GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Prediction of Heart Disease Using Machine Learning

A report for the Evaluation 3 of Project 2

Submitted by

VISHAL KUMAR

(1613101837/16SCSE101623)

In partial fulfilment for the award of the degree of

Bachelors of Technology

IN

Computer Science and Engineering

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

Dr. NITIN MISHRA, Professor

APRIL/MAY-2020

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "**PREDICTION OF HEART DISEASE USING MACHINE LEARNING"** is the bonafide work of "**VISHAL KUMAR (1613101837)**" who carried out the project under the work under my supervision.

SIGNATURE OF HEAD

Dr. MUNISH SHAHARWAL.,
PhD (Management),
PhD (CS) Professor &Dean
School of Computer Science
& Engineering

SIGNATURE OF SUPERVISOR

Dr. NITIN MISHRA,
Professor
School of Computer Science
& Engineering

TABLE OF CONTENTS

ABSTRACT

The prosperous test of information mining in exceptionally obvious fields like promoting, e-business, and retail has prompted its application in different divisions and enterprises. Medicinal services is being found among these territories. There is an extravagance of information accessible inside the social insurance frameworks. In any case, there is a shortage of auxiliary investigation execute to discover obnubilated connections in information. This exploration expects to give an itemized depiction of Naïve Bayes and choice tree classifier that are applied in our examination solidly in the forecast of Heart Disease. Some test has been directed to coordinate the execution of prescient information handling strategy on a proportional dataset, and in this way the result uncovers that bring Tree beats over Bayesian assignment.

Keywords — Data mining, Heart Disease Prediction, Naïve Bayes Classifier, Decision tree Classifier.

LIST OF FIGURES

INTRODUCTION

Information mining is the PC predicated procedure of extricating auxiliary data from tremendous arrangements of databases. Information mining is generally auxiliary in an explorative examination in light of nontrivial data from sizably voluminous volumes of proof. Clinical information digging has incredible potential for investigating the secretive examples in the informational indexes of the clinical area. These examples can be used for human services determination. Notwithstanding, the accessible crude clinical information are cosmopolitan , voluminous and heterogeneous in nature. These information should be collected in a sorted out structure. This amassed information can be then incorporated to make a clinical data framework. Information mining gives an utilizer-situated way to deal with novel and obnubilated designs in the information

The information digging executes are auxiliary for responding to business questions and procedures for foretelling the various illnesses in the human services field. Infection augur assumes a significant job in information mining.

This paper investigates the coronary illness forecasts using transfer calculations. These undetectable examples can be used for wellbeing finding in medicinal services information. Information mining innovation bears a proficient way to deal with the most recent and inconclusive examples in the information. The data which is distinguished can be used by the social insurance heads to show signs of improvement housing. Coronary illness was the most essential explanation behind casualties in the nations like India, Coalesced States. Information mining strategies like bunching, Sodality Rule Mining, Relegation calculations, for example, Decision Tree [2], C4.5 calculation, Ingenuous Bayes [4] are accustomed to investigate the various

types of heart - predicated binds. These calculations can be accustomed to upgrade the information stockpiling for down to earth and licit purposes.

EXISTING SYSTEM

Various works in writing related to the conclusion of Heart illness using information mining strategies have boosted this work. A concise writing overview is introduced here.

A model Perspicacious Heart Disease Prognostication System worked with the help of information mining strategies specifically, Neural Network, Naïve Bayes, and Decision Tree. Results show that every strategy has its rare power in understanding the targets of the characterized mining objectives. IHDPS can answer many-sided "imagine a scenario in which" questions which ordinary choice emotionally supportive networks can't be proposed by Sellappan Palaniappan et al. [2]. The outcomes outlined the boorish life of every one of the philosophies in fathoming the objective of the assigned mining targets. IHDPS was fit for reacting questions that the typical choice emotionally supportive networks weren't prepared to . It encouraged the establishment of critical savviness, for example, designs, connections in the midst of clinical components associated with coronary illness. IHDPS remains salubrity web-predicated, utilizer-warm, dependable, versatile and expandable.

The conclusion of Heart Disease, Blood Pressure and diabetes with the profit of neural systems was presented by Niti Guru et al. [7]. Analyses were completed on a tested informational collection of patient's records. The Neural Network is prepared and tried with 13 information factors like essential sign , Age, Angiography's report and along these lines the like. The managed arrange has been urged for finding of heart sicknesses. Preparing was done with the benefit of back engendering calculation. At whatever point new

information was embedded by the medico, the framework recognized the obscure information from examinations with the prepared information and caused a list of likely maladies that the patient is weakly helpless to.

In 2014, M.A.Nishara BanuB.Gomathy Pedagogia, Department of registering and Engineering has distributed an inquiry paper "Sickness Forecasting System Utilizing handling Methods"[8].In this article, the preprocessed information is grouped using bunching calculations as K-indicates to gather relevant information in a database. Maximal Frequent Item set Algorithm (MAFIA) is applied for mining maximal incessant model in coronary illness database. The customary examples can be consigned into various classes using the C4.5 calculation as preparing calculation using the idea of data entropy. The outcome exhibits that the planned forecast framework is equipped for soothsaying the coronary failure prosperously.

In 2012, T.John Peter and K. Somasundaram Preceptor, Dept of CSE introduced a paper, "An Empirical Study on Prognostication of Heart Disease using assignment information mining technique"[5]. In this examination paper, the utilization of example apperception and information digging methods are used for augur of danger in the clinical space of coronary illness medication is proposed here. A portion of the circumscriptions of the customary clinical scoring frameworks are that there is a nearness of inherent direct blends of factors in the info set, and subsequently they are not gifted at displaying nonlinear involute collaborations in clinical spaces. This imperative is dealt with in this examination by utilization of assignment models which can verifiably recognize involute nonlinear connections among free and ward factors just as the staff to distinguish every single imaginable cooperation between diviner factors.

In 2013, Shamsher Bahadur Patel, Pramod Kumar Yadav, and Dr. D. P.Shukla introduced an inquiry paper,"Predict the Diagnosis of heart

condition Patients Utilizing Relegation Mining Techniques "[6].In this examination paper, the medicinal services industry, the data digging is particularly used for the forecast of coronary illness. The goal of our attempts to soothsay the conclusion of coronary illness with a limited number of traits using Naïve Bayes, Decision Tree.

- **DATA SOURCE**

  Clinical databases have aggregated a bounty of data about patients and their ailments. The term Heart malady holds bunches of various conditions that influence the heart. Coronary illness is the main source of loss of individuals on the planet. The term cardiovascular illness contains an abundant scope of conditions that influence the heart and the veins and the manner by which blood is siphoned and coursed through the body. Records set with clinical traits were gotten from the Cleveland heart condition database. With the profit of the dataset, the examples related to the respiratory failure finding are separated. The records were part equipollently into two datasets: preparing dataset and testing dataset. A sum of 340 records with 78 clinical characteristic were gotten. All characteristics are numeric-esteemed. We are dealing with a repressed arrangement of qualities, for example just 15 properties. The accompanying table shows the rundown of qualities on which we are working.

PROPOSED SYSTEM

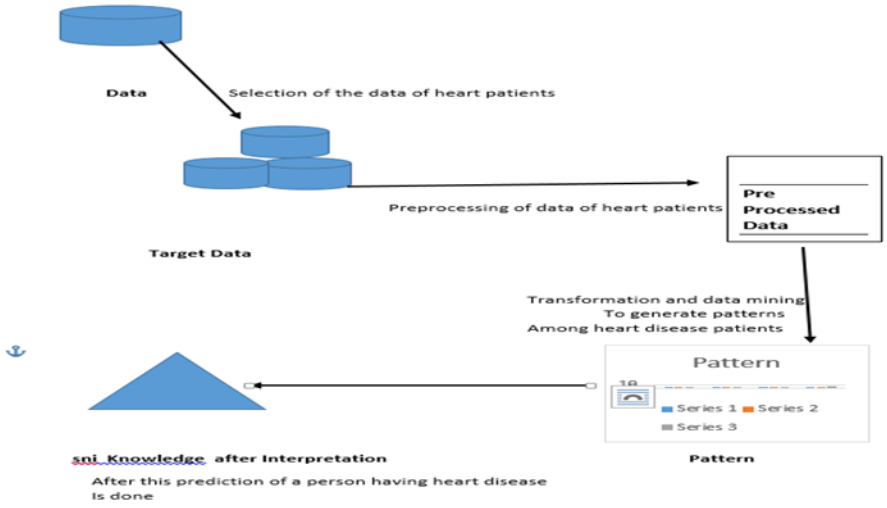


Figure 1.1 Working architecture

Classification procedures of Data Mining and Machine Learning Algorithms have significant influence in anticipating just as information mining. Likewise, side effects of cardio sickness may not be significant and subsequently are commonly disregarded. Enormous Data like records of patients is dealt with utilizing Hadoop Map Reduce technique.[12] Different kinds of investigation of the AI calculations is done and graphical portrayal of the outcomes is accommodated simpler comprehension.

The working of this framework is depicted in a bit by bit:

1. Amassment of information is done in which it contains tolerant subtleties.
2. Cull of qualities which is required for the visualization of coronary illness.
3. After distinguishing the accessible information assets, they are additionally separated, cleaned, made into the ideal structure.
4. Different assignment methods as verbalized will be applied on pre-handled information to forecast the accuracy of coronary illness.
5. Precision measure analyzes the accuracy of various classifiers.

IMPLEMENTATION:

Data Collection from heath care unit:

Information is being amassed from social insurance unit to play out this task. The amassed is comprise old enough, sex, circulatory strain and so forth. The information amassed is as exceed expectations sheet(.csv) group.

Charactestic	Description
Sex	Male or female
Chest Pain Type	value 1: typical type 1 angina, value 2: typical type angina, value 3: non- angina pain; value 4: asymptomatic
Fasting Blood Sugar	value 1: > 120 mg/dl; value 0:< 120 mg/dl
Age	In year
Height	In cms
Weight	In kg
Serum Cholesterol	In mg/dl
Thalach	maximum heart rate achieved
Oldpeak	ST depression induced by exercise relative to rest
Blood Pressure	In mm hg
Electrocardiograph	value 0:normal, value 1:having st t wave abnormality, value 2: showing definite left ventricular
Induced Angina	value 0: no, value 1: yes

Figure 1.2 Entities in dataset

Data Cleansing:

Information Cleansing is the procedure where of recognizing or redressing or abstracting degenerate or incorrect information from a record set, table, or database and alludes to distinguishing deficient, mistaken or superfluous parts of the information and afterward supplanting, altering, or erasing the filthy or coarse information. In the wake of purifying, an informational collection ought to be steady with other related informational indexes in the framework.

Algorithm Used:

k-Nearest Neighbor Algorithm

The k-most proximate neighbors computation (k-NN) is a non-parametric strategy used for arrange and return. In the two cases, the information involves the k most proximate planning cases in the segment space . The yield depends on upon whether k-NN is used for gathering or return, k-NN is a hardly model predicated realizing, where the limit is simply approximated locally and all computation is surrendered until course of action. The k-NN figuring is among

the most effortless of all AI counts. Both for order and return, it very well may be sufficiently fundamental to distribute weight to the responsibilities of the neighbors, so the more proximate neighbors contribute more to the unremarkable than the more far away ones. For example, a run of the mill weighting plan includes in giving each neighbor a load of 1/d, where d is the disseverment to the neighbor.

Naive Bayesian Classifier

In data mining we use guileless Bayesian game plan, in which we take various sorts of educational hoard and as betokened by that doubt we get the yield required most. It is a simple probabilistic classifier. This transfer requires some outrageous and versatile direct parameters in kind of variable.The Bayesian Relegation outlines a coordinated learning methodology and a quantifiable procedure for request .In this we can use most extraordinary likelihood planning strategy which should be conceivable in evaluating a closed shape articulation which takes straight time rather than iterative guess.

Bayes rule:

It is an unexpected probability which communicates that there is a probablity of some yield when given some data where a dependence relationship remains alive among observation and end.

Allow respect be "c" and data be "e" at that point probability is betokened by

$$p(c/e) = (p(e/c)p(c))/p(e)$$

Decision Tree

A tree which uses a tree like structure which implicatively implies it wires with a root center point, branches and a leaf center point. Each inward centers assigns a test on a trademark, each branch betokens the consequence of a test, and each leaf center holds a class category. By and cosmically tremendous DST are used as a

part of choosing informative winnows which can get in touch with us to our goal in most beneficial manner adventitiously a convincing invention for AI. The ways from root to leaf verbalizes with injuctively approve runs the show. Tree models where the target variable can take a compelled game plan of characteristics are called portrayal trees; in these tree structures, leaves verbalize with class checks and branches verbalize with conjunctions of segments that quick those class denominations.It takes after ravenous estimation that is, it picks the gainful way which can be optically perceived from the start and besides it takes after recursive and segment and defeat methodology. This estimation begins with whole game plans of getting ready educational files among which DST separates the one which has most data for gathering and incites a test center point. Classifier time won't work if all the arrangement instructive records radiates from a related class or in case it isn't worth to propagate advance with an additional separation when we optically observe that after further portrayal in like manner it incites request just with pre-gathering limit.DST figuring forms information get for every trademark and each round, the one with the most information get is winnowed for the testing reason.

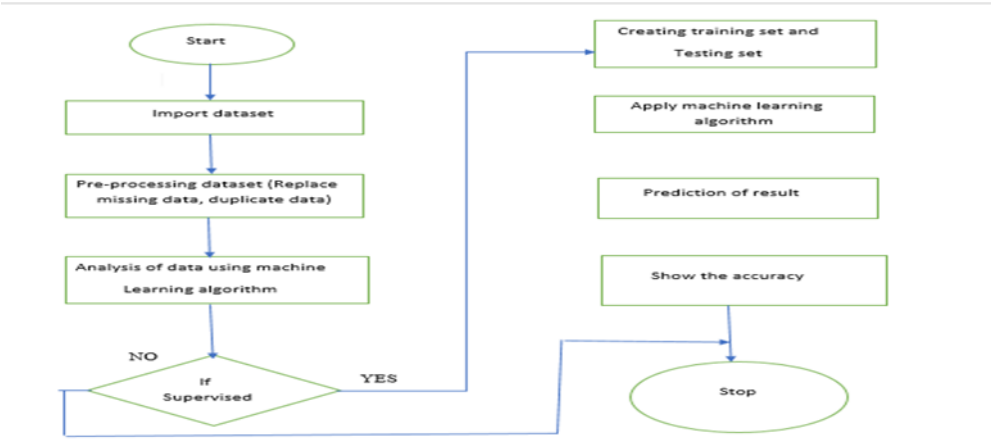
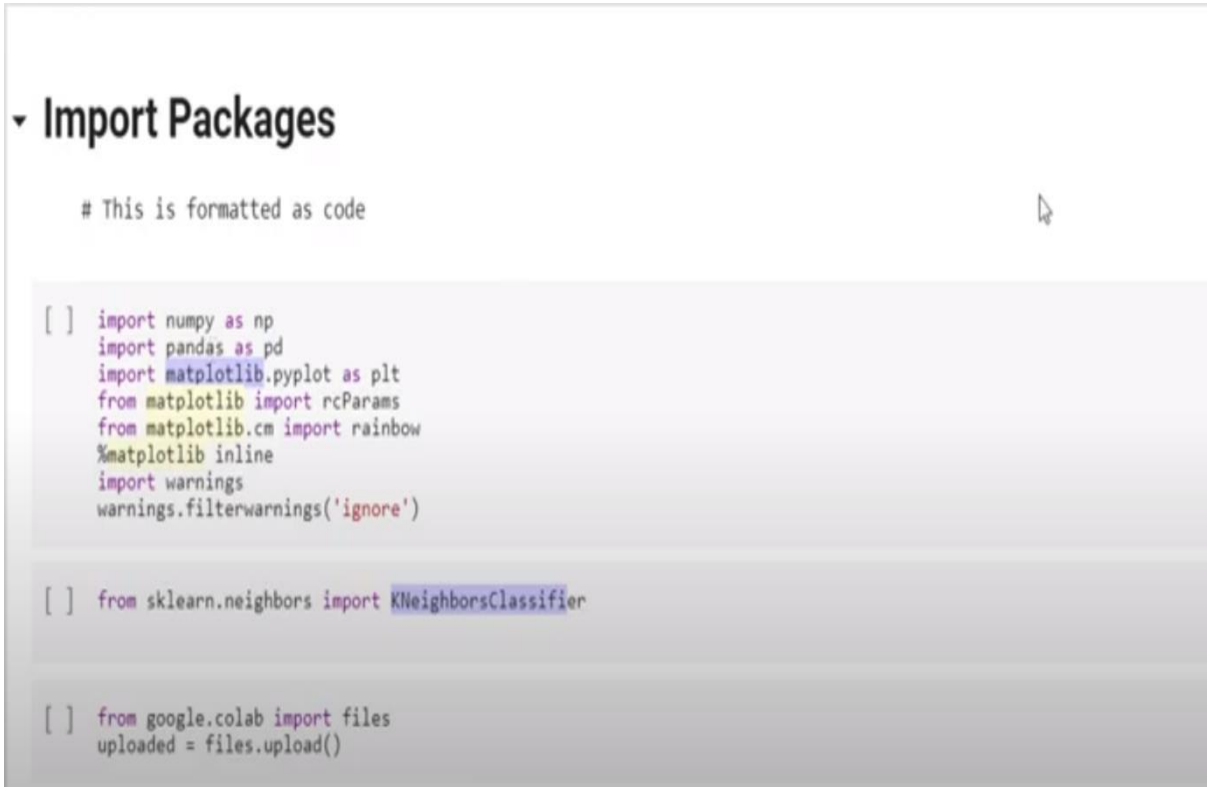
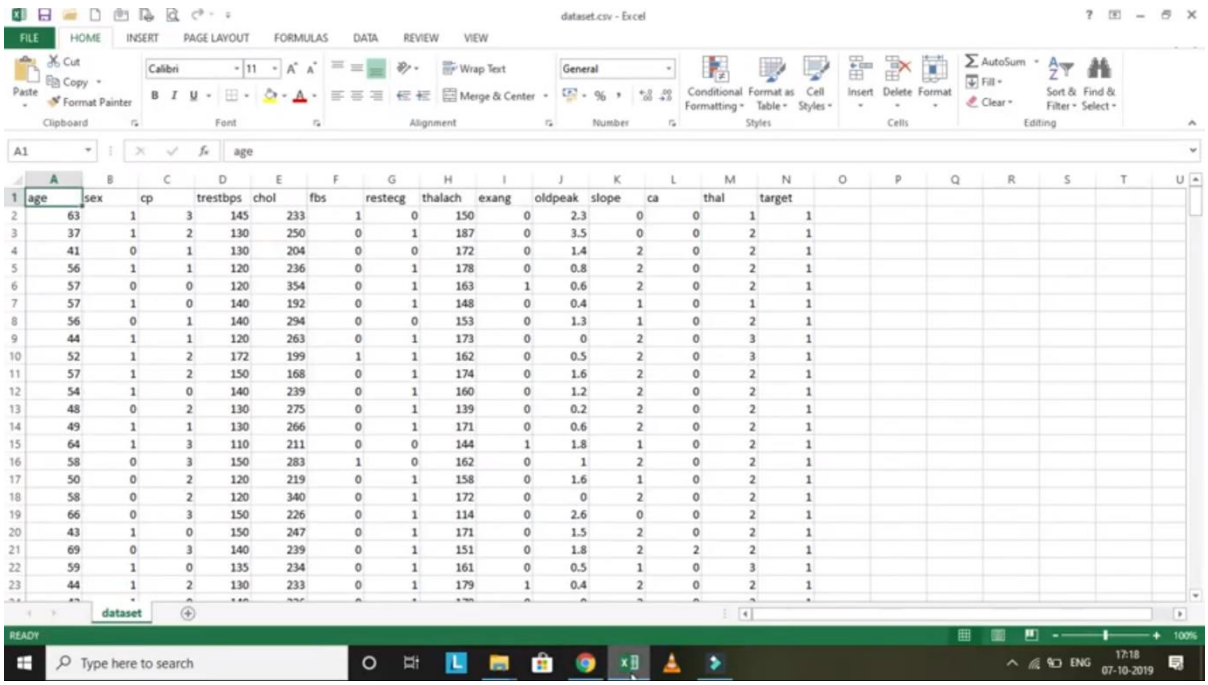


Figure 1.3 Working dataflow diagram

OUTPUT

Screenshots of the code implementations:





age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1

Data Processing

```
[ ] dataset = pd.get_dummies(df, columns = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])
```

```
[ ] from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import StandardScaler
    standardScaler = StandardScaler()
    columns_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
    dataset[columns_to_scale] = standardScaler.fit_transform(dataset[columns_to_scale])
```

```
[ ] dataset.head()
```

	age	trestbps	chol	thalach	oldpeak	target	sex_0	sex_1	cp_0	cp_1	cp_2	cp_3	fbs_0	fbs_1	restecg_0	r
0	0.952197	0.763956	-0.256334	0.015443	1.087338	1	0	1	0	0	0	1	0	1	1	
1	-1.915313	-0.092738	0.072199	1.633471	2.122573	1	0	1	0	0	1	0	1	0	0	
2	-1.474158	-0.092738	-0.816773	0.977514	0.310912	1	1	0	0	1	0	0	1	0	1	

```
[ ] y = dataset['target']
    X = dataset.drop(['target'], axis = 1)

[ ]
    from sklearn.model_selection import cross_val_score
    knn_scores = []
    for k in range(1,21):
        knn_classifier = KNeighborsClassifier(n_neighbors = k)
        score=cross_val_score(knn classifier,X,y,cv=10)
        knn_scores.append(score.mean())

[ ]

[ ] plt.plot([k for k in range(1, 21)], knn_scores, color = 'red')
    for i in range(1,21):
        plt.text(i, knn_scores[i-1], (i, knn_scores[i-1]))
    plt.xticks([i for i in range(1, 21)])
    plt.xlabel('Number of Neighbors (K)')
    plt.ylabel('Scores')
    plt.title('K Neighbors Classifier scores for different K values')
```

Screenshots of the output:

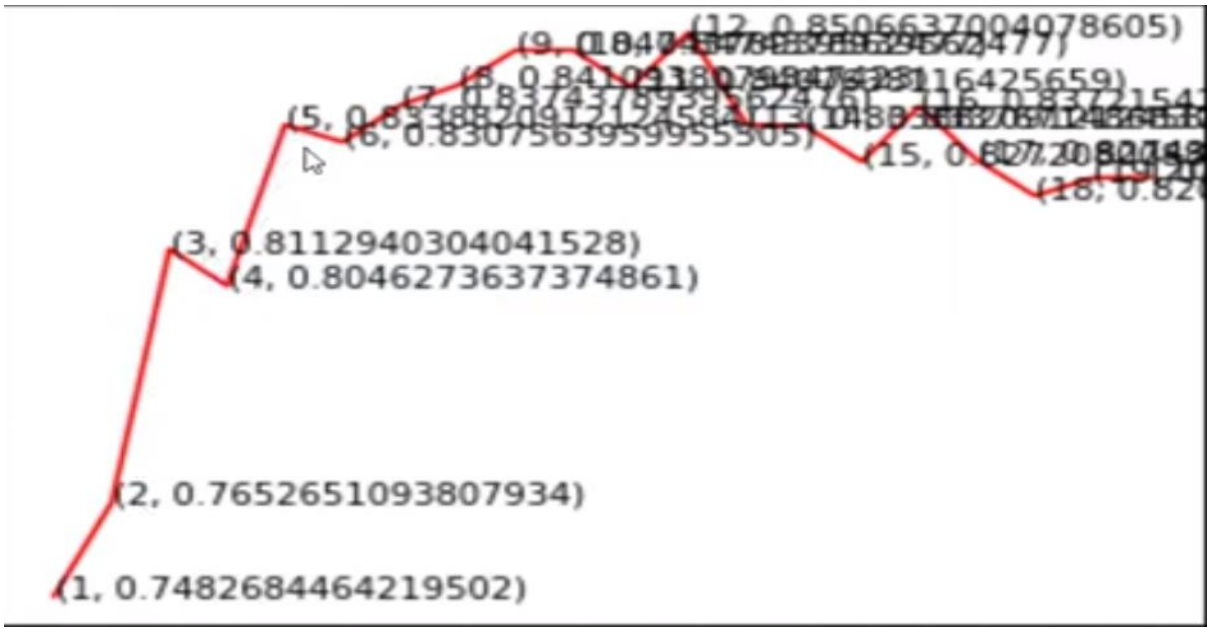


Fig 1.4 Output

CONCLUSION

The objective of our work is to give an examination of various data mining procedures that can be used in modernized coronary disease prospect structures. Various procedures and data mining classifiers are described in this work which has created as of late for gainful and puissant coronary disease end. The assessment exhibits that particular advances are used as a segment of the extensive number of papers with taking assorted number of properties. Along these lines, special progressions used showed the unmistakable precision to one another. In a couple of papers it is shown that that Decision Tree has performed well with 96.2% accuracy by using 13 properties. Hence, exceptional headways used and showed the differing accuracy depends on number of properties taken and instrument used for usage. Pushed by the general growing mortality of coronary ailment patients consistently and the openness of brobdingnagian proportions of data, researchers are using data mining techniques in the finding of coronary sickness. Regardless of the way that applying awareness profundity methodologies to give human housing experts in the finding of cardio ailments is having some accomplishment, the usage of data mining frameworks to recognize a consistent treatment for coronary ailment patients has gotten less thought.

We can incorporate a couple of segments with the goal that it's accuracy rate can be extended and it's degree of locale can in like manner be widened. These can be significantly auxiliary in pending days of our future as it will lessen crafted by experts patients can in like manner be recovered progressively speedy. In any case, there is in like manner a couple of slip-ups if which is done it can cost amazingly.

REFERENCES

[1] V. Manikantan and S. Latha, "Anticipating the investigation of coronary illness side effects utilizing therapeutic preparing techniques", International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013.

[2] Sellappan Palaniappan and Rafiah Awang, "Astute coronary illness expectation framework utilizing preparing methods", International Journal of registering and Network Security, vol.8, no.8, pp. 343-350,2008.

[3] K.Srinivas, Dr.G.Ragavendra and Dr. A. Govardhan," A Survey on expectation of heart dreariness utilizing information preparing techniques",International Journal of information Mining and Knowledge Management Process (IJDKP)vol.1, no.3, pp.14-34, May 2011.

[4] G.Subbalakshmi, K.Ramesh and N.Chinna Rao," Decision support in coronary illness expectation framework utilizing Naïve Bayes", ISSN: 0976-5166, vol. 2, no. 2.pp.170-176, 2011.

[5] T.John Peter , K. Somasundaram, "An Empirical Study on Prediction of heart condition utilizing order handling procedure" IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012.

[6] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla, "Foresee the Diagnosis of heart condition Patients Using Classification Mining Techniques",IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS),Volume 4, Issue 2 (Jul. - Aug. 2013).

[7] Niti Guru, Anil Dahiya, Navin Rajpal, "Choice system for coronary illness Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. I (January - June 2007).

[8] M.A.Nishara Banu, B.Gomathy, "Illness Forecasting System Using handling Methods," International Conference on Intelligent Computing

Applications,2014.

[9] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, "Foreseeing Disease By Using handling upheld Healthcare Information System", IEEE International Conference on Granular Computing,2012.

[10] Chotirat "Ann" Ratanamahatana and Dimitrios Gunopulos,"Scaling up the Naive Bayesian Classifier:Using Decision Trees for Feature.