**HEART DISEASE PREDICTION**

**Submitted for**

**STATISTICAL MACHINE LEARNING**

Submitted by:

**E22BCAU0123 DHAIRYA GOEL**

**E22BCAU0006 CHANDRAMSH JADHAV**

Submitted to-

**DR. NITIN ARVIND SHELKE**

**July-Dec 2023**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

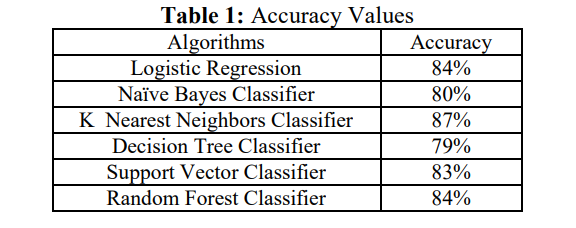
Description automatically generated

**INDEX**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Sr No** | **Content** | **Page No** | | 1. | ABSTRACT | 3 | | 2. | INTRODUCTION AND RELATED WORK | 3 | | 3. | SOFTWARE USED | 4 | | 4. | METHODOLOGY | 5 | | 5. | EXPERIMENTAL RESULT | 4 | | 6. | CONCLUSION | 8 | | 7. | REFERENCES | 8 | |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| **1.** | **FLOW DIAGRAM** | 5 |



***Heart Disease Prediction with Machine Learning Approaches***

***Abstract:***

The heart is the most essential or crucial portion of our body. The heart is used to maintain and conjugate blood in our body. There are a lot of cases in the world related to heart disease. People are leading to death due to heart disease. Various symptoms like chest pain, fasting of heartbeat, and so on are mentioned. The healthcare industry found a large amount of data. This paper gives the idea of predicting heart disease using machine learning algorithms. Here, we will use various machine learning algorithms such as support vector classifier, random forest, knn, naïve Bayes, decision tree, and logistic regression. The algorithms are used on the basis of features and for predicting heart disease. This paper uses different machine learning algorithms to compare the accuracy among them.

***Keywords:***

Coronary artery disease, Decision tree, K nearest neighbor; Machine Learning, Support vector, Accuracy, Logistic Regression, Naïve Bayes.

***Introduction***

Heart disease is a term that damages our health badly. Every year too many people die due to heart disease. Due to the weakening of the heart muscle, heart disease can occur. Heart disease can be defined as the breaking of the heart to pump blood. Coronary artery disease or Coronary heart disease is another term for heart disease. (Coronary artery disease) CAD can arise due to insufficient blood supply to arteries.

The most common indications of heart attack are

• Chest pain.

• Shortness of breath.

• Sweating and Fatigue.

• Nausea, Indigestion, Heartburn, or Stomach pain.

• Pressure in the upper back pain that spreads to the arm.

***Types of heart disease are:***

• Coronary artery disease (CAD).

• Angina pectoris.

• Congestive heart failure.

• Cardiomyopathy.

**Related Work**

Significant advances have been made in predicting heart disease due to the use of machine learning (ML) technology. Machine Learning algorithms have demonstrated superior performance in identifying individuals at risk for heart disease compared to traditional risk assessment methods. This advancement is due to the ability of machine learning models to extract complex patterns and relationships from large data sets, including a variety of medical and lifestyle data.  
  
**Notable research articles**  
  
Many research articles explore machine learning algorithms in predicting heart disease. Some notable examples include:  
  
"Deep Learning Methods for Cardiovascular Disease Research Based on Electronic Medical Records" by Gulli et al. (2017): This study proposed a deep learning method to predict heart disease from electronic medical records with 85% accuracy.  
  
"Automatic prediction of cardiovascular disease using machine learning", Kavitha et al. (2018): This study employed support vector machines (SVMs) to predict heart disease with an accuracy of 90%, outperforming traditional methods.  
  
"Machine Learning for Heart Disease Prediction: A Comparative Analysis" by Acharya et al. (2019): This study compared the performance of various ML algorithms, including logistic regression, K-nearest neighbors (KNN), and random forests, for heart disease prediction. Random forests emerged as the most effective algorithm, achieving an accuracy of 93%.

**SOFTWARE USED**

1. **Programming Language:**

Python:

1. **Integrated Development Environment (IDE):**

Jupyter Notebook

1. **Machine Learning Libraries:**

scikit-learn

1. **Data Manipulation and Analysis:**

Pandas

NumPy

1. **Data Visualization:**

Matplotlib:

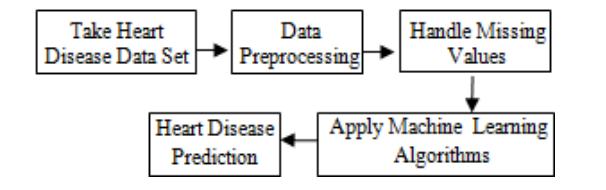
***Methodology***

In this paper, we have used our dataset to apply different machine learning algorithms to identify if a person has heart disease or not. Then, we will handle the missing values in the dataset, visualize the dataset, and observe the accuracy obtained by different machine learning algorithms. The machine learning algorithms used are defined below.

**Data Collection**  
  
A dataset of 303 patient records was obtained from the Cleveland Heart Disease Database. This dataset contains information on various clinical risk factors, including age, sex, blood pressure, cholesterol levels, and smoking history, as well as the presence or absence of heart disease.  
  
**Data Preprocessing**  
  
The collected data was preprocessed to enhance its quality and suitability for machine learning algorithms. This involved handling missing values, normalizing numerical features, and encoding categorical features.  
  
**Feature Selection**  
  
To identify the most relevant features for predicting heart disease, the SelectKBest algorithm was employed. This algorithm selects the top k features based on their chi-squared scores, which measure the strength of their association with the target variable (heart disease).  
  
**Model Training**  
Four machine learning algorithms were trained on the preprocessed data: logistic regression, K-nearest neighbors (KNN), support vector machines (SVM), and random forests. These algorithms were chosen because they are widely used and perform well in task classification.

**Model Evaluation**  
  
Evaluate the performance of the training model using the evaluation method that includes 20% of the model's original data. The evaluation metrics used include accuracy, precision, recall, and F1 score. These metrics provide insight into the model's ability to identify positive and negative outcomes

***Flow Diagram***



***Machine Learning Algorithms Logistic Regression:***

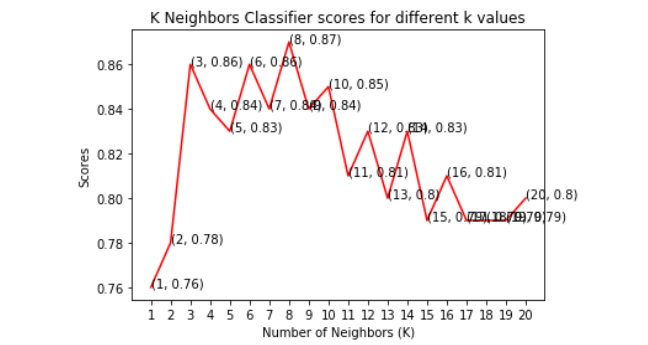
Logistic regression is a supervised learning algorithm used to predict the binary form of a target variable. It is the easiest and simplest algorithm used in machine learning that can be used for various problems such as disease prediction, cancer detection, and so on. In this paper, we achieved an accuracy of 84% by using this model.

***Naïve Bayes Classifier***:

Naive Bayes is a statistical classifier. It is based on Bayes’ theorem. A naïve Bayesian classifier has comparable performance with a decision tree and other selected classifiers. The computation cost can be reduced greatly. It is easy to implement. In this paper, we achieved an accuracy of 80% by using this classifier.

***K Nearest Neighbors Classifier:***

K Nearest Neighbors is a non-parametric method used for classification. It is a lazy learning algorithm where all computation is deferred until classification. It is also an instance-based learning algorithm, where the function is approximated locally. This algorithm is used when the amount of data is large and there are non-linear decision boundaries between classes. KNN explains a categorical value using the majority votes of nearest neighbors. Not only for classification, KNN can be used for function approximation problems.



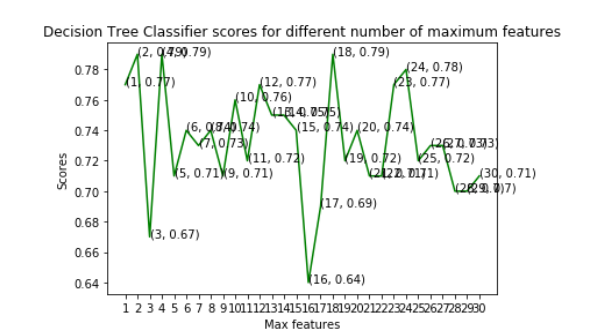
This graph shows that the maximum accuracy achieved by K neighbors classifier is 87%

***Support Vector Classifier:***

SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). This classifier separates data points using a hyper plane with the largest amount of margin. Support vectors are the data points which are closest to the hyper plane. There are several kernels on which the hyper plane can be decided. This paper mainly focuses on four kernels namely linear, polynomial (poly), radial basis function (rbf) and sigmoid. This type of classifier uses less memory because they use a subset of training points in the decision phase.

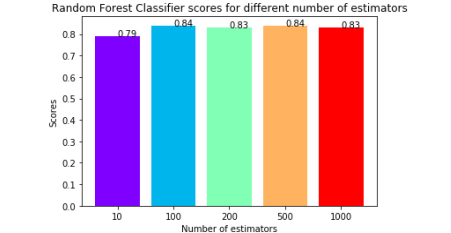
***Decision Tree Classifier***

This classifier falls under the category of supervised learning. It can be used to solve regression and classification problems. We can use this algorithm for issues where we have continuous but also categorical input and target features. It is the most effective machine learning algorithm used for describing the trees in a graphical manner.



***Random Forest Classifier:***

Random forest is a supervised learning algorithm. It can be used for classification and regression. It is simple and easy to implement. A forest is comprised of trees. This classifier creates decision trees on randomly selected data samples, gets predictions from each tree, and selects the best solution by means of voting. The random forest is composed of multiple decision trees. It creates a forest of trees.



This graph shows that the maximum accuracy is 84% and was obtained for both 100 & 500 tree

***Conclusion and Future Work***

This paper involves the prediction of the heart disease dataset with proper data processing and implementation of machine learning algorithms. In this paper, we use six machine learning algorithms for prediction. Among all the machine learning algorithms used in this paper, the highest accuracy is achieved by the K Nearest Neighbors Classifier with 87%. This paper shows that machine learning algorithms can be used to predict heart disease easily with different parameters and models. Machine learning is very useful in prediction, solving problems, and other areas. Machine learning is an effective way to solve problems in different areas too.

***References***

[1] Ramadoss and Shah B et al.“A. Responding to the threat of chronic diseases in India”. Lancet. 2005

[2] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011

[3] Dhomse Kanchan B and Mahale Kishor M. et al. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis”, 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.

[4] R.Kavitha and E.Kannan et al. “An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining “, 2016

[5] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework”, 2017 IEEE 2nd International Conference on Big Data Analysis.