ISYE7406 GROUP 8 PROJECT PROPOSAL

Title: Detection Of Heart Failure- A Life-Saving Model

Authored By: (1) Seraphina Toh (Gtid:xxxxxx780, ztoh6@gatech.edu) ; (2) Alissa Ong (Gtid: xxxxxx352, aong9@gatech.edu) ; And (3) Lim Wei Xuan (Gtid:xxxxxx174, lxuan7@gatech.edu)

1. Introduction

1.1 Problem Description

Accurate prediction and early detection of heart failures are important for timely interventions. In this data science project, various classification methods will be employed to predict heart failure using dataset sourced from various hospitals such as Cleveland, Switzerland, Long Beach combined. Several machine learning algorithms including (1) LDA, (2) QDA, (3) Naïve Bayes, (4) Logistics Regression, (5) KNN, (6) XGBoost Classifier, (7) Decision Tree, (8) Random Forest and (9) SVM will be used and evaluated using appropriate performance metrics. The model performance will be assessed through 10-fold cross validation and compared to determine the most effective approach for heart failure prediction.

Commented [AO1]: Similar to abstract

Commented [AO2R1]: For abstract will add on to this para to include our findings (e.g which features are important, which model is the best etc)

1.2 Motivation

Cardiovascular Disease (CVD) is the number one cause of death; accounting for about one-third of all deaths worldwide. Heart attack and stroke are the primary causes for four out of five CVD deaths, with one-third of these deaths occurring prematurely in people under 70 years of age. Heart failure is a common condition attributed to CVDs, often occurs due to factors such as narrowed arteries and high blood pressure. Heart failure is a life-threatening condition if left untreated. Thus, individuals with CVD or at high risk of CVD (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

1.3 Data Mining Challenges

One challenge in this problem would be to secure adequately-sized dataset as not all datasets possess the same predictors for merging. In addition, given advances in medical technology, there might be more predictors that could offer better explanatory power which are not captured in currently available datasets. Given the vast number of classification models, we would also have to qualitatively shortlist models that offer good prediction accuracy.

1.4 Problem Solving Strategies

We would test both simple and complex models, with a preference for simple models. This meant that we prefer to adopt simple models instead of complex models if the models offer statistically similar performance. This is because a simpler model would likely be more easily understood by medical professionals i.e. which factor poses higher risk to heart disease and allows them to translate to the patients.

2. Data Source

The data set is available at Kaggle, as detailed on < https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>. The combined dataset consisted of 918 rows (i.e., 918 heart failure datapoints), and 12 columns:

S/N	Predictors	
1	**Age**	Age of the patient [years]
2	**Sex**	Sex of the patient [M: Male, F: Female]
3	**ChestPainType**	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4	**RestingBP**	Resting blood pressure [mm Hg]
5	**Cholesterol**	Serum cholesterol [mm/dl]
6	**FastingBS**	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7	**RestingECG**	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8	**MaxHR**	Maximum heart rate achieved [Numeric value between 60 and 202]
9	**ExerciseAngina**	Exercise-induced angina [Y: Yes, N: No]
10	**Oldpeak**	Oldpeak = ST [Numeric value measured in depression]
11	**ST_Slope**	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12	**HeartDisease**	Output class [1: heart disease, 0: Normal]

3. Scientific Research Questions

3.1 Feature Importance

a) Which predictors are most significant for accurate prediction of heart failure?

3.2 Model Performance

a) Is it possible for any classification models to correctly detect heart failure (accuracy) at least 95% of the time?
b) How do the different machine learning algorithms perform in terms of accuracy and sensitivity for predicting heart failure, with more weight given to reducing number of false negatives.
c) What is the best classification model that balances accuracy with sensitivity for heart failure?
d) Can simple models maintain the same or provide better accuracy than more complex models?
e) Do certain demographic groups (e.g age, sex) have higher risk of heart failures?

3.3 Interpretability of Models

a) How interpretable are the models in explaining prediction of heart failure?

Commented [LW3]: Not sure if we should do this since Age and Sex are already predictors. However, we might want to convert Age from numeric to categorical e.g. Age groups. This could improve the outcome.

Commented [AO4]: This could be part of our write up on how interpretable these models are for clinical uses.

4. Proposed Methodology

4.1 Data Preparation

We use the original dataset from Kaggle for detection of heart failure. We check for missing data and impute the values with the median of the respective variable. Numerical variables will be scaled for models which are distance-based such as KNN and SVM.

4.2 Exploratory Data Analysis

We perform data exploration to identify any notable patterns or anomalies that are noteworthy. We would investigate the distribution of the predictors to check for normality and then check for correlation of variables to the response as preliminary methods to identify predictors with significant explanatory power to the response. We check for potential outliers and remove them from the dataset.

4.3 Data Mining and Statistical Learning Methods

We performed the following classification methods and compute the Cross-Validated ("CV") errors:

1. LDA: LDA models and classifies the categorical response with a linear combination of predictors using Bayes thereom. It assumes that the covariance matrix across classes is the same.
2. QDA : QDA models and classifies the categorical response with a non-linear combination of predictors using Bayes thereom. It does not assume constant covariance matrix across classes.
3. Naïve Bayes : It uses the Bayesian theorem to derive the probability to be in a class given a set of features x based on the likelihood of x to be in the class as well as the probability of that class. It assumes that given the response, the predictors are conditionally independent.
4. Logistics Regression: Logistic regression models the probability of success given the predicting variables. It assumes that the link function between the probability of success to the predictors is the logit function, in a way that this function of the probability of success is a linear model of the predictions.
5. KNN with the optimal k value: KNN classifies data points based on the majority vote of the K nearest neighbors. It assumes that the closer points are to each other, the more related they are.
6. XGBoost Classifier: It is a machine learning algorithm under the Gradient Boosting framework. XGBoost provides a parallel tree boosting in a fast and accurate way. It assumes that the objective function is continuous, differentiable, and convex.
7. Classification Decision Tree: A decision tree is built by splitting the variables in a way that minimises the sum of squares within each region of the tree of one splitting. The splitting is decided by a split point; decided based on which feature at that split provides the best separation of data. As a non-statistical approach, it makes no assumptions on the training data or residuals.
8. Random Forest Classifier: To deal with the sensitivity in the Decision Trees, one approach is to build many trees then aggregate across them. It is based on the idea of model averaging; averaging across multiple models to reduce over-fitting. It assumes that the predictions from each tree have very low correlations.
9. Support Vector Machines (SVM) : SVM performs classification by estimating a hyperplane that best separates the data, where "best" here is based on some criterion or a classification rule, for example, maximizing the margin between the training points for the two classes if binary data.

4.4 Determination of best performing model

The underlying principle is that we must be conservative in evaluating the reliability of the model as the consequences of a false negative will be more detrimental than a false positive: a false negative would result in an undetected health risk to the individual whereas a false positive would only result in loss of time or resources to perform further tests on the individual.

To determine the best performing model, we will be using the following evaluation metrics:

1. Accuracy: It measures the percentage of correct predictions.
2. Sensitivity: It measures the rate of the true positives out of all new responses that were observed as successes
3. AUC-ROC: It measures the likelihood that the model estimates a random "yes" point higher than a random "no" point.
4. F1-Score: It measures the harmonic mean of precision and recall.