International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Diabetes Detection Using Supervised Machine Learning Classifiers

Neelam Barak[1]

[1]Department of ECE, Maharaja Surajmal Institute of Technology Delhi

---

Abstract - Diabetes mellitus is a chronic metabolic disease that has grown to be a major global health issue. Effective treatment and the avoidance of serious consequences depend on an early and precise diagnosis. A promising method for automated and precise diabetes detection in recent years has been machine learning (ML) approaches. The use of supervised machine learning techniques, such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Random Forest, for diabetes detection is examined in this study. A popular benchmark dataset in the field of diabetes research, the Pima Indian Diabetes Dataset, is used to assess the efficacy of these algorithms. With Logistic Regression obtaining the best accuracy, experimental results demonstrate the effectiveness of machine learning algorithms to predict diabetes. The study's conclusions demonstrate ML's potential as a valuable tool for early detection and prevention of diabetes.

***Key Words*:** Machine learning, Diabetes detection, Support vector machine, Logistic regression, Random forest

1. INTRODUCTION

The chronic illness known as diabetes mellitus is typified by high blood glucose levels. It is mainly divided into two categories: Type 1 diabetes, an autoimmune disease and Type 2 diabetes, a metabolic disease linked to insulin resistance. Diabetes can cause blindness, kidney failure, heart disease, and stroke, among other serious health issues. Conventional diabetes diagnostic techniques rely on expert views and clinical testing, which can be laborious and prone to human error. A viable substitute is machine learning, which analyses big patient data sets to find trends and forecast the risk of diabetes. Several studies have explored the application of ML techniques for diabetes detection. For instance, Mitushi Soni et al. used ML classification and ensemble techniques to make predictions about diabetes. They employed K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest algorithms. Sivaranjani S et al. used SVM and Random Forest (RF) methods for identifying potential risks of diabetes related diseases. After data preprocessing and implementing forward & backward stepwise feature selection was utilized to identify the most impactful features. They employed Principle Component Analysis to reduce dimensionality. Shejal Kale et al. applied ML Classification & ensemble techniques to make predictions about diabetes on a given dataset. Ashwini R et al. trained ML ALGO such as KNN, Random Forest(RF), Logistic Regression(LR), and SVM using various datasets. They used preprocessing techniques to improve the accuracy of their models and prioritized risk factors by employing various feature selection approaches. K.VijiyaKumar et al. proposed Random Forest algorithm for the prediction of diabetes and developed a system which can perform early prediction of diabetes for a patient with a higher accuracy. Nonso Nnamoko et al. presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a metaclassifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi et al. presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease. Deeraj Shetty et al. proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

In this work, we have made a machine learning based diabetes detection system for early detection of diabetes. We have used the Pima Indian Diabetes dataset, which is a well-known dataset in the field of diabetes research for early prediction of Diabetes.

2. METHODOLOGY

2.1 Dataset: The Pima Indian Diabetes Dataset, a widely used benchmark dataset in the field of diabetes research, was used for this study. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. The dataset contains 768 records of female patients. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. Second is the value of Plasma glucose concentration a 2 hours in an oral glucose tolerance test and then is the Diastolic blood pressure (mm Hg), fourth in line is the Triceps skin fold thickness (mm), then is the 2-Hour serum insulin (mu U/ml), sixth is Body mass index (weight in kg/ (height in m) ^2) and then seventh is the Diabetes pedigree function and the second last value is the that of the Age (years). The ninth column is that of the Class variable (0 or 1), 0 for no diabetes and 1 for the presence. The dataset contains information about 768 patients, including demographic data, medical history, and laboratory test results. Figure 1 shows the snippet of data heads of the Pima Indians Dataset.

	Pregnanci	Glucose	BloodPres	SkinThick	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Figure 1. Snippet of Pima Indians Dataset

2.2 Data Preprocessing

The process of preprocessing data is of utmost importance, especially for data concerned with healthcare, which may contain missing values and other contaminants that may affect the effectiveness of data mining. This process is essential to achieve accurate results and successful predictions. The PIMA dataset, which is quoted above, has lapsed and has shed data. To make the dataset serviceable and obtain knowledge from it, we have performed data preprocessing. The dataset was pre-processed to handle missing values and normalize the features. Missing values were imputed using mean imputation. Feature scaling was performed to ensure that features with different scales contribute equally to the model. Figure 2 shows the process of data preprocessing.

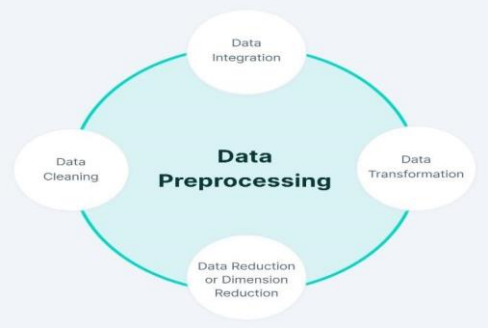

Figure 2. Data Pre-processing

A value of zero are removed since it is not possible to have a value of zero for certain features. This process helps in feature subset selection by eliminating irrelevant features or instances, reducing the dimensionality of the data and enabling faster processing. After cleaning the data, it is normalized and split into training and testing sets. The algorithm is trained on the training dataset, and the test dataset is kept aside. This training process produces a model based on logic, algorithms, and feature values in the training data. Normalization is used to bring all attributes to the same scale. The split the modelling dataset into training and testing sets is to assign 2/3 data points to the former and the remaining one-third to the latter. Therefore, we train the model using the training set and then apply the model to the test set. In this way, we can evaluate the performance of our model. For instance, if the training accuracy is extremely high while the testing accuracy is poor then this is a good indicator that the model is probably over fitted. A flow diagram of process methodology is shown in Figure 3.

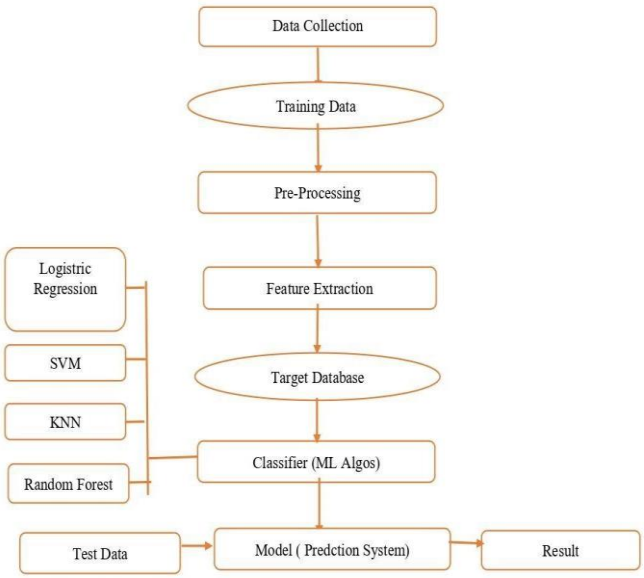

Figure 3. Diabetes Prediction Flow Diagram

2.3 Distribution of Diabetic patients- In our attempt to develop a diabetes prediction model, we encountered a slightly imbalanced dataset. Out of the total 768 samples, around 500 were designated as 0, denoting the nonexistence of diabetes, while 268 were designated as 1, denoting the existence of diabetes.

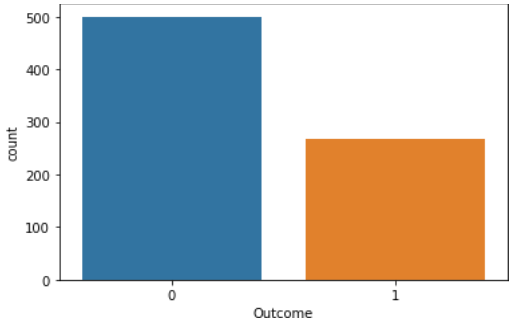

Figure 4. The proportion of patients with diabetes compared to those without diabetes

2.4 Supervised machine learning:

With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output.

2.4.1 Machine Learning Models:

2.4.1.1 Logistic Regression:

Logistic regression is a statistical method used to predict the probability of an event occurring. It's particularly useful for binary classification problems, where the outcome can be one of two categories e.g., yes/no. The input data consists of features and a binary target variable. The logistic regression model learns the relationship between the features and the probability of the target variable being 1. The trained model can then be used to predict the probability of the target variable being 1 for new data. In logistic regression, a hyperplane is used to separate data points into different classes. Data points on one side of the

hyperplane belong to one class, while those on the other side belong to the other class. The logistic regression model finds the optimal hyperplane that best separates the data as shown in Fig. 5.

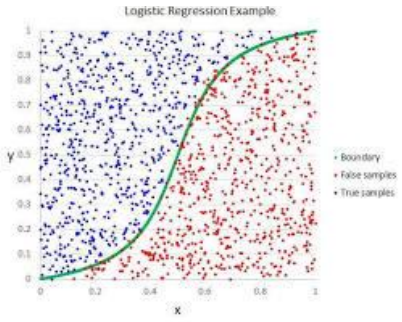

Figure 5. Logistic regression curve

2.4.1.2 K-Means Clustering:

K-means clustering is an unsupervised machine learning algorithm used to divide a dataset into a predetermined number of clusters (groups) based on feature similarity. It aims to group data points such that points within each cluster are as similar as possible to each other and as different as possible from those in other clusters. K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the cluster based on its distance from centroid of the cluster. After assigning each point to one of the cluster, new cluster centroids are assigned. This process runs iteratively until it finds good cluster. In the analysis we assume that number of cluster is given in advanced and we have to put points in one of the group. Fig.6 explains the working of the K-means Clustering Algorithm:

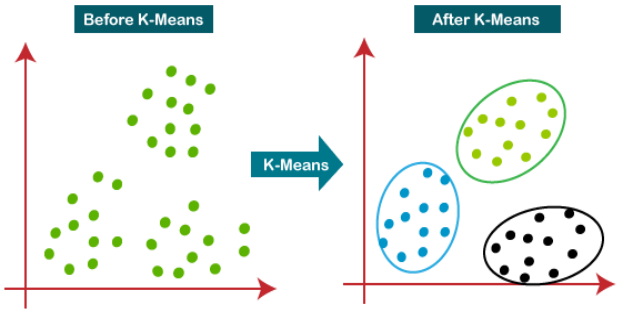

Figure 6. K-Means Clustering Algorithm

2.4.1.3 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Fig.7 explains the working of the Random Forest algorithm.

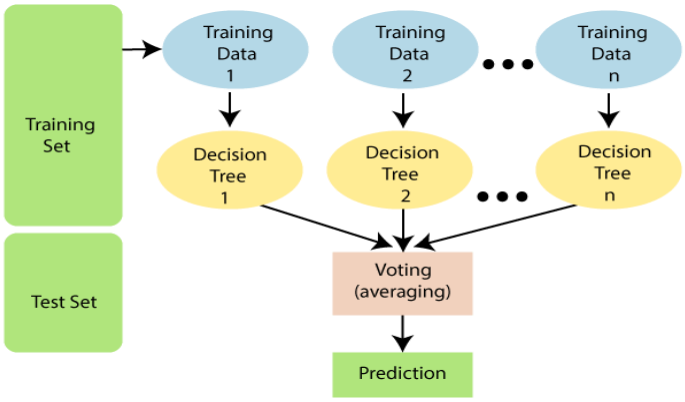

Figure 7. Random Forest Classifier

2.4.1.4 Support Vector Machine Algorithm

A support vector machine is a supervised learning algorithm that sorts data into two categories. It is trained with a series of data already classified into two categories, building the model as it is initially trained. The task of an SVM algorithm is to determine which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier. It is a supervised machine-learning problem where we try to find a hyperplane that best separates the two classes. SVM does this by finding the maximum margin between the hyperplanes that means maximum distances between the two classes as shown in Fig.8.

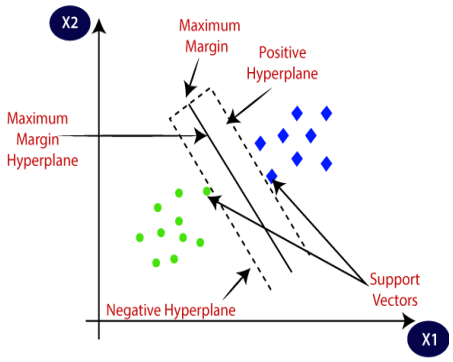

Figure 8. Support Vector Machine

2.5 Model Evaluation: The performance of the aforementioned machine learning models was evaluated for the given diabetes dataset. The models were fitted and the evaluation parameters were calculated. Using the standard metrics such as accuracy, precision, recall, and F1-score, the model's ability to correctly classify positive and negative instances was calculated.

2.6 Performance Analysis

In this paper, various machine learning algorithms like SVM, Decision Tree, Random Forest, and Logistic regression are used to predict diabetes. Diabetes Prediction dataset, has a total of 9 attributes, out of those only 9 attributes are considered for the prediction of Diabetes Prediction. Various attributes of the patient like Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, etc. are considered for this project. The accuracy of individual algorithms has to be measured and whichever algorithm gives the best accuracy, is considered for the diabetes prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

3. RESULTS AND DISCUSSION

The aim of this paper is to develop a system that can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. Algorithms like K nearest neighbor, Logistic Regression, Random Forest and Support vector machine are used. The accuracy of the models using each of the algorithms is calculated. The model with highest accuracy is considered as the optimal one for predicting diabetes.

3.1 Prediction Algorithm

The presence of disease has been identified using the appearance of various symptoms. However, the methods use different features and produce varying accuracy. The result of prediction differs with the methods/measures/ features being used. Towards diabetic prediction, a Disease Influence Measure (DIM) based diabetic prediction has been presented. The method pre-processes the input data set and removes the noisy records. In the second stage, the method estimates disease influence measure (DIM) based on the features of input data point. Based on the DIM value, the method performs diabetic prediction. Different approaches of disease prediction have been considered and their performance in disease prediction has been compared. The analysis result has been presented in detail.

The first step involves importing the necessary libraries and loading the diabetes dataset. In step two, the data is pre-processed to eliminate missing values. Step three involves splitting the dataset into training and test sets using an 80-20 percentage split. Next, the machine learning algorithms like logistic regression, support vector machine, Random forest, KNN, are selected in step four. Step five involves building the classifier model using the training set. The classifier model is then tested using the test set in step six. In step seven, a comparing and evaluating the performance results of each classifier is carried out. Finally, in step eight, after analysing the results based on various measures, the best performing algorithm is concluded. The figures showing relation between different attributes are shown in Figs.9 and 10.

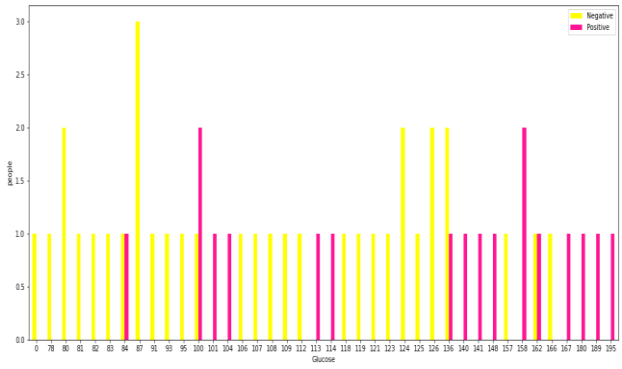

Figure 9. Comparing Glucose with the Outcome

Figure 10. Comparing Diabetes in function of age and Blood Pressure

After this, the pair plotting of different attributes with respect to each other are shown in Fig.11 to show the correlation of attributes. Figure 12 shows the plot of all columns (attributes) of the dataset when the patient is having diabetes i.e. when the outcome value is 1.

The experimental results demonstrate the effectiveness of ML techniques in accurately predicting diabetes. The table 1 displays the performance values of various classification algorithms, calculated using different measures. Logistic Regression achieved the highest accuracy of 77.78%, followed by SVM with 69.23%, Random Forest with 64.29% and KNN with 60%. These results indicate that ML algorithms can be a valuable tool for early detection and prevention of diabetes.

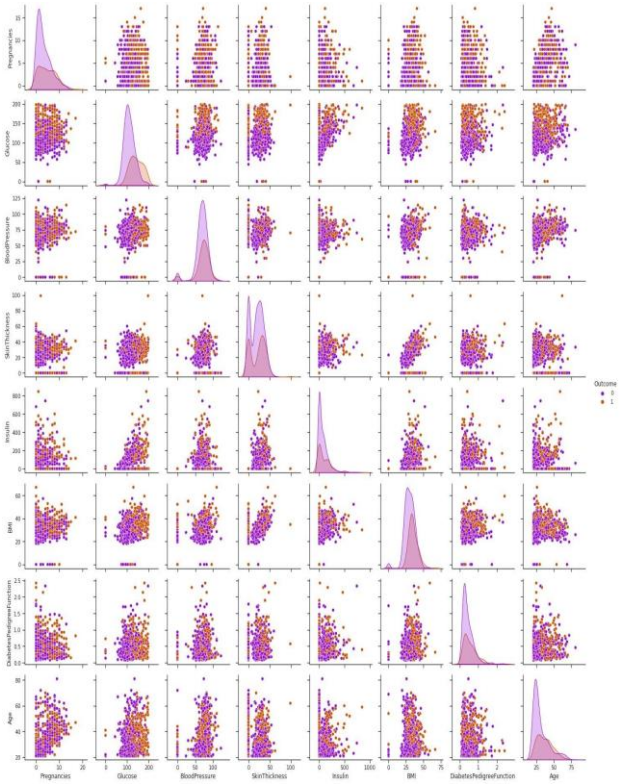

Figure 11. Pair plotting of data frame

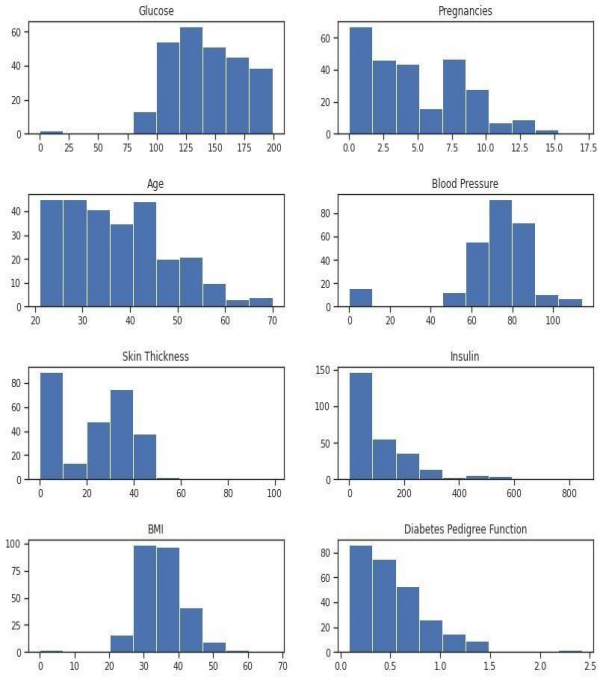

Figure 12. Comparing all columns when the Outcome is 1(has Diabetes)

Based on the table 1, it is observed that Logistic regression exhibits the highest accuracy. Therefore, the Logistic regression machine learning classifier is capable of predicting the likelihood of diabetes with greater precision than other classifiers. So with the help of above mentioned procedure, a Diabetes Prediction System using Machine Learning has been presented to improve diabetes detection process using machine learning. In this project, the data was formulated in different formulations and the model was trained with above 80% accuracy.

Table 4.1 Evaluation parameters for different classifiers

Classification Algorithm	Precision	Accuracy	F1-score	Recall
Logistic Regression (LR)	83.33 %	**77.78 %**	83.33%	83.33%
Support Vector Machine (SVM)	77.78 %	69.23 %	77.78%	77.78%
Random Forest (RF)	77.78 %	64.29 %	73.68%	70%
KNN	75%	60%	66.67%	60%

4. CONCLUSIONS

The efficiency of Logistic Regression with other linear classifiers including SVM , KNN, and Random Forest has been presented for early detection of diabetes. The results of the comparison revealed that Logistic Regression outperformed all the other classifiers. The accuracy of Logistic Regression was found to be the highest, at 77.78%. The proposed approach utilized ensemble learning and classification methods, which resulted in high accuracy levels. These experimental results can assist healthcare professionals by enabling early predictions and informed decisions. The main aim of this paper was to design and implement Diabetes Prediction Using Machine Learning Methods and performance analysis of these methods and it has been achieved successfully In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. These classifiers can aid in the treatment of diabetes and potentially save human lives.

REFERENCES

1. Arwatki Chen Lyngdoh, Nurul Amin Choudhury, Soumen Moulik, “Diabetes Disease Prediction Using Machine Learning Algorithms”, IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) 2020, DOI: 10.1109/IECBES48179.2021.9398759 .
2. Mitushi Soni, Dr. Sunita Varma " Diabetes Prediction using Machine Learning Techniques" , Journal of Engineering Research & Technology (IJERT) 2020,
3. Sivaranjani S, Ananya S, Aravinth J, Karthika R, " Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction”, 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021 DOI:10.1109/ICACCS51430.2021.9441935 .
4. Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal PatilB “Diabetes Prediction Using Different Machine Learning Approaches” IJSDR | Volume 7 Issue 5 , 2022.
5. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942928, 2018.
6. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
7. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.
8. G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1009- 1014, 2020.
9. M. F. Faruque, Asaduzzaman and I. H. Sarker, ”Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus”, 2019 International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 14, 2019.
10. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, ”Type 2 diabetes mellitus prediction model based on data mining”, Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.
11. Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of BigData ,6(1), p.13.(2019)
12. Sisodia,D. and Sisodia,DS,2018.Prediction of diabetes using classification algorithms. Procedia computer science,132, pp.1578-1585. (2018)
13. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
14. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to

Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

15. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942928, 2018.
16. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
17. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.
18. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.
19. Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.
20. A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.
21. Azra Ramezankhani, Omid Pournik, Jamal Shahrabi, Fereidoun Azizi and Farzad Hadaegh, "An Application of Association Rule Mining to Extract Risk Pattern for Type 2 Diabetes Using Tehran Lipid and Glucose Study Database", Int J Endocrinol Metab, April 2015.
22. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.
23. A Abbasi, LM Peelen, E Corpeleijn, YT van der Schouw, RP Stolk, AM Spijkerman et al., "Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study", BMJ, 2012.
24. "Mining constrained association rules to predict heart disease", IEEE 13th International Conference on Data Mining, pp. 433, 2010.