Informatics in Medicine Unlocked 17 (2019) 100179

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques

Changsheng Zhu[a,*], Christian Uwa Idemudia[a], Wenfang Feng[b]

[a] School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China
[b] School of Economics and Management, Lanzhou University of Technology, Lanzhou, 730050, China

ARTICLE INFO

Keywords:
PCA
K-means
Diabetes
Data mining
Logistic regression

ABSTRACT

Diabetes causes a large number of deaths each year and a large number of people living with the disease do not realize their health condition early enough. In this study, we propose a data mining based model for early diagnosis and prediction of diabetes using the Pima Indians Diabetes dataset. Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers which determine the final cluster result, which either provides a sufficient and efficiently clustered dataset for the logistic regression model, or gives a lesser amount of data as a result of incorrect clustering of the original dataset, thereby limiting the performance of the logistic regression model. Our main goal was to determine ways of improving the k-means clustering and logistic regression accuracy result. Our model comprises of PCA (principal component analysis), k-means and logistic regression algorithm. Experimental results show that PCA enhanced the k-means clustering algorithm and logistic regression classifier accuracy versus the result of other published studies, with a k-means output of 25 more correctly classified data, and a logistic regression accuracy of 1.98% higher. As such, the model is shown to be useful for automatically predicting diabetes using patient electronic health records data. A further experiment with a new dataset showed the applicability of our model for the predication of diabetes.

1. Introduction

Diabetes stands among the top 10 causes of death for 2016. Diabetes killed 1.6 million people in 2016, up from less than 1 million in 2000. With this figure diabetes replaced HIV/AIDS as the seventh top cause of death [1]. The number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014, with the global prevalence of diabetes among adults over 18 years of age rising from 4.7% in 1980 to 8.5% in 2014 [2].

By 2040, 642 million adults (1 in 10 adults) are expected to have diabetes. Also, 46.5% of those with diabetes have not been diagnosed [3]. In order to reduce the number of deaths attributable to diabetes, it is essential that methods and techniques that will aid in early diagnosis of diabetes be devised, because a large number of deaths in diabetic patients are due to late diagnosis.

In order to achieve cutting-edge techniques for the early diagnosis of diabetes, we need to utilize advanced information technology, and data mining is a suitable field for this. Data mining offers the ability to extract and discover previously unknown, hidden, but interesting patterns from a large database repository. These patterns can aid medical diagnosis and decision-making.

Various techniques and algorithms have been designed for application in extracting knowledge and information in the diagnosis and treatment of disease from medical databases. PCA is a simple, non-parametric method for extracting relevant information from confusing data sets [4]. When a large dataset is to be clustered into a user specified number of clusters (k), which are represented by their centroids, k-means will cluster the data by minimizing the squared error function [5], and often misclassifies some data due to outliers; also the time complexity will be greater. To overcome these problems, principal components analysis (PCA) can be used to reduce the dataset to a lower dimension, while ensuring that the least information is lost, and providing a better centroid point for clustering. K-means clustering partitions a dataset into different groups of similar objects. Clusters that are highly dissimilar from the others are regarded as outliers and discarded. Logistic regression is an efficient regression predictive analysis algorithm. Its application is efficient when the dependent variable of a dataset is dichotomous (binary). Logistic regression is used in the description and analysis of data in order to explain the relationship between one dependent binary variable and one or more independent

* Corresponding author.
E-mail addresses: Zhucs_2008@163.com (C. Zhu), xtianidemudia@yahoo.co.uk (C.U. Idemudia), 1036784024@qq.com (W. Feng).

https://doi.org/10.1016/j.imu.2019.100179
Received 20 January 2019; Received in revised form 27 March 2019; Accepted 4 April 2019
Available online 05 April 2019

variables.

This research work proposes PCA for dimensionality reduction, which helps to define suitable initial centroids for our dataset when the k-means algorithm is applied. K-means is then used to find outliers and to cluster the data into similar groups, with logistic regression as a classifier for the dataset. In this paper, section 2 provides a review of related work done by other researchers in the area of diabetes prediction and diagnosis. Section 3 shows details of the experimental procedures. Section 4 describes the experimental result, while section 5 concludes the work while suggesting possible direction for future work.

2. Related study

Diabetes is a standout amongst the most well-known non-transmittable diseases in the world. It is assessed to be the seventh leading cause for death [6]. It is predicted that the diabetes rate in adults worldwide will become 642 million in 2040 [3]. The early diagnosis of diabetes in patients has been a major goal for medical researchers and professionals. With the availability of vast technological innovation in computer science, collaborative studies have shown that by applying computer skills and algorithms (such as data mining), efficient, cost effective and rapid techniques can be derived for the diagnosis of diabetes.

Many researchers have developed various prediction models using data mining to predict and diagnose diabetes. Iyer [15] in their study proposed the use of the Naïve Bayes algorithm to predict the onset of diabetes. The study gave an accuracy result of 79.56%. Tarun [13] used PCA and a support vector machine for the classification of diabetic patients. Experimental result from the study showed that the previous level can be improved upon as they had a classification accuracy of 93.66%. Mustafa S. Kadhm [18] proposed the use of a Decision Tree (DT) to assign each data sample to its appropriate class after applying the K-nearest neighbor algorithm for eliminating undesired data. Han et al. [3] designed a model that uses the k-means algorithm and the logistic regression algorithm for predicting diabetes. The model attained a 95.42% accuracy.

In Ref. [14], the authors used k-means clustering in identifying and eliminating outliers, a genetic algorithm and correlation based feature selection (CFS) for relevant feature extraction, and finally used k-nearest neighbor(KNN) for classification of diabetic patients. Patil [16] proposed a hybrid prediction model that applied k-means clustering to the original dataset and then used C4.5 algorithms in building the classifier model. The classification accuracy result was 92.38%. Anjali [7] proposed a methodology based on Principal Component Analysis (PCA) to reduce the dimension of extracted features with Neural Network (NN) as the classifier. The accuracy result was 92.2%.

The studies all used a common dataset (the Pima Indian Diabetes Dataset) from the University of California, Irvine (UCI) machine learning database. Considering the need for an effective prediction algorithm, improving the already existing prediction algorithm will be a major task of our research whilst using the same dataset as other researchers. While great result has been achieved by various researchers, their data preprocessing step limited the amount of data available for their final prediction and classification. Therefore, we need to propose a model for enhanced data preprocessing that will produce a large amount of useable data and also enhance the classification algorithm.

3. Methodology

This section is comprised of the following steps: the data description, preprocessing technique and the classification algorithm. The proposed model is designed and implemented by combing the benefit of applying PCA, K-means and Logistic regression. A new methodology is then proposed by using PCA to transform the initial set of features, thereby solving the problem of correlation, which makes it difficult for the classification algorithm to find relationships among the data. The

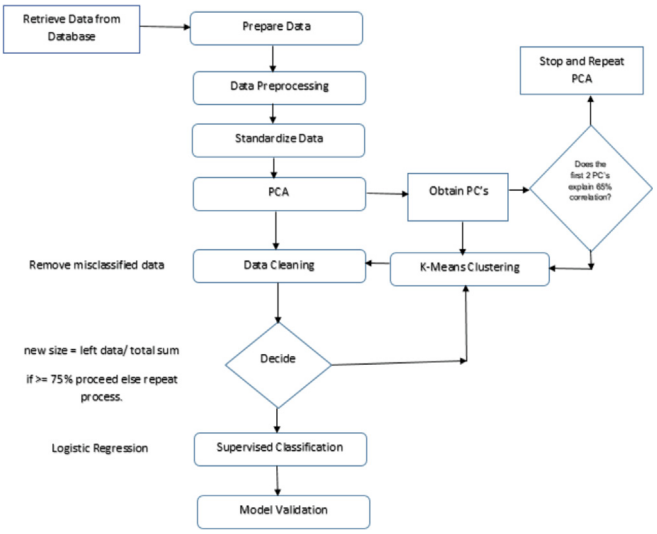

Fig. 1. Proposed algorithm model.

PCA application helps to filter out irrelevant features, thereby lowering the training time, cost, and also increases model performance [10]. After performing PCA analysis, the result is then passed for unsupervised clustering using K-means because of the ability of k-means to address outliers [11]. The K-means cluster result is cleaned and Logistic Regression is applied to build our supervised classification for the dataset. The proposed model flowchart is shown in Fig. 1.

3.1. Data mining toolkit

Anaconda is a free and open Python programming language toolkit. It consists of over 250 popular packages for data science and machine learning related application. Applying this package, we are able to perform related data mining tasks on our dataset and implement (design) our proposed model.

By efficiently preprocessing the original dataset, performing PCA, and simulating the same experiment as other researchers, we show that improvement to the accuracy of diabetes diagnosis using data mining techniques can be done.

3.2. Dataset description

The Pima Indian Diabetes dataset obtained from UCI repository of machine learning was utilized for this study. The dataset is comprised of 768 sample female patients from the Arizona, USA population who were examined for diabetes. The dataset has a total of 8 attributes (representing medical diagnosis criteria) with one target class (which represents the status of each tested individual). In the dataset there is a total of 268 tested positive instances and 500 tested negative instances. The attributes in the dataset include the following:

- Number of times pregnant (Preg)
- Plasma glucose concentration at 2hr in an oral glucose tolerance test (Plas)
- Diastolic Blood pressure (Pres)
- Triceps skin fold thickness (Skin)
- 2-hr serum insulin (Insu)
- Body mass index (BMI)
- Diabetes pedigree function (Pedi)
- Age (Age)
- Target Variable (Diag)

Table 1
Original and preprocessed dataset statistics.

Statistics	Dataset	Preg	Plas	Pres	Skin	Insu	BMI	Pedi	Age
COUNT	*Original*	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000
	Preprocess	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000
MEAN	*Original*	0.8554	120.8945	69.1054	20.5364	79.7994	31.9925	0.4718	33.2408
	Preprocess	0.8554	121.6867	72.4051	29.1534	155.5482	32.4574	0.6718	33.2408
STD	*Original*	0.3518	31.9726	19.9522	15.9522	115.2440	7.8841	0.3313	11.7602
	Preprocess	0.3518	30.4359	12.0963	8.7909	85.0211	6.8751	0.3313	11.7602
MIN	*Original*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0780	21.0000
	Preprocess	0.0780	44.0000	24.0000	7.0000	14.0000	18.2000	0.0780	21.0000
25%	*Original*	1.0000	99.0000	62.0000	0.0000	0.0000	27.3000	0.2437	24.0000
	Preprocess	1.0000	99.7500	64.0000	25.0000	121.5000	27.5000	0.2437	24.0000
50%	*Original*	1.0000	117.0000	72.0000	23.0000	30.5000	32.0000	0.3725	29.0000
	Preprocess	1.0000	117.0000	72.2025	29.2534	155.5482	32.4000	0.3725	29.0000
75%	*Original*	1.0000	140.2500	80.0000	32.0000	127.2500	36.6000	0.6262	41.0000
	Preprocess	1.0000	140.2500	80.0000	32.0000	155.5482	36.6000	0.6262	41.0000
MAX	*Original*	1.0000	199.0000	122.0000	99.0000	846.0000	67.1000	2.4200	81.0000
	Preprocess	1.0000	199.0000	122.0000	99.0000	864.0000	67.1000	2.4200	81.0000

3.3. Data preprocessing

Today's real world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge sizes and their likely origin from multiple, heterogeneous sources [13]. Data quality is an important factor in the data mining process for disease prediction and diagnosis, because low quality data may lead to inaccurate or low prediction result. In order to make our original dataset more productive and applicable for predicting diabetes, we applied several preprocessing techniques using various packages offered within the Anaconda integrated development environment.

First, we took a closer look at the various attributes, and discussing with a professional dietician, analyzed the medical relevance of each attribute to diabetes prediction and diagnosis. It was discovered that "number of times pregnant" has less significance to the current research direction. We decided to apply the same technique used by Han Wu [12] by transforming this numeric attribute into a nominal attribute of value 0 and 1, with 1 indicating a patient previously pregnant and 0 indicating a patient was never pregnant. This helps to reduce the complexity of analyzing the dataset [12] (see Table 1).

Secondly, statistical analysis of our dataset suggested the presence of missing values. Table 2 below shows the statistical result for our dataset. From the statistical result, it is observed that Plasma glucose concentration, Diastolic blood pressure, Skin fold thickness, 2hr serum insulin and Body mass index have a min value of 0. Medical knowledge explains that such attributes (medical result) cannot be 0; therefore it suggests that the dataset contains a missing value that if not handled can impair the quality of our model result and accuracy. Various methods have been suggested for handling missing values in datasets. In our case we replaced missing values with the mean such attribute.

As part of our data preprocessing, the original data values are scaled so as to fall within a small specified range of [0, 1] values by performing normalization of the dataset. This will improve speed and reduce runtime complexity. Using the Z-Score we normalize our value set V to obtain a new set of normalized values V' with the equation below:

$$V' = \frac{V - Y}{Z} \tag{1}$$

Where V' = New normalized value, V = previous value, Y = mean, Z = standard deviation.

Table 2
Confusion matrix.

0 (Negative)	1 (Positive)	Class
391	3	Predicted Negative
13	207	Predicted Positive

3.4. Model algorithm design

Our model algorithm will be made up of 3 sub stages. In the first stage of the design we will perform dimensionality reduction on the already processed dataset (using PCA). Then we will cluster the selected principal component using K-means to address outliers and remove any incorrectly classified data. Finally, the correctly clustered and classified data will be used as input for our supervised classification using logistic regression.

3.5. Principal component anaylysis

During data analysis it is often very difficult to find all the relationships among attributes. PCA allows a huge amount of information enclosed in initially correlated data to be transformed into a set of new orthogonal components, thereby making it possible to discover concealed relationships, enhance data visualization, detection of outliers, and classification within the newly defined dimensions [5]. The application of PCA on a dataset can be of great help when unsupervised learning is required to be performed on such a dataset, as it will aid in efficiently initializing centroids for clustering.

Because PCA yields a feature subspace that maximizes the variance along the axes, we first standardize the dataset onto a unit scale (mean = 0 and variance = 1) to improve the PCA result which is a requirement for the optimal performance of many machine learning algorithms.

Our objective here is to transform our dataset X of p dimension into a new sample set Y of smaller dimension L ($L < p$), where Y is the Principal component of X i.e.

$$Y = PC(X) \tag{2}$$

We proceed as follows:

(a) Organize our dataset:

With X having a set of n vectors $(x_1, x_2, ..., x_n)$ where each x_i element is an instance of our dataset.

(b) Find the mean using the equation:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3}$$

(c) Calculate Variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \tag{4}$$

(d) Calculate Covariance:

$$X^{n\times n} = (x_{i,j},\ x_{i,j} = cov(Dim_i, Dim_j)) \tag{5}$$

Where $X^{n\times n}$ is our data matrix with n rows and n columns and Dim_i is the *ith* dimension.

(e) Calculate Eigenvalues and Eigenvectors:

The core of a PCA is the eigenvector and eigenvalues of the covariance matrix. The eigenvectors will determine the directions of the new feature space while the eigenvalues determine the magnitude.

If A is an $n \times n$ matrix, then a nonzero vector x in $\mathbb{R}^n$ is called an eigenvector of A (or of the matrix operator T_A) if Ax is a scalar multiple of x; that is,

$$Ax = \lambda x \tag{6}$$

for some scalar λ. The scalar λ is called an eigenvalue of A and x is said to be an eigenvector corresponding to λ. Since the eigenvectors corresponding to an eigenvalue of a matrix A are the nonzero vectors that satisfy the equation

$$(\lambda \mathrm{I} - \mathrm{A})x = 0 \tag{7}$$

we define the set E to be all vectors x that satisfy equation (7) as our corresponding Eigen space.

$$E = \{x: (A - \lambda \mathrm{I})x\} = 0 \tag{8}$$

(f) Once the Eigen space is found from the covariance matrix, the next step is to order the eigenvectors by eigenvalue, highest to lowest. This eliminates less significant components and we are left with the principal components that provide a good approximation of the original data. Our new principal components will be used as input for our k-means clustering in the next stage of our algorithm design.

3.6. K-MEANS clustering

K-means is one of the simplest and efficient unsupervised classification algorithms. K-means is a well-known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids [5]. It is a typical distance-based clustering algorithm, in which the distance is used as a measure of similarity, i.e. the smaller distance between objects shows greater similarity [12]. Fig. 2 shows the graphical procedure for the k-means clustering by applying the following steps:

(a) Step (a) in Fig. 4 shows our entire dataset. Initialize k = 2 since the target variable contains two possible outcomes (positive and negative).
(b) Next is to determine for each input data the cluster center that it is nearest to by using equation (9) extracted from Ref. [9] (step b)

$$S_i^{(t)} = \{x_p: ||x_p - m_i^{(t)}||^2 \le ||x_p - m_i^{(t)}||^2\ \forall_j, 1 \le j \le k\} \tag{9}$$

(c) Applying equation (10) from Ref. [9], update the cluster centers by recalculating the mean of each input data assigned to the cluster. (step c)

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|}\sum_{x_j \in S_i^{(t)}} x_j \tag{10}$$

(d) To bring our k-means cluster to a stop, we loop through step (b) and (c) until there is a convergence in the mean value of the clusters. (step d)

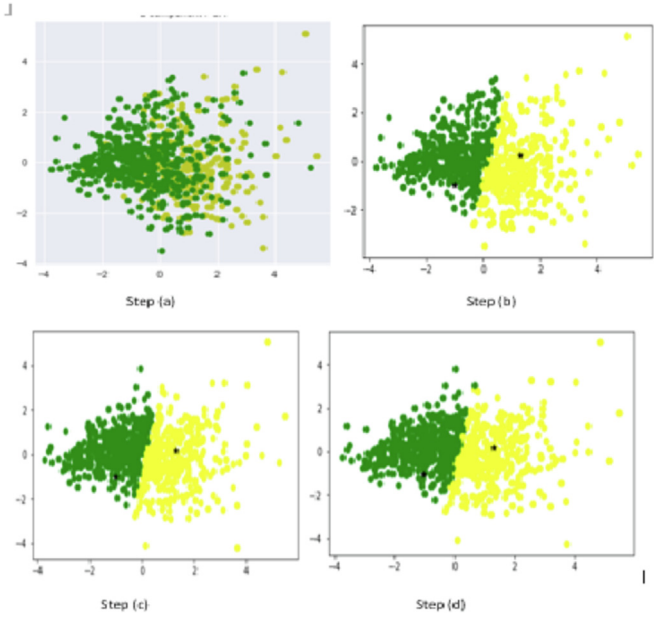

Fig. 2. K-means clustering procedure.

Thereafter, we cleaned our k-means cluster result by removing incorrectly clustered data and make a decision to find our new dataset for classification by using equation (11). If the new data size is above 75%, then we proceed with supervised classification, else we repeat the k-means step until a suitable size is determined.

$$new\ size = \frac{left\ data}{total\ sum} \tag{11}$$

After cleaning the clustered data, we obtained 614 correctly clustered patients, which is used as input to train the logistic regression algorithm.

3.7. Logistic regression algorithm

The application of the Logistic regression model has featured prominently in many domains such as the biological sciences. The Logistic regression algorithm is used when the objective is to classify data items into categories. Usually in logistic regression the target variable is binary, which means that it only contains data classified as 1 or 0, which in our case refers to a patient that is positive or negative for diabetes. The purpose of our logistic regression algorithm is to find the best fit that is diagnostically reasonable to describe the relationship between our target variable and the predictor variables.

The logistic regression algorithm is based on the linear regression model given in equation (12) below

$$y = h_\theta(x) = \theta^T x \tag{12}$$

Equation (12) will be highly inefficient to predict our binary values ($y^{(i)} \in \{0, 1\}$), therefore we introduce the function in equation (13) to predict the probability that a given patient (with given attributes) belongs to the "1" (positive) class versus the probability that it belongs to the "0" (negative) class.

$$P(y = 1|x) = h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)} \equiv \sigma(\theta^T x)$$

$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - h_\theta(x) \tag{13}$$

Applying equation (14), known as the sigmoid function, we are able to keep the value of $\theta^T x$ within the [0, 1] range. Then we search for a value of θ such that the probability $P(y = 1|x) = h_\theta(x)$ is large when x belongs to the "1" class and small when x belongs to the "0" class (i.e. $P(y = 0|x)$ is large.)

$$\sigma(t) = \frac{1}{(1 + e^{-t})} \tag{14}$$

Once our logistic regression algorithm has been successfully modelled and implemented; the output and result is discussed in the next section.

4. Experimental result

A major result discovered from the use of PCA is that the process helped in minimizing the drawback of having redundant features which are of no help for clustering. Since the reduction in the number of variables in the original data set assisted in handling noisy and outlier data, PCA therefore improved our k-means result. The main advantage of PCA is that once we have found these Principal Components from the data and we can compress the data i.e., by reducing the number of dimensions without much loss of information, it became an essential process in order to determine the number of clusters and provide a statistical framework to model the cluster structure [5].

The efficiency and accuracy of any predictive and diagnostic model is of paramount importance and should be ensured before such a model is deployed for implementation. We analyzed and evaluated our model output using different evaluation metrics, and the result is shown in Fig. 3.

First, to determine the performance of our model, we utilized the k-fold cross validation technique, which allows us to determine how well our model will perform when given new and previously unlearned data. Our choice of the 10-fold cross validation meant that our dataset was divided into 10 subsets. On each trial, one subset is used as the test set and the other nine subset formed the training set. Then, the average error across all 10 trials was computed to get the total performance of our model. This method helps solve two issues, first is that it reduces the problem of bias as almost all of the data is used for fitting, and secondly, the problem of variance is greatly reduced.

The confusion matrix is a popular way to provide a summarized representation of predictive findings. The confusion matrix gives the result of the following indices: true positive (TP), true negative (TN), false positive (FP), false negative (FN). Table 2 shows the confusion matrix for our model.

The performance metrics of our model are represented in Table 3 below:

```
=====Model Performance Summary=====
Correctly Claasified Instances===>       598     97.40%
Incorrectly Claasified Instances===>      16      2.60%
Kappa Statistic===>                        0.942
Mean squared error===>                     0.026
MCC===>                                    0.943
ROC Value===>                              0.967
Kappa Statistic===>                        0.942
Total Number of Instance===>             614
=====Detailed Accuracy By Class=====
                precision    recall  f1-score   support

Tested Negative       0.97      0.99      0.98       394
Tested Positive       0.99      0.94      0.96       220

    avg / total       0.97      0.97      0.97       614

=====Confusion Matrix=====

  |   0   1 |
--+---------+
0 |<391>  3 |
1 |  13<207>|
--+---------+
(row = reference; col = test)
```

Fig. 3. The result of the experiment.

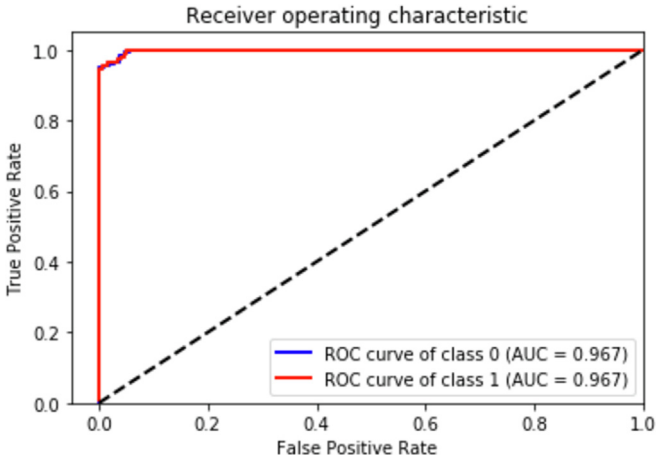

Fig. 4. The ROC curve.

Table 3
Performance metrics.

Performance Measure	Score
Recall	0.97
Precision	0.97
Accuracy	0.9739
MCC	0.94

The ROC Curve is a graphical plot that represent the performance of a classifier. The ROC allows us to look at the performance of our model across all possible thresholds. In our experiment, the ROC value was 0.967 and the ROC curve is shown in Fig. 4. The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The Kappa statistic normally holds a value between 0 and 1. Our experiment had a kappa statistic value of 0.942.

4.1. Comparison using other algorithms

To further evaluate how our model performs, we modelled our dataset with four different algorithms using the following variations: original dataset, PCA processed data, PCA + Kmeans processed data and Kmeans only. The result is shown in Table 4 below:

From the above table, the PCA and Kmeans integration technique improved the performance accuracy of the different algorithms we modelled our dataset with, an exception in performance is when XGBoost algorithm is used. Though there is improvement from applying just XGBoost on the original dataset, the result shown in Table 4 indicates a decline in accuracy from 95% when only Kmeans is integrated with XGBoost as against our proposed PCA and Kmeans technique to a value of 93%. Furthermore, Kmeans alone was shown to be a good procedure to improve the accuracy of each of the algorithms, while PCA reduced the accuracy result when applied alone.

5. Discussion

The experimental results showed that employing PCA enhances the k-means clustering algorithm, as we obtained 614 correctly clustered dataset, versus other studies (See Table 5). The closest result to ours is that of Han Wu et al. [12] which had an accuracy of 95.42% from a sample size of 589 obtained from their k-means clustering. With reference to our experimental result, we can clearly illustrate that the proposed PCA and K-means technique improved the classification accuracy of logistic regression for the Pima Indian diabetic dataset. A comparison with the classification result reported by other researchers

Table 4
Model comparison with different algorithms.

Algorithm	Original dataset	PCA processed	Kmeans only clustered dataset	PCA + KMEANS processed dataset
Logistic Regression	0.77	0.71	0.82	**0.97**
KNN	0.75	0.69	0.93	0.96
XGBoost	0.76	0.66	0.95	0.93
SVM	0.76	0.72	0.83	0.92
Naïve Bayes	0.74	0.73	0.86	0.90

Table 5
Comparison of K-means clustering result.

Author(year)	Methodology	Correctly clustered data	Accuracy percent
Our proposed method	**PCA + K-means**	**614**	**79.94%**
T. Santhanam et al. (2015) [11]	K-means	511	66.53%
Han et al. (2017) [12]	K-means	589	76.69%
Patil B.M et al. (2016) [16]	k-means	433	56.38%
Asha Gowda et al.(2012) [17]	Cascaded K-means	299	38.93%
Mustafa Kadhm(2018) [18]	K-means	570	74.21%

The bold number indicate the best result.

Table 6
Accuracy comparison with other experiments.

Author(year)	Methodology	Accuracy percent
Our proposed method	**PCA + K-means + Logistic regression**	**97.40%**
Han et al. et al. (2017) [12]	K-means + Logistic regression	95.42%
Patil B.M et al. (2016) [16]	k-means + C4.5	92.38%
Sanakal S. et al. (2014) [17]	SVM + Fuzzy C-means clustering	94.30%
Iyer A et al. (2015) [15]	Naive Bayes	79.56%
Kumari A.V. et al. (2013)	SVM	78%
Anjali Khandegar et al. (2017) [6]	PCA + NN	92.2%
Motka et al. (2013) [8]	PCA-ANFIS	89.2%
Tarun et al. (2014) [4]	PCA + SVM	93.66%
Han et al. et al. (2017) [12]	K-means + Logistic regression	95.42%

The bold number indicate the best result.

Table 7
Description of new dataset features.

Attributes	Description
age	The age of each patient
body mass index	The measure of body fat based on height and weight
family history	Indicates if patient have any close family relate ever diagnosed of diabetes.
increased urination	Does patient feel the need to urinate often?
fatigue	Does patient feel tired often?
increased appetite	Is there an abnormal increase in how often patient wants to eat?
weight loss	Any reported case of weight loss?
increased thirst	Did patient report increased thirst?
eating pattern	Do patient have a controlled eating pattern/diet?
regular exercise	Do the patient engage in exercise often?
sex	Gender of patient
blood pressure	Blood pressure of the patient at the time of test.
fasting blood glucose	Diabetes fasting blood glucose test score for patient
oral glucose tolerance test	Test for incidence of diabetes in patient using the ogtt method

Table 9
Clustering result for new dataset.

Methodology	Correctly clustered data	Accuracy percent
PCA + K-means	**773**	**51.53%**
K-means only	737	49.13%

is shown in Table 6.

A key issue solved by our study is the improvement in the accuracy of the prediction model. The PCA technique we proposed contributed much to the improvement of the prediction model. The kappa statistics value of the proposed model is 0.942 (which is almost equal to 1) which indicates that there is a match between the proposed classifier and the real world output.

5.1. New dataset evaluation

A major concern surrounding the development of machine learning

Table 8
Algorithm comparison.

Algorithm	Original dataset	PCA processed	Kmeans only clustered dataset	PCA + KMEANS processed dataset
Logistic Regression	0.47	0.48	0.75	**0.89**
KNN	0.53	0.48	0.67	0.78
XGBoost	0.51	0.49	0.74	0.85
SVM	0.50	0.47	0.45	0.58
Naïve Bayes	0.54	0.52	0.64	0.82

algorithms for medical application is the reliability of such a model when deployed practically. To evaluate the performance of our model, we used a more practical dataset collected from a known population. In collaboration with the Specialist hospital, Benin City, we extracted information from patient records who were tested for diabetes at the medical facility. We formed a dataset from 1500 random records while considering only those records that had no missing values for the features that we needed. The dataset consists of 13 attributes, namely age, body mass index (bmi), family history, increased urination, fatigue, increased appetite, weight loss, increased thirst, eating pattern, regular exercise, sex, blood pressure, fasting blood glucose (fbg) and oral glucose tolerance test (ogtt). The class variable has a distribution of 760 negative and 740 positive cases. A description of the dataset is given in Table 7.

We subjected this dataset to the same preprocessing steps as we did the pima-indian dataset and then used the output to experiment on the performance of our proposed model. To further demonstrate the applicability of our model, we compared its result with that of other algorithm using the new dataset. The result is shown in Table 8. The performance accuracy of our model stood at 89%. This shows that the new method proposed is reliable even when used with a more practical dataset.

In addition, the application of PCA on our new dataset also improved the Kmeans clustering algorithm as shown in Table 9.

6. Conclusion and future work

The aim of this work was to design an efficient model for the prediction of diabetes. After a careful study of other published work, we proposed a novel model, which consists of using PCA for dimensionality reduction, k-means for clustering, and logistic regression for classification. With the intent to improve the k-means result of other researchers, we first applied the PCA technique to our dataset. Though PCA is a well-known technique, its efficiency in improving k-means clustering and in turn the logistic regression classification model has not been given sufficient attention. Through our experiment we have shown that an improved logistic regression model for predicting diabetes is possible through the integration of PCA and k-means. The novelty achieved in the study includes, the ability to obtain an enhanced k-means cluster result far above what other researchers have obtained in similar studies. Also the logistic regression model performed at an improved level in predicting diabetes onset, as compared to the results obtained when other algorithms where used in our study and that of other studies. Another advantage is the fact that our model has the ability to model a new dataset successfully.

Declarations

Availability of data and materials

The Specialist hospital dataset can be requested through edosauwalia@yahoo.com.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Funds for Distinguished Young Scientists of Lanzhou University of Technology (grant number 201304).

Authors' contributions

CSZ and CUI conceived and designed the research. CUI conducted the literature review, developed the code, carried out the experiments and manuscript writing, CSZ and CUI helped interpreting the results. WFF instructed CUI for dataset processing. CSZ instructed CUI for model validation. All authors read and approved the final manuscript.

Ethical statement

There are no extra ethical statement to make.

Acknowledgements

Not applicable.

References

[1] retrieved http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accessed date: 27 July 2018.
[2] http://www.who.int/news-room/fact-sheets/detail/diabetes retrieved 27/07/2018.
[3] https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes/.
[4] Tarun Jhaldiyal, Pawan Kumar Mishra Analysis and prediction of diabetes mellitus using PCA, REP and SVM 2014 Int J Eng Tech Res (IJETR) ISSN: 2321-0869, Volume-2, Issue-8.
[5] Prabhu P, et al. Improving the performance of K-means clustering for high dimensional data set. Int J Comput Sci Eng June 2011;3(6). ISSN: 0975-3397.
[6] Khandegar Anjali. Khushbu Pawar diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm. Int J Digital Appl Contemp Res 2017;5(6).
[7] Novakovic J, Rankov S. Classification performance using principal component analysis and different value of the ratio R. Int J Comput Commun Control 2011;Vol. VI(2):317–27. ISSN 1841-9836, E-ISSN 1841-9844.
[8] Motka Rakesh, Parmarl Viral, Kumar Balbindra, Verma AR. Diabetes mellitus forecast using different data mining techniques. IEEE 4th international conference on computer and communication technology (ICCCT). IEEE; 2013. p. 99–103.
[9] https://en.wikipedia.org/wiki/K-means_Clustering.
[10] Seyed S, Mohammad G, Kamran S. Combination of feature selection and optimized fuzzy apriori rules: the case of credit scoring. Int Arab J Inf Technol 2015;12(2).
[11] Santhanam T, Padmavathi MS. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput Sci 2015;47:76–83.
[12] Wu Han, Yang Shengqi, Huang Zhangqin, He Jian. Xiaoyi Wang Type 2 diabetes mellitus prediction model based on data mining. Inf Med. 2018;10:100–7. Unlocked.
[13] Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd USA: Morgan Kaufmann Publishers; 2012.
[14] Gowda Karegowda Asha, Jayaram MA, Manjunath AS. Cascading K-means clustering and K-nearest neighbor classifier for categorization of diabetic patients. Int J Eng Adv Technol 2012;1(3). ISSN: 2249 – 8958.
[15] Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Manag Process (IJDKP) 2015;5(1).
[16] Patil BM, Joshi RC, Durga Toshniwal. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37:8102–8.
[17] Gowda Karegowda Asha, Punya V, Jayaram MA, Manjunath AS. Rule based classification for diabetic patients using cascaded K-means and decision tree C4.5. Int J Comput Appl 2012;45(12). (0975 – 8887).
[18] Kadhm Mustafa S, Ghindawi Ikhlas Watan, Mhawi Duaa Enteesha. An accurate diabetes prediction system based on K-means clustering and proposed classification approach. Int J Appl Eng Res 2018;13(6):4038–41. ISSN 0973-4562.