HEART DISEASE PREDICTION SYSTEM

Bhavana Vangala
Student NetID:<bvang004>
University of California, Riverside
Graduate Student ID:862121697
bvang004@ucr.edu

Shreya Reddy Seelam
Student NetID:<sseel005>
University of California, Riverside
Graduate Student ID:862121969
sseel005@ucr.edu

Suchitra Pithavath
Student NetID:<spith001>
University of California, Riverside
Graduate Student ID:862121654
Spith001@ucr.edu

ABSTRACT

Data Mining Techniques are used in the various clinical decision support systems for the prediction and analysis of the multiple diseases with the vast amount of data using different algorithms and compares the performance of the prediction systems as its accuracy would matter utmost with the lives of people. The Heart Disease Prediction system is of one such application of Data Mining techniques which enables us to predict the Heart Disease of the user using various Risk factors like age, smoking habits, body mass index, hereditary etc. The Approach would be collecting the data from the UCI machine learning repository for four different regions. The raw data that is collected is pre-processed and arranged the data using different data mining tools. Then data is analyzed using supervised machine learning algorithms like simple K-means, Random Forest and Naïve Bayes to the data. The obtained results had considerable amount of error, so we further analyzed the data using more sophisticated algorithms like Artificial Neural Networks and Multi variant Regression to predict whether a person has a heart disease or not. The project gives the optimum artificial neural network model by splitting the data into 70% training set, 20% validation set and 10% as the testing set, the data is selected randomly, and errors are evaluated. The model which has all the three errors minimum is selected as optimal model. Number of experiments has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Artificial neural networks with 20 neurons and three hidden layers with minimum of 200 iterations. Then the accuracy, precision and recall of the model is predicted to evaluate the model and visualization is performed on the results.

Keywords

Data Mining, Artificial Neural Networks, Simple K-Means, Random Forest, Naïve Bayes, Neurons, Hidden Layers, WEKA, Python IDLE, Classifiers, Classification, Clustering, Overfitting, Training Data, Validation Set, Testing Data.

1. INTRODUCTION

The Data Mining holds great value for the health industry to enable health systems to orderly use data and analytics to identify best practices to reduce costs and improve care. Like in analytics and business Intelligence. Data Mining follows a method of discovering patterns in large data sets which involves methods at the intersection of machine learning, Statistics and database Systems. The goal of Data Mining is to extract large amount of data and converts them into useful and knowledgeable patterns which is used to predict something which helps the mankind and takes the technology to next level. Data mining in healthcare industry mean different things to different people.

Heart disease is the leading cause of death for people of most ethnicities around the world. We seek to predict the occurrence of heart disease by analyzing the internal patterns of the attributes that are available. There is a large amount of data in the medical sectors that is computerized, but they are not put to use. By efficiently using the available data heart disease can be predicted. Researchers have been using Classifications, Clustering's, Regressions, Artificial Intelligence, Decision Trees, K-Nearest Neighbor method to help healthcare professionals improve their efficiency and accuracy in heart disease diagnosis.

According to Statistics, 610000 people die of heart disease in the United States every year that is one in four deaths. In the rapidly growing world, people are running behind the goals, but they are neglecting to take care of themselves. The whole lifestyle is being changed as their food habits are changing. Because of stress and unhealthy food habits people are prone to diseases like diabetes and blood pressure. These diseases move towards major treat namely heart disease, the most vital organ of the body that affects all the other organs of an individual. Heart disease is the major cause for mortality. According to W.H.O in next ten years about 23.6 million individuals will be suffering with cardiac arrest. Thus, to overcome this disaster detection of illness should be done. Various factors exist that are useful for detecting the heart disease like diabetes, family history, obesity etc. Every human being is not equally skilled and so as doctors. And skilled doctors are not available in all the places. But in most cases, identification of the disease is done by the doctors and based on test results and this decision is based on the intuition. Demonizing the illness is a challenging task as it needs experience and expertise. An automatic system in medical diagnosis would enhance medical care and it can also reduce cost. The goal is to use some data and analyze other than using intuition. There is a large amount of data in the medical sectors that is computerized, but they are not use. By efficiently using the available data and data-mining techniques key information can be extracted to predict the heart disease. These techniques assist the patient as well as doctor to decide which as a solid proof. According to the statistics half of the victims have no prior indications of heart attack. Analysis of different attributes is done to investigate the heart disease. Generally, physicians make conclusion by seeing current test result of the patients but the confidence in the decision increases with the expertise. This fast-moving world with growing technology makes mankind more lethargic due to which heart diseases are seen more prominently. Heart disease prediction is a risky and challenging task. Predicting it accurately is the most important task as it is related with the life of people. The above heartbroken statistics drove us to make such an application where people are facilitated to know about their heart disease even before they consult a doctor. This application is a software-based application where it is built on the software tools like WEKA and python.

2. RELATED WORK

We have referred few papers related to the medical field, Heart Disease Prediction System, Artificial Neural Networks, Hidden Layers, Neurons. We have taken into consideration on how different algorithms work on different data. The research papers we referred from the various sites are:

Paper 1: Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors (IEEE)

This paper primarily describes how data mining techniques are used in clinical decision support systems for the prediction and diagnosis of various diseases with high accuracy. These Data Mining techniques are used in the effective designing of clinical support systems as their ability to find the hidden patterns and relationships in medical data for that matter any data is very high. This paper actually gives us an idea about one of the essential applications of Data Mining techniques based on Data mining tools, Neural Networks and genetic algorithms (hybrid model) which predicting the heart disease based on the very common risk factors which not only helps patients in getting the warning even before they visiting the doctors but also helps the medical professionals on diagnosing and treating the patients. The data in this paper is collected from 50 people. This paper actually put forth different hybrid models of predicting systems built using Data mining and neural networks which are proposed by a) extracting significant patterns from K-Means Clustering and MAFIA algorithm to mine frequent patterns b) Hybrid fuzzy and K nearest neighbor algorithm with 87% accuracy c) using neural ensemble with accuracy of 89% and d) using CANFIS and genetic algorithms which had a very low MSE. The data is divided into 70% for training and 15% each for testing and validating the data. They used Genetic Algorithm as Search method to optimize the neural network weights and produced Confusion matrix and determined accuracy using TP, TN, FP, FN. In conclusion, this paper’s primary objective is not to limits its use as clinical decision support but also helps people reduce the heart diseases.

Pros/Cons: According to me the most significant advantage of this paper is its scope on how it is used and its accuracy in the results. One of the significant limitations of this paper is its decidedly fewer data.

Possible Extension: The data should be collected in large amounts, and it should be subjected to the evaluations using data mining techniques.

Paper 2: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction (IJCA)

This paper shows concern on lack of effective analysis tools to discover hidden relationships and trends in data. This paper provides a comparison of the performance of predictive data mining technique on the same dataset using different supervised machine learning algorithms such as decision tree, Bayesian, etc. The approach in this paper is that the three above described algorithms have been used to analyze the

dataset. Tanagra tool is a data mining suite which is a graphical user interface algorithm which is used to classify the data. Data is processed and evaluated using 10- fold cross-validation. The results obtained by using a) decision tree classifier is that it is easier to read and interpret with an accuracy of 52% b) Naïve Bayes classifier is that it works well in many complex real-world situations with an accuracy of 52.33% c) KNN classifier is that it is degraded by the presence of noise data with an efficiency of 45.67%. The results obtained by processing the data for intelligent Heart Disease prediction for 909 records with 15 medical attributes by applying genetic algorithm and by using different algorithms in .NET interface such as a) Decision Tree gives accuracy of 89% b) Naïve Bayes with an accuracy of 86.53% c) ANN with accuracy of 85.53%. In conclusion, the paper says that the accuracy of the Decision tree and Bayes classifier improves more after applying the genetic algorithm.

Pros/Cons: The most significant advantage of this research is that it evaluates the accuracy. The limitation is that the experiment should be conducted very precisely.

Possible Extension: The data can be conducted from the different health care centers, and techniques should be compared with utmost accuracy.

Paper 3: Detection of Cardiac Disease using Data Mining Classification Techniques

The objective of this paper is to detect weather a person has a Cardiac disease or not by learning the internal patterns of the data that is present with most accuracy. Accuracy is the most vital factor in the prediction because medical diagnostic errors are dangerous and costly. This research is needed to predict the accurate results. This paper classifies the data set that is available using decision tree algorithm. Classify the data into one of the class eight attributes are considered here, some of them are age, blood sugar, chest pain etc. In this paper authors assumes that decision tree technique is the best technique for classification and he performed all the experiments using UCI heartbeat data. The accuracy of the results is almost 80 percent. The main goal of the research was to detect patients having heart disease more precisely and more accurately with minimum number of tests. The practitioners need expertise to detect the heart disease accurately. This research plays an important role in the cost reduction of treatment, diagnose disease and additional enhancement of the medical studies. This paper only predicts the heart disease and we can similarly detect other chronic disease in advance. According to me the paper only considered few attributes this can also consider few more attributes. More sophisticated algorithms like ANN can also be used to detect the results.

Paper 4: Comparing Data Mining Techniques Used for Heart Disease Prediction

This paper also deals with the prediction of heart disease. Predicting heart disease is essential these days. Data mining has become a more popular field of research among health care. This paper aims to detect heart disease because it is the leading cause of death in the past ten years. Predicting the disease can save the lives is the primary motivation for the research. Hospitals collect a lot of data regarding heart disease, but it is not mined to predict the patterns correctly this data can be extracted for effective data making. This paper uses different techniques like neural networks, decision trees, and just based on methods to predict the decisions. In neural networks, the hidden features of the data from different dimensionalities are learned. The decision tree gives the best visualization where every non-leaf node is a test and branch is an outcome of the trial. Whereas naive based can be used if the attributes are independent and to get more efficient output. In this paper along with the essential attributes, additional attributes like smoking and heart disease history are also used. Then the risk level of heart disease is assigned. Among all the techniques available neural networks with 15 attributes gave the best results. Then there is the decision tree and then the naive Bayes. In this paper prediction is made using all the three techniques this improves the scope of the result and builds the trust in the decision. I feel all the attributes are not appropriately mentioned and the necessary procedure of the experiments is also not specified.

Paper 5: Intelligent Heart Disease Prediction System Using Data Mining Techniques

This paper talks about the prototype intelligent heart disease prediction system(IHDPS) using data mining techniques namely, decision trees, naive Bayes, neural network. Using the medical profile such as blood sugar, age, sex, etc. can predict the likelihood of getting heart disease. IHDPS uses CRISP-DM (Cross industry standard process for data mining) methodology to build the prediction system which consists of six phases namely business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase has its own functionality to help in building the prediction system. total of 909 records with 15 attributes were used from Cleveland heart disease database. the attribute "diagnosis" is used as predicted attribute with value 1 as patient with heart disease and "0" as without heart disease. The effectiveness of models was evaluated using life chart and classification matrix. From the evaluation, the paper concludes that the most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees. five evaluating goals are defined against trained models provide decision support to doctors for discovering medical factors associated & diagnosing patients with heart disease.

Pros/Cons: IHDPS helps in answering the complex queries for diagnosis patients with heart disease which assist health care practitioners to make intelligent decision support. IHDPS used only 15 attributes & only categorical data is used. Another limitation is that it applies only three data mining techniques.

Possible extension: IHDPS can be further expanded by incorporating more attributes, considering continuous data and additional data mining techniques such as Time Series, Association Rules and clustering. Another area is to text mining to get knowledge out of the vast amount of unstructured data available in healthcare databases.

Paper 6: Human Heart Disease Prediction System using Data Mining Techniques

This paper gives the survey of different data mining classification techniques for predicting the risk level of heart disease of each person based on attributes. It mainly focuses on KNN classification technique to predict the heart disease of person. KNN is a non-parametric method used for regression and classification. In this algorithm, K is a user-defined constant. The test data are classified by assigning a constant value which is chronic among the K-training samples nearest to the point. the dataset consists of input, key and prediction attributes. Commonly used input attributes are age, gender, blood pressure, pulse rate and cholesterol of which age and gender are non-modifiable attributes. gender is static and constant where age is continuous and dynamic. To get more appropriate results, smoking and history of heart disease are also included. Patient ID is used as the key attribute. the prediction attribute is found to predict the risk level of heart disease, and risk level is classified in to three levels high which is more than 50%, low which is less than 50% and normal which is 0.the proposed method first checks all the input attributes and classify using KNN algorithm called as classifier module. the classes are analyzed with standard values, then the risk rate of heart disease is calculated using ID3(Iterative Dichotomiser 3) algorithm which is also known as prediction module.

Pros/Cons: The accuracy of the predicted system is increased by considering additional attributes such as smoking and a history of heart disease. Only one data mining technique is used that is KNN algorithm.

Possible extension: The accuracy of the prediction system can be increased by considering other data mining techniques such as ANN, naive Bayes, and decision tree.

3. PROPOSED METHOD

3.1 Dataset and its Processing

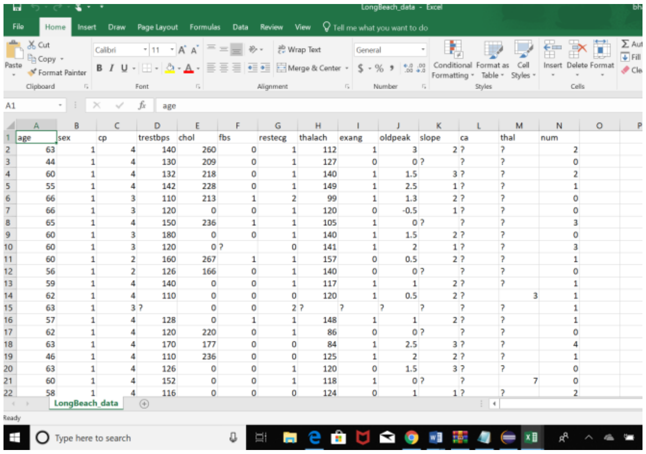
This section contains proposed methodology of prediction whether a person has heart disease or not. For this we have taken many different algorithms into consideration for us to make sure we arrive at utmost accuracy. Because accuracy in this prediction model is the ultimatum and lowering the accuracy would cost the lives of people more specifically lives of users.

3.1.1 Dataset

The first and the foremost important task is to collect the data. The data is extracted from the UCI repository. Out of the 76 attributes that were available in the repository we have taken only fourteen (14) attributes into consideration based on the importance of that attribute to predict the heart disease. The fourteen attributes that were taken into consideration are: age (age of the patient), sex (Gender of the patient), cp (chest pain type), trestbps (resting blood pressure which is measured in mm Hg on admitted to hospital), chol (serum cholesterol in mg/dl), fbs (fasting blood sugar > 120 mg/dl where 1 = true; 0 = false), restecg (resting electrocardiographic results where Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria), thalach (maximum heart rate achieved), exang (exercise induced angina where 1 = yes; 0 = no), old peak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment where Value 1: upsloping, Value 2: flat, Value 3: down sloping), ca (number of major vessels (0-3) colored by fluoroscopy), thal (3 = normal; 6 = fixed defect; 7 = reversible defect), num (the predicted attribute). This data is collected from four different locations namely: Long beach, Cleveland, Hungary and Switzerland.

3.1.2 Preprocessing the Dataset

The next task is to clean the data as the data contained many noisy junk values. All the irregular characters and noisy data is removed for the results to be accurate. This processing is done in the Open Refine tool where you remove all the unwanted or noisy data accurately. We have performed this processing for all the four datasets namely for Cleveland Dataset, Hungarian Dataset, Switzerland Dataset and Long Beach Dataset.



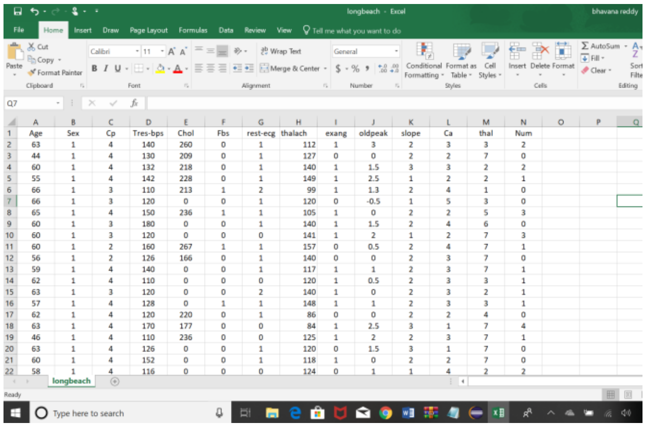
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	4	140	260	0	1	112	1	3	2	?	?	2
44	1	4	130	209	0	1	127	0	0	?	?	?	0
60	1	4	132	218	0	1	140	1	1.5	3	?	?	2
55	1	4	142	228	0	1	149	1	2.5	1	?	?	1
66	1	3	110	213	1	2	99	1	1.3	2	?	?	0
66	1	3	120	0	0	1	120	0	-0.5	1	?	?	0
65	1	4	150	236	1	1	105	1	0	?	?	?	3
60	1	3	180	0	0	1	140	1	1.5	2	?	?	0
60	1	3	120	0	?	0	141	1	2	1	?	?	3
60	1	2	160	267	1	1	157	0	0.5	2	?	?	1
56	1	2	126	166	0	1	140	0	0	?	?	?	0
59	1	4	140	0	0	1	117	1	1	2	?	?	1
62	1	4	110	0	0	0	120	1	0.5	2	?	3	1
63	1	3	?	0	0	2	?	?	?	?	?	?	1
57	1	4	128	0	1	1	148	1	1	2	?	?	1
62	1	4	120	220	0	1	86	0	0	?	?	?	0
63	1	4	170	177	0	0	84	1	2.5	3	?	?	4
46	1	4	110	236	0	0	125	1	2	2	?	?	1
63	1	4	126	0	0	1	120	0	1.5	3	?	?	0
60	1	4	152	0	0	1	118	1	0	?	?	7	0
58	1	4	116	0	0	0	124	0	1	1	?	?	2

Fig 1: Unprocessed Data

Next, we processed data by taking only 14 attributes out of 76 attributes that are available in the repository.

3.1.3 Handling Missing Data

The next important step was to handle the missing data. We handled missing data by replacing the missing attribute value by most frequently occurred value among that attribute values. That's how we cleaned, preprocessed and handled the missing values of the data.



Age	Sex	Cp	Tres-bps	Chol	Fbs	rest-ecg	thalach	exang	oldpeak	slope	Ca	thal	Num
63	1	4	140	260	0	1	112	1	3	2	3	3	2
44	1	4	130	209	0	1	127	0	0	2	2	7	0
60	1	4	132	218	0	1	140	1	1.5	3	3	2	2
55	1	4	142	228	0	1	149	1	2.5	1	2	2	1
66	1	3	110	213	1	2	99	1	1.3	2	4	1	0
66	1	3	120	0	0	1	120	0	-0.5	1	5	3	0
65	1	4	150	236	1	1	105	1	0	2	2	5	3
60	1	3	180	0	0	1	140	1	1.5	2	4	6	0
60	1	3	120	0	0	0	141	1	2	1	2	7	3
60	1	2	160	267	1	1	157	0	0.5	2	4	7	1
56	1	2	126	166	0	1	140	0	0	2	3	7	0
59	1	4	140	0	0	1	117	1	1	2	3	7	1
62	1	4	110	0	0	0	120	1	0.5	2	3	3	1
63	1	3	120	0	0	2	140	1	0	2	3	2	1
57	1	4	128	0	1	1	148	1	1	2	3	3	1
62	1	4	120	220	0	1	86	0	0	2	2	4	0
63	1	4	170	177	0	0	84	1	2.5	3	1	7	4
46	1	4	110	236	0	0	125	1	2	2	3	7	1
63	1	4	126	0	0	1	120	0	1.5	3	1	7	0
60	1	4	152	0	0	1	118	1	0	2	2	7	0
58	1	4	116	0	0	0	124	0	1	1	4	2	2

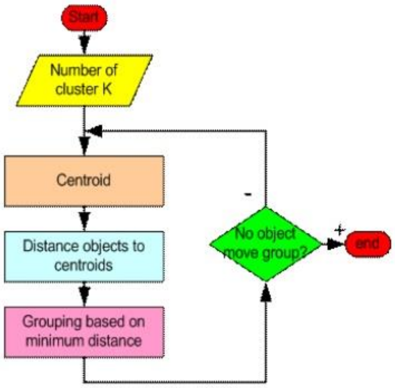
Fig 2: Processed Data

3.2 Algorithms

The algorithms that we propose to get efficient results is Artificial Neural Network(ANN) and Multi Variant Regression. Firstly, using ANN we predict the model and then compare it with the Multi Variant Regression prediction model to get the accuracy.

3.2.1 Simple K-Means Clustering Algorithm

Simple K-Means is a unsupervised learning algorithm which is used to solve the clustering Algorithms. This algorithm divides the entire dataset into several clusters which is defined by a variable "k" which is given by the user by his/her choice. The clusters that are formed from the above defined method are then put as or positioned as points and all the observations or the data points are associated with the nearest cluster, computes and then gets adjusted until the desired result is obtained.



Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Fig 3: Simple K-Means Algorithm

3.2.2 Random Forest

It is an ensemble of decision trees and it is trained with the bagging method. Random Forest gives an extra randomness to the discrete model while building the trees. Instead of searching for the best splitting attribute for each node, it selects a random subset of features and among them searches for the best feature. This gives the variety of trees that generally results in a better model. Even a random subset of feature values can be considered instead of the best split of feature value. We implemented Random forest in Weka.

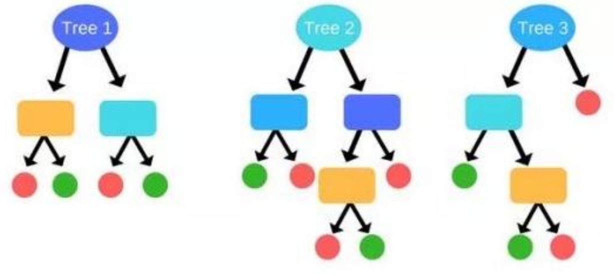
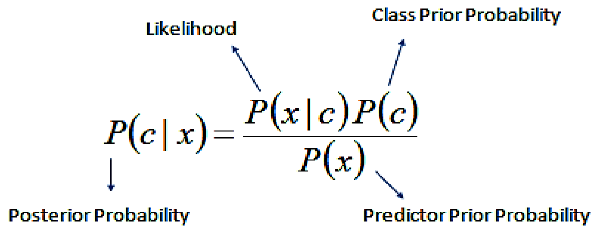


Fig 4: Random Forest Algorithm

3.2.3 Naïve Bayes Algorithm

This classification is based on Bayes theorem. The main assumption is that all the features are independent on each other. Naive Bayes classifier assumption is that that the presence of a feature in a class is unrelated to the presence of any other feature. Even if the features are related we assume that each of them contributes independently. This naive Bayes algorithm is executed in WEKA and the results are analyzed accordingly.



$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

$$P(c\mid \mathrm{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Fig 5: Naïve Bayes Algorithm

3.2.4 Artificial Neural Network (ANN)

In Artificial Neural Networks we have used multi-Layer neural network. This network is having 13 input nodes and one output node. The number of input nodes is based on the related set of risk factors that are described in the above section. The number of hidden layers is decided for which the training is fast, and the output is best. The number of neurons in each layer is decided by calculating the complete error of the model with the different number of neurons and among them the number of neurons which gave the least error. After the experiments we have decided there should be 3 hidden layers and each layer should have 20 neurons for the best results. The first step is to initialize the weights of the neural network randomly. Then these weights are passed through the genetic algorithm for optimization according to the cost function. Once the weights are optimized backpropagation is used for training the model. The functionality of the model is obtained by forward propagation of training values, backward propagation of error and updating the weights and biases accordingly. The maximum number of epochs to train the model is 1000. The learning stops after the thousand epochs, at this stage the classification is accurate. The weights that are present in this stage are responsible for the functionality of the model. Also, the activation function that is used in each neuron is Relu. The output that is predicted gives whether a person has heart disease or not.

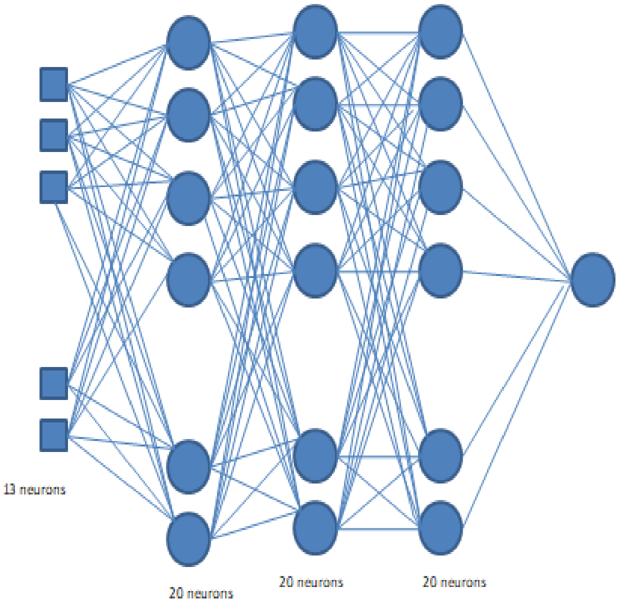


Fig 6: Artificial Neural Network with three hidden layers consisting of 20 neurons each

3.2.5 Multivariate Regression Model

This multivariate regression is a technique that estimates a single regression model which has more than one outcome. So, when there is more than one predictable variable in a multivariate regression model, then it gives a multivariate regression model.

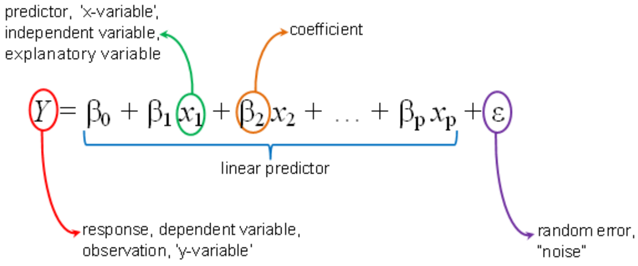

predictor, 'x-variable',
independent variable,
explanatory variable
coefficient
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$
linear predictor
response, dependent variable,
observation, 'y-variable'
random error,
"noise"

Fig 7: Multivariate Regression Model

So here, Multivariate regression coefficient indicates the change in the dependent variable associated with the change in the independent variable, in question holding constant the other independent variables in the equation. In brief it is an equation with more than one independent variable. This regression model takes 13 attributes as the independent variables and fits an equation that can best predict the dependent variable that is whether a person has heart disease or not. The coefficients of the equation are learned by the given training set.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

4.1.1 Open Refine

Open Refine is one powerful tool which enables us to convert our messy and noisy data into one structured form. It also enables us to convert one form of data to another in large scale. The below figure shows on how the User Interface of Open Refine tool looks like and how does it work. We used this tool to clean the data and make it free from noisy and junk data.

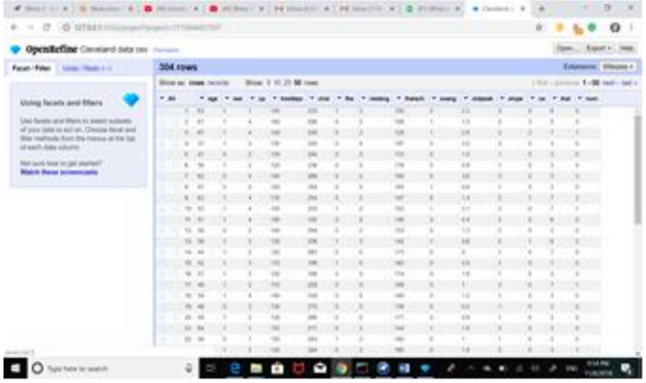


Fig 8: Open Refine Tool

4.1.2 WEKA

Weka is an open source Java-based platform which consists of various machine learning algorithms. These machine learning algorithms are used for data mining tasks. In this platform, the algorithms can directly be applied to a dataset. Weka contains tools for various stages in the KDD process. It is used to detect the hidden patterns in our dataset and extract knowledge from the dataset. We loaded our dataset and applied different algorithms like simple K-means, Random Forest and Naïve Bayes to it and analyze the results obtained. The below is the figure of how WEKA interface looks when we first load the data.

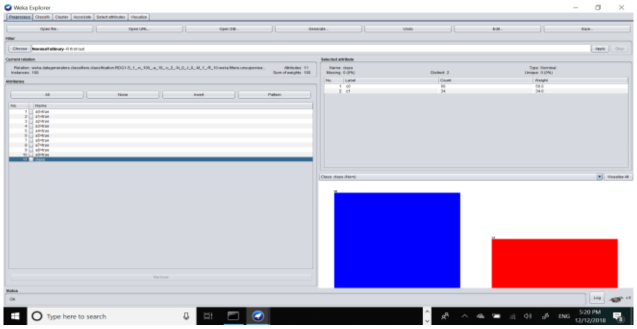
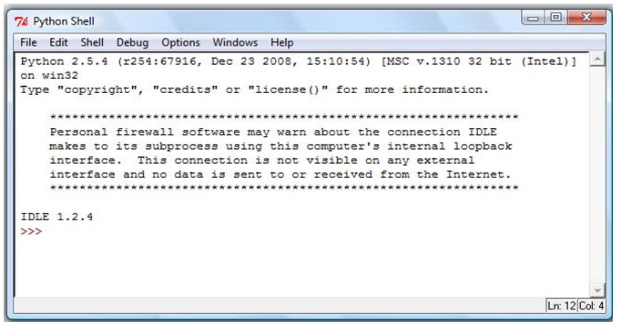


Fig 9: WEKA Tool

4.1.3 *Python IDLE*

IDLE is integrated development environment for the scripting language which is called as python. It is packaged with optional part of python with many linux distributions. We have executed out ANN and Multivariant Regression model algorithms in this IDLE using Python language.



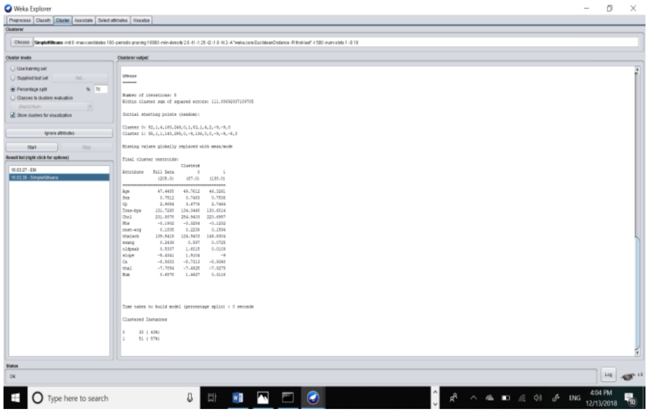
```
Python 2.5.4 (r254:67916, Dec 23 2008, 15:10:54) [MSC v.1310 32 bit (Intel)]
on win32
Type "copyright", "credits" or "license()" for more information.

    ****************************************************************
    Personal firewall software may warn about the connection IDLE
    makes to its subprocess using this computer's internal loopback
    interface.  This connection is not visible on any external
    interface and no data is sent to or received from the Internet.
    ****************************************************************

IDLE 1.2.4
>>>
```

Fig 10: Python IDLE Tool

4.2 Comparison of the Algorithms & Results

4.2.1 *Simple K-Means Clustering Algorithm*

We tested this algorithm with Dataset on WEKA tool by splitting the data into 70% training data and 30% testing data. By applying this algorithm to the above selected datasets we got sum of the squared errors to around 112 which is not feasible to predict the model accurately. Hence we have not based our prediction model on Simple K-means Clustering.

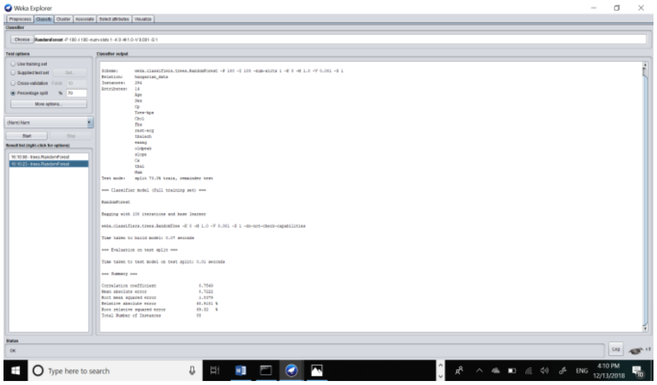


```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 111.89492937169705

Initial starting points (random):

Cluster 0: 52,1,4,160,246,0,1,82,1,4,2,-9,-9,3
Cluster 1: 55,1,1,140,295,0,-9,136,0,0,-9,-9,-9,3

Missing values globally replaced with mean/mode

Final cluster centroids:
                          Cluster#
Attribute    Full Data          0          1
               (205.0)     (67.0)    (138.0)
=============================================
Age            47.4488    49.7612    46.3261
Sex             0.7512     0.7463     0.7536
Cp              2.9854     3.4776     2.7464
Tres-bps      131.7268   134.0448   130.6014
Chol          231.8878   254.9403   220.6957
Fbs            -0.1902    -0.3284    -0.1232
rest-ecg        0.1805     0.2239     0.1594
thalach       139.9415   124.9403   146.6304
exang           0.2439      0.597     0.0725
oldpeak         0.5307     1.0015     0.0109
slope          -9.4341     1.9104         -9
Ca             -8.8683    -8.7313    -8.9348
thal           -7.7854    -7.4925    -7.9275
Num             0.6878     1.8627     0.3116



Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0      35 ( 43%)
1      51 ( 57%)
```

Fig 11: Simple K-Means Clustering output

4.2.2 *Random Forest Algorithm*

We also tested this algorithm on the WEKA tool with the dataset by splitting the data into 70% training data and 30% testing data. By applying this algorithm to the above loaded data set we get relative absolute error to be 60%. As we want to better the accuracy we did not base our prediction model on the Random Forest.

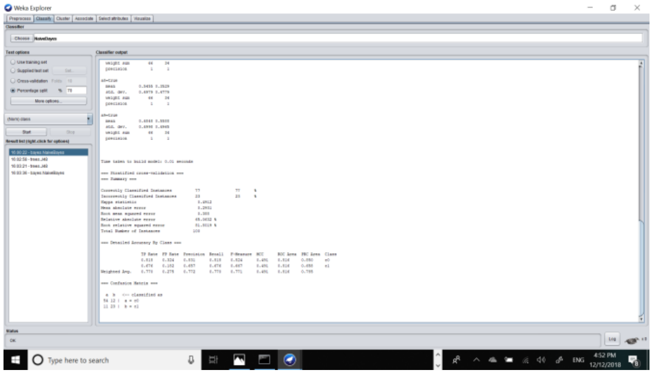


```
Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     hungarian_data
Instances:    294
Attributes:   14
              Age
              Sex
              Cp
              Tres-bps
              Chol
              Fbs
              rest-ecg
              thalach
              exang
              oldpeak
              slope
              Ca
              thal
              Num
Test mode:    split 70.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.07 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correlation coefficient                  0.7549
Mean absolute error                      0.7222
Root mean squared error                  1.0379
Relative absolute error                 60.9181 %
Root relative squared error             69.32   %
Total Number of Instances               88
```

Fig 12: Random Forest output

4.2.3 *Naïve Bayes Algorithm*

We even tested this algorithm on the WEKA tool with the dataset by splitting the data into 70% training data and 30 % testing data and we converted the nominal data to binary data. We noticed that the precision through this algorithm is around 77% which is not satisfying number atleast in this medical field. Hence we have decided to not base our prediction on this. The below figure is the output of the algorithm which gives the statistics when applied with the dataset.



```
  weight sum        66      34
  precision          1       1

A8=true
  mean          0.3455 0.2529
  std. dev.     0.4979 0.4779
  weight sum        66      34
  precision          1       1

A9=true
  mean          0.4848 0.5588
  std. dev.     0.4998 0.4965
  weight sum        66      34
  precision          1       1



Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          77               77      %
Incorrectly Classified Instances        23               23      %
Kappa statistic                          0.4912
Mean absolute error                      0.2931
Root mean squared error                  0.388
Relative absolute error                 65.0632 %
Root relative squared error             81.8019 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.818    0.324    0.831      0.818    0.824      0.491    0.816     0.880     r0
                 0.676    0.182    0.657      0.676    0.667      0.491    0.816     0.658     r1
Weighted Avg.    0.770    0.275    0.772      0.770    0.771      0.491    0.816     0.785

=== Confusion Matrix ===

  a  b   <-- classified as
 54 12 |  a = r0
 11 23 |  b = r1
```

Fig 13: Naïve Bayes Output

4.2.4 *Artificial Neural Network Algorithm*

The neural network system was developed using tensor flow. All the weight matrix is initialized with zeros and the data set is split into three sets training set, validation set, and testing set. The training set has 70% of the data, the validation set has 20% of the data and training set has 10%. For each set, the data is selected randomly. After splitting the data we trained the data and calculated the different errors such as validation error, training error and testing error by changing the number of neurons and hidden layers. The entire prediction is done in python IDLE. The error graphs for the different number of neurons and the hidden layers are shown below.

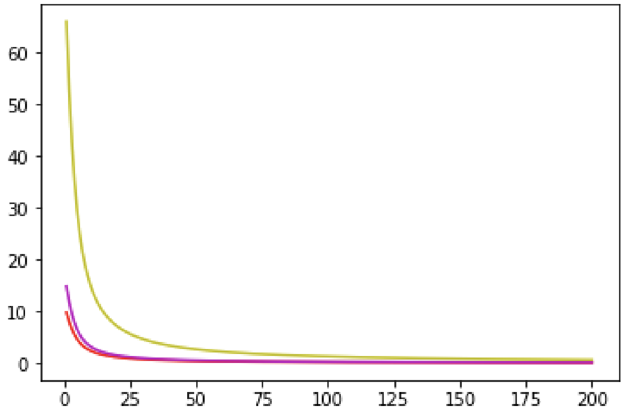


Fig 14: Error graph of Artificial Neural Networks with one hidden layers.

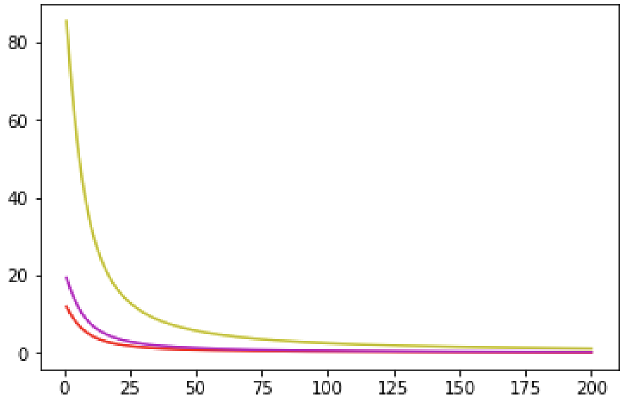
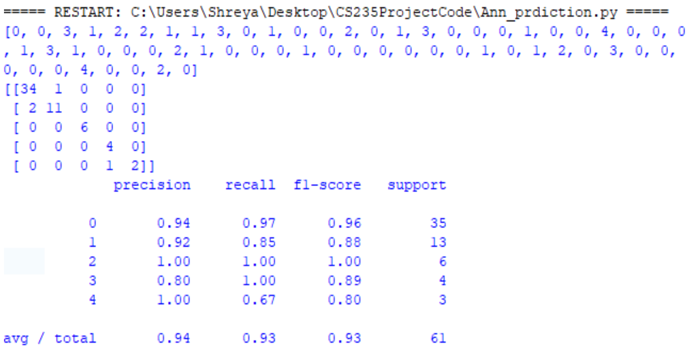


Fig 15: Error graph of Artificial Neural Networks with two hidden layers.

Next we found the least error by combining all the above obtained errors and selected the neurons and hidden layers to be three hidden layers and 20 neurons per each hidden layer. corresponding to the least error. The precision that is achieved through this algorithm is 94% which is a great number.



```
===== RESTART: C:\Users\Shreya\Desktop\CS235ProjectCode\Ann_prdiction.py =====
[0, 0, 3, 1, 2, 2, 1, 1, 3, 0, 1, 0, 0, 2, 0, 1, 3, 0, 0, 0, 1, 0, 0, 4, 0, 0, 0
, 1, 3, 1, 0, 0, 0, 2, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 2, 0, 3, 0, 0,
0, 0, 0, 4, 0, 0, 2, 0]
[[34  1  0  0  0]
 [ 2 11  0  0  0]
 [ 0  0  6  0  0]
 [ 0  0  0  4  0]
 [ 0  0  0  1  2]]
             precision    recall  f1-score   support

          0       0.94      0.97      0.96        35
          1       0.92      0.85      0.88        13
          2       1.00      1.00      1.00         6
          3       0.80      1.00      0.89         4
          4       1.00      0.67      0.80         3

avg / total       0.94      0.93      0.93        61
```

Fig 16: Prediction Model with the Evaluation Measures using ANN

Thus we predicted the model as 0's and 1's basing on this Artificial Neural Network.

4.2.5 Multi-Variate Regression Algorithm

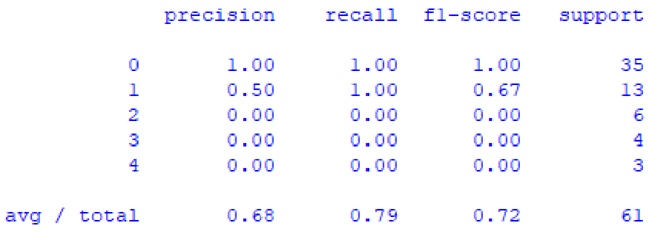
We split the entire data into training and testing data. Out of the 14 attributes we have 13 of them are actual inputs and 14th attribute is the output. We fit the regression line using 13 attributes using the training dataset. We then predicted the model using the testing data using python IDLE being the platform.

The output of prediction

[0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1,

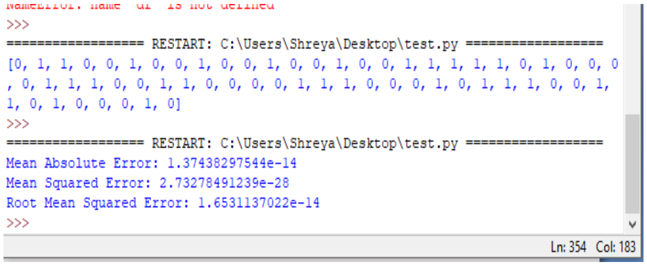
1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0]

The precision was around 68% but the errors were very minimum. That is why we have taken this algorithm in predicting the model. The below are the figures that shows the precision and all the other evaluation measures with the error rate.



```
             precision    recall  f1-score   support

          0       1.00      1.00      1.00        35
          1       0.50      1.00      0.67        13
          2       0.00      0.00      0.00         6
          3       0.00      0.00      0.00         4
          4       0.00      0.00      0.00         3

avg / total       0.68      0.79      0.72        61
```

Fig 17: Evaluation Measures using Multi-Variate Regression Model



```
NameError: name 'df' is not defined
>>>
================= RESTART: C:\Users\Shreya\Desktop\test.py =================
[0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0
, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1,
1, 0, 1, 0, 0, 0, 1, 0]
>>>
================= RESTART: C:\Users\Shreya\Desktop\test.py =================
Mean Absolute Error: 1.37438297544e-14
Mean Squared Error: 2.73278491239e-28
Root Mean Squared Error: 1.6531137022e-14
>>>
```

Fig 18: Error Rates and Prediction model using Multi-Variate Regression Algorithm

4.3 Visualization

4.3.1 Visualization of ANN prediction model

This visualization helps us in visualizing the prediction model that is X-axis describes the number of data points and Y-axis scale defines the output that is 0's and 1's for each data point. This is how the entire prediction model is graded. The below is the figure which is called Scatter plot and describes the visualized version of the output.

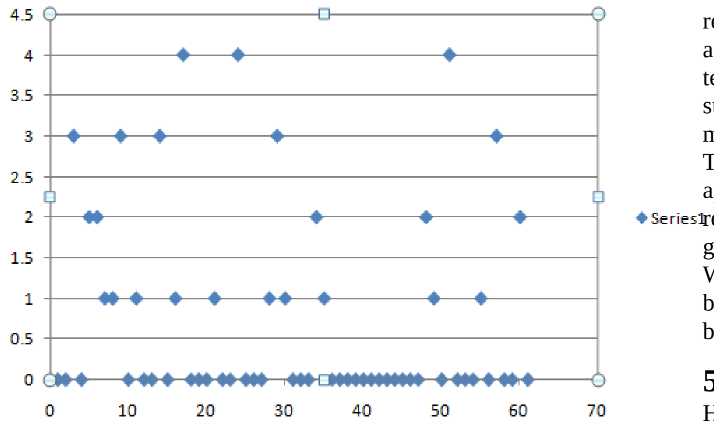


Fig 19: Visualization of ANN Prediction Model using Scatter Plot

4.3.2 Visualization of Multi-Variant prediction model

This visualization helps us in visualizing the prediction model that is X-axis describes the number of data points and Y-axis scale defines the output that is 0's and 1's for each data point. This is how the entire prediction model is graded. The below is the figure which is called Scatter plot and describes the visualized version of the output.

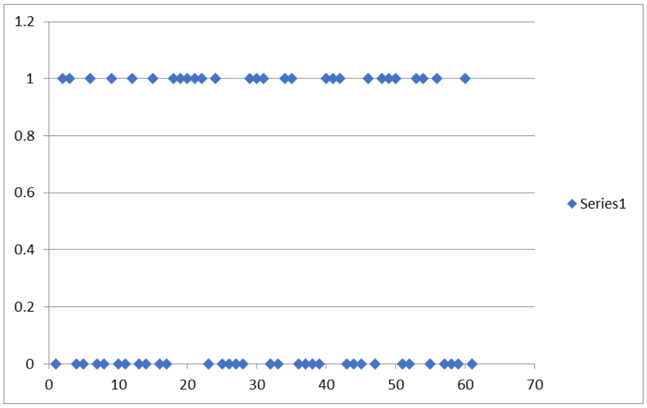


Fig 20: Visualization of Multi-Variate Prediction Model using Scatter Plot

5. DISCUSSION & CONCLUSION

5.1 Conclusion

Data mining techniques and methods are applied in a patient medical dataset has resulted in decision support, standards and innovations system that have significant success in improving the overall quality of medical services and the health of patients. But we still Need systems which could predict heart diseases in early stages so that patients can plan and monitor his/her health on own and take preventive measures and treatment at early stages of the disease.

The heart disease prediction system is developed using two data mining modeling techniques Such as artificial neural network and Multivariate regression. The hidden knowledge is extracted by the system using historical heart disease database from the UCI repository. Tensor flow library and python is used to build and access the models. The models are validated and tested against a test dataset. Confusion matrix, precision, recall, f1-score, and support are used to evaluate the effectiveness of the models. Two models can extract patterns in response to the predictable state. The most effective model to predict heart disease of a patient appears to be an artificial neural network followed by Multivariate regression. We combined both the results and intersected them to get the output that is prediction model with an accuracy of 92%. We did follow the above process to ensure that accuracy is at its best and treating people well before their visit to the doctor because any issue with the prediction will cost lives of the people.

5.2 Future Work

Heart disease prediction system can be further enhanced by incorporating other data mining techniques like clustering, time series, and association rules. The system can include other medical attributes besides 14 listed. We can also use continuous data instead of just categorical data. Text mining can also be done to mine the vast amount of unstructured data available in healthcare databases, but the biggest challenge is to integrate text mining and data mining.

6. ACKNOWLEDGMENTS

We are very thankful to our mentor Vagelis Papalexakis for his persistent guidance and support throughout our work. It was incredibly exciting for us to work under his mentorship.

7. REFERENCES

[1] Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors (IEEE) Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, Department of Computer Science & Engineering Integral University, Lucknow, India

[2] Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction (IJCA) Jyoti Soni, Ujma Ansari, Dipesh Sharma, India

[3] Detection of Cardiac Disease using Data Mining Classification Techniques. Abdul Aziz Faculty of Computer Science & IT Superior University, Lahore, Pakistan, Aziz Ur Rehman Faculty of Computer Science & IT Superior University Lahore, Pakistan

[4] Comparing Data Mining Techniques Used for Heart Disease Prediction. Prof. Mamta Sharma, Farheen Khan, Vishnupriya Ravichandran

[5] Intelligent Heart Disease Prediction System Using Data Mining Techniques, Sellappan Palaniappan Rafiah Awang, Department of Information Technology Malaysia University of Science and Technology Block C, Kelana Square, Jalan SS7/26 Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia

[6] Human Heart Disease Prediction System using Data Theresa Princy. R Research Scholar Department of Information Technology Christ University faculty of engineering, Bangalore, India-560060. princy.aida@gmail.com J. Thomas, Department of Computer Science and Engineering Christ University faculty of engineering, Bangalore, India

[7] https://stackoverflow.com

[8] UCI Repository

[9] Open Refine tool

[10] WEKA tool

[11] Hai H.Dam, Hussain A.Abbass and Xin Yao, "Neural –Based Learning Classifier Systems", IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.1, pp.26-39, 2008.

[12] Shantakumar B.Patil and Y.S.Kumaraswamy, "Intelligent and EffectiveHeart Attack Prediction System Using Data Mining and Artificial NeuralNetwork", European Journal of Scientific Research, Vol.31, No.4,pp.642-656, 2009.

[13] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of Data MiningTechniques in Healthcare and Prediction of Heart Attacks," InternationalJournal on Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp.250-255, 2010.

[14] Wood D, De Backer, Prevention of coronary heart disease in clinicalpractice: recommendations of the Second Joint Task Force of Europeanand other Societies on Coronary Prevention. Atherosclerosis 140: 199–270, 1998.

[15] R. Das, I. Turkoglu, and A. Sengur, Effective diagnosis of heart diseasethrough neural networks ensembles, Expert Systems with Applications, Elsevier, pp. 7675–7680, 2009.

[16] Polat , K., S. Sahan, and S. Gunes, Automatic detection of heart diseaseusing an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour based weighting preprocessing. Expert Systems with Applications 2007. 32 p.625–631.

[17] D. Isern, D. Sanchez, and A. Moreno, "Agents Applied in Health Care:A Review", International Journal of Medical Informatics,79(3),pp.146-166, oi:10.1016/j.ijmedinf201O.01.003, 2010.

[18] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent andEffective Heart Attack Prediction System Using Data Miningand Artificial Neural Network, European Journal of ScientificResearch ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656

[19] Fadi Thabtah, A review of associative classification mining,The Knowledge Engineering Review, Volume 22 , Issue 1(March 2007),Pages 37-65, 2007.

[20] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDB'94, Santiago, Chile, Sept.1994. pp. 487-49

[21] N.A. Setiawan, P.A. Venkatachalam, and Ahmad FadzilM.H. Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set, International Journal ofRecent Trends in Engineering, Vol 2, No. 5, November 2009.

[22] WHO Europe (2007) The challenge of obesity in the WHO European Region and the strategies for response. Copenhagen: WHO RegionalOffice for Europe.

[23] Berenson GS, Srinivasan SR, Bao W, Newman WP III, Tracy RE, et al.(1998) Association between multiple cardiovascular risk factors and a the rosclerosis in children and young adults. The Bogalusa Heart Study.N Engl J Med 338:1650-1656.

[24] Dinarević S, Mulaosmanović V (2005) Primary prevention ofHypertension in Sarajevo Children: Role of Adiposity. 29th UMEMPS Congress Union of Middle Eastern and Mediterranean PaediatricSocieties, pp. 154-156.

[25] Muhammad Subhi Al- Batah "Testing the probability of Heart Diseaseusing Classification and Regression Tree" Annual Research & Reviewin Biology 4 (11): 1713-1725,(2014).

About the authors:

Bhavana Vangala is a Graduate Student Bearing Graduate Student ID: 862121697 whose major is Computer Engineering at University of California, Riverside.

Shreya Reddy Seelam is a Graduate Student Bearing Graduate Student ID: 862121969 whose major is Computer Science at University of California, Riverside.

Suchitra Pithavath is a Graduate Student Bearing Graduate Student ID: 862121654 whose major is Computer Science at University of California, Riverside.