# Performance Evaluation of Machine Learning Models for Diabetes Prediction

**Aishwarya Jakka, Vakula Rani J**

***Abstract : Diabetes is one of the prevalent diseases all over the world. As per the International Diabetes Federation (IDF) report of the year 2017, diabetes is prevalent in about 8.8% of the Indian adult population and is one of the top ten causes of death in India. In untreated and unidentified diabetes could cause fluctuations in the sugar levels and extreme cases, damage organs such as kidneys, eyes, and arteries in the heart. By using Machine learning algorithms to predict the disease from the relevant datasets at an early stage could likely save human lives. The purpose of this investigation is to assess the classifiers that can predict the probability of disease in patients with the greatest precision and accuracy. Experimental work has been carried out using classification algorithms such as K Nearest Neighbor (KNN), Decision Tree(DT), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest(RF) on Pima Indians Diabetes dataset using nine attributes which is available online on UCI Repository. The performance of classifier is evaluated based on precision, recall, accuracy and is estimated over correct and incorrect instances. The results proved that Logistic Regression (LR) performs better with the accuracy of 77.6 % in comparison to other algorithms.***

***Keywords: Classification; K-Nearest-Neighbor (KNN) Decision Tree (DT), Naïve Bayes (NB), Support-Vector Machine (SVM), Logistic Regression (LR) ,International Diabetes Federation (IDF).***

## I. INTRODUCTION

Globally, many chronic diseases are prevalent in developing and developed countries. Diabetes mellitus (DM) often known as diabetes, is one of the metabolic disorders which may cause blood sugar, by producing more significant or a smaller amount of insulin. Diabetes affects the different parts of the human body parts like eyes, kidneys, heart, and nerves [18]. Hence, the prevention and detection of disease in the early stages could likely save human lives. Most common are Type-I diabetes and Type-II diabetes. In Type I desease, insulin will not be produced by the human body and 10% of diabetes caused by this type. Type II diabetes is non-insulin-dependent diabetes caused because of not enough production of insulin by the pancreas or body cells are resistant to absorb it. Therefore, there is a need to prevent and diagnose the disease to save human life from an early death. Researchers have proved that data mining and machine learning models Decision Tree, Support Vector Machine etc. works better in diagnosing diseases [1], [22], [23].

**Revised Manuscript Received on September 05, 2019**.

***** Correspondence Author

**Dr.Vakula Rani J ***, Dept. of MCA, CMRIT, Bangalore, India. Email: vakula.r@cmrit.ac.in

**Aishwarya Jakka ,** Graduate student, Information Science at Pittsburgh, USA. Email: aishwarya.jakka@hotmail.com

Machine learning methods support researchers to retrieve new facts from large health-related data sets, which improves health care distribution, decision making and disease supervision. According to CDA (Canadian Diabetes Association) between 2010 and 2020, the expected increase of the disease would be from 2.5 million to 3.7 million [18]. The major aim of this study is to investigate the performance of various classification models, that can anticipate the probability of disease in patients with the greatest precision. One of the issues in bioinformatics investigation is to accomplish the right conclusion of certain important data. Multiple lab tests could complicate the primary identification procedure and which results in delay in the diagnosis, especially for a situation where numerous lab tests might be required. The present work is carried out based on the secondary data set taken from PIDD and machine learning classifiers are used to the predict presence of disease inpatient. The experimental results, accuracy, misclassification rate, precision, recall and F1-score of these classifiers are presented. This paper is organized into six sections. Section I and II briefs the Introduction and related work. Section III and IV describe the Methodology and Data set description. Section-5describes the classification models, results and its interpretation, the last section presents the conclusion and future work of this work.

## II. RELATED WORK

Praveen et al. [9] examine the importance of Bagging and Ada-boost strategies using J48 DT as base classifier [7] for identifying the presence of disease and also in light of its risk factors. Orabi et al. [8] proposed a diabetes disease prediction framework, which can predict the likeliness of diabetes at a specific age. The proposed strategy is structured dependent on the idea of applying a decision tree (DT). The results obtained were satisfactory for the prediction of the diseases at a specific age, with the highest precision using Decision tree model[11], [3]. Pradhan.P et. al. [2] have applied Genetic programming (GP) technique to train and test the PID data set for diabetes from the UCI store. Results accomplished by applying Genetic Programming [10], [20] gave ideal accuracy when compared with other strategies. There can be a noteworthy improvement in the accuracy by setting aside less effort for classifier age. Rashid et.al. [21] proposed an algorithm with sub-modules for the prediction of diabetes. Artificial—Neural-Network(ANN) was applied in the main algorithm. Decision Tree (DT)[4], [9] classifier is used to distinguish the indications of diabetes disease in patients.

Nongyao et al [6] proposed a classifying technique for the risk of diabetes mellitus. Author's have applied four techniques, in particular, DT, SVM, LR and NB. Experimentation results demonstrate that the LR gives ideal outcomes among others.

## III. METHODOLOGY

The Exploratory Data Analysis (EDA) is conducted to understand the various data sources and collected the contextualized Pima Indians Diabetes dataset from the UCI repository as step-1. Cleaned the dataset, that is, removal of the duplicate and missing values as a step of data preprocessing. Then, identified a set of candidate classification models for the experiment on the training dataset to fit a model. Assessed the models based on various performance metrics such as Accuracy, Recall, f1-Score, Misclassification Rate and ROC- AUC-Score. Compared the models on the various performance metrics to analyze the accuracy of a model for the dataset. Represented the performance evaluation of the candidate classification models based on the metrics graphically.

The proposed model diagram is presented in the below figure-1. This framework presents the sequence and flow of the experimental process followed in this work.

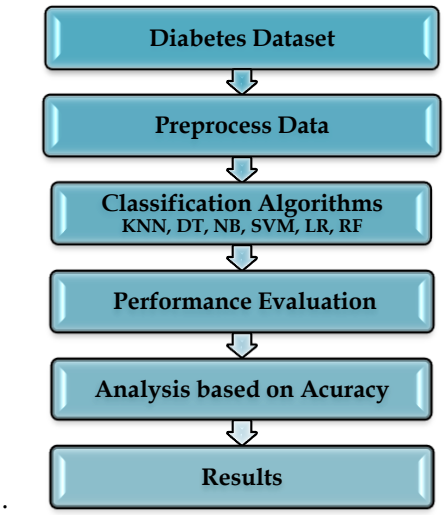


**Figure-1: Proposed Framework**

**Description of Classifiers**

### A. KNN Classification Technique

Supervised Learning model, K-NearestNeighbors is a simple technique used for regression and classification problems. This model stores every available state and classifies as the latest instance based on a similarity measure of distance function ' K' . These distance measure could be calculated by using Euclidean, Manhattan, Hamming, and Minkowski distance for categorical variables. The neighbours classify a data point by the KNN distance measured function and will be using average rather than voting from its nearest neighbours. Given, a points $X = (x_1,x_2,\ldots x_k)$, and $Y = (y_1,y_2,\ldots y_k)$, the Euclidean distance between them is given by equation [12]

$$Euclidean = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

The k is a positive integer and its value is guessed by observing the data set. These boundaries will segregate the 'k' value of each class. The boundaries separate the two classes with different values of 'k'. If the value of 'k' equal to 1, then this belongs to the class of neighbour which is nearest. One of the challenges in KNN modelling is choosing 'k' value. KNN classifier performs well on underlying recognition problems, and in spite of its straightforwardness, KNN gives excellent results for applications like data compression and economic forecasting. However, KNN can be computationally expensive for prediction in real-time applications where there are a large number of training examples and noisy data.

### B. DT Classification Technique:

Decision Tree (DT) classification technique is a machine learning model which can be used classification problems where the data is divide/split according to a specific parameter continuously. The primary goal of applying Decision Tree is to predict the object/target class using decision rules which are dawn from prior information. With the help of its internal and external nodes, it does the task of classification and prediction. Instance classification with different features could be done with the help of Root node and Internal nodes, whereas the leaf nodes represent classification. The selection of each node is done by evaluating the maximum information gain among all other features in each stage [5]. Decision Trees algorithm is comprehensible, resulting in a set of policies or rules followed the same as human decision makings. However, there is a high chance of overfitting in the Decision Tree, and prediction accuracy is lesser when compared to other classifiers.

### C. NB Classification Technique:

Baye's conditional probability theorem is the base for Naive Bayes (NB) classification technique, that requires every feature of the data set to be independent and unrelated to each other. NB handles a better way for the attributes with missing data or unbalancing values[15]. The main advantage of this algorithm is that with modest RAM and CPU requirement, the training is quick and may provide viable solutions for massive problems (many rows and columns) that are too compute-intensive for other methods. However, it cannot incorporate feature interactions and performance is sensitive to skewed data.

The posterior probability $P(X|A)$ could be calculated from $P(A/X)$, $P(X)$, and $P(A)$[13][14].

Therefore,

$$P(X|A) = (P(A|X) * P(X)) / P(A)$$

Where, $P(X|A)$ –posterior probability of target/object class.

P(A |X)-predictor class probability
P(X) -True probability of class-X.
P(A) -predictor prior probability

### D. SVM Classification Technique:

SVM is a supervised machine learning model employed in classification and regression as well. The algorithm determines the best hyperplane separator between the two classes for a given training data set[16]. However, the hyperplane should not be very nearer to the data points of the another class for generalization. In other words, the margin should be chosen such that the data points are far away from each other class. The data point which is near to the hyperplane is called support vectors[17]. The optimal margin can be determined by increasing the distance between the two decision boundaries.

The equation to optimized hyper plane distance is given by :

$w^T x + b = -1 \quad \& \quad w^T x + b = 1$

If the distance value is equal to $2/||W||$ then it needs to be optimized again. Also, the x(i) should be classified by model , that is $yi * (w^T xi + b) >= 1$, for all i $\in \{1, \notin \notin, N\}$.

### E. Logistic Regression

The logistic regression model basically uses the technique of attribute data mapping to a known class, based on the existing data attributes.

A threshold value will be set by the model to classifies the data items. It classifies data items based on the predicted probability over a threshold value. i.e. predicted value ≥ 0.5, then classify into one class else the other. Decision boundary can be linear or non-linear. In many problems, the target variable of the logistic regression belongs to is binary classification. The mathematical representation of logistic regression is given by,

$$\ln \frac{P(y=1)}{1-P(y=1)} = b_0 + b_1 x_1 + b_2 x_2 ...$$

Where y=1 is the case of yes/true. LHS is called logit (P(y=1)), Figure-2 shows logit(P(y=1)) that is inverted by the sigmoid function [19]

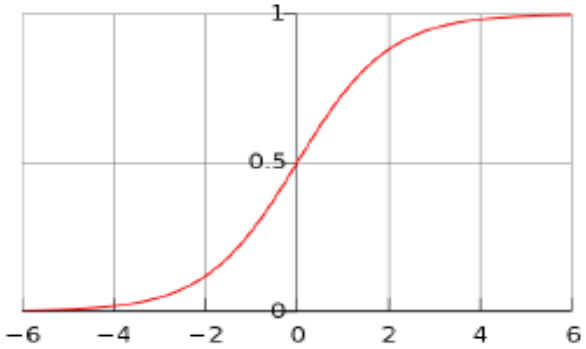


Figure-2 : sigmoid Function[19]

### F. Random Forest Classification Technique

Random Forest (RF) technique is an ensemble machine learning model used for regression and classification problems. It also used for the tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. It is a bagging method to enhance the overall result by adding the additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

## IV. DESCRIPTION ABOUT THE DATA SET

The Diabetic Dataset attributes, data type and its statistics are presented in the table-1 and table-2. Based on the data set attribute the classifier predicts that the individual patient is diabetes positive or not. The Dataset used is a secondary diabetic data set taken from the UCI repository for ML database. The data was collected from a female patient who is at the age of 21 and above living near Phoenix, Arizona. This is a two-class problem with class value 1 being interpreted as “tested positive for diabetes”. The data set contains 768 observations in that 500 samples of class 1 and 268 of class 2 with no missing values. It is also observed that the data set contains an impractical value such as zero bodies mass index and zero plasma glucose. There is a total of nine attributes, in that eight are independent attributes and one dependent. All the attributes are either discrete or continuous and numeric. The attribute descriptions and statistics are shown in the table-I and II respectively. RangeIndex: 768 entries, 0 to 767.

**Table-I: Attribute Description**

| Sl.No | Attribute | Description |
|---|---|---|
| 1 | No.of times Pregnant | Discrete data of type int64 |
| 2 | Plasma Glucose Concentration | Discrete data of type int64 |
| 3 | Diastolic Blood Pressure (mm Hg) | Discrete data of type int64 |
| 4 | Skin Thickness (mm) | Discrete data of type int64 |
| 5 | Insulin mu U/ml | Discrete data of type int64 |
| 6 | BMI ( Weight/ Height ) ($kg/m^2$) | Continuous data of type int64 |
| 7 | Diabetes Pedigree Function | Continuous data of type int64 |
| 8 | Age | Discrete data of type int64 |
| 9 | Outcome Class | Discrete data of type int64 |

**Table-II: Data Set Statistics**

| Data-set | PIDD |
|---|---|
| Samples | 768 |
| I/P Attributes | 8 |
| O/P Classes | 2 |
| Total Attributes | 9 |
| Missing values | Nil |
| Noisy-Attributes | Nil |

## V. RESULTS & MODEL EVALUATION

Comparison of Various machine learning Classifier models is evaluated to the Diagnosis of Diabetes. Performance accuracy of the classifiers is evaluated based on Incorrectly and Correctly Classified Instances out of a total number of instances. Corresponding classifiers performance is measured over Accuracy and values in terms % and is shown in the table-4. The classifier performance is based on the classified instances and cab be calculated by eqn(1). The input Attribute sample distribution graphs of diabetes, glucose, insulin, pregnancy and skin thickness are presented in the figure-3. The Pearson correlation between age and Glucose r is 0.26 and p is 1.2e-13 and shown in the figure-4. From Table-3 it is observed that first Logistic Regression (LR) with accuracy 77.6%and next Naive Bayes (NB) with 75.525 showing the maximum accuracy and SVM is showing minimum accuracy of 65.62%. So the accuracy probability of Logistic Regression is more when compared with other classification techniques.Figure-5 shows the accuracy comparison Graph of the classifications models which are considered for the study. Finally, the results showed that Logistic Regression (LR) is considered as the best classification technique of this experiment because it has given the highest accuracy of 77.6% when compared with other classification techniques.

Model evaluation of the classifiers is essential to check for output results are satisfactory.

The performance of the model can be described by Confusion Matrix as False Negative (FN), False Positive (FP), True Negatives (TN), True Positives (TP). The correctness of the model in predicting the instances is measured in terms of accuracy. It can be computed as :

**Accuracy (Acc) = (True Positives + True Negatives) / Total Samples count ……..(1)**

**Total Samples count (N) = TP +TN + FP +FN**

Precision is the number of correct/ true positive instances divided by the number of actual (true positive + false positive) instances predicted by the classifier. The recall is the number of true positive instances divided by the number of predicted (true positive + false negative ) instances by the classifier. Both precision and recall are therefore based on an understanding and measure of relevance.

F1-Score is a simple measure of a test's accuracy which takes into account both precision and recall. It is simply the harmonic mean of precision and recall. Also, It aims to maximize this score to make the model better, that is, F1-Score reaches its best value at 1 (perfect precision and recall) and worst at 0. High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The classification accuracy, misclassification rate, precision, recall and F1-score are used to measure the performance of the models and are presented in the table-. Figure 3 and 4 depicts the comparison of accuracy and misclassification rate of the classifiers models. Figure-5 shows the comparison of precision, recall, f1-score, misclassification rate and ROC-AUC score of all the classifiers. Logistic regression out-performs with highest f1-score and ROC-AUC score of 75 and 73.6 respectively and least Misclassification rate as 23.8. Table III shows Measures of Classification Models

**Table-III: Measures of Classification Models**

| Classifier | Accuracy % | Recall | f1-Score | Misclassification Rate | ROC-AUC-Score |
|---|---|---|---|---|---|
| KNN | 73.43 | 69 | 69 | 31.1 | 69.8 |
| DT | 70.31 | 72 | 72 | 28.57 | 69.2 |
| NB | 75.52 | 74 | 74 | 25.97 | 70.2 |
| SVM | 65.63 | 64 | 49 | 36.36 | 60.5 |
| LR | **77.6** | **76** | **75** | **23.8** | **73.6** |
| RF | 74.30 | 69 | 69 | 29.0 | 70.1 |

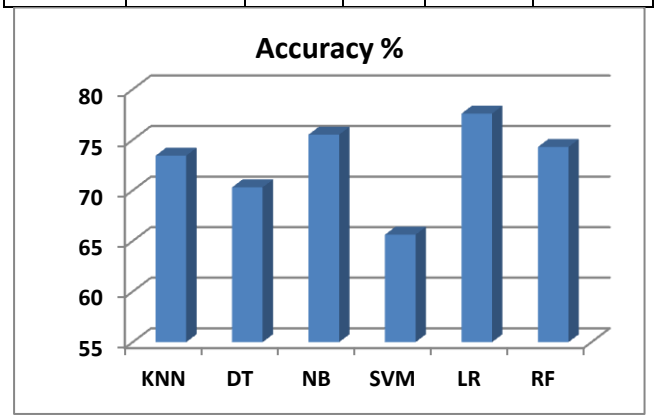


**Figure-3 : Accuracy Comparison of Classifiers**

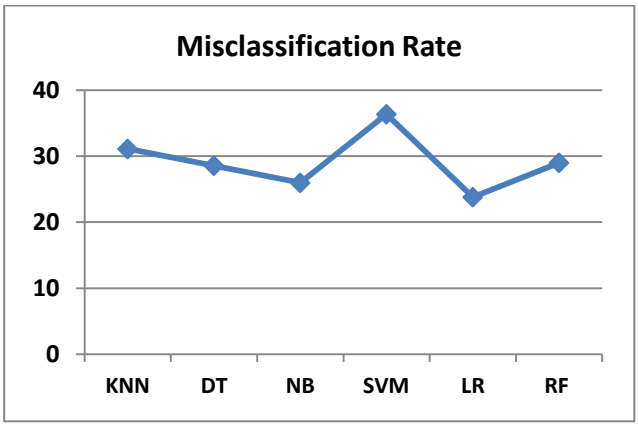


**Figure-4 : Comparison of Misclassification Rate**

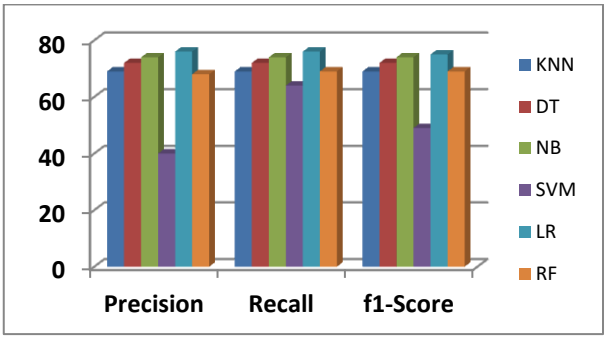


**Figure-5: Comparison of Precision, Recall & F1-score**

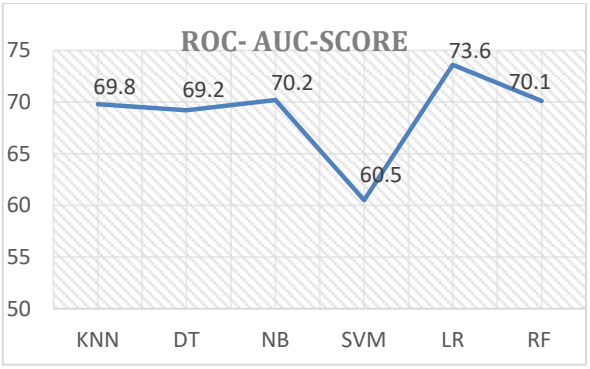


**Figure-6 : ROC-AUC-Scores**

## VI. CONCLUSION

Diabetes detection and prediction are one of the prevalent medical problems in real-world. The persistence of it in the human body for a long time leads to the microvascular complications of diabetes. In this paper, a systematic experimental study was conducted using various Machine Learning classifiers to predict the likeliness of diabetes in a human being. These models that were used to fit the training set were compared and estimated their performance in terms of precision, recall , accuracy and ROC-AUC-Score. The results show that Logistic Regression (LR) classifier performed better with a with highest accuracy of 77.6 %, f1-score of 75 , ROC-AUC score of 73.6 and with least Misclassification rate as 23.8 in comparison to other algorithms. This work can be extended to improve the prediction accuracy using ensemble machine learning techniques.

## REFERENCES

1. Aishwarya, R., "A Method for Classification Using Machine Learning Technique for Diabetes," in International Journal of Engineering and Technology (IJET), 2013, vol. 5, pp 2903–2908.
2. Pradhan P, "Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming," in Advances in Intelligent Systems and Computing ( AISC) , 2014, vol. 1, pp 763–770.
3. Esposito F., "A comparative analysis of methods for pruning decision trees," in IEEE Transactions on Pattern Analysis and Machine Intelligence 1997, vol.19, pp 476–491.
4. Han J., "Discovering decision tree based diabetes prediction model," in International Conference on Advanced Software Engineering and Its Applications, 2008, Springer. pp. 99–109.
5. Iyer A., "Diagnosis of Diabetes Using Classification Mining Techniques," in International Journal of Data Mining & Knowledge Management Process , 2015, vol. 5, pp 1–14.
6. NaiArun, "Comparison of Classifiers for the Risk of Diabetes Prediction," in Proceedia Computer Science,
7. 2015, vol. 69, pp 132–142.
8. NaiArun, "Ensemble Learning Model for Diabetes Classification", in Advanced Materials Research, 2014,
9. pp931-932.
10. Orabi," Early Predictive System for Diabetes Mellitus Disease," , in Industrial Conference on DataMining, 2016, Springer.pp.420–427.
11. Perveen S., "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," in Proceedings Computer Science, 2016, vol. 82, pp 115–121.
12. Pradhan P., "A Genetic Programming Approach for Detection of Diabetes,", in International Journal of Computational Engineering Research, 2012, Vol. 2, pp 91–94.
13. Priyam A., "Comparative Analysis of Decision Tree Classification Algorithms," in International Journal of Current Engineering and Technology, 2013, vol. Vol.3, pp 334–337.
14. Pradeep Khanhasamy J., "Performance Analysis of classifier models to predict diabetes Mellitus," in Elsevier journal, 2016, vol. 82, pp 115 – 121.
15. Deepti Sisodiaa, Dilip Singh Sisodiab, 2018, Prediction of Diabetes using Classification Algorithms, International Conference on Computational Intelligence and Data Science - ICCIDS 2018 ,Science Direct Procedia Computer Science 132 (2018) 1578–1585.
16. RayS., " Easy Steps to Learn Naïve Bayes Algorithm (with code in Python)," 2017
17. Rish I., "An empirical study of the naïve Bayes classifier," workshop on empirical methods in artificial intelligence, IBM., IJCAI 2001 , .pp.41 -46
18. Sisodia D., "SVM for face recognition," in Proceedings of International Conference on Computational Intelligence and Communication Networks, 2010, doi:10.1109 /CICN. 2010.109 pp 554–559.
19. Sisodia D., "Fast and Accurate Face Recognition Using SVM and DCT", in Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS2012), 2012, Springer.pp.1027–1038.
20. Han Wu, "Type 2 diabetes mellitus prediction model based on data mining," in Elsevier journal, 2017, 2352-9148.
21. 19." Data Science And Big Data Analytics" , EMC2 Proven Professional Course material 2012.
22. Sharief,A, "Mathematical Model to Detect Diabetes Using Multigene Genetic Programming", in International Journal of Advanced Research in Artificial Intelligence IJARAI, 2014 , vol. 3, pp 54–59.
23. Tarik A. Rashid, "An Intelligent Approach for Diabetes Classification, Prediction and Description." in Advances in Intelligent Systems and Computing, 2016, doi:10.1007/978-3-319-28031-8, pp323–335.
24. Dhomse Kanchan B., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis," in International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE, 2016, pp. 5–10.
25. Aljumah, A.A., "Application` of data mining: Diabetes health care in young and old patients", in Journal of King Saud University, Computer and Information Sciences, vol. 25, pp 127–136. doi:10.1016/j.jksuci.2012.10.003.
26. Pima Indians Diabetes dataset ( PIDD) from the UCI repository

## AUTORS PROFILE

**Aishwarya Jakka** completed B.E. in Computer Science and Engineering from CMRIT affiliated with VTU. She has worked as a Software Developer for about 2 years on Java, Oracle and Machine Learning. Her research interests include Machine Learning, Artificial Intelligence and Internet of Things. She is currently pursuing her Masters in Information Science at Pittsburgh, USA.

**Dr. Vakula Rani J** is working as a professor in Dept of Computer Applications at CMR Institute of Technology, Bangalore, India. She has completed MCA from Andhra University, M.Phil from Bharatidasan Univeristy and received Ph.D in Computer Science from RayalaSeema University, Kunool, Andhra Pradesh. She has about twenty years of teaching and research experience. Her areas of specialization include Distributed and Parallel computing , Cloud Computing an Machine Learning and Data Science .She has published papers in the international journals and presented research papers in international and national conferences.