Cardiovascular Disease Prediction using Supervised Machine Learning

AUTHORS KEVIN ALKINDY FAISAL MD ASHIQUR RAHMAN ELMI ABDIKADIR HUSSEIN

Abstract

Cardiovascular disease (CVD) or heart disease accounts for the **leading cause of death worldwide**. It is a matter for us to be concerned about today's highly chaotic lifestyle that leads to various diseases. **Early prediction** and identification of heart-related diseases can **bring down** the **mortality** rate. Consequently, many studies explore the realm of classification algorithms to find the best models. This paper follows the CRISP-DM Methodology to solve the problem of heart disease prediction. A relatively large dataset is selected from Kaggle, and it is cleaned by various preprocessing methods like imputation, normalization, encoding, transformation, and resampling. Although many ML models can be used, **this paper focuses on Logistic Regression** and **Random Forest**. In order to obtain a comprehensive result, the dataset is **prepared using five different approaches** where the combination of preprocessing methods has been changed. Since the **dataset** used in this study is **highly imbalanced,** a combination of various evaluation metrics like Confusion Matrix, Precision, Recall, and ROC curve have been used to analyze the performance of the two ML models. Finally, a detailed analysis of the result has been presented to explain how **Random Forest is more appropriate for predicting heart diseases**.

Problem Understanding

Background of Study

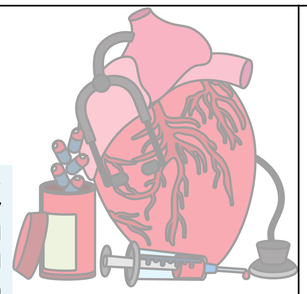
The number of people suffering from cardiovascular disease (CVD) or so-called **heart disease is significantly rising every year**. People do less physical work because everything is automated. **The medical equipment** that exists today still **takes time**, **less accurate**, and **costly**. **Machine learning can be used to predict CVD**. It can help the people take the precaution steps and the medical officers to personalize medical treatments for the people who have the probability to have CVD earlier.

Problem Statement

Manual diagnosis and complex treatment. The study of the patient's medical history manually is still used, reports of physical examinations performed by medical professionals are less accurate and take a long time.

Inaccurate detecting technique. It is critical to recognize the presence of this condition as soon as possible.

Lack of data. Accuracy could not improve due to a lack of large amounts of data or the nature of the dataset.



Objective & Motivation

The main objective of this study is to predict the probability of patients to have cardiovascular disease (CVD) in the future by using machine learning. Random forest and logistic regression algorithms are being employed to predict CVD outcomes. **The focus of this project is to maximize recall score by at least 95%**, for it is risky to misdiagnose patients and could lead to death.

Data Understanding

15 Categorical Attributes

- HighBP
- HighChol
- CholCheck
- Smoker
- Stroke
- Diabetes
- PhysActivity
- HvyAlcoholConsump
- Veggies
- Fruits
- AnyHealthcare
- GenHlth
- DiffWalk
- Sex
- Age group

3 Numerical Attributes

- BMI
- MentHlth
- PhysHlth

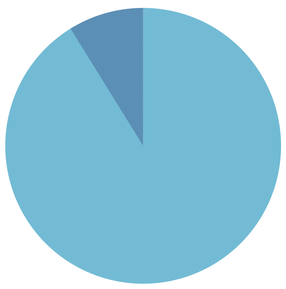
Kaggle

Dataset source :
BRFSS 2015 : https://rb.gy/unj6fw
Alex Teboul : https://rb.gy/8thljp

1 Target

2 Class Labels

Have Heart Disease
8.8%



No Heart Disease
91.2%

Highly Imbalanced

Data Preparation

Data Cleaning

Missing Values

34% of data

Two terms to call missing values in this dataset. Given missing values that can be detected by function, like NaN. Another one is values that have similar meaning with missing values, such as refused, not sure, & never checked.

Handle : Listwise deletion, Statistical Imputation

Outliers

Outliers exist in all numerical features. In feature BMI, outliers are under the lower bound, while the other two features are on the top of upper boud.

Handle : Remove with IQR, Capping outliers

Data Pre-Processing

Scaling / Normalization

The data is scaled using the minimum-maximum value-based method. It scales the value of the features within the range of 0 to 1. This **method** is **suitable for feature BMI, MentHlth, and PhysHlth**, since they **have skewed distribution**.

Log Transformation

Some linear algorithms assume the data follow a normal distribution. A log transform is helpful when dealing with positive numbers with a heavy-tailed distribution. All numerical features have positively skewed distributions.

Resampling methods

The imbalanced dataset is commonly found in medical diagnosis data. In order **to balance** the dataset, **3 resampling methods were applied** :

Random Undersampling
Random Oversampling
SMOTE

Categorical Encoding

There are 2 methods applied to encode the data.

Label Encoding

The target variable and all categorical features, except Sex, are encoded using this technique.

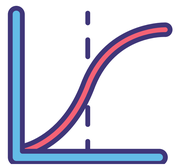
One-hot Encoding

Feature Sex was applied to this method. It consists of two unique values : male and female.

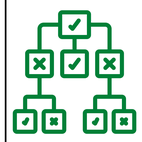
Algorithms

Logistic Regression

Logistic regression can solve classification problems through fits an S-shaped curve logistic function using the maximum likelihood In this case, the Logistic Regression model can show the relationship between the target variable and features, which features contribute the most to predict whether the people have a high probability to get heart disease or not.

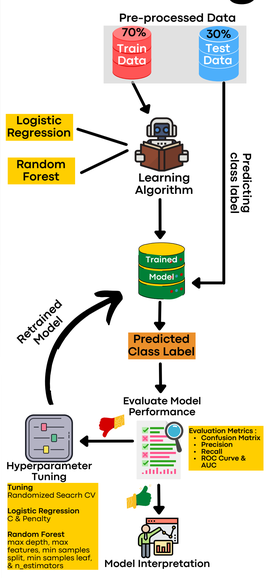


Random Forest



Random forest is one of the supervised learning algorithms which can handle both linear and non-linear data. Random forest predicts the new instance made by the trees that have low correlations. It is faster to fit the train data and predict the new instance with a large dataset. Furthermore, the combination of Random Forest and the resampling method can overcome the imbalanced class problem. Hence this is one of the proposed algorithms used in this project.

Modeling



Experiment Setup

Tools

- Python 3.9
- Jupyter Lab
- Pandas
- Scikit Learn
- Seaborn
- Matplotlib

Experiment 1

- Remove all missing values
- Remove outliers-IQR
- Normalization
- Log Transform
- Encoding

Experiment 2

- Statistical imputation for all missing values
- Capping outliers-IQR
- Normalization
- Log Transform
- Encoding

Experiment 3

Experiment 3 is the same with experiment 2, but **Random Undersampling (RUS)** is applied to **handle imbalanced class.**

Experiment 4

Experiment 4 is the same with experiment 2, but **Random Oversampling (ROS)** is applied to **handle imbalanced class.**

Experiment 5

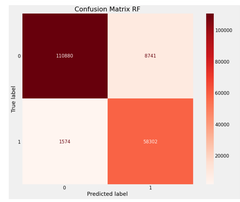
Experiment 5 is the same with experiment 2, but **SMOTE** is applied to **handle imbalanced class.**

Results & Analysis

Comparison of Model Performance

Model Experiment	Precision	Recall	AUC
LR EXP 1	53.90%	12.50%	55.70%
LR HYP EXP 1	55.40%	9.10%	54.10%
RF EXP 1	35.50%	14.50%	55.80%
RF HYP EXP 1	62.90%	5.60%	52.60%
LR EXP 2	54.30%	12.30%	55.70%
LR HYP EXP 2	56.20%	9.80%	54.50%
RF EXP 2	35.10%	14.50%	56.00%
RF HYP EXP 2	62.60%	5.30%	52.50%
LR EXP 3	69.20%	61.50%	74.00%
LR HYP EXP 3	69.50%	61.00%	73.90%
RF EXP 3	64.10%	62.10%	72.50%
RF HYP EXP 3	69.20%	62.00%	74.20%
LR EXP 4	69.90%	61.00%	74.00%
LR HYP EXP 4	70.00%	61.00%	73.90%
RF EXP 4	**86.90%**	**97.30%**	**94.90%**
RF HYP EXP 4	70.30%	63.80%	75.20%
LR EXP 5	77.00%	81.30%	78.50%
LR HYP EXP 5	76.90%	81.60%	78.50%
RF EXP 5	87.50%	92.10%	89.50%
RF HYP EXP 5	86.90%	90.10%	88.30%

Confusion Matrix of Random Forest



Feature Importances

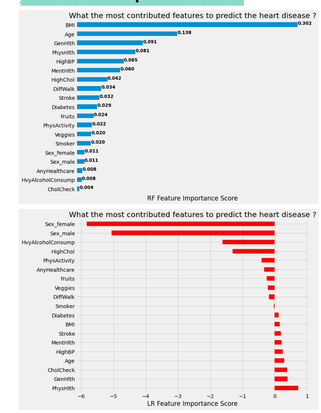


Figure above **shows the importance of each features** in both LR and RF respectively. These graphs shows how the two algorithms uses very different approach to solve the same problem. For **LR**, the **gender** of the patients **played** the most **important roles** in the classification process. But for **RF**, **BMI** and **Age** has the most influence in the model prediction.

Logistic regression (LR) has significantly higher precision than Random Forest (RF) in the first two experiments, where the imbalanced data is not handled. However, both Recall and AUC score was slighting higher for RF in these two experiments. Moving onto experiment 3 where random undersampling is applied, all the scores are significantly increased from the first and second experiments. **The results from experiments 4 and 5 show that RF clearly outperforms LR in all Precision, Recall and AUC when the imbalanced dataset is handled with oversampling.** Moreover, we can see that the random oversampling method give the best best for all the metrics in RF, whereas SMOTE-overasampling gives the best result for LR.

Conclusion

From 5 experiments were done. It can be concluded that the **combination Random Forest algorithm and random oversampling (ROS) method achieve the best score** for all evaluation metrics, with the **precision of 86.9%, recall of 97.3%, and AUC score of 94.9%**. That means a combination of Random Forest and ROS can do the classification properly. In comparison, Logistic Regression achieved the highest score with SMOTE technique. However, it is not as good as the Random Forest score. For future works, it is suggested to use TensorFlow to handle large-scale data in machine learning so that the model can be run faster. It also might be possible to improve the score using another algorithm and combination of resampling method.