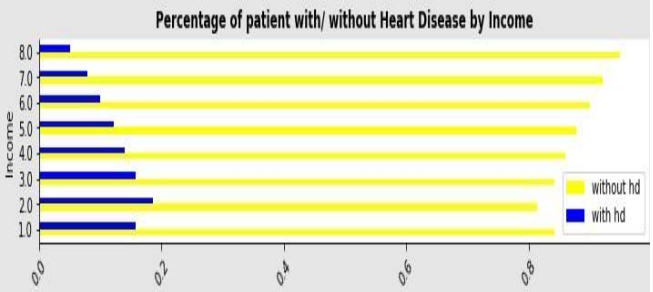

Figure 7: Distribution of Income

inverse relationship between income level and heart disease. Groups with higher incomes are less likely to get heart disease when compared to other groups. Low-income people have a higher risk of many diseases, including heart disease, diabetes, and cancer. Part of this risk comes from reduced access to healthcare. Lifestyle factors also play a role.

III. Methodology

This research builds on existing research on risk factors influencing heart disease to construct binary classifiers using Neural Networks, Random Forests, KNN, AdaBoost, and Bagging classifiers. Different models were tested on datasets with the train set, test set, and imbalanced data set, 50-50 dataset, 60-40 dataset, and not imbalanced dataset. Performance is measured in terms of accuracy score, confusion matrix, roc curve, precision, and recall curve using the sklearn metrics library.

A. Neural Network

Artificial neural networks are often used for efficient data mining and transforming raw data into actionable information. In this project, the performance of a simple neural network with 3 Dense layers is implemented. The dense layer is a layer that is deeply linked to its previous layer; this means that the neurons of the layer are connected to every neuron of the previous layer. This layer is the most used layer in artificial neural networks. The multilayer perceptron classifier man solvers used here are tested with cross-entropy loss function, sigmoid and relu activation functions, 0.001 learning rate, and 100 epochs. The Adam solver is the optimization that implements the Adam algorithm. Man optimization is a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments. Cross-entropy is a loss function that measures the performance of a classification model whose output has a probability value between 0 and 1. Activation functions are generally used to convert the linear outputs of a neuron to nonlinear outputs and enable a neural network to learn. The learning rate simply controls how much the weights are updated.

We have already emphasized that we split the data to avoid class imbalance, and we tested the model on these datasets. The results are described in Table 1.

Dataset	Accuracy	Loss	Runtime
Imbalance	0.9077	0.2378	≈ 1070 sec
50 50	0.7635	0.4933	≈ 200 sec
60 40	0.7785	0.4510	≈300 sec
Not imbalance	0.7809	0.4461	≈300sec

Table 1: Neural network performance

The ANN model performed well. Specifically, it performs well on the imbalance dataset with 90% accuracy. As expected, there was a decrease in accuracy for the 50–50, 60–40, and imbalance datasets, although the runtime performed well than the imbalance dataset.

B. KNN

K-Nearest Neighbor Algorithm is a popular supervised machine learning algorithm that can solve both classification and regression problems. The algorithm is quite intuitive and uses Euclidean distance measures to find the k nearest neighbors to a new, unlabeled data point to make an estimate. Finding the value of k in KNN is not easy. A small value of k means that noise will have a higher impact on the result and a large value will make it computationally expensive. Therefore, we initialize the empty array first and find where the k value has minimum error at 50 iterations. Then, we perform knn classification to reach higher accuracy with this k value. The performance of the Knn algorithm performance on imbalanced data sets is explained in Table 2.

	precision	recall	f1-score
0(not have disease)	0.91	0.99	0.95
1(have disease)	0.38	0.08	0.14
accuracy			0.99
macro avg	0.65	0.53	0.54
weighted av	0.86	0.90	0.87

Table 2: KNN on imbalance dataset

Precision – What percent of your predictions were correct?
As expected, those with heart disease have high precision and those without heart disease have low precision. This is because of the imbalance class. It also has a weighted average precision of 0.86.
Recall – What percent of the positive cases did you catch?
Recall has the same issue with precision due to the class imbalance. The weighted average recall is 0.90.

	precision	recall	f1-score
0(not have disease)	0.71	0.74	0.72
1(have disease)	0.73	0.69	0.71
accuracy			0.72
macro avg	0.72	0.72	0.72
weighted av	0.72	0.72	0.72

Table 3: KNN on 50-50 dataset

	precision	recall	f1-score
0(not have disease)	0.76	0.87	0.81
1(have disease)	0.63	0.45	0.53
accuracy			0.73
macro avg	0.70	0.66	0.67
weighted av	0.72	0.73	0.72

Table 4: KNN on 60-40 dataset

	Precision	recall	f1-score
0(not have disease)	0.77	0.87	0.81
1(have disease)	0.63	0.46	0.53
accuracy			0.73
macro avg	0.70	0.66	0.67
weighted av	0.72	0.73	0.72

Table 5: KNN on not imbalance dataset

From the above tables, we get higher precision, recall, and accuracy on the imbalance dataset. However, this imbalance causes imbalance precision and imbalance recall. The 60-40 dataset and the not imbalance dataset have pretty much the same result. The KNN model performed best on the 50-50 dataset.

C. Random Forest

Random forest, as its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The random forest classifier performance described below tables.

	Precision	recall	f1-score
0(not have disease)	0.91	0.98	0.95
1(have disease)	0.42	0.11	0.18
accuracy			0.90
macro avg	0.67	0.55	0.56
weighted av	0.87	0.90	0.88

Table 6: Random forest on imbalance dataset

	Precision	recall	f1-score
0(not have disease)	0.78	0.72	0.75
1(have disease)	0.74	0.80	0.77
accuracy			0.76
macro avg	0.76	0.76	0.76
weighted av	0.76	0.76	0.76

Table 7: Random forest on 50-50 dataset

	Precision	recall	f1-score
0(not have disease)	0.81	0.84	0.83
1(have disease)	0.66	0.62	0.64
accuracy			0.77
macro avg	0.74	0.73	0.73
weighted av	0.76	0.77	0.76

Table 8: Random forest on 60-40 dataset

	Precision	recall	f1-score
0(not have disease)	0.82	0.84	0.83
1(have disease)	0.65	0.62	0.63
accuracy			0.76
macro avg	0.73	0.73	0.73
weighted av	0.76	0.76	0.76

Table 9: Random forest on not imbalance dataset

We tested different datasets on a random forest classifier as described above table. The model performed well on the imbalanced dataset. Also, detailed information can be accessed through this notebook including the confusion matrix, roc curve, precision, and recall curve.

D. Ada boost

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as the Ensemble Method in Machine Learning. This is called Adaptive Boosting as the weights are reassigned to each sample and higher weights are assigned to the misclassified samples. Reinforcement is used to reduce variance as well as bias for supervised learning. It works on the principle that students grow in order. Except for the first, each subsequent student is trained from previously trained students. In simple words, weak learners are transformed into strong ones. The AdaBoost algorithm works on the same principle as boosting with a small margin.

We did all the reporting operations as we did before. We trained and tested the model on each dataset. We have reported all detailed information such as accuracy, precision and recall as stated above. In summary, we get better results on the imbalance dataset as expected. The weighted average precision and recall are 0.88 and 0.91, respectively. Detailed information can be accessed using the notebook such as precision, recall, roc curve, confusion matrix, precision, and recall curve.

E. Bagging

Bagging is the abbreviation of Bootstrap Aggregation, is a simple and very powerful ensemble method. An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART). Decision trees are sensitive to the specific data on which they are trained. If the training data is changed the resulting decision tree can be quite different and in turn the predictions can be quite different. Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees.

In this section, we implement a bagging classifier as a decision tree selected as the weak learner. We got the same results on the 50-50 dataset, the 60-40 dataset, and not imbalanced dataset. We achieved greater accuracy in the imbalance dataset than in the others.

IV. Conclusion

As stated above, we applied data mining techniques to examine risk behavioral factors associated with heart disease using BRFSS data. We compared classification accuracy for neural network, KNN, random forest, adaboost and bagging modeling approaches. Our analysis of outcomes included alcohol consumption, income, gender, age, smoking, etc. associated with reported heart disease including demographic, social and behavioral factors.

The purpose of this article was to determine the usefulness of using survey data from the BRFSS dataset to build predictive models for heart disease risk. Machine

learning models for heart disease created using BRFSS dataset were found to be highly predictive of heart disease with demographic, social and behavioral factors, as evidenced by high-accuracy models.

A. Best Overall Performance

The best overall performance is hard to pin down given the nearly identical performance between the Neural Networks, KNN, Random Forest, AdaBoost, and Bagging. That said, on the basis of accuracy alone, the Neural Network beat out the other classifier models on the imbalance dataset with slight difference for each dataset. It also had the lowest loss than the other models. The code notebook, including the classification report, is available at this link.

REFERENCES

[1] Johannes Christopher Eichstaedt. 2017, May 24. *Predicting And Characterizing The Health Of Individuals And Communities Through Language Analysis Of Social Media*.UPENN.

[2] Madhumita Pal, Smita Parija. 2021, December 17. *Prediction of Heart Diseases using Random Forest*. IOP

[3] Alex Teboul. Jun 20, 2020. *Building Predictive Models for Heart Disease*. MEDIUM

[4] Karan Bhanot. Feb 13, 2019. *Predicting presence of Heart Diseases using Machine Learning*. MEDIUM

[5] Scott Robinson. November 21st, 2021. *K-Nearest Neighbors Algorithm in Python and Scikit-Learn*. STACKABUSE

[6] Joseph Gatura. February 22, 2022. Bagging algorithms in Python. SECTIONIO.

[7] Elena Gritsenko Toth, Alexander McLeod, David Gibbs, Jacqueline Moczygemba. May 1, 2019. *HEALTH CARE ANALYTICS: MODELING BEHAVIORAL RISK FACTORS ASSOCIATED WITH DISEASE*.TXSTATE

[8] Debmalya Chatterjee, Saravanan Chandran. April 5, 2019. *Prediction and Classification of Heart Diseaseusing AML and Power BI*. SCITEPRESS

[9] Alma Pochini, Ben M. Williamsy, Hasanboy M. Isomitdinovz, Gongzhu Hux. July 1, 2015. *A Data Mining Analysis of Asthma Risk Factors*.

[10] Hardik Deshmukh. Jun 18, 2020. *Heart Disease UCI-Diagnosis & Prediction*.MEDIUM

[11] Centers for Disease Control and Prevention. Behavioral risk factor surveillance system. http://www.webmd.com/asthma/features/loweringcosts-asthma-treatment,.2015

[12] Harleen Kaur and Siri Krishan Wasan. *Empirical study on applications of data mining techniques in healthcare*. Journal of Computer Science, 2(2):194–200, 2006.

[13] Hian Chye Koh and Gerald Tan. *Data mining applications in healthcare*. Journal of Healthcare Information Management, 19(2):64–72, 2011.

[14] Ahsan, M.M.; Siddique, Z. *Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review*. arXiv 2021,arXiv:2112.06459.

[15] Lebedev, A.;Westman, E.; VanWesten, G.; Kramberger, M.; Lundervold, A.; Aarsland, D.; Soininen, H.; Kłoszewska, I.; Mecocci, P.; Tsolaki, M.; et al. *Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness*. NeuroImage Clin. 2014, 6, 115–125.