© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed

Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

Heart Diesease Prediction Using Machine Learning

Sanket Bhujade
Department of Computer Engineering
St. Vincent Pallotti College of Engg & Tech
Sanketbhujade2002@gmail.com

Shivam Pandey
Department of Computer Engineering
St. Vincent Pallotti College of Engg & Tech
spshivampandey8948@gmail.com

Dr. Samir Ajani
Department of Computer Engineering
St. Vincent Pallotti College of Engg & Tech
sajani@stvincentngp.edu.in

Yash Bokade
Department of Computer Engineering
St. Vincent Pallotti College of Engg & Tech
Yashbokade777@gmail.com

***Abstract*—** The human heart is among the body's most vital organs. For living a healthy life, it is very important to make our hearts healthy. Diagnosis and prediction of heart disease is very difficult and complicated. It requires more perfection and accuracy to predict any result. In the modern era, we can predict heart disease with the use of various technologies and models. Machine learning is one of the best modules to use to predict heart disease. Machine learning is one of the branches of the artificial intelligence that gives the more accurate outcomes. The proposed work predicts the health of the heart disease by performing different algorithms like K-NearestNeighbour, random forest, and logistic regression So the given paper gives the information about the project that we create and also tells which algorithm predicts heart disease more accurately.

Introduction

Heart Disease is a type of disease which is related to Heart and Blood. More people die from heart diseases than from any other disease. It is estimated, about 28000 people die from heart disease every year in India [2]. And about 12 million people die from heart disease per year globally. So, it is very important to predict heart disease in its earlier phase.

The four main types of heart disease (CVDs) are:

- coronary heart disease.
- stroke.
- peripheral arterial disease.
- aortic disease.

Machine Learning is one of the efficient techniques for prediction of heart disease, in which we have to train and test our model. it is a branch of Artificial Intelligence, in which the machine acts as human and can make decision itself. In which machine learning systems needs to be trained for data processing and efficient use of data.

In machine learning, our model learns from the natural phenomenon, natural things. In this project we use the biological parameter as for testing data such as cholesterol, Blood pressure, age, sex, etc. and based on this, a comparison of the algorithms' accuracy is made, for example, in our project we utilised three algorithms: logistic regression, KNN, and Random Forest [3].

The accuracy of three alternative machine learning algorithms is calculated in this study, and the result is used to determine which strategy is the most accurate.

This Paper Consist of the following points:

- Introduction about Machine Learning
- Methodology used for prediction

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed
Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

- Algorithms used in the project
- Visualization
- Result and conclusions
- References

I. Machine Learning

One effective technique is machine learning, which relies on two concepts: testing and training. The system learns from data and experience directly, and based on this training, tests should be applied to various needs in accordance with the necessary algorithms.

There are three type of machine learning algorithms:

A. Supervised Learning
B. Unsupervised Learning
C. Reinforcement

A. Supervised Learning

Supervised learning is defined as learning under proper supervision or as learning while a teacher is present [9]. We have a training dataset that serves as the teacher for making predictions on the given dataset, therefore there is always a training dataset when testing a dataset. The "train me" principle underpins supervised learning.

Supervised learning has following processes:

- Classification
- Random Forest
- Decision tree
- Regression

The following are the types of Supervised Learning Algorithms

1. Linear Regression
2. Logistical Regression
3. Neural Networks
4. Random Forest
5. Decision Tree

B. Unsupervised Learning

Unsupervised learning is defined as learning without instructor supervision, where no teacher is providing direction. When a student undertakes unsupervised learning dataset is provided [9]. When new data is provided, it automatically classifies it and stores it in one of the relationships after analyzing the information to identify patterns and correlations between them. Unsupervised learning is built on the idea of being "self-sufficient."

Here, Some of the Unsupervised Learning Algorithms are:

1. T-SNE
2. k- means clustering
3. PCA

C. Reinforcement

The ability of an agent to interact with the environment and ascertain the result is known as reinforced learning. It is built on the "hit and trial" principle. Each agent receives positive and negative points in reinforced learning, and on the basis of the positive points, reinforced learning produces the dataset output that was trained on the basis of the positive awards, and on the basis of this training, performs the dataset testing [10].

II. Methodology used for prediction

Data gathering is the first step in processing any system, and we use the Kaggle repository dataset for this because it has been thoroughly vetted by numerous researchers.

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed
Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

- *Collection of Data*
  For prediction of heart disease, it is necessary to train model first and test it. so, for this purpose we need data to train and test our machine learning model [1].
  In this project we have take 80% training dataset and 20% testing dataset.
- *Attribute Selection*
  Dataset attributes are characteristics of the dataset that are used for systems. some of attributes of dataset taken are Age of Person, Gender of person, Heart bit rate of a person, etc.

- Preprocessing of data
  Preprocessing needed for achieving prestigious result from the machine learning algorithms. For example, Random Forest algorithm does not support null values dataset and for this we have to manage null values from original raw data.

- Data Balancing
  Since the data balancing graph shows that both the target classes are equal, data balancing is crucial for accurate results. Fig. shows the goal classes, with "0" standing for a patient with heart illness, whereas "1" denotes a patient without heart disease [12].

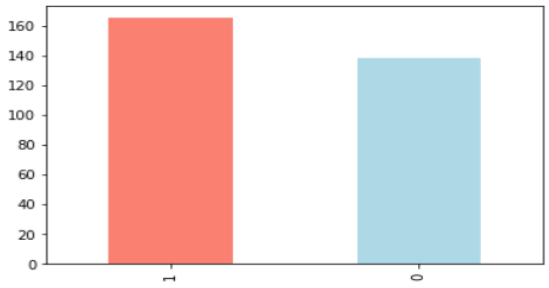


Table of Dataset:

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed
Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

- Prediction of disease

For classification, a variety of machine learning algorithms are employed, including KNN, Naive Bayes, Decision Trees, Random Forest, Logistic Regression. In order to forecast cardiac disease, various algorithms are compared, and the method that provides the highest accuracy is chosen.

III. Machine Learning Algorithm

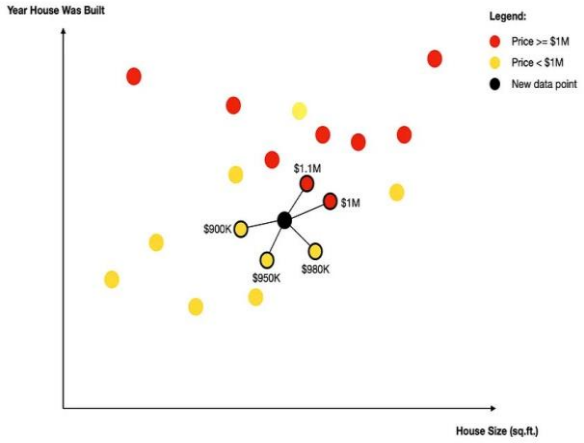
Powerful technology known as "machine learning" is the methodical study of various algorithms that gives the system the ability to mimic human learning processes without being explicitly programmed. Three categories further split machine learning: supervised learning, reinforcement learning, and unsupervised learning, Supervised Learning, and Reinforcement Learning.

1. *K-Nearest Neighbors Algorithm*

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. Pause! Let us unpack that [5].

First, let's examine the letter "k" in the kNN. We must specify to the algorithm the precise number of neighbors we wish to take into account because the algorithm bases its predictions on the nearest neighbors. As a result, "k" stands for the number of neighbors and is only a hyperparameter that can be adjusted.

Let's now assume that we have chosen k=5 and that we have a dataset with the square footage (sq. ft.), construction year, and price of each house. With the use of this data, we wish to train a kNN model, which we will then apply to forecast the cost of a different home that interests us [5].



Once the neighbors are found, one of the two things will happen depending on whether you are performing classification or regression analysis.

Classification: The algorithm selects the label for the new data point by using a simple majority vote. In our scenario, there are primarily 3 neighbors and a price of $1M. As a result, the new data point's projected label is $1M.

2. *Random Forest*

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed
Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

A supervised learning method used in machine learning is the Random Forest classifier. It can be applied to machine learning tasks including classification and regression [11]. It is based on ensemble learning, a technique for solving complicated problems and enhancing the performance of models by merging many classifiers. In order to increase the dataset's predicting accuracy, Random Forest uses multiple decision trees on different subsets of the input data. Instead of relying on just one thing, using a decision tree as a model, the random forest gathers the forecasts from each tree and predicts the outcome depending on which predictions received the most support. Better accuracy and overfitting are both prevented by the larger number of trees in the forest [11].
For a classification problem, the final output is obtained by utilizing the majority voting classifier, whereas for a regression problem, the final output is the mean of all the outputs.

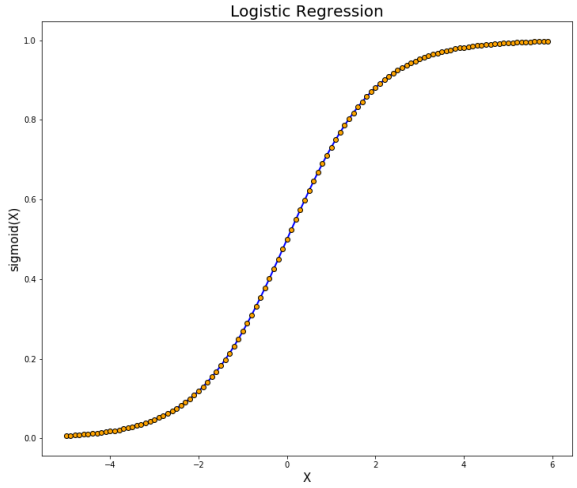
3. *Logistic Regression*

A machine learning (ML) algorithm for supervised learning - categorization analysis is logistic regression [4].

We have a labelled training dataset for classification issues that consists of categorical output variables (y) and categorical input variables (X). We can determine the best-fitting logistic function to explain the relationship between X and y using the logistic regression algorithm.

Y is a binary variable with two possible values for the traditional logistic regression, such as win/loss or good/bad. We frequently label classes as either 1 or 0, with 1 denoting the intended class of prediction, because y is binary [6].

$P(y = 1 \mid X) = p$



There are three main types of logistic regression:

1. Binary Logistic Regression

In the instance to detect an object as an animal or not, binary logistic regression was described earlier; it is an either/or approach [4]. There are only two viable responses for the outcome. In coding, this concept is frequently represented by a 0 or a 1.

2. Multinominal Logistic Regression

An object can be classified into numerous classes in a multinomial logistic regression model [6]. Before the model is run, a set of three or more preset classes is created.

3. Ordinal Logistic Regression

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed
Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

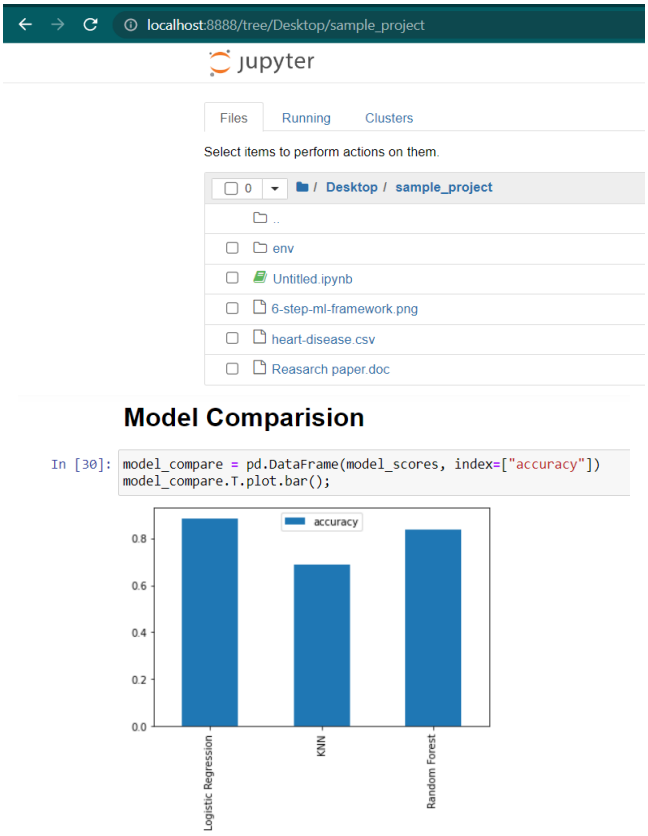
An arrangement of the classes is necessary in the case of ordinal logistic regression, which is also a model where an item can be divided into numerous classes. Class proportions are not had to be equal. Varying distances can exist between classes [6].

IV. Result Analysis

A. About Jupyter Notebook

To create and distribute documents with live code, equations, visualizations, and text, you can use the free and open-source Jupyter Notebook web application. Project Jupyter staff members are in charge of maintaining Jupyter Notebook [6].

The IPython project, which formerly had an IPython Notebook project of its own, gave rise to Jupyter Notebooks. The primary programming languages it supports are Julia, Python, and R, hence the name Jupyter. There are presently more than 100 additional kernels available, however Jupyter comes with the IPython kernel, which enables Python programming.



C. Age vs Max heart rate for heart disease

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed
Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

V. Conclusion

The heart is one of the most significant and crucial organs in the human body, and predicting heart disease is a major concern for people, thus algorithm accuracy is key.

Three ML classification modelling techniques have been used to create a model for the detection of cardiovascular disease. By extracting the patient medical history that results in a fatal heart illness from a dataset that contains patients' medical history such as chest pain, sugar level, blood pressure, etc., this method predicts persons with cardiovascular disease. Based on clinical information about a patient's prior heart disease diagnosis, this heart disease detection system helps the patient. The given model was constructed using the algorithms of KNN, Random Forest Classifier, and Logistic Regression. Our model has an accuracy rate of 87.5%.

B. Accuracy Calculation

Following the implementation of the machine learning approach for training and testing, we discover that accuracy of the Logistic Regression is higher than the other algorithms [1].

References

[1] Mrs. G. Pranitha
Assistant professor
Department of computer science and engineering
Anil neerukonda institute of technology and sciences
(Ugc autonomous)

[2] https://news.abplive.com/science/world-heart-day-2022-70-of-heart-attack-deaths-last-year-occurred-in-30-60-age-group-1555818.

[3] 1Rishabh Magar, 2Rohan Memane, 3Suraj Raut 1Prof. V. S. Rupnar 1Computer Department, 1MMCOE, Pune, India.

[4]https://www.mastersindatascience.org/learning/machine-learning-algorithms/logistic-regression/#:~:text=There%20are%20three%20main%20types,%3A%20binary%2C%20multinomial%20and%20ordinal.

[5] https://towardsdatascience.com/k-nearest-neighbors-knn-how-to-make-quality-predictions-with-supervised-learning-d5d2f326c3c2

[6]https://realpython.com/jupyter-notebook-introduction/#:~:text=The%20Jupyter%20Notebook%20is%20an,the%20people%20at%20Project%20Jupyter.

[7]https://realpython.com/jupyter-notebook-introduction/#:~:text=The%20Jupyter%20Notebook%20is%20an,the%20people%20at%20Project%20Jupyter.

[8]https://github.com/mrdbourke/zero-to-mastery-ml/blob/master/section-3-structured-data-projects/end-to-end-heart-disease-classification-video.ipynb

[9]https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

[10] Richard S. Sutton and Andrew G. Barto c 2014, 2015, The MIT Press Cambridge, Massachusetts London, England

[11] Leo Breiman Statistics Department University of California Berkeley, CA 94720

[12]https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset