REVIEW PAPER ON DISEASE PREDICTION USING DATA MINING TECHNIQUES

[1]Kruti S Khara, [2]Hemant D. Vasava, [3]Mosin I Hasan
[1]P.G. Student, [2]Assistant Professor, [3]Assistant Professor
[1]Dept. of Computer Engineering,
[1] Birla Vishwakarma Mahavidyalaya, V.V. Nagar, Gujarat, Anand, Gujarat

Abstract—Medical Diagnosis System plays vital role now a days. They are used for treatment of Disease[8]. In present time there are many technologies and innovation like scanning system supporting the doctors in their clinical decision making. But this is much costly and thus due to this poor and needy people are not able to use this services.[15] This leads to increasing the death rate. The healthcare industries collect large amount of data that are not feasible to handle manually. Due to advancement in technologies today many hospitals use information and generate relation among them.[15] Data mining is the techniques used for prediction of disease. In this papers there are different techniques used for prediction of Disease.

***IndexTerms*- KNN Algorithm, Genetic Algorithm, Naïve Bayes, Neural Network, Prediction.**

I. INTRODUCTION

Data mining itself is derived from the name of searching important information from large number of database[1]. Classification is one of the data mining techniques used for prediction. In this paper there are different methods used for prediction of disease. The techniques are neural network, KNN Algorithm, Naïve Bayes algorithm and Genetic Algorithm. The accuracy of different algorithms is defined and different datasets are used to predict the disease[1].

1.1 Introduction to data mining

Data mining computational process of finding patterns in large data sets including methods at the intersection of machine learning, artificial intelligence, statistics and database systems. Used to analyze the rich collection of data from different perspectives and deriving useful information. The main focus of data mining process is to obtain information from the data and converted it into and knowledgeable and reasonable structure for further use. [9]

Extracting significant information from huge collection of data which can be in various forms is the process called data mining. Data mining is also termed as knowledge discovery in databases (KDD). [10]

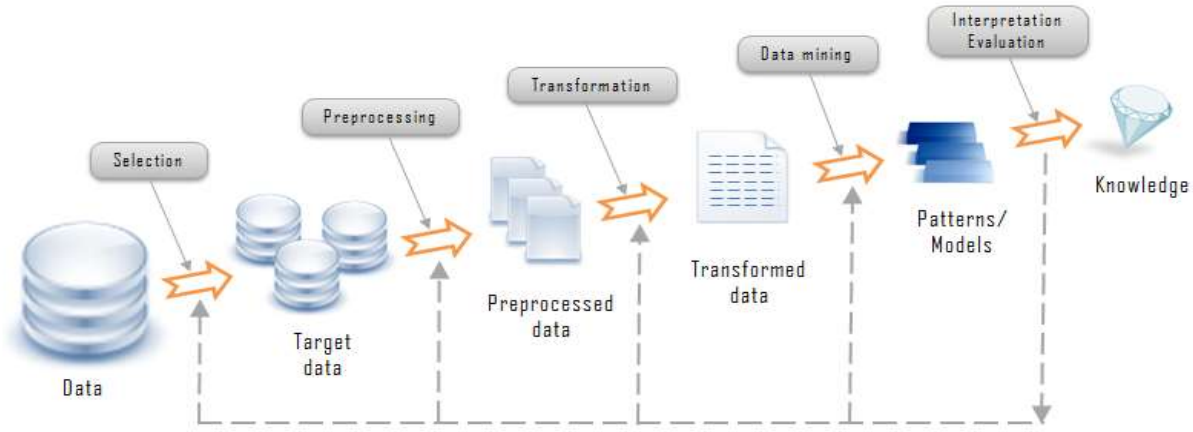

Fig 1 KDD Process

1.2 Medical data mining

Medical Diagnosis System plays vital role now a days. They are used for treatment of Disease. In present time there are many technologies and innovation like scanning system supporting the doctors in their clinical decision making. But this is much costly and

thus due to this poor and needy people are not able to use this services. This leads to increasing the death rate. The healthcare industries collect large amount of data that are not feasible manually. Due to advancement in technologies today many hospitals use information and generate relation among them. Data mining Involves lot of accuracy and uncertainty it include massive data for making decision. [11]

1.3 Classification

Classification means to predict something from the given output. For prediction the algorithm consist of training and testing sets. First the data is trained and then it is tested using testing sets. If the obtain output is correct then we can say the algorithm is accurate [12]. Mining databases for knowledge can be of any type like the relational databases, object oriented databases and other kind of databases. Classification based on the knowledge mined utilizes the data mining functionalities like classification, prediction, outlier clustering and many others. The system categorization based on techniques makes use of methods of data analysis employed like machine learning, statistics, and visualization. Classification of data mining systems can also be done based on kinds of applications adapted like finance, stock markets and many other applications. [13]

Data mining can be considered as a synonym for knowledge Discovery of data or as a step in the process of knowledge discovery. Knowledge discovery is an iterative process and it consists of seven basic steps. The first four steps are used to preprocess the data i.e. data is prepared for mining. The steps can be elaborated as follows: [13]

• Data cleaning • Data integration • Data selection • Data transformation • Data mining • Pattern evaluation • Knowledge presentation [13].

II. PROPOSED ALGORITHMS

2.1 KNN ALGORITHM

KNN is a strategy which is used for viewing of nearest in the component space. KNN is the most essential kind of occurrence based learning. A separation measure is expected to decide the "closeness" of examples. KNN orders a case by discovering its closest neighbors and picking the most mainstream class among the neighbors.KNN is a lazy supervised algorithm as it takes more time to train data until Datasets are used for classification. This is used for classification and prediction. [1]

1. Take a sample dataset of n columns and m rows named as R. In which n-1th columns are the input vector and nth column is the output vector. [5]
2. Take a test dataset of n-1 attributes and y rows named as P.[5]
3. Find the Euclidean distance between every S and T by the help of formula[5].

   Euclidian distance= root of (R-P) square
4. Then, Decide a random value of K. K is the no. of nearest neighbors.[5]
5. Then with the help of these minimum distance and Euclidean distance find out the nth column of each.[5]
6. Find out the same output values.[5]

If the values are same then the patient is suffering from disease otherwise not. Then the accuracy and error rate is calculated. Accuracy obtain was 82% by using this algorithm and the dataset was taken from**.**
www.stanford.edu/~hastie/Papers/LARS/diabetes.data this is used to predict diabetes [5]. The dataset comprises of 11 attributes which are as follows: [5]

1) Age (years)
2) Sex
3) Body mass index
4) Blood Pressure (mm Hg)
5) Plasma Glucose Concentration (Glucose tolerance test)
6) Triceps Skin folds
7) 2-Hour serum insulin
8) Diabetes Pedigree function
9) Cholesterol Level
10) Weight (kg)
11) Class variable (0 or 1)

2.2 Naive Bayes Algorithm

It is used in medical domain for solving diagnosis problem. According to Bayes theorem of probability theory[7]

$$P(ck|\{A\}) = P(\{A\}|ck)P(ck)/P(\{A\}) \quad (1)$$

The steps followed by it are. [7]

1. The Symptoms are given to the program.
2. The probability of each are calculated and stored. The Equation used to calculate probability is

$$P(Ck) = |Ck|+1/m+ \Sigma|Cj| \quad (2)$$

Here m is number of class and p is probability. Ck is probability of attributes

$$P(Ck)=P(A1|Ck)*P(A2|Ck)*P(A3|Ck)*\ldots\ldots*P(AM|Ck) \quad (3)$$

3. Then the classes are divided into three parts[7]

A=Absolutely Normal

B=Less Probability

C= Suffering from Disease

Using this algorithm it was obtain that the efficiency is 70% and the dataset was and the disease was swinflu [7]

1Fever- (Value 0:98-99.9f, Value 1:100-102 f, Value 2:100-102 f)
2. Body ache-(Value 0: No, Value 1: Yes, Value 2: Severe)
3. Blood Pressure-(Value 0:120/80, Value 1: Abnormal
4. Color of Nails-(Value 0: Pink, Value 1: Blue)
5. Breathlessness-(Value 0: No, Value 1: Moderate, Value 2: Severe)
6. Diarrhea-(Value 0: No, Value 1: Sometimes)
7. Vomiting-(Value 0: No, Value 1: Occasionally)
8. Cough-(Value 0: No, Value 1: Yes, Value 2: Severe
9. Skin Color-(Value 0: Not Applicable, Value 1: Blue)
10. Sore Throat-(Value 0: No, Value 1: Occasionally, Value 2: Severe)
11. Chills-(Value 0: No, Value 1: Occasionally, Value 2: Severe)
12. Age-(Value 0:0-18 years, Value 1:18-30 years, Value 2:31-45 years, Value 3: Above 45)
13. Gender-(Value 0: Male, Value 1: Female)
14. Lung Disease-(Value 0: No, Value 1: Yes)
15. Chest Pain-(Value 0: No, Value 1: Yes)
16. Pandemic Area-(Value 0: No, Value 1: Yes)
17. Service in the health industry-(Value 0: No, Value 1: Yes)

2.3 Neural Network [3]

1. In first step neural network is trained using the dataset. This data must be pre-processed by pre-processing steps
2. Then the input layer, hidden layer and output layer are given and data is trained.
3. The output of trained data is then compared with original data. If the output is similar then it is declared as trained data.
4. Now it comes of testing it. For testing the data is given and the result is obtained from it.

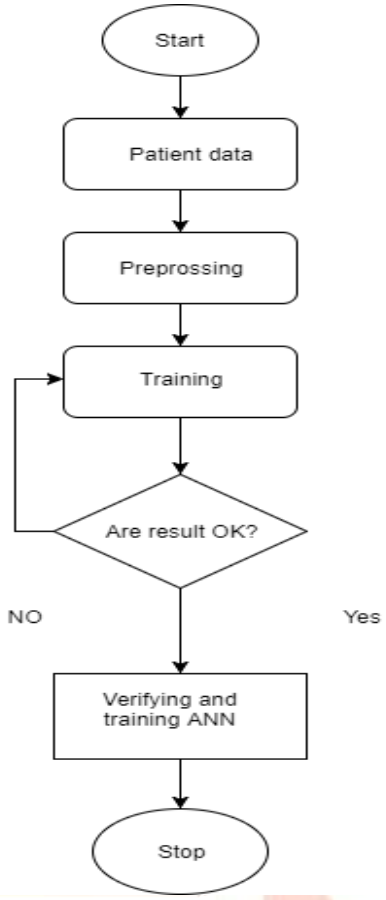

Fig 2 Work Flow [3]

For this the datasets taken were of UCI and the disease was heart disease and it gave 95% of accuracy [3]

1. Sex
2. Age
3. Blood Cholesterol
4. Blood pressure
5. Hereditary
6. Smoking
7. Alcohol
8. Physical activities
9. Diabetic
10. Obesity
11. Stress
12. Heart disease

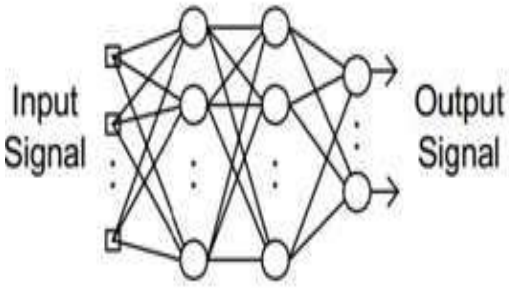

Fig 3Neural Network Architecture [6]

2.4 Genetic Algorithm

Genetic algorithms are computing methodologies constructed in analogy with the process of evolution. It closely resembles the natural process of regeneration, reproduction, inheritance evolution. Genetic algorithms are typically used for problems that cannot be solved efficiently with traditional techniques. Genetic algorithms seem to be useful for searching very general spaces and optimization problems. Each solution generated in Genetic algorithms is called a chromosome (individual).Each chromosome is made up of genes, which are the individual elements (alleles) that represents the problem. The collection of chromosomes is called a population. [14]

The internal representation of the chromosomes is known as its genotype. This can be either bit strings or gray codes or hexadecimal codes. The external manifestation of the genotype or the real world representation of the genotype is known as the phenotype. [14] Basically there are three genetic operators are used for generating new strings. The functions of genetic operators are as follows: [14]

1) Selection: selection deals with the probabilistic survival of the fittest in that, more fit chromosomes are chosen to survive.

2) Crossover: This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point. Various types of crossover operators are a) single point b) two point c) uniform d) half uniform e) reduced surrogate crossover f) shuffle crossover g) segmented crossover[14].

3) Mutation: mutation alters the new solutions so as to add stochasticity in the search for better solution. The most common method way of implementing mutations is to select a bit at random and flip (change) its value. There are 2 types of mutations use in genetic network programming 1) mutating the judgment node 2) mutating the value of the judgment node. In associative classification attributes and their values are taken as judgment nodes and class values as processing nodes[14].

Fitness function: Ideally the discover rules should have a) high predictive accuracy b) be comprehensible c) be interesting. The accomplishment of a genetic algorithm is directly linked to the accuracy of the fitness function. [14]

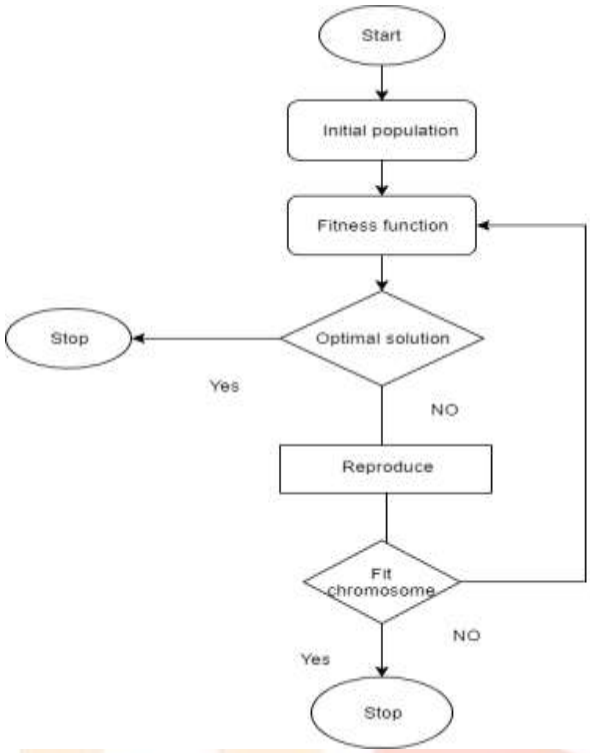

Fig 4 Genetic Algorithm [2]

III. RESULT

Disease	Algorithm Used	Accuracy
Diabeties	KNN Algorithm	82%
Swine Flu	Naïve Bayes Algorithm	70%
Heart Disease	Neural Network	95%

IV. CONCLUSION

In this paper we have determine many algorithm to predict disease and from this paper we conclude that Neural network has most accuracy of 95% as compared to other algorithm

REFERENCES

[1] Diagnosis of heart disease using Average K-nearest neighbor Algorithm using data mining technique by C.Kalaiselvi, PhD and Tirupur,Tamilnadu,India,2016 International Conference on Computing for Sustainable Global Development

[2] Analytical Study of Heart Disease PredictionComparing With Different Algorithms by Sana Bharti M.Tech Scholar, ASET-CSE Amity University Noida, India. and Dr.Shaliendra Narayan Singh Associate Professor Amity University Noida, India, International Conference on Computing, Communication and Automation (ICCCA2015), 2015 IEEE 7

[3] Artificial Neural Network Approach for Classification of Heart Disease Dataset byManjusha B. Wadhonkar ,Prof. P.A. Tijare and Prof.S.N.Sawalkar3, Volume 3, Issue 4, April 2014, International Journal of Application or Innovation in Engineering & Management

[4] Implementation of Genetic Algorithm in Predicting Diabetes byS.Sapna., Dr.A.Tamilarasi and M.Pravin Kumar, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

[5] Diagnosis of Diabites melltus using K nearest neighbor algorithm by Krati Saxena, Dr. Zubair Khan, Shefali Singh, International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4, July-Aug 2014.

[6] Prediction of Heart Disease Using BP-Neural Network & Genetic Algorithm by Harshal Yeole, Sayali Ukirde, Sushma Khadse, Priyanka Pednekar, International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue National Conference "NCPCI-2016", 19 March 2016

[7] Health Care Decision Support System For Swine Flu Prediction Using Naïve Bayes Classifier By Binal A. Thakkar, Mosin I. Hasan and Mansi A. Desai, 2010 International Conference on Advances in Recent Technologies in Communication and Computing

[8] Cardiovascular Disease Prediction System Using Genetic Algorithm And Neural Network by Bhuvaneswari Amma N.G. in Computing, Communication and Application, 2102 international Conference.
[9] https://en.wikipedia.org/wiki/Data_mining
[10] https://en.wikipedia.org/wiki/KDD
[11] https://en.wikipedia.org/wiki/Health_informatics
[12] https://en.wikipedia.org/wiki/Statistical_classification
[13] Survey on Classification Techniques Used in Data Mining and their Recent Advancements by Saranya Vani.M1 , Dr.S. Uma2 ,Sherin.A3 , Saranya.K in International Journal of Science, Engineering and Technology Research, Volume 3, Issue 9, September 2014
[14] Heart Disease Prediction System using Associative Classification and Genetic Algorithm by M.Akhil jabbar,Dr.Priti Chandrab, Dr.B.L Deekshatuluc in International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012
[15] Genetic Neural Network Based Data mining in prediction of Heart Disease Using Risk Factor by Syed Umar Amin, Kavita Agarwal, and Dr. Rizwan Beg in IEEE Conference on Information and Communication Technologies 2013