Predicting Heart Disease from Health and Lifestyle Factors

Data Mining Project

Kavya Kalapala
Data Science
University of Colorado - Boulder
Boulder, Colorado, USA
kaka2920@colorado.edu

ABSTRACT

According to the CDC, heart disease is the leading cause of death globally. In 2020, around 697,000 people died from heart disease in the United States alone. Early diagnosis of heart disease and other heart-related issues can help improve chances of survival of individuals. Machine learning in healthcare is growing research field with many potential applications. Hospitals generate mass amounts of data every day. By using data mining techniques and machine learning algorithms on the data collected from healthcare systems, medical professionals can identify trends or patterns that can help in early recognition of various diseases, including heart disease.

For this project, I've used a publicly available dataset from Kaggle which is a cross-section of the original data called the 2015 Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a health survey carried out by the CDC to find out various risk factors to a variety of health problems. The target variable was heart disease which had to be predicted by training the supervised learning models using the 21 risk factors such as age, diabetes, etc., as features. There was high class imbalance where negative (no heart disease) label was 90% more than the positive (has heart disease) label. I've used under-sampling as a way to solve the imbalance problem and created two balanced datasets where one is resampled in the ratio 50:50 and the other in ratio 60:40. I've trained four models using the imbalanced dataset and balanced datasets. The accuracy of models was pretty high when trained on imbalanced data. Linear regression had 90.0% accuracy, but AUC score of 0.56. Random forest had 89.25% accuracy and AUC score of 0.55. K-NN had 89.01% accuracy and AUC score of 0.53. Gaussian NB had 80.63% accuracy and AUC score 0.69. After looking at the precision and recall metrics, the model performances weren't great. I've the trained the models using the balanced datasets. The model with best accuracy was logistic regression with 76.59% (50:50 dataset). It had a better precision score when trained with the 60:40 dataset (70.18%). However, the Gaussian NB had the best recall for both datasets. Lastly, I trained the models using a subset of features to see if the model performances would improve. However, this proved to be futile. Overall, the best performing model was linear regression. It had an accuracy of 76.65% (60:40) and 76.73% (50:50), precision of 71.12% (60:40) and 67.51% (50:50), recall of 63.74% (60:40) and 56.87% (50:50) and an AUC score of 0.74 (60:40) and 0.72 (50:50). The second best performing model is random forest. It had an accuracy of 75.44% (60:40) and 75.67% (50:50), precision of 67.81% (60:40) and 64.53% (50:50), recall of 65.91% (60:40) and 58.43% (50:50) and an AUC score of 0.74 (60:40) and 0.71 (50:50).

KEYWORDS

Heart Disease, Data mining, Python, BRFSS, Supervised Machine Learning, Classification, Imbalanced Data, Linear Regression, Random Forest, k-NN, Gaussian Naïve Bayes

ACM Reference format:

Kavya Kalapala. 2023. Predicting Heart Disease from Health and Lifestyle Factors: Data Mining Project. In Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1234567890

1 Introduction

High blood pressure, high cholesterol and smoking are key risk factors for heart disease. According to the CDC, about 47% of Americans have at least one of the three risk factors. Several other conditions and lifestyle choices, such as diabetes, obesity, etc., also can put people at a higher risk of heart disease.

Diagnosis of heart disease is difficult as it requires vast amount of knowledge and experience. Various medical tests must be run and medical practitioners rely on medical-related attributes such as fasting blood sugar, cholesterol, ECG features, etc., to determine whether a patient is suffering from heart disease. By applying machine learning techniques, we can establish a relationship between existing health conditions, lifestyle choices and heart disease. It can help in early intervention and diagnosis.

The primary objective of this project is to build and compare supervised learning models for predicting heart disease risk, based on the given health and lifestyle factors. The models' performances will be compared using various statistical parameters. A well-performing model can help give an early indication of heart disease and also help raise awareness among individuals who may be at a high risk.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WOODSTOCK'18, June, 2018, El Paso, Texas USA
© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00

Additionally, I will analyze the problem of using datasets with high class imbalance when training machine learning models. We will compare model performance when using imbalanced and balanced datasets for training.

2 Related Works/Literature Review

There are many contributions made to the diagnosis of heart disease using machine learning.

Many of these works use the heart disease dataset from University of California-Irvine repository for building predictive and diagnostic models. This dataset is comprised of 303 patients with recorded medical features such as age, gender, fasting blood sugar, cholesterol, etc. Using this dataset, a 2022 paper published in NLM compared a multilayer perceptron neural network model and k-Nearest neighbors model using accuracy and AUC scores. The multilayer perceptron model achieved 82% accuracy and an AUC score of 86% while the k-NN model was 74% accurate and had an AUC score of also 86%.

A recent study from February 2023 used a real-world dataset consisting of 70,000 instances and 12 features to train various machine learning models and evaluate how accurately these models can predict risk of cardiovascular diseases. They used decision trees, XGBoost, random forests and multilayer perceptron. They trained each model with and without cross-validation. They found that decision tree achieved accuracies of 86.37% with cross-validation and 86.53% without cross-validation. It also had an AUC score of 0.94. The XGBoost model achieved accuracies of 86.87% with cross-validation and 87.02% without cross-validation and had an AUC score of 0.95. Random forest achieved accuracies of 87.05% with cross-validation and 86.92% without cross-validation. Lastly, multilayer perceptron achieved accuracies of 87.28% with cross-validation and 86.94% without cross-validation. The random forest model and multilayer perceptron both had an AUC score of 0.95. They concluded that the multilayer perceptron was the best performing model.

In 2019, a study by CDC researchers Zidian et. Al used the 2014 BRFSS data to develop a predictive model to identify risk factors for Type II diabetes. They found that for a cross-sectional data 138,146 survey responses, neural networks had the highest accuracy (82%), specificity (90%) and AUC score (79%).

This project uses the 2015 BRFSS dataset to build a predictive model that uses various lifestyle and health factors such as high cholesterol, high blood pressure, diabetes, education, income, etc., to predict whether an individual has a high risk for heart disease or not.

3 Proposed Work

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey conducted annually by the CDC. The original dataset consists of 441,455 responses and 330 features.

3.1 Dataset

I've downloaded a csv of cross-section of the data, containing 253,680 responses and 21 features. This dataset is publicly available on Kaggle.

#	Features	Data Type	Description
1	HeartDiseaseorAttack	Binary	0 – No heart disease 1 – Has heart Disease
2	HighBP	Binary	0 – No high blood pressure 1 – Has high blood pressure
3	HighChol	Binary	0 – No high cholesterol 1 – Has high cholesterol
4	CholCheck	Binary	0 – Hasn't checked cholesterol in the last 5 years 1 – Has checked in the last 5 years
5	BMI	Numeric	Data Range – (12, 98)
6	Smoker	Binary	0 – Has not smoked at least 100 cigarettes in entire life 1 – Has smoked at least 100 cigarettes in entire life
7	Stroke	Binary	0 – Never suffered from a stroke 1 – Has suffered from a stroke
8	Diabetes	Numeric	0 – No diabetes or only during pregnancy 1 – Pre-diabetes or borderline 2 – Has diabetes
9	PhysActivity	Binary	0 – No physical activity in last 30 days 1 – Has physical activity in the last 30 days
10	Fruits	Binary	0 – No fruit consumption per day 1 – Consumes one or more fruits per day
11	Veggies	Binary	0 – No vegetable consumption per day 1 – Consumes one or more vegetables per day
12	HvyAlcoholConsump	Binary	0 – No heavy drinking 1 – Heavy drinking
13	AnyHealthCare	Binary	0 – No healthcare access 1 – Has healthcare access
14	NoDocbcCost	Binary	0 – Has met a doctor in last 12 months 1 – Has not met a doctor in last 12 months due to cost
15	GenHlth	Numeric	1 – Excellent, 2 – Very Good, 3 – Good 4 – Fair, 5 – Poor
16	MentHlth	Numeric	Number of bad mental health days in a month. Range – (0, 30)
17	PhysHlth	Numeric	Number of bad physical health days in a month. Range – (0, 30)
18	DiffWalk	Binary	0 – No difficulty while walking 1 – Has difficulty while walking
19	Sex	Binary	0 – Female 1 – Male
20	Age	Numeric	Age categories: 1 – 18-24, 2 – 25-29, 3 – 30-34, 4 – 35-39, 5 – 40-44, 6 – 45-49 7 – 50-54, 8 – 55-69, 9 – 60-64, 10 – 65-69, 11 – 70-74, 12 – 75-79, 13 – 80 and older
21	Education	Numeric	Categories: 1 – Never attended school or only kindergarten, 2 – Grades 1 to 8, 3 – Grades 9 to 11, 4 – High school graduate, 5 – College 1 to 3 years, 6 – College graduate or higher
22	Income	Numeric	1: <$10 K, 2: $10–$15 K, 3: $15–$20 K, 4: $20–$25 K, 5: $25–$35 K, 6: $35–$50 K, 7: $50–$75 K, 8: >$75 K

3.2 Tools

The project was implemented in a Jupyter Notebook, using the programming language, Python. The following libraries were utilized in this project:

- Pandas
- Numpy
- Scikit-Learn
- Seaborn
- Matplotlib
- Imbalanced-Learn

3.3 Main Tasks

3.3.1 Data Loading and Cleaning. The dataset needs to be cleaned before further analysis. I will check for missing values and duplicated values. If found, both the missing and duplicated values will be removed.

3.3.2 Exploratory Data Analysis. I will perform a brief EDA. This is to analyze the relationship between various risk factor features such as BMI, High Cholesterol and the target attribute, Heart Disease. For the final step of EDA, I will check for highly correlated features.

3.3.3 Feature Selection. During this step, I will create two sets of features for later model training. One set will consist of all the features. The second set will consist of a subset of features. This subset of features will contain the uncorrelated features that were obtained when checking for highly correlated features. This is to see if removing the correlated set features will improve model performance.

3.3.4 Data Preprocessing. Before training the data, I need to scale the features. This is because distance-based algorithms such as k-Nearest Neighbors, etc. are heavily impacted by features that have different scales from each other. Features with higher magnitude may be given more weightage. To make sure that every feature will contribute equally towards the model performance, the features must be scaled. I will use a scaling technique called Standardization. In standardization, data values are centered around the mean with unit standard deviation.

3.3.5 Data Classification. I will train my machine learning models using various supervised learning algorithms, namely: Logistic Regression, Random Forest Classifier, k-Nearest Neighbors and Gaussian Naïve Bayes. Each model will be trained with both the full set of features and a subset of features.

3.3.6 Model Analysis. For each supervised machine learning model, I will analyze the performance by evaluating various statistical parameters such as accuracy, precision, recall and AUC score. The objective is to find the model with the best performance.

4 Evaluation

In order to analyze the results of this project, we need to evaluate each model performance using various metrics such as accuracy, precision, recall, etc.

4.1 Experimental Setup

The dataset will be divided into two portions, 80% for training and 20% for testing. Then, each supervised model will be trained with a full set of features from imbalanced dataset, 50:50 balanced dataset and 60:40 balanced dataset. The process will be repeated using a partial set of features. The predicted responses will be compared to the actual responses in order to obtain the required evaluation metrics.

4.2 Methods to Compare

I will analyze the performance of each model using the following metrics.

4.2.1 Confusion Matrix. It has four important parameters that can be used to evaluate the model performance.

- True Positive (TP). Number of correctly predicted positive values.
- True Negative (TN). Number of correctly predicted negative values.
- False Positive (FP). Number of incorrectly predicted positive values.
- False Negative (FN). Number of incorrectly predicted negative values.

4.2.2 Accuracy. This can be interpreted as representing the baseline performance of a model. It is defined as the number of correct predictions divided by total number of predictions. It can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.2.3 Precision. It is the ratio of correct positive predictions and total positive predictions. It can be calculated as:

$$Precision = \frac{TP}{TP + FP}$$

4.2.4 Recall. It is the ratio of correctly predicted positive values and the total number of actual positives. It can be calculated as:

$$Recall = \frac{TP}{TP + FN}$$

4.2.5 AUC Score. After plotting the true positive rate and the false positive rate in a single graph, we get the resulting curve which is known as the ROC curve. The area under this ROC curve is the metric obtained known as the AUC score.

$$TPR = \frac{TP}{TP + FN} \quad , \quad FPR = \frac{FP}{FP + TN}$$

The resulting evaluation metrics of each supervised model will be compared to find the model with the best performance.

4.3 Key Results

4.3.1 Training Models using Imbalanced Dataset.

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.900864	0.548474	0.131003	0.559411
1	Random Forest Classifier	0.892530	0.398073	0.115137	0.547736
2	K-NN Classifier	0.890136	0.335891	0.084477	0.532806
3	Gaussian NB Classifier	0.806254	0.270910	0.537521	0.687064

Table 1: Model Performance Using All Features

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.898209	0.511161	0.048848	0.521771
1	Random Forest Classifier	0.897687	0.483871	0.044795	0.519684
2	K-NN Classifier	0.886481	0.245428	0.054394	0.517698
3	Gaussian NB Classifier	0.853450	0.305751	0.343643	0.627503

Table 2: Model Performance Using Uncorrelated Features

4.3.2 Training Models using Balanced Dataset.

4.3.2.1 Balanced Dataset with Resampling Ratio 50:50.

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.767269	0.675088	0.568753	0.716926
1	Random Forest Classifier	0.756728	0.645275	0.584291	0.712999
2	K-NN Classifier	0.710351	0.615849	0.325883	0.612853
3	Gaussian NB Classifier	0.733539	0.588098	0.643678	0.710751

Table 3: Model Performance Using All Features

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.719064	0.640112	0.340358	0.623027
1	Random Forest Classifier	0.712810	0.588576	0.432099	0.641624
2	K-NN Classifier	0.675146	0.513446	0.304811	0.581231
3	Gaussian NB Classifier	0.710702	0.580094	0.447850	0.644045

Table 4: Model Performance Using Uncorrelated Features

4.3.2.1 Balanced Dataset with Resampling Ratio 60:40

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.766543	0.711200	0.637416	0.740830
1	Random Forest Classifier	0.754368	0.678069	0.659091	0.735396
2	K-NN Classifier	0.725512	0.642159	0.608375	0.702187
3	Gaussian NB Classifier	0.727963	0.632451	0.658670	0.714165

Table 5: Model Performance Using All Features

	Model	Accuracy	Precision	Recall	AUC Score
0	Logisitic Regression	0.702585	0.653536	0.443392	0.650973
1	Random Forest Classifier	0.701399	0.605497	0.588805	0.678979
2	K-NN Classifier	0.677682	0.580223	0.513678	0.645025
3	Gaussian NB Classifier	0.696577	0.615813	0.511364	0.659696

Table 6: Model Performance Using Uncorrelated Features

5 Discussion

5.1 Project Timeline

5.1.1 Week 1 – Data Loading, Cleaning and Exploratory Data Analysis. I've chosen a cross-section of the 2015 BRFSS dataset which is available on Kaggle. The data set had no missing values, but had 23,899 duplicated values which were removed. After data cleaning, there were 229,781 observations and 22 attributes. I did a brief EDA to see if I could identify trends between the various health factors and heart disease. I've identified that the following features could pertain to a high risk of heart disease: high blood pressure, high cholesterol, diabetes, cholesterol check, smoking habit, difficulty in walking, age and gender.

5.1.2 Week 2 – Feature Selection, Data Preprocessing and Balancing the Dataset. I will use the correlation matrix to select features that are not highly correlated for each dataset. The result was 6 features: HighBP, HighChol, CholCheck, BMI, Stroke, Diabetes. These are all high risk factors mentioned by the CDC. There's a heavy class imbalance in the dataset. To deal with this, I used a resampling technique called under-sampling to create two datasets using the resampling ratios, 50:50 and 60:40. For my analysis, I used three sets of data. One is the original imbalanced dataset and the other two are the balanced sets. This is so that I can compare the effect of imbalanced data on model performance.

5.1.2 Week 2.5 – Train Models using Supervised Machine Learning Algorithms. I've used four classification algorithms, Logistic Regression, Random Forest Classifier and K-Nearest Neighbors and Gaussian Naïve Bayes.

5.1.3 Week 3 – Evaluating Model Performances. I've obtained preliminary results for the above mentioned supervised learning models. I've created tables for each model performance with aforementioned feature sets and datasets.

5.1.2 Week 3.5 – Results and Conclusion. Attain results for all the models and compare performances. Identify the model and dataset that can best predict heart disease risk. Summarize the project and findings.

5.2 Potential Challenges

The dataset is heavily imbalanced. It might lead to the model being biased. In such cases, accuracy cannot be relied on as a proper evaluation metric of the predictive models. Other metrics such as precision, AUC score, and recall should be taken into consideration when analyzing the model performances.

5.3 Alternative approach

Balance the dataset using the under-sampling technique given in the Python library, Imbalanced-Learn. A balanced dataset consists of 10-20% variation. For comparison purposes, I will create 2 balanced datasets using resampling ratios of 50:50 and 60:40 to use for training models and then to analyze which resampling ratio gives the models better performances.

5.4 Changes/Lessons

Initially, I was only going to use accuracy and AUC score to evaluate model performances. The accuracy of models trained with imbalanced data was actually pretty high with each model having above 88% accuracy. However, the AUC score was only between 0.51 and 0.56 for all models. When training models with balanced datasets, the accuracy was less, but the AUC scores were higher. This was surprising to me and not what I expected. I realized I needed to use other performance metrics like precision and recall to get a better picture of how the model was performing. The lesson learned is that accuracy is not a good measure of performance for models trained on imbalanced data.

While trying to train the Support Vector Classifier model, the training time was taking too long. This is probably because I'm using a very large dataset and after some research, I found out that SVMs are not suitable for training on large datasets. I will probably use Naïve Bayes algorithm as the final machine learning model since it's known to be useful for massive datasets.

6 Conclusion

6.1 Project Summary

Heart Disease is a prevalent problem around the world. Using machine learning to identify key indicators of heart disease can help medical practitioners as well as the patients to be more informed about the health factors and lifestyle choices that may lead to increased risk of heart disease.

I have used a cross-section of 2015 BRFSS dataset to build predictive models using high risk factors to predict whether an individual has a risk of heart disease or not. After some data cleaning, I have done a brief exploratory data analysis to examine the relationships between various risk factors and heart disease. During this process, I've realized that there was a class imbalance problem. The negative class label formed around 90% of the dataset. To solve this problem, I used imblearn under-sampling technique to create two balanced datasets using resampling ratios 50:50 and 60:40. Furthermore, I selected a set of features that are not highly correlated with each other. I've used four supervised machine learning algorithms for my models and trained each model on the balanced datasets and imbalanced datasets. Furthermore, I've also trained them using a full set of features and a subset of features. I've used the metrics, accuracy, precision, recall and AUC score, to evaluate each model performance.

6.2 Key Findings

- Accuracy is not a good metric to evaluate models trained on imbalanced dataset. The other evaluation metrics were skewed. When training on data with high class imbalances, it is necessary to use other metrics like precision, recall, f1 score and AUC to analyze the model performance
- It is interesting to note that the models performed better when using the full set of features. This probably means that the even if there were a few highly correlated features, they probably do convey important information.
- The best performing model based on precision, recall and AUC score is the Logistic Regression model trained on balanced dataset with the resampling ratio 60:40 and using all features.
- However, Gaussian NB model has a better recall than all the models when trained with the three datasets.

6.3 Future Work

The work done in this project can be taken as a good starting point for further analysis. There are many potential areas for improvement. The CDC has stated that ethnicity is also an important risk factor, because different ethnic groups are more prone to heart disease than others. Including this feature in the dataset may help the models be more accurate with their predictions.

Using correlation to select features may not be the best method as we've seen in our analysis. There are many better methods like using variance threshold and fisher score for feature selection. Experimenting with different feature selection methods may ultimately lead to having better model performance if the right method is chosen.

In this project, I mainly focused on using supervised learning algorithms. For future work on this dataset, the use of deep learning algorithms can be explored. They are efficient and self-adaptive. Artificial neural networks can be implemented to obtain better analyses and results.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Qin (Christine) Lv, Associate Professor of Computer Science and instructor for Data Mining Foundations and Practice.

REFERENCES

[1] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019; 16:190109. DOI: http://dx.doi.org/10.5888/pcd16.190109

[2] Madhumita Pal, Smita Parija, Ganapati Panda, Kuldeep Dhama, Ranjan K. Mohapatra. Risk Prediction of Cardiovascular Diseases using Machine Learning Classifiers Open Med. 2022;17(1):1100–1113. DOI: https://doi.org/10.1515/med-2022-0508

[3] https://medium.com/@alexteboul17/building-predictive-models-for-heart-disease-using-the-2015-behavioral-risk-factor-surveillance-b786368021ab

[4] Fernando, C.D., Weerasinghe, P.T. and Walgampaya, C.K., 2022. Heart Disease Risk Identification using Machine Learning Techniques for a Highly Imbalanced Dataset: a Comparative Study. KDU Journal of Multidisciplinary Studies, 4(2), pp.43–55. DOI: http://doi.org/10.4038/kjms.v4i2.50

[5] Wilson PWF, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97(18):1837–1847; DOI: 10.1161/01.cir.97.18.1837

[6] Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*. 2023; 16(2):88. https://doi.org/10.3390/a16020088