Heart Diseases Prediction using Machine Learning Algorithms

Mrs. Garima Singh [1], Nikhil Kaintura [2], Mayank Bartwal [3], Kumar Gaurav [4], Ashutosh Dwivedi [5]
[1,2,3,4,5]*Inderprastha Engineering College Ghaziabad, India*

***Abstract*— We are living in the modern Technological era, and we are trying to solve the problem using technology for that we are striving for more knowledge with help of our technology, the more our lifestyle is changing. In the fast growing world everyone is neglecting their Health. Result of this Health diseases are increasing day by day due to our lifestyle and other hereditary reasons. Especially Heart Diseases have become a major issue these days. WHO report that over 11 million deaths are caused each year worldwide due to heart related complications. Using this we can understand that the condition is very serious. There is an urgent need for some technological solutions which can help in educating people and early diagnosis of these disease. This requires a string infrastructure including a large enough work force, which is a slow process. In the recent year machine learning has evolved a lot.**

To resolve this problem, we can use Machine Learning that can help us in this. Sudden increase in the number of patients depict the need for scalability in our medical system. While work force may be limited, we can look for algorithmic solutions for screening and diagnosis stages, due to their correlation with Our project "Heart Health Classification and Early Diagnosis of Heart Disease" is a system that uses predictive capability of machine learning models based on similar data points collected in past. We are going to use the previous medical report data to predict the risk of heart disease.

I. INTRODUCTION

Increasing population, changing life patterns, new diseases, pandemics, and change in the natural environment has surely changed the way we take care of our health. Maintaining good health is a major factor associated with any individual's efficiency and it has always been a big concern for all the countries of the world. Providing a good medical service across the whole country proves to be challenging for developing and poor countries.

The condition of medical systems across the world is different for different countries. The ratio of healthcare workers (doctors especially) to the population is a good way to determine the strength of any healthcare system. Developed countries (the USA and Western European) have this ratio to a good level (~25:10000), for Southeast Asian countries like India are way below the world average (India has a ratio of 12.21 to the 10000, whereas the world average is closer to 10:1000). Increasing this ratio of doctors to patients is a complex process and takes a lot of time in order of years or maybe decades and medical buildings equipped with modern machinery are needed.

It is also a general foreseen that people avoid going to doctors for preliminary stages of their diseases and other risks. Designing a suitable system for them so that they can pre-screen themselves can be very helpful hence A basic solution is to design a good scalable system here. Scalability here means that the technological stacks used can expand the scope of health care without needing more workforce i.e. Use software services.

Software services generally are beneficial for the fact that they can detect Early Disease and their Speed & Precision in Medical Diagnostics, enabling the Provision of Quality Care. It also promotes Improved Patient Participation & Management and improves the chances of safer Treatment Solutions.

II. LITERATURE REVIEW

Sushmita Roy Tithi et al discussed about ECG data analysis and heart disease prediction using machine learning algorithms. [6]

In this paper they have used 6 supervised machine learning algorithms to distinguish between normal and abnormal ECG. also they wanted to find a particular disease. They divided there dataset into 2 parts 75% for training and rest 25% for testing. They used - ECG, Machine Learning, Logistic Regression, Decision Tree, Nearest Neighbor, Naïve Bayes, Support Vector Machine, Artificial Neural Network, Right bundle branch block, Myocardial infarction,

Sinus tachycardia, Sinus Bradycardia, Coronary Artery disease, Abnormal ECG. ECG provides us with series of sinus rhythm which defines the condition of heart. Used to detect certain kind of diseases.

Disease Name	Best Algorithm	Score
CAD	Naïve Bayes	94%
Sinus Bradycardia	Decision Tree	95%
Sinus tachycardia	All except NN	95%
Myocardial infarction	Decision Tree	96%
Right bundle branch block	Logistic Regression	96%

Table 2.1 Best Algorithms for Various Disease

Bo Jin et al discussed about predicting the Risk of Heart Failure with EHR Sequential Data Modeling.[7] Their aim was to provide heart patients an early diagnosis and treatments. Because now a heart failure is really common among people age of 65, overweight people and those with previous heart attack. This paper develops a new approach to this vital task using and enhanced long short- term memory networks (LSTM) method and a data-driven framework. In this paper they proposed a novel method for diagnosis event modeling that includes one-hit encoding and word vectors and employs LSTM approach for this. This paper used electronic health record (EHR) data from real-world databases regarding congestive heart disease. Dataset had 2 parts A- diagnostic records of 5000 patients who have been diagnosed with heart failure. B- diagnostic records for 15000 patients who have not been diagnosed.

Based on the research above, we will be using 13 attributes:

Age: The probability of heart related issues after 60 years of age is very high, so it's important to include age in the heart disease prediction model. Over 80% of heart disease deaths occur in people over the age of 65, according to empirical estimates.

Sex: Cardiovascular risk factors are more pronounced in women with clinically manifest heart disease than in men. Females are more likely than males to have their first AMI after smoking.

Resting Blood Pressure:(trestbps,Integer, in mmHg) An increased resting heart rate was related to heart disease in both sexes in an annotated study, even after controlling for confounders such as abdominal obesity and general obesity. Resting blood pressure is one of several factors contributing to the risk of heart disease, so it is important to control it.

Angina: Induced during exercise (exang, 0: no: yes) Reduced blood flow to the heart causes this type of chest pain. The condition is associated with coronary artery disease. Angina pectoris is another name for the symptom. Pain in the chest is commonly described as being squeezing, heavy, or tight.

Highest ST-segment (slope, 1: upward slope, 2: flat, 3: downward slope

Resting ECG (aberration in ST/T-wave): Your heart is affected when the electrical impulses produced by other muscles interfere with the heart's electrical impulses. Heart disease patients' ECGs are characterized by naturalness, making them a good indicator to detect the disease.

Classification of Chest-pain:(cp,Integer, myocardial infarction: {1(typical), 2(atypical)} 3: pain, 4: asymptotic)} Shown to be closely related to heart related risk factors

Fasting Blood Sugar (fbs, Integer, 1: if fasting blood sugar over 120mg/dl, 0: otherwise) Heart related disease risks and fasting glucose levels tend to follow J shape curves. A glucose level of 85 to 99 mg/dL carries the lowest risk. The risk of heart disease, ischemic heart disease, myocardial infarction, and thrombotic stroke increased progressively when fasting glucose levels reached more than 110 mg/dL, but not the risk of hemorrhagic stroke. A fasting glucose level below 70 mg per dL was associated with increased stroke risk in 15 of 17 patients (hazard ratio 1.0, 95% CI 1.01-1.11) in men, and 1.10, 1.05-1.18 for
women.

Fluoroscopy colored vessels (ca, Integer, 0,1,2,3)Serum Cholesterol (chol, Integer, in mg/dl) A correlation between serum cholesterol level and coronary heart disease was found to range from 1.3 (95% CI 0.7-2.3) in those with a level between 4.7 and

5.1.However, the difference was less for those with 6.2 mmol/L or more (95 % CI 1.0, 2.7), compared to those who had 4.7 mmol/L or less. In contrast, more active individuals did not experience this. Individuals who participated in lesser physical activity had significantly inverse relative risks for all levels of cholesterol, including cholesterol levels exceeding 6.2 mmol/L (Relative risk = 0.4) (95% CI 0.2, 0.7).

Thalassemia (thal, Integer, defect: {1 (non-fixable), 2(reversible), otherwise: 3} Your body has less hemoglobin than normal as a result of this inherited blood disorder. Basically, hemoglobin carries oxygen. If you have thalassemia, you will feel fatigued.

Max heart rate achieved (thalach, Integer) Despite controlling for age in the study, the maximum heart rate is correlated with heart failure in the annotated study. Likewise, max heart rate, negatively correlated with access adipose tissue, is an indicator of fitness. So it is a good tool, presumably associated with an inverse relationship between heart disease risk and exercise.

Exercise induced ST depression (oldpeak, Integer) In an ECG, exercise-induced ST segment depression is defined as a horizontal or downsloping ST depression greater than 1.0 mm at 80 ms after the J point or any ST depression of greater than 1.0 mm.

Highest ST-segment (slope, 1: upward slope, 2: flat, 3: downward slope.

III. METHODOLOGY

1. Data Collection

The dataset which we have used is available on Kaggle website and is available for public download [10]. It is processed from UCI's dataset and contains valid tuples for further processing.

Data Characters	Multivariate
Number of Tuples	1025
Number of Attributes	14
Attribute Datatype	Categorical, Integer, Real
Source	Kaggle

Table 3.1 Dataset Detail

Next after downloading datasets is generally to clean the data if some missing values, noise in data is present. First the data is imported from downloaded csv files using pandas libraries of Python, to RAM in data type called data frame which is optimized adequately to handle two dimensional array data. The dataset does not have any null values. 13 of our attributes are the attributes which are used to predict the result, while the last attribute "target" is the result, i.e., whether or not the individual was suffering from heart disease.

2. Study of Dataset

We will generate and understand correlation between our attributes and target class. Correlation matrix will be generated and plotted using matplotlib.

For a correlation matrix, the more positive the value of correlation, the more increase in value of one variable causes the other to increase i.e. more directly proportional. The higher a negatively correlated variable gets, the lower the value of the target becomes.

3. User Interactive Front End

There will be a multipage website containing a homepage, a page for the user in which he enters details to predict the presence or absence of heart disease and a page about us. Flask has been used to connect the frontend with trained models of the backend.

Flask is a Python web framework that was created with a philosophy in mind. Armin Ronacher conceived and developed Flask as an April Fool's Day hoax in 2010. Despite its comedic beginnings, the Flask framework has grown in popularity as a viable alternative to Django projects' monolithic structure and dependencies.

There are some advantages of flask over other frameworks as per requirement of our project. They are, Flask is a Python web framework built for rapid development of small projects**,** Flask offers a diversified working style while Django offers a Monolithic working style.

IV. EXPERIMENT RESULT

1. Logistic Regression:

This is the one of the most common model used in ML, Logistic Regression is often applied in the actual manufacturing context the fields such as data mining, automatic disease diagnosis and economic prediction. For our model, we use Logistic regression to know the

risk factors for heart disease and forecast the probability of disease occurrence based on risk factors. This model is most frequently applied for classification, primarily two-category issues (that is, there are only two types of output, each representing one category), and can indicate the probability of occurrence of each classification event.

This technique used is also known as sigmoid function. Sigmoid function helps in the easy representation in graphs. Logistic regression also provides better accuracy. By using equation the logistic regression algorithm is represented in the graphs showing the difference between the attributes. The process of modeling the probability of a discrete outcome given an input variable is known as the Logistic Regression. The most common logistic regression, as its name suggests is not regression rather it is a classification algorithm that classifies something that can take two values such as true/false, yes/no, and so on. Logistic regression identifies a hyperplane in a manner that when it is passes through a function whose value ranges between 0 and 1 (typically we use sigmoidal), it optimizes cost function. Bases on closeness to 0 or 1, it predicts a Boolean output. Here vector parameters is used for training. σ(.) is usually a sigmoid function, with output between 0 and 1.

In the logistic function equation, *x* is the input variable. Let's **feed in values** −20 to 20 into the logistic function.

```
In [146]: logistic_model = LogisticRegression()
          logistic_model.fit(X_train.values, y_train.values)
          logistic_model_prediction=logistic_model.predict(X_test.values)
          print(accuracy_score(y_test.values,logistic_model_prediction))
          print(classification_report(y_test.values,logistic_model_prediction))

          0.8131868131868132
                        precision    recall  f1-score   support

                     0       0.85      0.71      0.77        41
                     1       0.79      0.90      0.84        50

              accuracy                           0.81        91
             macro avg       0.82      0.80      0.81        91
          weighted avg       0.82      0.81      0.81        91
```

Figure. 4.1 Logistic Model Result

2. KNN:

K-nearest neighbors is an online processing algorithm having capabilities of both classification and regression. When the input datapoint is provided, it calculates its distance from all the available training data, using any distance metric like Euclidean, Manhattan, etc. It finds k examples closest to the input. Based on majority vote, it predicts the target result for the input. It is nonparametric algorithm. The training is done by passing train set to the function fit () of the object K-NeighborsClassifier from sklearn.neighbors. For prediction, we use predict () function of the same object. The value of k, i.e., the number of neighbors has to be experimentally checked. The one with the best results is selected. After increasing the number of neighbors to a certain extent, the accuracy either starts dropping or the increase is negligible. This point of called knee point. Since the larger number of neighbors means more calculation for evaluation, we tend to choose the knee point, as it provides the minimum number of neighbors that will give good enough results. In practical scenarios, if k turns out to be less than 5, we choose 5, which is also the default value provided by sklearn.

Since we are getting the best value at k = 11 and k = 15, it was decided in favour of k = 11, which is usually the default value. Finally, prediction is done on test data and the result table is prepared

```
#dated::10/0/2022
knn_scores = []
for k in range(2,21):
    knn_classifier = KNeighborsClassifier(n_neighbors = k)
    knn_classifier.fit(X_train.values, y_train.values)
    knn_score=round(knn_classifier.score(X_test.values, y_test.values),2)
    knn_scores.append(knn_score)


knn_classifier = KNeighborsClassifier(n_neighbors = 5)
knn_classifier.fit(X_train, y_train)
knn_score=knn_classifier.predict(X_test)
print(classification_report(y_test,knn_score))
```

```
              precision    recall  f1-score   support

           0       0.62      0.49      0.55        41
           1       0.64      0.76      0.70        50

    accuracy                           0.64        91
   macro avg       0.63      0.62      0.62        91
weighted avg       0.64      0.64      0.63        91
```

Fig 4.2 Classification report of KNN

3. **SVM:**

It belongs to the supervised machine learning algorithm. This is the type of algorithm that can be used for both classification and regression challenges. SVM is mostly used in classification problems. In the SVM algorithm, we set each data object as a point in the n- dimensional space (where n is the number of attributes you have) by

the value of each element which is the value of a particular combination. Then, we do the splitting by finding a hyper-plane that separates the two sections very well. Support Vectors are simply links to individual points.

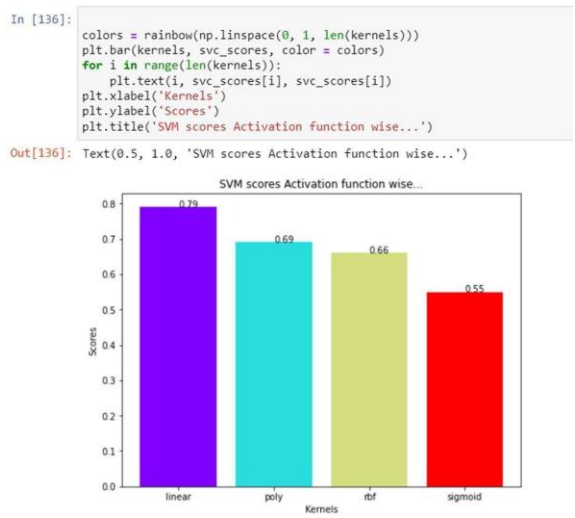

Fig 4.3 SVM Scores against various kernels

KERNEL	ACCURACY
Linear	79
Poly	69
RBF	66
Sigmoid	55

Table 4.1 SVM Scores against various kernels

4. Decision Tree:

A decision tree contain a flowchart-like structure. In the structure of DT (decision Tree) in test on an attribute is represent using internal node. The result of the test is represent using the branch. To represent a class label leaf node is used. To represent classification rules paths from root to leaf is used.

This is the type of analysis in which for A visual and analytical root , closely related influence diagram are used. where the expected values of competing alternatives are calculated. Architecture of Decision Tree.

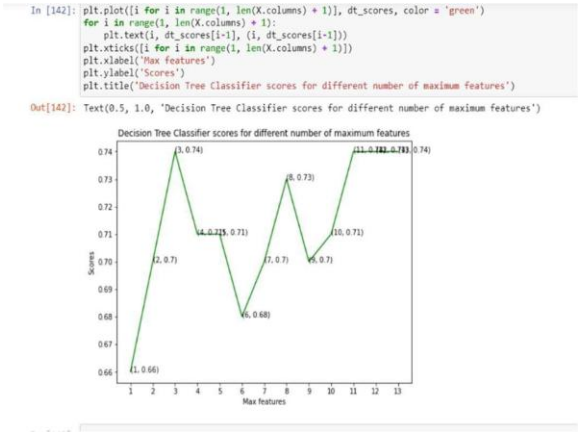

Fig 4.7 Decision tree Result Graph

5. Random Forest Classifier:

It is based on the decision Tree. It has a group of various Decision Tree. Various decision trees are used internally. When we try to make a classification for the given Input data we feed same data into all Decision trees. Now we collect all the votes from Decision trees and majority of votes is going to the result for input.

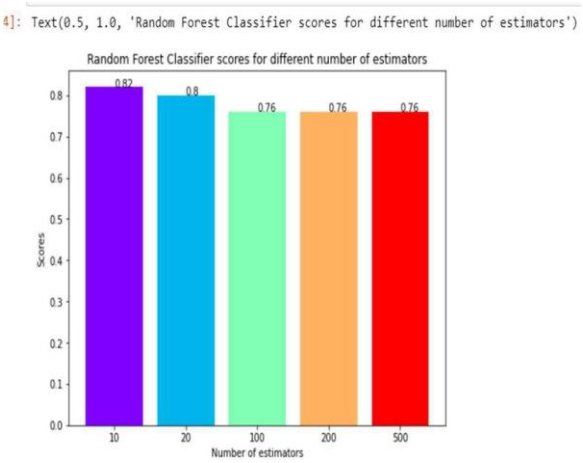

Fig 4.9 Random Forest Result Image

V. APPLICATION

1. Working Model Diagram:

The working front End has a page which contains form where user will enter all the required medical information. The form page screenshot is attached as:

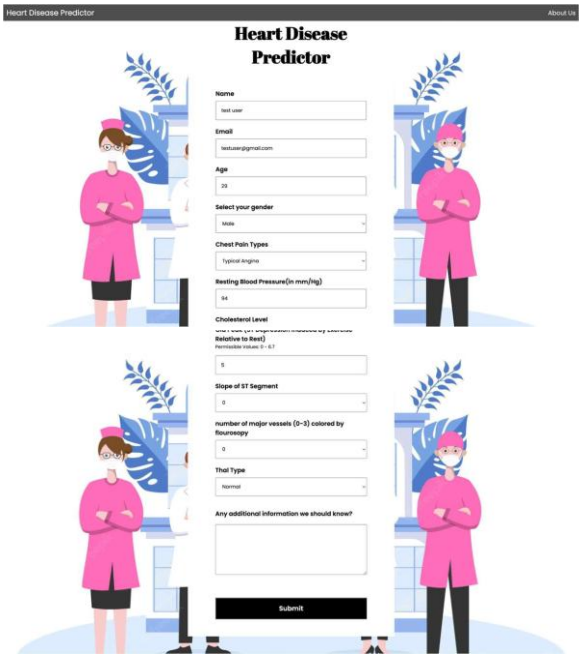

Figure 5.1 Input Form

The result page screenshot:

Figure. 5.2 Result Page

It also has a button to print the report generated so that user can have a record of data entered byhim/her and the respective result.

VI. BENEFITS OF PROJECT

Below are the advantages one can get from this project:

1. Anyone can check the presence or absence of heart disease right from their devices only with help of medical records.
2. Medical professionals can use this tool to directly determine the results.
3. Decrease in engagement of medical professionals for decision taking.
4. Project can be launched on large scale and be adapted in every hospitals related to heart diagnosis centers.

VII. PROJECT LIMITATIONS

This project models require 13 attributes for their prediction. If we analyze the attributes closely, then most of the attributes are not available to any normal person until he/she takes some medical tests which will cost them more money, time and for medical professionals, equipment. The attributes required are also more in a medical term than in general language which almost everyone can understand. It would be more friendly and easy for a user if the attributes which require more medical tests, can be decreased significantly and more common attributes which are responsible for heart disease can be included. Some attributes which can work for this are whether the user is a smoker, alcoholic, exercise frequency etc.

VIII. FUTURE WORKS

Future work for this project can be included as deploying after modifying the parameters. Models can also be trained with reduced parameters. One can increase the accuracy if possible, as there is always a chance of improvement. One can also use other classification algorithms with reduced parameters so that in case of absence of some attributes, users will be able to check, however with reduced accuracy.

REFERENCES

[1] Ahmad AA, Polat H. "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm." *Diagnostics* (2023).

[2] Bhatt C. M., Patel P. "Effective Heart Disease Prediction Using Machine Learning Techniques." *Algorithms* (2023).

[3] Subramani S, et al. "Cardiovascular Diseases Prediction by Machine Learning Incorporation with Deep Learning." *Front Med* (2023).

[4] Hajiarbabi M. "Heart Disease Detection Using Machine Learning Methods." *J Med AI* (2024).

[5] Nayeem M. J., Rana S. "Prediction of Heart Disease Using Machine Learning Algorithms." *Eur J AI ML* (2022).

[6] Sushmita Roy Tithi ,AfifaAktar , Fahimul Aleem , Amitabha Chakrabarty ‘’ECG data analysis and heart disease prediction using machine learning algorithms’’. Proceedings of 2019 IEEE Region 10 Symposium

[7] Bo Jin ,Chao Che, Zhen Liu, Shulong Zhang, XiaomengYin, AndXiaopeng Wei, “Predicting the Risk of Heart Failure WithEHR Sequential Data Modeling” ,IEEE Access Volume 6 2018.

[8] Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor, Redhwan Nour4, Samad Wali And Abdul Basit ,“An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection” , IEEE Access 2017.This work is licensed under a Creative Commons Attribution 4.0 License Volume 4 2016.

[9] Muhammad, Y., Tahir, M., Hayat, M. et al. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Sci Rep 10, 19747 (2020). https://doi.org/10.1038/s41598-020-76635-9.[https://www.nature.com/articles/s41598-020-76635-9]

[10] Yar Muhammad et al. "Early Detection and Diagnosis of Heart Disease Using Intelligent Computational Models." *Healthcare Informatics Journal* (2024).

[11] Ashir Javeed et al. "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection." *ML Applications Journal* (2024).

[12] Yar Muhammad et al. "Intelligent Computational Models for Heart Disease Detection." *Advanced Healthcare Systems Journal* (2024).

[13] Ankit Sharma et al. "A Hybrid Deep Learning Model for Heart Disease Prediction." *Journal of Healthcare Informatics* (2024).

[14] Priya R. et al. "Random Forest and Decision Trees in Heart Disease Diagnosis." *AI in Medicine* (2023).

[15] Ramesh Kumar et al. "Gradient Boosting for Cardiovascular Risk Prediction." *Heart Health Journal* (2023).

[16] Sofia Martinez et al. "Deep Neural Networks for Heart Disease Classification." *Machine Learning Applications in Health* (2024).
https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[17] https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc- 20353118