B.PUBLICATION

Diabetes prediction using decision Tree J48

BODLA VIKRAM
Department of Computer Science And Engineering
Sathyabama Institute of Science and Technology
Chennai, India
Email:bodlavikram92@gmail.com

CHINNI NIKHIL
Department of Computer Science And Engineering
Sathyabama Institute of Science and Technology
Chennai, India
Email:nikhilchinni06@gmail.com

Dr.A.MARY POSONIA
Department of Computer Science And Engineering
Sathyabama Institute of Science and Technology
Chennai, India
Email:soniadelicate@gmail.com

Abstract—
Gestational diabetes is found among majority of the Indian pregnant women, when un-attended may give birth defects to child. Diabetes, which is caused by the rise in level of glucose in blood, has many latest devices to identify from blood samples. Diabetes, when unnoticed may bring many serious diseases like heart attack, kidney disease. In this way there is a requirement for solid research and learning models enhancement in the field of gestational diabetes finding and analysis. In this work, we proposed machine learning knowledge, for example, Decision Tree J48 calculation for diabetes forecast. Decision Tree is one of the powerful classification models. The dataset considered of 768 patients data with major 8 features and a target column with result "Positive" or "Negative". Experiment is done with weka, outcome of our demonstration shows that Decision Tree J48 calculation gives more efficiency with less processing time.

Keywords— Diabetes, Decision Tree J48, Machine Learning, Classification algorithm, Weka tool

I. INTRODUCTION

Information mining is a powerful technique for extraction of data on immense dataset. Data mining can be exploited in hospital dataset where we need clustering, classification, pattern recognition, machine learning such as prediction, applying and in identifying statistical techniques.

These days, diabetes is a typical disease influences the individual who has imbalance in blood glucose levels and furthermore pregnant women confronting diabetes issues. Diabetes is a general cause brought by high blood glucose level.

Numerous investigations and research demonstrates that pregnant women with diabetes increasingly inclined to have a child with birth defects than ladies without diabetes. The child may influence by illness condition, for example, coronary illness and spina bifida. Diabetes mellitus mainly sorted as three types such as diabetes mellitus, insulin resistance and third one is gestational diabetes generally found in pregnant ladies.

The primary aim of our work is to classify gestational diabetic or non-gestational diabetic. Proposed work is based on decision tree j48. Major study of classification systems give high accuracy with high handling time, though a few strategies give low precision even with enormous dataset. Along these lines, our work goes for high accuracy with immense dataset and less process time invited articles in the journals.

The below figure, Figure 1, we represent the machine learning carried out over medical dataset. Nowadays, many efficient analysis techniques are available for affordable cost. These data analysis improves detection accuracy in modern hospitals.

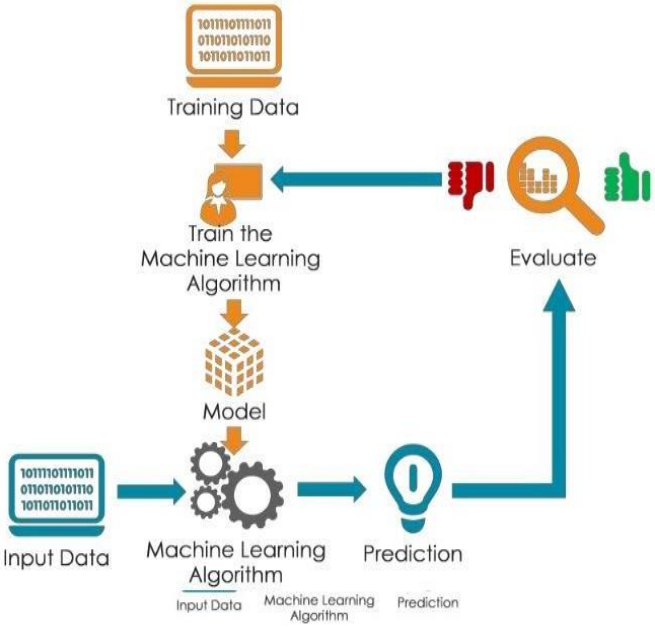

Fig. 1 Machine learning Model

Gestational diabetes, when early detected can avoid serious complications and also controlled by a healthy

diet. This study focuses on gestational diabetes by applying Decision Tree j48 classifier on our dataset.

The remainder of this paper provides an overview of existing research handled by various authors in diabetes detection using machine learning approaches. Section 3 provides complete view of our implementation details. We end this study with a conclusion and future reference in section4.

II. EXISTING WORK

Numerous works has been proposed on diabetes discovery systems exploiting machine learning clustering, classification, information mining and learning. In this study, few of them is discussed with their concise proposition. One of the works done by Sajida et al. [1] in machine learning utilizing Decision tree J48 for Diabetes Mellitus dependent on risk factors. In their system they demonstrated that Ad boost outperforms as far as efficiency is concerned than bagging and Decision tree J48.

Deepti et.al. in [2] examined execution performance of classification algorithms specifically Decision Tree, SVM and Naive Bayes. In their work they considered Pima Diabetes Database (PIDD), though the highest accuracy accomplished by their work is around 76.3%. In our proposed work, we attempted to accomplish over 80% by considering the equivalent dataset.

The classification technique on machine learning namely SVM for diabetes detection is exploited by Santi Waulan [3], they proposed improved version of SVM namely Smooth SVM (SSVM) and MKS-SSVM. They demonstrated over Pima dataset. In their results, they achieved about high accuracy for MKS-SSVM than SSVM.

Faezeh et al [4] considered Fuzzy Clustering method (FACT), which decides the quantity of fitting clusters dependent on density. The proposed algorithm is insensitive to initial number of clusters, while initial cluster numbers are less than threshold number of clusters. Their strategy discovered number of cluster by making new cluster focuses through outlier detection. In their work, they demonstrated experimentally that proposed heuristic algorithm exhibit a superior performance than conventional K-means calculation.

Radha et al, in [5] proposed fuzzy logic based application to analyze diabetes. In their model, they proposed two correlation among symptoms and diseases, specifically occurrence relationship and a confirm ability relationship. Occurrence relationship confirms recurrence of appearance of a symptom; confirm ability relationship portrays the intensity of symptoms for disease presence. Likewise they proposed Fuzzy Logic with minimum and maximum relationship. They made use of real world dataset of 40 patients and determined the fuzzy relationship.

Nongyao et al. [6] in their research considered the risk of diabetes by order procedures. In their work, they proposed four machine learning procedures in particular Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes. They likewise created as a web application with PHP as front end and backend MySQL, in which they utilized ROC curve method for diabetes forecast. The data are fed in the application display they predicted the output with actual and forecasting. They experimentally proved that Random Forest accomplishes great accuracy.

III. PROPOSED WORK

In our proposed work, we considered classification algorithm, Decision Tree J48 and applied over Pima Indians Diabetes Database. This dataset is analyzed using weka tool. Many data mining tools were available, whereas weka is found to be efficient and easy to use for research and analysis purpose. Out of many available classification algorithms, Decision Tree J48 is preferred for its better accuracy on prediction part. The below table, Table1 represents the characteristics of PIMA dataset. Missing attributes and Noisy attributes are not considered in our dataset.

Table I. Dataset Characteristics

Data set	PIMA
Number of samples	768
Feature Attributes	8
Output classes	2
Total number of feature attributes	9
Missing attribute status	None
Noisy attribute status	None

Table II. Gestational diabetes dataset description

Features description	**Features Symbol**
Number of times pregnant	Preg
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Plas
Diastolic blood pressure (mm Hg)	Pres
Triceps skin fold thickness (mm)	Skin
2-Hour serum insulin (mu U/ml)	Insu
Body mass index (weight in kg/(height in m)^2)	Mass
Diabetes pedigree function	Pedi
Age (years)	Age
Class variable (Positive or Negative)	Class

The above table, Table2 represents the feature attributes of our PIMA dataset along with their represented symbols

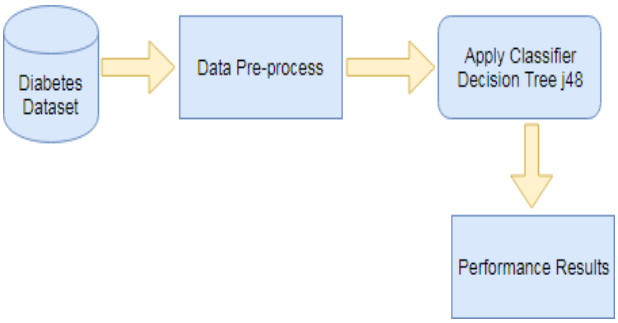

Fig 2. Overall Architecture of Proposed work

The above figure, Fig 2 is the overall process handled in our proposed system. First we load the dataset, second it is pre-processed for any null values, third, pre-processed data is applied our algorithm DT J48 and finally generating the output results with efficiency and elapsed time for the evaluations.

The above pre-processing steps make dataset ready to use for experiment.

The dataset is ready to use as the feature values are converted to numeral values. For example, the class value is converted to 0 or 1 for tested _positive and tested _negative.

Feature selection is done using CfsSubsetEval algorithm, and considered attributes are given below

1. Plasma glucose concentration
2. Body mass index (kg/m2)
3. Diabetes pedigree function
4. Age (years)
5. Class Variable (nominal) - tested _positive and tested negative

The Proposed work is executed with DT J48 classifier algorithm and discussed in section. The above figure represents the Decision Tree model for gestational diabetes, where we represented the root nodes and its flow to leaf nodes. This sample visualization of proposed algorithm shows how the algorithm works for PIMA database with 8 features. The above tree shows plasma, pedigree and number of times pregnancy are found to be the important features in our classification model. For example, patients with plasma value above 116 mg/dl and pedigree value above 0.2 are more prone to diabetes.

Similarly, for the patients with plasma value above 116 mg/dl and number of time pregnancy is above two are also more prone to diabetes.

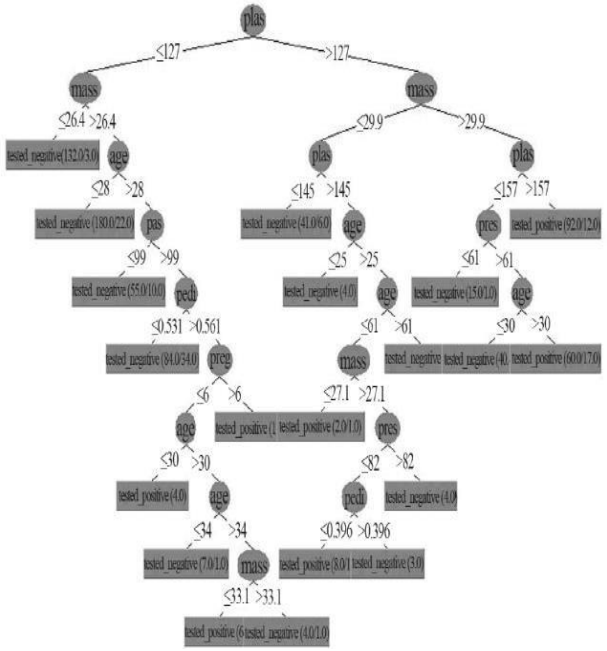

Fig 3. Decision Tree J48 Algorithm for Gestational diabetes with Weka

```
Relation:     pima_diabetes.
Instances:    768
Input Attributes: 5: plas, pres, mass, pedi, age

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561: tested_positive (34.0/9.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
```

```
| | | | | | | pedi <= 0.396: tested_positive (8.0/1.0)
| | | | | | | pedi > 0.396: tested_negative (3.0)
| | | | | | pres > 82: tested_negative (4.0)
| | | | age > 61: tested_negative (4.0)
| mass > 29.9
| | plas <= 157
| | | pres <= 61: tested_positive (15.0/1.0)
| | | pres > 61
| | | | age <= 30: tested_negative (40.0/13.0)
| | | | age > 30: tested_positive (60.0/17.0)
| | plas > 157: tested_positive (92.0/12.0)
```

IV. RESULTS AND DISCUSSION

The experiments result shows that number of leaves is 20 and size of tree is 39. Time taken to build the model is 0.12 seconds. The below table shows precision metrics and other error values.

Accuracy (Ac):

Correctness determines the precision of the algorithm over foreseeing instances.

Ac = (TPS+TNS) / (Total number ofsamples).

Precision (PN):

Classifiers correctness / accuracy is measured by Precision.

PN = TPS / (TPS+ FPS).

Recall (RC):

To measure the classifiers completeness.

RC =TPS / (TPS+FNS).

F-Measure (FM):

F-Measure is the weighted average of precision and recall

FM=2*(PN*RC) / (PN+RC)

Precision	0.804
Recall	0.780
F-Measure	0.792
Mean absolute error	0.312
Root Mean square error	0.4528

We provided the confusion matrix for our experiments below

Tested _negative	390	110
Tested _positive	95	173

Experimental results shows that our proposed work achieve 91.2% efficiency

```
Algorithm Efficiency
Efficiency : 91.19892499659267
Elapsed Time - -5213
```

Fig 4. Experimental results achieved by Decision Tree J4 using Weka tool

V. CONCLUSION

Medical data need to be processed to find out the pattern and extraction of data for analysis purposes, data mining and machine learning were used. In different sectors of medical, these techniques were found useful including medical image processing like brain tumour, cancer disease detection, diabetes, lever disease and heart disease, Parkinson disease identification, early detection of leukaemia etc. In this work, we considered diabetes, as it found to be a common and major disease amongst Indians. We considered Pima Indian diabetes database, evaluated in weka tool using Decision tree J48 algorithm. Our work achieved 91.2% efficiency. Applying convolution model of deep neural network is our interest of further study on this research. Also this research can be extended to apply feature selection method before training the model.

Using the web application we can predict whether a person is having diabetes or not. User need to enter the details through the web application

REFERENCES

[1] Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib, Karim Keshavjeec, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes" Procedia Computer Science 82 (2106) 115 – 121.

[2] Deepti Sisodia, Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms", International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

[3] Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain and S.P. Rahayu, "A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis", Journal of Computer Science 5 (12): 1003-1008, 2009

[4] Faezeh Ensan, Mohammad Hossien Yaghmaee, Ebrahim Bagheri, "FACT: A new Fuzzy Adaptive Clustering Technique", The 11th IEEE Symposium on Computers and Communications, Sardinia, 26-29 June 2006

[5] Aishwarya, Gayathri, Jaisankar, "A Method for Classification Using Machine Learning Technique for Diabetes", International Journal of Engineering and Technology 2013.

[6] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE.

[7] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and MachineIntelligence19, 476–491.

[8] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications.

[9] Kayaer, K., Tulay, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the international conference on artificial neural networks and neural information processing.

[10] Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7.