Prediction of Type 1 Diabetes Mellitus using Datamining Techniques

K.Saminathan

Abstract**:** ***Type1 diabetes is a sickness occurs when your immune system fighting against infection, affects and erode the insulin generating beta cells of the pancreas. In general, when the blood sugar stage increases, the pancreas makes more insulin. Insulin helps to go sugar out of the blood so it can be used for liveliness. Type 1 diabetes occurs due to the immune system which affects cells in the pancreas that make insulin. The pancreas cannot make adequate insulin, so the blood sugar level continues to increase. According to the children history of type 1 diabetes may enhance risk of their life. Type 1 diabetes cannot be cured, but it can be controlled and managed. In this study we use Naive Bayes, linear regression and k-means algorithm for data analysis and prediction. It predicts the diabetes affected children with maximum level of accuracy 96% by using of data mining algorithms.***

Keywords **:** ***Type 1 Diabetes, Naive Bayes, Linear Regression, K-Means, Decision Stump.***

I. INTRODUCTION

Diabetes is a chronic disease that comes when the pancreas is not able to make enough insulin, or when the body cannot fulfill good use of the insulin it produces [1]. It has three types.

- Type 1
- Type 2
- GDM

Type 1 diabetes is a disease in which the body makes little insulin or not to control blood sugar levels. Type 1 diabetes is kown as an insulin-dependent diabetes. It is otherwise named as juvenile diabetes. Digestion process converts the food which we intake is broken down into vital components. In general one of the major energy producing food material is carbohydrate. It is further refined into sugars and glucose. Human body cells obtain energy mainly from glucose. Glucose usually invades into cells to give energy to the body [2]. Insulin is a hormone made by the pancreas that acts like a key to control blood sugar. If our body doesn't produce enough insulin to control the sugar, it may develop hyperglycemia (high blood sugar). Hyperglycemia provides long term complications. To avoid those complications, the body needs insulin. But People with type1 diabetes cannot make insulin because the pancreas are damaged and destroyed. So they need insulin injection to control blood sugar.Insulin disagreement actually had symptoms like extreme hunger, weight loss, fatigue, irritability or behavior changes [3].

Diabetes patients' count grows day by day in all over the world. Classification techniques are preferred and extensively applied in the medical field. The collection of data are classified into various classes according to required constrains that helps to predict the disease. Researchers were conducted experiments to diagnose the diseases using different classification algorithms of machine learning approaches like J48, Support Vector Machine, Naive Bayes, Decision Tree, Decision Table. Researchers have determined that machine learning algorithms works better in diagnosing different diseases. Data Mining and Machine learning algorithms have the competence of handling abundant amount of data from several different sources and integrating the background information for study of possible cases. The proposed work focuses on children who are suffering from diabetes. In this work classification algorithms and also clustering algorithms

- Linear Regression algorithms
- Naive Bayes
- Decision Stump
- K-Means

are used to identify the prediction of diabetes from patients' data set. Experimental performances of all the algorithms are compared on various measures and achieved good accuracy.

II. RELATED WORKS

Prediction of diabetes type-1 using classification algorithms discuss the multiple and different strategies, which are widely used in the medical field. The classification process of data into diverse classes according to derived constrains comparatively an individual classifier [4]-[7]. Diabetes prediction system has the main objective of prediction of diabetes among particular or certain age of people suffered by type–I [8]. Machine learning system is applied to design the diabetes system. The decision tree algorithm is applied. The experimental results were good enough as the designed system facilitates well in predicting the diabetes incidents at group of age, with better accuracy level using Decision tree [9]. Dataset on significant risk factors for Type 1 Diabetes are reviewed [10]. In this study explain dataset and detailed data analysis results of Type-1 Diabetes have been given. Comparison of Classifiers for the Risk of Diabetes Prediction. The diabetes mellitus prediction model based on data mining is designed to get prediction [11].

Revised Manuscript Received on October 15, 2019
***** Correspondence Author

Dr.K.Saminathan*, Assistant Professor, Department of Computer Science, A.V.V.M Sri Pushpam College, Poondi. arksami@gmail.com
(Affiliated to Bharathidasan University)

Four well known classification models that are Artificial Neural Networks, Decision Tree, Naive Bayes and Logistic Regression were tested. This model uses data mining algorithms and ensures that the dataset quality is sufficient. It gets a better accuracy level of prediction.

III. RESEARCH TOOLS AND METHODOLOGY

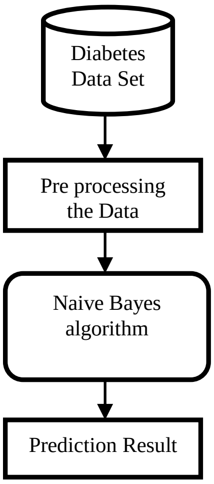


Fig.1. Proposed work Architecture

The proposed work architecture is shown in figure 1. In this research many approaches such as Linear regression, Naive Bayes, Decision Stump and K-Means were experimented to find the best result.

A. WEKA Tool

In this work WEKA tool software is used to experiment. This tool is designed at University of Waikato in the country of New Zealand. The designed tool contains a collection of enormous machine learning techniques for regression, data classification, clustering, visualization, Association etc. The major advantage of using WEKA is that it supports customization as per the requirements of the problem. The primary goal of this study is the prediction of the patient suffered by type1 diabetes using the WEKA tool in the medical database Indian pediatrics.

Table-I: Diabetes Data set Description

Age	Height	Weight	BMI	Pressure	Class	Insulin
6	133.4	23	22.1	110.7	1	8.7
9	151.8	28.3	27	115.3	1	6
7	139.3	22	23.5	109.3	0	9.1
5	125.2	23.7	21.7	104.5	1	9
8	144.4	20.7	24.8	112.7	1	9.2
7	138.9	28.4	27.3	109.5	1	9.1
10	156.4	30.4	28.4	119.4	0	5
6	133.9	24.3	22.4	106.9	1	9.1
7	139	22	23.7	109.3	1	9.1
5	150.9	29.3	27.4	108.2	1	9

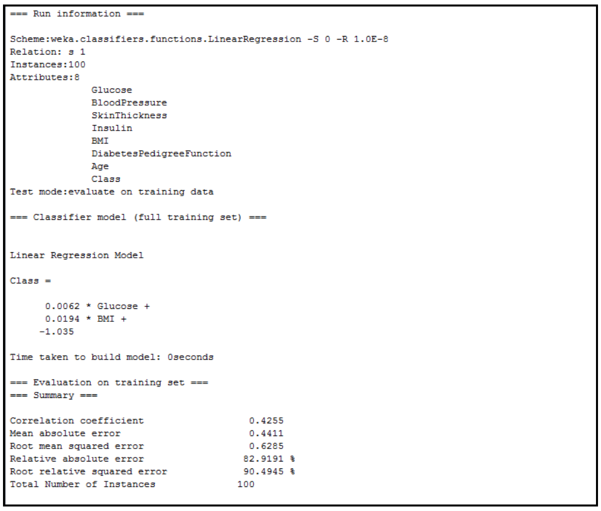
In this research 102 instances with 7 attributes are experimented and verified. Four data mining algorithms were applied to verify the best which results good result.

B. Linear Regression

Linear regression algorithm tries to model the relationship between two variables by fitting a linear equation to observed data [11]. A linear regression line has an equation form

$$L = m + nX,$$

Where X is the explanatory variable, L is the dependent variable.

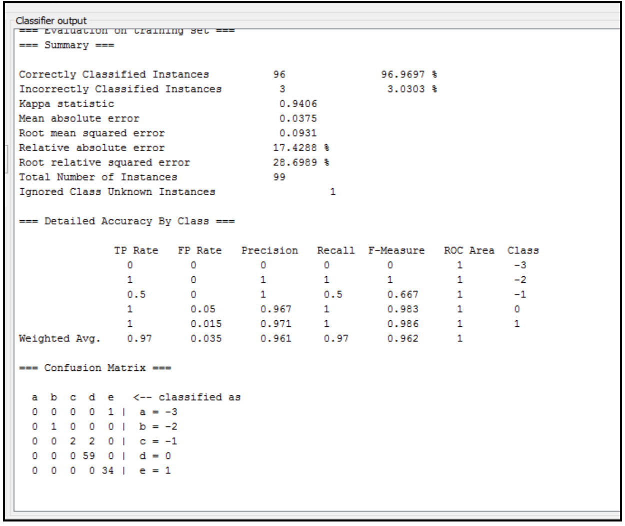


```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation: s 1
Instances:100
Attributes:8
              Glucose
              BloodPressure
              SkinThickness
              Insulin
              BMI
              DiabetesPedigreeFunction
              Age
              Class
Test mode:evaluate on training data

=== Classifier model (full training set) ===


Linear Regression Model

Class =

      0.0062 * Glucose +
      0.0194 * BMI +
     -1.035

Time taken to build model: 0seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.4255
Mean absolute error                      0.4411
Root mean squared error                  0.6285
Relative absolute error                 82.9191 %
Root relative squared error             90.4945 %
Total Number of Instances              100
```

Fig.2. Result of Linear Regression Algorithm

C. Bayesian Classification

One of the widely used statistical classifiers is Bayesian classifier. The class membership probabilities are easily predicted by this classifier. The probability of a given tuple fit to a particular class. The Naive Bayes classification algorithm is a probabilistic classifier algorithm. It is based on probability models that associate strong independence assumptions.

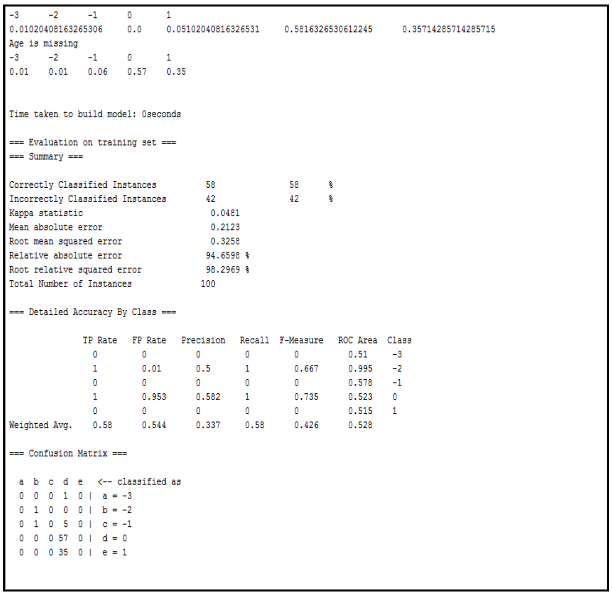


```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          96               96.9697 %
Incorrectly Classified Instances         3                3.0303 %
Kappa statistic                          0.9406
Mean absolute error                      0.0375
Root mean squared error                  0.0931
Relative absolute error                 17.4288 %
Root relative squared error             28.6989 %
Total Number of Instances               99
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0         0          0         0          1        -3
                 1         0         1          1         1          1        -2
                 0.5       0         1          0.5       0.667      1        -1
                 1         0.05      0.967      1         0.983      1         0
                 1         0.015     0.971      1         0.986      1         1
Weighted Avg.    0.97      0.035     0.961      0.97      0.962      1

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  0  1 |  a = -3
  0  1  0  0  0 |  b = -2
  0  0  2  2  0 |  c = -1
  0  0  0 59  0 |  d = 0
  0  0  0  0 34 |  e = 1
```

Fig.3. Result of Naive Bayes Algorithm

D. Decision Stump

A machine learning method, which consist of a one-level decision tree is decision stump. “That is, it is a decision tree with one internal node which is root node is immediately connected to the terminal nodes such as leaf nodes.” A decision stump predictions are decided based on the value of mono input.



```
-3      -2      -1      0       1
0.01020408163265306     0.0     0.05102040816326531     0.5816326530612245      0.35714285714285715
Age is missing
-3      -2      -1      0       1
0.01    0.01    0.06    0.57    0.35


Time taken to build model: 0seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          58               58      %
Incorrectly Classified Instances        42               42      %
Kappa statistic                          0.0481
Mean absolute error                      0.2123
Root mean squared error                  0.3258
Relative absolute error                 94.6598 %
Root relative squared error             98.2969 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.51     -3
                 1         0.01       0.5       1         0.667      0.995    -2
                 0         0          0         0         0          0.578    -1
                 1         0.953      0.582     1         0.735      0.523    0
                 0         0          0         0         0          0.515    1
Weighted Avg.    0.58      0.544      0.337     0.58      0.426      0.528

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  1  0 |  a = -3
  0  1  0  0  0 |  b = -2
  0  1  0  5  0 |  c = -1
  0  0  0 57  0 |  d = 0
  0  0  0 35  0 |  e = 1
```

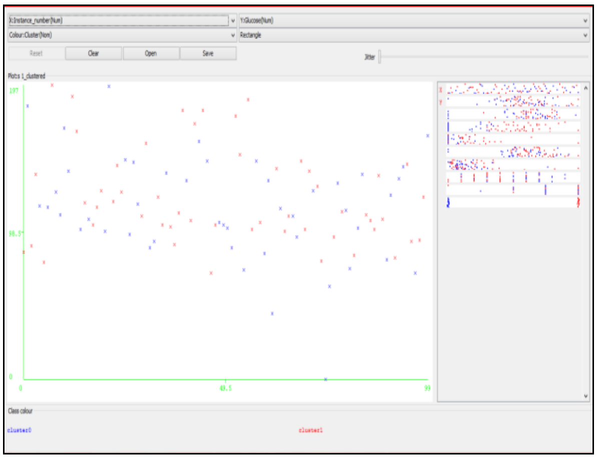
Fig.4. Result of Decision Stump Algorithm

E. K-Means algorithm

A cluster is a collection of similar data objects. It groups the similar data for analysis [12]. K-means clustering algorithm is an unsupervised learning algorithm. The k-means clustering algorithm tries to split a given anonymous data into k clusters. Initially k chooses centroids. A centroid is a center data point (imaginary or real).

Table-II: Cluster Formation and Number of Clusters

Number	Label	Count
1	Cluster 0	36
2	Cluster 1	64



X:Instance_number(Num)
Y:Glucose(Num)
Colour:Cluster(Nom)
Rectangle

Fig.5. Result of K-Means Algorithm

The following table explains the accuracy level of different algorithms.

Table- III: Accuracy Levels from Different Algorithms

No	Algorithm	Accuracy
1	Linear Regression	82
2	Naive Bayes	96
3	Decision Stump	58
4	K- Means	77

IV. PERFORMANCE OF VARIOUS ALGORITHMS

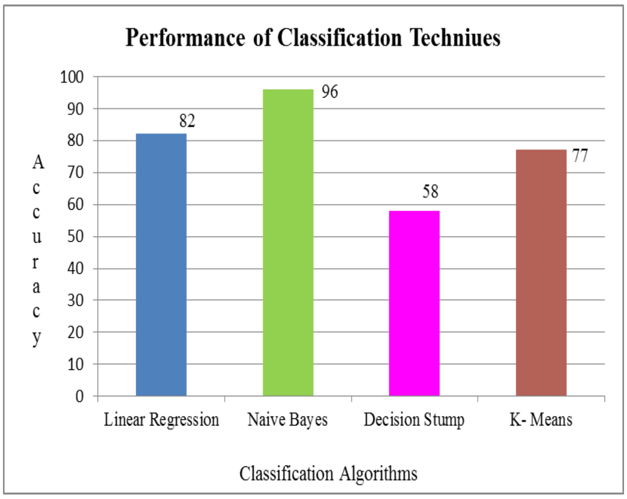


Fig.6. Performance of Classification Algorithms

Performance of four classification algorithms are depicted clearly in Figure 6. It is clear the eperimental work of algorithms as Linear regression, Naive Bayes, Decision Stump and K-Means results 82, 96, 58 and 77 percentage of accuracy respectively.

V. CONCLUSION AND FUTURE WORK

One of the most wide spread real-world medical hurdle is the detection and prediction of diabetes at its primary stage. In this research work, proficient efforts are used in designing a system which results in the prediction of emerging ruin out disease like diabetes. In this system design the data mining algorithms are used for diabetes prediction with maximum accuracy level of 96% is obtained. The Maximum accuracy level is obtained from Naive Bayes algorithm compared with other algorithms.

For future work, it is suggested and expected to use the hospital's real and recent patients' data for continuous training and development of our proposed model. The size of the dataset is preferred to be large volume in training phase and predicting. In future work will use Big Data Technology for analyzing large amount of data. Use effective new algorithms for accurate prediction.

REFERENCES

1. Anastasia Katsarou, et.al. "Type1 diabetes mellitus", National Rev. Dis. Prim. 3, 2017 PP.17016.
2. Aditi Narsale, et.al. "Type1 Diabetes Trial Net Study Group, Data on correlations between Tcell subset frequencies and length of partial remission in type1diabetes" , Data Brief8, 2016, PP.1348–1351.
3. K.Konrad, et.al. "Current practice of diabetes education in children and adolescents with type1 diabetes in Germany and Austria: analysis based on the German/Austrian DPV database", Pediatr. Diabetes 2016, PP.483–491.
4. Aishwarya, R., et.al. "A Method for Classification Using Machine Learning Technique for Diabetes", International Journal of Engineering and Technology, 2013, Vol. 5(3): PP..2903-2908.
5. Aljumah, A.A., et.al. "Application of data mining: Diabetes health care in young and old patients", Journal of King Saud University - Computer and Information
Sciences 25, 2013, PP. 127–136.

6. Dayna E.McGill, et.al. “Management to hypoglycemia in children and adolescents with type1diabetes mellitus”, Curr. Diabetes Rep. 2016, Vol.16 (9), PP.88.
7. Jennifer L.Sherr, et.al. “Use of insulin pump therapy in children and adolescents with type1 diabetes and its impact on metabolic control: comparison of results from three large, transatlantic paediatric registries”, Diabetologia, 2016, Vol. 59(1), PP. 87 – 91.
8. K.Ahmed, et.al. “Early prevention and detection of skin cancer risk using datamining”, International Journal Computer Application. 2013, Vol. 62(4).
9. Priyam A., et.al. “Comparative Analysis of Decision Tree Classification Algorithms”, International Journal of Current Engineering and Technology, 2013, Vol.3, PP. 334–337.
10. Orabi K.M., et.al. “Early Predictive System for Diabetes Mellitus Disease”, Industrial Conference on Data Mining, Springer. 2013, PP.420–427.
11. Nongyao Nai-aruna, et.al. “Comparison of Classifiers for the Risk of Diabetes Prediction”, Procedia Computer Science, 2015, Vol. 69, PP.132 – 142.
12. Han Wu, et.al. “Type 2 diabetes mellitus prediction model based on data mining”. Informatics in Medicine Unlocked”, 2018, Vol. 10, PP.100-107.

AUTHORS PROFILE

Dr.K.Saminathan, Assistant Professor, Department of Computer Science, A.V.V.M Sri Pushpam College, Poondi. He has thirteen years of experience and guided more than 20 projects. He has completed his Post Graduate degree in Computer Science at Periyar E.V.R College, Trichy in 2006. Master of Philosophy in Computer Science at A.V.V.M. Sri Pushpam College, Poondi in 2007. Ph.D in Computer Science at Bharathidasan University and degree awarded in the year of 2017. He published more than 10 research articles in various domains such as Image Processing, Computer Networks and Machine Learning in peer reviewed International Journals.