PROJECT TITLE:

Visualizing and Predicting Heart Diseases with an Interactive Dash Board

TEAM ID:

PNT2022TMID12709

TEAM MEMBERS:

1. Sabarishkumar T - 19L139 (Team Lead)
2. Mariyon Sunil Larsen J - 19L131
3. Srinath S P - 19L144
4. Shivakumar V - 19L141

CHAPTER 1

INTRODUCTION

The World Health Organization (WHO) claims heart disease kills 17.7 million people each year [14]. Heart disease is a common and serious ailment that people face, accounting for 80 percent of all fatalities in the country. By 2030, this is anticipated that 25 million people will die of heart disease. Coronary heart disease was responsible for an estimated 7.2 million fatalities [15]. In India, heart disease is currently the top cause of death, posing a considerable danger to both men and women. According to the Indian Heart Association (IHA), four individuals die of heart disease in India per minute, with the majority of casualties will be between the ages of 30 and 50. One-fourth of all heart failure mortality is attributed to those below the age of 40.

In India, nine hundred persons under the age of 30 die every day from various cardiac problems [18]. Heart disease is thought to be caused by several factors. The diagnosis of heart disease would be difficult because of the certain high prevalence of diabetes, high blood pressure, excessive triglycerides, improper pulse rate, and several other problems. More people die of heart disease after their first heart attack than for any other reason. Consequently, various challenges in the areas of breast cancer, emphysema, the ventricle, and heart attacks have already been addressed. Increased blood viscosity is a major risk factor for cardiovascular disease. Blood's very viscous composition prevents it from flowing freely, causing blood flow resistance [16]. Heart disease is on the rise as a result of our contemporary lifestyle.

Our way of living had a significant influence on our health, resulting in heart disease and other health issues [17]. As a result, assessing heart illness is a difficult process. Even back then, various researchers tried several methods to predict heart disease development. Traditional methodologies, on the other hand, are restricted in their capacity to evaluate such a big quantity of data, but sophisticated technologies like visual analytics may help anticipate cardiac disease and pinpoint the most important aspects. Over examining the system, the majority of these studies chose pattern recognition or a deep learning-based technique [11]. Although much study is being done to diagnose heart illness utilizing more advanced technology, we're also looking towards a more advanced technique that uses an intuitive graphical dashboard to show the best way to evaluate the key parts of heart disease.

As a result, this research is critical because it allows for the detection of heart disease and significant risk factors. As a consequence, to solve the constraints noted above, we created HDVis, an interactive visual system (Vis) for understanding cardiac illness. The visualization approach is being used to provide a more complete solution, assisting end-users in assessing heart problems and identifying critical components that are substantially linked with heart disease. To accomplish so, we employed a qualitative research method and learn more about the factors that influence heart.

Heart disease is the major cause of deaths globally. More people die annually from CVDs than from any other cause, an estimated 12 million people died from heart disease every year. Heart disease kills one person every 34 seconds in the United States. Heart attacks are often a tragic event and are the result of blocking blood flow to the heart or brain. People at risk of heart disease may show elevated blood pressure, glucose and lipid levels as well as stress. All of these parameters can be easily measured at home by basic health facilities. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are the categories of heart disease.

The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) causes many diseases, disability and death. Diagnosis of the disease is important and complex work in medicine.

Medical diagnosis is considered as crucial but difficult task to be done efficiently and effectively .The automation of this task is very helpful. Unfortunately all physicians are not experts in any subject specialists and beyond the scarcity of resources there some places. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision making. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important.

One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest to predict the heart disease based on some health parameters.

Heart disease is the major cause of deaths globally. More people die annually from CVDs than from any other cause, an estimated 12 million people died from heart disease every year. Heart disease kills one person every 34 seconds in the United States. Heart attacks are often a tragic event and are the result of blocking blood flow to the heart or brain.

People at risk of heart disease may show elevated blood pressure, glucose and lipid levels as well as stress. All of these parameters can be easily measured at home by basic health facilities. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are the categories of heart disease. The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) causes many diseases, disability and death.

Diagnosis of the disease is important and complex work in medicine. Medical diagnosis is considered as crucial but difficult task to be done efficiently and effectively .The automation of this task is very helpful. Unfortunately all physicians are not experts in any subject specialists and beyond the scarcity of resources there some places. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision making. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public.

The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest to predict the heart disease based on some health parameters.

1.1 ALGORITHMS USED

1.1.1 Logistic Regression

Classification techniques are an essential part of machine learning and data mining applications. Approximately 70% of problems in Data Science are classification problems. There are lots of classification problems that are available, but logistic regression is common and is a useful regression method for solving the binary classification problem. Another category of classification is Multinomial classification, which handles the issues where multiple classes are present in the target variable. For example, the IRIS dataset is a very

famous example of multi-class classification. Other examples are classifying article/blog/document categories.

Logistic Regression can be used for various classification problems such as spam detection. Diabetes prediction, if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on a given advertisement link or not, and many more examples are in the bucket.

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. Its basic fundamental concepts are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Where, y is a dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

$$p = \frac{1}{1 + e^{-y}}$$

Apply Sigmoid function on linear regression:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}$$

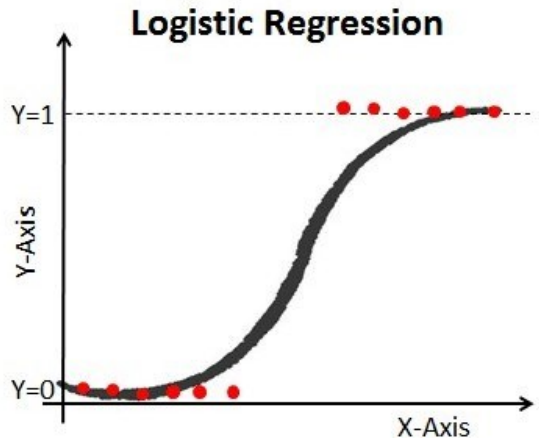

Fig.1: Logistic Regression

1.1.2 Naïve Bayes

Bayes' Theorem provides a way that we can calculate the probability of a piece of data belonging to a given class, given our prior knowledge. Bayes' Theorem is stated as:

P(class|data) = (P(data|class) * P(class)) / P(data)

Where P(class|data) is the probability of class given the provided data.

Naive Bayes is a classification algorithm for binary (two-class) and multiclass classification problems. It is called Naive Bayes or idiot Bayes because the calculations of the probabilities for each class are simplified to make their calculations tractable.

Rather than attempting to calculate the probabilities of each attribute value, they are assumed to be conditionally independent given the class value.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

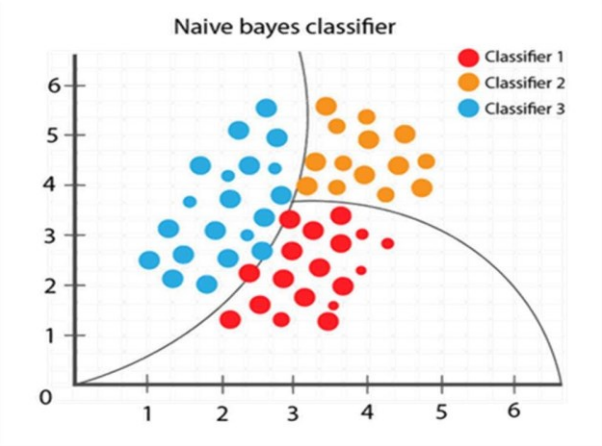

Fig.2: Naïve Bayes Algorithm

1.1.3 KNN Algorithm

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

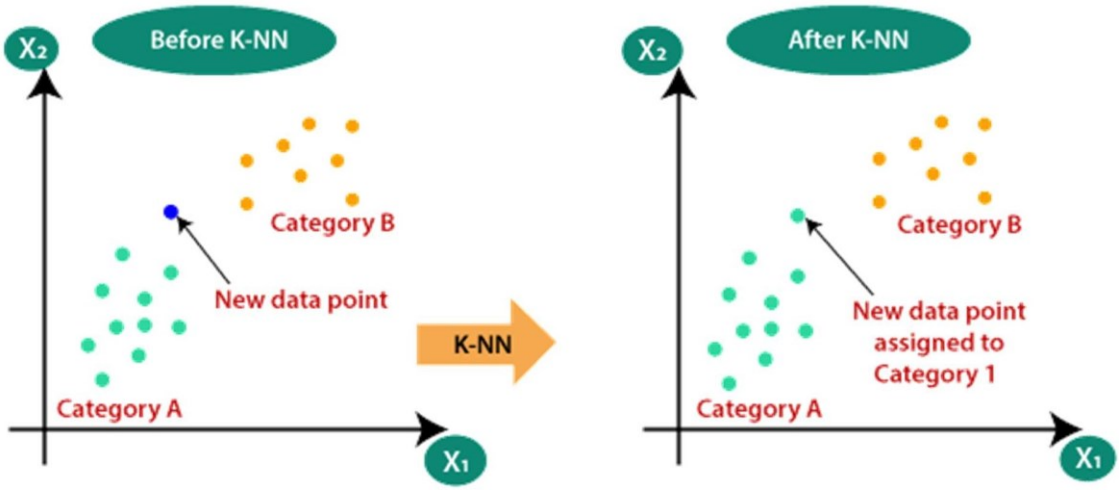

Fig.3: KNN Algorithm

CHAPTER 2

LITERATURE SURVEY

2.1 Prediction of heart disease and classifiers' sensitivity analysis

Author: Khaled Mohamad Almustafa

Publisher: Bioinformatics (2020)

Heart disease (HD) is one of the most common diseases nowadays, and an early diagnosis of such a disease is a crucial task for many health care providers to prevent their patients for such a disease and to save lives. In this paper, a comparative analysis of different classifiers was performed for the classification of the Heart Disease dataset in order to correctly classify and or predict HD cases with minimal attributes. The set contains 76 attributes including the class attribute, for 1025 patients collected from Cleveland, Hungary, Switzerland, and Long Beach, but in this paper, only a subset of 14 attributes are used, and each attribute has a given set value. The algorithms used K- Nearest Neighbor (K-NN), Naive Bayes, Decision tree J48, JRip, SVM, Adaboost, Stochastic Gradient Decent (SGD) and Decision Table (DT) classifiers to show the performance of the selected classifications algorithms to best classify, and or predict, the HD cases.

It was shown that using different classification algorithms for the classification of the HD dataset gives very promising results in term of the classification accuracy for the K-NN (K = 1), Decision tree J48 and JRip classifiers with accuracy of classification of 99.7073, 98.0488 and 97.2683% respectively. A feature extraction method was performed using Classifier Subset Evaluator on the HD dataset, and results show enhanced performance in term of the classification accuracy for K-NN (N = 1) and Decision Table classifiers to 100 and 93.8537% respectively after using the selected features by only applying a combination of up to 4 attributes instead of 13 attributes for the predication of the HD cases.

2.2 Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering

Author : D.Kumar

Publisher: Semantic scholar on 2015

The diagnosis of diseases is a crucial and difficult job in medicine. The recognition of heart disease from diverse features or signs is a major issue which is not free from false presumptions often accompanied by unpredictable effects. The healthcare industry gathers enormous amounts of heart disease data that unfortunately, are not mined to determine concealed information for effective diagnosing. Due to this rapid growth is the main motivation for researchers to mine useful information from these medical databases. As the volume of stored data increases, data mining techniques play an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Heart disease prediction suffers from the problem of missing data, statistical tests will lose power, results may be based, or analysis may not be feasible at all. There are several ways to handle the problem, for example through imputation.

To overcome this problem initially, the data set containing 13 medical attributes were obtained from the Cleveland heart disease database missing attributes data is replaced with the help of imputation method. With imputation, missing values are replaced with estimated values according to an imputation method or model. In this paper, preprocessed dataset from EM is given as input to clustering method for heart disease prediction. In this paper, an efficient approach non negative matrix factorization with hierarchical clustering methods (NMF-HC) is proposed for the intelligent heart disease prediction. The dataset is clustered with the aid of NMF-HC clustering algorithm. The NMF-HC is trained using the preprocessed data sets. The proposed NMF-HC works as promising tool for prediction of heart disease.

2.3 Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification

Author: Ashir

Publisher: Hindawi on 2020

Diagnosis of heart disease is a difficult job, and researchers have designed various intelligent diagnostic systems for improved heart disease diagnosis. However, low heart disease prediction accuracy is still a problem in these systems. For better heart risk prediction accuracy, we propose a feature selection method that uses a floating window with adaptive size for feature elimination (FWAFE). After the feature elimination, two kinds of classification frameworks are utilized, i.e., artificial neural network (ANN) and deep neural network (DNN). Thus, two types of hybrid diagnostic systems are proposed in this paper, i.e., FWAFE-ANN and FWAFE-DNN. Experiments are performed to assess the effectiveness of the proposed methods on a dataset collected from Cleveland online heart disease database.

The strength of the proposed methods is appraised against accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and receiver operating characteristics (ROC) curve. Experimental outcomes confirm that the proposed models outperformed eighteen other proposed methods in the past, which attained accuracies in the range of 50.00–91.83%. Moreover, the performance of the proposed models is impressive as compared with that of the other state-of-the-art machine learning techniques for heart disease diagnosis. Furthermore, the proposed systems can help the physicians to make accurate decisions while diagnosing heart disease.

2.4 Analysis of Heart Disease Using Parallel and Sequential Ensemble Methods With Feature Selection Techniques: Heart Disease Prediction

Author: Dhyan Chandra Yadav

Publisher : Researchgate , Jan - 2021

This paper has organized a heart disease-related dataset from UCI repository. The organized dataset describes variables correlations with class-level target variables. This experiment has analyzed the variables by different machine learning algorithms. The authors have considered prediction-based previous work and finds some machine learning algorithms did not properly work or do not cover 100% classification accuracy with overfitting, underfitting, noisy data, residual errors on base level decision tree. This research has used Pearson correlation and chi-square features selection-based algorithms for heart disease attributes correlation strength.

The main objective of this research to achieved highest classification accuracy with fewer errors. So, the authors have used parallel and sequential ensemble methods to reduce above drawback in prediction. The parallel and serial ensemble methods were organized by J48 algorithm, reduced error pruning, and decision stump algorithm decision tree-based algorithms. This paper has used random forest ensemble method for parallel randomly selection in prediction and various sequential ensemble methods such as AdaBoost, Gradient Boosting, and XGBoost Meta classifiers. In this paper, the experiment divides into two parts: The first part deals with J48, reduced error pruning and decision stump and generated a random forest ensemble method. This parallel ensemble method calculated high classification accuracy 100% with low error.

The second part of the experiment deals with J48, reduced error pruning, and decision stump with three sequential ensemble methods, namely AdaBoostM1, XG Boost, and Gradient Boosting. The XG Boost ensemble method calculated better results or high classification accuracy and low error compare to AdaBoostM1 and Gradient Boosting ensemble methods. The XG Boost ensemble method calculated 98.05% classification accuracy, but random forest ensemble method calculated high classification accuracy 100% with low error.

2.5 Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms

Author: Kaushalya Dissanayake

Publisher : Hindawi ,Nov - 2021

Heart disease is recognized as one of the leading factors of death rate worldwide. Biomedical instruments and various systems in hospitals have massive quantities of clinical data. Therefore, understanding the data related to heart disease is very important to improve prediction accuracy. This article has conducted an experimental evaluation of the performance of models created using classification algorithms and relevant features selected using various feature selection approaches. For results of the exploratory analysis, ten feature selection techniques, i.e., ANOVA, Chi-square, mutual information, ReliefF, forward feature selection, backward feature selection, exhaustive feature selection, recursive feature elimination, Lasso regression, and Ridge regression, and six classification approaches, i.e., decision tree, random forest, support vector machine, K-nearest neighbor, logistic regression, and Gaussian naive Bayes, have been applied to Cleveland heart disease dataset. The feature subset selected by the backward feature selection technique has achieved the highest classification accuracy of 88.52%, precision of 91.30%, sensitivity of 80.76%, and f-measure of 85.71% with the decision tree classifier.

2.6 Early Prediction of Heart Diseases Using Data Mining Techniques

Author: Vikas Chaurasia

Publisher : Researchgate,2013

Largest-ever study of deaths shows heart diseases have emerged as the number one killer in world. About 25 per cent of deaths in the age group of 25-69 years occur because of heart diseases. If all age groups are included, heart diseases account for about 19 per cent of all deaths. It is the leading cause of death among males as well as females. It is also the leading cause of death in all regions though the numbers vary. The proportion of deaths caused by heart disease is the highest in south India (25 per cent) and lowest -12 per cent -in the central region of India. The prediction of heart disease survivability has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease survivability. We used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a large dataset. We also used 10-fold cross-validation methods to measure the unbiased estimate.

2.7 Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks

Author : Vanisree K , Jyothi Singaraju

Publisher : Academic Journals Database

Congenital Heart Disease is one of the major causes of deaths in children. However, a proper diagnosis at an early stage can result in significant life saving. Unfortunately, all the physicians are not equally skilled, which can cause for time delay, inaccuracy of the diagnosis. A system for automated medical diagnosis would enhance the accuracy of the diagnosis and reduce the cost effects. In the present paper, a Decision Support System has been proposed for diagnosis of Congenital Heart Disease. The proposed system is designed and developed by using MATLABs GUI feature with the implementation of Backpropagation Neural Network. The Backpropagation Neural Network used in this study is a multi layered Feed Forward Neural Network, which is trained by a supervised Delta Learning Rule. The dataset used in this study are the signs, symptoms and the results of physical evaluation of a patient. The proposed system achieved an accuracy of 90 percent.

2.8 Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis

Author : Sameh Ghwanmeh

Publisher : Journal of Intelligent Learning Systems and Applications , 2013

Heart diagnosis is not always possible at every medical center, especially in the rural areas where less support and care, due to lack of advanced heart diagnosis equipment. Also, physician intuition and experience are not always sufficient to achieve high quality medical procedures results. Therefore, medical errors and undesirable results are reasons for a need for unconventional computer-based diagnosis systems, which in turns reduce medical fatal errors, increasing the patient safety and save lives. The proposed solution, which is based on an Artificial Neural Networks (ANNs), provides a decision support system to identify three main heart diseases: mitral stenosis, aortic stenosis and ventricular septal defect.

Furthermore, the system deals with an encouraging opportunity to develop an operational screening and testing device for heart disease diagnosis and can deliver great assistance for clinicians to make advanced heart diagnosis. Using real medical data, series of experiments have been conducted to examine the performance and accuracy of the proposed solution. Compared results revealed that the system performance and accuracy are acceptable, with a heart diseases classification accuracy of 92%.In this paper, it is proposed to build an ANN-based decision support system for heart diseases diagnosis.

The system comprises two main components: a software part and a hardware part, which consists of a simple electronic stethoscope with a matching impedance electronic circuit. It is expected that the proposed system would benefit medical physicians to diagnose heart sound signals and checking the up-normality. Also, it is anticipated that the proposed system would provide innovative diagnostic tool to classify the heart diseases: mitral stenosis (MS), aortic stenosis (AS) and ventricular septal defect (ASD). Also, the system would offer a promising opportunity to develop an operational screening and testing device for heart disease diagnosis.

The performance, classification accuracy and effectiveness of the proposed NN-based decision support system have been examined. The testing experiments have been conducted to provide a clear comparison between the proposed system and other systems and techniques. Performance evaluation process has been conducted to reveal the best value of heart diseases classification accuracy.

2.9 Medical diagnostic systems: a case for neural networks

Author : C N Schizas

Publisher : National Library of Biomedical information , 2019

Recent advances in computer technology offer to the medical profession specialized tools for gathering medical data, processing power, as well as fast storing and retrieving capabilities. Artificial intelligence (AI), an emerging field of computer science is studying the issues of human problem solving and decision making. Furthermore, rule-based systems and knowledge-based systems that are other fields of AI have been adopted by many scientists in an effort to develop intelligent medical diagnostic systems. In this study artificial neural networks (ANN) are introduced as a tool for building an intelligent diagnostic system; the system does not attempt to replace the physician from being the decision maker but to enhance ones facilities for reaching a correct decision.

An integrated diagnostic system for assessing certain neuromuscular disorders is used in this study as an example for demonstrating the proposed methodology. The diagnostic system is composed of modules that independently provide numerical data to the system from the clinical examination of a patient, and from various laboratory tests that are performed. The examination procedure has been standardized by developing protocols for each specialized area, in cooperation with experts in the area. At the conclusion of the clinical examination and laboratory tests, data in the form of a numerical vector represents a medical examination snapshot of the subject.

Artificial neural network (ANN) models were developed using the unsupervised self-organizing feature maps algorithm. Data from 71 subjects were collected. The ANN models were trained with the data from 41 subjects, and tested with the data from the remaining 30 subjects. Two sets of models were developed; those trained with the data from only the clinical examinations; and those trained by combining the clinical and the laboratory test data. The diagnostic yield that was obtained for the unknown cases is in the region of 73 to 93% for the models trained with only the clinical data, and in the region of 73 to 100% for those trained by combining both the clinical and laboratory data. The pictorial representation of the diagnostic models through the self organized two dimensional feature maps provide the physician with a friendly human-computer interface and a comprehensive tool that can be used for further observations, for example in monitoring disease progression of a subject.

2.10 Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data

Author : JanainaMourão-Miranda

Publisher : NeuroImage , 2015

In the present study, we applied the Support Vector Machine (SVM) algorithm to perform multivariate classification of brain states from whole functional magnetic resonance imaging (fMRI) volumes without prior selection of spatial features. In addition, we did a comparative analysis between the SVM and the Fisher Linear Discriminant (FLD) classifier. We applied the methods to two multisubject attention experiments: a face matching and a location matching task. We demonstrate that SVM outperforms FLD in classification performance as well as in robustness of the spatial maps obtained (i.e. discriminating volumes). In addition, the SVM discrimination maps had greater overlap with the general linear model (GLM) analysis compared to the FLD. The analysis presents two phases: during the training, the classifier algorithm finds the set of regions by which the two brain states can be best distinguished from each other. In the next phase, the test phase, given an fMRI volume from a new subject, the classifier predicts the subject's instantaneous brain state.

CHAPTER 3

SYSTEM SPECIFICATION

3.1 HARDWARE REQUIREMENTS

Hard Disk : 250GB

RAM : 2GB

3.2 SOFTWARE REQUIREMENTS

Storage File : CSV

Tool Kit : Jupyter Notebook, Tableau

Operating System : Windows 10

CHAPTER 4

SYSTEM DESIGN AND ARCITECTURE

System design is the systematic process of defining the components of the purposed system that consists of model, architecture, and interface of different elements. It describes the operation of a system that demonstrates data flow structure and a link between the database tables (Odhiambo, 2018). In this system design, we are going to design a procedure programming system design related to our project. They can be a Context diagram, DFD Level-1, DFD Level-2. It is an approach to design the system to organize a way to easily understand the whole system.

The system architecture (Kautish et al, 2016, 2018, 2019) is like a blueprint of any object. It is a conceptual model to integrate between business logic and physical system in an organized way. It demonstrates the structure, view, behavior, features, and functionalities of the system. It is the way of portraying the desired system in visualizing a way to well understand for people. The System architecture is the foundational orchestrate of a system that incorporated in its elements, their relationships of elements, and the science of its design (MITRE, 2019).

HDPS is a web-based application that runs on the browser. This system is embodied in a web application. The web application architecture of the HDPS is to define the communication between applications, and database on the web. It also helps to know about other third-party application required like python's packages. It represents the represent the relationship between them and visualizes how they work together simultaneously.

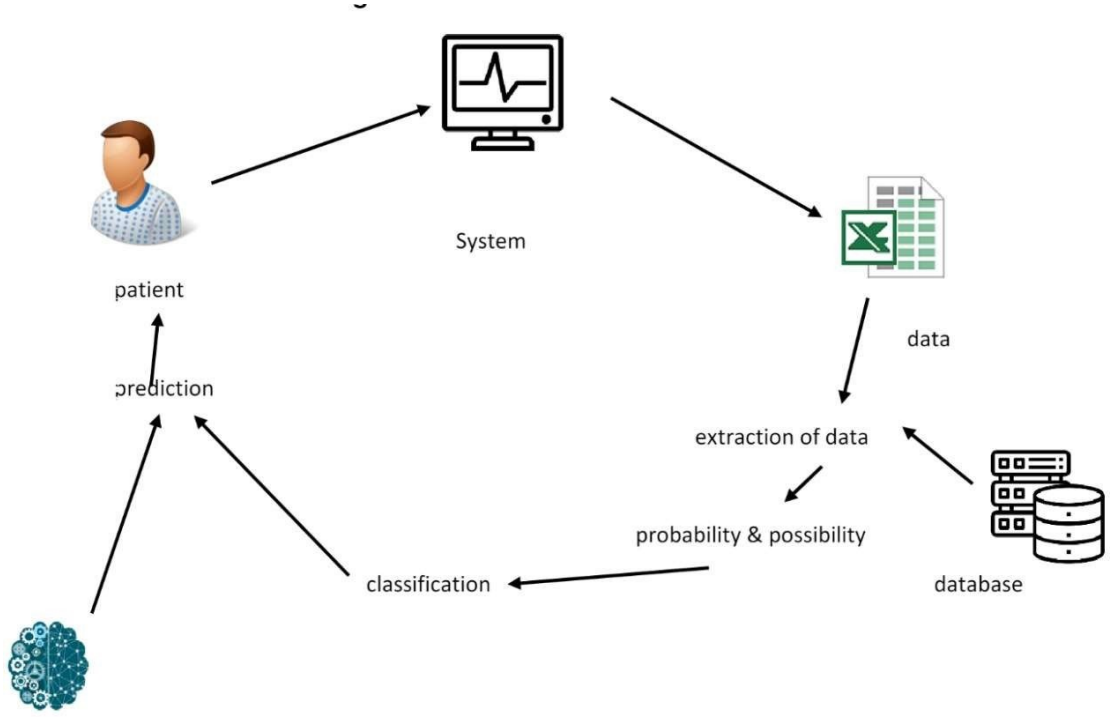

Fig.4: System Design

4.1 SOLUTION ARCHITECTURE

System design is the systematic process of defining the components of the purposed system that consists of model, architecture, and interface of different elements. It describes the operation of a system that demonstrates data flow structure and a link between the database tables (Odhiambo, 2018). In this system design, we are going to design a procedure programming system design related to our project. They can be a Context diagram, DFD Level-1, DFD Level-2. It is an approach to design the system to organize a way to easily understand the whole system

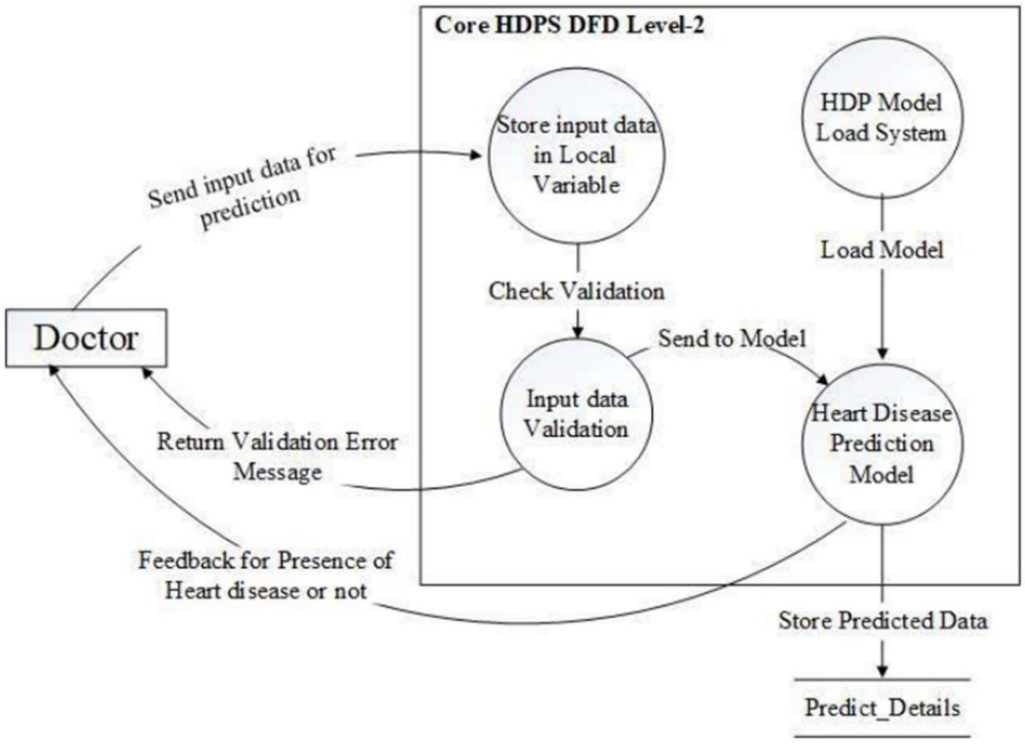

Fig.5: Solution Architecture

4.2 MODULE DESCRIPTION

4.2.1 COLLECTION OF DATA

Data collection is a systematic process of gathering observations or measurements. Whether you are performing research for business, governmental or academic purposes, data collection allows you to gain first-hand knowledge and original insights into your research problem.

While methods and aims may differ between fields, the overall process of data collection remains largely the same. Before you begin collecting data, you need to consider:

- The aim of the research
- The type of data that you will collect
- The methods and procedures you will use to collect, store, and process the data

ML depends heavily on data. It's the most crucial aspect that makes algorithm training possible and explains why machine learning became so popular in recent years. But regardless of your actual terabytes of information and data science expertise, if you can't make sense of data records, a machine will be nearly useless or perhaps even harmful.

The thing is, all datasets are flawed. That's why data preparation is such an important step in the machine learning process. In a nutshell, data preparation is a set of procedures that helps make your dataset more suitable for machine learning. In broader terms, the data prep also includes establishing the right data collection mechanism. And these procedures consume most of the time spent on machine learning. Sometimes it takes months before the first algorithm is built!

All data preparation should be done by a dedicated data scientist. And that's about right. If you don't have a data scientist on board to do all the cleaning, well… you don't have machine learning. But as we discussed in our story on data science team structures, life is hard for companies that can't afford data science talent and try to transition existing IT engineers into the field. Besides, dataset preparation isn't narrowed down to a data scientist's competencies only. Problems with machine learning datasets can stem from the way an organization is built, workflows that are established, and whether instructions are adhered to or not among those charged with recordkeeping.

In this module, the data has been collected from the user using interactive web page and dataset has been collected from a repository for training the ML model is the basic step in machine learning. The predictions made by ML systems can only be as good as the data on which they have been trained.

4.2.1.1 Google Data Search:

There are tens of millions of datasets on the web, with content ranging from sensor data and government records, to results of scientific experiments and business reports. Indeed, there are datasets for almost anything one can imagine, be it diets of emperor penguins or where remote workers live. More than two years ago, we undertook an effort to design a search engine that would provide a single entry point to these millions of datasets and thousands of repositories. The result is Dataset Search, which we launched in beta in 2018 and fully launched in January 2020. In addition to facilitating access to data, Dataset Search reconciles and indexes datasets using the metadata descriptions that come directly from the dataset web pages using schema.org structure.

As of today, the complete Dataset Search corpus contains more than 31 million datasets from more than 4,600 internet domains. About half of these datasets come from .com domains, but .org and governmental domains are also well represented. The graph below shows the growth of the corpus over the last two years, and while we still don't know what fraction of datasets on the web are currently in Dataset Search, the number continues to grow steadily.

4.2.1.2 Kaggle:

Kaggle is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve data science challenges. The aim of this online platform (founded in 2010 by Anthony Goldbloom and Jeremy Howard and acquired by Google in 2017) is to help professionals and learners reach their goals in their data science journey with the powerful tools and resources it provides. As of today (2021), there are over 8 million registered users on Kaggle.

One of the sub-platforms that made Kaggle such a popular resource is their competitions. In a similar way that HackerRank plays that role for software developers and computer engineers, "Kaggle Competitions" has significant importance for data scientists; you can learn more about them in our Kaggle Competiton Guide and learn how to analyze a dataset step-by-step in our Kaggle Competition Tutorial. In data science competitions like Kaggle's or

DataCamp's, companies and organizations share a big amount of challenging data science tasks with generous rewards in which data scientists, from beginners to experienced, compete on their completion. Kaggle also provides the Kaggle Notebook, which, just like DataCamp Workspace, allows you to edit and run your code for data science tasks on your browser, so your local computer doesn't have to do all the heavy lifting and you don't need to set up a new development environment on your own.

Kaggle provides powerful resources on cloud and allows you to use a maximum of 30 hours of GPU and 20 hours of TPU per week. You can upload your datasets to Kaggle and download others' datasets as well. Additionally, you can check other people's datasets and notebooks and start discussion topics on them. All your activity is scored on the platform and your score increases as you help others and share useful information. Once you start earning points, you will be placed on a live leaderboard of 8 million Kaggle users.

Kaggle is suitable for different groups of people, from students interested in data science and artificial intelligence to the most experienced data scientists in the world. If you are a beginner, you can take advantage of the courses provided by Kaggle. By joining this platform, you will be able to progress in a community of people of various levels of expertise, and you will have the chance to communicate with many highly experienced data scientists. As you earn Kaggle points and medals, which are proof of your progress, it is quite possible that you may even end up attracting headhunters and recruiters, and unlock new job opportunities.

Last but not least, when applying for jobs in data science, mentioning your Kaggle experience definitely makes a positive impact. It goes without saying that all these benefits also apply to highly experienced data scientists. No matter how experienced you are, this platform offers continuous learning and improvement possibilities, and, of course, the cash rewards that can come with the competitions are just as interesting.

4.2.1.3 Data.Gov :

Data.gov is primarily a federal open government data site. However, state, local, and tribal governments can also publish metadata describing their open data resources on Data.gov for greater discoverability. Data.gov does not host data directly (with a few exceptions), but rather aggregates metadata about open data resources in one centralized location. Once an open data source meets the necessary format and metadata requirements, the Data.gov team can harvest the metadata directly, synchronizing that source's metadata on Data.gov as often as every 24 hours.

Under the OPEN Government Data Act and the Open Data Policy, federal agencies are required to publish an enterprise data inventory, provided as a data.json file,using the standard Project Open Data metadata schema. The machine readable listing, as a standalone JSON file on the agency's website at agency.gov/data.json. This data.json file is what gets harvested to the Data.gov catalog.

Federal agencies that do not have a platform to inventory their metadata can make use of a free service hosted by Data.gov called inventory.data.gov (see the separate guide). Contact the Data.gov team via email if you're interested in using this service. You can find more information and tools on resources.data.gov.

When an agency is ready for Data.gov to harvest its data.json for the first time, the agency should notify Data.gov via email and the Data.gov team will create a new Data.gov harvest source for the data.json. The Data.gov team is available to assist agencies in generating the data.json file and provide tools that may help agencies prepare their data listings.

4.2.3 DATA PREPROCESSING

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn't have a regular, uniform design.

Machines like to process nice and tidy information – they read data as 1s and 0s. So calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis.

Data has been taken from its dataset repository which was obtained. The collected data is recorded in the form of a dataset. We cleanse the data by taking relevant fields (columns) by data cleaning. After cleaning the data, preprocessing is done in order to remove null values using Python.

4.2.3.1. Data Cleaning:

Data cleaning is the process of removing incorrect, duplicate, or otherwise erroneous data from a dataset. These errors can include incorrectly formatted data, redundant entries, mis-labelled data, and other issues; they often arise when two or more datasets are combined together. Data cleaning improves the quality of your data as well as any business decisions that you draw based on the data.

There is no one right way to clean a dataset, as every set is different and presents its own unique slate of errors that need to be corrected. Many data cleaning techniques can now be automated with the help of dedicated software, but some portion of the work must be done manually to ensure the greatest accuracy. Usually this work is done by data quality analysts, BI analysts, and business users.

Data cleaning has fill the null data and converted some data types,and omitted some unwanted columns.

a. Missing Data:

- It means it fills the nan data or null data and in that place, if it is numeric it can be stored and the mean value if it is a text column it can be removed.
- The data which cannot be read by the machines is known as noisy data.
- We can replace it by average values of the respecting columns or by omitting the entire rows.

b. Noisy Data:

Binning Method:

- The data is sorted either in ascending or descending order for smoothening the process

Regression:

- The data is made smooth by fitting it to a regression. The regression used may be linear or multiple .

Clustering:

- This groups the data into clusters based on their similarity. The dissimilarities(outliers) are mostly outside the clusters.

4.2.3.2. Data Integration:

Data Integration is involved in data analysis task which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. The issue to be considered in Data Integration is schema integration. It is tricky.

How can real-world entities from multiple data sources be 'matched up'? This is referred as entity identification problem. For example, how can a data analyst be sure that customer_id in one database and cust_number in another refer to the same entity? The answer is metadata. Databases and data warehouses typically have metadata. Simply, metadata is data about data.

Metadata is used to help avoiding errors in schema integration. Another important issue is redundancy. An attribute may be redundant, if it is derived from another table. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

4.2.4. DATA TRANSFORMATION:

Data transformation is the process of changing the format, structure, or values of data. For data analytics projects, data may be transformed at two stages of the data pipeline. Organizations that use on-premises data warehouses generally use an ETL (extract, transform, load) process, in which data transformation is the middle step. Today, most organizations use cloud-based data warehouses, which can scale compute and storage resources with latency measured in seconds or minutes. The scalability of the cloud platform lets organizations skip preload transformations and load raw data into the data warehouse, then transform it at query time — a model called ELT (extract, load, transform).

Processes such as data integration, data migration, data warehousing, and data wrangling all may involve data transformation.

Data transformation may be constructive (adding, copying, and replicating data), destructive (deleting fields and records), aesthetic (standardizing salutations or street names), or structural (renaming, moving, and combining columns in a database).

An enterprise can choose among a variety of ETL tools that automate the process of data transformation. Data analysts, data engineers, and data scientists also transform data using scripting languages such as Python or domain-specific languages like SQL.

To transform the data into required forms or to perform data mining operations data transformation is used

1. Normalization:

   Normalization is used to scale the values within a specified range

2. Attribute Selection:

   For mining process new attributes are built with the existing set of attributes

3. Discretization:

   To replace the raw values Discretization is used.

4. Concept Hierarchy Generation:

   The attributes are converted from lower level to higher level .

4.2.5. ANALYSING AND VISUALIZING THE DATA

The next step in the process is to analyze the preprocessed data. We analyze the data by using machine learning algorithms. By using algorithms we will be able to analyze the signal strength in various regions. The algorithm predicts the efficiency levels, speed, strength, and other criteria which match the signal analysis. With the help of the algorithm's prediction, we will be able to update the Admin of the region with what needs to be done for better performance and efficiency. This will be better understood by visualization.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise.

It’s hard to think of a professional industry that doesn’t benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

While we’ll always wax poetically about data visualization (you’re on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it’s also one of the most useful professional skills to develop. The better you can convey your points visually, whether in a dashboard or a slide deck, the better you can leverage that information. The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs the who, what, when, where, and how.

While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

CHAPTER 5

IMPLEMENTATION

CODE:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.metrics import accuracy_score, log_loss, roc_auc_score, confusion_matrix,
roc_curve, ConfusionMatrixDisplay
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier,
GradientBoostingClassifier
from sklearn import metrics
from xgboost import XGBClassifier

df = pd.read_csv("heart_disease_health_indicators_BRFSS2015.csv")
df.head()

cols=list(df.columns)

plt.figure(figsize=(15,40))
for i,column in enumerate(cols):
    plt.subplot(len(cols), 2, i+1)
    plt.suptitle("Plot Value Count", fontsize=20, x=0.5, y=1)
    sns.countplot(data=df, x=column)
    plt.title(f"{column}")
    plt.tight_layout()


plt.figure(figsize=(15,40))
for i,column in enumerate(cols):
    plt.subplot(len(cols), 2, i+1)
    plt.suptitle("Plot Value Proportion", fontsize=20, x=0.5, y=1)
```

```
    plt.pie(x=df[column].value_counts(), labels=df[column].unique(), autopct='%.0f%%')
    plt.title(f"{column}")
    plt.tight_layout()
# convert categorical variables into dummy variables
df=pd.get_dummies(df, drop_first=True)
# show firt five records and data size
display(df.head())
display(df.shape)


plt.figure(figsize=(15,10))
corr=df.corr()
mask = np.triu(np.ones_like(corr, dtype=bool))
sns.heatmap(corr,mask=mask, cmap='Blues',annot=False)
plt.show()

np.sum(df.corr()>0.6)

Check Outliers
plt.figure(figsize=(15,60))
for i,column in enumerate(cols):
    plt.subplot(len(cols), 2, i+1)
    plt.suptitle("Plot Value Count", fontsize=20, x=0.5, y=1)
    sns.boxplot(data=df, x=column)
    plt.title(f"{column}")
    plt.tight_layout()

Convert categorical variables to dummy variables
# find the categorical variables and  encoded.

# categorical variables
catcols = ['HighBP', 'HighChol', 'CholCheck',
       'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'Fruits', 'Veggies',
       'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost',
       'DiffWalk', 'Sex',
       ]
```

```
#convert the data type of categorical variables
for cat in catcols:
    df[cat]=pd.Categorical(df[cat])

df.dtypes
Checking the distribution of target variable
df.HeartDiseaseorAttack.value_counts()

Data splitting
#select HeartDiseaseorAttack as target variable:
y = df['HeartDiseaseorAttack']

#select all the other columns minus HeartDiseaseorAttack as the feature variables:
X = df.drop(['HeartDiseaseorAttack'],axis=1)

X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size = 0.2, random_state=0)
print('Dimensions: \n x_train:{} \n x_test{} \n y_train{} \n y_test{}'.format(X_train.shape,
X_test.shape, Y_train.shape, Y_test.shape))


Model Selection and Evaluation
Logistic Regression
from sklearn.linear_model import LogisticRegressionCV

logreg = LogisticRegressionCV(penalty='elasticnet', # Type of penalization l1 = lasso, l2 =
ridge, elasticnet
                              Cs = [0.1, 1, 10, 100],
                              tol=0.0001, # Tolerance for parameters
                              cv = 3,
                              fit_intercept=True,
                              class_weight='balanced', # Weights, see below
                              random_state=0, # Random seed
                             # max_iter=100, # Maximum iterations
                              verbose=2, # Show process. 1 is yes.
                              solver = 'saga', # How to optimize.
                              n_jobs = 2,      # Processes to use. Set to number of physical  cores.
```

```
                        refit = True,     # If to retrain with the best parameter and all data after
finishing.
                        l1_ratios = np.arange(0, 1.01, 0.1), # The LASSO / Ridge ratios.
                       )

logreg.fit(X_train, Y_train)

# show the coefficients for logistic regression
coef_df = pd.concat([pd.DataFrame({'column': X_train.columns}),
                 pd.DataFrame(np.transpose(logreg.coef_))],
                 axis = 1
                )

coef_df


# predict test data
pred_test=logreg.predict(X_test)
pred_test_prob=logreg.predict_proba(X_test)
# Let's evaluate logistic regression
conf = confusion_matrix(Y_test,pred_test)
ConfusionMatrixDisplay(conf).plot()
plt.show()

# Calculate the ROC curve points
fpr, tpr, _ = roc_curve(Y_test, pred_test_prob[:,1]) #just take yprob of positive class

# Save the AUC in a variable to display it. Round it first
auc_logreg = np.round(roc_auc_score(y_true = Y_test, y_score = pred_test_prob[:,1]),
decimals = 3)

# Create and show the plot
plt.plot(fpr, tpr,label=f"Logistic Regression: , auc={auc_logreg}")
plt.legend(loc=4)
plt.show()
```

```
#save the trained logistic regression model
import pickle
filename = 'logisticRegression_model.sav'
pickle.dump(logreg, open(filename, 'wb'))
```

CHAPTER 6
RESULTS AND DISCUSSIONS

The results are compared with different algorithms and the algorithm with the best accuracy has been taken and implemented. The algorithms used in this paper are KNN algorithm, SVM algorithm and NAIVE BAYES Classification.KNN algorithm gives 89% accuracy and it cannot be used for online learning and so it can be used to calculate using the distance formula.

Since the accuracy is not perfect , the next algorithm is taken into analysis (i.e) SVM. SVM algorithm gives 94% due to its data conversion into vector representations. Because of its ONline learning mechanism it seems to be better than KNN in terms of its accuracy. Since the best accuracy is not reached , some reference in the classification algorithms are taken and concluded that Logistic regression seems to have better accuracy than SVM Classifiers because of its implementation using Naives Bayes theorem and it can also be used for online learning. By analysing the data using Naives Bayes, the accuracy is 97% .

Hence for the HEART DISEASE PREDICTOR with the current data we predict that Naives Bayes is the best suited algorithm with the best accuracy percent.

CHAPTER 7

CONCLUSION

Visualizing and Prediction of Heart disease has been implemented and the results have been found out. We can see that it did very well since it has a pretty high precision score. Of the 35 people that had a head problem, the model misdiagnosed just 1. However, we can see that the number of people who did not have a disease but were predicted to have a heart problem is also relatively high (at 20). This means that we save many people, but we save them by recommending too many tests, which may lead to a waste of resources.

Our study mainly focused on the use of data mining techniques in healthcare especially in the detection of heart disease. Heart disease is a fatal disease which may cause death. Data mining techniques were implemented using the following algorithm, KNN, Neural Networks, Decision Tree, and Naive Bayes and Random Forest. We measured performance on the basis of Accuracy, TN, FP, FN and TP rate and in some algorithm. We conducted five experiments with the same data set to predict heart disease. The result of all the implemented algorithm are shown in tabular form for better understanding and comparisons .The experiment shows that Naive Bayes gives the highest accuracy which is 88% followed by ANN and KNN with accuracy of 87% .Our findings indicate that data mining can be used and applied in the healthcare industry to predict and diagnose the disease at early stages.

CHAPTER 8

REFERENCES

1. V. Manikantan & S.Latha,"Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering, Volume-2, Issue-2, pp.5-10, 2013.

2. Dr.A.V.Senthil Kumar, "Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering", International Journal of Advanced Trends in Computer Science and Engineering, Volume-4, No.6, pp.07-18, 2015.

3. Uma.K, M.Hanumathappa, "Heart Disease Prediction Using Classification Techniques with Feature Selection Method", Adarsh Journal of Information Technology, Volume-5, Issue-2, pp.22-29, 2016

4. Himanshu Sharma, M.A.Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms:A Survey",International Journal on Recent and Innovation Trends in Computing and Communication,Volume5,Issue-8,pp.99-104, 2017.

5. S.Suguna, Sakthi Sakunthala.N ,S.Sanjana, S.S.Sanjhana, "A Survey on Prediction of Heart Disease using Big data Algorithms", International Journal of Advanced Research in Computer Engineering & Technology,Volume-6,Issue-3,pp.371-378,2017.

6. A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nature Reviews Cardiology, vol. 8, no. 1, pp. 30–41, 2011.

7. J.Mourão-Miranda,A.L.W.Bokde,C.Born,H.Hampel,and M. Stetter, “Classifying brain states

and determining the discriminatingactivationpatterns:supportvectormachineon

functionalMRIdata,”NeuroImage,vol.28,no.4,pp.980–995, 2005.

8. S.Ghwanmeh,A.Mohammad,andA.Al-Ibrahim,“Innovative artificial neural networks-based decision support system for heartdiseasesdiagnosis,”JournalofIntelligentLearningSystems and Applications, vol. 5, no. 3, pp. 176–183, 2013.

9. Q. K. Al-Shayea, “Artificial neural networks in medical diagnosis,” International Journal of

Computer Science Issues, vol. 8, no. 2, pp. 150–154, 2011.

10. K. Vanisree and J. Singaraju, “Decision support system for congenital heart disease diagnosis

based on signs and symptoms using neural networks,” International Journal of Computer

Applications, vol. 19, no. 6, pp. 6–12, 2011.