Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning

L. Arockiam, S. Sathyapriya, V.A. Jane, A. Dalvin Vinoth Kumar

Abstract: ***Machine learning is a part of AI which develops algorithms to learn patterns and make decision form the massive data. Recently, Machine learning has been used to resolving various critical medical problems. Diabetes is one of the dangerous disease, which can lead to more complicated, including deaths if not timely treated. The study is designed for providing the prevalence of Diabetes Mellitus in Tiruchirappalli district using machine learning algorithms and it was detected that the polluted air causes diabetes disease and also increases the risk of that disease. This proposed work helps the people in preventing diabetes disease using various diabetic attributes with an aim to enhance the quality of healthcare and lessen the diagnoses cost of the disease. In future, the work done may be extended by considering many other attributes and by implementing it through various algorithms to improve the prediction accuracy of diabetes mellitus.***

Index Terms: Diabetes Mellitus, Machine Learning, Prediction, WEKA .

I. INTRODUCTION

Machine Learning plays an efficient role in medical especially diabetes research. Diabetes is a widely spreading disease in this modern society due to exercise gap, increased obesity rates, food habits and environment pollutants etc. Research on diabetes plays an important role in the field of medicine, and the number of daily data in this field is high. Continuous measurements are best suited for implementation of these data using data mining methods and can be handled immediately and these methods differ from other traditional methods and also one of the best ways in diabetes research when handle massive amounts of data related to diabetes. The main difference between them is more complicated than statistical approaches. Every day vast amount of data are stored in the various domains like finance, banking, hospital, etc. and rapidly increasing day by day. Such a Database may contain potential data that can be useful for decision making. Extraction of this valuable information manually from large volume of data is extremely difficult task. From the rapidly growing data, it is very hard to find useful knowledge without using ML techniques. Discovered knowledge can be useful in making prominent decisions. Data mining is widely used in fields such as business, medicine, science, engineering and so on [1-5].

Revised Manuscript Received on July 20, 2019.

Dr.L.Arockiam, Associate Professor. Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

S.Sathyapriya, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

V.A.Jane, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

A. Dalvin Vinoth Kumar, Assistant Professor, REVA University, Bangalore.

II. RELATED WORKS

Himansu Das et al., [6] proposed a framework for predicting diabetes mellitus. Diabetes Mellitus was predicted by classification algorithms such as j48, Naïve Bayes and these two were implemented using the weka tool. Questionnaire based data collection was done and data cleaning was performed to remove the unwanted data. The diabetes mellitus had been diagnosed by using j48 and Naïve Bayes. The final stage in the proposed framework generated the report of diabetes.

N.Vijayalakshmi and T.Jenifer [7] analysed risk factors of diabetes through data mining and statistical analysis techniques. The experiment for diabetes prediction was done by using classification algorithms, clustering, and subset of evaluation, association rule mining and statistics analysis. J48 provided better accuracy of 81% to the given dataset than the other techniques.

C.Kalaiselvi and G.M .Nasiria [8] predicted whether people with diabetes may have cancer and heart disease. Diabetes dataset was classified by using ANFIS and AGKNN algorithm and gained good accuracy level. The performance of algorithms was evaluated by using performance metrics. The proposed method reduces the complexity than the exiting methods.

Swaroopa shastri et al., [9] proposed a system to predict whether type 2 diabetes influences kidney disease. Here by the data mining algorithms were utilized. The proposed system generated the report of a patient, it assisted doctors, and also suggested precautions to the patient from kidney disease.

Huwan- chang et al., [10] developed a model for predicting postprandial blood glucose to undiagnosed diabetes cases in a cohort study. For this purpose, there were five data mining algorithms that were utilized and compared each other in this work. The data set used in this model was collected from Landseed Hospital in northern Taiwan over the period of 2006 to 2013 and also evaluated the performances of the data mining algorithms. The overall result of the proposed model provided the accurate reasoning and prediction; it could be useful to assist doctors to improve the skill of diagnosis and prognosis diseases.

Aiswarya Iyer et al., [11] utilized Decision Tree and Naïve Bayes algorithms for predicting diabetes in pregnant women. Training and test data was separated by 10 fold cross validation technique and J48 algorithm was employed on the Pima Indians Diabetes Database of "National Institute of Diabetes and Digestive and Kidney Diseases" using WEKA. The proposed work concluded that both algorithms were efficient for the diagnosis of diabetes and Naïve Bayes technique gave the result with least error rate.

A.A. Aljumah et al., [12] recommended a model based on regression technique for diabetes treatment. The proposed model predicted the diabetes disease by Oracle Data Miner tool and results were employed for experimental analysis on collected Datasets by support vector machine algorithm (SVM).

Mohammed et al., [13] presented a survey on application using Map Reduce programming framework which was discussed in early work and discussed Hadoop implementation in clinical big data related to healthcare fields.

N.M. Saravana Kumar et al., [14] proposed a Predictive Analysis System Architecture with various stages of data mining. Prediction approach carried out on Hadoop / Map Reduce environment. Predictive Pattern matching system was used to compare the threshold value analyzed with the estimated value after the analyzed reports were presented by the system.

III. METHODOLOGY

The proposed Model plays a significant role in predicting diabetic patients and produces the prevalence report of diabetes.

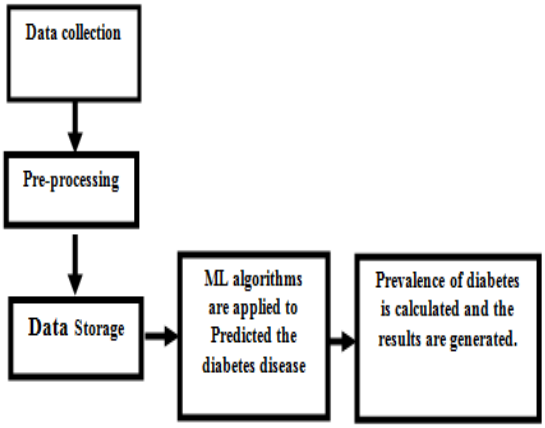


Fig 1: Work flow of proposed methodology

The work flow for diabetic prediction is shown in fig 1. In the initial step, the data collection is performed and it done through various ways such as questionnaire based data collection, sensor based data and some data from clinical report. Cloud storage is used where the electronic records are stored securely and cloud computing can be utilized for: data processing, data analysis and predictive analysis. These are carried out by statistical tools and data mining techniques. The predicative analytic stage sends the report of diabetes prevalence in Tiruchirappalli.

1. Data collection: It is one of the most initial steps in the proposed model and plays a major in data related research. In this paper there were following three types of data format collected from sensors, clinical and questionnaire.

2. Questionnaire: The data collected through questionnaire is called as the primary data. There were two types of data that were collected namely medical data and personal details. The questionnaire was prepared and given to various people who are living in Tiruchirappalli district. The question was developed using Google Form with 22 questions based on various factors such like gender, habits which spoils their health like smoking and alcohol drinking, food habit, BMI, medication taken by individual, blood pressure, family history , sleeping time, normal health problem , work type , educational background, environment pollutants and physical activity. Some of the questions were in yes/no format and some were in answer format. The model of the questionnaire sheet is given below in fig 2.

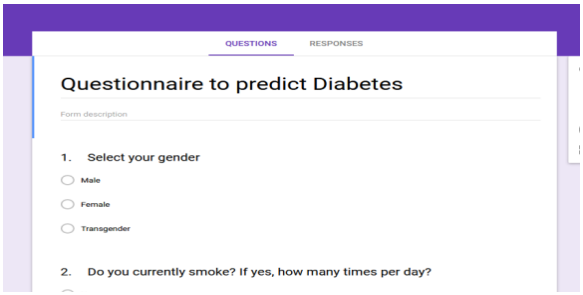


Fig 2: Questionnaire model based data collection

3. Sensor Data: Some data were collected by using sensor and also by using medical devices. In this thesis, Honeywell HPm Particle Sensor is used to find out the PM 2.5 and PM10 in the air and it is shown in fig 3. PM means particulate matter it used to find out the particles level in the air. PM 2.5 means particles with a size below 2.5 microns and PM10 includes particles with 10 microns and below. PM 2.5 is very serious than PM10 because PM2.5 contain very small particles it can travel to our lungs deeply and then causes more harmful effects. Further, it can lead to diabetes. In this paper particle matter is considered as a factor to predict the diabetes disease because air is an important factor for the people to survive in the world.

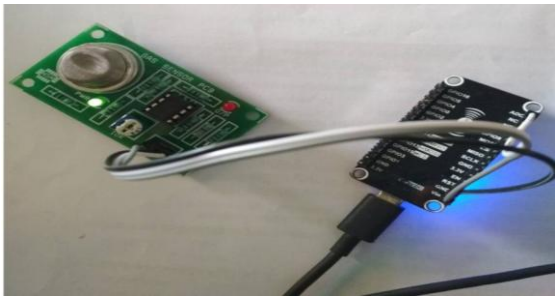


Fig 3: Data collection from sensor

4. Pre-Processing: Data Pre-processing is an important step during knowledge discovering. The collected data may contain missing, fault and outliers etc., Removal of these kinds of invalid data may produce misleading outcomes and makes knowledge discovery a challenge. Data is pre-processed by different ways such as cleaning, normalization, transformation, feature extraction and selection, etc. The major obstacle with clinical data is that redundant records and these records are eliminated to enhance the detection accuracy. Data transformation and data validation are two important pre-processing techniques.

5. Data Storage: The data stored in a cloud storage system with remote servers that accessible by internet and it managed, operated, and maintained by service provider. This proposed approach, the collected data are stored in ThingSpeak which is a cloud service provider. The flow of storage is showed in the fig 4.

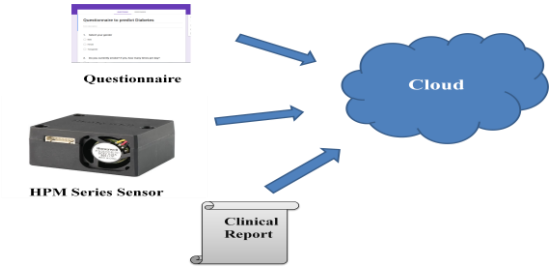


Fig 4: Collection of various data)

IV. PREDICTION OF DIABETES

The study made on various classification algorithms used in existing methods, three algorithms play major role in predicting Diabetes mellitus. They are J48, KNN, and Naïve Bayes. The PIMA Indian Dataset was applied to these 3 algorithms in which J48 algorithm predicts results with better accuracy [15]. So in this study J48 is used and the collected data is applied in WEKA to classify Diabetes Mellitus based on different attributes like age, sex, income, education, work type, blood pressure (diastolic and systolic), body mass index (BMI), dietary history, physical activity, pattern and Pm (Pm2.5& Pm10). The outcome of predicting Diabetes Mellitus is represented as a class variable 1 or 0, depending on whether the person has diabetes or not respectively.

The nature of the collected data has described in this section. The overall male and female from the total study population has been separated based on their age with a percentage of the population and it is listed below in the table 1.

Table 1: Distribution of population based on their age and sex

Age	No. Male Population (%)	No. Female Population (%)	Total Population (%)
< 30 years	42(59.15 %)	29(40.84%)	71 (5.81)
30- 35 years	31 (58.49%)	22(41.50)	53(4.34)
36- 40 years	172(64.66)	94 (35.33)	266(21.78)
41- 50 years	612 (48.84)	310 (51.15)	606 (49.63)
51- 60 years	118(68.20)	55 (31.79)	173 (14.16)
>60 years	21(40.38)	31(59.61)	52(4.25)

A. **Family and Income:** From the study of population, people are separated based on their family and income. They were grouped into four categories based on their income style such as below 50,000, 50,000 to 1,50,000, 1,50,000 to 2,00,000 and above 2,00,000. According to these categories, people were separated like diabetic and non-diabetic and tabulated as shown in table 2.

Table 2: population separated based their monthly income

Income	Total	Percentage of total (%)
Below 50,000	341	27.92
50,000 to 1,50,000	662	54.21
1,50,000 to 2,00,000	161	13.18
above 2,00,000	57	4.66

B. Education: In Tiruchirappalli district, people are living with various education levels, such as school, college, and illiterate. These survey details are given in the fig 5.

Fig 5: Education level based division

C. Work Type: According to the physical work of individuals, the work is categorized as easy, medium, and hard and based on their work type the details about diabetic patients were represented in the fig 6.

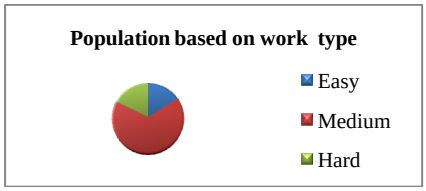


Fig 6: Population divided by work type.

D. Awareness of Diabetes Test: People who have diabetes are certainly aware of the disease and also will be aware of the precautions to be taken. The evaluation of awareness among people is depicted as a graph in fig 7.

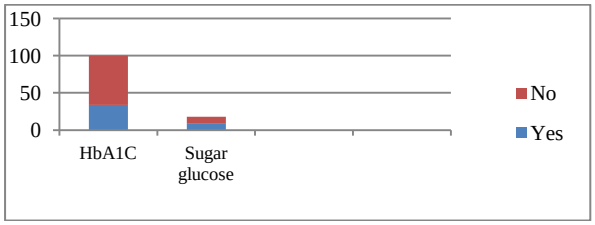


Fig 7: Awareness about Diabetes Mellitus

Furthermore, sugar count helps to find out the sugar level of an individual, suppose if a person has a sugar count below 140 then it is known as low sugar level, or if the sugar count is above 140 to 180 then the sugar level is normal, which is also called as pre-diabetic but if the sugar count exceed above 180 then the count is high. The surveyed result is shown in Table3.

Table 3: Sugar level based on the sugar test.

	low sugar	pre-diabetes	high sugar
below 140	37.2		
140 - 180		42.6	
above 180			20.2

E. Blood Pressure and Work Type: Blood pressure varies based on the people's work type. There are three categories of works such as easy, medium and hard. The pressure level is also divided into high, medium and normal. Figure 8 depicts the list of people who have blood pressure, which is separated based on easy, medium and hard type of work.

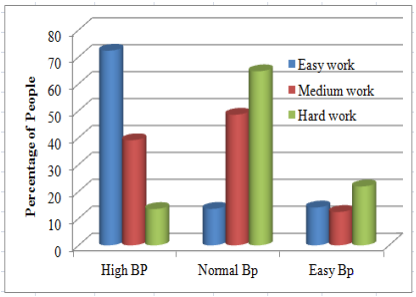


Fig 8: Work type vs Blood pressure level

F. Smoking and Liquor Drinking Habits: People, who are smoking, consuming alcohol, both smoking & consuming alcohol are 314, 193 and 178 respectively

Table4: List of data related with smoking and drinking.

7. Air Quality: Air quality is as an important factor in this study because it also one of the reason for diabetes mellitus. The air quality level is measured through the $PM_{2.5}$ and PM_{10} level in the air and fixed into the area to evaluate the particle level. From this the PM level is measured and separated among diabetes people that showed in table 5.

Table 5: Air quality and Diabetes

Air Quality	Diabetic	Non-Diabetic
High	68	36
Medium	15	47
Low	17	17

V. CONCLUSION

In Machine Learning data patterns are extracted by applying intelligent methods. These methods provided the great opportunities to assist physicians deal with this large amount of data. This study provided a view about the prevalence of diabetes mellitus using classification techniques. It helps the patients to prevent themselves from the disease. Decision tree model has outperformed than naïve Bayes and KNN techniques. The proposed work detected that the polluted air causes the diabetes and also increases the risk of diabetes. The proposed work can be further enhanced and expanded with stacking techniques to increase the accuracy of prediction..

ACKNOWLEDGMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).

REFERENCES

1. Arun K Pujari, "Data Mining Techniques", Universities Press (India) Private Limited2001
2. Krochmal, Magdalena, and Holger Husi, "Knowledge discovery and data mining" Integration of Omics Approaches and Systems Biology for Clinical Applications , 2018, pp. 233-247.
3. Qi Luo. "Advancing Knowledge Discovery and Data Mining", IEEE Workshop on Knowledge Discovery and Data Mining,2008.
4. S.D.Gheware, A.S.K. ejkar,S.M.Tondare,"Data mining: Task, Tools techniques and applications", International Journal of Advanced Research in Computer and Communication Engineering, Vol.3, Issue.10,2014, pp. 8095 -8098.
5. Krishnaiah, V. Narsimha, G. and Subhash Chandra, N. "A Study on Clinical Prediction using Data Mining Techniques", International Journal ofComputer Science Engineering and Information Technology Research (IJCSEITR), Vol.3, Issue.1, 2013, pp.239-248.
6. Himansu Das, Bighnaraj Naik and H. S. Behera,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", Springer Nature Singapore Pte Ltd,2018,pp:539-549.
7. Miss. N. Vijayalakshmi, Miss. T. Jenifer,"An Analysis of Risk Factors for Diabetes Using Data Mining Approach",International Journal of Computer Science and Mobile Computing, Vol.6, Issue.7, 2017, pp:166 – 172.
8. Kalaiselvi, C., and G. M. Nasira. "Prediction of heart diseases and cancer in diabetic patients using data mining techniques." Indian Journal of Science and Technology ,Vol.8, Issue. 14 ,2015.
9. Swaroopa Shastri, Surekha, Sarita, " Data Mining Techniques to Predict Diabetes Influenced Kidney Disease", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol.2, Issue. 4, 2017, pp. 364-368.
10. Chang, Huan-Cheng, Pin-Hsiang Chang, Sung-Chin Tseng, Chi-Chang Chang, and Yen-Chiao Lu. "A comparative analysis of data mining techniques for prediction of postprandial blood glucose: A cohort study." International Journal of Management, Economics and Social Sciences (IJMESS) , Vol.7 ,2018, pp. 132-141.
11. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, Issue.1, 2015, pp. 1-14.
12. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes healthcare in young and old patients", Journal of King Saud University - Computer and Information Sciences, 2013, Vol.25, pp. 127-136.
13. Emad A Mohammed, Behrouz H Far and Christopher Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends", BioData Mining 2014, Vol.7, pp.1-23, http://www.biodatamining.org/content/7/1/22
14. Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S, " Predictive Methodology for Diabetic Data Analysis in Big Data", Procedia Computer Science 50, 2015, pp. 203 - 208, Available online at www.sciencedirect.com.
15. Dr. L. Arockiam, A. Dalvin Vinoth Kumar, S. Sathyapriya, "Performance Analysis of classification Algorithms for Diabetic Prediction Using Pima- Indian dataset", Journal of Emerging Technologies and Innovative Research (JETIR), Vol. 5, No.12, 2018, pp.563-569.

AUTHORS PROFILE

First Author Dr. L. Arockiam is working as Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Thiruchirapalli, Tamil Nadu, India. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.

Second Author S.Sathyapriya is doing her Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. Her research area is IoT Data Analytics.

Third Author V. A. Jane is doing his Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. His research area is IoT Data Analytics.

Fourth Author Dr. A. DalvinVinoth Kumar is working as Assistant Professor in the Department of Computer Science, Kristu Jayanthi College Bengaluru, Karnataka, India. His research interests are: MANET, Routing and IoT

Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India

Dr. L. Arockiam, S. Sathyapriya, V.A. Jane, A. Dalvin Vinoth Kumar

Abstract: ***Diabetes mellitus is a non-communicable disease, however it may lead to other health problems such as blood pressure, heart attack, vision problem, slow healing sores to patients with arthritis etc. Diabetes disease is caused due to lifestyle, food habits, and low level of fabrication of insulin and pedigree factors of individual. According to the study, there will be 552 million people around the world will be affected by diabetes at 2030. This paper estimates the total populations of type 2 diabetes patients in the central region (Cuddalore, Thanjavur, Perambalur, Tiruchirappalli, Ariyalur, Karur, Nagapattinam, Thiruvarur, Pudukottai, and Karaikal) of Tamil Nadu. Diabetes patients have been diagnosed with the help of various parameters such as blood pressure, body mass index, dietary history, physical activity and pollution level in the air. The Honeywell HPm series particle sensor is used to access the PM 2.5, PM 10 levels in the air. Considering the air quality as a parameter, there are lots of illnesses caused by air pollutants and also cause additional problems for people who are already suffering due to disease. This review work provides the knowledge about the prevalence of type-2diabetes and it will help people to take precautions about diabetes disease and its risk.***

Index Terms: Diabetes, Air Quality, Sensor, PM2.5, PM10.

I. INTRODUCTION

Diabetes mellitus is one type of non-communicable disease. The prevalence of diabetes is rapidly increasing all over the world at a tremendous rate [1].It occurs when the glucose level increases in the blood. Blood glucose is the main source which produces energy to human body. The high blood sugaris defined as a medical syndrome, which is also called as hyperglycemia, which is caused due to an inadequacy of insulin in the human body. The level of blood sugar is standardized by a hormone, which is done by the insulin generated by the pancreas. The pancreas is a very tiny organ which is placed between the stomach and liver that helps to digest the food. According to the report of World Health Organization (WHO)[2], the highest number of diabetes affected people are living in India. The total number of diabetes patients in the year 2016 is 7.8 million it will exceed 79.4 million by 2030. The International Diabetes Federation (IDF)[3] in the world has reported on diabetes that it has proved 425 million adults living with diabetes. According to the report of IDF, 5.2 % of Indian people are not aware that they are suffering from high blood sugar. In specific, the Madras diabetes research [4] foundation instructed that about 42 lakhs individuals are suffering from diabetes and 30 lakh people are in pre-diabetes.

A. *Types of diabetes disease:*

There are various ways to detect the presence of diabetes in the human body. There are three categories in diabetes mellitus. They are Type-1 diabetes, Type-2 diabetes and Gestational diabetes[5]. The early stage of diabetes is identified using the following factors such as long-lasting blood sugar, blood sugar fasting, diabetes history of genes, measuring waist and the ratio of height waist of individuals. In this paper type 2 diabetes is considered.

a. Type 2 Diabetes

Type 2 diabetes is called as non-insulin dependent diabetes[6]. In type 2 diabetes, pancreas produces sufficient insulin but the beta cells do not use it properly and that's why insulin resistance is caused. In such case, insulin tries to get glucose into the cell but it can't maintain instead of this the sugar level may increase in the blood. People may get affected by the type 2 diabetes at any age even in childhood. Type 2 [7] diabetes is caused by overweight and inactivity which leads to insulin deficiency. These types of diabetes can be controlled by weight management, regular exercise and nutrition. The symptoms of type2 diabetes are same as type 1diabetes except itching skin and the problem in vision. This type of diabetescan'tbe cured but can be controlled by medicine and injection which is given for diabetes, physical exercise, blood monitoring and glucose controlling.

B. *PubMed NCBI*

Over the past few years, awareness about diabetes is growing and the possibility also growing in this field. According to PubMed NCBI, referred as a journal for publishing MEDLINE papers, indexed by PubMed has computed diabetes related details which are surveyed from the year of 1983 and 2018 by using the keyword "Prediction and Diabetes". The surveyed results are shown in the form of graph, which is displayed in Fig1. The count for 2018 is extrapolated till June 27, 2018.

Revised Manuscript Received on July 20, 2019.

Dr.L.Arockiam, Associate Professor. Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

S.Sathyapriya, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

V.A.Jane, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

A. Dalvin Vinoth Kumar, Assistant Professor, REVA University, Bangalore.

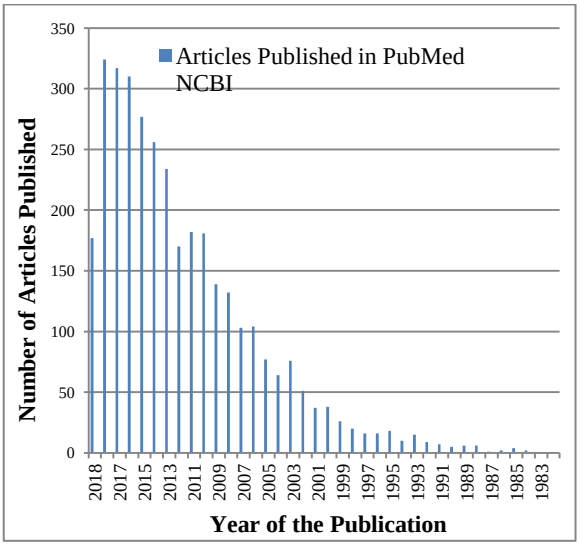


Fig 1: Number of publications index by Pub Med with keywords "Predictive and Diabetes." The 2018 count is extrapolated based on the number from June 27, 2018

II. REVIEW OF LITERATURE

Cheng lin et al [8]discussed about the classification and prediction in data mining by analyzing the information based on diabetes data. This paper partitioned the sets of data for classifying by using decision tree and data prediction was done through the linear regression, multiple regressions and non-linear regression whereas evaluated the classification accuracy. The process of classification and prediction of data mining also discussed about similarities and differences between them.

K.Lakshmi et al [9] proposed a System Architecture for diagnoses diabetes disease using clustering and classification algorithm such as decision tree and KNN. The proposed system has stored data into a server which was collected based on different diseases of patients. Here, they considered 11 attributes of diabetessuch as Age (years), Sex, Body mass index, Blood Pressure (mm Hg), Plasma Glucose Concentration (Glucose tolerance test) Triceps Skin fold 2-Hour serum insulin Diabetes Pedigree function, Cholesterol Level, Weight (kg) and Class variable (0 or 1) to predict the diabetes.The proposed method consists of some basic components such as admin, user (doctor, patient, physician etc), server, database, application, and data mining techniques. In the first step of the proposed system, KNN and Decision tree were applied for training the dataset after receiving the request from the user, which are like a supervised classification model.Admin received the inputs from requestor. In the final step DM approach was used to predict the result and send back to the user. Time and cost are reduced to diagnoses in this approach.

Dr.Prof.Neeraj et al [10] described the J48 algorithm for predicting recurrence of cancer-based data set to breast cancer. Recurrent cancer can be analyzed in three ways and they are: cancer comes back after treatment or it is in the same place, where it started first whether in any portion of the body. Hereafter J48 algorithm was used on the data set of breast cancer and implemented by WEKA tool and generated the decision tree by using 10 fold cross-validation method to predict the recurrent event due to its attributes such as tumor size, the degree of malignancy, age, node-caps, menopause etc.UCI machine learning repository provided the data set for predicting recurrence cancer of undergone treatment to patients. A result of experiment was tabulated and the decision tree was shown in the figure. Furthermore, results were concluded accurately and specific range value was used to find out the changes of recurrence cancer.

Manal Abdullah et al [11] proposed a method for finding five types of anemia is one of the hematological diseases and predicted what type of anemia hold by patient using classification algorithms. This paper proposed an algorithm for classification with the help of complete blood count test. The data sets were collected from patients and were filtered. Multiple experiments were conducted using various algorithms namely naive Bayes, neural network, J48 decision tree, and SVM. Compared with other algorithms J48 decision tree provided the best potential classification of anemia types. J48 decision tree algorithm provided better performance with accuracy, recall, true positive rate, false positive rate, precision and F-measure and it was proved by weka experiment. The tested results were tabulated in percentage (like 20%, 40%,60%). The anemia types can be detected with the help of given algorithms but this paper concentrated only on five types of anemia for finding accuracy and prediction of preferred results.

Himansu Das et al [12] focused on Diabetes Mellitus Disease. They used two data mining technique such as J48 and Navie Bayesian for predicting diabetes. The proposed technique was quicker and efficient for diagnosis the disease. The dataset was collected from medical college hospital by providing set of questions that about particular patient name, age, sex, blood, sugar level, and plasma glucose and as well as online repository. After thatthe data cleaning was performed to remove the unnecessary data and was stored in the warehouse. The proposed method predicted whether the patient has diabetes or not, by classification technique. The two classification techniques were implemented through WEKA software and the experimental results were tabulated. Navie Bayes better than J48 and also the outcome was proved by its productivity.

N.Vijayalakshmi and T.Jenifer [13] worked on data mining and statistical analysis for identifying diabetes disease. The data source contained pertaining diabetes which has taken from nursing home research center. The collected data divided as diabetic patients and non-diabetic patients. WEKA tool was used for analyzing the most important factors causing diabetics and also used to perform statistical analysis method on every single attribute. Tow classification techniques such as J48 pruned tree technique and the Random tree provided the validation result and the detailed accuracy on datasets by class. Hence this paper proved J48 pruned tree is a better technique compared with other classifying techniques and the accuracy of the predicted result was 81%.

III. SURVEY AREA

Tamil Nadu is one of the states in India. Based on the direction of the districts located, it is divided into 4 Regions namely central region, western region, southern region and Chennai city region. Each region has at least more than 4 districts. The central region has 10 districts such as Cuddalore, Thanjavur, Perambalur, Tiruchirappalli, Ariyalur, Karur,

Nagapattinam, Thiruvarur, Pudukottai, and Karaikal. The western region has 6 districts which are Coimbatore, Erode, Nammakal, Salem, Dharmapuri and the Nilgiris. The southern region has 9 districts that are Dindigul, Madurai, Theni, Sivaganga, Virudunagar, Ramanathapuram, Tirunelveli, Thoothukudi and Kanyakumari. Finally, Chennai, Thiruvalluvar, Kancheepuram, Vellore, Tiruvannamalai, and Puducherrydistricts have come under the Chennai city region.

A. *Central region*

According to the census report at 2011, the Central region's total population is 12,212,084 where the men and women are in the frames of 7,031,520 and 7,194,867 . The total taluk in the central region of all districts are 66 whereas total revenue villages and panchayat villages are 4638 and 3154 respectively. From the report, the total number of literate people in that region is 7,369,787. Men and women in this category are 3,982,437 and 3,432,656. The total number of children (age between 0-6) in this region is 1,042,373, from this total number of male children and female children are 3,982,437 and 3,432,656.

IV. MATERIALS AND METHODS

All the study samples were randomly collected from states in the central region of Tamilnadu. The total study population is 10115 among them 5566 were male and 4549 femlae which is 55.1% and 44.9% respectively. The population was screened for blood pressure (diastolic and systolic) and blood sugar along with their screening data, the body mass index (BMI), dietary history, physical activity, pattern and Pm2.5 (pm & Pm10). The population screened for diabetics by random Blood Sugar Meter(RBS). The Blood pressure is screened using Arm Bp digital monitor. The dietary history, physical activity are assessed by a set of stored questions. The air pollutants (Pm2.5 & Pm10) are assured using Honeywell HPm series particle sensor. The number of total study population for male and female percentage has separated based on their age wise and listed in the table 1.

Table 1: Age and sex wise distribution of the study population

Age	No. Male Population (%)	No. Female Population (%)	Total Population (%)
< 30 years	1422 (72.9)	526 (27.1)	1948 (100)
30-35 years	1658 (60.7)	1071 (39.3)	2729 (100)
36-40 years	1427 (69.3)	632 (30.7)	2059 (100)
41-50 years	612 (36.25)	1076 (63.75)	1688 (100)
51-60 years	376 (23.7)	1213 (76.3)	1589 (100)
>60 years	71 (69.7)	31 (30.3)	102 (100)

V. RESULTS AND DISCUSSION

According to the report of total study population, people have separated based on their age and sex. From this, the total number of male and female has displayed in Fig2 in the form of graph. The age of both gender classified as, Below 30, 30 to 40, 41 to 50, 51to 60 andAbove 60.

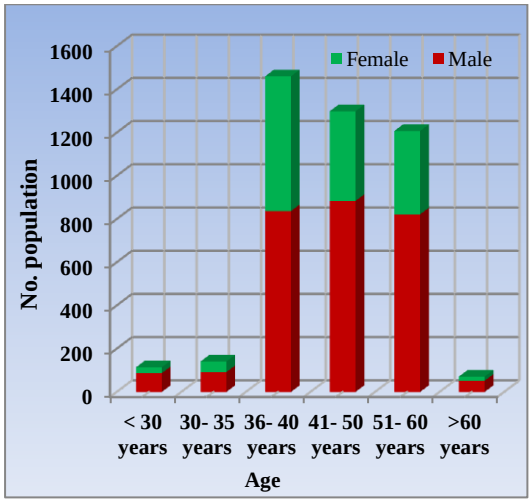


Fig 2: Distribution of Age and Sex

A. *Diabetic and Age*

Among the major factors of diabetes, age is considered like one kind of major factor. The total number of diabetes patients derived from total study population has given in the graph with its percentage. Fig3 represents the above mentioned details as a graph.

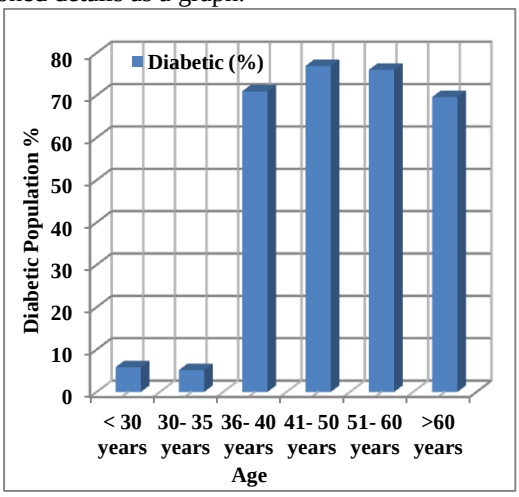


Fig 3: Diabetic and Age

B. *Education and Diabetic*

Diabetes awareness between literate and illiterate were surveyed. Totally 36.50 % percentage of illiterate people has lived in Tamilnadu, 41% percentage of people completed their schooling,22.50% percentage completed graduation. The comparison is between these categories of people represented in the form of graph in Fig4.

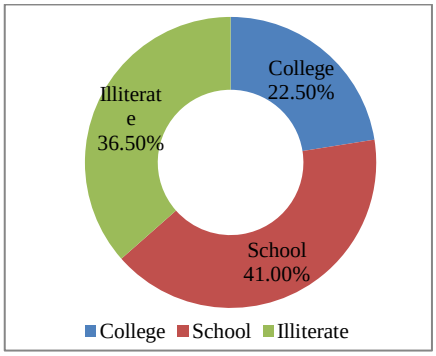


Fig 4: Education and Diabetic

C. *Air Pollution and Diabetic*

When the pancreas function decreases, the function of insulin is reduced. Diabetes occurs when the pancreas does not produce sufficient insulin. Today Air pollution is increasing throughout the world, and the air is most often polluted by the urban area so air pollution may affect the pancreas as well as the **livelihood may be affected to diabetic patients**. Here using Honeywell HPm series particle sensor, the air pollution(Pm 2.5 and Pm10) detail was collected and displayed. An average of air pollutants level is given as a graph in fig 5.

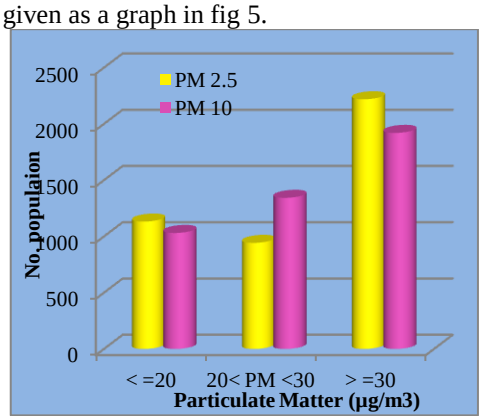


Fig 5: Air Pollution and Diabetic

D. *Diabetic Control measure*

According to this study, diabetes people have followed insulin or treatment taken from required government hospital or have followed any diet to control their diabetes or awareness about HCA1C test and carbohydrate count. In order to the study of total population has described and the number of population based on diabetic control measures which is represented in fig 6 as a graph.

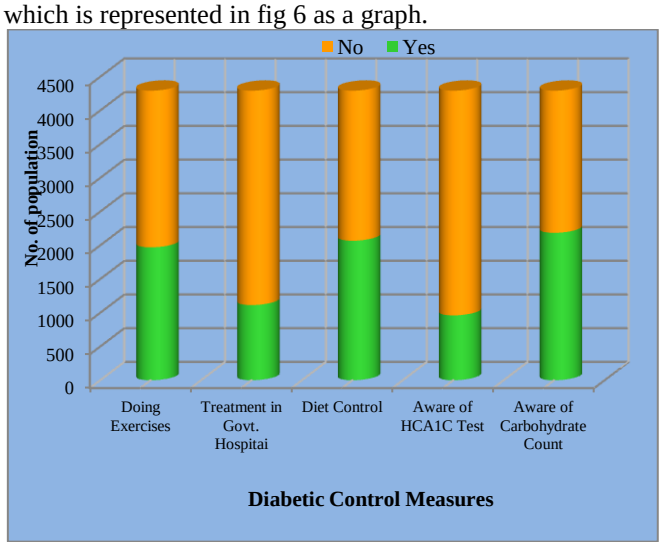


Fig 6: Diabetic Control measure

VI. CONCLUSION

Diabetes Mellitus is a chronic disease that can affect human life. Massive data was collected from census report at 2011, questionnaire and IoT devices. Tamilnadu has been separated into four regions with respect to the location. This study concentrated mainly on the central region of Tamilnadu and total population was surveyed in that region. From this, total number of male and female population was also reviewed. The number of people living with diabetes in the central region was calculated by using various parameters. The results of experiments exhibited the number of diagnosis made for diabetic patient and were computed individually on the basis of their age, education, physical activity, dietary history, and air pollution. This paper will help to spread the awareness about diabetes among people. In future, these experiments may be conducted all over Tamilnadu and it may improve the accuracy level with the help of various parameters

ACKNOWLEDGMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).Authors thankful to Dr. Ravi MD, CEO,MHealth,Trichy Tamil Nadu India and MrPeriyasamy, Meyer pharmaceuticals for helping us in data collection and also thankful to all the patients who have participated in this survey.

REFERENCES

1. P. Agrawal and A. Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus," Int. Res. J. of Eng. and Tech. IRJET, Vol. 02, pp. 1039-1043, June-2015.
2. NDTV Food Desk, Updated: November 14, 2017 13:06 IST, available at: https://www.ndtv.com/food/world-diabetes-day-2017-number-of-diabetics-to-double-in-india-by-2023-1775180.
3. World Health Organization. "Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus", 1999, pp. 17-21.
4. American Diabetes Association. "Diagnosis and classification of diabetes mellitus", Vol.37, Issue.1, 2014, pp : 81-90.
5. Olson, Brooke. "Applying medical anthropology:Developing diabetes education and prevention programs in American Indian cultures", American Indian Culture and Research Journal, Vol.23, Issue.3, 1999, pp: 185-203.
6. Alberti, G., Zimmet, P., Shaw, J., Bloomgarden, Z., Kaufman, F. Silink, M. "Type 2 diabetes in the young: the evolving epidemic Diabetes care", Vol.27, Issue.7,2004, pp: 1798-1811.
7. Recognising Type-2 Diabetes', Feb 2016 Available: https://www.healthline.com/ health/type-2 -diabetes /recognizing-symptoms, [Accesed : 15-jan-2018].
8. Lin, C., & Yan, F. "The study on classification and prediction for data mining" In Measuring Technology and Mechatronics Automation (ICMTMA), 2015 Seventh International Conference ,2015,pp. 1305-1309.
9. Dr Prof. NeerajBhargava, N., Sharma, S., Purohit, R., &Rathore, P. S. , "Prediction of recurrence cancer using J48 algorithm" Proceedings of the 2nd International Conference on Communication and Electronics Systems, 2017,pp:386-390.
10. K. Lakshmi,D.Iyajaz Ahmed & G. Siva Kumar, " A Smart Clinical Decision Support System to Predict diabetes Disease Using Classification Techniques" IJSRSET,vol.4,Issue.1,2018,pp: 1520-1522.
11. Abdullah, M., & Al-Asmari, S.," Anemia types prediction based on data mining classification algorithms", Communication, Management and Information Technology–Sampaio de Alencar (Ed.),2017,pp:615-621

12. Himansu Das, BighnarajNaik and H. S. Behera,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", Springer Nature Singapore Pte Ltd,2018,pp:539-549.
13. Miss. N. Vijayalakshmi, Miss. T. Jenifer,"An Analysis of Risk Factors for Diabetes Using Data Mining Approach",International Journal of Computer Science and Mobile Computing, Vol.6, Issue.7, 2017, pp:166 – 172.

AUTHORS PROFILE

First Author Dr. L. Arockiam is working as Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Thiruchirapalli, Tamil Nadu, India. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.

Second Author S.Sathyapriya is doing her Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. Her research area is IoT Data Analytics.

Third Author V. A. Jane is doing his Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. His research area is IoT Data Analytics.

Fourth Author Dr. A. DalvinVinoth Kumar is working as Assistant Professor in the Department of Computer Science, Kristu Jayanthi College Bengaluru, Karnataka, India. His research interests are: MANET, Routing and IoT

PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DIABETIC PREDICTION USING PIMA- INDIAN DATASET

Dr. L. Arockiam[1], A. Dalvin Vinoth Kumar[2], S. Sathyapriya[3],
Associate Professor[1], Assistant Professor[2], M.Phil. Scholar[3]
Department of Computer Science[1,3], School Of CSA[2]
St. Joseph's College (Autonomous), Tiruchirapalli, India[1,3] , REVA University, Bangalore, India[2] .

Abstract : Data Mining is a process of collecting, extracting the data from various data warehouse and summarizing the data as a useful one. Essentially, data mining is referred to as "Knowledge Discovery from Data" (KDD) that is an extraction of knowledge automatically or in a convenient way . The predictive analytics is one such branch of advanced analytics in data mining, which is used to predict the future events. The predictive analytics uses different techniques such as machine learning, statistics, data mining and AI for analyzing massive data. This paper provides a review on predictive analytics and elaborates the predictive techniques with their application. Using the Pima Indian diabetes dataset, the prediction for diabetes is done with the help of various classification algorithms such as J48, Naïve Bayes and KNN. The accuracy of the algorithms is compared and found that J48 algorithm gives the highest accuracy.

IndexTerms **- Classification, Data mining, J48, KNN, Naïve Baiyes, WEKA.**

I. INTRODUCTION

Data mining is the process of discovering the knowledge from various data sets. which also defined as "Knowledge Discovery in data bases". Data mining techniques are applied on large volume of data for finding hidden models and relationship among the patterns which are helpful in decision making [1].Data mining techniques can be classified as descriptive model and predictive model [3]. The descriptive model represents the data in a brief form and applied to discover patterns in data and to analyze the relationship between attributes. The descriptive techniques include association mining, sequence discovery, clustering and summarization. The predictive model operates by predicting the values of unknown data from the known results. This includes regression, classification, analysis and prediction. The figure 1 depicts the Data mining techniques.

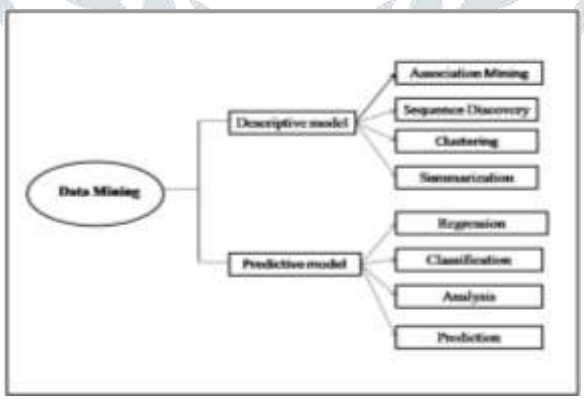


Figure 1: Various techniques of data mining

II. EXPLANATION OF DATA MINING TECHNIQUES

Classification

Classification categorizes data into one of the predefined classes. It has two processes. First one examines the objects and builds a model using training data which describes predetermined set of data classes. Secondly , the objects are assigned to a predefined class and classification techniques are applied. It mostly uses classification techniques such as Bayesian classifiers, Support Vector Machines, K-Nearest Neighbor, decision trees and neural networks.

Regression

Regression is the oldest and most popular statistical technique used for numeric prediction. This is used to map a data item to a real valued prediction variable. Regression analysis is used to identify the relationship between independent variables and dependent variables.

Time Series Analysis

Time series analysis encompasses methods and techniques for analyzing time series data in order to extract meaningful statistics. The values usually are obtained at uniform time intervals (hourly, daily, weekly, etc.).

Prediction

Prediction is used to predict future data which are relevant to past and current data. A few applications of prediction include speech recognition, machine learning, and pattern recognition.

Clustering

Grouping of objects is where similar objects exist in the same cluster and dissimilar objects exist in different clusters is called Clustering. It is also known as unsupervised classification. The similarity is calculated using Euclidian distance. The different types of clustering include, Hierarchical clustering, Partition clustering, Categorical clustering, Density based clustering and Grid based clustering.

Summarization

Summarization is also called as characterization or generalization. It summarizes a subset of data. The information about the database is collected by retrieving portions of the data. The resulting information is a set of aggregate information.

Association Rule Mining

Exploration of association rules between attributes in a transactional database is done and the association rules are used to find the frequency of items occurring together. These extracted rules are defined based on user defined minimum support value with minimum confidence value. This enables effective decision making. The algorithms used for association mining are Apriori algorithm, FPgrowth algorithm, Partition algorithm, Pincer-search algorithm and Dynamic Itemset Counting algorithm.

Sequence Discovery

Sequential discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. The patterns identified are most likely to have similar data and the relationship is based on time.

III. RELATED WORKS

Jan Andrzej Napieralski [7] discussed the different algorithms for predicting the items given in database. The process of prediction was accomplished by using various statistical methods. In addition to that, the appropriate preprocessing methods were implemented.Later the statistical method was applied.At the end of the experiment, probability of data prediction was calculated by using logistic regression and R programming language. Furthermore, additional mathematical methods were used while pre-processing the data on the dataset for better performance.

Satr et al [8] presented the review of decision tree data mining algorithms such as CART and C4.5. This paper provided the comparative study of both CART and C4.5 algorithm. Commonly decision tree algorithm can be used to predict the target value of its inputs. From the experiment, it is proved that the C4.5 algorithm is better than the CART algorithm.

Pragati et al [9] focused in the diagnosis of diabetes Mellitus using data mining techniques and analyzed k-fold cross validation, classification method, class wise K- Nearest Neighbor[CKNN], Support Vector Machine [SVM], LDA Support Vector Machine and Feed Forward Neural Network, Artificial Neural Network, Statistical Normalization and Back propagation methods for diabetic diagnosis. And presented that, SVM provided better accuracy on diabetic dataset.

Priyanka Chandrasekar et al [10] presented the method for improving the accuracy of decision tree mining with preprocessing data. Preprocessing method presented the benefits of classification accuracy performance tests. In this paper, the supervised filter discretization was applied with J48 algorithm. The process of proposed model classified the data by both

training and tested dataset. Classification accuracy was improved by entropy-based discretization method. Finally, the performance of this approach was compared with the J48 algorithm.

SwaroopaShastri et al [11] discussed about the type 2 diabetes disease and predicted using data mining algorithms whether diabetic patients had the diabetic kidney disease (DKD). The DKD patients information was collected who were affected by diabetic and prediction was done based on given attributes. The AES algorithm and Apriori algorithm were used for correlating and mining the set of items from the database. It established the correlation between diabetes and kidney disease patients. It helped the doctors to suggest the best medicine to the patient.

Various Data mining techniques, tools and data sets surveyed are presented in table 1.

Table 1: comparison between classification algorithms with various applications

Techniques & Algorithm	**Dataset & Tool**	**Parameter**	**Application domain**
J48 , LAD tree , MCC,MAE,RAE,RMSE,RRSEC	NBSS & LUB dataset (National Bureau of Soil Survey and Land use Planning. WEKA Tool	Accuracy, sensitivity, specificity	Agriculture [12]
J48 , SMO , Naïve Bayes , Multi Layer Perception	Complete B Blood count data set WEKA tool	Performance Accuracy Precision, Recall, True Positive rate, false Positive rate, F- measure	Medical [13]
K Nearest Neighbor, Decision Tree	Diabetic Dataset	Time Reduction, reduced cost, Accuracy	Medical [14]
J48, Naïve Bayesian	Diabetic dataset from medical college hospital. WEKA tool	Accuracy , Productivity	Medical [15]
J48, K-Means, Clustering, Decision Tree, Classification algorithms	Diabetes data set. WEKA TOOL	Accuracy	Medical [16]
J48, Deceision Tree, Multilayer perception, Naïve Bayes, Sequential Minimal Optimization	Turkey student evaluation records	Accuracy	Education [17]
PSO (Particle Swarm Optimization Algorithm), ANFIS (Adaptive Neuro Fuzzy Inference System), AGKNN (Adaptive Group Based K-Nearest Neighbor)	Diabetic Patients Record Dataset. MATLAB	Performance, Accuracy, Efficiency, Reduce Complexity	Medical [18]
Logistic Regression, Naïve Bayes	Heart patient dataset	Accuracy	Medical [19]
J48, Decision Tree algorithm , Meta – Technique	Private soil testing laboratory in pure(India)	Accuracy	Agriculture [20]
Artificial Neural Network,	Climate & Soil Dataset	Efficient	Agriculture

Back Propagation Training Method , Feed Forward Algorithm			[21]

IV. DATA DESCRIPTION AND ANALYSIS OF PIMA DATA SET IN WEKA

The data set collection is one of the important processes in data mining. The most relevant data is chosen from a particular domain for further analysis. The derived values can be more flexible and informative in that domain. In this study, PIMA Indian diabetic data set was used and it having nine attributes which are considered to predict diabetes. These sets of data obtained from UCI repository and the data set contains the basic knowledge about individuals such as age, BMI, BP and pregnant ladies, etc. The above mentioned attributes are numeric values with continuous data type. Totally, the data set having 768 instances, 9 attributes that have shown in table 2.

Table 2: Dataset Description

S.No	Attribute	Description	Maximum level
1.	Pregnancies	Total number of pregnant times	17
2.	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	199
3.	BloodPressure	Diastolic blood pressure (mm Hg	122
4.	SkinThickness	Triceps skin fold thickness (mm)	99
5.	Insulin	2-Hour serum insulin (mu U/ml)	846
6.	BMI	Body mass index (weight in kg/(height in m)^2)	67.1
7.	DiabetesPedigree Function	Diabetes pedigree function	2.42
8.	Age	Age (years)	81
9.	Outcome	Class variable (0 or 1)	-

In order to use WEKA tool, the data set available in .csv format was converted to .arff file format. Weka 3.6.13 is the latest version which is used in this study [26]. Weka consists of many machine learning algorithms which are capable to solve problems of data mining and machine learning. The converted data set from .csv is applied in weka for classification. The overview of weka environment after applied data set in it is shown in figure 2.

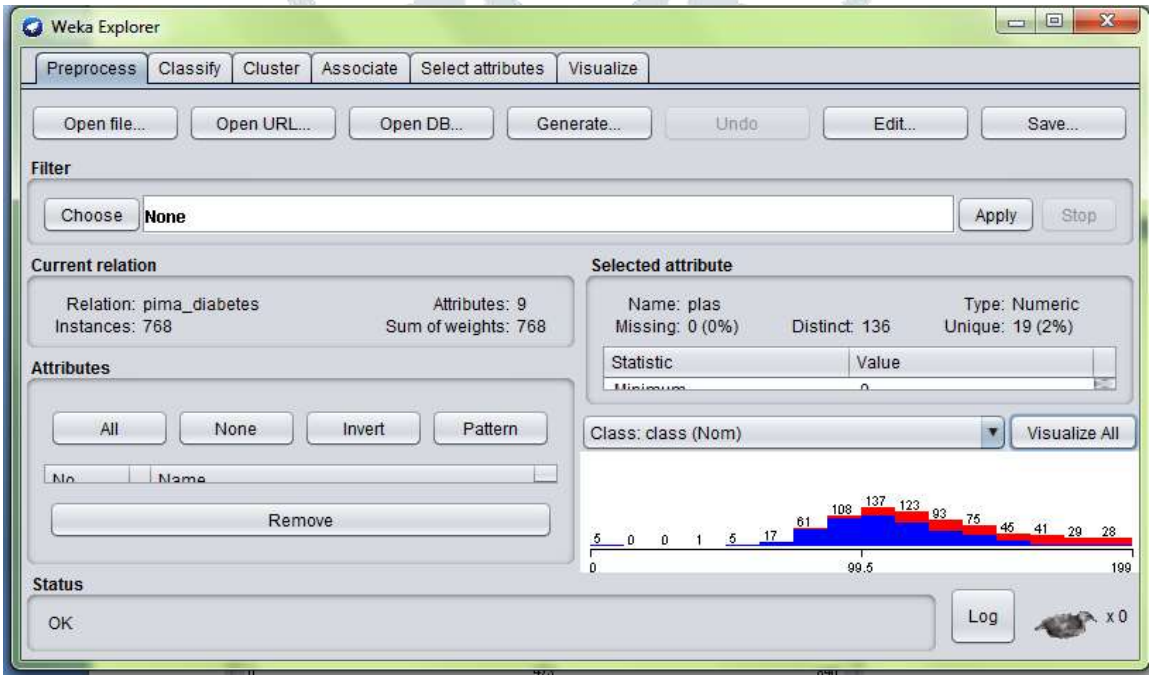


Figure 2: Overview of the weka tool environment

V. EXPERIMENTAL RESULTS

Pima Indian diabetes dataset is used for this study. The preprocessing techniques were applied on the instances of data sets. The Principle Component Analysis (PCA) is applied for reducing the dimensionality of dataset and it returned six attributes to be used for training the classifiers. The resample filter was used for omitting the replication of data. Hereby, the classifiers and cluster algorithms were applied to the Pima data set. The flow of attributes in Pima diabetic dataset is shown in figure 3.

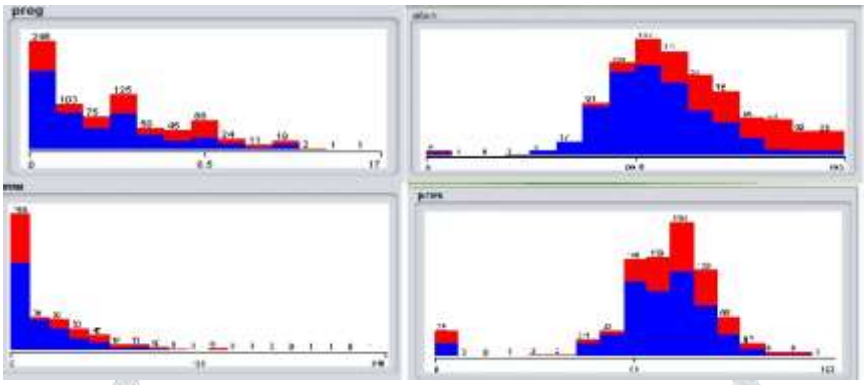


Figure 3: Process flow for attributes of PIMA Indian dataset

The results of classified instances are evaluated by comparing them in terms of Correctly Identified Correct Instances (CICI), Correctly Identified In-correct Instances (CIII), In-correctly Identified Correct Instance (IICI) and In-correctly Identified In-correct Instances (IIII). The most important operation in this work is to find accuracy, recall and precision. The f -measure and ROC were applied from this. F measure means average of precision and recall. In this study, three performances were considered viz Accuracy (Auc), Precision (Pcn) and Recall (Rcl). Auc is an arbitrary performance measure and deals with ratio of correctly predicted observation. If the class is balanced then the accuracy is best to measure. The formula used to calculate the accuracy is shown in equation 1,

$$\boldsymbol{Auc} = \frac{CICI+CIII}{TotalIdentifiedInstancese} \qquad (1)$$

Precision denotes the number of True Positives that are divided by the number of True Positives and False Positives. Therefore, it predicts the number of positive predictions divided by the total number of positive class values. Precision is also named as the Positive Predictive Value (PPV). The formula to calculate the precision is given in equation 2. Similarly, recall denotes the number of True Positives which are divided by the number of True Positives and the number of False Negatives. From this, the number of positive predictions divided by the number of positive test data class values. Recall is also known Sensitivity or the True Positive Rate. The formula to calculate the recall is given in equation 3. In this study, three algorithms such as J48, Naïve Bayes and KNN are used on the PIMA Indian dataset to classify the data. The comparison of precision and Recall in PIMA Indian Dataset for NB, J48 and KNN is shown in figure 4.

$$\boldsymbol{PPV} = \frac{CICI}{CICI+\text{IICI}} \qquad (2)$$

$$\boldsymbol{Rcl} = \frac{CICI}{CICI+\text{IIII}} \qquad (3)$$

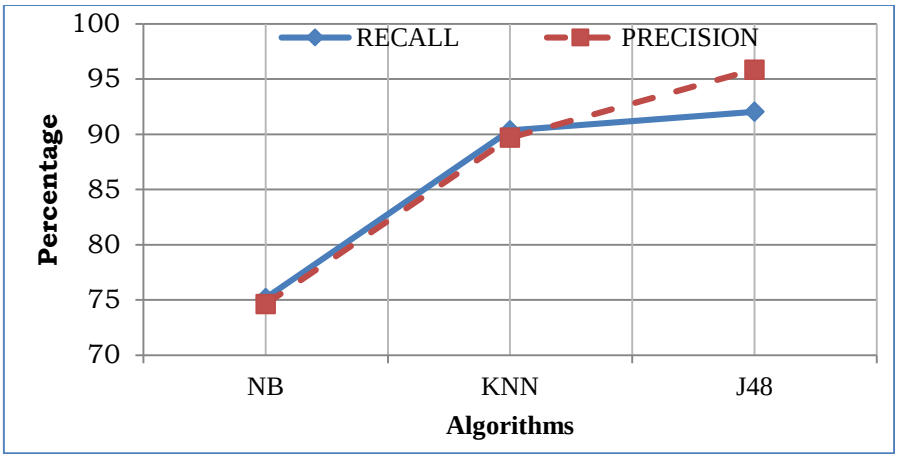


Figure 4: Precision and Recall for classification algorithm.

The accuracy is computed for three classification algorithms with PIMA Indian dataset and each accuracy is compared within themselves which is shown in the figure 5. It is observed that, J48 gives the higher accuracy (94.39 %) than Naïve bayes and KNN.

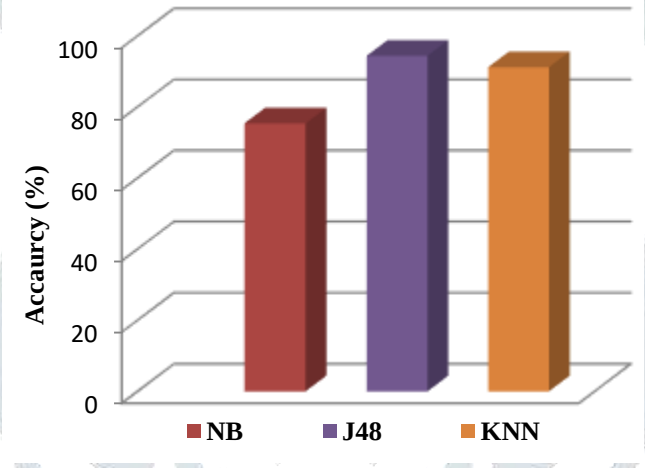


Figure 5: Accuracy for classification algorithms.

VI. CONCLUSION

Predictive analytic is the most essential and widely used technique. The process of predictive analytics is done with the help of various techniques such as data mining, machine learning and statistical tools. In this study, data mininig techniques was used for predicting diabetes disease which uses PIMA Indian diabetes data set. In order to diabetes disease was predicted by using three different algorithms like KNN, Naïve Bayes and J48. Finally, the result obtained from the experiment and J48 provided the better accuracy than other algorithms.

ACKNOWLEDGEMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).

REFERENCES

[1] J. Han and M. Kamber, ”Data Mining: Concepts and Techniques”, *Third Edition (The Morgan Kaufmann Series in Data Management System`s)*, 2000, 3rd Edition.

[2] K.S Deepikashri and Ashwinikamath, “Survey on Techniques of Data Mining and it’s Applications”, *International Journal of Emerging Research in Management &Technology*, Vol. 6, Issue: 2, 2017, pp: 198-201.

[3] Pradnya P. Sondwale, “Overview of Predictive and Descriptive Data Mining Techniques” *IJournals: International Journal of Software & Hardware Research in Engineering*, Vol.5, Issue. 4, 2015, pp: 53-60.

[4] Smita, Priti and Sharma, "Use of Data Mining in Various Field: A Survey Paper", *IOSR Journal of Computer Engineering*, Vol.16, Issue.3, 2014, pp: 18-21.

[5] Sayyed Muzammil Ali, Prof. Ms. R.R Tuteja," Data Mining Techniques", *International Journal of Computer Science and Mobile Computing*, Vol.3, Issue. 4, 2014, pp: 879 – 883.

[6] Brijesh Kumar Baradwaj, SaurabhPal" Mining Educational Data to Analyze Students Performance",(*IJACSA) International Journal of Advanced Computer Science and Applications*, Vol.2, Issue.6, 2011, pp: 63-69.

[7] Jan AndrzejNapieralski, " Statistical Methods for Data Prediction", *Department of Microelectronics and Computer Science, Lodz University of Technology*, 2016, pp: 20-24.

[8] SatbirKaur and Harjit Kaur," Review of Decision Tree Data mining Algorithms: CART and C4.5", *International Journal of Advanced Research in Computer Science*, Vol.8, Issue.4, 2017, pp: 436-439.

[9] Agrawal, Pragati, and Amit kumar Dewangan. "A brief survey on the techniques used for the diagnosis of diabetes-mellitus." *Int. Res. J. of Eng. and Tech. IRJET* , Vol. 2, Issue. 3, 2015, pp. 1039-1043.

[10] Priyanka Chandrasekar, Kai Qian, HossainShahriar and Prabir Bhattacharya," Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing", *IEEE 41st Annual Computer Software and Applications Conference*, 2017, pp: 481-484.

[11] Swaroopa Shastri, Surekha, Sarita, " Data Mining Techniques to Predict Diabetes Influenced Kidney Disease", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*,vol.2,Issue.4,2017,pp:364-368.

[12] Niketa Gandhi, leisa J.Armstrong, "Application of data mining techniques for predicting rice crop yield in Semi-Arid climate zone of India",*IEEE International conference on technological Innovations in ICT for Agriculture and rural Development*,2017,pp:116-120.

[13] Manal Abdullah and Salma Al-Asmari,"Anemia types prediction based on data mining classification algorithms",*Communication, Management and Information Technology – Sampaio de Alencar (Ed.)*,2017,pp:615-621.

[14] K. Lakshmi, D.Iyajaz Ahmed, G. Siva Kumar," A Smart Clinical Decision Support System to Predict diabetes Disease Using Classification Techniques", *IJSRSET*,vol.4,Issue.1,2018,pp: 1520-1522.

[15] Himansu Das, BighnarajNaik and H. S. Behera,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", *Springer Nature Singapore Pte Ltd*, 2018, pp: 539-549.

[16] Miss. N. Vijayalakshmi, Miss. T. Jenifer,"An Analysis of Risk Factors for Diabetes Using Data Mining Approach",*International Journal of Computer Science and Mobile Computing*, Vol.6, Issue.7, 2017, pp:166 – 172.

[17] B. Ahmed Mohamed Ahmed, AhmetRizanerc, Ali HakanUlusoyc," Using data mining to predict instructor performance",*12th International Conference on Application of Fuzzy Systems and Soft Computing*, 2016, pp: 137 – 142.

[18] C. Kalaiselvi and G. M. Nasira, "Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques", *Indian Journal of Science and Technology***,** Vol: 8(14), 2015.

[19] Mr. A. Amol , Wghmode, Mr. DarpanSawant, Prof. Deven D. Ketkar, "Heart Disease Prediction Using Data mining Techniques", *International Journal of Engineering Technology Science and Research*,vol:4,Issue:10,2017,pp:366-369.

[20] Jay Gholap," Performance Tuning of J48 Algorithm for Prediction of Soil Fertility", *Department Of Computer Engineering, College Of Engineering, Pune, Maharashtra, India*, 2017.

[21] Miss.SnehalS.Dahikar, Dr.SandeepV.Rode**,"** Agricultural Crop Yield Prediction Using Artificial Neural Network Approach**",** *International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering*, Vol.2, Issue .1, 2014, pp: 683-686.

AMODEL TO PREDICT AND PRE-TREAT DIABETES MELLITUS

Arockiam L[1], Dalvin Vinoth Kumar[2]
1,2Department of Computer Science, St. Joseph's College (Autonomous), Thiruchirapalli, Tamil Nadu, India

Abstract:

Diabetes Mellitus (DM) is one of the non-communicable diseases and it causes major health problems. According to a study, there will be 552 million diabetic patients by 2030 all over the world. The Things embedded with sensors that are connected to the internet is referred as Internet of Things (IoT). The collection, storage and analysis of data from IoT devices facilitate effective monitoring diabetic patients. In this paper, a model for prediction of diabetes is proposed. This prediction model consists of layer of sensors for data collection, layer for storage and layer for analytics. The diabetic data collection may include the data from other sources such as clinical experiments and questionnaire. The collected data are cleaned using pre-processing techniques. In the storage layer, the preprocessed data are stored in the warehouses. The predictive analytics is performed using statistical, data mining and machine learning algorithms in the analytical layer. This model provides an approach to predict the diabetic mellitus

Keywords—**Diabetes Mellitus, Predictive model, Diabetes predication, Clinical data analytics, Diabetics in India.**

I. INTRODUCTION

Diabetes is referred as diabetic mellitus in which blood sugar levels are too high. It is defined as a clinical syndrome characterized by hyperglycaemia, due to inadequacy of insulin in the human body. High levels of blood glucose can damage the blood vessels in kidney, heart, eyes and entire nervous system. Lack of awareness about diabetes can lead to these complications. According to WHO, India is the residence of the highest number of diabetics with the population of 79.4 million by 2030 [1]. The International Diabetes Federation said that nearly 52 % of Indians not aware that they are suffering from high blood sugar [2]. In particular, Madras Diabetes Research Foundation suggested that about 42 lakh individuals have diabetes and 30 lakh people are in pre-diabetes.

II. TYPES OF DIABETES MELLITUS

There are three types of diabetes. They are Type-1 diabetes, Type-2 diabetes and Gestational diabetes [3]. The presence of diabetics is identified using the following factors long-term blood sugar (HbA1C), fasting blood sugar, fotal triglycerides, Family history of high blood sugar, Waist measurement, Height and Waist-to-hip ratio[4].

i. **Type-1 Diabetes:** This type of diabetes is also called as insulin dependent diabetes. It will start from childhood. It is immune mediated and idiopathic forms of b cell dysfunction, which lead to absolute insulin deficiency [5]. This is also an auto-immune mediated disease process which gives rise to absolute deficiency of insulin and therefore total dependency upon insulin for survival. It increases the risk of heart disease and stroke. The symptoms are very thirsty, urinating frequently, rapid weight loss, feeling very hungry, feeling extreme weakness and fatigue, Nausea, vomiting and irritability. The treatments of type -1 diabetes are injections of insulin, oral medications or dietary modifications, physically activity, regular

check-up of blood sugar levels, controlling blood pressure and monitoring cholesterol levels [6].

ii. **Type-2 Diabetes:** It is also called as non-insulin dependent diabetes and adult onset diabetes. It is the most common form of diabetes. People may be affected by type-2diabetes at any stage, even during childhood [7]. Being overweight and inactive increases thechances of developing type-2 diabetes. It may originate from insulin resistance andrelative insulin deficiency. It can be controlled with weightmanagement, nutrition and exercise. The symptoms are very thirsty, urinating frequently,rapid weight loss, feeling very hungry, feeling extreme weakness, fatigue, nausea,vomiting, irritability, blurred vision, excessive itching, skin infections, sores that healslowly and dry and itchy skin [8]. Treatments such as using diabetes medicines, insulininjections, healthy food choices, exercise, Self Monitoring of Blood Glucose (SMBG),controlling blood pressure and monitoring cholesterol levels are some measures to control Diabetes Mellitus [9].

iii. **Gestational Diabetes:** According to the National Institutes of Health, the reported rate of gestational diabetes is between 2% to 10% of pregnancies [10]. Gestational diabetes usually resolves itself after pregnancy. It is caused by the hormones of pregnancy or a shortage of insulin. It causes risks to the life of baby which include abnormal weight gain before birth, breathing problems at birth, and higher obesity and diabetes risk later in life.

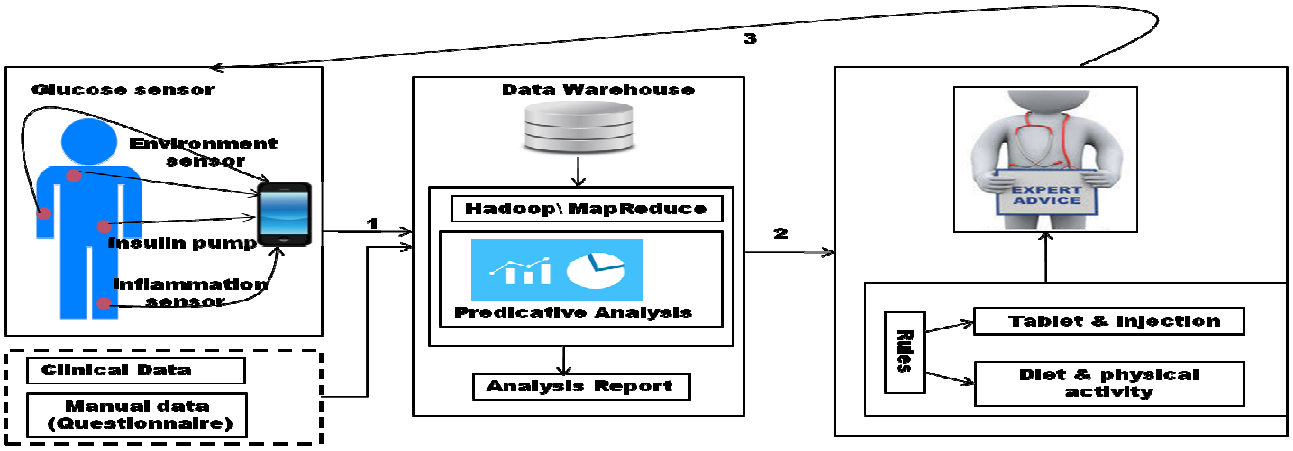
III. Problem Definition

In this modern era, human beings encounter different health issues. Most of the health issues are due to the food habits of the individuals. Based on the questionnaire, clinical data and sensor data, a predictive model is proposed to prevent the Diabetic Mellitus.

IV. The Proposed IDPM Model

A Diabetes Based Prediction Model plays animportant role in predicting diabetes and pre treatingdiabetic patients. It consists of three layers namelystorage layer and analytics and action layer. The layers in the proposed model for diabetic prediction are presented in Fig. 1.

The device layer consists of body sensors or wearables such as inflammation sensor, insulin pump, glucose sensor and environment sensor. These sensors are implanted in the body. These sensors are monitoring the health status of the human body.



Each sensor has its own capability to sense the data. The data gathered from these body sensors are electronic health record vitals, lab results and medical history of patients from hospitals or clinics. The second layer is cloud storage where the electronic records are stored securely. Cloud computing can be used for: data processing, data analysis and predictive analysis using statistical and data mining techniques and tools. Distributed Data analysis is performed by Hadoop/MapReduce.The predicative analytic engine sends report to doctors for consultation. The decision making at the cloud level is based on some rules such as diet pattern, physical fitness, current medicine intake etc. This layer is called as analytics and action layer. The rules for taking actions are set by physicians and medical experts for various health related issues. The rule based consultation will also consider previous health records and medical actions already taken. Here, the text pre-processing techniques are used. The doctor checks the analysis report of the patient. The doctor sends the treatment details from the prediction report to the insulin pump actuator as shown in Fig. 2.

Fig. 1. The proposed IDPM Model

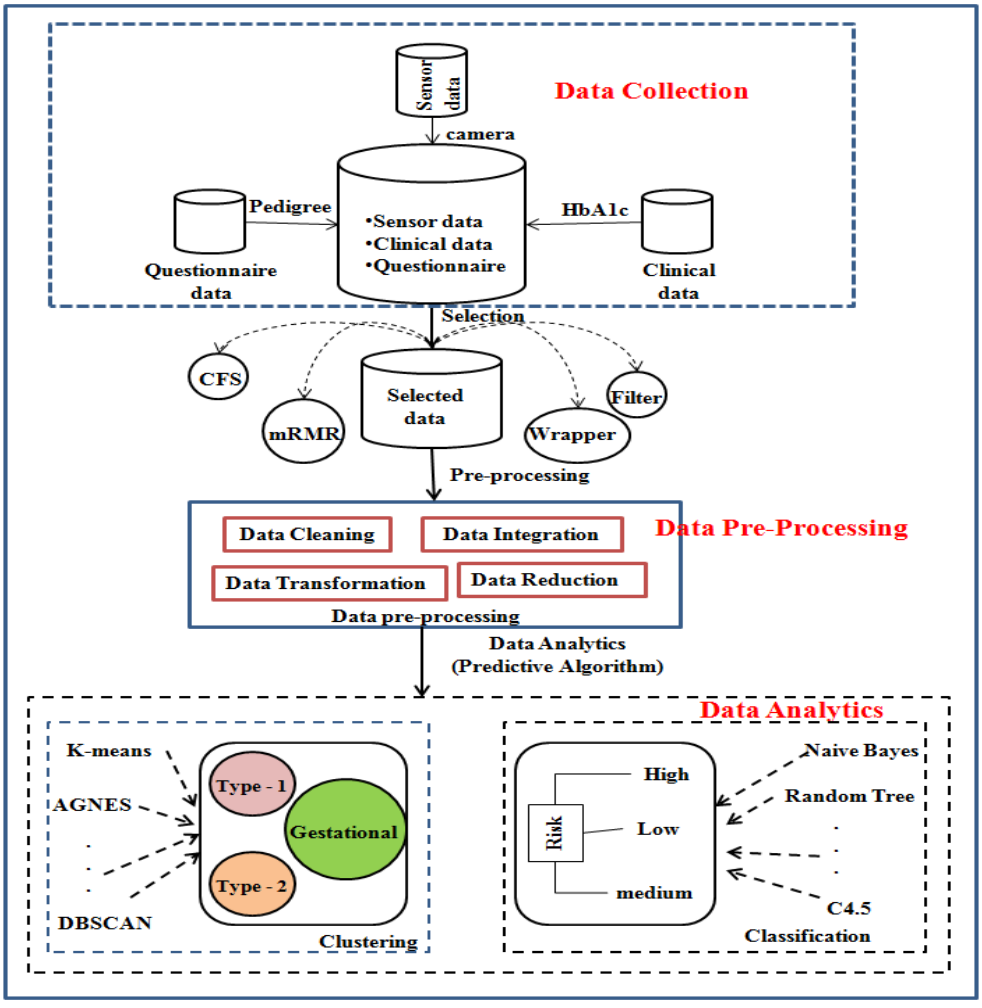


Fig. 2. The Diabetes Prediction

A. Data collection

Data collection is one of the most important stages of a research. Data collection is very demanding job which needs thorough planning, hard work, endurance, resolve and more to be able to complete the task successfully. Data Collection has two critical components. They are information gathering and decision making. Data collection is divided into two types. They are qualitative and quantitative data. Qualitative data are mostly non-numerical in nature. This means data will collect in the form of sentences or words. Quantitative data is numerical in nature and can be mathematically calculated.. In this proposed model, three types of data involved namely sensor data, clinical data and questionnaire data. The blood glucose level, body temperature, sleep time etc are collected from sensors. The clinical data like HbA1C test data are collected from clinical data. The family blood sugar history, number of time got pregnant etc., are collected from Questionnaire. Some of the data collected from sensors are qualitative and some are quantitative in nature.

B. PRE-PROCESSING

In real-time, it is very tedious to process massive amount of medical datasets containing information of individual patient health records so as to identify the disease pattern and to find out the causal association between them for planning curative actions. As a result, pre-processing of large volume of clinical data becomes essential. Data pre-processing is an important step in the data mining process. Data collection methods are largely loosely controlled, resulting in out-of-range values, missing values, etc. Observing the data which has not been carefully examined for such issues can produce misleading outcomes. If there is ample amount of incorrect and redundant information or noisy and unreliable data, then knowledge discovery becomes challenging.

Data pre-processing includes various steps such as cleaning, normalization, transformation, feature extraction and selection, etc. The outcome of data pre-processing is the complete data set with reduced attributes. The major drawback with clinical data set is the existence of redundant records. These redundant records cause the learning algorithm to be biased. So, eliminating redundant records is essential to enhance the detection accuracy. During pre-processing, the raw data is supplied as input and several suitable data pre-processing methods are applied thereby decreasing the invalid instances in the dataset.Data transformation and data validation are two important pre-processing techniques. They are explained below.

i. Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve Normalization, Smoothing, Aggregation and Generalization of the data.

a. *Normalization:* It is the process of converting the given values into a smallest range such as -1.0 to 1.0 or 0.0 to 1.0
b. *Smoothing*: Smoothing refers to removal of the noise from data. Smoothing techniques include binning, regression and clustering.
c. *Aggregation*: The Process of gathering information and expressing a summary form for the purpose of statistical analysis.
d. *Generalization*: Generalization is the process where low level or primitive data are replaced by higher level concepts through the use of concept hierarchies.

ii. Data validation

Data validation is defined as the assessment of all the collected data for entirety and reasonableness, and the elimination of error values. This step changes the raw data into validated data.. Data validation may be simple or complex depending on the way it is performed. Data validation can be updated either automatically or manually. The data validation helps to control the invalid data being entered into the system.

C. Predictive Analytics

The predictive analysis is carried out using classification and clustering algorithms. The clustering algorithms like K-means, DBSCAN etc. are used to cluster the pre processed data. The population is clustered based on urban/ village, male / female, educated / un-educated, diabetic / non diabetic etc. The clustered population is classified into DM high, low and medium using the classification algorithms like Bayesian network, J48, random tree etc. Some of the classification algorithms are explained below.

Bayesian network is statistical model which represents a set of variables and conditions. The relationship between the variables is carried out using Directed Acyclic Graph (DAG). In Bayesian network the nodes in DAG represent variables and edges between the nodes represent conditional dependency. Diabetic prediction using Bayesian network is influenced by the parameters. The Bayesian network for diabetic prediction with three variables (Blurred Vision (BV), Hunger and Fatigue (HF) and Family History (FH) of high blood sugar) implies Diabetes Mellitus (DM) with conditional dependency as shown in the fig. 3. The conditional dependency between the variables has two possibilities true and false respectively. The probability (pr) of occurrence of diabetics with respect to FS, BV and FV is given in equation 1.

$$Pr\,(FS\ BV\ FH) = Pr\left[\frac{FH}{(BVHF)}\right] * Pr\left[\frac{BV}{(HF)}\right] * Pr\,[HF]\ (1)$$

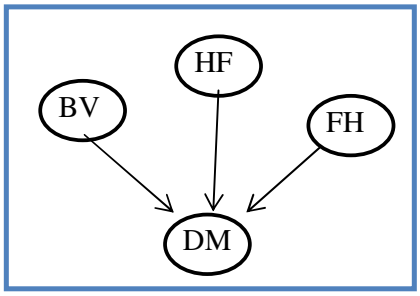


Fig. 3. Bayesian Network for Diabetes predication

The logistic regression is a statistical model used in various fields including machine learning. The decision is made by the influence of dependent variable (DV). In diabetic prediction, logistic regression utilizes general characteristics like age, Walking steps per day, Body Mass Index (BMI), Blood sugar level etc. Table 1 gives information about a set of patients (p1, p2, p3, and p4) walking steps per day and Blood sugar status. Logistic regression is suitable for the data in the table 1. The reason is, dependent variable Blood sugar status value is 1 or 0 represented to reduced or not reduced.The relationship between the steps walked per day and blood sugar status as shown in Fig. 4.

Table 1 Diabetes predication variable

Patient	Walking steps/day	Blood sugar Status
P1	4400	1
P2	3300	1
P3	2200	0
P4	0000	0

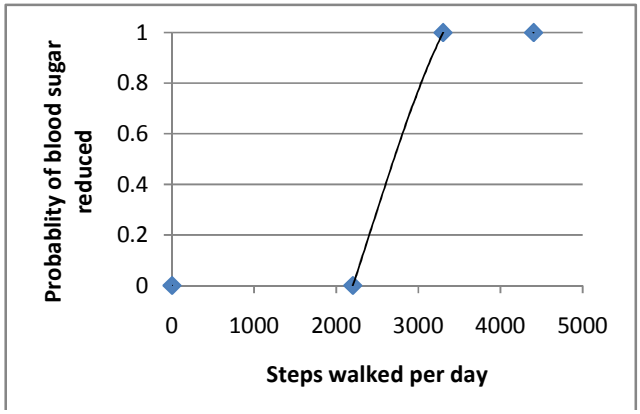


Fig. 4 Probability of blood sugar reduced versus number of steps walked.

V. CONCLUSION

Diabetes Mellitus is a chronic non communicable disease which has impact on human life span. A lot of datais collected from diabetic patients using IoT devices, clinical experiments and questionnaire, etc. The doctors can find value and make decisions when analysing the data accumulated from these sources. Early prediction of the deficiency will help the doctors to decide the treatment methods. Hence, a model for prediction of diabetes is proposed enabling pre-treatmentof the patients. .

ACKNOWLEDGMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).

REFERENCES

1. NDTV Food Desk, Updated: November 14, 2017 13:06 IST, available at: https://www.ndtv.com/food/world-diabetes-day-2017-number-of-diabetics-to-double-in-india-by-2023-1775180

2. Olson, Brooke. "Applying medical anthropology:Developing diabetes education and prevention programs in American Indian cultures", American Indian Culture and Research Journal, Vol:23, No:3, 1999, pp: 185-203.

3. World Health Organization. "Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus.", 1999, pp. 17-21.

4. American Diabetes Association. "Diagnosis and classification of diabetes mellitus", Vol:37, No:1 , 2014, pp : 81-90.

5. Gavin III, James R, "Report of the expert committee on the diagnosis and classification of diabetes mellitus." Diabetes care, Vol: 20, No:7, 1997, pp :963-967.

6. Diabetes Control and Complications Trial. "Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes." The New England journal of medicine, Vol: 353, No: 25, 2005, pp: 26-43.

7. Alberti, G., Zimmet, P., Shaw, J., Bloomgarden, Z., Kaufman, F. Silink, M. Type 2 diabetes in the young: the evolving epidemic. Diabetes care,Vol:27,No:7,2004, pp: 1798-1811.

8. Recognising Type-2 Diabetes', Feb 2016 Available : https://www.healthline.com/ health/type-2 -diabetes /recognizing-symptoms, [Accesed : 15-jan-2018].

9. Czupryniak, Leszek, "Self-monitoring of blood glucose in diabetes: from evidence to clinical reality in Central and Eastern Europe—recommendations from the international Central-Eastern European expert group.",Diabetes technology & therapeutics , Vol:16, No:.7, 2014, pp: 460-475.

10. Alatab, S., Fakhrzadeh, H., Sharifi, F., Mirarefin, M., Badamchizadeh, Z., Ghaderpanahi, M. Larijani, B.,." Correlation of serum homocysteine and previous history of gestational diabetes mellitus.",Journal of Diabetes and Metabolic Disorders, Vol : 12, No :1,2004, pp: 34-36.

11. Gillman, M. W., Rifas-Shiman, S., Berkey, C. S., Field, A. E., &Colditz, G. A. "Maternal gestational diabetes, birth weight, and adolescent obesity".Pediatrics, Vol:111, No: 3, 2003, pp:221-226.