Early Stage Diabetes Risk Prediction

FARAZ SIKDER SHAFIN
21101171
Department of Computer Science
BRAC University
Dhaka, Bangladesh
faraz.sikder.shafin@g.bracu.ac.bd

RITISHA ISLAM
20201222
Department of Computer Science
BRAC University
Dhaka, Bangladesh
ritisha.islam@g.bracu.ac.bd

MD MAHMUDUL HASAN
19201026
Department of Computer Science
BRAC University
Dhaka, Bangladesh
md.mahmudul.hasan@g.bracu.ac.bd

SHARTHAK DAS
19201025
Department of Computer Science
BRAC University
Dhaka, Bangladesh
sharthak.das@g.bracu.ac.bd

Abstract :

Diabetes is a chronic condition characterized by elevated blood glucose levels, which, if left unmanaged, may have detrimental effects on several organ systems. It is associated with the development of heart disease, renal complications, impaired peripheral nerves, compromised blood vessels, and vision impairment. The timely prediction of diseases has the potential to significantly reduce mortality rates and empower healthcare professionals to effectively manage various medical disorders. Many individuals diagnosed with diabetes had little knowledge of the risk factors associated with the condition prior to their diagnosis. In the present day, hospitals use rudimentary information systems that produce substantial volumes of data that lack the ability to be effectively transformed into meaningful and actionable information. Consequently, these data sets are unable to adequately facilitate clinical decision-making processes. Various automated approaches are now available for the early prediction of diseases. Ensemble learning is a data analysis methodology that integrates numerous methodologies into a unified predictive system with the aim of assessing bias and variance, as well as enhancing prediction accuracy. The dataset on diabetes, including 17 variables, was collected from the UCI repository, which has a diverse range of datasets. This research utilizes predictive models such as AdaBoost, Bagging, and Random Forest to assess the precision, recall, classification accuracy, and F1-score. Ultimately, the Random Forest Ensemble Method exhibited the highest level of accuracy, reaching 97%. In contrast, the AdaBoost and Bagging algorithms had comparatively lower accuracy, precision, recall, and F1 scores.

Introduction:

In this paper, we will examine the topic at hand and explore its various aspects in an
The increasing prevalence of diabetes in individuals' daily lives may be attributed to the upward trajectory of living standards. Diabetes mellitus, often known as diabetes, is a chronic medical illness characterized by elevated levels of blood glucose. A variety of physical and chemical tests may be used to identify and diagnose this particular illness. Untreated and undiscovered diabetes has the potential to adversely affect critical organs such as the eyes, heart, kidneys, feet, and nerves, ultimately resulting in mortality.

Diabetes is a persistent medical illness that has the capacity to significantly impact world health. Recent investigations undertaken by the World Health Organization (WHO) have shown a notable surge in both the prevalence and fatality rates among individuals diagnosed with diabetes on a worldwide scale. According to projections made by the World Health Organization (WHO), it is expected that diabetes will be ranked as the seventh most prevalent cause of mortality by the year 2030. Based on statistics provided by the International Diabetes Federation (IDF), the global population now comprises 537 million individuals diagnosed with diabetes, with projections indicating an anticipated increase to 643 million by the year 2030.

The only approach to mitigating the difficulties associated with diabetes is to promptly detect and manage the condition. The timely identification of diabetes is crucial due to the progressive escalation of its associated consequences.

The prediction of diabetes has significant importance in facilitating appropriate treatment strategies aimed at mitigating the potential consequences associated with the condition. A plethora of research has been undertaken pertaining to illness prediction, including areas such as diagnosis, prognostication, classification, and therapeutic interventions. According to recent research, a variety of machine learning (ML) algorithms have been used to detect and predict illnesses. These advancements have resulted in a significant enhancement in the effectiveness and progress of both traditional and machine-learning methodologies. A wide range of machine learning algorithms and ensemble approaches have been used in the domain of illness categorization. However, based on the existing body of research, none of these methods have shown the ability to achieve high levels of accuracy, namely over 80%. In 2022, Saxena et al. conducted a comprehensive analysis of existing studies pertaining to the classification model for predicting diabetes. This examination revealed a research gap that our current study aims to address. The authors reached the conclusion that the dataset underwent the application of commonly used machine learning algorithms, with just one author using the AdaBoost and gradient boost approaches. Hence, a system that has the capability to provide discoveries of higher accuracy exhibits enhanced processing speed, thereby becoming more valuable for predictive endeavors. The objective of this research is to enhance the precision of machine learning ensemble standard algorithms, such as Support Vector Machines, Logistic Regression, Decision Tree, KNN, Naïve Bayes, and Random Forest, via an analysis of the UCI diabetes dataset and a comparative evaluation of their respective performances.

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Upon careful analysis of the works of many writers and researchers, it becomes apparent that determining the influential characteristics within a dataset is a challenging task. Furthermore, it is vital to note that achieving 100% accuracy via optimal feature selection is not guaranteed. A wide range of classification approaches is often used by academics, including support vector machines, decision trees, random forests, k-nearest neighbors, multilayer perceptron, and logistic regression. A limited number of academics have devised a methodology that effectively predicts occurrences by using recurrent neural networks or deep learning. Table 1 presents a comprehensive comparison of the research studies that have been taken into consideration in this study. The primary findings of this study highlight the notable characteristics that mitigate the risks associated with early-stage diabetes, as well as the ensemble approaches that exhibit superior accuracy, particularly the Random Forest method with a precision rate of 97%.

Table 1

List of characteristics with their standards.

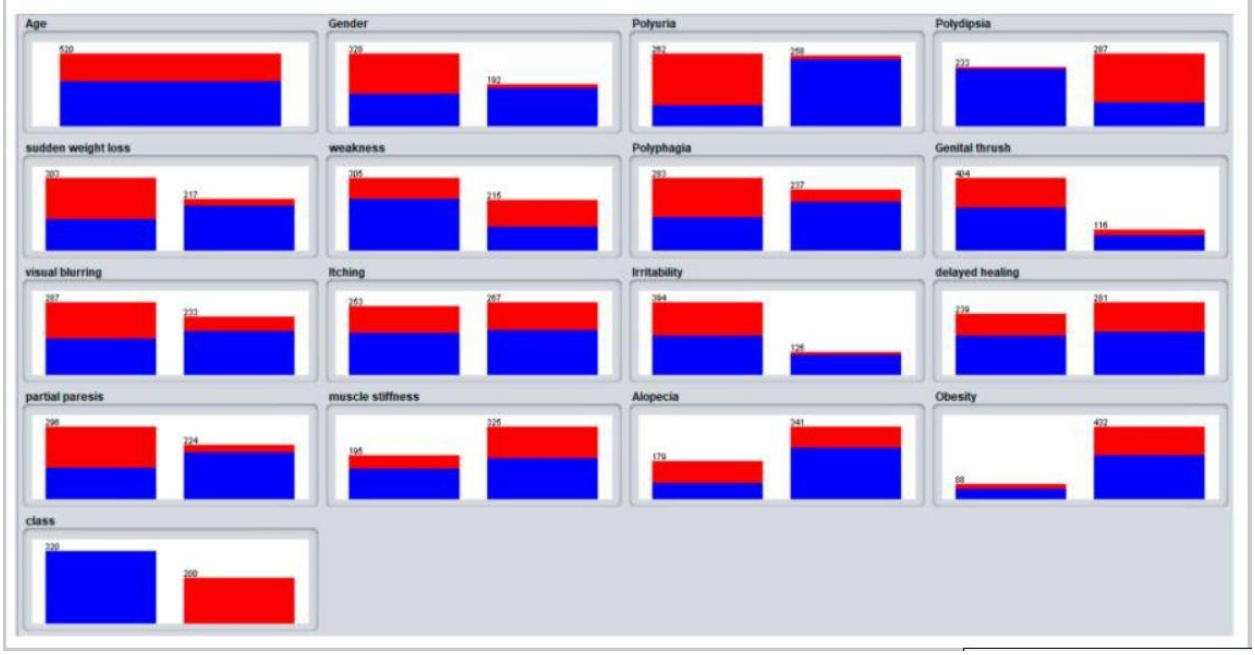
ATTRIBUTES	VALUE	
Age	Numeric	
Gender	Men = 328, Women = 192	
Polyuria	✓ = 258, × = 262	
Polydipsia	✓ = 233, × = 287	
Sudden weight loss	✓ = 217, × = 303	
Weakness	✓ = 305, × = 215	
Polyphagia	✓ = 237, × = 283	
Genital thrush	✓ = 116, × = 404	
Visual blurring	✓ = 233, × = 287	
Itching	✓ = 253, × = 267	
Irritability	✓ = 126, × = 394	
Delayed healing	✓ = 239, × = 281	
Partial paresis	✓ = 224, × = 296	
Toughness of muscle	✓ = 195, × = 325	
Alopecia	✓ = 179, × = 341	
Overweightness	✓ = 88, × = 432	
Class	Positive = 320, Negative = 200	

Data preprocessing:

The process of data preprocessing plays a vital role in the field of data mining, particularly when confronted with data that is incomplete, noisy, or inconsistent. This step involves transforming the data into a form that is both useable and ideal for further analysis. In order to consistently organize data in a logical and accurate manner, the process of data preparation encompasses several tasks like data cleansing, data discretization, data integration, data reduction, data transformation, and other related operations. In this particular case study, a dataset consisting of diabetes data with 17 variables was obtained from the UCI repository, which serves as a repository for various datasets. The dataset used in this study consists of 17 variables that represent various patient and hospital outcomes. The use of ensemble methods has been utilized to evaluate the predictive accuracy of a model. This model is constructed using clinical treatment data obtained from direct surveys conducted among patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh. The validation of this data has been performed by medical professionals.

Certain data mining approaches have a greater ease in handling discrete attributes. Discrete attributes, sometimes referred to as notional attributes, are qualities that serve to categorize data. Ordinal features refer to certain attributes that define a particular category and have relevance in terms of the arrangement or hierarchy of these categories. Discretization refers to the process of converting a continuous attribute with real-valued data into an ordinal attribute or bin. A discretization filter was used in this context due to the presence of real-valued input data, hence facilitating their organization into distinct bins.

This research uses a dataset consisting of 520 instances, each characterized by 17 qualities. One of these attributes serves as the class attribute, which is employed to predict the likelihood of an individual having diabetes or not. Table 1 displays the qualities along with their corresponding values, whereas Figure 1 illustrates the preprocessing outcomes for each individual attribute.



The Chi-Square characteristics selection approach [35] is used in this study to evaluate the relevant qualities. The primary characteristics of interest make a connection between two categorical variables, namely, a contingency, which represents the association between observed and expected frequencies. The Chi-Square approach is used to compute the attribute scores for diabetes data. A 10-fold cross-validation was used. The use of a systematic division in percentages is a customary method employed in evaluation approaches. The dataset is partitioned into 10 distinct portions, and then, each segment is subjected to individual testing. Ten evaluation results are obtained and then averaged. During the first division in the "stratified" cross-validation, it ensures that each fold has an equal proportion of class values. The learning algorithm is iterated for the last time (11th iteration) using the whole dataset to generate the output after 10-fold cross-validation. This process will provide the evaluation results.

Ensemble approaches have been used in the analysis of diabetes data because of the escalating prevalence of diabetes among individuals. Consequently, it is essential to proactively ascertain the likelihood of developing diabetes in the future. Ensemble learning is a data mining methodology that combines many methodologies into a unified predictive model with the aim of reducing bias and volatility or improving prediction accuracy. In comparison to a single model,

this strategy has superior prediction performance. In this work, the ensemble approaches of Support Vector Machines, Logistic Regression, Decision Tree, KNN, Naïve Bayes and Random Forest were used to forecast the likelihood of early-stage diabetes risk.
The Random Forest algorithm is a machine learning methodology that combines decision tree models for classification and regression tasks. It is based on the Bagging ensemble method. One drawback associated with bagged decision trees is their reliance on a greedy algorithm for determining the ideal split point throughout each stage of the tree development process. Consequently, the resulting trees provide a comparable visual representation, hence reducing the variability of predictions across all the bags and diminishing the resilience of the predictions.

Methodology:

A. The Dataset

The dataset used in this study was obtained from the UCI repository. The dataset consists of 520 instances and 16 attributes, with a small number of missing values. These missing values were handled by pre-processing by excluding tuples that included partial values. The dataset has been summarized.

Following the pre-processing stage, the dataset was reduced to a final count of 520 occurrences. Among the total of 520 cases, 320 are characterized as positive values, while the remaining 200 instances are classified as negative values. The use of two class factors, namely positive and negative, is employed to determine the presence or absence of diabetes risk in a patient. The experimental procedure was conducted in accordance with established scientific protocols. The experimental approach is conducted in accordance with the following sequential steps: The dataset is divided into two sets, namely the training set and the test set, with a ratio of 80:20. This partitioning is achieved using a 10-fold cross-validation technique. The aforementioned seven classification methods are then used to classify the dataset into two distinct classes, namely positive and negative. Four evaluation measures, namely classification accuracy, F-score, ROC value, and calculation time, are computed in order to compare the performance of the provided methods. This process is conducted in order to identify the most optimal categorization method.

C. Algorithms

1) Support Vector Machines:

Support Vector Machines (SVMs) are a kind of machine learning algorithm that are widely used in several fields, including pattern recognition and classification tasks. SVMs are based on the

concept of finding an optimal hyperplane that separates different classes of data points. The introduction of Support Vector Machines (SVM) may be attributed to Vapnik. The Support Vector Machine (SVM) algorithm operates by identifying crucial instances, referred to as support vectors, from each class. It accomplishes class separation by constructing a function that maximizes the margin between the classes, using these support vectors. Hence, it can be said that Support Vector Machines (SVM) facilitate the creation of a mapping from an input vector to a space with large dimensionality. The objective of SVM is to identify the optimal hyperplane that effectively separates the dataset into distinct classes [21]. The primary objective of this linear classifier is to optimize the separation between the decision hyperplane and the closest data point, known as the marginal distance. This is achieved by identifying the most suitable hyperplane.

This study used the Radial Basis Function (RBF) kernel, often referred to as the Gaussian kernel, to categorize the data throughout the comparative analysis of different kernels. When conducting training for a Support Vector Machine (SVM) using the Radial Basis Function (RBF) kernel, it is essential to take into account two crucial hyperparameters, namely C and gamma. The hyperparameter C, which is universally applicable to all support vector machine (SVM) kernels, balances the trade-off between the misclassification of training instances and the complexity of the decision surface. A smaller value of the hyperparameter C results in a smoother decision surface, while a larger value of C is intended to achieve perfect classification of all training samples. The magnitude of impact exerted by an individual training instance is determined by the parameter gamma. As the value of gamma increases, the proximity of additional instances required to be influenced becomes greater.

The measurement of the distance between data points is conducted using the Gaussian kernel:

$$K_{rbf}(x_i,x_j)=\exp(-\gamma|x_i-x_j|^2)$$

In this context, xi and xj represent individual data points, whereas ‖ xi – xj ‖ refers to the Euclidean distance between them.

The selection of kernel functions is contingent upon the characteristics of the data and the particular domain issue at hand. The validation set is used to evaluate the performance of different values of C and gamma. The selection of the optimal value of C for achieving the highest accuracy on the validation set is determined by evaluating its performance. The optimal validation score is achieved when the values of gamma and C are set to 0.1 and 1.7, respectively.

2) Decision Tree:

The Decision Tree classifier is a very effective supervised machine learning technique used for classification tasks. The process entails making judgments by drawing upon existing facts. The Decision Tree classifier consists of specific features that serve as nodes inside the tree structure. At each level of the process, a node is selected based on the assessment of the attribute with the biggest information gain.

The determination of the priority of nodes in a decision tree is established by using either the Gini index or the Entropy measure. These metrics serve as indicators to identify the most effective classifier from the available characteristics. The decision tree may be conceptualized as a collection of sequential questions that aid in the process of classification. The algorithm's efficiency is enhanced by adjusting hyperparameters such as the maximum depth of a tree and the criteria, which may be specified as either 'Gini' or 'Entropy'. The dataset yields the highest accuracy when the maximum depth is set to 7 and the 'Gini' criteria are used.

3) Random Forests:

A random forest may be described as an aggregation of many decision trees. The underlying principle of random forest is to aggregate diverse subsets of training data in order to construct decision trees, therefore mitigating the risks of overfitting and misclassification via the process of averaging the outcomes of many decision trees. The model proposed in the research utilizes a maximum depth of 13 and 100 estimators for the purpose of classifying the data points. When the maximum depth was increased, there was a noticeable decline in the accuracy of the categorization.

4) Naive Bayes:

The naive Bayes classifier is widely used as a probabilistic classifier in several domains. The implementation of Bayes' Theorem [26] is used, disregarding the order and rules while assuming independence among the characteristics. Therefore, the naïve character of the phenomenon may be derived. There are many kinds of Naive Bayes Algorithms, including multinomial Naive Bayes, Gaussian Naive Bayes, Bernoulli Naive Bayes, and others. A Gaussian Naive Bayes classifier was used for the model. The Gaussian Naive Bayes algorithm is often used in the analysis of datasets with a large number of dimensions. The Naive Bayes algorithms have efficient training and prediction capabilities.

5) KNN classifier:

The K-nearest neighbor (KNN) classifier is a straightforward and non-probabilistic technique often used in machine learning. The training dataset is kept, and the prediction process entails identifying the nearest data point inside the training set.

$$\text{Euclidean} = \left(\sum_{i=1}^{k} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

The equation provided serves as a means to compute the distance between two data points, xi, and yi, in a dataset with k dimensions. The value of k is dictated by the characteristics of the dataset. We assign the value of p as 2, representing the Euclidean Distance formula. In the model described in the research, the authors have established a configuration of three neighboring data points that yield satisfactory accuracy. However, increasing the number of neighbors above this threshold leads to improved classification performance, but also introduces a notable risk of misclassification. Consequently, the model becomes susceptible to overfitting.

6) Logistic Regression:

Logistic Regression is a binary classification technique that is governed by the equation:

$f(x)= 1 /1+e-x$

The sigmoid function is bounded inside the interval (0,1). By default, logistic regression uses L2 regularization. The significance of the parameter C is pivotal in the training process of the model, as it serves to minimize misclassification and improve the overall accuracy of the model. A lower value of the hyperparameter C results in a smoother decision surface, while a higher value of C is intended to achieve perfect classification of all training data. The optimal value for achieving the highest level of accuracy is obtained when C is set to 1.7. The approach does not need extensive fine-tuning of hyperparameters. Favorable outcomes may be seen while working with extensive datasets. There was no observed effect on the accuracy of the model as a result of increasing the number of iterations.

7) In order to assess the efficacy of the classification methods used in this work, the dataset was partitioned into training and test datasets, comprising 75% and 25% of the total data, respectively. The datasets were randomly partitioned.

The use of the confusion matrix facilitated the evaluation of the classification performance. Table 1 illustrates the four components of the matrix, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

The samples that have been labeled as non-diabetes and diabetes are classified as the positive class and negative class, respectively. In this context, TP refers to the instances when non-diabetes samples are accurately categorized as "non-diabetes". FP represents the instances where samples labeled as "diabetes" are incorrectly classified as "non-diabetes". TN denotes the instances where diabetes samples are properly classified as diabetes. Lastly, FN represents the instances where samples labeled as "non-diabetes" are predicted as "diabetes". The assessment of performance was conducted using the metrics of precision, recall, F1 score, and accuracy, which were acquired from the confusion matrix. These metrics are represented by equations (1), (2), (3), and (4) correspondingly.

$$Precision\ (PPV) = \frac{TP}{TP + FP} \tag{1}$$

$$Recall\ (TPR) = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ Score = 2 * \frac{PPV * TPR}{PPV + TPR} \tag{3}$$

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

This section presents the classification accuracy and F-score of the seven algorithms used for the diabetes dataset. Table II provides a comprehensive summary of all the assessment measures.

TABLE II. COMPARISON OF ACCURACY, F-SCORE:

Model	**Accuracy**	**F-Score**
K-Nearest Neighbors	92.543%	0.93
Logistic Regression	92.543%	0.93

Support Vector Machines	94.471%	0.95
Naive-Bayes	90.638%	0.92
Decision Tree	94.227%	0.95
Random Forests	98.077%	0.98

Table II presents a comparative analysis of the performance of seven distinct machine learning methods. Random forests provide the highest classification accuracy, with a validation set accuracy of 98.0778%. The subsequent optimal outcomes are shown by the Multilayer Perceptron (MLP) model, followed by the Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel, achieving accuracies of 95.44% and 94.47% respectively. Random forests have the highest F-score performance, with a score of 0.98. Following random forests, support vector machines (SVM), multilayer perceptron (MLP), and decision tree models show F-scores of 0.9529, 0.9507, and 0.9503, respectively. Based on the findings shown in Figure 1, it can be seen that among the seven algorithms discussed in the research, Random Forest exhibits the highest level of performance in terms of classification accuracy and F-score values.

Conclusion:

The integration of technology into the medical field is a notable achievement so far. Machine learning models have the potential to accurately forecast a range of significant medical conditions, such as diabetes, in individuals at the initial stages when intervention and treatment are most effective. This study included the use of several classification methods to analyze a dataset pertaining to diabetes. A total of seven classification methods have been applied to the validation set of the dataset in question. The findings derived from the experimentation with many machine learning models provide convincing evidence that the Random Forest Classifier emerged as the most effective model among those used in the study for the given dataset. It achieved an accuracy score of 98.0778%, an ROC score of 0.9979, and an F-score of 0.9790. The three most effective classifiers for the given dataset are the Random Forests classifier, the Multi-layer perceptron, and the Support Vector Machine. Despite the fact that the other portions of the algorithm exhibited an accuracy of over 90%, as well as an F-score and ROC value surpassing 0.9, the random forest classifier distinguishes itself by achieving the highest score across all three assessment metrics. Therefore, based on the acquired findings, it can be confidently said that the Random Forest Classifier method is very successful when used in binary classification datasets. In order to achieve optimal accuracy, it is important to provide a larger quantity of training data to the Multi-layer Perceptron. This is a significant factor contributing to its poor performance in the dataset in question. In order to get a more exact and accurate categorization of the illness, it is essential to gather further data from various regions throughout the globe in the future. Subsequent research endeavors will prioritize the identification of additional variables that possess the capacity to induce diabetes, with the aim of incorporating

these possible components into the existing dataset to enhance the accuracy of categorization. This has the potential to facilitate the improvement and streamlining of illness diagnosis. Future research in the field of diabetes may be enhanced by investigating the condition via the use of diverse data mining and machine learning algorithms. This approach has the potential to improve the accuracy and timeliness of early diabetes prediction.

In further investigations, it is vital to prioritize the progression of algorithms inside interconnected fields and use innovative and efficient approaches to address existing challenges, including diverse deep-learning models. There exists a must to acquire supplementary data, encompassing indicators of standard of living and visual data, in order to augment the quality of data collection. Furthermore, it is essential to modernize the existing system and develop more precise models.

Reference:

[1] https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at an early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

[4] Kandhasamy, J Pradeep & Balamurali, Saminathan. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Computer Science. 47. 45-51. 10.1016/j.procs.2015.03.182.

[5] Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013 Feb;29(2):93-9. doi: 10.1016/j.kjms.2012.08.016. Epub 2012 Oct 16. PMID: 23347811.

[6] Nongyao Nai-arun, Rungruttikarn Moungmai, Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science, Volume 69, 2015, Pages 132-142, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.10.014. (https://www.sciencedirect.com/science/article/pii/S1877050915031786) [

7] Kavakiotis, Ioannis & Tsave, Olga & Salifoglou, Athanasios & Maglaveras, N. & Vlahavas, I. & Chouvarda, Ioanna. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. 15. 10.1016/j.csbj.2016.12.005.

[8] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017). Application of data mining methods in diabetes prediction. 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 1006-1010.

[9] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.

[10]Pradeep, K.R., & Naveen, N. (2016). Predictive analysis of diabetes using J48 algorithm of classification techniques. 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 347-352.

[11]Saxena, K., Khan, Z., & Singh, S. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm.

[12]T R, Prajwala. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. IJARCCE. 196-199. 10.17148/IJARCCE.2015.4142.

[13]Li, L. (2014, November). Diagnosis of diabetes using a weight-adjusted voting approach. In 2014 IEEE International Conference on Bioinformatics and Bioengineering (pp. 320-324). IEEE.

[14]Kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. International Journal of Engineering and Applied Sciences, 2(5), 257905.

[15]Perveen, Sajida & Shahbaz, Muhammad & Guergachi, Aziz & Keshavjee, Karim. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science. 82. 115-121. 10.1016/j.procs.2016.04.016.

[16]Ramzan, M. (2016). Comparing and evaluating the performance of WEKA classifiers on critical diseases. 2016 1st India International Conference on Information Processing (IICIP), 1-4.

[17]Saravananathan, K & T, Velmurugan. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i43/93874.

[18]Zheng, Tao & Xie, Wei & Xu, Liling & He, Xiaoying & Zhang, Ya & You, Mingrong & Yang, Guixin & Chen, You. (2016). A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records. International Journal of Medical Informatics. 97. 10.1016/j.ijmedinf.2016.09.014.

[19]Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. ARPN Journal of Engineering and Applied Science, 10(1), 8-13.