International Journal of
Science Inventions Today
www.ijsit.com
ISSN 2319-5436
Research Article

# A DATA MINING APPROACH FOR PREDICTION AND TREATMENT OFDIABETES DISEASE

VelidePhani Kumar [1*] and Lakshmi Velide[2]

[1*] *Principle client consultant, TELE 9 Technologies Limited, Hyderabad Andhra Pradesh,India.*

[2]*Asst Professor, Department of Biotechnology, GokararajuRangaraju Institute ofEngineering And Technology, Kukatpally, Hyderabad, Andhra Pradesh ,India.*

**ABSTRACT**

The advancement in computers provided large amount of data. The task is to analyse the input data and obtain the required data which can be done by various data mining techniques. The diagnosis of diabetes is a significant and tedious task in medicine. So the present work focus on analysis of diabetes data by various data mining techniques which involve,Naive Bayes, J48(C4.5) JRip ,Neural networks, Decision trees, KNN, Fuzzy logic and Genetic Algorithms based on accuracy and time. The 9 selected attributes wereSex, Diastolic B.P, Plasma glucose, Skin fold thick, BMI, Diabetes Pedigree type, No. of times Pregnant, 2 hr Serum Insulin and Diabetes probability**.**J48(C4.5) reported simple, efficient classifier of diabetes data.

**Keywords**: Dibetesdata,Attributes,Naïve Bayes, J48 (C4.5), JRip, Neural networks, Decision trees, KNN, Fuzzylogic and Genetic algorithm.

## INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used for industrial, medical and scientific purposes. As such the process of data mining involves sorting through large amounts of data and discovering patterns in the data [1]. Medical, Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods[2].Medical reports always gives useful information for diagnosis and also facilitates therapeutic improvement. The medical knowledge management is shown as cycle among clinical research, guidelines, quality indicators, performance measures, outcomes and concepts [3]. Thus huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analysed by traditional methods. Medical data mining is used in the knowledge acquisition and analyses the information obtained from research reports, medical reports, flow charts, evidence tables, and transform these mounds of data into useful information for decision making[4].Diabetes is a major health problem in India.There is a long history of diabetic registries and databases with systematically collected patient information. This disease has many side effects such as higher risk of eye disease, higher risk of kidney failure, and other complications. However, early detection of the disease and proper care management can make a difference. The main purpose of identifying a suitable diabetes data system guides the diabetic patients during the disease. Diabetic patients could benefit from the diabetes data system by entering their daily glucoses rate and insulin dosages; producing a graph from insulin history; consulting their insulin dosage for next day. The system is not only for diabetic patient, but also for the people who suspect if they are diabetic. The present work has taken up to analyse the obtained data of diabetic patients by various data mining algorithms which can be helpful for medical analysts or practitioners for accurate diabetes diagnosis.

## METHODOLOGY

It includes examining the publications, journals and reviews in the field of computer science, engineering, data mining and diabetes reports in recent times. A five year sample dataset is created to mine for knowledge discovery. The actual dataset contains 865 instances. The data mining tool Weka 3.6.6 is used for experiment. Initially missing values were identified in the data set and they were replaced with appropriate values using Replace missing values filter from 3.6.6[5]. Following data mining techniques have been applied on diabetes data base.

### Supervised machine learning algorithm:

The obtained data is classified based on various supervised machine learning algorithms, like Naive Bayes, Decision List,KNN,JRip and J48(C4.5).TANAGRA a data mining tool for academic and research purpose used to classify the obtained data and evaluated using 10-fold cross validation[6]. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. It provides the user an easy analysis either real or synthetic data. This tool also allows the users the easy

addition of their own data mining methods, to compare their performances. It is a wide set of data sources, direct access to data warehouses and databases, data cleansing, interactive utilization

## Naïve Bayes:

Naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. It can be trained very efficiently in a supervised learning. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) [7].

## J48 (C4.5):

It is an open source algorithm in Weka data mining tool. A decision tree can be generated from the input data by C4.5 proramme. It is an algorithm used to generate a decision tree and is an extension of Quinlan's earlier ID3 Algorithm. The decision trees generated by this can be used for classification and so referred to as statistical classifier [8].

## JRip:

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William Cohen (1995) as an optimized version of IREP. It is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms [9].

## Decision tree:

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labelled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree based models which include classification and regression trees, are the common implementation of induction modelling[10].Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications.

Experiments are conducted by using the training data set of 865 instances with 15 different attributes. Depending upon the attributes, the dataset is classified into two parts, i.e. 60% of the data is used for training and 40% is used for testing. Performance of each algorithm is determined and comparison is made based on the accuracy and evaluation time of calculation for each algorithm [11].

## Neural Network:

An artificial neural network (ANN), often just called a "Neural network" (NN), is a mathematical model or computational model based on biological neural network. Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements (neurones) working in parallel to solve a specific problem [12]. In medicine, ANNs have

been used to analyse blood and urine samples, track glucose levels in diabetics, determine ion levels in body fluids and detect pathological conditions [13]. Artificial Neural networks are well suited to tackle problems that people are good at solving, like prediction and pattern recognition. Neural networks have been applied within the medical domain for clinical diagnosis, image analysis and interpretation [14], signal analysis and interpretation and drug development [15].

## Fuzzy logic and genetic Algorithm:

Fuzzy set theory and fuzzy logic are highly suitable for developing knowledge based systems in healthcare for diagnosis of diseases [16].Experiments are conducted in Mat lab using fuzzy tool. For this, Mamdani model of fuzzy system is used. The fuzzy rules are generated based on experts' knowledge in this domain..The dataset from UCI machine learning repository is used,and only 6 attributes are found to be effective and necessary for diabetes prediction. In the proposed system, the input is the set of all the selected features and the output of the system is to achieve a value 0 or 1 that indicates the absence or presence of diabetes in patients. In fuzzy logic process, initially fuzzification is performed bycollecting the crisp set of input data and converting it to a fuzzy set using fuzzy linguistic variables, fuzzy linguistic terms and membership functions. After that, an inference is made based on a set of rules and lastly, de fuzzification step is performed [16].

## STATISTICAL ANALYSIS

Each analysis was replicated thrice by the above said classifiers which were used to compare and evaluate the data based on accuracy and time.

## RESULTS AND DISCUSSION

For better understanding results for each data mining technique have shown in different tables. Various classifiers are used in combination with different data mining techniques for diabetes dataset analysis. Table I gives details about various attributes selected for diabetes data analysis. It also shows application of various data mining techniques to study whether a patient can be diagnosed high, low or medium for diabetes. Table II depicts the outcome of the research work by comparison done with various classifiers. It was reported that J48 (C4.5) had outplayed over other techniques by showing 100% accuracy.J48 is very simple and accurate classifier to make a decision tree over other cassifiers[17].Table III shows that the fuzzy and genetic algorithm generates fuzzy rules based on support set. The results obtained by using supervised machine learning had shown that the time taken for data analysis was high in KNN. The accuracy was high and time taken was least in J48 (C4.5).This shows that the computational cost for data analysis was low in J48(C4.5) and so the performance is accurate(Table IV).

| ID | SEX | Diastolic B.P mm Hg | Plasma glucose mg/dL | Skin fold thick mm | BMI Kg/m$^2$ | Diabetes Pedigree type | No. of times pregnant | 2 hr Serum Insulin muU/ml | Diabetes probability |
|---|---|---|---|---|---|---|---|---|---|
| 1 | F | 100 | 182.5 | 27.76 | 31.75 | 2 | 0 | 140 | High |
| 2 | M | 68 | 98.30 | 35.75 | 28.12 | 1 | - | 54 | Low |
| 3 | M | 88 | 111.36 | 35.25 | 28.95 | 2 | - | 78 | Low |
| 4 | F | 52 | 131.18 | 27.68 | 28.75 | 2 | 1 | 122 | Medium |
| 5 | F | 73 | 142.2 | 28.64 | 28.55 | 1 | 0 | 105 | Medium |
| 6 | M | 92 | 172.28 | 33.25 | 32.65 | 2 | - | 138 | High |
| 7 | M | 82 | 189.25 | 34.56 | 30.25 | 2 | - | 138 | High |
| 8 | F | 115 | 175.68 | 28.25 | 31.25 | 2 | 0 | 135 | High |
| 9 | M | 95 | 135.25 | 31.65 | 29.56 | 2 | - | 102 | Medium |
| 10 | M | 86 | 112.45 | 36.76 | 30.25 | 2 | - | 75 | Low |
| 11 | F | 90 | 156.25 | 30.15 | 27.68 | 2 | 1 | 110 | Medium |
| 12 | M | 59 | 160.54 | 30.75 | 29.35 | 2 | - | 142 | High |
| 13 | M | 55 | 166.34 | 35.62 | 31.25 | 2 | - | 132 | High |
| 14 | F | 82 | 102.52 | 28.45 | 26.75 | 2 | 2 | 72 | Low |
| 15 | M | 95 | 125.75 | 33.45 | 29.54 | 2 | - | 102 | Medium |

**Table I:** Sample Data set

| Classifiers | Accuracy |
|---|---|
| Naïve Bayes | 95.85% |
| JRip | 96.54% |
| J48(C4.5) | 100% |
| Decision Trees | 98.48% |
| Neural Networks | 97.85% |

**Table II**: Comparison of various Classifiers

| S.No | Attributes | Support Set | |
|---|---|---|---|
| | | Diabetic | Non Diabetic |
| 1 | Diastolic B.Pmm Hg | 52-115 | 80-85 |
| 2 | Plasma glucosemg/dL | 131-189 | 98-112 |
| 3 | Skin fold thickmm | 27-35 | 26-36 |
| 4 | BMIKg/$m^2$ | 27-32 | 26-30 |
| 5 | No. of timespregnant | 0-1 | 0-2 |
| 6 | 2 hr Serum InsulinmuU/ml | 102-140 | 54-75 |

**Table III**: Values of various attributes in support set

| Classifiers | Time Taken | Accuracy |
|---|---|---|
| Naïve Bayes | 845min | 55.85% |
| JRip | 765min | 65.48% |
| J48(C4.5) | 658min | 68.58% |
| Decision Tree | 875min | 52.58% |
| KNN | 956min | 50.68 |

**Table IV**: Analysis of classifiers Performance

## CONCLUSION

Thus in conclusion it is shown that various data mining techniques were employed to analyse the obtained diabetes data. J48(C4.5) with 9 attributes had shown accurate and better performance with least time taken for analysis of Diabetes data.

## REFERENCES

1. Witten,I. and Eibe,F. Data mining practical machine learning tools and techniques.2nded,Sanfrancisco: Morgan Kaufmann series in data management systems.,2005.
2. Cunningham,S.J. and Holmes, G. Developing innovative applications in agriculture using data mining.In the proceedings of the south east Asia, Regional computer confederation conference., Newzealand,1999.
3. McCourt,B.,Harrington,R.A.,Fox,K.,Hamilton,C.D.,Booher,K.,Hammond,W.E.,Walden,A. and Nahm,M.Data Standards: At the Intersection of Sites, Clinical Research Networks, and Standards Development Initia-tives. Drug Information Journal.,2007,41(3): 393-404.
4. Wang,X.S.,Nayda,L. and Dettinger,R. Infrastructure for a Clinical Decision–Intelligence System. IBM Systems Journal.,2007,46(1), pp. 151-169.

5. ChaitraliDangare, S. and SulabaApte,S.Improved study of disease prediction using data mining classification techmiques. Int.J.Comp.Appl.,2012,47(10):75-88.
6. http://eric.univ-lyon2.fr/ricco/tanagra/
7. "Naïve Bayes", Wikipedia, March 2013.
8. "C4.5(J48)", Wikipedia, March 2013.
9. "JRip",Wikipedia,March 2013.
10. Han, J. and M. Kamber, M.DataMining:Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.,2001.
11. AshaRajkumar and Sophia Reena.G.Diagnosis of heart disease using data mining algorithm. Global J.Comp.Sci.Tech.,2010,38(10):38-43.
12. "Neural Network", Wikipedia, March 2013.
13. Stanfford, G.C.,Kelley,P.E.,Syka, J.E.P.,Reynolds,W.E and Todd,J.F.Recent improvements in and analytical applications of advanced ion-trap technology. Intl. J. Mass Spectrometry Ion Processes.,1984,60: 85-98.
14. Miller, A., Blott,B. and Hames, T. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp.,1992, 30: 449-464.
15. Weinstein, J., Kohn,K. and Grever,M.Neural computing in cancer drug development: Predic ting mechanism of action. Science., 1992, 258: 447-451.
16. Ephizibah,E.P. Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis.Int.J.on soft computing.,2011,2(1):1-10.
17. Baskar,S.S.,Arokiam,L.andCharles,S.Applying data mining techniques on soil fertility predictions. Int.J.Comp.Appl.Tech.Res., 2013,2(6):660-662.