International Journal of Engineering & Technology, 7 (4.19) (2018) 397-401

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predicting the Blood Sugar using Machine Learning Approach

V.V.Ramalingam[1*], A.Pandian[2], K.Manikandan[3], Neela Mega Samala Roobanvaikundaraja[4]

[1, 2, 4] Department of Computer Science, SRMIST Kattankulathur, Chennai, Tamilnadu
[3] Department of Information Technologies, SRMIST Kattankulathur, Chennai, Tamilnadu
**Corresponding Author Email: ramabi1976@gmail.com*

Abstract

Background/Objective: This study is performed for predicting the blood sugar level by machine learning classifier techniques and identifies the best techniques among the techniques. Diabetes mellitus known as Diabetes caused due to increase in the blood glucose level. This will be impacting the pancreas and will be affecting the body beta cells. The beta cells play the major role in converting the glucose into energy. If the blood glucose remain undiagnosed for longer tenure, can cause various health complications like cardiovascular diseases neuropathy, nephropathy, organ failures and Eye disease. Diagnosing the sugar level in the earlier stage can help to prevent and control the health issues and save our life. Methods/Statistical Analysis: There are many classifiers already available in the computerized system. This analysis is to compare and identify the best classification technique between the classifier techniques like Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Classification and Regression Tree (CART), K – Nearest Neighbour (K-NN) and C5. Findings: In 2017 the International Diabetes Federation had anticipated that 50% of the people in the world with the age group of 20 to 79 are not worried about their Diabetes and not aware that they may be impacted by Diabetes in near future. It also explained that 76.5% of Diabetic patients are from very low-income countries. These are not yet diagnosed if you really speaking that they don't know that they have Diabetes. So, there is a need to diagnose and monitor the diabetes to support and help them to cure it. Keeping that in mind there are various classification techniques used to predict the diabetes. The main objective of this study is to make a comparative study and identify the best classifier which is consistent among the various datasets. The datasets used for this comparative study is University of California, Irvine machine learning repository (UCI) and PIMA Indian Diabetes Dataset. Application/ Improvements: In this study, the R-Studio application tool is used for developing and comparing the classifiers.

Keywords: *Diabetes, Statistical Classifiers, Decision Tree Classifier, Support Vector Machine, Classification and Regression Tree, Naïve Bayes Classifier, K-Nearest Neighbour, C4.5, C5.*

1. Introduction

Data mining is the statistical methodology that applied in the large and complex data or database which involves the data clustering, data classification and discovers the hidden patterns. There are many researches happening in the Artificial Intelligence that tries to extract the hidden patterns in the Biomedical and Healthcare domains. The one of the major life killing disease is Diabetic, caused due to various factors, like the type of food that we eat, the changes in the life style, the time that we sleep, less physical activity and Family history. The Diabetes can be classified as three major types, they are Type I Diabetes, Type II Diabetes and Gestational Diabetes Mellitus.

1.1 Type I Diabetes

The Type I diabetes are caused by autoimmune reaction, meaning that the body immune system affects the beta cells. This affects in predicting the level of insulin. The major symptoms of the Type I diabetes are Abnormal Thirst and Dry Mouth, Sudden Weight Loss, Frequent Urination, Lack of Energy and Fatigue, Frequent Hunger and Blurred Vision.

1.2 Type II Diabetes

The Type II diabetes are common across all the people. Almost 90% of the diabetes fall under this type. The Hyperglycaemiais the main reason for this type of diabetes. The body refuses to produce the adequate insulin is the cause for this type of diabetes. Mostly this type of diabetes will affect the older people, however now a day young adult and children also getting this type of diabetes due to change in life style and type of food that they eat. The symptoms of Type II diabetes are Excess Thirst and Dry Mouth, Frequent Urination, Lack of Energy and Tiredness, Blurred Vision and Numbness in Hand and Foot.

1.3 Gestational Diabetes Mellitus (GMD)

The pregnant women with increased in blood glucose level are classified under Gestational Diabetes Mellitus. Usually when women get pregnant second or more times this can happen. The babies born to the women having GMD have the higher chances of getting the Type II Diabetes.
The Knowledge on diabetes help to identify the factors that can be used to cluster and classify the data from UCI and PIMA Indian Dataset. There are various Glucose meters available in the market that identify the glucose level using various classifiers.
This study is to make a case study and compare the classifiers like Naïve Bayes, Support Vector Machine, Decision Tree,

Classification and Regression Tree, K- Nearest Neighbour and C5 and identify the best and consistent classifier among them.

2. Related Works

Machine learning algorithms are used in various predictions like Heart attack, Tumour, Diabetes and Blood Pressure in health care domain. Artificial Neural Network (ANN) is helping to transform the medical diagnose through computer systems.

- ✓ Classification of data
- ✓ Recognizing the patterns
- ✓ Eliminating the unwanted data

2.1 Naïve Bayes Classifier

This classification technique is based on the Bayes theorem, it is one of the inductive learning algorithm for data mining. In 2014 Bum Ju Lee and Jong Yeol Kim made an analysis to identify the Type II Diabetes using the factor Anthropometry and Triglycerides. A detailed study has been made using eleven thousand nine hundred and thirty-seven subjects among those four thousand nine hundred and six subjects were male and seven thousand and thirty-one subjects were female in the age limit of 31-80. To diagnose the Type II diabetes and Hypertriglyceridemia of all the subjects the factors used were Fasting Glucose Plasma (FGP) and Triglyceride (TG). The main objective of the study was to identify the relationship with the Hypertriglyceridemia Waist Phenotype and Type 2 diabetes and predict the relationship between the other phenotypes and Type 2 diabetes using the anthropometric measurement and Triglyceride for Korean Adults. The outcomes of the analysis explained that the Hypertriglyceridemia Waist phenotype was having the best association with the Type II diabetes even after making the necessary pre-processing on the site and age.

2.2 Decision Tree (DT)

This is one of the best classification used to make the decision. This works fast and derive the IF-THEN decision tree. The input will be any one of the factors like Age, Blood Pressure, Glucose Level, Insulin Dosage, Family History so on. The output will be the decision like Positive or Negative. The Decision Tree classification technique uses the classifiers like Classification and Regression Tree called as CART, the classification technique of C4.5, the QUEST known as Quick Unbiased Efficient Statistical Tree and the CHAID known as Chi-Squared Automatic Interaction Detector. In 2015 Ayush Anand and Divya Shakti used the CART algorithm to predict the diabetes with 75% accuracy. The blood pressure, eating junk foods, sleeping late hours, family history and less physical activity are the major factors used for the prediction.

2.3 Support Vector Machine (SVM)

In 2010 the authors of Nahla H Barakat, Mohamed Nabil HBarakat and Andrew P Bradley have used the machine learning classifier the Support Vector Machine called as SVM and have used the additional explanation module known as the Black Box Model. After the detailed analysis found that the prediction based on the accuracy and specificity as 94% and based on sensitivity as 93%.

3. Collection of Data

There are lots of analysis done in predicting the diabetes. The related works also explained various classification algorithms used to identify the features that affects the blood sugar level. Before using the classification with or without applying the pre-processing gives the accuracy with 70% to 80%. This practice includes the various techniques like Data Collection, Pre-Processing, Clustering and Classification.

In numerous analysis the authors were using various datasets like KHGES referred as Korean Health and Genome Epidemiology Study DB, Historical Electronic Medical Records (EMRs) from Canadian Primary Care Sentinel Surveillance Network (CPCSSN), University of California, Irvine (UCI) and PIMA Indian Datasets. The datasets from UCI and PIMA Indian datasets are reliable and many predictions has been made with the datasets.

The main features used in the PIMA datasets are

Number of pregnancy till date

- ✓ BMI Known as Body Mass Index measured as Kg/m^2
- ✓ Skin Width or Fatness measured in mm
- ✓ Oldness of the Patient
- ✓ BP known as Blood Pressure measured in mm hg
- ✓ Blood Glucose level (Based on 2 Hours Oral Glucose Test)
- ✓ Serum Insulin – 2 Hours (mu U/ml)
- ✓ Diabetes Pedigree Function (DPF)
- ✓ The Result – 0 for Negative Result and 1 For Positive Result

The sample dataset used for this study from PIMA is shown in the Fig 1

Number of Pregnancy	Glucose Level	Blood Pressure (BP in mm hg)	Body Mass Index (BMI)	Fatness of the Skin	Serum Insulin 2 Hrs (mu U/ml)	Diabetes Pedigree Function	Age of the Patient	Output Variable (0-Negative/1 Positive)
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1

Fig. 1: Sample PIMA Dataset

The control values of the subject and its Normal limit and Diabetes values used are shown in the Fig 2

Parameter	Diabetic value for men	Normal value for men	Diabetic value for women	Normal value for women
Body Mass Index	25	24	25.3	23.8
Oldness of the Patient	58.5	55.8	61.4	54.4
Fasting Plasma Glucose	138	93.4	141	91.2
Patient Weight	69	65	60	58
Circumference of Waist	89.9	85	88.3	83.59
Triglycerides	184.1	146.5	158.2	119.7
Number of time Patient get Pregnant	-	-	>2	2
Hemoglobin A1c Test	> 5.7%	4 - 5.6%	> 5.7%	4 - 5.6%
Blood Pressure - Systolic	>125	120- 125	>124	118-124
Blood Pressure - Diastolic	>81	80-81	>78	76-78

Fig. 2: The control values of the subject

To handle the missing values of the attributes, we keep the values unchanged. In case the major attributes like Blood Pressure, BMI and DPF is zero or null the whole example will be removed.

4. Pre-Processing Methods

4.1 Discretize

This method decreases the large attributes into small number of intervals and label them which are
used in classification or association of the data complexity. This process can be executed through below major steps.

- ✓ First sort the continuous values of the features to be discretized
- ✓ Evaluate an interval to split
- ✓ Merge the intervals of the continuous values
- ✓ Stop the process.

4.2 Principle Component Analysis

With large volume of data the distribution of the matrix will be too large to study and understand correctly. So, we need to apply the correlation between the variables to consider. As the data are more projections than clatter, we need to remove the clatter data and new set of data to be used to improve the data variability.
This is calculated using the covariance of the matrix X the dataset

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_p \end{pmatrix}$$

The data variability is determined using the covariance of the matrix X using the formula

$$\mathbf{var}(\mathbf{X}) = \sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

This can be used in the R-Studio program using the pre-defined function as below

```
> cov(Glucose, Pregnancies)
[1]  13.94713
>
```

5. Classification Methods

5.1 Support Vector Machine (SVM)

This is one of the most commonly used data mining algorithm which comes under the supervised learning. The major functionality and usage of SVM is, avoiding the overfit of the dataset and improves the correctness of the prediction techniques. This method separates the data using the linear hyperplanes such that the data space divided into segments and every segment contains only one attribute of data. Here the linear segments will be positive or negative. This can be used in where the data is non-regulatory means that the distribution of data is unknown. We can use the SVM method available in the R-studio to predict the blood glucose.

5.2 Naïve Bayes Classifier (NB)

We also consider the Naïve Bayes classification using the same set of pre-processed and discretized data as input. This classification is one of the mountable classification algorithm. This will consider all the attributes as independent each other, so that it can evaluate the conditional probability. Assume that the PIMA dataset has M attributes and c class labels
M= $\{m_1, m_2, m_3 \ldots\ldots m_n\}$ where n is the total number of datasets available in the PIMA Indian datasets.
Then the Naïve Bayes conditional independence is derived using the formula
$P(M|C=c) = \pi^{n}_{i=1} P(M_i|C_{=c})$
As an alternative to calculate each combinations of M, Find the M_i given c. After the pre-processing data the prediction probability can be extracted using the R-Studio by the formula

$$P(C|M) = \frac{P(C)\ \pi^{d}_{i=1} P(M_i|C)}{P(M)}$$

Sample R Program Shows the Summary of Naïve Bayes Classification

```
> summary(model)

Call:
glm(formula = Outcome ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4366  -0.7741  -0.4312   0.8021   2.7310

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.3461752  0.8157916 -10.231  < 2e-16 ***
Pregnancies              0.1246856  0.0373214   3.341 0.000835 ***
Glucose                  0.0315778  0.0042497   7.431 1.08e-13 ***
Insulin                 -0.0013400  0.0009441  -1.419 0.155781
BMI                      0.0881521  0.0164090   5.372 7.78e-08 ***
DiabetesPedigreeFunction 0.9642132  0.3430094   2.811 0.004938 **
Age                      0.0018904  0.0107225   0.176 0.860053
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 700.47  on 539  degrees of freedom
Residual deviance: 526.56  on 533  degrees of freedom
AIC: 540.56

Number of Fisher Scoring iterations: 5

>
```

Fig. 3: Summary of Naïve Bayes classification using R

5.3 Decision Tree

The Decision Tree is one is also one of the best classification technique which can be applied with the pre-processed dataset along with the missing or errored dataset. The major steps involved in the Decision Tree for predicting the diabetes are

- Find the base cases of the diabetes factor
- Study each diabetic factor and find out gain
- Find the factors that has the evidence gain
- Allocate the factor as root node
- Form the sub nodes as the children of the root node

The R-Studio is used to construct the Decision Tree once the datasets are pre-processed. If the BMI is considered as the root node, the BMI value is greater than the threshold value then we predict as positive and if the BMI value is below the threshold limit then the prediction will be negative.

The Confusion Matrix of DT Implementation is shown below

```
> confusionMatrix(test_pred, dt_testing$Result)
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative      151       52
  Positive       24       41

               Accuracy : 0.7164
                 95% CI : (0.6584, 0.7696)
    No Information Rate : 0.653
    P-Value [Acc > NIR] : 0.016031

                  Kappa : 0.3268
 Mcnemar's Test P-Value : 0.001954

            Sensitivity : 0.8629
            Specificity : 0.4409
         Pos Pred Value : 0.7438
         Neg Pred Value : 0.6308
             Prevalence : 0.6530
         Detection Rate : 0.5634
   Detection Prevalence : 0.7575
      Balanced Accuracy : 0.6519
```

After training the Decision Tree classifier with criterion as gini index, the confusion matrix received as below

```
> test_pred_gini <- predict(dtree_fit_gini, newdata = dt_testing)
> #check accuracy
> confusionMatrix(test_pred_gini, dt_testing$Result )
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative      151       52
  Positive       24       41

               Accuracy : 0.7164
                 95% CI : (0.6584, 0.7696)
    No Information Rate : 0.653
    P-Value [Acc > NIR] : 0.016031

                  Kappa : 0.3268
 Mcnemar's Test P-Value : 0.001954

            Sensitivity : 0.8629
            Specificity : 0.4409
         Pos Pred Value : 0.7438
         Neg Pred Value : 0.6308
             Prevalence : 0.6530
         Detection Rate : 0.5634
   Detection Prevalence : 0.7575
      Balanced Accuracy : 0.6519

       'Positive' Class : Negative

>
```

5.4 K- Nearest Neighbour (K-NN)

The K-NN algorithm is the simple algorithm which supplies all presented cases and classifies new case by majority vote of its K neighbours. This technique converts the unlabelled data points into well-defined groups.

Choosing the number of the nearest neighbour means finding the k value plays the important role in this classifier. The selection of the k value is the one determines how the dataset can be used to generalize the results of K-NN algorithm. If the k value is large which can give benefit of reducing the variance due to noisy data. In the Diabetic prediction we divide the data into two portions with the ratio of 65:35 here 65% are considered as training dataset and 35% are considered as test dataset. The major steps involved in the K-NN are

- ✓ Collect the data
- ✓ Prepare and explore the dataset
- ✓ Normalize the dataset
- ✓ Create Training and Test dataset
- ✓ Training the model on the Diabetes data
- ✓ Evaluate the model performance

The KNN- Cross Table Implementation using R

```
> CrossTable(x = dbs.testLabels, y = dbs_pred, prop.chisq=FALSE)


   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  246


                | dbs_pred
dbs.testLabels  |  Negative |  Positive | Row Total |
----------------|-----------|-----------|-----------|
       Negative |       127 |        33 |       160 |
                |     0.794 |     0.206 |     0.650 |
                |     0.713 |     0.485 |           |
                |     0.516 |     0.134 |           |
----------------|-----------|-----------|-----------|
       Positive |        51 |        35 |        86 |
                |     0.593 |     0.407 |     0.350 |
                |     0.287 |     0.515 |           |
                |     0.207 |     0.142 |           |
----------------|-----------|-----------|-----------|
   Column Total |       178 |        68 |       246 |
                |     0.724 |     0.276 |           |
----------------|-----------|-----------|-----------|
```

6. Conclusion

As part of this study a comparison will be made and identify the best classification technique which can predict the blood glucose with more accuracy. Here the R-Studio will be used, as it has all the classification methods considered for this study. Various pre-processing will be applied in the PIMA Indian dataset and UCI dataset to pre-condition the data to compare the classifiers. This study will be applied in the various pre-conditioned datasets and identify the best features which cause for the blood glucose level increase. The same will be applied in various datasets to ensure the consistency of the result.

References

[1] Veena Vijayan V, Anjali C "Decision Support Systems for Predicting Diabetes Mellitus – A Review" IEEE proceedings of 2015 Global Conference on Communication Technologies (GCCT 2015)

[2] R Bellazzi and B. Zupan "Predictive Data Mining in clinical Medicine:Current Issues and Guidelines", International Journal of Medical Informatics, Vol 77, pp 81-97, 2008

[3] Ning Wang, Guixia Kang, "Monitoring System for type 2 diabetes mellitus", IEEE conference on e-Health Networking, pp 62-67, 2012

[4] Sonu Kumari, Archana Singh, "A data mining approach for the diagnosis of diabetes mellitus", IEE conference on Intelligent systems and control, pp-373-375, 2013

[5] Rakesh Motka, Viral Parmar, "Diabetes Mellitus Forecast Using Different Data Mining Techniques", IEEE International Conference on Communication and Computer Technology (ICCCT), 2013

[6] Abid sarvwar, Vinod Sharma, "Intelligent Naïve Bayes approach to Diagnose Diabetes Type 2" Special issue of International Journal of Computer Applications on Issues and Challenges in Networking, Intelligence and Computing Technologies, November 2012

[7] Veena Vijayan V, Aswathy Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus" International Journal of Computer Applications, Vol 94 pp 12-16, June-2014

[8] C. Kalaiselvi, G.M Nasira, "A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS", IEE Computing and Communicating Technologies, pp-188-190, 2014

[9] Nirmala Devi M, Appavu alias Balamurugan S, Swathi U.V, "An Amalgam KNN to predict Diabetes Mellitus", IEEE International Conference on Emerging Trends in Computing, Communication and Nano technology (ICECCN), pp 691-695, 2013

[10] Velu C.M, K.R Kashwan, "Visual Data Mining Technologies for Classification of Diabetic Patients", IEEE International Advance Computing Conference (IACC), pp 1070-1075, 2013

[11] Asma A, AlJarullah, "Decision Discovery for the Diagnosis of Type II Diabetes", IEEE Conference on Innovations on Information Technology, pp 303-307, 2011

[12] Xiaoran Zhang, Ruoyu Ding, "Predicting Patients with Diabetes Type II from HER data" Computer Science Engineering.

[13] I. Kononenko, "Machine Learning for Medical Diagnosis: History State of the Art and Perspective", Artificial Intelligence in Medicine, Vol 23, No 1, pp 89-109, 2001

[14] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining" Pearson Education, Inc 2006

[15] Vinod Chandra S.S, Anand Hareendran S, "Artificial Intelligence and Machine Learning" PHI Learning Private Limited, Delhi 110092, 2014

[16] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat, "Intelligible Support Vector Machine for Diagnosis of Diabetes Mellitus", IEEE Transactions on Information Technology in Biomedicine, Vol 14, No 4, July 2010

[17] Krzysztof J. Cios, J William Moore, "Uniqueness of Medical Data Mining" Artificial Intelligence in Medicne Journal pp 1-19, 2002

[18] DeFronzo RA, Ferrannini E, Zimmet P, et al. International Text Book of Diabetes Mellitus 2 Volume Set, 4th Edition, Wiley Blackwell, 2015, IDF - DIABETES ATLAS Eighth edition 2017, pp 16-17, 2017

[19] Bum Ju Lee, Jong Yeol Kim, "Identification of Type II Diabetes Risk Factor Using Phenotypes Consisting of Anthropometry and Triglycerides Based on Machine Learning" IEE Journal of Biomedical and Health Informatics pp 4-7

[20] Paul Teetor, "R Cookbook", Sixth Indian Reprint, pp 195-311, August 2014

[21] Nina Zumel, John Mount, "Practical Data Science With R" Edition 2014, pp 175-251. 2016, Reprint Edition 2016.