Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

DIABETICS PREDICTION USING MACHINE LEARNING TECHNIQUES

CH. KRANTHIREKHA
Department of Electronics and Communications Engineering, Vignana Bharathi Institute of Technology, Gathkesar, Telangana Hyderabad. Email: Kchennaboina20@gmail.com

V. SHARMILA
Department of Electronics and Communications Engineering, Vignana Bharathi Institute of Technology, Gathkesar, Telangana Hyderabad. Email: sharmilavallem@gmail.com

S. POTHALAIAH
Department of Electronics and Communications Engineering, Vignana Bharathi Institute of Technology, Gathkesar, Telangana Hyderabad. Email: pothalaiahs@gmail.com

VEERLAPATI SIRI
Department of Electronics and Communications Engineering, Vignana Bharathi Institute of Technology, Gathkesar, Telangana Hyderabad. Email: Siriharshu04@gmail.com

U. MADHURI
Department of Electronics and Communications Engineering, Vignana Bharathi Institute of Technology, Gathkesar, Telangana Hyderabad. Email: ushannagarimadhuri@gmail.com

SUNKIREDDY AKSHITHA
Department of Electronics and Communications Engineering, Vignana Bharathi Institute of Technology, Gathkesar, Telangana Hyderabad. Email: akshithasunkireddy242@gmail.com

Abstract

Diabetes, characterized by elevated glucose levels in the human body, poses significant health risks if left untreated, including heart issues, kidney dysfunction, hypertension, eye damage, and harm to other organs. Early prediction and intervention are critical for effective diabetes management. In pursuit of this objective, this project employs various machine learning techniques to enhance the accuracy of diabetes prediction. Machine learning, a potent tool for predictive analytics, leverages patient datasets to construct models. In this study, we apply classification and ensemble techniques within the machine learning framework to predict diabetes. The selected algorithms encompass Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF). Each machine learning model exhibits distinct accuracy levels when compared to its counterparts. The central aim of this project is to identify the most accurate model, highlighting its proficiency in predicting diabetes effectively. Our findings demonstrate that the Support Vector Machine (SVM) model outperforms other machine learning techniques, showcasing its potential for early diabetes prediction. This research seeks to contribute to improved healthcare outcomes by enabling early intervention and enhanced disease management.

Keywords: Ensemble learning, classification, dataset, techniques.

I. INTRODUCTION

Diabetes, recognized as one of the most pernicious diseases worldwide, emerges primarily from factors such as obesity and elevated blood glucose levels. This condition disrupts the delicate balance of the hormone insulin, leading to abnormal carbohydrate metabolism and heightened blood sugar levels. The root cause of diabetes often lies in the insufficient production of insulin by the body.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

According to the World Health Organization (WHO), approximately 422 million individuals grapple with diabetes, with a significant prevalence in low-income or economically idle countries. This number is projected to escalate to a staggering 490 million by the year 2030. Notably, diabetes casts its shadow across diverse nations, including countries like Canada, China, and India. In a country as populous as India, the number of diabetics surpasses 40 million, given its population exceeding 100 million. Alarmingly, diabetes stands as a major contributor to global mortality rates. The importance of early disease prediction, especially in the case of diabetes, cannot be overstated, as it holds the potential to save countless lives. To address this critical need, our research delves into the prediction of diabetes by examining various attributes associated with the disease. We harness the power of the Pima Indian Diabetes Dataset and employ a spectrum of Machine Learning classification and ensemble Techniques to forecast diabetes occurrences. Machine Learning serves as an invaluable method for training computers and machines to understand patterns and relationships in data explicitly. The diverse arsenal of Machine Learning Techniques offers efficient tools for extracting knowledge by crafting intricate classification and ensemble models from meticulously gathered datasets. These datasets, enriched with pertinent information, serve as the foundation for predicting diabetes.

While various Machine Learning techniques stand capable of making predictions, the challenge lies in selecting the most suitable approach. To overcome this hurdle, we systematically apply popular classification and ensemble methods to our dataset, aiming to identify the most effective means of predicting diabetes.

This research endeavors to contribute to the early detection and control of diabetes, thereby enhancing the quality of life and longevity of individuals at risk of this debilitating condition.

II. LITERATURE REVIEW

K. Vijiya Kumar introduced a Support Vector Machine algorithm aimed at achieving early diabetes prediction with heightened accuracy. By harnessing the power of SVM within the realm of machine learning, the proposed model emerges as a robust system for effectively and efficiently predicting diabetes in patients. The results of this study underscore the system's capability to deliver prompt diabetes predictions.

Nonso Nnamoko and collaborators put forth an ensemble supervised learning approach for predicting diabetes onset. In this approach, five widely used classifiers are employed to form ensembles, and a meta-classifier aggregates their outputs. The study's results are presented and compared with similar research efforts utilizing the same dataset. The outcomes highlight the ability of this method to predict diabetes onset with enhanced accuracy.

Tejas N. Joshi and colleagues presented a study focusing on diabetes prediction using three distinct supervised machine learning methods: Support Vector Machine (SVM), Logistic Regression, and Decision Trees (DT). The project's objective is to propose an

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

effective technique for the early detection of diabetes, leveraging the capabilities of these machine learning algorithms.

Muhammad Azeem Sarwar and his team conducted a study on diabetes prediction using various machine learning algorithms in the healthcare domain. They applied six different machine learning algorithms and discussed their performance and accuracy. The comparison of these techniques provided insights into which algorithm is best suited for the prediction of diabetes.

The collective body of research in diabetes prediction is a testament to the growing interest in leveraging machine learning and classification techniques to identify diabetic patients accurately. These studies underscore the significance of developing robust systems for diabetes prediction, as early detection holds the potential to address critical healthcare challenges. Given the wealth of prior research, this area continues to be a vital focus for computational approaches aimed at improving healthcare outcomes.

III. PROPOSED METHODOLOGY

The central aim of this paper is to research and identify a model that can enhance the accuracy of diabetes prediction. To achieve this goal, we conducted a series of experiments involving diverse classification and ensemble algorithms specially tailored for diabetes prediction. Below, we provide a concise overview of the research phases.

A. Dataset Description: Our research leveraged the "Pima Indian Diabetes Dataset," sourced from the UCI repository. This dataset encompasses a wide array of attributes collected from a total of 768 patients.

S. No	Attributes
1.	Pregnancy
2.	Blood Pressure
3.	Glucose
4.	Skin Thickness
5.	Insulin
6.	BMI
7.	DPF
8.	Age

The ninth attribute within each data point of the dataset serves as the class variable. This variable plays a pivotal role in indicating the outcome of diabetes prediction. It employs a binary classification, with values of 0 and 1 signifying different conditions:

0: This class label signifies a negative outcome, indicating the absence of diabetes.

1: On the other hand, this class label represents a positive outcome, indicating the presence of diabetes.

Distribution of Diabetic Patients

While our objective was to create a predictive model for diabetes, it's worth noting that the dataset displayed a slight class imbalance. Specifically, the distribution of classes was as follows:

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

Class 0: Approximately 500 instances were labeled as 0, indicating a negative outcome, or no diabetes.

Class 1: Around 268 instances were labeled as 1, indicating a positive outcome, or diabetic patients.

This class imbalance is a crucial consideration in model development and evaluation, as it can impact the model's performance and necessitate the use of techniques to address class imbalance issues during the analysis.

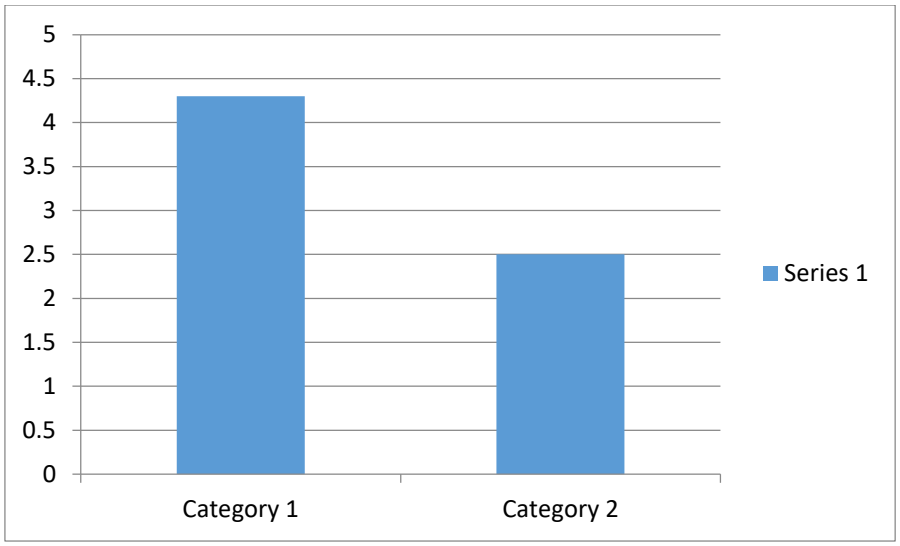

Figure 1: Ratio of Diabetic and Non Diabetic Patient

B. Data Preprocessing: Data preprocessing is a critical phase, particularly in healthcare-related data analysis. Healthcare datasets often contain missing values and other imperfections that can significantly impact data quality and the effectiveness of subsequent mining processes. To ensure the accurate application of Machine Learning Techniques to our dataset, data preprocessing is an essential step for achieving precise results and successful predictions.

For the Pima Indian Diabetes Dataset, we undertook data preprocessing in two key steps:

1. Missing Values Removal: The first step involves the removal of instances that contain zero (0) as a value. In the context of this dataset, having zero as a value for certain attributes is not plausible or meaningful. Therefore, these instances are deemed irrelevant and are consequently eliminated from the dataset. This process of eliminating irrelevant features or instances is known as feature subset selection. By reducing the dimensionality of the data, it facilitates faster and more efficient data analysis.

2. Splitting of Data: After the data cleaning phase, the dataset is divided into two subsets: a training set and a testing set. This division enables us to train machine learning algorithms on the training dataset while keeping the test dataset separate. During the training process, the algorithm learns from the training data, building a model based on

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

logic and algorithms that capture the relationships and patterns within the features of the training data.

Normalization is a crucial aspect of this step, as it aims to bring all attributes to the same scale. Ensuring that attributes have consistent scales is essential for the accurate functioning of machine learning algorithms, as it prevents attributes with larger scales from dominating the learning process. Normalization enhances the effectiveness of the subsequent modeling and prediction phases.

In summary, data preprocessing plays a pivotal role in preparing the dataset for machine learning. It involves the removal of irrelevant instances and the appropriate division of data into training and testing sets, all while ensuring that attributes are on a consistent scale. These preprocessing steps lay the foundation for accurate and successful diabetes prediction using machine learning techniques.

C. Apply Machine Learning Techniques: Once the dataset has undergone preprocessing, the subsequent step involves the application of various Machine Learning Techniques to predict diabetes. Our primary objective is to analyze the performance of these techniques, determine their accuracy, and identify the most important features contributing to prediction.

The Techniques Utilized:

1. Support Vector Machine (SVM): SVM is a renowned supervised machine learning algorithm used extensively in classification tasks. It is known for creating hyperplanes that effectively separate two classes. SVM operates efficiently in high-dimensional spaces and can classify instances not explicitly represented in the data. The separation is achieved by identifying a hyperplane that maximizes the margin between classes, minimizing the chance of misclassification.

Algorithm Steps:

- Selection of the hyperplane for optimal class separation.
- Calculation of the margin, representing the distance between classes.
- Selection of the class with the highest margin, where margin = distance to positive point + distance to negative point.

2. Decision Tree: Decision trees are fundamental classification models that are supervised and suitable for categorical response variables. They utilize a tree-like structure based on input features to describe the classification process. Each node in the tree represents an input feature, and the tree construction involves selecting the feature that provides the highest information gain to predict the output.

Steps for Decision Tree Algorithm:

- Constructing the tree with nodes representing input features.
- Selection of the feature for predicting the output based on the highest information gain.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

- Repeating the process to create sub-trees using unused features.

3. Logistic Regression: Logistic regression is a supervised classification algorithm used for estimating the probability of binary responses based on one or more predictors. It is employed when the goal is to classify data items into categories, such as determining whether a patient is positive or negative for diabetes. Logistic regression aims to establish the relationship between target and predictor variables, utilizing the sigmoid function to predict probabilities of positive and negative classes.

Sigmoid Function:

P = 1 / (1 + e^(-a - bx))

(P = probability, a and b = parameters of the model)

4. Random Forest: Random Forest is an ensemble learning method suitable for both classification and regression tasks. It is known for its accuracy, particularly with large datasets. Random Forest constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of individual trees. It mitigates variance and improves the performance of decision trees.

Algorithm Steps:

- Selection of a subset of features from the total feature set.
- Choosing the node using the best split point from the selected features.
- Splitting the node into sub-nodes based on the best split.
- Repetition of steps until a certain number of nodes are reached.
- Building the forest by repeating the process to create multiple trees.

Ensembling techniques like Bagging and Gradient Boosting were also utilized in this research to predict diabetes effectively, enhancing the accuracy of the models. These ensemble methods address errors, noise, bias, and variance to improve overall performance.

In summary, our research involves the comprehensive exploration and application of these machine learning and ensemble techniques on the Pima Indian Diabetes Dataset. Our goal is to evaluate their performance, ascertain accuracy, and identify the most influential features in diabetes prediction.

IV. MODEL BUILDING

The model building phase is pivotal and represents the heart of this research, involving the implementation of various machine learning algorithms discussed earlier for the prediction of diabetes. Below, we outline the procedure for our proposed methodology:

Step 1: Importing Required Libraries and Dataset

Begin by importing the necessary libraries and loading the diabetes dataset.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

Step 2: Data Preprocessing

Preprocess the dataset to handle missing data, ensuring it is clean and ready for analysis.

Step 3: Data Splitting

Perform a percentage split, typically 80% for the training set and 20% for the test set. This division enables the model to learn from one portion of the data and evaluate its performance on another.

Step 4: Algorithm Selection

Choose the machine learning algorithm(s) to be employed for diabetes prediction. Options include Support Vector Machine, Decision Tree, Logistic Regression, and Random Forest, as discussed in earlier sections.

Step 5: Model Building

Build the classifier model(s) for the selected machine learning algorithm(s) based on the training set. The model is trained to recognize patterns and relationships within the training data.

Step 6: Model Testing

Evaluate the classifier model(s) using the test set. This step involves applying the trained model(s) to unseen data to assess their predictive accuracy.

Step 7: Performance Comparison

Conduct a comprehensive comparison and evaluation of the experimental performance results obtained for each classifier. This comparison encompasses various evaluation metrics and measures to gauge the effectiveness of each algorithm.

Step 8: Algorithm Selection

After analyzing the results based on different evaluation criteria and performance measures, conclude the best-performing machine learning algorithm for diabetes prediction. The algorithm demonstrating the highest accuracy and suitability for the task is selected as the primary model.

This model building phase forms the core of the research, allowing for the selection of the most effective algorithm for diabetes prediction based on thorough evaluation and performance.

V. EXPERIMENTAL RESULTS

This research involved a series of meticulously planned steps, culminating in the proposed approach that leverages a variety of classification and ensemble methods implemented in Python. These methods represent established and widely-recognized machine learning techniques, meticulously selected to extract the highest accuracy from the dataset. The experimental findings are presented below:

Method Selection: In the course of this research, various classification and ensemble methods were harnessed, all with the common goal of optimizing accuracy in diabetes prediction.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

Python Implementation: The chosen methods were implemented using the Python programming language, a versatile and widely-used platform for machine learning and data analysis.

Performance Comparison: A detailed examination of the results revealed that the Random Forest classifier consistently outperformed other methods in terms of prediction accuracy. Random Forest proved to be the most effective technique among those employed.

High Performance Accuracy: The research underscored the importance of selecting the most appropriate machine learning techniques for diabetes prediction, and the application of these techniques led to notably high levels of performance accuracy.

Visual Representation: A figure was used to visually illustrate the outcomes of the various Machine Learning methods, offering a clear and concise presentation of the results.

In conclusion, the experimental results demonstrate that the Random Forest classifier emerged as the most effective approach for diabetes prediction in this research. Overall, this study emphasizes the significance of selecting the best-suited Machine Learning techniques to achieve remarkable performance accuracy in predictive modeling for healthcare applications.

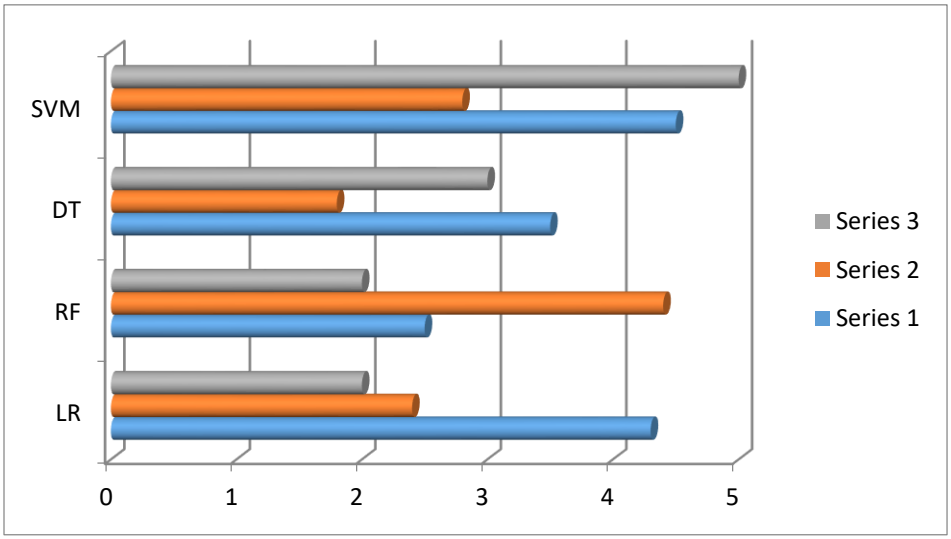

Fig.2: Accuracy results of machine learning methods

In the presented results, the importance of each feature in diabetes prediction is showcased, with a focus on the Support Vector Machine algorithm. The importance values of each feature, which significantly contribute to diabetes prediction, have been plotted. Here's a recreated description of the visualization:

Feature Importance for Diabetes Prediction using Support Vector Machine:

In this visualization, we highlight the pivotal role that individual features play in predicting

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

diabetes, with a particular emphasis on the Support Vector Machine (SVM) algorithm. The importance of each feature is represented on

the X-axis, while the names of these important features are displayed on the Y-axis. This graphical representation provides a clear and insightful view of the contribution of each feature towards accurate diabetes prediction, as assessed by the SVM algorithm. It allows for the identification of the most influential features that significantly impact the prediction outcome.

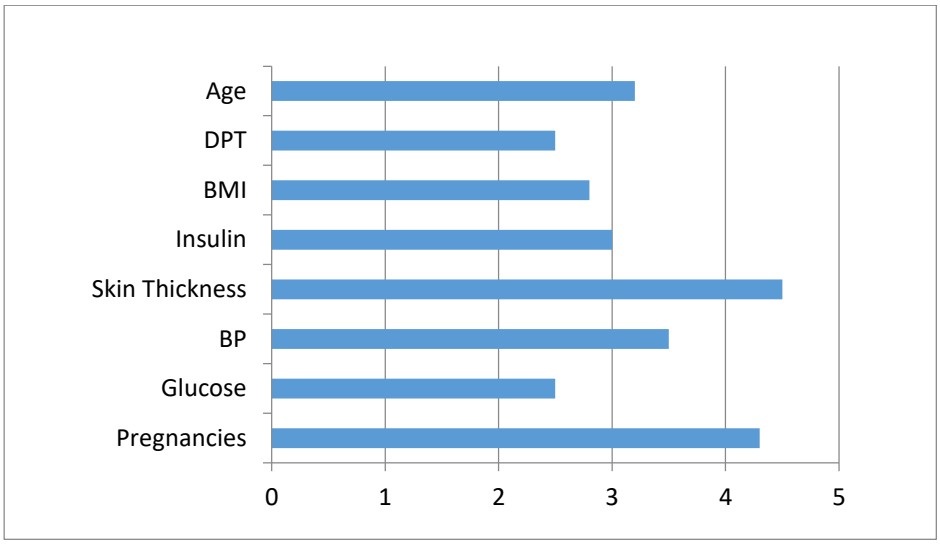

Fig.3: Features importance

VI. CONCLUSION

The primary objective of this project was the development and implementation of a Diabetes Prediction System utilizing Machine Learning techniques, followed by a thorough performance analysis of these methods. We have successfully achieved this objective. The proposed approach harnesses a range of classification and ensemble learning methods, including SVM, Random Forest, Decision Tree, and Logistic Regression classifiers.

In conclusion, the successful design and implementation of a Diabetes Prediction System using Machine Learning techniques, coupled with a comprehensive performance analysis, demonstrate the promising potential of these methods in healthcare. The accuracy achieved and the insights gained from this research underscore the importance of early diabetes prediction as a means to enhance patient care.

References

1) PriyankaIndoria, Yogesh kumar Rathore. A survey: Detection and Prediction of diabetics using machine learning techniques. IJERT, 2021.

2) Khaleel, M.A., Pradhan, S.K., G.N Dash. A Survey of Data Mining Techniques on Medical Data for finding frequent diseases. IJARCSSE, 2019.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10200729

3) K. Vembandasamy, R. Sasipriya, E. Deepa. Heart DiseaseS Detection using Naïve Bayes Algorithm. IJISET, 2017.

4) Tawfik Saeed Zekia, Mohammad V. Malakootib, Yousef Ataeipoorc, S.Talayeh Tabibid. An Expert System for Diabetics Diagnosis. American Academic & Scholarly Research Journal special Issue Vol.4,No.5, September 2012.

5) Vishali Bhandari and RajeevKumar. Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis. International Journal of Computer Applications (0975 – 8887) Volume 132 – No.6, December 2015.

6) Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, “Machine Learning and Data Mining Methods in Diabetics Research”, Jan 8, 2017.

7) Eka Miranda, EdyIrwansyah, AlowisiusY. Amelga, MarcoM. Maribondang, MulyadiSalim. Detection of cardiovascular Disease Risk’s Level for Adults using naïve Bayes Classifier, The Korean Society of Medical informatics (KOSMI), 2016.

8) ZhengT, XieW, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. A Machine Learning-Based Framework to identify Type 2 Diabetics through Electronic Health Records, International Journal of medical informatics (IJMI),2017, Vol9, pages120-127.