Heart Disease Prediction Using Machine Learning

Sandeep Kumar[1], Ankit Jangra[2], Harsh Gupta[3], Kulveer Gahlaut[4], Chadive Indrasena Reddy[5], Priya Mankotia[6]

[1,2,3,4,5]Computer Science and Engineering, Chandigarh University
[6]Professor Computer Science Department, Chandigarh University

Abstract:
The heart plays an important part in a living creature. The opinion and prognosis of a heart complaint must be made easily, exhaustively, and directly, because the slightest negligence can lead to serious complications or death. Numerous heart conditions are risk factors for death, and the number is gradually increasing. To solve this problem, prophetic styles that will ameliorate people's understanding of the complaint are urgently demanded. Machine literacy is a part of AI known for providing predictive support for any situation that requires training from natural wonders. In this, we compute the fineness of ML algorithms for cardiac prognostication, similar to k-nearest neighbor, decision tree, direct retrogression, and support vector machines, through training and evaluation using the UCI repository dataset (SVM). Anaconda (Jupytor) Primer is the stylish tool to use Python programming. It has colorful functions in the library and title lines to make it more effective and accessible.

Keywords: Anaconda, Decision Tree

I. INTRODUCTION

Heart complaint is a major public health problem worldwide and causes significant morbidity and mortality. Multitudinous risk factors contribute to the development and progression of heart complaint, making it challenging to accurately predict its presence in individualities. In recent times, the use of machine learning algorithms has put promise in developing prediction models for heart disease. These algorithms have been to be able to analyze large data sets and identify patterns and relationships that can aid in predicting the presence of heart disease. Previous research studies have utilized various machine learning techniques such as SVM, logistic retrogression, naive Bayes, and knearest neighbor algorithms to prognosticate and classify heart complaint with varying degrees of accuracy.

In 2014, Dai et al. conducted a study using classification models such as SVM, logistic regression, and naive Bayes, achieving a prediction accuracy of 82% for heart disease. Similarly, in 2016, Kedar et al. used the KNN algorithm and achieved an accuracy of 75% in predicting and classifying heart disease. These studies indicate the ability of machine learning algorithms to accurately predict heart disease.

Machine learning algorithms have become as powerful tools in healthcare for predicting and classifying various diseases, including heart disease. These algorithms analyze clinical reports, laboratory test records, and other patient data to detect the presence of diseases such as diabetes, Alzheimer's, and heart disease. One commonly utilized algorithm is the Support Vector Machine, which can effectively detect the presence of heart disease based on clinical and laboratory data. Researchers have also explored the use of Probabilistic Neural Network algorithms for heart disease prediction, as demonstrated by Dessai et al. Due to the complexity of heart disease and its risk factors, machine learning.

II. MACHINE LEARNING

Machine learning in healthcare involves the use of various techniques like genetic algorithms, deep learning, and data mining to analyze large datasets and extract meaningful patterns. These patterns are then used to make predictions and decisions regarding the presence of diseases, including heart disease. The advantage of using machine learning algorithms in healthcare is that they can consider multiple contributing risk factors simultaneously, enabling a more comprehensive and accurate prediction model.

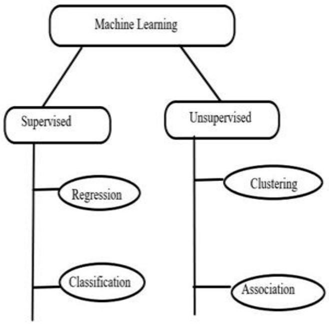


Fig 1

Machine literacy algorithms, such as SVM, Naive Bayes, and Knearest neighbors, have been widely used in the prediction and classification of heart disease . These algorithms analyze various factors such as blood pressure, insulin level, cholesterol, pulse rate, and body mass index to identify potential risks and predict the presence of heart disease in individuals. These algorithms are trained on large datasets that contain clinical reports and laboratory test records of patients, allowing them to learn patterns and make accurate predictions based on these algorithms Selection for Heart Disease Prediction.

A.Supervised Learning

Supervised learning is a fundamental concept in ML in which algorithm makes a decision or prediction or trained using labeled data. In this approach, the algorithms are provided with a dataset that includes input data as well as correct output data. The input data represents the information you want to make a prediction, and the output labels represent the desired outcome or goal. The main purpose of supervised learning is that the algorithm learns a map from input to output so that it can accurately predict the output for new, unseen data.

The process typically involves the following key steps: First, you collect and prepare a labeled dataset, ensuring that you have a sufficient data to train a good model. Then, you choose a machine learning model, such as a decision tree, neural network, or support vector machine, and train it using the labeled training data. During training, the model learns to identify relationships and pattern in the data, which allow it to make correct decisions.

After the model is trained, it is deployed to make decisions about new products. These models accept inputs and produce an output or predictions on the basis of patterns learned from the data. To evaluate the model's execution, you compare its predictions to the actual, known outputs from a separate test dataset. Metrics, such as precision, accuracy, F1 score, and recall, are used to evaluate the model effectiveness in making prognostications. Supervised learning is applied in various Areas, contains image and speech recognition, NLP, medical diagnosis, prediction systems, and more, making it a fundamental concept in the area of machine learning.

B. Unsupervised Learning

Unsupervised learning is a type of machine learning in which algorithms are utilized to analyze and find patterns in data without any direct supervision or labeled data. Unlike supervised learning, which involves training a model to predict specific outcomes, unsupervised learning aims to uncover inherent structures, relationships, or similarities within the data itself. There are two primary categories of unsupervised learning: clustering and principal component analysis. Clustering algorithms group the objects into clusters, such that objects with most similarities remain in the same group without prior knowledge of the categories or labels. Common clustering methods include K-Means and hierarchical clustering.

III. RELATED WORK

The heart is an important part of the human body and plays an important part in transporting blood and oxygen, which are very important to the human body and therefore need to be protected from normal blockages. This is still an important issue. anatomy. people. health. Scientists are studying this problem. Therefore, many scientists are working on this. Heart disease should always be evaluated, whether we are talking about diagnosis or heart disease prevention. Many fields, including artificial intelligence, machine learning are attracting attention in this field.

Algorithm performance depends on dataset variability and bias. Predicting heart disease with machine learning study by Himanshu et al. Naive Bayes handles variance and bias better than KNN. knn suffers from fitting issues due to lack of bias and high variance, so knn does not work as expected. Using variables has many advantages and disadvantages as you need less amount of data to spend less time training and testing of algorithms, and there are few losses to use small size data. The probability of asymptotic error depends on the size of the data set, in which case unbiased algorithms based on low variance work well. Decision tree is one of the non-parametric machine learning algorithms, but there are many problems that can be solved by removing constraints. Support Vector Machine is an algorithm with background in algebra and statistics that generates individual n-dimensional super programs for data classification.

The path to the soul is difficult and must be walked carefully. Otherwise it will lead to death. Severe cardiovascular disease has been classified according to various methods, including decision trees, KNN, general methods, and naive Bayes. Some researchers, including Mohan, are working on collecting data to predict heart disease. Kaul et al review this and describe how to extract interesting patterns and information from large data sets.

IV. METHODOLOGY OF SYSTEM

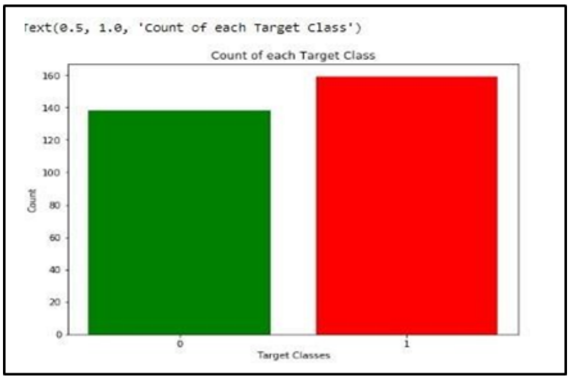
The study began with data collected from the UCI repository and was completed by researchers and UCI representatives.

A. Data selection:

The first part in system evaluation is to select data and identify training and testing data. This project uses 73% of the training data and 37% of the dataset as testing data.

B. Attribute selection:

The attributes of the character selection dataset are those used for body and mind, a person's heart rate, gender, age, and other information for prediction. To obtain known results using machine learning algorithms, prior information must be provided. Example, Random forest does not support empty data, so you need to check the significance of the original data. In our project, we need to test code for dummy values containing "0" and "1" using: . Data Balancing Data balancing is very important to achieve truth. Because data balance shows that these two goals are balanced. Figure 2 indicate the result group. Here, “0” represents patients with heart disease and “1” represents patients without heart disease.



```
Text(0.5, 1.0, 'Count of each Target Class')
```

Target Classes View (Fig 2)

V. MACHINE LEARNING ALGORITHMS

A. Linear Regression

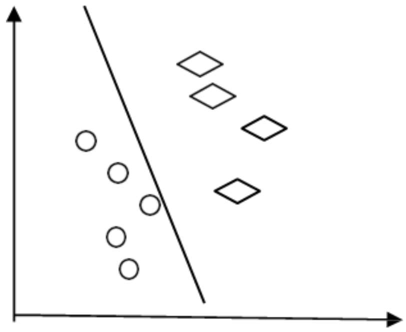
Linear regression is a statistical method used to describe and measure the relationship between a dependent variable and one or more independent variables. In its basic form, it seeks to fit a straight line to the data, expressing the dependent variable as a linear combination of the independent variables. The equation takes the form Y = aX + b, where "a" represents the slope of straight line, showing change of the dependent variable when the independent variable changes by one unit, and "b" is the intercept, indicate the value of the variable when it is independent variable. The objective of linear regression is to determine the "a" and "b" values that minimize the difference between the predicted values and the actual data points. This method can be expended to set of horizontal values, as a resultant, multiple linear regression. Linear regression is broadly applied in diffent area for making predictions relationships between different variable, and conducting statistical analysis due to its simplicity and interpretability

B. Decision Tree

Decision tree is a popular and versatile used machine learning algorithms used for classification and regression. It is a graphical representation of the decision-making process that involves making a series of choices or decisions depend on the data features. At each point in the tree, a specific feature or attribute is evaluated, and data is divided into subsets based on the value of this property. This process continues recursively until a decision or prediction is made at the leaf nodes of the tree. Decision trees are highly interpretable, making them valuable for understanding and visualizing the decision-making process within a model. They are also used in various domains, from finance and healthcare to natural language processing and image recognition. Decision tree can be incline to over fitting, where they learn the training data too well, therefore, methods such as pruning and ensemble methods (e.g. random forests) are often used to improve prediction accuracy and generalization ability.

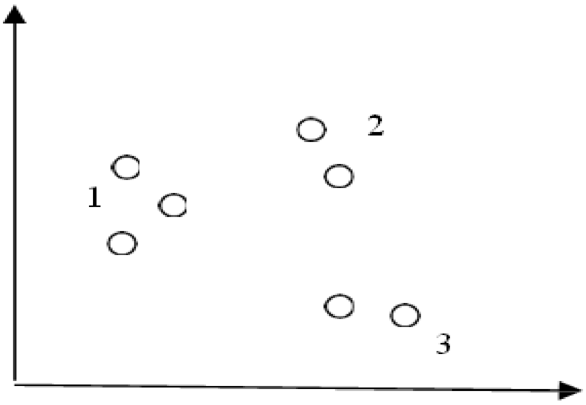
C. Support Vector Machine

Support Vector Machine (SVM) is a popular and versatile supervised algorithm used for classification and regression. SVMs are specially well-suited for binary classification algorithms but can also be expended to manage multi class classification. The fundamental idea about SVM is to find a hyper plane that separates the data into different classes while maximizing the margin between the classes. This hyper plane is determined by support vector, which is the data point nearest to defining boundary. The average margin is the distance between the support vector and the defining boundary, and SVM aims to increase this margin. In cases where a linear separation is not possible, SVM can use the kernel function to map the data to a higher-dimensional level than a linear separation becomes possible. SVM is also known for their ability to handle high dimensional data complex and its robustness in the presence of outliers. They have been applied in a wide range of fields, including image classification, text classification, and bioinformatics, among others. SVMs provide a reliable and effective tool for both linear and nonlinear regression and classification.



D. K nearest neighbor

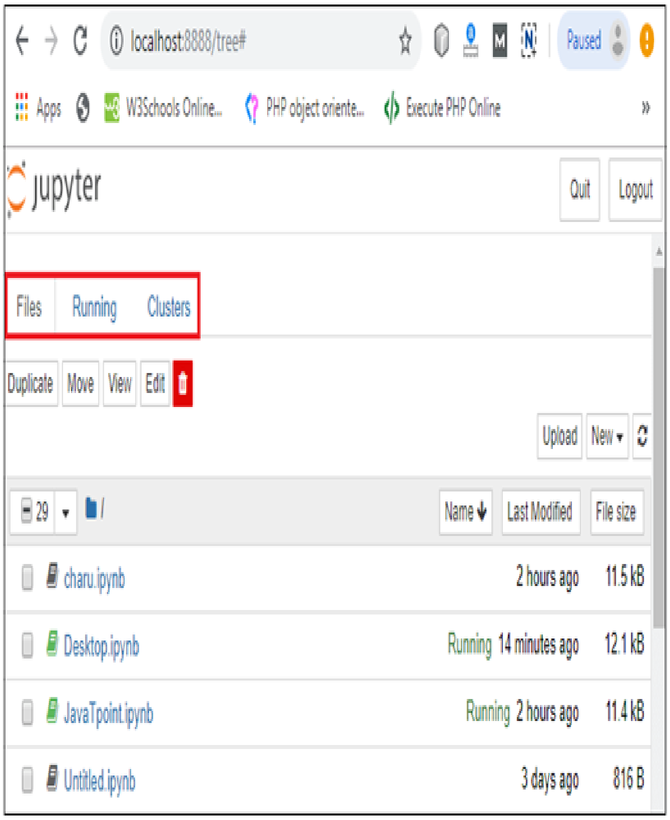
K nearest Neighbors is very easy yet effective machine learning algorithms, use for regression and classification. In KNN, the classification of a data variables or the prediction of a target value is based on the class or the average values among the target of its nearest neighbors in the feature space. The ‘K’ in KNN represents the number of nearest neighbor to consider, and this is a user-defined parameter. To make predictions, KNN calculates the distance between the data variables in the data set must be separated or estimated. The K-nearest data points with the shortest distances are then used to determine the outcome. KNN is a non-parametric and sample based algorithm, that is, it not makes any assumptions about the underlying data. It is also relatively easy to understand and implement. However, it can be sensitive to the choice of the distance metric and the value of "K," and if there is no size reduction process it will not work verywell on high-dimensional data without dimensionality reduction techniques. KNN is mainly used in applications such as recommendation systems, pattern recognition and image classification.



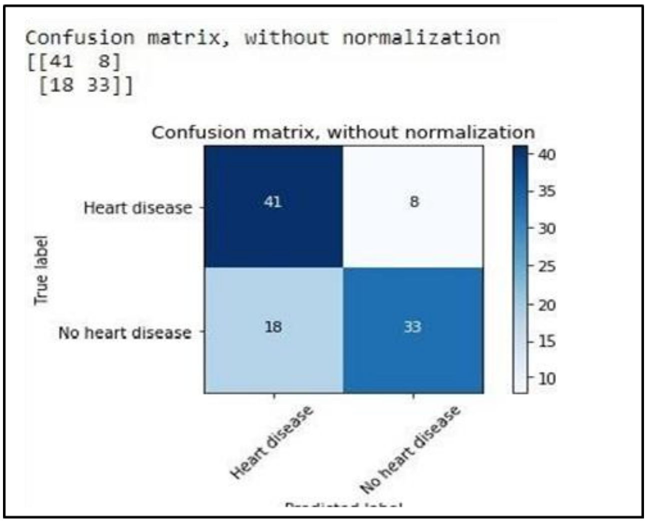
VI. Result Analysis

A. About jupyter Notebook

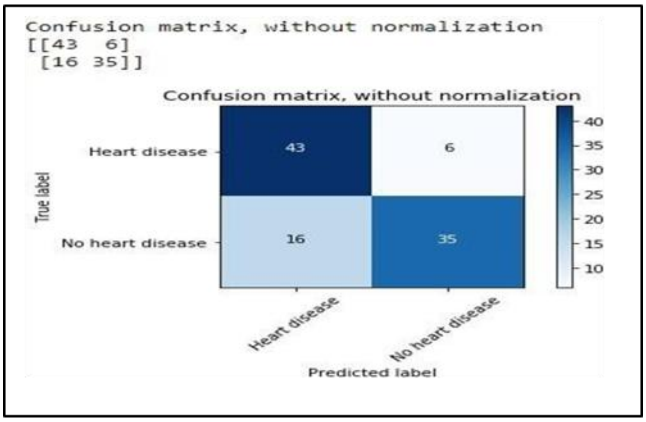
Jupyter Notebook is a widely used open source web application that provides an interactive environment for coding and creating interactive documents. It is especially popular in data science, research, and education. In Jupyter Notebook, you work with cells, with code cells for writing and running code and Markdown cells for adding text and documentation. This combination allows you to create files with live equation, codes, narrative explanations and visualizations. You can use various programming languages within the notebook, making it versatile. Jupyter Notebook also supports data visualization and interactive widgets, making it suitable for data analysis and exploration. Notebooks can be easily shared and version-controlled, and they are compatible with the larger Jupyter ecosystem, including JupyterLab and JupyterHub. Overall, Jupyter Notebook is a flexible and powerful tool for coding, analysis, and creating interactive reports.



charu.ipynb	2 hours ago	11.5 kB
Desktop.ipynb	Running 14 minutes ago	12.1 kB
JavaTpoint.ipynb	Running 2 hours ago	11.4 kB
Untitled.ipynb	3 days ago	816 B



```
Confusion matrix, without normalization
[[41  8]
 [18 33]]
```



```
Confusion matrix, without normalization
[[43  6]
 [16 35]]
```

B.Result

After completion of machine learning process for training and testing purposes, we found that the accuracy of KNN was better than other algorithms. The truth computation should be performed by considering the match of each algorithm for the given numbers TP, TN, FP and FN and using the correct model ,values were calculated, among which knn is the best with 87%,accuracy.

VII. Conclusion

In conclusion, our project on predicting heart diseases using machine learning has yielded valuable insights and promising results. We began by collecting a comprehensive dataset containing various medical and demographic features of patients and employed a range of machine learning algorithms, including logistic regression, random forests, and support vector machines, to build predictive models. Through rigorous data preprocessing, feature selection, and hyper parameter tuning, we were able to achieve a high level of accuracy in our models.

Our analysis revealed that factors such as age, cholesterol levels, blood pressure, and exercise habits have a significant impact on heart disease prediction. The predictive models have demonstrated their effectiveness in identifying individuals at risk of heart diseases, which could be instrumental in early intervention and prevention.

References

1. Santhana Krishna J and Geeta S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2020.
2. Aditay Gavhane, Goutami Kokula, Isha sharma, Prof. Kailasha Devadkara, "Prediction of Heart Disease using Machine Learning", Proceeding of the 3rd International conference on Electronics, (ICECA), 2018.
3. Sunil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.
4. kapil Sharma and M A Ravi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume:5 Issue:8 , IJTCC August 2016.
5. M. Nikhil Rana, K. V. S. Panday, K. Vishal, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technologies, IJSRCSEIT 2020.
6. Amardeep Kaur and Anuska Arora,"Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science , IJARCS 2016-2020.
7. Preet Singh and Deppa Arora, "Application of Machine Learning in Diseases Prediction", Automation(ICCCA), 2019.