International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

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].

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.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
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.

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.

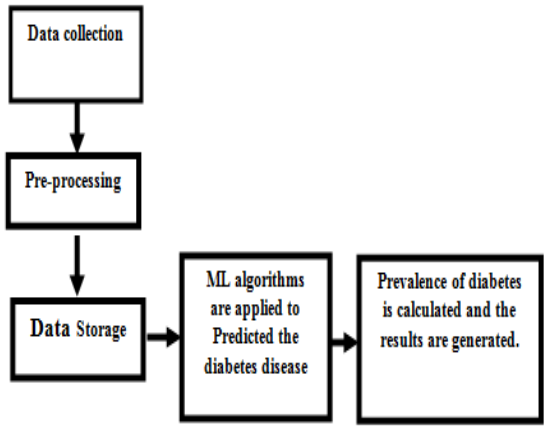

Fig 1: Work flow of proposed methodology

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.

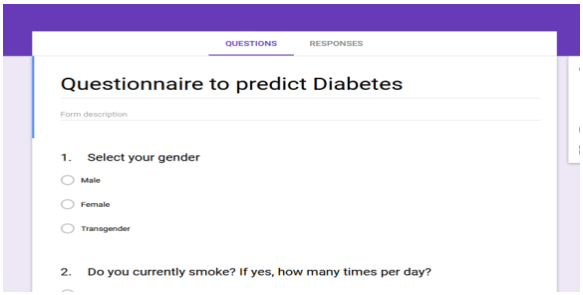

Fig 2: Questionnaire model based data collection

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.

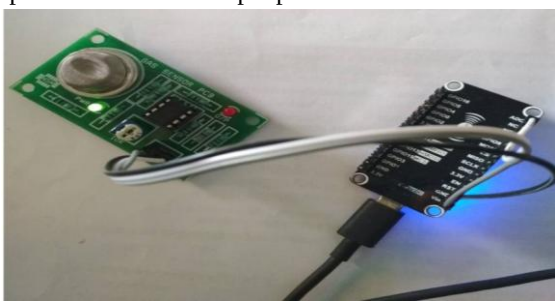

Fig 3: Data collection from sensor

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.

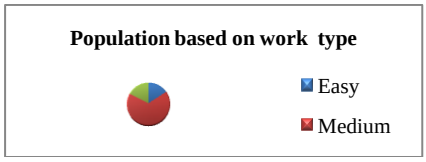

Fig 6: Population divided by work type.

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.

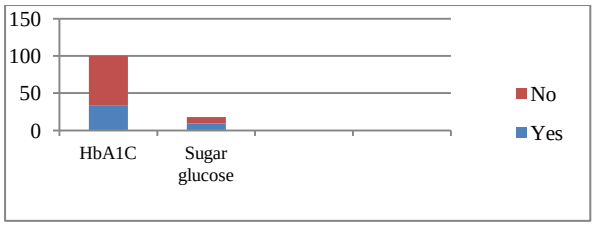

Fig 7: Awareness about Diabetes Mellitus

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.

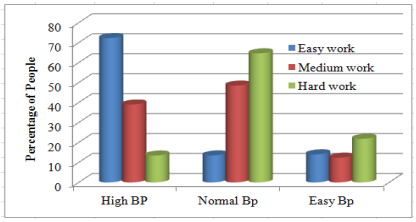

Fig 8: Work type vs Blood pressure level

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), Tiruchirapalli, 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), Tiruchirapalli, 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), Tiruchirapalli, 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