Prediction of Heart Attack Using Machine Learning

Akshit Bhardwaj[1], Ayush Kundra[2], Bhavya Gandhi[3], Sumit Kumar[4], Arvind Rehalia[5], Manoj Gupta[6]

[1,2,3,4,5,6]*Department of Instrumentation & Control Engineering*
Bharati Vidyapeeth's College of Engineering Delhi-110063

Abstract- Cardiovascular diseases are one of the biggest reasons for death of millions of people around the world only second to cancer. A heart attack occurs when a blood clot blocks the blood flow to a part of the heart. In case this blood clot cuts off the blood flow entirely, the part of the heart muscle begins to die as a result. Going by the statistics, a heart problem can gradually start between the age of 40-50 for people with unhealthy diet and bad lifestyle choices. So, an early prognosis can really make a huge difference in their lives by motivating them towards a healthy and active life. By changing their lifestyle and diet this risk can be controlled. This Project intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using machine learning. The machine learning model predicts the likelihood of patients getting a heart disease trained on dataset of other individuals. As the result, the probability of getting a heart disease based on current lifestyle and diet is calculated. The model was trained with Framingham heart study dataset.

Keywords:-Heart Disease, Machine Learning, logistic regression, Cross-validation

I. INTRODUCTION

Machine Learning is one of the most rapidly evolving fields of AI which is used in many areas of life, primarily in the healthcare field. It has a great value in the healthcare field since it is an intelligent tool to analyse data, and the medical field is rich with data. In the past few years, numerous amounts of data were collected and stored because of the digital revolution. Monitoring and other data collection devices are available in modern hospitals and are being used every day, and abundant amounts of data are being gathered. It is very hard or even impossible for humans to derive useful information from these massive amounts of data that is why machine learning is widely used nowadays to analyse these data and diagnose problems in the healthcare field.

A simplified explanation of what the machine learning algorithms would do is, it will learn from previously diagnosed cases of patients. A heart Problem must be diagnosed quickly, efficiently and correctly in order to save lives. Due to this Researchers are interested in predicting risk of heart disease and they created different heart risk prediction systems using various machine learning techniques. The presence of missing and outlier data in the training set often hampers the performance of a model and leads to inaccurate predictions. So, it is critical to treat missing and outlier values before making a prediction.

1. Missing: For continuous variable, we can find the missing values using isnull() function. Mean of the data can also help identify. We can also write an algorithm to predict the missing variables.
2. Outlier: We can use a scatter plot to identify and as per need, delete the data, perform transformation, binning, Imputation or any other method. Diagnosis of heart disease using K-fold cross validation method will be used to evaluate the data and the result would be more accurate. 80% data of the patients will be used for training and 20% for testing. Parameter tuning is also necessary if accuracy is not close to 80 %. Logistic regression is the suitable regression analysis to perform when the dependent variable „y" is either 0 or 1. Like all regression models, the logistic regression is a type of predictive analysis. Logistic regression is used to explain the relationship between dependent variable usually „y" and various nominal, ordinal, interval or ratio-level independent variables (array of x features).
3. Features have higher odds of explaining the variance in the dataset. Thus, giving improved

model accuracy. The dependent variable or target variable should be binary/dichotomous in nature.

4. There should be no missing value/outlier in the data, which can be assessed by or converting the continuous values to standardized scores.
5. There shouldn‟t be a high correlation among the predictors. This can be interpreted by a correlation matrix among the predictors. The regression analysis is the task of estimating the log of odds ratio of an event.
6. Statistical tools easily allow us to perform the analysis for better results. Adding independent variables to a logistic regression model will always increase the amount of variance which would reduce the accuracy.

II. LITERATURE SURVEY

In the past few years, a lot of projects related to a heart disease risk prediction have been developed. Work carried out by various researchers in the field of medical diagnosis using machine learning analysis has been discussed in this section of the paper.

Das et al [1] worked on Deep Learning technology to find odds ratio or the prediction values from various different analytical models and with K-nearest neighbours got 89.00% classification accuracy on the Cleveland dataset for heart study.

Anbarasi et al [2] used 3 binary classifiers such as Naive Bayes, K-means clustering and Random forest for heart attack risk prediction using 13 features and then applied feature engineering for algorithm tuning and got great prediction results. They discovered that Random forest outperforms the other two binary

Classifiers with an accuracy of 99.2% for binary classification. The accuracy of K-means was 88.3% and Naive Bayes was about 96.5%.

Zhang et al [3] suggested an effective heart attack prediction model using Support Vector Machine (SVM) algorithm. In this, Principal Component Analysis was applied to retrieve the imperative features and different kernel functions. The highest accuracy was found with Radial Basis Function. To get the optimum parameter values, Grid search in SVM was brought to use and optimum values were found. The maximum classification accuracy touched about 88.64%.

Vadicherla and Sonawane et al [4] proposed a minimal optimization technique of SVM for coronary heart disease prediction. This technique helps in training of SVM by looking for the optimal values during training period. This shows minimal optimization technique provided good results even on a big dataset and execution time was also reduced significantly.

Elshazly et al [5] presented a classifier called Genetic algorithm SVM method Bio-Medical diagnosis in which 18 features were reduced to 6 features via dimensionality reduction. Different kernel functions were put up for use and performance was compared in terms of measures like accuracy, precision, recall, area under curve and f1 score. The results showed that linear SVM classifier managed an 83.10% accuracy with 82.60% true positive rate, 84.90% AUC and 82.70% f1 score.

III. PROPOSED SYSTEM

The machine learning technique used for the prediction of heart attack is Logistic Regression. The Dataset used for analysis and training is taken from Framingham Heart Study. It is a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham and is now on its third generation of participants. This Dataset can be found at www.framinghamheartstudy.org

This research intends to pinpoint the relevant/risk factors of heart disease as well as predict the overall risk using logistic regression. Mathematically, we can say that the logistic regression uses a Sigmoid function. Logistic regression values are categorical unlike linear regressions which are continuous. The logistic function is a sigmoid function, which takes any real value between 0 and 1. Mathematically,

$S(y) = 1 / 1 + e\text{-}y$

Or p (probability) = $1 / 1 + e - (\beta 0 + \beta 1x)$ Consider „y‟ as a linear function in a regression analysis,

$y+= \beta 0 + \beta 1x$

Putting y in s(y) sigmoid function, it becomes a logistic function after solving, $logit(p) = \log(p/(1-p)) =$

$\beta 0+\beta 1* \text{ Sexmale}+\beta 2* \text{ age}+\beta 3* \text{ cigsPeryear}+\beta 4* \text{ tot}$

Chol+β5∗ BP+β6∗ heartrate+β7∗ BMI

Here, β0 = Regression Constant p/1-p = odds ratio of the event βk = coefficient of x (predictors) where k = 1,2...

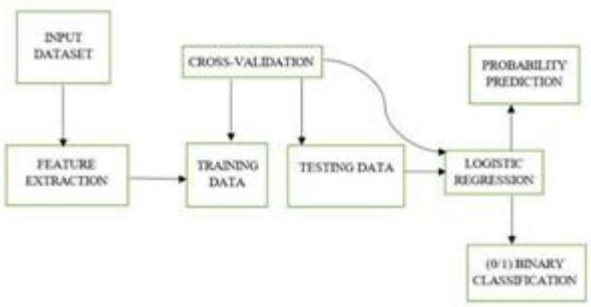

Fig 3.1: Block Diagram of the Model

Diagnosis of heart disease using K-fold cross validation method will be used to evaluate the data and the result would be more accurate. 80% data of the patients will be used for training and 20% for testing. This shows us better accuracy score and takes the least amount of execution time.

IV. IMPLEMENTATION

a. Feature Engineering

Feature Engineering aids in extracting information from the current data. Information is extracted in form of new features. These features might have a higher chance of explaining the (variance) in the data. Thus, giving better model accuracy.

b. Feature transformation

There are various cases where feature transformation is required

- Changing the variable from original scale to a scale between 0 and 1. This is known as normalization.
- Some algorithms work well with normal distribution. So, we have to remove skewness of the variables. There are techniques such as square root or log transformation to remove skewness.

c. Feature Selection

Feature Selection is a process of looking out for the best attributes which better define the relationship of an independent variable with target or dependent variable. The data is sliced into x and y training-testing datasets using cross-validation.

d. Algorithm Tuning

The aim of parameter tuning is to find the best value for each parameter to improve the accuracy of the ML model. To tune them, we must have a good knowledge about their impact on the output. We can repeat this process for other algorithms.

e. Results and Analysis

The machine learning model and implementation of a heart disease risk predictor for patients with risk of future heart disease using a logistic regression algorithm was successful. Accuracy was calculated as the ratio of total number of correct predictions to the total number of predicted outputs. And written as (TP+FN)/(TP+TN+FP+FN) Where, TP = True positive, TN = True negative, FN = False negative, FP = False positive.

Cost function f(x) is the sum of the squares of the difference between the actual value and the predicted value and iterations are the number of times the code will be executed to obtain the lowest value of f(x). Here, the Global minimum was calculated using gradient descent.

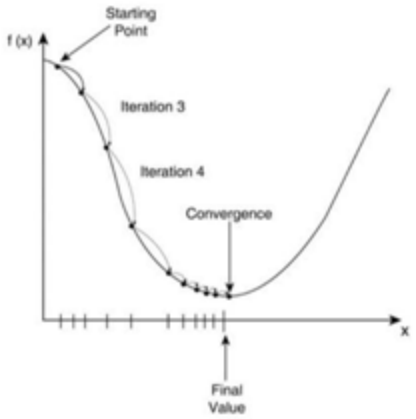

Fig 5.1: cost function

We also used a Receiver Operating Characteristic curve which tells us about how good the model can differentiate between TN and TP.

If the area under the ROC curve is more than 0.8 or it is bent more towards the left, the model will give better results and predictions would be precise.

Accuracy was 87% during testing as calculated by

the expression for the model on predicted and test values.

Details of 5 individuals and their corresponding probabilities. Threshold is 0.2, that is if probability risk > 0.2 would mean there is chance of a heart attack in future

Table 5.1: Medical data of five individuals

age	Sex_male	cigsPerMonth	totChol	sysBP	diaBP	BMI	heartRate
39	1	0.0	195.0	106.0	70.0	26.97	80.0
46	0	0.0	250.0	121.0	81.0	28.73	95.0
48	1	20.0	245.0	127.5	80.0	25.34	75.0
61	0	30.0	225.0	150.0	95.0	28.58	65.0
46	0	23.0	285.0	130.0	84.0	23.10	85.0

In Table 5.1 we can see the parameters use in the dataset for the analysis. Here several parameters are considered which include controllable and non-controllable factors. The parameters are Age, Sex i.e. male or female, Cigarettes count per month, Total cholesterol, Systolic and diastolic blood pressure, Body Mass Index and Heart rate of the individual.

Table 5.2: Probabilities of the individuals

No Risk of Heart Attack (%)	Risk of Heart Attack (%)
76.694188	23.305812
82.202523	17.797477
78.380359	21.619641
70.451408	29.548592
66.165425	33.834575

In Table 5.2 we can see the result of the system. Here the probability of the individual getting a heart attack is calculated based on parameters discussed in table 5.1. Here the percentage of NOT getting heart attack is depicted in first column followed by the percentage of getting heart attack in second column.

V. CONCLUSION

Men seem to be more susceptible to heart disease than women. Every 1 in 4 men [8] are likely to have heart disease whereas in case of women every 1 out of 5 women is likely to have heart disease [9]. Increase in Age, number of cigarettes and systolic Blood Pressure also show increasing odds of having heart disease. Interestingly total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol‟ in the total cholesterol reading. The early prognosis of cardiovascular diseases would aid in making better decisions on lifestyle changes in high risk patients and in turn reduce any future heart problems.

VI. FUTURE SCOPE OF THE PROJECT

At some point in future, the machine learning model will make use of a larger training dataset, possibly more than a million different data points maintained in electronic health record system. Although it would be a huge leap in terms of computational power and software sophistication but a system that will work on artificial intelligence might allow the medical practitioner to decide the best suited treatment for the concerned patient as soon as possible [6]. A software API can be developed to enable health websites and apps to provide access to the patients free of cost. The probability prediction would be performed with zero or virtually no delay in processing [7].

REFERENCES

[1] Das, Turkoglu, and Sengur, “Efficient diagnosis of heart disease via machine learning models”, Expert systems with applications, 2009.

[2] Vanisree and Jyothi, “Decision Support model for Heart Disease prognosis based on early signs of 8–51 patients using binary classification”, International Journal of Computer Applications, 2011.

[3] Y. Zhang, “Studies on application of Support Vector Machines in coronary heart disease prediction model”, Electromagnetic Field Problems and Applications, Sixth International Conference (ICEF), IEEE 2012.

[4] Vadicherla and Sonawane, “Decision Support for coronary Heart Disease analysis Based on

[5] Minimal Optimization technique”, International Journal of Engineering Sciences and Emerging Technologies, 2013.

[6] H. Elshazly, Hassanien and Elkorany, “Lymph diseases prediction based on support vector machine algorithm”, Computer Engineering &

Systems 9th International Conference (ICCES), 2014.

[7] Bhupender Kumar & Yogesh Paul, "Medical Applications of Machine Learning Algorithms", UIET, Kurukshetra University, 2016.

[8]]Ram Avatar & Vineet Kumar, "Deep Learning in healthcare", UIET, Kurukshetra University, 2018.

[9] Xu, JQ, Murphy, SL., Kochanek, KD, Bastian, BA. Deaths: Final data for 2013. National Vital Statistics Report. 2016.

CDC. Million Hearts™: strategies to reduce the prevalence of leading cardiovascular disease risk factors. United States, 2011. MMWR 2011.