Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 9749579, 11 pages
https://doi.org/10.1155/2022/9749579

Research Article

Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach

Rasool Jader and Sadegh Aminifar

Computer Science, Faculty of Science, Soran University, Soran 44008, Kurdistan, Iraq

Correspondence should be addressed to Sadegh Aminifar; saminifar@yahoo.com

Received 1 June 2022; Accepted 8 October 2022; Published 22 October 2022

Academic Editor: Ridha Ejbali

Gestational diabetes is a type of high blood sugar that develops during pregnancy. It can occur at any stage of pregnancy and cause problems for both the mother and the baby, during and after birth. The risks can be reduced if they are early detected and managed, especially in areas where only periodic tests of pregnant women are available. Intelligent systems designed by machine learning algorithms are remodelling all fields of our lives, including the healthcare system. This study proposes a combined prediction model to diagnose gestational diabetes. The dataset was obtained from the Kurdistan region laboratories, which collected information from pregnant women with and without diabetes. The suggested model uses the clustering KMeans technique for data reduction and the elbow method to find the optimal k value and the Mahalanobis distance method to find more related cluster to new samples, and the classification methods such as decision tree, random forest, SVM, KNN, logistic regression, and Naïve Bayes are used for prediction. The results showed that using a mix of KMeans clustering, elbow method, Mahalanobis distance, and ensemble technique significantly improves prediction accuracy.

1. Introduction

According to the World Health Organization (WHO), over 1.5 million people die yearly from diabetes. Gestational diabetes is one of the most prevalent pregnancy complications, affecting approximately one in six babies worldwide [1]. According to the International Diabetes Federation, gestational diabetes mellitus (GDM) is a severe and underrecognized danger to mother and infant health. Many women with gestational diabetes will experience complications during their pregnancies, including high blood pressure and birth weights. Within five to ten years following childbirth, around 50% of women with a history of GDM develop type 2 diabetes. [2]. GDM is a prevalent metabolic illness that is typically a temporary pregnancy disorder. Women with gestational diabetes mellitus are at an increased risk for poor pregnancy outcomes that compromise a normal birth [3]. All international healthcare organizations urge that women should be evaluated for hyperglycemia risk at the initial prenatal exam, as this allows for early detection of the condition. Women with diabetes in pregnancy or GDM must carefully maintain and monitor their blood glucose levels with the assistance of their healthcare professionals to avoid the risk of bad pregnancy outcomes. Unfortunately, there are only periodic tests available for pregnant women in the Kurdistan region of Iraq, and the necessary attention has not been paid to this issue. Many previous papers have worked on data from other regions. This encouraged us to collect data in this area, and we were able to obtain diabetes tests from 1012 pregnant women. Of these, 217 tests were suffering from GDM, which is not a good result. The collected data's characteristics, which include age, weight, height, number of pregnancies, heredity, and diabetes tests, reveal when and under what conditions pregnant women are more likely to develop gestational diabetes.

This study will attempt to create a model that employs techniques for machine learning to examine both new and old cases of diabetes and diagnose new problems. Both the patients and hospital management will benefit from this. The proposed model employs the clustering KMeans technique for data reduction, the elbow method for determining the optimal k value, the Mahalanobis distance method for identifying the cluster most closely related to new samples, and the classification methods of decision tree, random forest, SVM, KNN, logistic regression, and Naive Bayes for prediction. Classification algorithms used the ensemble max voting approach to get the optimal outcome. In this strategy, predictions are created for each data point using classification models, and the ensemble method makes the final decision based on all the mentioned classification methods. Each data test goes through the ensemble operation. Using a combination of KMeans clustering, elbow technique, Mahalanobis distance, and ensemble technique considerably enhances prediction accuracy, as seen in the results. The remains of this study are structured as follows. Section 2 explains several studies of related work. Section 3 describes the methodology of the proposed model and some discussion of the results. Finally, Section 4 expresses the conclusion and what we want to do in future work.

2. Literature Review

This section briefly discusses several works on intelligent systems and machine learning methods for modelling and predicting different types of diabetes diagnoses. Table 1 illustrates the summaries and differences between the related works.

For diabetes prediction, Al-Zebari and Sengur [4] analyzed and compared the outcomes of different machine learning strategies. In this study, the MATLAB classification learner tool was utilized, and several machine learning techniques such as decision trees, support vector machines, K-nearest neighbor, and logistic regression were utilized. When it comes to performance measurements, the outcomes are judged according to how accurately they are classified.

Rising blood plasma sugar levels cause diabetes. According to [5], various intelligent systems used classifiers to predict diabetes using machine learning methods such as decision tree, SVM, Naive Bayes, and ANN. This study suggests a decision assistance system that incorporates AdaBoost and the decision stump. AdaBoost uses SVM, NB, and DT as basis classifiers to assess correctness. AdaBoost is more accurate than support vector, with a decision stump as the basic classifier. This research proposes a diabetes prediction system that leverages a decision stump in AdaBoost. The study employs a 768-instance, 9-attribute global training dataset from UCI's machine learning repository. They utilize the Kerala data for validation. The AdaBoost decision stump (DS) classifier can predict diabetes with an accuracy of 80.729% and a 19.27% error rate. The accuracy of the decision support system might be enhanced by adding other powerful classifiers, such as artificial neural network and K-nearest neighbor, or by combining several classifiers using local datasets from different states.

In another way [6], the different machine learning methods for predicting gestational diabetes for pregnant women were displayed using the PIMA Indian dataset. To make it more accurate, the author cleans the data at first. The accuracies of all methods were compared using receiver operating characteristic (ROC) and area under the curve (AUC) scores. The result of the confusion matrix of algorithms was illustrated to determine the efficacy and errors of the models. This research proves that the accuracy of the machine learning algorithms may be improved by adjusting their parameters. In this study, the researcher used the ensemble method, which is a combination of several machine learning methods. As shown in the result, the accuracy of the techniques used in ensemble learning used XGBoost had the best accuracy and result of 77.5%.

Alehegn et al. [7] analyzed 768 samples of data from the (PIDD) Pima Indian Diabetes dataset by using predictive algorithms such as K-nearest neighbor, Naive Bayes, random forest, and J48 to create an ensemble learning by merging particular machine learning techniques within one to enhance the accuracy of the suggested system's performance. Researchers used different data mining methods and machine learning algorithms to examine different medical datasets. This showed that machine learning methods work differently with different sets of data and talked about how the single algorithm was less accurate than the ensemble algorithm. The authors discussed that, in most studies, the accuracy of the decision tree was high. In their research, they use tools to predict diabetes datasets and a hybrid system made up of Weka and Java.

From another perspective [8], the study aimed to estimate a patient's diabetes risk more precisely. A model construction uses classification techniques such as decision tree, ANN, Naive Bayes, and SVM. The decision tree, Naive Bayes, ANN, and SVM models all have an accuracy of 74%. Results show the procedures accuracy. After receiving the input dataset, the authors' recommended model would forecast the data using ML algorithms and offer a comparison to estimate the greatest accuracy for treating diabetes. Support vector machines are great when they do not know the data. SVMs do well when they do not know the data. SVM works effectively with semi-structured and unstructured data, including text, pictures, and trees. Some parameters must be set to get the best classification results using SVM. The algorithm needs this. The decision tree is easy to understand and implement. Instability in the decision tree may be discovered by making minor changes to the optimal decision trees data structure. They are often somewhat off. Naive Bayes skips probability estimate calculation for missing values. It is suitable for huge datasets. Expanding training datasets increases bias. An ANN is accurate and easy to use. Processing detailed data is difficult and time-consuming.

In the other way [9], the Mahalanobis distance metric is employed by the authors as a restricted optimization problem with the ratio of distances as the goal function. To overcome this issue, an optimization technique is suggested. A lower limit and an upper bound, including the optimal, are computed directly and then utilized to provide the beginning value for iterations. Experiments indicate that their solution is

Table 1: Summarizes and differences of the related works.

No	Title	Reference	Advantages	Outcomes
1	"Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection"	[4]	Found that the LR has the best accuracy because of categorical data	DT 75.3% LR 77.9%% SVM 77.6%% KNN 76%
2	"Prediction and Diagnosis of Diabetes Mellitus: A Machine Learning Approach"	[5]	The results show that adaptive boosting with the decision stump's base as a classifier is more accurate	Adaboost% DT-Base 77.6%% SVM-Base 77.6%% DS-Base 80.72%
3	"Prediction of Gestational Diabetes by Machine Learning Algorithms"	[6]	They proved that the ensemble learning used XGBoost has the greatest accuracy	Adaboost 76.2% GBM 76.5% XGBoost 77.5%
4	"Analysis and Prediction of Diabetes Diseases using Machine Learning Algorithm: Ensemble Approach"	[7]	It is found that different datasets affect machine learning algorithms' accuracy and single algorithm is less accurate than ensemble learning	Each algorithm has different accuracy on different datasets
5	"Diabetes Prediction Using Different Machine Learning Approaches"	[8]	It has been shown that the SVM method will gain more accuracy when we have no prior knowledge of the data	DT 74% SVM 82%% NB 80%% ANN 81%
6	"The Mahalanobis distance"	[9]	They show the effect of data variance	The MD's results are fewer and more accurate than ED
7	"Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis"	[2]	The meta-analysis and findings of heterogeneity were done with the help of the Meta Disc software	Age, heredity, BMI, and fasting blood glucose were the most common features used to build models
8	"A comparative analysis of KNN, GA, SVM, DT, and LSTM algorithms in machine learning"	[10]	The performance of five essential machine learning algorithms is compared and demonstrated	SVM algorithm has provided one of the best results in predictive analytics in real-time applications
9	"Discovering Tree Based Diabetes Prediction Model"	[11]	This study focuses on essential features; this puts a lot of effort into the data mining and reduces the complexity of predicting model	Feature selection, and prediction model by DT algorithm, obtain a good result

an excellent performance distance metric with a limited number of paired constraints. In this study, techniques based on Mahalanobis distance and its effects in different areas of chemometrics, such as multivariate calibration, pattern recognition, and process control, were explained and talked about. And they discussed how the Mahalanobis distance was different from the Euclidean distance.

Zhang et al. [2] measured the risk of bias in the machine learning models with the new prediction model risk of bias assessment tool. The meta-analysis and findings of heterogeneity were done with the help of the Meta Disc software. They also did sensitivity analyses, meta-regressions, and subgroup analyses to reduce the effect of heterogeneity. Twenty five studies with more than 18 years old women who had no history of serious illness were looked at. The pooled area under the receiver operating characteristic curve for machine learning models predicted gestational diabetes was 84%, the recall was 69%, and the precision was 75%. Logistic regression, one of the most commonly used ML methods, had an overall proportion of 81%. In contrast, nonlogistic regression models did better, with an overall pooled of 88%. Also, the age of the pregnant women, a history of diabetes in the family which is called heredity, body mass index (BMI), and fasting blood glucose were the four most common features used to build models using the different feature selection methods.

We compare KNN, GA, SVM, DT, and LSTM algorithms with recent applications [10]. These machine learning methods have several uses. This study discusses the novel applications created using them. An in-depth overview of algorithms and related topics is given, from their origins to inventive uses. This study explains when and how algorithms are utilized for real-time forecasting and other applications. This study discusses how to apply these algorithms and their results and performance in new and novel research. Their findings are described using quantitative and qualitative criteria. After much investigation and study, they reached some crucial conclusions about the LSTM network. The SVM algorithm provides one of the most satisfactory predictive analytical outcomes in real-time applications such as medical, bank fraud, facial recognition, student performance prediction, and energy consumption prediction. Deep learning with feedback is one of the greatest LSTM algorithms. It remembers key information, allowing for accurate predictions. The study's results illustrate how heavily machine learning and AI will be employed in the future. Machine learning and AI are projected to help people accomplish their jobs or replace them, unleashing a wave of automation.

They have created a diabetes prediction model using decision tree categorization. The data include diabetic and non-diabetic women [11]. Twenty or older, number of

pregnancies, blood glucose test rate, insulin test, BMI, and family history of diabetes were used to predict diabetes in this research. This research emphasizes the above-mentioned traits and excludes others. This complicates data mining. It simplifies and complicates model analysis without sacrificing accuracy. The authors chose 723 records for the prediction model. They normalized data using the Min-Max method. The Rapid Miner tool was used to build the decision tree model. Rapid mining software found that high blood sugar patients were more likely to acquire diabetes. Two hundred forty-eight diabetic patients and 475 nonpatients were surveyed. It projected 231 diabetics and 499 non-diabetics. Two hundred thirty-one patients were wrongly predicted. The confusion matrix has 88.50% accuracy, 79.83% sensitivity, and 93.15% specificity.

3. Methodology and Proposed Model

The proposed model uses a combination of data mining and machine learning algorithms. As shown in the flow diagram in Figure 1, the model involves feature extraction in the first step after data collection and exploration. In the second stage, data preprocessing tries to normalize the data by the z-score method. In the third stage, the KMeans clustering technique tries to cluster the dataset into an optimal number of clusters with the help of the elbow method. The elbow method shows the best number of clusters by finding the knee value. The next step is the Mahalanobis distance, which assigns a new sample or patient to a more relevant cluster. The MD was used to find the best or nearest cluster instance of the Euclidean distance, which is the default technique of the KMeans algorithm. Mahalanobis works better than Euclid because it also calculates the variance of the data, not just the distance from the data center. The relevant or chosen cluster is copied to a new data frame and fitted to classification algorithms. The final stage is classification techniques such as decision tree, random forest, SVM, KNN, logistic regression, and Naïve Bayes for forecasting by hard voting in the ensemble method.

3.1. Data Collection. Healthcare systems and laboratories have generated large amounts of data worldwide, and advanced applications rely on that data for better results [12]. The training data for our models were collected from public and private laboratories in the Iraqi Kurdistan Region. The dataset included 1012 instances and seven attributes.

3.2. Feature Extraction. Feature extraction comes in helpful when you need to reduce the number of features required for processing without losing essential or relevant information. Feature extraction can also reduce duplicate data in the dataset [8]. The weight and height dataset represents and makes body mass index (BMI) in our dataset. There was a power correlation between BMI and weight and a negative correlation between BMI and height.

3.3. Data Preprocessing. One of the data processing techniques is data normalization [13], which is used to transform incomprehensible data into comprehensible data collection. Normalize data are a scaling or mapping technique for converting unnormal data to standard data [14–16]. We applied the z-score technique in our model and its normalized values in the dataset, which is a numerical data type. The formula for z-score normalization is

$$z = \frac{(x_i - \mu)}{\sigma}. \tag{1}$$

3.4. Clustering Algorithm. Clustering is a machine learning method that divides a group or collection of data points into numerous groups. Data points in the same group are more similar than data points in other groups. We put it in another way. The idea is to classify groups with similar features into clusters [17]. In data science, clustering is a helpful technique. It is a method for detecting cluster structure in a dataset based on the most remarkable, significant dissimilarity between clusters and the highest similarity within each cluster. [18]. Our research used the KMeans technique for data reduction, which is the most well-known and often used clustering algorithm. Various KMeans extensions have been suggested in the literature. Initializations always impact the KMeans approach and its expansions with a required number of clusters a priori. However, it is unsupervised learning to cluster in pattern recognition and machine learning [19]. The ideal number of clusters into which the data may be grouped is crucial in an unsupervised technique. To define the optimum k of KMeans clusters, we employed the elbow method, one of the most prominent approaches for determining this ideal value of k. In the elbow method, we are counting the number of clusters represented by (K) and computing the WCSS (within-cluster sum of square) points for each value of K, as shown in Figure 2.

3.5. Mahalanobis Distance. The Mahalanobis distance is a popular chemometrics, or multivariate statistics, measure. This work uses this feature to identify whether a sample is an outlier, whether a process is under control, or if a sample belongs to a group [3]. The formula is shown in Equation (2). We used the Mahalanobis distance to recognize the new sample or patient more relevant to which cluster is classified by the KMeans algorithm:

$$\mathrm{MD} = \sqrt{\left(\frac{\mu - \overline{\mu}}{\sigma_1}\right)^2 + \left[\left\{\left(\frac{\upsilon - \overline{\upsilon}}{\sigma_2}\right) - \rho_{12}\left(\frac{\mu - \overline{\mu}}{\sigma_1}\right)\right\}\frac{1}{\sqrt{1 - \rho_{12}^2}}\right]^2}. \tag{2}$$

3.6. Classification Algorithms (Ensemble Learning). The results of the clustering stage, in which data are reduced and grouped into a particular group, act as an input for the classifier techniques, as illustrated in Figure 3, where a sample of the data from the previous stage is allocated

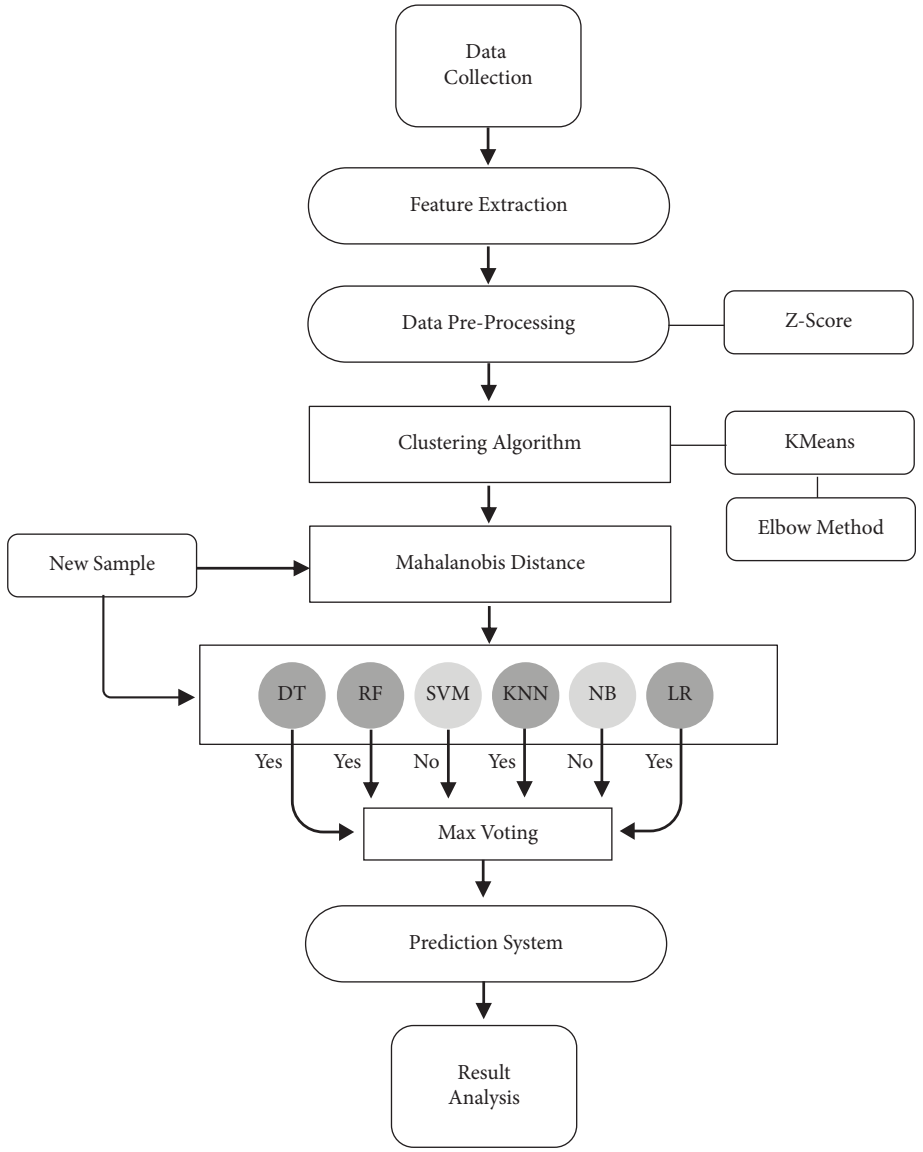

Figure 1: The flowchart of the proposed model.

to a similar cluster of new input data. A proposed model uses the maximum voting technique of the ensemble method for decision-making. The result of each method has been compared and analyzed, called hard voting. The classification algorithms include decision tree, random forest, SVM, KNN, logistic regression, and Naive Bayes.

3.6.1. Decision Tree. A decision tree is an algorithmic strategy for partitioning data on specific parameters, one of the algorithms for supervised learning. The purpose is to learn basic decision tree instructions to develop a concept that predicts the value of a target variable. It is ideal for continuous learning. The decision tree often follows the rules in the form of if-then-else expressions [8]. Classification is performed using decision trees, which do not require a lot of computing. Continuous data may be handled using decision trees [20].

3.6.2. Random Forest. A random forest is a dimensionality reduction technique that uses many decision trees to create a categorization. It is an example of an ensemble technique for classification and other tasks. [21]. It may be used to rank variables in their significance [22].

3.6.3. SVM (Support Vector Machine). The action of support vector machines is to find a hyperplane line that separates the negative and positive samples with the most significant margin [23]. The support vector machine objective is to find a hyperplane in multidimensional features that classifies data points. An SVM model in which the instances are represented as points in space mapped such that the examples of the various categories are separated by as much space as feasible. New instances are then mapped into the same space and given a category based on their position in the gap. Some methods for doing that use the fuzzy clustering method [15, 24, 25].

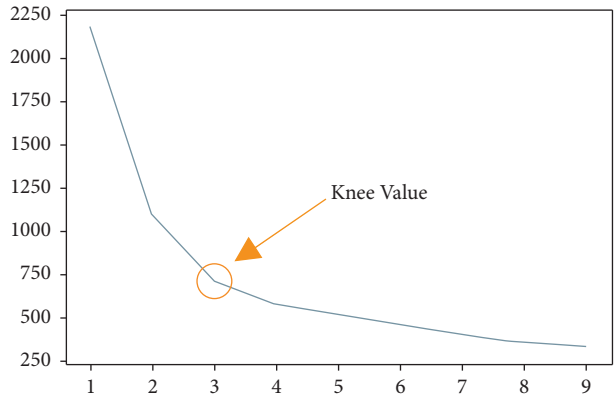

Figure 2: Determine the knee value of the elbow method.

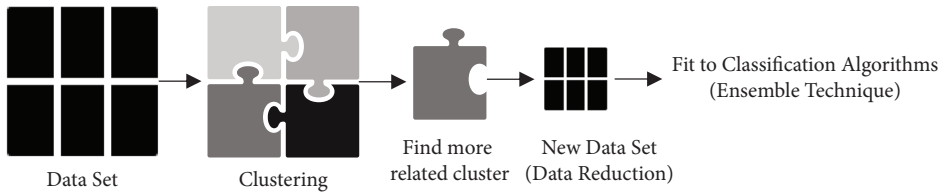

Figure 3: Data reduction from clustering stage to classification stage.

3.6.4. KNN (K-Nearest Neighbors). KNN is only simulated locally, and calculations are postponed until classification is complete. The KNN algorithm is one of the most basic machine learning algorithms available [26].

3.6.5. Logistic Regression. Logistic regression is another name for the generalized linear model. Nonlinear functions are divided into the linear component and the link function. The linear component of the classification model's output is delivered via the link function. A logistic function is used to handle the linear output in a logistic regression. The logistic function only returns values between 0 and 1 [4].

3.6.6. Naïve Bayes. The Naive Bayes method is a classification technique that can manage missing values during classification. A supervised learning method was used to classify them. The fundamental idea of Naive Bayes is that it operates based on conditional probability. For connected qualities, conditional independence fails, and Naive Bayes performance suffers as a result [5]. Naive Bayes is robust against noise points when taking conditional probabilities.

3.7. Prediction System. The proposed model code was written in *Python*, and a confusion matrix was used to illustrate the results for measuring classification methods and the ensemble technique. Some of the comparing criteria employed in our study were accuracy, precision, and recall percentage. The quantity of correctly recognized predictions is determined by accuracy. The formula is shown in equation (3). Precision is the ratio of properly recognized true positives to total positive samples. The total number of positive samples equals the sum of correctly and incorrectly classified samples, as shown in equation (4). The recall percentage of correctly classified positive samples, total positive samples, and total false-negative samples is shown in equation (5):

$$\text{Accuracy} = \frac{\text{total currect predictions}}{\text{total predictions}}, \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

3.8. Result Analysis. This study proposes an accurate model prediction approach for detecting gestational diabetes using some techniques of data mining and machine learning algorithms. We attempt to use a mix of clustering and classification algorithms (ensemble approach). The input variables or attributes that were used in the proposed model came from Iraqi Kurdistan Region laboratories. The dataset included 1012 instances and 7 attributes. Table 2 illustrates the attribute description.

To illustrate and discuss the power of the proposed model, we first employed only classification methods on the collected data. The results are shown below in Table 3. However, it was not a pleasant accuracy, so we attempted to find a better one by combining clustering and classification methods. KMeans clustering was used for data reduction and cleaning. When a cluster was predicted as the more related cluster to the new sample, which was reduced and cleaned, it was fitted to the classification algorithms as a new

Table 2: The attributes' description.

	Attributes' name	Attributes' detail	Data type	Mean	Standard deviation
1	Age	Age (years) of pregnant women	Numeric	30.36	7.02
2	Weight	Weight (kg)	Numeric	72.99	12.55
3	Height	Height (cm)	Numeric	158.15	7.36
4	BMI	Body mass index (weight/(height)^2)	Numeric	29.26	5.09
5	Pregnancy number	Number of times pregnant	Numeric	2.52	1.53
6	Heredity	Family history of diabetes	Boolean	783 samples (0), 229 samples (1)	
7	Blood sugar test	Plasma glucose concentration	Boolean	795 samples (0), 217 samples (1)	

Table 3: Comparison of each algorithm and the ensemble technique with different values of k in the KMeans algorithm.

No. of K	DT (%)	RF (%)	SVM (%)	KNN (%)	LR (%)	NB (%)	Ensemble (%)
1	81	83	86	89	85	85	87
2	87	88	88	86	89	87	88
3	92	92	90	92	90	92	92
4	89	92	92	91	92	90	92
5	87	89	89	87	89	87	89
6	90	92	92	92	92	93	92
7	87	89	92	87	91	89	89
8	87	89	89	87	90	89	89
9	90	92	92	90	92	90	92
10	94	96	96	96	96	94	96

data frame. The outcomes and results are illustrated in Table 3. As you can see, most algorithms did not produce better accuracy or results.

The proposed model uses a clustering approach to partition the data without missing any. KMeans methods are used instead of other clustering techniques because the DBSCAN algorithm may identify samples as noise and discard these objects, while KMeans generally cluster all the objects. It is worth mentioning that the numeric dataset prevents us from using K-modes clustering. To improve the KMeans clustering algorithm, the proposed model uses the elbow technique to find the optimal k for clustering the data, which is presented in Table 3. As illustrated, each algorithm has a different accuracy from the different numbers of k. The elbow method was proposed as number 3 for categorizing the data into 3 clusters. In another way, it uses the Mahalanobis technique to find the best or more related cluster to the new sample. The default method for the KMeans algorithm is the Euclidean distance to find the distance of a new sample to a cluster and predict it. However, because of the similarity between clusters and the variance of data, the proposed model tries to use the Mahalanobis distance, which has better performance than the Euclidean distance, as presented in Figure 4. The MD calculates the distance by the effect of data variance. Still, the Euclidean distance calculates the center of the cluster and does not care about data variance. Still, the Mahalanobis works differently and calculates the effect of data variance in each cluster. Table 4 illustrates the results of each distance method's performance. The MD findings are fewer and more accurate than those of the ED.

Finally, after using the mentioned techniques in the clustering part of the proposed model and finding the best and more related cluster by Mahalanobis, we compare the classification algorithms and use the ensemble max voting technique to get the best result. Most of the time, the maximum voting method is used for classification problems to sort things into groups. In this technique, predictions are made for each data point by using more than one model.

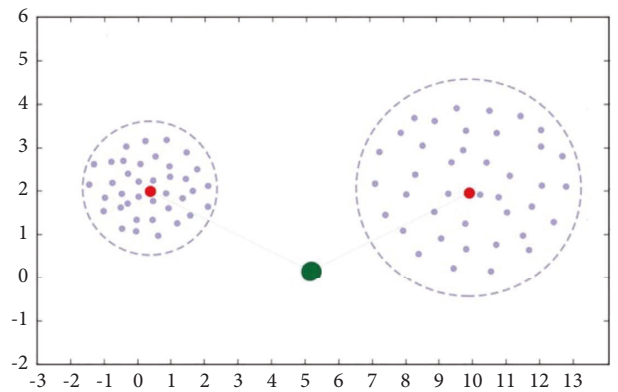

Figure 4: Calculate distance of new sample to clusters by MD.

Table 4: The comparison of the calculated distance between the new sample and each cluster by Mahalanobis and the Euclidean distance.

Features	Cluster 1		Cluster 2		Cluster 3	
	MD	ED	MD	ED	MD	ED
Age	1.32	13.62	10.76	33.16	40.95	27.83
BMI	1.40	15.31	0.29	12.25	2.56	16.13
Pregnancy no.	0.43	7.91	3.21	17.68	7.96	28.02
Heredity	0.05	3.87	0.24	7.34	1.40	7.21
Distance	1.79	22.31	3.81	40.66	7.27	42.85

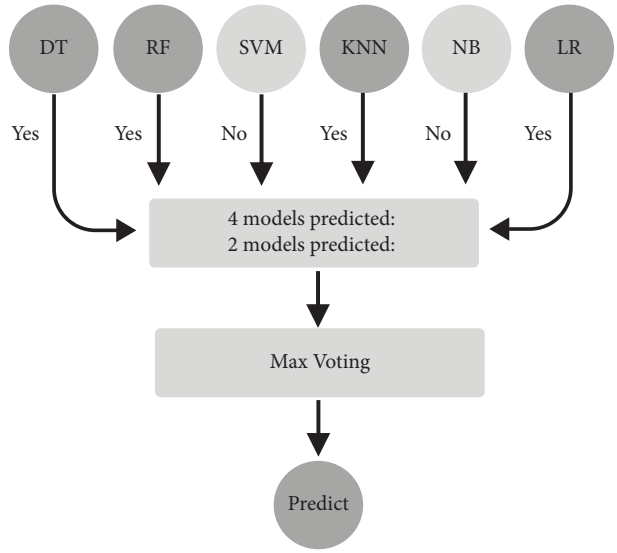

Figure 5: Decision-making by ensemble max voting technique.

Table 5: Classification accuracy and each technique of the proposed model comparison.

	Algorithms	Accuracy of each classification algorithm (fitting the origin data) (%)	Accuracy of each classification algorithm (proposed model) (%)	Ensemble technique (proposed model) (%)
1	Decision tree	81	92	
2	Random forest	83	92	
3	SVM	86	90	
4	KNN	89	92	92
5	Logistic regression	85	90	
6	Naïve Bayes	85	92	

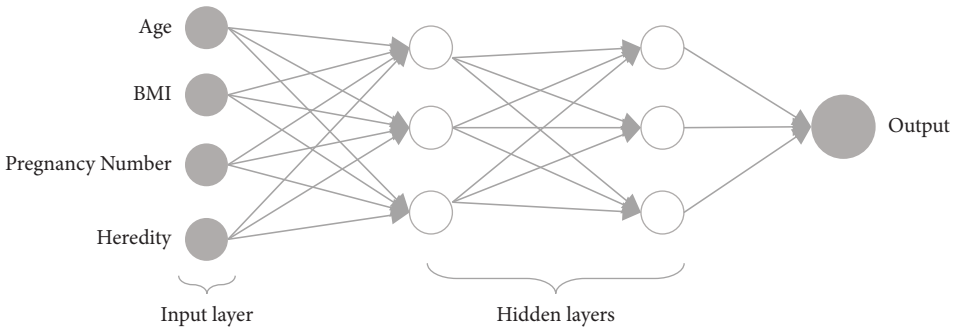

Figure 6: Artificial neural network binary classification.

Figure 5 is a flowchart illustrating how the ensemble method makes the decision. Each test data goes through the ensemble operation shown in Figure 5, and a confusion matrix is made from the obtained results. The final result is obtained from the confusion matrix and calculates precision, accuracy, recall, and f1 score.

Each model's predictions are treated as a "vote." The final prediction is based on what most models will do, and the majority of results will make the decision. [27]. Table 5 shows that the suggested model is more accurate than the method discussed in this study.

The results obtained from the proposed model were compared with the artificial neural network model by applying the same data. Because our application is a binary classification, we will have only two output classes (1 and 0) and employ only one neuron with a sigmoid activation function, and the optimal epochs (how many times neural networks will be trained [28]) are assigned

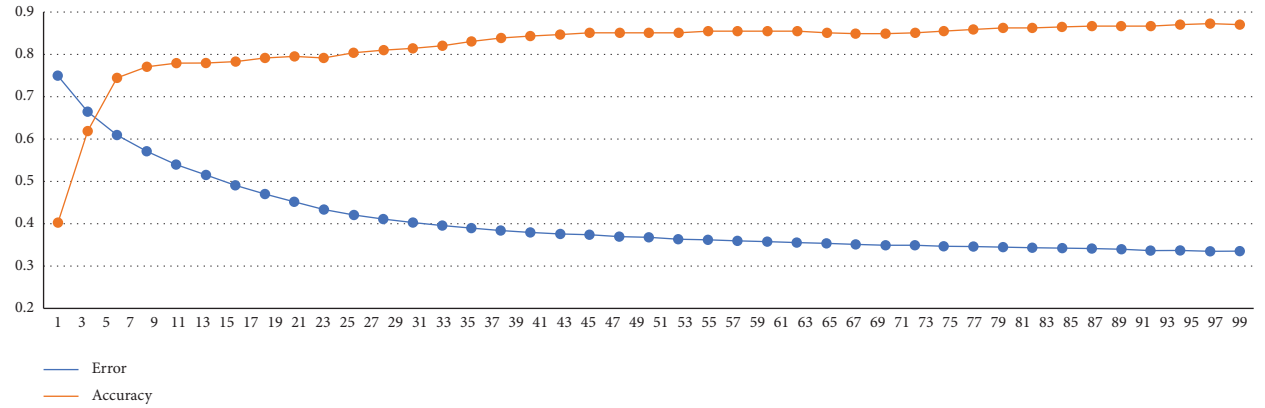

Figure 7: The flowchart of error rates and accuracy of 100 epochs of the ANN model.

Table 6: The machine learning accuracy of existing works.

	Algorithms	Accuracy (%)	References
1	Decision tree	78.1768	[20]
2	Random forest	91	[29]
3	SVM	82	[8]
4	KNN	77	[30]
5	Logistic regression	77.9	[4]
6	Naïve Bayes	79.84	[26]

Table 7: Hyperparameters of classification algorithms.

	Algorithms	Hyperparameters	Description	Value
1	Decision tree	max_depth	Maximum depth of the tree	"none"
		min_samples_split	Minimum number of samples required to split an internal node	2
		min_samples_leaf	Minimum number of samples required to be at a leaf node	1
		criterion	Function to measure the quality of a split	"gini"
		max_features	Number of features to consider when looking for the best split	"sqrt '
2	Random forest	n_estimators	Number of trees in the forest.	100
		max_depth	Maximum depth of the tree	"none"
		min_samples_split	Minimum number of samples required to split an internal node	2
		min_samples_leaf	Minimum number of samples required to be at a leaf node	1
		criterion	Function to measure the quality of a split	"gini"
		max_features	Number of features to consider when looking for the best split	"sqrt"
3	SVM	C	Strength of the regularization is inversely proportional to C	1.0
		Kernel	Specifies the kernel type to be used in the algorithm	"rbf"
4	KNN	n_neighbors	Number of neighbors for decision making to classification	5
5	Logistic regression	Solver	Algorithm to use in the optimization problem	"lbfgs"
		penalty	Specify the norm of the penalty	"elasticnet"
6	Naïve Bayes	priors	Prior probabilities of the classes (depend of data)	"n_classes"
		var_smoothing	Portion of the largest variance of all features	1e-9

to 100. In each epoch, the loss decreases, and the accuracy increases. At the first epoch, the accuracy was 0.5791, and at the last epoch, it was 0.8674, which is pretty remarkable for a neural network. And another important hyperparameter of the ANN is the number of hidden layers. The assigned hidden number of our model is 2 layers. Figures 6 and 7 illustrate the effect of both (epochs and layers) hyperparameters on neural network errors and accuracy.

We might compare the suggested model and our research to some obtained accuracy from previous intelligent models that were created by existing research. Table 6 shows that, in most cases, the proposed model is more accurate than previous work.

In the last part of the result analysis, we want to discuss the importance of hyperparameters to improve the performance of machine learning methods and to continue improving the ensemble result. The better performance of

KMeans helps to choose the best cluster. As shown in Table 3, the number K is one of the essential hyperparameters of KMeans and how it affects the accuracy of the model. We presented the accuracy of the model for each number of k. On the contrary, the effect of hyperparameters tunning was discussed in ANN modelling regarding the number of hidden layers and epochs. To improve the results of the classification models which is used in this research, we have shown some of the hyperparameters in Table 7.

4. Conclusion

The data for the proposed model were collected from laboratories in the Iraqi Kurdistan Region. The dataset includes 1012 instances and 7 attributes: age, pregnancy number, weight, height, BMI, heredity, and blood sugar test. A mixed prediction model in the proposed model has been developed to identify gestational diabetes. With the help of the elbow technique, the KMeans algorithm was used to cluster the data into an optimal number of clusters. The Mahalanobis distance method is used to select the most related cluster which is most closely connected to the new samples. In the prediction section, classification techniques such as DT, RF, NB, and KNN with 92% accuracy and SVM and LR with 90% accuracy were used for ensemble techniques. The obtained accuracy for the ensemble max voting method is 92%. Finally, the findings show that using a mix of KMeans clustering, elbow method, Mahalanobis distance, and ensemble learning significantly improves prediction accuracy. In future work, we will try developing the proposed model for an adaptive healthcare application to predict diabetes instances, especially in the mentioned geographic area for which the necessary attention has not been paid to this issue. The application will be designed to get a new sample and add it to the dataset. This method updates the database daily, so each time the model is trained, it will have more data to work with.

Data Availability

All codes and data used can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] N. Ali, A. S. Aldhaheri, H. H. Alneyadi et al., "Effect of gestational diabetes mellitus history on future pregnancy behaviors: the Mutaba'ah study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 1, pp. 1–12, 2021.

[2] Z. Zhang, L. Yang, W. Han et al., "Machine learning prediction models for gestational diabetes mellitus: meta-analysis," *Journal of Medical Internet Research*, vol. 24, no. 3, Article ID e26634, 2022.

[3] R. F. Jader, S. Aminifar, and M. H. M. Abd, "Diabetes detection system by mixing supervised and unsupervised algorithms," *Journal of Studies in Science and Engineering*, vol. 2, no. 3, pp. 52–65, 2022.

[4] A. Al-Zebari and A. Sengur, "Performance comparison of machine learning techniques on diabetes disease detection," in *1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation*, pp. 2–5, Ankara, Turkey, 2019.

[5] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus—a machine learning approach," in *2015 IEEE Recent Advances in Intelligent Computational Systems*, pp. 122–127, Trivandrum, India, 2016.

[6] I. Gnanadass, "Prediction of gestational diabetes by machine learning algorithms," *IEEE Potentials*, vol. 39, no. 6, pp. 32–37, 2020.

[7] M. Alehegn, R. Joshi, and M. Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: ensemble approach," *International Research Journal of Engineering and Technology*, vol. 4, no. 10, pp. 426–436, 2017.

[8] P. Sonar and K. Jaya Malini, "Diabetes prediction using different machine learning approaches," in *Proceedings of the 3rd International Conference on Computing Methodologies and Communication*, pp. 367–371, Erode, India, 2019.

[9] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "Related papers the Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, 2000.

[10] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, Article ID 100071, 2022.

[11] J. Han, J. C. Rodriguez, and M. Beheshti, "Discovering decision tree based diabetes prediction model," *Communications in Computer and Information Science*, vol. 30, pp. 99–109, 2009.

[12] F. Jiang, Y. Jiang, H. Zhi et al., "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[13] S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," pp. 20–22, 2015, https://arxiv.org/abs/1503.06462.

[14] S. Aminifar, "Uncertainty avoider interval type II defuzzification method," *Mathematical Problems in Engineering*, vol. 2020, Article ID 5812163, 16 pages, 2020.

[15] S. Aminifar and A. Marzuki, "Uncertainty in interval type-2 fuzzy systems," *Mathematical Problems in Engineering*, vol. 2013, Article ID 452780, 16 pages, 2013.

[16] N. M. Saravana Kumar, T. Eswari, and P. Sampath, "Predictive methodology for diabetic data analysis in big data," *Procedia Computer Science*, vol. 50, pp. 203–208, 2015.

[17] Y. Alapati and K. Sindhu, "Combining clustering with classification: a technique to improve classification accuracy," *International Journal of Computer Science and Engineering*, vol. 5, no. 06, pp. 336–338, 2016.

[18] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.

[19] A. Likas, N. Vlassis, and J. Verbeek, "The global k-means clustering algorithm Intelligent Autonomous Systems," *ISA technical report series*, vol. 36, 2011.

[20] A. A. AlJarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *2011 International Conference on Innovations in Information Technology*, pp. 303–307, Abu Dhabi, United Arab Emirates, 2011.

[21] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," *Advances in*

Intelligent Systems and Computing, Springer, Berlin, Germany, 2019.

[22] S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," in *2019 International Conference on Computer and Information Sciences*, pp. 1–4, Sakaka, Saudi Arabia, 2019.

[23] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114–1120, 2010.

[24] A. Hamad, S. Aminifar, and M. Daneshwar, "An interval type-2 FCM for color image segmentation," *International Journal of Advanced Computer Research*, vol. 10, no. 46, pp. 12–17, 2020.

[25] A. Marzuki, S. Y. Tee, and S. Aminifar, "Study of fuzzy systems with Sugeno and Mamdanitype fuzzy inference systems for determination of heartbeat cases on Electrocardiogram (ECG) signals," *International Journal of Biomedical Engineering and Technology*, vol. 14, no. 3, pp. 243–276, 2014.

[26] Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P. K. Abhilash, K. Sandeep, and D. Indira, "Prediction of diabetes using machine learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 7, pp. 2547–2551, 2019.

[27] G. Tuysuzoglu, D. Birant, and A. Pala, "Majority voting based multi-task clustering of air quality monitoring network in Turkey," *Applied Sciences*, vol. 9, no. 8, pp. 1–21, 2019.

[28] R. Jader and S. Aminifar, "Fast and accurate artificial neural network model for diabetes recognition," *NeuroQuantology*, vol. 20, no. 10, pp. 2187–2195, 2022.

[29] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[30] M. A. Sarwar, N. Kamal, W. Hamid, and M. Ali Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in *ICAC 2018 - 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing*, pp. 1–6, Newcastle Upon Tyne, UK, 2018.