International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 10, Issue 2, Jul–Dec 2021; 1–8

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

COMPARISON OF DATA MINING ALGORITHMS FOR DIAGNOSIS OF DIABETES MELLITUS

Ahmed Sami Jaddoa[1] & Ziyad Tariq Mustafa Al-Ta'i[2]

[1]Research Scholar, Business Informatics College, University of Information Technology and Communications, Iraq

[2]Professor, Department of Computer Science, Science College, Diyala University, Iraq

ABSTRACT

Diabetes is specified as the most chronic and deadliest disease that results in increasing blood sugar. The medical data mining approaches were utilized for detectingun observed patterns in the medical field sof sets of data for medical diagnosis and treatment. Data classification for diabetes mellitus is quite significant. Where utilizing two types of data sets, the first is local, collected from consulting laboratories at Baqubah General Hospital, and the second is global, which is the Pima India Diabetes Database. The experiment on the Local dataset shows that the accuracy if K-NN is 90 %, the accuracy of the SVM has been 98 %, the accuracy of the NB is 98 % and the accuracy of RF is 98 %. The experiment on the Pima dataset shows that the accuracy of K-NN is 81 %, the accuracy of SVM has been82 %, the accuracy of NB is 84 % and the accuracy of RF is 82 %.

KEYWORDS: *Diabetes Mellitus, Data Mining, Diagnosis, K Nearest Neighbors, Classification, Support Vector Machine, Naive Bayes, Random Forest*

Article History

Received: *21 Jun 2021* **|** ***Revised:*** *12 Jul 2021* **|** ***Accepted:*** *16 Jul 2021*

INTRODUCTION

Diabetes can be defined as one of the chronic diseases marked by elevated blood glucose levels in the body. Also, diabetes causes damage to the heart, kidneys, and eyes over time [1]. Diabetes mellitus (DM) is one of the world's deadliest yet most frequent diseases. It has a global impact on millions of individuals. It is caused by a high sugar diet as well as other lifestyle choices and unhealthy eating habits, including the lack of regular physical activity. The disease's onset is also influenced by genetics [2]. The International Diabetes Federation states that a total of 463 million individuals worldwide have diabetes in 2019, with that number projected to increase to 578 million and 700 million by 2030 and 2045 respectively [3]. DM is a common disease that happens when the pancreas has no ability for producing sufficient amount of insulin or when the cells in the body gain insulin resistance. Diabetes impairs the human body's ability to use energy contained in food [4].

The following are the most common diabetes types: Type-1 diabetes is caused by the body's inability for producing insulin. Juvenile diabetes, early onset diabetes or Insulin-dependent diabetes is all terms used to describe such type of diabetes. Type-1 diabetes typically strikes prior to the age of 40, commonly in teenage years or early adulthood. It is most likely to happen in young people and children. Type-2 diabetes happens in the case where the body does not produce sufficient amount of insulin for functioning properly, or when the body cells don't respond to insulin

(insulin resistance (IR)). It is most common in people over the age of 40. Gestational diabetes is the Type-3 diabetes that affects women throughout pregnancy. A few females' blood glucose levels are extremely high, while their bodies have no ability for producing sufficient amount of insulin for transporting all glucose into their cells, leading to steadily increasing glucose levels. Gestational diabetes is diagnosed throughout pregnancy [5]. The process of extracting knowledge is known as data mining (DM). As a result, disease predictability is going to improve, and the disease's early detection will aid patient care [6]. Classification is a crucial task in ML and DM, since it aims to categorize each one of the instances in a data set into distinct groups on the basis of the information identified via its features. Furthermore, one of the major DM tasks is data classification. We are attempting to create a classifier which identifies diabetes at the lowest cost and with the best performance [7].

RELATED WORK

Different Related Works Were Suggested in the Field of Diabetes, Such As:

MiteshWarke et al.[8], proposed a system for Diabetes Diagnosis by means of ML Algorithms, which utilized Decision Tree, KNN, Naïve Bayes, and SVM. In addition, the efficiency regarding the developed model was assessed through a database from the Pima India Diabetes dataset. The suggested approach for records classification with NB achieved an accuracy of 72 %, whereas the Decision Tree algorithm achieved 68 % accuracy, the SVM algorithm achieved 62 % accuracy, and the KNN algorithm achieved 66 % accuracy.

Md. Faisal Faruque, Asaduzzaman, and Iqbal H. Sarker[9], Suggested a system for Performance Analysis of ML approaches for predicting DM, which used SVM, Naïve Bayes, KNN, and C4.5 algorithms. The suggested technique for records' classification with NB achieved an accuracy of 68 %, whereas with the C4.5algorithm achieved 73 % accuracy, the SVM algorithm achieved 70 % accuracy, and the KNN algorithm achieved 71 % accuracy.

Kucharlapati Manoj Varma and Dr B S Panda[10], Compared the performance analysis of RF, KNN, C5.0, and SVM to predict diabetes using Machine Learning Techniques. The developed model’s efficiency was assessed via a database from the Pima India Diabetes dataset. The suggested technique for records' classification with KNN achieved an accuracy of 73.57 %, whereas the Random Forest algorithm achieved 74.67 % accuracy, the C5.0 algorithm achieved 74.63 % accuracy and the SVM achieved 72.17 % accuracy.

Aswan Supriyadi Sunge et al.[11], Proposed a system use C4.5 algorithm to predication diabetes mellitus. The developed model’s efficiency is accessed via a database from the Pima India Diabetes dataset. This algorithm gives the accuracy of 72.08 %.

NazimRazali et al.[12], Proposed a system using many DM approaches, like NB, Rep Tree, Sequential Minimal Optimization (SMO), and Simple Logistic Regression to classify if negative or positive result of diabetes diagnostic. The developed model’s efficiency is accessed via a database from Pima India Diabetes dataset. These techniques give the accuracy of 73.60 % for Naive Bayes, whereas give the accuracy of 75.70 % for Simple Logistic Regression, give the accuracy of 75.10 % for Rep Tree and give the accuracy of 74 % for Sequential Minimal Optimization (SMO).

Dr. GarimaVerma and Dr. HemrajVerma[13], Proposed a system that uses the Multilayer Perception Neural Network model (MLP) to predication diabetes mellitus. The developed model’s efficiency was assessed via a database from the Pima India Diabetes dataset. This model gives an accuracy of 82 %.

DATA COLLECTION

Local Dataset

The information was gathered from Baqubah General Hospital's consulting labs in Iraq's Diyala Governorate. There are approximately 250 instances in this data collection. In the dataset, each person is identified by ten attributes. Both male and female genders were included in the data collection, which was at least 5 years old. The answer variable is conditional and takes the values 0 or 1, with 0 indicating negative results for DM and 1 indicating a positive result for DM. In class 1, there are 152 cases and in class 0, there are 98 cases. Table (1) describes the collected local dataset.

Table 1: Description of Collected Local Dataset

Att. No	Attribute	Description	Attribute Specification
1	Gender	Gender for each person	1=Male 0=Female
2	Age	Age for each person-years	5-74 years
3	HbA1C	Hemoglobin A1C for each person 3-month plasma glucose concentration	4.80-5.90 %
4	Glucose	Glucose levels in the blood for each person	4.11-6.05 mmol/L
5	HDL Cholesterol	High-density lipoprotein cholesterol for each person	0.78-1.6 mmol/L
6	LDL Cholesterol	Low density lipoprotein cholesterol for each person	0-2.6 mmol/L
7	Total Cholesterol	Total amount of cholesterol in the blood	<5.2
8	Triglyceride	Triglyceride for each person	0.86-1.9 mmol/L
9	Creatinine	Creatinine for each person	62-106 µmol/L
10	Class	Diagnosis of disease	1=True 0=False

Global Data

The National Institute of Diabetes and Digestive and Kidney Diseases developed the Pima Indian Diabetic Database (PIDD). The information was gathered from Data World, Sets of data. According to open records, the whole patients are Pima Indian women who are at least 21 years old. There are a total of 768 cases, 268 of which are diabetic and 500 of which are not diabetic. The Pima class label categorized into (0 = False) indicates absence and (1 = True) presence of diabetes disease. Table 2 shows the Pima attributes description.

Table 2: Description of Pima Indian Diabetic Dataset

Att. No	Attribute	Description
1	Pregnant	Number of pregnancy
2	Glucose	Concentration of the plasma glucose a 2h interval in oral glucose tolerance test
3	Thickness of the Skin	Triceps skin fold thickness (mm)
4	Blood pressure	The pressure of diastolic blood(mm Hg)
5	Insulin	2h serum insulin (mu U/ml)
6	Diabetes pedigree function	The function of diabetes pedigree
7	Age	Age (in years)
8	BMI	weight in kg/ (height in m) ^2)
9	Outcome	Class variable ((0=False) for tested negative for the diabetes and (1=True) for tested positive for the diabetes)

Proposed Model

The suggested methodology in this paper has been summarized in the diagram shown in Figure 1.

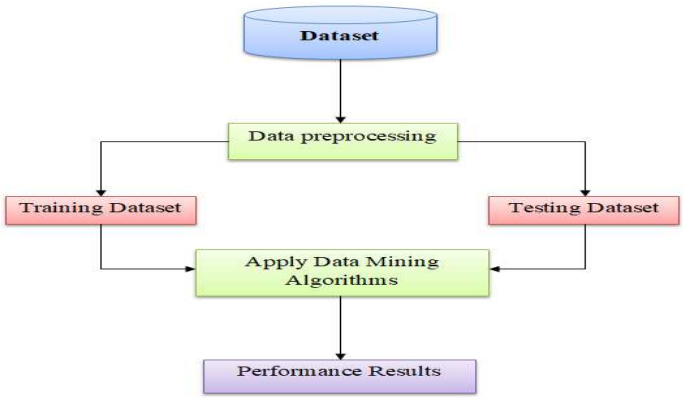

Figure 1: Block Diagram of the Suggested Model.

Data Preprocessing

Preprocessing of data represents an essential initial step applied to raw data to prepare them for the analysis. Tools of analytic could give wrong results and be misled if impurities are included in the data like missing data. Hence, preprocess the data is essential before implementing the process of data analysis.

Data Cleaning

Data cleaning can be defined as the process that is utilized to ensure that data is clear and ensure that it is prepared for additional processing. Filling the missing data is a data cleaning process. There are many ways available to fill absent values such as removing records that have absent values or replacing them with casual values or substituting those absent values by the average value of the obtainable ones. In this work, record deletion to handle missing values was utilized.

Feature Normalization

This step is often adopted before the design of the classifier because it considers as a precaution when the feature values vary in different dynamic ranges. If normalization is not used, attributes with large values have a considerable impact on the design of the classifier. So the normalization role is to put all values within specific ranges features values are normalized by using the Min-Max method.

The initial data is transformed linearly by min-max normalization. Assume that $min_A$and $max_A$are maximum and minimum values for attribute A, respectively. Equation (1) indicates that the min-max normalization maps a value v of A to V′ in the range [new_min_A, new_max_A].

$$V' = \frac{V - min_A}{max_A - min_A}(new_{max_A} - new_{min_A}) + new_{min_A} \quad (1)$$

Where V′ represent features that normalized.

Data Mining Algorithms

These types of algorithms fall under a supervised learning method which might be achieved on any data type. Also, classification learns from input data and after that based on it categorize the new data. Various classification algorithms are follows.

K-Nearest Neighbors

K-NN can be defined as one of the simple classification algorithms on the basis of finding closest K neighbors throughout training phase. In addition, a similarity metric for calculating the distance between the value of K and objects, which is representing the number of the nearest neighbors, in which the value of K is (11).

Support Vector Machines

SVM model has been specified as a supervised learning algorithm which is utilized for regression and classification, but mainly it is applied for classification problems. In SVM, the major aim is dividing the data-sets into classes for finding the maximum marginal hyper plane (MMH). In addition, the SVM algorithm was achieved with a kernel that is transforming the input data space into the needed form. The linear kernel was utilized in this study for implementing SVM.

Naive Bayes

NB classification algorithm, can be defined as a probabilistic classifier which is on the basis of Bayesian theorem with the assumption of the independence between the predictors. In addition, NB technique takes the dataset as input, performing analysis as well as predicting the class label with the use of Bayes Theorem. Also, it is calculating a probability of class in input data and allows predicting the class regarding the unknown data sample. It is a significant classification approach adequate for large data sets.

Random Forest

Random forest can be defined as one of the supervised learning algorithms that is primarily utilized to solve classification issues. A forest is comprised of trees, and the more trees there are, the more robust the forest becomes. Comparably, random forest constructs decision trees from the data samples, extracts predictions from each, and after that vote on optimal solution. It's an ensemble method that's better compared to a single decision tree since it performs the averaging of the results in order to reduce overfitting.

Algorithm 1: Apply Classification Algorithms

Begin

Step 1: Input the dataset after preprocessing

Step 2: Classify the dataset into training and testing dataset

Step 3: Specify algorithms that are used in model (KNN, SVC, NB and RF)

Step 4: algorithms = [KNN (), LinearSVC (), Random Forest Classifier (), GaussianNB()]

Step 5: Apply algorithms

Step 6: Find (Accuracy, confusion_matrix, classification_report)

End

EVALUATION CRITERIA

The accuracy concern is evaluated via using a confusion matrix that is composed of the concepts "False Positive" (FP),"True Positive" (TP), "False Negative" (FN) and "True Negative" (TN) [14].

Accuracy: refers to the ratio of total number of samples that have been correctly classified byclassifier to the total number of the samples.

$$Accuracy = \frac{TP+TN}{TP+TN+TN+FN} \quad (2)$$

Recall: refers to the TP and is expressed as,

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Precision: represents the TN and is expressed as,

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

F- Measure: is the one has the combination of both precision and recall which is used to compute the score.

$$F\text{- }Measure = \frac{2*Recall*Precision}{Recall+ Precision} \quad (5)$$

RESULTS

Following the selection of the two datasets, they were subjected to four classification algorithms (KNN, SVM, NB and Random Forest) in order to determine which technique produces the best performance (accuracy) on the same dataset.

Evaluation results are given in Table 3 and Figure 2 on Local Dataset.

Evaluation results are given in Table 4, table and Figure 3 on Pima Dataset.

Table 3: Performance Evaluation of Classification Algorithms

Model	Time (s)	Precision	Recall	f-Score	Accuracy	Confusion Matrix	
K-NN	0.03 sec	90 %	90 %	90 %	90 %	19	2
						3	26
SVM	0.02 sec	98 %	98 %	98 %	98 %	25	0
						1	24
NB	0.02 sec	98 %	98 %	98 %	98 %	21	0
						1	28
RF	0.22 sec	98 %	98 %	98 %	98 %	22	1
						0	27

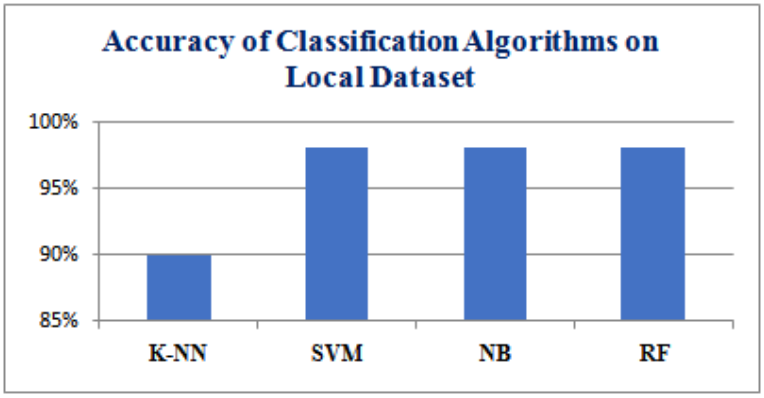

Figure 2: Accuracy of LR and KNN on Local Dataset

Table 4: Performance Evaluation of Classification Algorithms

Model	Time (s)	Precision	Recall	f- score	Accuracy	Confusion Matrix	
K-NN	0.03 sec	75 %	77 %	76 %	81 %	43	8
						5	12
SVM	0.03 sec	78 %	80 %	79 %	82 %	42	7
						5	14
NB	0.02 sec	84 %	79 %	81 %	84 %	42	3
						8	15
RF	0.23 sec	86 %	78 %	80 %	82 %	40	1
						11	16

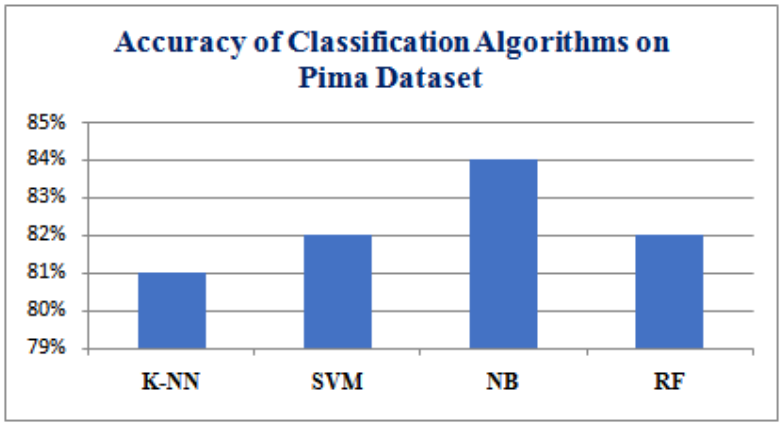

Figure 3: Accuracy of LR and KNN on Local Dataset

CONCLUSIONS

DM is one of the world widespread and complex diseases. By combining two datasets, the proposed model will diagnose diabetic and non-diabetic individuals. Result of classification on the Local dataset shows that the accuracy of K-NN is 90 %, the accuracy of SVM has been98 %, the accuracy of NB is 98 % and the accuracy of RF is 98 %.On the Pima dataset shows that the accuracy of K-NN is 81 %, the accuracy of SVM has been 82 %, the accuracy of NB is 84 % and the accuracy of RF is 82 %. This proves classification algorithms on Local dataset gives better performance than Pima dataset.

REFERENCES

1. *F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018, no. May, pp. 414–418, 2018, doi: 10.1109/ICOEI.2018.8553959.*
2. *S. M. Jacob, K. Raimond, and D. Kanmani, "Associated machine learning techniques based on diabetes based predictions," 2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019, no. Iciccs, pp. 1445–1450, 2019, doi: 10.1109/ICCS45141.2019.9065411.*
3. *P. Saeedi, I. Petersohn, P. Salpea, B.Malanda, S.Karuranga, N. Unwin, S.Colagiuri, L.Guariguata, A.A. Motala, K.Ogurtsova, J.E. Shaw, D.Bright, R.Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition", Diabetes Res. Clin. Pract., vol. 157, p. 107843, 2019, doi: 10.1016/j.diabres.2019.107843.*

4. A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.

5. B. S. Lal, "DIABETES : CAUSES, SYMPTOMS AND TREATMENTS DIABETES : CAUSES, SYMPTOMS AND TREATMENTS," no. December, 2016.

6. S. Mirza, S. Mittal, and M. Zaman, "Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree," Int. J. Appl. Eng. Res., vol. 13, no. 11, pp. 9277–9282, 2018, [Online]. Available: https://www.ripublication.com/ijaer18/ijaerv13n11_73.pdf.

7. M. Zhang, S. Member, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach", Ieee Trans. Cybern., pp. 1–16, 2012.

8. M. Warke, V. Kumar, S. Tarale, P. Galgat, D. C.- Diabetes, andundefined 2019, "Diabetes Diagnosis using Machine Learning Algorithms," Academia.Edu, pp. 1470–1476, 2019, [Online]. Available: http://www.academia.edu/download/60380576/IRJET-V6I327720190824-111158-dpcyom.pdf.

9. M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 1–4, 2019, doi: 10.1109/ECACE.2019.8679365.

10. K. M. Varma and Dr. B.S. Panda, "Comparative analysis of Predicting Diabetes Using Machine Learning Techniques," Jetir, vol. 6, no. 6, pp. 522–530, 2019, [Online]. Available: www.jetir.org.

11. A. S. Sunge, H. L. H. S. Warnar, Y. Heryadi, E. Abdurachman, B. Soewito, and F. L. Gaol, "Prediction Diabetes Mellitus Using Decision Tree Models," 2019 Int. Congr. Appl. Inf. Technol. AIT 2019, 2019, doi: 10.1109/AIT49014.2019.9144971.

12. N. Razali, A. Mustapha, S. Z. S. Idrus, M. H. A. Wahab, and S. A. F. Madon, "Analyzing Diabetic Data using Classification," J. Phys. Conf. Ser., vol. 1529, no. 2, 2020, doi: 10.1088/1742-6596/1529/2/022105.

13. G. Verma, H. Verma, I. Technology, and M. Studies, "A Multilayer Perceptron Neural Network Model For Predicting Diabetes," vol. 13, no. 1, pp. 1018–1025, 2020, doi: 10.13140/RG.2.2.23203.89126.

14. J. B. Raja and S. C. Pandian, "PSO-FCM based data mining model to predict diabetic disease," Comput. Methods Programs Biomed., vol. 196, 2020, doi: 10.1016/j.cmpb.2020.105659.