IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 5, March 2019 E-ISSN: 2347-2693

# Type 2 diabetes mellitus prediction model based on ensemble boosting method with Principal Component Analysis

**M.Sornam[1*], M.Meharunnisa[2]**

[1] Department of Computer Science, University of Madras, TamilNadu, India
[2] Department of Computer Science, University of Madras, TamilNadu, India

[*]*Corresponding Author: madasamy.sornam@gmail.com, Tel.: +91-9841166535*

DOI: https://doi.org/10.26438/ijcse/v7si5.124130 | Available online at: www.ijcseonline.org

***Abstract***— In the recent years, data mining has been employed in the medical field for extracting and manipulating information, and aids within the higher process. There is a growing want for the medical establishments to be extra suggested and knowledgeable concerning the diseases and to understand the risk factors before diagnosis. Predicting the results of a process with a high level of accuracy is a difficult task. In this study we took the advantage of the data mining models to predict the Type – 2 Diabetes mellitus. The benchmark dataset, "Pima Indian Diabetes" dataset is used for this study. The main objective of this study is to propose the extensive data pre-processing such as imputation of missing values and a feature engineering technique namely 'Principal Component Analysis' are used to transform the dataset into a compressed form. Ensemble or classifier combination method called boosting method such as Gradient boosting machine and Random Forest are used. The most downside that's attempting to be resolved isn't solely to extend the accuracy however additionally to retain all the information in the data set while not removing the missing data. The missing data are imputed by a method called 'predictive mean matching'. The results show that the ensemble learners, once used alongside PCA attained 100% accuracy of prediction. Moreover, it ensures that no missing information must be removed and might be imputed to confirm the data quality is enough. As a result, the model is shown to be helpful for the real time prediction of Diabetes Mellitus.

***Keywords***—Ensembles,Gradient boosting machine,Random Forest,Principal Component Analysis.

## I. INTRODUCTION

The global escalation of diabetes in developed and developing nations poses a great health challenge. Diabetes mellitus (or diabetes or DM) is a chronic disease, deep rooted condition that influences our body's capacity to utilize the vitality found in food. There are three noteworthy types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes. All the types have a commonality, i.e., body separates the sugars and starches into an uncommon sugar called glucose. Glucose powers the cells in our body. But the cells require insulin, a hormone, in our circulation system keeping in mind the end goal to take in the glucose and utilize it for vitality. With DM, either our body doesn't make enough insulin; it can't utilize the insulin it produces, or a mix of both [1].

As indicated by a recent report by Indian Council of Medical Research, states with higher per-capita total national output (GDP) have a higher pervasiveness of diabetes. As per the November 2017 report by the council, diabetes prevalence has increased by 64% across India [2]. The quantity of individuals with diabetes worldwide has dramatically increased amid the previous 20 years. A standout amongst the most stressing highlights of this fast increment is the rise of Type 2 diabetes in children, adolescents, and young adults [3]. In order to lower the influence of DM in the society, it is important for us to identify the high risk group of people with DM. In order to identify the high risk group, we need to utilize the latest data mining methodologies which help to predict them. What are required are techniques that allow the doctors to distil the most valuable information from mountains of accumulated data. The field of data mining provides such techniques. Data mining or knowledge discovery in databases (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for handling a large amount of information, including missing information and imbalanced datasets. Data mining strategies might be utilized for littler measures of information, yet the bigger the information the better the chance of discovering something novel and intriguing. The relationship between data mining and machine learning is not new and has existed for decades. Machine learning algorithms have been successfully applied to many healthcare applications such as diagnosis of breast cancer [22], ovarian disease, leukemia, brain cancer [23], lung cancer [24], and lymphoma.

 

Data pre-processing including data cleaning may be needed. In some cases, the sampling of data and testing of various hypotheses may be required before data mining can start. Data mining is neither a simple nor an inexpensive process that anyone with a database can carry out. The following sections deals with the present work in data pre-processing including PCA. Section 3 explains the proposed model. Section 4 will be devoted to exploratory data analysis. Section 5 discusses about the performance of classifiers and its results.

## II. LITERATURE REVIEW

In recent years, data mining technique has been used extensively within the medical field to predict the possibility of the disease.

One of the commonly used dataset for the research is Pima Indians Diabetes dataset from the University of California at Irvine from UCI machine learning database. Ahmad [4] studied the prediction accuracy of multilayer perceptron in neural network against the decision tree based algorithm such as ID3 and J48 algorithm. During data pre-processing, the dataset is transformed through generalization process in which low level concept are transformed into higher level concept. The dataset is pre-processed using min-max normalization. The results showed that J48 performed with higher accuracy of 89.3%. The prediction accuracy is improved to 89.7% after pruning number of times pregnant attribute.

Chen [5] proposed a model, in which all the impossible and missing values are replaced by mean and then used K-means clustering algorithm to remove incorrectly classified sample. Finally, 532 out of 768 samples are left, which are then classified using J48 algorithm. The classification accuracy obtained was 90.04%.Yang et al., [6] developed a double level algorithm using K-means to remove incorrectly clustered data. After the removal procedure, only 589 samples were left which are then classified using Logistic regression classification algorithm with the accuracy of 95.42%. The missing and incorrect values are replaced by means from the training data.

Kavakiotis et al [7] studied a systematic review of application of data mining and machine learning techniques in the field of diagnosis of diabetes. It concluded that 85% of the methods used are characterized by supervised learning and 15% were unsupervised ones. Specifically, the most widely used Support vector machines are the successful approach. Choubey [8] proposed the genetic algorithm which requires selection, crossover and mutation to select the attribute on PIDD. The selected attributes are then used and classified using Naïve Bayes classifier with the accuracy of approximately 77%.With genetic algorithm , the attributes are reduced from eight to three. Dewangan and Agrawal [9] proposed an ensemble model by combining Bayesian classification and multilayer perceptron. Information gain feature selection technique is used to rank the features. The accuracy of 81.89% is achieved with 6 feature subset after removing diastolic blood pressure and diabetes pedigree function features.

Zehra and Asmavathy [10] used the process of data discretization followed by data pre-processing, in which they removed all the instances which had the value of zero. They compared the accuracies of non-pre-processed and pre-processed data, which showed that the classification accuracy increases in pre-processed data.Vijayan.V [11] reviewed a pre-processing technique named principal component analysis (PCA) and discretization, which removed unwanted raw data and attributes. The classifiers used are SVM, Decision tree and naïve Bayes, where the accuracy level of Naïve Bayes and Decision tree are increased, but for SVM it decreases. Sowjanya [12] had developed a mobile application, MobDBTest which acts as a tool to predict the probability of diabetes, developed using decision tree machine learning algorithm. The classification rules extracted are integrated into the smart phone android application.

Gang [13] proposed a diabetes risk assessment model based on mobile devices to predict 5-year risk of diabetes. Logistic regression analysis was used to screen the diabetes. No pre-processing has been made. In summary, the related work shows that model established for Diabetes mellitus prediction used classification techniques. The most commonly used pre-processing techniques in the models were removing the instances that has missing values or irrelevant information, replacing those information with median or mean value of the training set. We need to propose a novel method for prediction while not removing any instances, which can successively, ends up in information loss. Therefore we chose principal component analysis for pre-processing the data, which supplies principal components. The principal components are used along with the ensemble learners GBM and random forest to predict DM.

## III. PROPOSED MODEL

The proposed model comprises of following steps:

1. Pima Indian diabetes dataset is taken from UCI machine learning repository.
2. Outliers and missing data are handled and imputed.
3. PCA is used for data pre-processing.
4. PCs are used as inputs to the ensemble boosting algorithm.
5. The performance of the boosting classifier can be evaluated by its accuracy.
6. Evaluate the performance of the classifier by measuring its accuracy.

 

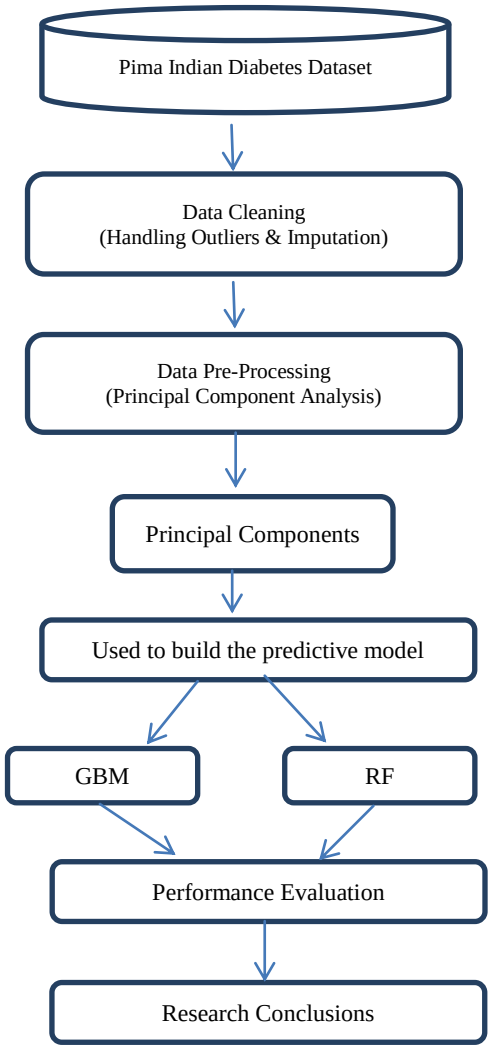


Figure 1. *The Proposed Model*

Figure 1. shows the diagrammatic representation of the proposed model.

## IV. EXPLORATORY DATA ANALYSIS

Exploratory data analysis refers to the critical process of initi al data research in order to detect patterns, detect anomalies, test hypotheses and verify assumptions using summary statistics and graphical representations.

### *A. Data Exploration*

The Pima Indian Diabetes dataset was obtained from UCI machine learning repository. It consists of 768 rows and 9 columns. Each instance consists of 8 predictor variable and 1 class variable, which is independent. There are 500 instances which are tested negative for diabetes and 268 are tested positive.

Table 1. Description of the dataset

| Features | Description |
|---|---|
| Pregnancies | Number of times pregnant |
| Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| Blood Pressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index (weight in kg/(height in m)^2) |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Age(Years) |
| Outcome | Class Variable (0 or 1) |

### *B. Data Cleaning*

*a)Unexpected Outliers*

When analysing the data, we can identify that there are some outliers in some columns.

**Blood Pressure:** By observing this attribute, it is found that there are 35 counts where the value is 0, which seems wrong because a living person cannot have diastolic blood pressure of zero. The wrong interpretation may be due to the malfunctioning of the instrument [14].

**Glucose Level:** After fasting, glucose level would not be as low as zero. Therefore zero is an invalid reading; it is found that there are 5 counts where the value is 0.

**Skin Fold Thickness:** The skin fold thickness cannot be less than 10mm for a normal people. The skin thickness for a new born on the first day of life is found be approximately 172.4 µm [15]. Total count of attribute where the value 0 is 227.

**BMI:** Unless the person is really underweight which could be life threatening, BMI should not be 0 or close to zero. It is found that there are 11 counts where the value is 0.

Insulin: A living person cannot have zero insulin except in a rare situation [16]. But the dataset consists of 374 counts where the value is 0.

*b) Handling Outliers*

There are several ways to handle invalid data values which may be ignore/remove the cases which would mean losing valuable information, average/mean values which would sometime send a wrong signal to the model, avoid using the features with lot of invalid values for the model, but it's sometimes hard to predict. The most commonly used method for handling the missing data are deleting the instance itself, which results in data loss and bias. Another way is to replace with a constant such as average or mean, but it may not be an optimal way to handle missing data.

 

Table 2. Summary of the missing data

| Features | Percentage of Missing data |
|---|---|
| Glucose | 0.65% |
| Blood pressure | 4.5% |
| Skin Thickness | 29.5% |
| Insulin | 48.6% |
| BMI | 1.4% |

From Table 2, it is clear that 0.65% of data from the glucose attribute, 4.5% of data from the attribute blood pressure, 29.5% of data from the attribute Skin Thickness, 48.6% of data from the attribute Insulin and 1.4% of BMI data are missing.

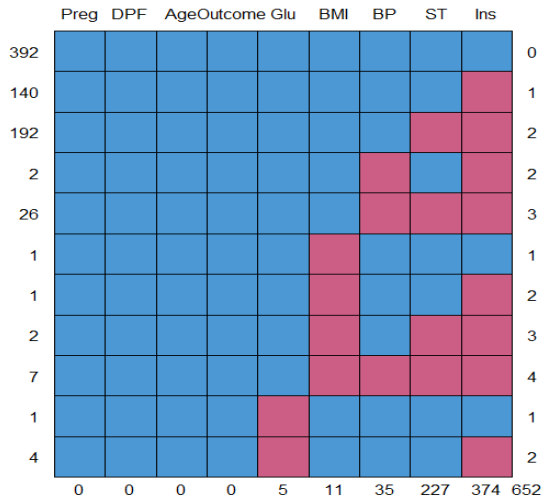


Figure 2. *Visual representation of missing data*

Figure 2. shows the diagrammatic representation of the missing data.

*C) Imputation*

The missing values are imputed using MICE (Multivariate Imputation via Chained Equation). It imputes data on a variable by variable basis by using linear regression to predict continuous values and logistic regression to predict categorical missing values. Predictive mean matching is a technique used for numeric variables [17].

## V. FEATURE ENGINEERING – PRINCIPAL COMPONENT ANLYSIS

The goal of PCA is to find a new set of attributes that better captures the variability of the data. Principal components are linear combination of normalized variables. Since Principal components are orthogonal to each other, the correlation between them becomes zero. This solves the problem of multi collinearity when there is more number of highly correlated independent variable. This technique is based on some mathematical concepts such as standard deviation, variance, covariance, Eigen vectors and Eigen values.

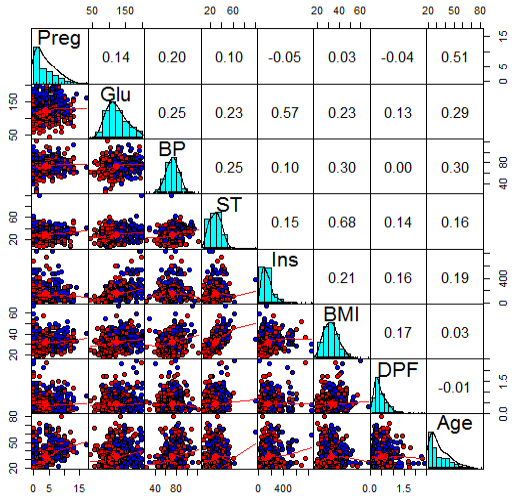


Figure 3. *ScatterPlot and Correlation among the variables.*

The Figure 3 shows the correlation among the independent variables. The correlation among Pregnancy and Glucose, Skin thickness and BMI are more than 0.5.While developing the predictive model, high correlations among independent variables lead to "Multicollinearity" problem. The estimates for the model become unstable. The solution to this problem is using Principal component analysis.

Table 3. PCA Analysis Report

| Principal Components | Std.dev | Proportion of Variance | Cumulative Proportion |
|---|---|---|---|
| PC1 | 1.5692 | 0.3078 | 0.3078 |
| PC2 | 1.2242 | 0.1873 | 0.4951 |
| PC3 | 1.1170 | 0.1560 | 0.6511 |
| PC4 | 0.9534 | 0.1136 | 0.7647 |
| PC5 | 0.85662 | 0.09172 | 0.85644 |
| PC6 | 0.69298 | 0.06003 | 0.91646 |
| PC7 | 0.62604 | 0.04899 | 0.96546 |
| PC8 | 0.52570 | 0.03454 | 1.00000 |

As can be observed from Table 3, PCA analysis reports as many PCs as the number of independent variables in the dataset. The standard deviation of each component is shown in the second column of the table. The last column shows the cumulative proportion of variance. It appears that first six PCs account for 91% of the total variance.

 

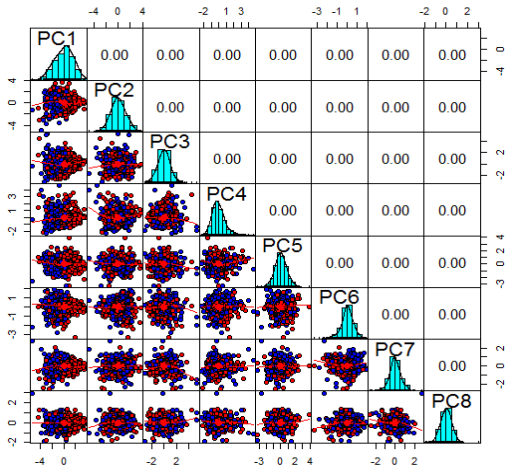


Figure 4. *Correlation among the PC's*

Figure 4 shows that PCs are used to remove Multicollinearity problem.

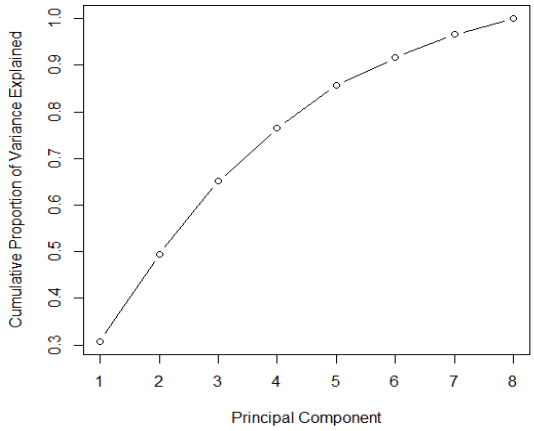


Figure 5. *Screeplot representing the variances of all principal components.*

Typically a few of principal components clarify a high measure of the variation and some of them should be chosen. Figure 5, the scree plot is used to make this decision. The point where the slope of the curve is clearly leveling off indicates the number of PCs that should be taken for the analysis. Here, the plot starts to stabilize from PC6 onwards, indicating that first six PCs or five PCs collect most of the total variance.

## VI. BOOSTING ALGORITHMS

Ensemble techniques are used to improve the accuracy by aggregating the prediction of multiple classifiers.
Ensemble methods work better with unstable classifiers, i.e., base classifiers that are sensitive to minor noise in the training set. Examples of unstable classifiers include decision trees, rule-based classifiers and artificial neural network [18]. The different methods involved are boosting, bagging and stacking [19].

### *A. Gradient Boosting Machine (GBM)*

Boosting is an iterative procedure used to adaptively change the distribution of training examples so that base classifiers will focus on instances that are hard to classify. It assigns a weight to each training instance and may adaptively change the weight at the end of each round.

GBM works on a loss function that is to be optimized, a weak learner such as decision tree are constructed in a greedy manner to choose the best split points based on gini index. The tree can grow up to 8 levels. The trees are added up to one at a time and existing trees in the model are unchanged. After calculating the loss or error, the weights are updated to minimize the error.

The result for the new tree is then added to the result of the current succession of trees with a goal to enhance the final result of the model. A countable number of trees are included or learning stops once loss achieves a satisfactory level or never again enhances a validation dataset from outside. The performance of GBM can be improved with regularization [20].The results of our experiment (Table 4) shows that GBM needs minimum six principal components to achieve 100% accuracy.

### *B. Random Forest*

Random Forest is a class of ensemble methods specifically designed for decision tree classifiers. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors, which is generated from a fixed probability distribution [21]. The results of our experiment (Table 4) shows that RF needs minimum five principal components to achieve 100% accuracy.

### *C. Analysis of the result*

There were two phases of experiment for this study: (1) training phase – 75% of dataset. (2) Testing phase – 25% of dataset. One common measure discussed in the literature is accuracy, which is defined as correctly classified instances divided by total number of instances.

$$\text{Accuracy} = \frac{TP(True\ Positive) + TN(True\ Negative)}{TP + TN + FP(False\ Positive) + FN(False\ Negative)} \quad (1)$$

 

Table 4. Summary of experiment results on GBM and RF with PCs and without PCs

| Ensemble Learner | Accuracy with No PC's | Principal Component | Accuracy with PC's |
|---|---|---|---|
| GBM | 76.06% | PC1-PC4 | 75% |
| | | PC1-PC5 | 76% |
| | | PC1-PC6 | 100% |
| Random Forest | 77.13% | PC1-PC4 | 72.8% |
| | | PC1-PC5 | 100% |
| | | PC1-PC6 | 100% |

Table 4 shows that with the first six principal components, GBM shows 100% accuracy i.e., with 91.6% variance in the dataset whereas random forest shows 100% accuracy with first five principal components i.e., with 85.6% variance in the dataset.

Table 5: Comparison with existing experiments.

| Method | Accuracy | Reference |
|---|---|---|
| Proposed model | 100% | This Paper |
| J48 | 89.3% | Ahmad |
| K-means +J48 | 90.04% | Chen |
| K-means+ Logistic Regression | 95.42% | Yang et al., |
| NB +GA | 77% | Choubey |
| Multilayer Perceptron | 81.89% | Agrawal |
| HPM | 92.38% | Patil |
| GBM | 76.06% | R |
| Random Forest | 77.13% | R |

The Table 5 shows the accuracy results of our proposed model and also the accuracy results of others.

```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 379   0
         Y   0 201

               Accuracy : 1
                 95% CI : (0.9937, 1)
    No Information Rate : 0.6534
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.6534
         Detection Rate : 0.6534
   Detection Prevalence : 0.6534
      Balanced Accuracy : 1.0000

       'Positive' Class : N
```

Figure 6. *The results of the experiment in training set*

```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 121   0
         Y   0  67

               Accuracy : 1
                 95% CI : (0.9806, 1)
    No Information Rate : 0.6436
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.6436
         Detection Rate : 0.6436
   Detection Prevalence : 0.6436
      Balanced Accuracy : 1.0000

       'Positive' Class : N
```

Figure 7. *The results of the experiment in training set*

Figure 6 and Figure 7 shows the experimental results conducted on training and testing set.

## VII. CONCLUSION

In this paper, it is concluded that the proposed model contributed a lot to the prediction model without losing any information from the dataset due to outliers, noise and missing data. The proposed model also made use of ensemble learners which helps to create stable model. The proposed method can be very helpful to the physicians for their final decision on their patients as by using such an efficient model to make accurate decision. The same method can also be applied to other disease prediction such as heart disease, liver disease, and identification of breast cancer and so on.

## REFERENCES


[1] Ndisang, Joseph Fomusi, Alfredo Vannacci, and Sharad Rastogi. *"Insulin Resistance, Type 1 and Type 2 Diabetes, and Related Complications 2017."* Journal of diabetes research2017 ,2017

[2] Anjana, R.M., Deepa, M., Pradeepa, R., Mahanta, J., Narain, K., Das, H.K., Adhikari, P., Rao, P.V., Saboo, B., Kumar, A. and Bhansali, A., 2017. *Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR–INDIAB population-based cross-sectional study*. The Lancet Diabetes & Endocrinology,Vol. 5, Issue . 8, pp.585-596.

[3] Zimmet, P.Z., Magliano, D.J., Herman, W.H. and Shaw, J.E., 2014. Diabetes: a 21st century challenge. The lancet Diabetes & endocrinology,Vol. 2, Issue 1, pp.56-64.

[4] Ahmad, A., Mustapha, A., Zahadi, E.D., Masah, N. and Yahaya, N.Y., 2011. *Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus.* In Digital Information Processing and Communications ,Springer, Berlin, Heidelberg. pp. 537-545

[5] Chen, W., Chen, S., Zhang, H. and Wu, T., November. *A hybrid prediction model for type 2 diabetes using K-means and decision tree.* 8th IEEE International Conference on Software Engineering and Service Science, (ICSESS). pp. 386-390,2017


 

[6] Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., 2018. *Type 2 diabetes mellitus prediction model based on data mining. Informatics* in Medicine Unlocked,Vol. 10, pp.100-107.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., *Machine learning and data mining methods in diabetes research.* Computational and structural biotechnology journal,Vol. 15, pp.104-116,2017

[8] Choubey, D.K., Paul, S., Kumar, S. and Kumar, S., 2017, February. *Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection.* In Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016) pp. 451-455, 2017

[9] kumar Dewangan, A. and Agrawal, P., Classification of diabetes mellitus using machine learning techniques. International Journal of Engineering Appled Science, Vol.2, Issue.5, pp.145-148, 2015.

[10] Amatul, Z., Asmawaty, T., Kadir, A. and MAM, A., *A Comparative Study on the Pre-Processing and Mining of Pima Indian Diabetes Dataset*, 2013.

[11] Vijayan, V.V. and Anjali, C.,*Decision support systems for predicting diabetes mellitus—A Review*. Global Conference in Communication Technologies, IEEE, pp. 98-103,April 2015 .

[12] Sowjanya, K., Singhal, A. and Choudhary, C., 2015, June. *MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices.* IEEE International In Advance Computing Conference (IACC), pp. 397-402, 2015

[13] Gang, S., Shanshan, L. and Ding, Y., 2015, November. *Design and Implementation of Diabetes Risk Assessment Model Based on Mobile Things.* In,7th International Conference on Information Technology in Medicine and Education (ITME) pp.425-428, 2015.

[14] Choudhary, D., Suthar, O. P., Bhatia, P. K., & Biyani, G. *"Zero" diastolic blood pressure.* In The Indian Anaesthetists' Forum . Medknow Publications and Media Pvt. Ltd, Vol. 17, No. 1, pp. 32-32, January 2016.

[15] Vitral, G. L. N., Aguiar, R. A. P. L., de Souza, I. M. F., Rego, M. A. S., Guimarães, R. N., & Reis, Z. S. N. (2018). *Skin thickness as a potential marker of gestational age at birth despite different fetal growth profiles: A feasibility study*. PloS one, Vol. 13, Issue 4, e0196542, 2018.

[16] Gisela Wilcox, *Insulin and Insulin Resistance* , Clinic Biochem Rev ,Vol 26, pp 19-39, 2005.

[17] Buuren, S. van, and Karin Groothuis-Oudshoorn. *"mice: Multivariate imputation by chained equations in R."* Journal of statistical software pp. 1-68, 2015.

[18] Breiman, Leo. *"Bagging predictors."* Machine learning 24.2 (1996): 123-140.

[19] Valentini, G. and Masulli, F.,. *Ensembles of learning machines*,In Italian Workshop on Neural Nets Springer, Berlin, Heidelberg, pp. 3-20, May 2002.

[20] Freund, Y., & Schapire, R. E. *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of computer and system sciences, Vol. 55,Issue 1,pp. 119-139,1997

[21] Breiman, L. *Random forests. Machine learning*, Vol.45,Issue 1, pp. 5-32,2001.

[22] Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). *Prediction of benign and malignant breast cancer using data mining techniques.* Journal of Algorithms & Computational Technology, Vol.12, Issue 2, pp. 119-126.

[23] Singh, K., Lilhore, U. K., & Agrawal, N. *An Efficient Supervised Learning Technique for Tumour Detection and Analysis from MR Image Data Set*,2018

[24] Zhu, X. F., Zhu, B. S., Wu, F. M., & Hu, H. B. *DNA methylation biomarkers for the occurrence of lung adenocarcinoma from TCGA data mining*. Journal of cellular physiology, 2018.


## Authors Profile

Dr M. Sornam received her MSc in Mathematics from the University of Madras in the year 1987, Master's Degree in Computer Applications from the University of Madras in the year 1991 and received her Ph.D from the University of Madras in the year 2013. Since 1991–1996, she worked as a Lecturer in Computer Science at Anna Adarsh College, Chennai. Later, from 1996 to 2000 she worked as a Lecturer in Computer Science at T.S. Narayanasami College of Arts and Science, Chennai. Since 2001, she has been working in the Department of Computer Science, University of Madras. At present, she is working as an Associate Professor in Computer Science at the University of Madras. Her area of interest includes artificial intelligence and artificial neural networks , image processing, data mining, pattern recognition and applications.

*Mrs.M.Meharunnisa* pursed Bachelor of Computer Application in Justice Basheer Ahmed Sayeed College for Women from University of Madras, India in the year 2005, Master of Science in Computer Science, Justice Basheer Ahmed Sayeed College for Women for Women from University of Madras, India in the year 2013. She is currently pursuing Ph.D. in Department of Computer Science, University of Madras, India since 2016, and currently working as Assistant Professor in Department of B.C.A, Ethiraj College for Women,Chennai since 2013. Her main research work focuses on Data Mining,Machine learning and Deep learning.

 