Project Report:

Heart Disease Prediction Using Various Machine Learning Algorithms

Submitted To: Madam Satwat Bashir

Course Name: Machine Learning

Submission Date: 29/06/2021

Submitted By:

Khalil ul Rehman (F2020313010)

Shahzada Farhan (F2020313004)

Muhammad Ali (F2020313017)

Farhan Ahmad (F2020313023)

1. Problem Statement:

From the previous studies, it has been observed that various heart diseases are the fatal cause of death. From the research, it has been also identified that every 1 out of 10 deaths is due to the heart diseases. Many researchers have worked on machine learning algorithms to discover patterns of heart diseases that are still hidden. These patterns can enable us to identify if the patient has some kind of heart disease or not and how much severe it is.

Moreover, researchers have seen some issues while conducting their research such as no real-time data is processed and bulk quantity of data generated by medical centers have become a challenge for the researchers etc. However, enormous quantity of historical data was generated in order to process, store and analyze it which can be very helpful still for the future studies. In this regard, many researchers have opted big data tools to analyze above mentioned hurdles.

Hence, keeping in view the above problem and research gaps, we have also opted to work on previous data sets and our purpose to do analysis is same i.e., to analyze different hidden patterns in order to identify if a patient has heart disease or not but our research will cover some additional frameworks such as neural networks to check the accuracy of collected data.

2. Existing work:

2.1.Existing Literature Problem statement:

Researchers have moved to big data platforms and with the help of big data platforms they have participated much in the machine learning. As discussed and evident from the literature review that most of the heart diseases are fatal and patients with heard diseases are living with severe risks. ***Ahmed et al., (2020)*** discussed that many past studies have same as predicted results but somehow, due to some limitations their model was not successful. For example, ***Desai et al., (2019)*** applied logistic regression and back-propagation neural network technique in order to identify the patterns if the patient has heart disease or not. Similarly, ***Enriko et al., (2016)*** applied Naïve Bayes theorem along with Decision tree.

Selected research papers have discussed about similar issues and the main aim of the previous researches lie around the same research problem. ***Du, Yang et al., (2020)*** studied and worked on the hidden patterns of patients with hypertension that can also cause serious heart diseases. Their research aim was to work with the patients of hypertension and study if hypertension patients are open to coronary heart disease or not. Similarly, ***Pavithra et al., (2021)*** worked

on the heart disease diagnostics but their contribution was additional as they also worked on the patient health and identified the patient's proper diet in order to save the patients from the severity of the heart disease. Hence previous all studies were carried out regarding heart disease diagnostics.

2.2. Tools and Techniques in Literature:

Previous researches were conducted through various tools and techniques. Thought those techniques have different accuracy on same data sets but those were trained on clean data set. Their data collection source was mostly Kaggle and they sued the clean data from the Kaggle. Most of the data sets and in addition, one of them we are also using have 14 attributes including 1 target attribute. ***Pavithra et al., (2021)*** performed analysis using classification of the data set using entropy. In addition, ***Du, Yang et al., (2020)*** performed various techniques including extreme gradient boosting, SVM, logistic regression, Decision tree, KNN and Random Forest. Moving on, ***Ahmed et al., (2020)*** also performed analysis to identify heart disease but the novelty in their work was they had used data from the patient's social media and performed their analysis on spark.

2.3.Results:

Ahmed et al., (2020) worked on real-time heart disease detection and they developed their solution on Apache Kafka and Apache Spark. Machine learning classification algorithms used in their analysis were SVM, DT, RF, and LR. Their research concluded that Random Forest was good solution for the analysis with 11% more accuracy as compared to all others.

Moving on, Du, Yang et al., (2020) performed HER studies on the data set collected from hospitals and they analyzed the data without any medical examination and other investigations which not only facilitate their analysis process but also made it easy for them to analyze. They concluded in their study that accumulation with EHR through platforms that are centralized specially time-point changes that are multiple in nature provides good prediction and before the time prevention of different chronic diseases.

In addition, Pavithra et al., (2021) concluded after analyzing with the help of classification based on entropy and achieved 89% accuracy from the data set. Their proposed algorithm had higher accuracy than the previous algorithms and with the help of this algorithm they also helped to suggest patients with their diet and proper exercise. Hence, they concluded their study

with the remarks that using data mining algorithms, results can be obtained with higher accuracy.

3. Our Methodology:

In any research and execution, methodology plays a vital role. There are various classification algorithms that are used to analyze our research area i.e., predict the heart disease occurrences and in specific special algorithms will check the highest accuracy of the predictions.

3.1.Collection of data:

Before going to the methodology of data analysis, data collection will be discussed. Collection of data is done through Kaggle which is a vide platform for the data sources and a great platform for the data analysis discussions and dashboards to store data analysis tasks. In order to address the topic (our hypothetical statement) our focus will be on the same data set on which previous studies were performed but the novelty we have brought to our work is that we have also applied neural network technique on the heart.csv data set to see if the accuracy as compared to other big data tools has increased or not. Hence our primary data source will be CSV file obtained from the Kaggle and in case any other information is required, the same data set will be amended as per requirements.

3.2.Classification Algorithms:

a) Linear Regression analysis:

LRA (said as Linear Regression analysis) is the basic tool used to analyze the linear behavior between the two variables. In this type of analysis, using statistical equation, ***y=mx+c*** analysis is performed. In the above equation, y is said as the linear dependent variable which is depending on 'x' variable as 'x' increases, y will increase or decrease depends as per type of relation between the variables. This regression analysis technique is performed from the scratch on the data set obtained from the Kaggle and cleaned as per requirements (if needed).

b) Logistic Regression analysis:

Talking about the classification tools to analyze the said data set, we will start with the logistic regression after exploratory data analysis. In this part of data classification, we are aiming to preform our analysis from the scratch and in this regards we have done some scratch-based coding in which libraries are not used and coding is done on the basis of basic formulas. Hence, through the logistic regression analysis, we have obtained the accuracy score of the data set which is shown in the **table named -Results** below.

c) SVM (Support Vector Machine):

It is a model which is supervised machine learning based model and it uses classification for two groups classification problems. Talking about our problem, we will use SVM to classify either occurrence of the disease is predicted right or wrong and either the patient has heart disease or not. With the help of SVM modelling, decision boundary will be drawn to analyze the results. In case our results are not as per requirement and boundary line is not linear in nature, then as per the literature, we will add third axis in order to make the results convenient to understand. Hence using third axis, it will give us three-dimensional space.

Thus, using above models of classification, we will additionally analyze the previous data sets and shall interpret the results. In addition to the previous studies, we will check if the same data set has some different predictions about the disease occurrence or not.

d) KNN:

Using KNN machine learning technique, accuracy of our model is analyzed. This technique is used in this data set because in the present case, data and labels are known as we have further cleaned the data by checking non-zero rows (null values) etc. In this technique, we have compared test data with the train data to find the most look alike K instances and then the most similar data is summarized in order to test data classification.

e) Neural Networks:

This is a type of deep learning architecture. This is a bit difficult to train and to get the weights that are right for the modeling. This technique needs large amount of data and it computes powers and biases from the weights. Neural network consists of nodes which is actually in the form of collection of nodes.

This part is actually the novelty as it has not been seen in the previous computations and calculations. We have computed the accuracy using Neural Networks using 2-layers NN technique. In this technique, we have passed input layer, weights (8) and biases (1) to the code. Iterations to run the code is set as 100 and learning rate set in the function calling is set as 0.01. In the beginning, after checking shape etc. of the data set, we have checked for the null values in the data set. Moving on, we have split the data into test and train on the basis of 20:80 respectively. Further, user defined functions are made with the help of various platforms and scratch-based code is prepared for the neural network. In the end, tuning the model with different input, weight and bias values our model has shown average output around 88%.

f) PCA:

This technique dimensionally reduces the data set and performs necessary operations on the data set to do necessary calculations to see the projections of the data which is our primary data in our case. This technique transforms the larger data set into smaller ones and performs the operations on the smaller data sets to check the accuracy of the testing values. This modeling technique do so by creating variables that uncorrelated in nature which in result increase the variance.

g) Decision Tree:

It is a supervised machine-learning algorithm which is used for the classification problems. It is said as decision tree because it is actually same as the structure of the tree. It includes features and class labels which are actually internal nodes and leaf nodes respectively. In addition, the branches of this algorithm represent the branches of the tree. This technique splits the data that maximizes the sorting and separation of the data and hence forming the tree like structure. Most common test performed and used by the researchers is information gain which says that at every single split, entropy decreasing is maximum.

4. Tabular form of Results:

	Models	Accuracy Score
0	Linear Regression	54.464945
1	KNN	72.131148
2	Decision Tree	77.049180
3	SVM	80.327869
4	Logistic Regression	80.327869
5	PCA	83.000000
6	Neural Network	88.000000

Table 1: Results

Hence, the above Table "Results" has shown accuracy values for the different models and from the above models, it has been seen that after tuning the variables, our novelty model has shown best accuracy score of average 88%.

5. References:

Ahmed, H., Younis, E. M., Hendawi, A., & Ali, A. A. (2020). Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, 111, 714-722.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back-propagation neural network versus logistic regression in heart disease classification. In Advanced computing and communication technologies (pp. 133-144). Springer, Singapore.

Du, Z., Yang, Y., Zheng, J., Li, Q., Lin, D., Li, Y., ... & Cai, Y. (2020). Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation. JMIR medical informatics, 8(7), e17257.

Enriko, I. K. A., Suryanegara, M., & Gunawan, D. (2016). Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 8(12), 59-65.

Pavithra, M., Sindhana, A. M., Subajanaki, T., & Mahalakshmi, S. (2021). Effective Heart Disease Prediction Systems Using Data Mining Techniques. Annals of the Romanian Society for Cell Biology, 6566-6571.