Prediction of Heart Disease and Determining the Factors Influencing it

Smriti Tilak
Computer Science and Engineering
PES University
Bangalore, India
smrititilak@gmail.com

Kakumani Sai Deepika
Computer Science and Engineering
PES University
Bangalore, India
ksdrao7@gmail.com

Baddela Swathi Reddy
Computer Science and Engineering
PES University
Bangalore, India
swathireddybaddela@gmail.com

Deepa S
Computer Science and Engineering
PES University
Bangalore, India
deepadhruthi13@gmail.com

***Abstract*—This goal of this project is to determine if a patient could potentially have a heart disease or a failure on the basis of his or her previous medical records. It also aims to identify the factors that influence the patient's risk of having a heart failure using various statistical methods and picking the one which provides the most accurate results. This paper includes a brief introduction to our problem along with the review of other papers that discuss a similar problem. It also includes the visualizations for the dataset and our initial analysis and insights in order to accomplish a better understanding of it.**

Keywords—heart disease, heart failure, prediction, classification, data analysis, statistical model.

I. Introduction

Heart is a muscle that pumps blood received from veins into arteries throughout the body. As any other organ in the body, the heart is also prone to diseases, collectively called cardiovascular diseases. They affect the structures or functions of the heart, such as: Arrhythmias (abnormal heart rhythms), Aorta disease and Marfan syndrome, Congenital heart disease etc. Heart failure (HF), also known as congestive heart failure (CHF) and (congestive) cardiac failure (CCF), is a set of manifestations caused by the failure of the heart's function as a pump supporting the blood flow through the body; its signs and symptoms result from a structural and/or functional abnormality of the heart, that disrupts its filling with blood or its ejecting of it during each heartbeat. Heart Failure is a major cause of death throughout the world. As people age, the risk of heart failure increases. In addition to this, the lifestyle of the people plays a very important role in their lifespan. The economic burden is also predicted to rise particularly given the chronic nature of HF and high risk of hospitalisation Therefore, it would be beneficial to healthcare providers and payers if they can predict the risk of future outcomes associated with HF and a number of risk prediction models have been developed for the same.

This project considers a dataset [1] obtained from Kaggle. It has 12 attributes in total of which one of them is a binary attribute stating if the patient has a heart disease or not. It has 918 records which amounts to 11016 values. We aim to use various statistical models for classification and determining influential factors. In order to achieve this, our analysis started off with visualizing values of the various attributes and determining if any outliers were present. Insights were drawn from the data after its cleaning and pre-processing.

II. Literature Review

A. Machine Learning Techniques for Heart Failure Prediction

The aim of this research paper [2] was to determine which Machine Learning (ML) model was apt for early heart failure prediction. The paper compares four of the ML models – Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and Logistic Regression (LR). The performance of these techniques was measured by accuracy, precision, f1-score, recall, specificity and sensitivity. This study also investigates to find which factors influenced coronary heart failure the most, out of the 13 features that it considered. It checked to see if elimination of any of the factors still gave accurate predictions.The dataset for the study was obtained from Kaggle heart disease dataset containing 303 patients with 14 variables: age, sex, cp (chest pain experience), chol (cholesterol), target (1 - heart failure, 0 - no failure) etc.

Random Forest (RF) is a learning technique that makes use of Decision Trees (DT) for classification. It is known to work well on large datasets. A DT is constructed for every random sample taken from the dataset, hence multiple DTs will be constructed for a single RF model. Final prediction result is decided on the basis of votes given to every predicted result by the individual DTs. Support Vector Machine (SVM) is a supervised ML technique which works by creating a hyper-plane separating data into separate classes. The hyper-plane acts as the decision boundary. Naïve Bayes (NB) is an effective algorithm in classification using Bayes Theorem. It predicts the probability of different classes based on the attributes. The study made use of NB to detect presence of heart disease based on patients' previous records .Equation for Bayes Theorem used in NB is as follows:$P(x) \; = \frac{P(c)p(c)}{P(x)}$

$P(x) = P(c) \times P((c) \times \dots \dots \times P((c) \times P(c)$
P(x) is the posterior probability, p(c) is the class prior probability and P(c) is the likelihood.

Logistic Regression (LR) measures the relationship between the dependent variables and the independent variables. In this case, the dependent variable was the target variable stating if the patient will have a heart failure or not and the independent variables are the thirteen attributes.

All four models were used for building a prediction model. SVM with linear kernel function produced the best results among other kernel functions, hence, this result was compared with that of other techniques (RF, NB and LR). RF showed the best results for recall and sensitivity, which was 0.97 respectively. RF and LR achieved the same values of accuracy and f1-score which are 0.88 and 0.9 respectively. SVM achieved the highest validation accuracy which was 0.9 and NB didn't seem to outperform any of the models. Hence, the study concluded that RF achieved the highest average performance score followed by LR, NB and SVM. Another experiment was conducted to determine the least significant feature so as to eliminate these one at a time. Fasting blood sugar was found out to have the least influence on the prediction model and it was followed by Resting ECG value.

B. Evaluating the risk prediction models for adults with heart failure

The aim of this research paper [3] was to evaluate the risk prediction models for adults with heart failure.A number of models have been developed so that it would be beneficial to healthcare providers and payers to predict the risk of future outcomes associated with HF.Literature databases were searched from March 2013 to May 2018 to identify risk prediction models.However, the models were not based on uni variables in order to overcome the risk of bias(ROB). The models were based on multi variables so that there is no bias.Analysis of bias (PROBAST) was also used to assess the risk of bias.

Each model was assessed according to 3 or 4 domains respectively. These risk prediction models differed with regard to patient population, their statistical approach and the modelling applied. The agenda of these risk prediction models was to provide healthcare workers and payers a better approach to predict the risk of heart failure.According to PROBAST that is used for the analysis of bias, if the overall outcome was low then it was concluded that the model is a good match. However, if the overall outcome was high, it was not completely concluded that the model was a bad match but it was confirmed that there were some factors that influenced the overall outcome.

Out of 5425 citations, 4720 were non duplicated and among these, 40 publications were taken into consideration as they had relevant information. 250 other publications were excluded due to the uncertainty in the information. The outcomes from these publications included all cause mortality(n=17), HF hospitalisation(n=15), cardiovascular death(n=9) and other composite endpoints(n=14).

- CV mortality: Of the 9 CV mortality model outcomes, 3 reported c-statistic for model discrimination, 3 reported AUC-ROC, 1 used Kaplan-Meier assessment and 2 used the Therneau's survival concordance index.
- HF hospitalisation: C-statistics ranged between 0.59 and 0.8 for heart failure hospitalisation.
- Composite end points: C-statistics ranged between 0.620 and 0.745 for composite endpoints.
- All-cause mortality: Discriminatory value was assessed based on C-statistic.Internal Validation was carried out by 8 models and External Validation was carried out on 3 models.The median ranged between 0.677 and 0.826.

Some publications among these had a detailed report on handling the missing data.From these it was observed that, the discriminatory ability for predicting all-cause mortality, cardiovascular death, and composite endpoints was generally better than for HF hospitalization.Many distinct predictor variables(n=105) were identified.Of all the models it was observed that only 19% models had overall low ROB, 45% models discussed internal validation and 24% discussed about the external validation.Although 11 models were rated low ROB according to the assessment of PROBAST only 3 models had 'yes' or 'partial yes' in all domains.The other 8 models were considered overall low ROB according to PROBAST despite being rated unclear within at least one domain.

From the study it was observed that the majority of the 58 risk prediction models for HF represent concerns according to ROB assessment done based on PROBAST. This is mainly due to lack of validation and calibration. The utility of machine learning tools is yet to be determined.

C. Failure prediction using personalized models and an application to heart failure prediction

The main motive of this research paper [4] is to evaluate the failure in prediction of heart failure and conclude that the personalized models are better than generalized models in the case of predicting heart failures or any such physical systems. In this paper, they proposed a method for personalized modeling of a physical system for failure prediction (or, to be precise, predict the start of the degradation process of a system) based on time-series data produced by sensors and other measurement instruments. They then showed the application of this method to predict the "onset" of subsequent decompensated heart failure of three patients from the NIH (National Institute of Health) study.

When dealing with non-homogenous, disparate physical systems, personalized models can be better predictors of a phenomenon compared to generalized models based on data collected from an assortment of such physical systems.The advantage of personalized modeling is that it does not require large amounts of data collection about other similar systems. And for situations where it is difficult to generalize from diverse population characteristics, personalized models can be far more accurate. Heart failures are generally a slow degradation process and are similar to slow failure processes of many other physical systems. Thus, the method can be applied to failure prediction of machinery and production processes with similar characteristics. Although the proposed method uses time-series data produced by sensors and other instruments, it does not actually construct or use any time-series models. It simply examines the distribution of time-series data across specified time cycles to make predictions about the "onset" of failure. Prediction of the "onset" of slow degradation processes is also not strictly anomaly detection but is more about trend analysis.

In this paper they have talked about many methods that had been proposed earlier, few of them are Ross et al. [6] performed a systematic review of studies evaluating patient characteristics associated with hospital readmission for heart failure (HF). Choi et al. [7] used a recurrent neural network (RNN) model to predict the initial diagnosis of heart failure. The RNN model exploited temporal relations among events in electronic health records (EHRs) and used 3884 HF records of primary care patients. They also compared it with other models such as logistic regression, neural network, support vector machine and K-nearest neighbor classifiers. etc. After reviewing all the above papers they proposed a personalized model and the data collected and used to predict the "onset" of heart failure using such a model.

Here are the steps of the proposed 'onset' method:

- Select a set of time-series to use for modeling the physical system. This set may include both original sensor/device measurements (e.g. weight, blood pressure of a patient) and derived measurements (e.g. QRS complex, atrial premature complexes from an ECG).
- Determine a time interval *P* for recording time-series data and record the average values of the selected time-series within that time interval from streaming data. For example, the time interval *P* can be a millisecond, a minute or 5 min and would depend on how frequently one should observe the physical system for degradation or "onset" of failure or certain other events.
- Suppose we record the average values of *N* time-series in each time interval of duration *P*. Suppose *T* is the total number of observed time intervals during normal operations of the physical system, and $X\ (t,\ i)$, $i = 1 \ldots N$, $t = 1 \ldots T$, the average value of the *i*th time-series at the *t*th time interval. Thus, $X\ (t,\ i)$, $i = 1 \ldots N$, is an observation of the physical system across all *N* time-series at time *t*. Each time-series *i*, $i = 1 \ldots N$, represents a feature in the dataset.
- Using *T* collected observations of *N* time-series, create a clustering model to record the distribution of time-series data in that observation period *T*P*.
- Obtain a ranking of the features (time-series) using any unsupervised feature ranking method or from the clustering method.
- Monitor the physical system in subsequent periods of length *T*P* using the model created in point 4 and data from the *N* time-series. The monitoring tracks any changes in the distribution of time-series data. Monitoring essentially means passing the data through the clustering model and assigning each datapoint to the closest cluster. To understand the factors that cause a change from the normal behavior of the physical system, track some of the highest-ranked features (time-series) found in point 5
- If significant deviation in the distribution of the time-series data and in the trajectory of some of the highest ranked features occurs during the monitoring period, notify/alert a supervisory system of the change from normal.

Apply the above method to predict the "onset" of decompensated heart failure. After that they have concluded that In medicine, where we encounter non-homogeneous, disparate physical systems, generalized predictive models based on a population of patients may be less accurate compared to personalized models. Generalized models are also expensive to build because they require extensive data collection. And, because they require extensive data collection, there is a significant wait time for generalized models. On the other hand, a personalized model of the type proposed here can be built quickly because it needs data from just one individual patient whose model is to be built.

D. Congestive Heart Failure Symptoms in Patients with Preserved Left Ventricular Systolic Function: Analysis of the CASS Registry

In this research paper [5] the goals of this CASS registry analysis were twofold:

- To compare the clinical characteristics and long-term survival of patients with congestive heart failure and preserved ventricular function with those of patients without heart failure symptoms who also underwent coronary angiography; and
- To determine how the presence of coronary artery disease additionally affects survival in patients with heart failure and whether surgical revascularization offers a significant survival benefit over medical treatment alone.

Patients in the CASS registry with a history of moderate to severe functional impairment from congestive heart failure as assessed by the enrolling physician were eligible for inclusion in this study. Congestive heart failure was considered to be present if the patient demonstrated two or more clinical symptoms of pulmonary congestion including orthopnea, paroxysmal nocturnal dyspnoea and dyspnea on exertion <= 2 months before enrolment CASS. Only patients with class III or IV symptoms were included in this study. Two hundred eighty-four patients met these criteria and were selected for study. Patients in the registry without symptoms of heart failure who met the exclusion criteria used for the study group were included in the control group. The entire cohort of 13,071 patients who met the criteria was included in the control group.

The left ventricular ejection fraction, determined from a single-plane 45' right anterior oblique left ventricular angiogram at study entry, was required to be >0.45. An assessment of regional wall motion was also made from the left ventricular angiogram. Each of five cardiac segments (anterobasal, anterolateral, apical, diaphragmatic and posterobasal) were graded on a scale of I to 6 as follows: I = normal; 2 = moderately hypokinetic; 3 = severely hypokinetic; 4 = akinetic; 5 = dyskinetic; 6 = aneurysmal. The sum of the scores for each region is hereafter referred to as the left ventricular score. Thus, a score of 5 represents completely normal left ventricular systolic function. The number of diseased vessels in individual patients was determined by coronary angiography in the following manner. Stenosis >=70% in either the proximal or the distal portion of any of the three major coronary vessels was required by the CASS to classify a vessel as diseased. For the purpose of this study no distinction was made between proximal and distal lesions. In other words, a 701/c stenosis in the left anterior descending artery was classified as a single diseased vessel whether the steno-sis occurred proximal to the first septal perforator or in the mid or distal portion of the artery.

Comparisons of clinical characteristics between the two groups were performed with use of the chi-square and Student's t tests for categorical and continuous variables, respectively. Survival functions were computed with use of the Kaplan-Meier method and statistical comparisons between groups were made with use of the log-rank statistic. A p value < 0.05 is considered statistically significant; all p values are two-sided. Adjustments for differences in baseline variables were made with use of the Cox proportional hazards model. A stepwise procedure was used to screen the baseline factors that were significantly associated with mortality and then an adjusted p value was calculated in the model with use of those factors.

Clinical Characteristics: The two groups differ markedly across a wide range of clinical variables. Patients with congestive heart failure are older (56 vs. 52 years < 0.0001) and are more likely to be female (44% vs. 26%, p < 0.0001) than are patients without heart failure. Study patients also have a higher incidence of hypertension (p = 0.0003), diabetes (p < 0.0001) and chronic lung disease (p < 0.0001). Canadian Cardiovascular Society class III or IV angina severity was present in 70% of the patients with heart failure compared with 43% of those without heart failure (p < 0.0001). Patients with heart failure were also more likely to have had a prior myocardial infarction (53%Vs.38%, p < 0.0001). The differential use of nitrates (75% vs. 61%, p < 0.0001) was higher in the heart failure group. However, the use of beta-adrenergic blocking agents (46% vs. 47%) was not significantly different. The use of digitalis (4696vs.896, p <0.0001) and diuretic agents (55%vs. 18%, p < 0.0001) was predictably higher in the heart failure group.

Survival: After 6 years, 82% of patients in the heart failure group were alive compared with 91%of patients in the control group (p < 0-0001). Stepwise multiple regression analysis using the Cox proportional hazards model was perfumed for the combined study and control groups. The left ventricular score, the composite score for segmental wall motion abnormalities, was the strongest independent predictor of mortality (chi-square = 102.4).The presence of moderate to severe impairment from congestive heart failure was a significant univariate predictor of mortality in this study. However, after the stepwise multivariate analysis, the importance of congestive heart failure as an independent predictor of outcome was reduced to marginal significance. This analysis suggests that although the presence of heart failure in a CASS registry patient with preserved systolic function is associated with a decreased 6-year survival rate, it is the presence of associated clinical features and illnesses rather than the heart failure itself that primarily accounts for the decreased mortality. Nonetheless, in CASS registry patients the presence of heart failure appears to be an important and convenient marker for decreased long-term survival even among patients with preserved global systolic function.

On this basis a reasonable clinical argument could be made for surgical revascularization in patients with multivessel coronary disease and severe angina, heart failure and preserved Systolic Ventricular function. Whether coronary revascularization includes procedures including coronary angioplasty, would improve survival is not yet known. As the median age of the general Population continues to rise, an increasing number of patients with heart failure symptoms, advanced coronary disease and preserved systolic function will likely be seen. The appropriate management of their patients will require careful consideration of the concurrent physiologic abnormalities that predispose them to reduced long-term survival.

III. The Dataset

The dataset we've considered for our study is taken from Kaggle [1]. It contains 12 attributes with 918 observations. Each of these observations is a medical record of a patient indicating various health measures and it also includes a variable target that indicates if the patient has faced a heart failure. Using this data we aim to predict if a patient's medical record should be classified as one that can have a heart attack.

The variables included in the dataset are:

- Age: Current age of the patient
- Sex: Gender of the patient
- Chest Pain: Qualitative variable indicating the type of chest pain (TA - Typical Angina, ATA - Atypical Angina, NAP - Non Anginal Pain or ASY - Asymptomatic).
- RestingBP: Resting Blood Pressure of the patient
- Cholesterol: Indicates the cholesterol levels of a patient.
- FastingBS: Binary variable indicating fasting blood sugar (0 - if less than 120 mg/dL, 1 - if more than 120 mg/dL)
- RestingECG: Qualitative variable that can take the values - normal, ST (having ST-T wave

abnormality) and LVH (showing probable or definite left ventricular hypertrophy).

- MaxHR: Maximum heart rate achieved by the person.
- ExerciseAngina: Exercise induced angina (Y - yes or N - no)
- Oldpeak: ST (positions on the ECG plot) depression induced by exercise relative to rest.
- ST_Slope: Qualitative variable indicating slope of the peak exercise ST segment. It can take 3 values (Up - upsloping, Flat and Down - downsloping)
- HeartDisease: Binary variable indicating if the person had heart problem/failure or not

IV. Insights Obtained After Exploratory Data Analysis

Initially, we tried to visualize the various attributes through their histograms and pie charts. We noticed that the attribute RestingBP and Cholesterol had outliers as shown in the figures 1 and 2 respectively.

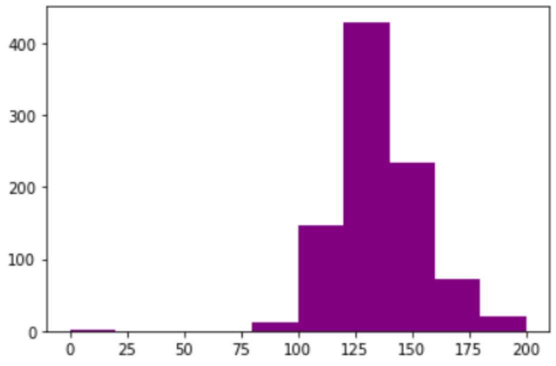


Figure 1

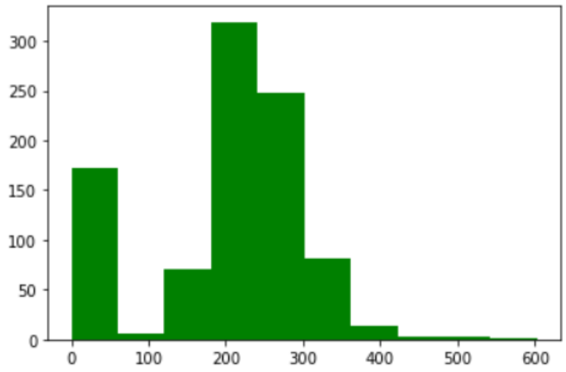


Figure 2

The presence of outliers in a prediction model can drastically affect its accuracy. Hence, it was important to nullify the effect of the outliers. These outliers were dealt with, using the interquartile range method (IQR method). In this method, the only accepted values fall within the range of the tenth and ninety fifth percentile. Rest of the values are replaced with the median of the accepted range. Once, the outliers are handled, we get the following graphs for RestingBP (Figure 3) and Cholesterol (Figure 4) respectively:

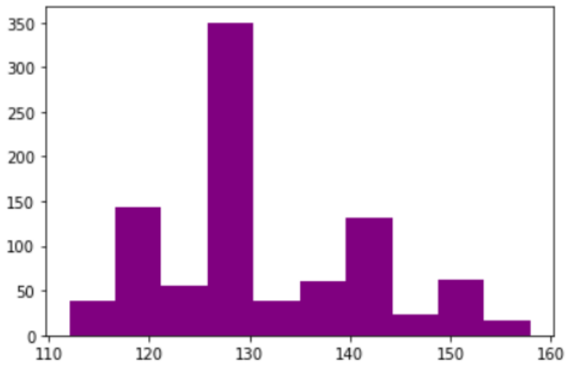


Figure 3

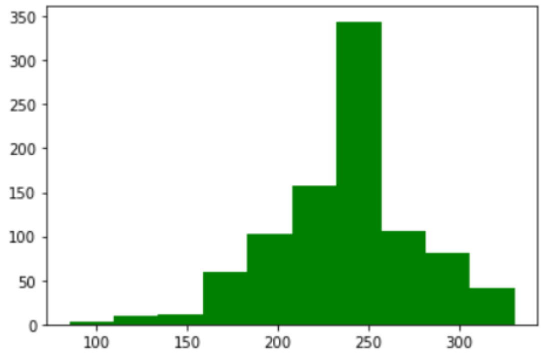
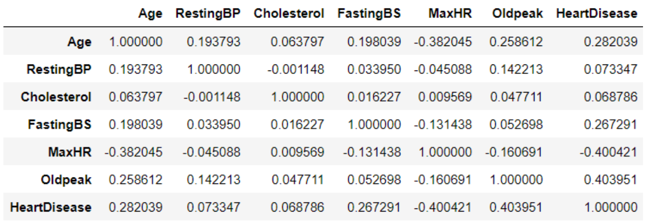


Figure 4

As it can be observed from these graphs, outliers seem to have been successfully removed. After the removal of outliers we analysed the correlation between the variables as we aim to understand the influence of these factors over the target variable. The following correlation matrix was obtained.



	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.000000	0.193793	0.063797	0.198039	-0.382045	0.258612	0.282039
RestingBP	0.193793	1.000000	-0.001148	0.033950	-0.045088	0.142213	0.073347
Cholesterol	0.063797	-0.001148	1.000000	0.016227	0.009569	0.047711	0.068786
FastingBS	0.198039	0.033950	0.016227	1.000000	-0.131438	0.052698	0.267291
MaxHR	-0.382045	-0.045088	0.009569	-0.131438	1.000000	-0.160691	-0.400421
Oldpeak	0.258612	0.142213	0.047711	0.052698	-0.160691	1.000000	0.403951
HeartDisease	0.282039	0.073347	0.068786	0.267291	-0.400421	0.403951	1.000000

From the correlation matrix, we can infer that MaxHR and the target variable (HeartDisease) have a negative correlation, indicating that a person who can achieve higher heart rates is less prone to heart failure. This is true in general, as younger people can achieve much higher heart rates compared to elders and heart failure is low or rare among the youth. The same is indicated by the negative correlation between the variables Age and MaxHR.

References

[1] URL to the dataset to be used for our analysis: Heart Failure Prediction Dataset | Kaggle

[2] https://www.researchgate.net/publication/353834616_Machine_Learning_Techniques_for_Heart_Failure_Prediction

[3] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224135

[4] https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-020-00044-2

[5] https://www.sciencedirect.com/science/article/pii/0735109791905892

[6] Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, Krumholz HM. Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch Intern Med. 2008;168(13):1371–86.

[7] Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2016;24(2):361–70.