IJCERT

International Journal of Computer Engineering in Research Trends

Multidisciplinary, Open Access, Peer-Reviewed and fully refereed

Research Paper Volume-8, Issue-12, 2021 Regular Edition E-ISSN: 2349-7084

Minimal Rule Based Classifier on Diabetic Dataset Using Machine Learning Techniques

[1]Madhavaram Swapna,[2]Dr.D.William Albert

[1,] M.Tech Student, Dr.K.V. Subba Reddy College of engineering for Women,Kurnool, Andhra Pradesh, India.

[2]Professor and HOD, Department of CSE, Dr.K.V. Subba Reddy College of engineering for Women,Kurnool, Andhra Pradesh, India.

**Corresponding Author: swapnasunilkp@gmail.com*

Available online at: http://www.ijcert.org

Received: 22/11/2021, Revised: 25/12/2021, Accepted: 26/12/2021, Published: 29/12/2021

Abstract: - Diabetes mellitus is a chronic, lifelong disorder that affects a large number of people. As a result, finding the most relevant clinical registries and performing fast computer-aided pre-diagnoses and diagnoses will become increasingly important in clinical practise. This paper investigates the use of basic rule-based classifiers over a diabetes dataset utilising PCA (Principal Component Analysis) in order to predict diabetic risk and enhance the classification performance of the classifiers. Specifically, PCA will compress the smallest feature correlation among the features and predict the disease in order to enhance classification performance. As a consequence, PCA increases the classification performance while simultaneously decreasing the computation time required by the system. The classification performance of the Pima Indians Diabetes Dataset is examined with and without PCA, and the performance assessment metrics of precision, recall, accuracy, and F1 Score are used to evaluate the classification performance.

Keywords: Diabetic Classification, Principal Component Analysis, Machine Learning, Support Vector machine, Linear Regression, Decision Tree.

1. Introduction

Diabetes mellitus is a category of metabolic disorders characterized by an unusually high level of fasting plasma glucose or hyperglycemia during oral glucose tolerance testing [1]. It induces insulin resistance combined with impaired insulin secretion by pancreatic B cells [2-3]. Diabetes raises the risk of kidney failure, blindness, nerve damage, damage to the blood vessel, and leads to cardiac conditions [4]. More than 200 million people currently experience type 2 diabetes worldwide. The total number of diabetes patients worldwide is projected to exceed 370 million by 2030[5]. The highest prevalence and occurrence of type 2 diabetes is observed in Pima Indians in Arizona in the world [4].Early identification of diabetes, treatment of hyperglycemia and associated metabolic disorders is crucial [6]. Although the interpretation of patient data and expert decisions are the most significant diagnostic variables in diabetes, vast volumes of patient-related data are held in the Diabetes database in modern medicine, and there is an ever growing distance between data collection and data comprehension (4-6). All available information can always not be analyzed and an informed judgment on fundamental patterns can be made. Intelligent data analysis such as data mining is therefore important to extract the valuable information from these data in order to support decision-makers. Data mining is the search for conations and patterns that are present in large databases but are concealed between a limited number of data, such as the correlation between patient data and its medical diagnosis. Data mining techniques on diabetes data are helpful in predicting the risk factors causing diabetes or in predicting individuals at risk for diabetes [4-9]. Classification is one of the popular applications of data mining techniques for medical diagnosis [7].

1. Various methods and algorithms have been developed in order to derive knowledge and information from medical records for the diagnosis and treatment of diseases. PCA is a simple, non-parametric way to extract relevant data from misleading data sets [4].
2. Classification systems offer shorter and more detailed medical data to be analyzed. The classification is based on its characteristics and the prediction is an indicator, based on findings, experiences or empirical reasons. The goal of classification is to optimize predictive precision; therefore, predictive accuracy is generally accepted by researchers and practitioners as the primary measure [4].
3. In addition, ML-based systems can be used both as techniques of feature selection (FST) and as classificatory. It also allows people to diagnose diabetes correctly and the best classifier is the main problem for the precise stratification of diabetes risk. Different ML-based systems were used to diagnose and predict diabetic disease for early diagnosis.

The main goal is to build a machine-based learning system (ML) for predicting diabetic patients, and propose Principal Component Analysis (PCA) that increases the efficiency of classification by compressing the minimum attributes. In the controlled machine learning classification including Linear SVM Naive Bayes, Decision Tree, K-Nearest Neighbor and Regression we compare classification output to PCA and PCA.

2. Related Work

A literature review is a critical analysis of published sources, or literature, on a particular topic. It is an assessment of the literature and provides a summary, classification, comparison and evaluation. According to my research subject line.

Berina Et Al. [8]: This article discusses the Artificial Neural Network and the Bayesian Network and its use to identify diabetes and CVD diseases. The goal is to demonstrate the comparison of machine learning techniques and to determine the best way to achieve highest classification performance accuracy. This includes the use of PCA for dimensional reduction, k-means for clustering and logistic classification. While PCA is well known, it has not been given sufficient attention to boost k-means clustering and the logistical regression classification model. Through this experiment we have shown that the combination of PCA and k-means makes an improved logistic regression model for predicting diabetes possible. The innovation achieved in this study involves the potential to obtain an improved K-means cluster result well beyond what other researchers in related studies have obtained.

K. Kourou Et.Al [9] The Author has proposed PCA for dimensionality reduction in this paper that helps define suitable initial centroids for diabetic data set while applying the k-means algorithm. K-means is then used to identify outliers and cluster data into related classes, with regression as a dataset classifier. The design and implementation of the proposed model is focused on the advantages of PCA, K-means and logistical regression. Then a new approach is suggested by PCA to transform the initial collection of features to solve the correlation problem, which makes it hard to find correlations between the data in the classification algorithm. PCA helps to out redundant features and thereby decreases the time, expense and efficiency of the product.

Song et al. [10] describe different classification Algorithms using different parameters such as Glucose, Blood Pressure, Skin Thickness, insulin, BMI, Diabetes Pedigree, and age. The researches were not incorporated pregnancy parameter to predict diabetes disease (DD). In this research, the researchers were using only a small sample of data for prediction of Diabetes. The algorithms were used by this paper were five different algorithms GMM, ANN, SVM, EM, and Logistic regression. Finally. The researchers conclude that ANN (Artificial Neural Network) was providing High accuracy for the prediction of Diabetes.

2.1 Problem statement: Diabetic classification is an important and difficult problem for diagnosis and diabetic perception. A rule-based classifier therefore plays an significant role in today's diabetic diagnosis. The noble classifier offers high-precision classification rules from past diagnoses. Meanwhile, each diagnosis consists of the enormous size of the data features; from such historical data it encounters minimally precise classification laws. In theory, the reduction of features will help reduce the number of classification rules. In the other hand, it decreases the efficiency of classification.

3. System Study

3.1 Existing System

Diabetes causes multiple deaths globally and many individuals who deal with the condition don't know their health early enough. Diabetes is more prevalent in people's daily lives as living conditions grow. Therefore, how to diagnose and evaluate diabetes easily and reliably is a worthy topic. Machine learning can allow users to make a tentative decision on diabetes mellitus based on regular physical test data which can act as a doctor's guide. For machine learning the most critical issues are how to pick the appropriate features and the proper classifier. A few algorithms have been used to forecast diabetes by using a maximum rule-based classification. The precision of the prediction is diminished because the consistency of medical knowledge is not sufficient

Pitfalls of the study

Conventional techniques for clinical decision support systems are based on a single classifier which results the low accuracy. Classification rules are derived from previous diagnosis with a large amount of features, it challenges to build a minimal number of rules with high performance while retaining all diagnosis information.

3.2 Proposed System

Diabetes mellitus is a communal and complex chronic lifelong disease. Henceforth, it is of high clinical significance to find the most relevant clinical directories and to perform efficient computer-aided pre-diagnoses and diagnoses. Classification rules are derived from previous diagnosis with a large amount of features; it challenges to build a minimal number of rules with high performance. In this project we proposed Principal Component Analysis (PCA) which improves the classification performance by compress the minimal attributes. We compare the classification performance with PCA and without PCA among the supervised machine learning classifier such as Linear SVM Naive Bayes, Decision Tree and Regression and also evaluate the prediction accuracy similarly Minimal rule based classifier improve the classification accuracy by reducing the correlation features among the large diabetic dataset. PCA helps in reduce the dimensions of raw data. Improves the classification performance and save the computation time.

4. Methodology

Dataset: The Prima Indian Diabetes Dataset has been used in this study, provided by the UCI Machine Learning Repository. The dataset has been originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of some medical distinct variables, such as pregnancy record, BMI, insulin level, age, glucose concentration, diastolic blood pressure, triceps skin fold thickness, diabetes pedigree function etc. The diabetes data set consists of 768 data points, with 9 features i.e 1. Number of times pregnant 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test. 3. Diastolic blood pressure (mm Hg) 4. Triceps skin fold thickness (mm) 5. 2-Hour serum insulin (mu U/ml) 6. Body mass index (weight in kg/(height in m)2) 7. Diabetes pedigree function 8. Age (years) 9. Class variable (0 or 1) each: where all the patients are female and at least 21 years old. The number of true cases is 268 (34.90%) and the numbers of false cases are 500 (65.10%), respectively, in the dataset. I used four classification techniques, Support Vector Machine (SVM), Decision tree (DT), K-Nearest Neighbor (KNN) and Logistics Regression (LR).

System Architecture:

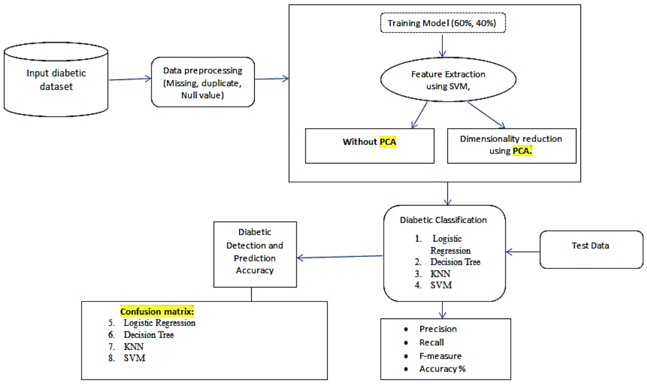


Figure 1. Block Diagram of the Project

From the above block diagram figure 1.operations begin with loading dataset for this project we have been used The Prima Indian Diabetes Dataset consists of 768 data points, with 9 features , after loading the dataset need to perform the data preprocessing is a crucial step for any data analysis problem. It is often a very good idea to prepare your data in such way to best expose the structure of the problem to the machine learning algorithms that you intend to use. This involves a number of activities such as: Assigning numerical values to categorical data; Handling missing values; and normalizing the features (so that features on small scales do not dominate when fitting a model to the data).

Build Training Model: Training and testing dataset

The simplest method to evaluate the performance of a machine learning algorithm is to use different training and testing datasets. Here I will

- Split the available data into a training set and a testing set. (70% training, 30% test)
- Train the algorithm on the first part,

- make predictions on the second part and
- Evaluate the predictions against the expected results.

Feature Extraction with PCA and without PCA

An important machine learning method for dimensionality reduction is called Principal Component Analysis. It is a method that uses simple matrix operations from linear algebra and statistics to calculate a projection of the original data into the same number or fewer dimensions. The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal. Importantly, the dataset on which PCA technique is to be used must be scaled. The results are also sensitive to the relative scaling.

Simple understanding of PCA:

As a layman, it is a method of summarizing data. Imagine some wine bottles on a dining table. Each wine is described by its attributes like colour, strength, age, etc. But redundancy will arise because many of them will measure related properties. So what PCA will do in this case is summarize each wine in the stock with fewer characteristics.

Algorithm-1: PCA Algorithm

Steps of Principal component Analysis:

Step 1 – *Transforming the data to a similar scale*
Step 2 – *Standardized the data.*
i.e., Re – center the original dataset to the origin at means zero
Step 3 – *Calculate eigenvalue and eigenvector of the covariance matrix.*
Step 4 – *Calculate trace and variance explained by principal components.*
Step 5 – *Derive the new data through the selected principal components.*
(*New* = *eigenvector* * *Data*).

Brief description of the PCA in the diabetic dataset, features range is very high (1 to 100) and some features range is very low (0 to 1) due to which high features have more effect on the predictions of output as compared to the low features data. Find the covariance of data to find the correlations between all features. After that find the Eigenvalues and eigenvectors of the covariance matrix. Next, sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues. Use this eigenvectors matrix to transform the samples onto the new subspace. It can also be used for data compression. If we need to transfer data, instead of sending n-dimensional data points, then we can ensure PCA and send m dimensional coordinates in a best-fit subspace (plus the subspace equation). It's often used in machine learning to discover and remove redundant variables in our dataset because many machine learning algorithms perform best when each variable contributes new information.

Machine Learning Classifiers

k-Nearest Neighbors : The k-NN algorithm[11] is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set—its "nearest neighbors." First, Let's investigate whether we can confirm the connection between model complexity and accuracy:

Decision Tree: The accuracy on the training set is 100%, while the test set accuracy is much worse [12]. This is an indicative that the tree is over fitting and not generalizing well to new data. Therefore, we need to apply pre-pruning to the tree. We set max_depth=3, limiting the depth of the tree decreases over fitting. This leads to a lower accuracy on the training set, but an improvement on the test set.

Support Vector Machine: The model over fits quite substantially, with a perfect score on the training set and only 65% accuracy on the test set. SVM [13] requires all the features to vary on a similar scale. We will need to re-scale our data that all the features are approximately on the same scale: Scaling the data made a huge difference! Now we are actually under fitting, where training and test set performance are quite similar but less close to 100% accuracy.

Logistic regression : Logistic Regression is one of the most common classification algorithms[14].

Diabetic diseases prediction

Diabetic disease prediction using machine learning based classifier and compares the prediction accuracy.

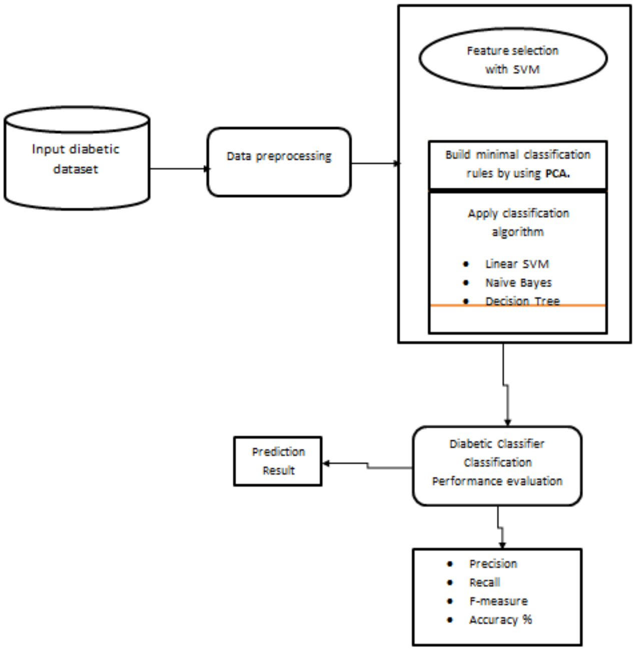

Input diabetic dataset
Data preprocessing
Feature selection with SVM
Build minimal classification rules by using PCA.
Apply classification algorithm
Linear SVM
Naive Bayes
Decision Tree
Prediction Result
Diabetic Classifier Classification Performance evaluation
Precision
Recall
F-measure
Accuracy %

Figure 2.Flow diagram

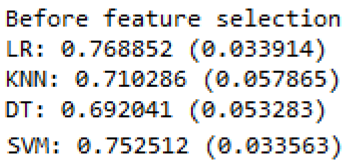
Classifier Evaluation

The most important part after the completion of any classifier is the evaluation to check its accuracy and efficiency. There are a lot of ways in which we can evaluate a classifier. Let us take a look at these methods listed below.

5. Result and Analysis

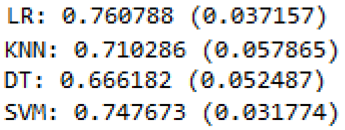
For this experiment we have been used Anaconda framework is an open source distribution of Python and R. It is used for data science, machine learning, deep learning, etc. IDE: Spyder 3.6 is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts.

Feature Selection:
Before feature Selection:



```
Before feature selection
LR: 0.768852 (0.033914)
KNN: 0.710286 (0.057865)
DT: 0.692041 (0.053283)
SVM: 0.752512 (0.033563)
```

PCA feature Selection



```
LR: 0.760788 (0.037157)
KNN: 0.710286 (0.057865)
DT: 0.666182 (0.052487)
SVM: 0.747673 (0.031774)
```

Table 1. Overall Classification performance without PCA

Classifiers	**Training (Accuracy)**	**Testing (Accuracy)**	**Accuracy rate**	**Miss Classification rate**
DT	0.85	0.72	0.72	0.28
KNN	0.79	0.71	0.71	0.29
LR	0.77	0.77	0.85	0.15
SVM	0.81	0.73	0.73	0.27

The best classifier with high accuracy: Linear Regression (85%) .

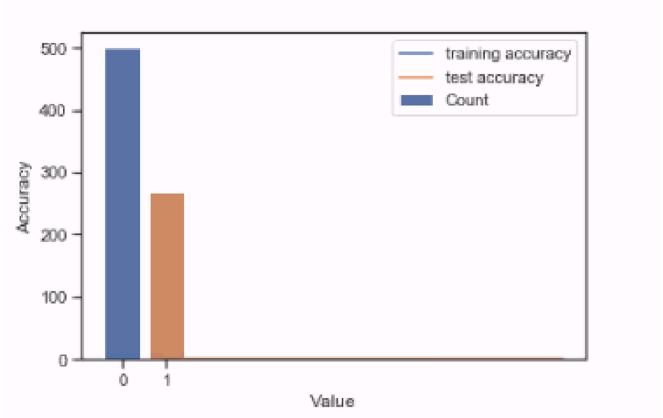


Figure 3. Training and Testing Accuracy count

The best classifier with high accuracy: Linear Regression (85%) .

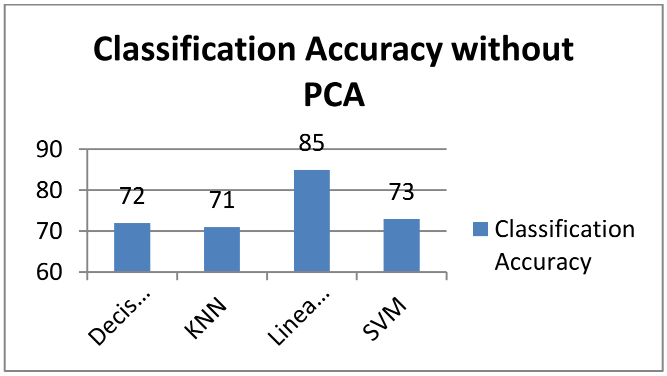


Figure 4. Classification accuracy without PCA

Above experimental results prove that Classification performance without Principal component Analysis is achieved highest with linear regression: 85%.

Classification performance With PCA:

Table 2. Classification performance with PCA

Classifiers	Accuracy (%)
Decision tree	87.17
KNN	77.92
Linear regression	82.93
SVM	77.48

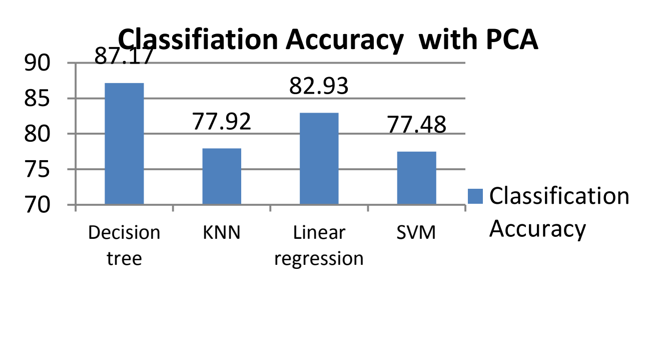


Figure 5. Classification accuracy with PCA

Above experimental results prove that Classification performance with Principal component Analysis is achieved highest with Decision Tree Classifier: 87.17%

6. Conclusion

Diabetes mellitus is a communal and complex chronic lifelong disease. Henceforth, it is of high clinical significance to find the most relevant clinical directories and to perform efficient computer-aided pre-diagnoses and diagnoses. In this project, we demonstrate the performance of Minimal rule based classifier on diabetic dataset using Machine learning based classifiers such as Linear SVM Naive Bayes, Decision Tree and Regression. Finally our project performs the high classification accuracy on PCA based classifiers. **Future Enhancement:** We can extend this work to other dimensionality reduction techniques such as kernel discriminant analysis (KDA), linear discriminant analysis (LDA), and nonparametric discriminant analysis (NDA) to measure the classification performance with some other disease diagnosis datasets.

References

[1] D. Soumya and B Srilatha, Late stage complications of diabetes and insulin resistance, J Diabetes Metab. 2(167) (2011) 2- 7.

[2] K. Papatheodorou, M. Banach, M. Edmonds, N. Papanas, D. Papazoglou, Complications of Diabetes, J. of Diabetes Res. 2015 (2015), 1-5.

[3] L. Mamykinaa, et al., Personal discovery in diabetes self-management: Discovering cause and effect using self-monitoring data, J. Biomd. Informat. 76 (2017) 1–8.

[4] A. Nather, C. S. Bee, C. Y. Huak, J. L.L. Chew, C. B. Lin, S. Neo, E. Y. Sim, Epidemiology of diabetic foot problems and predictive factors for limb loss, J. Diab. and its Complic. 22 (2) (2008) 77-82.

[5] Shiliang Sun, A survey of multi-view machine learning, Neural Comput. & Applic. 23 (7–8) (2013) 2031–2038.

[6] M. I. Jordan, M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science. 349 (6245) (2015) 255-260.

[7] P. Sattigeri, J. J. Thiagarajan, M. Shah, K.N. Ramamurthy, A. Spanias, A scalable feature learning and tag prediction framework for natural environment sounds , Signals Syst. and Computers 48th Asilomar Conference on Signals, Systems and Computers.(2014) 1779-1783.

[8] Alic, Berina & Gurbeta Pokvic, Lejla & Badnjevic, Almir. (2017). Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases. 10.1109/MECO.2017.7977152.

[9] K. Kourou, T. P.Exarchos, K. P.Exarchos, M. V.Karamouzis, D. I.Fotiadis, Machine learning applications in cancer prognosis and prediction, Computation. and Struct. Biotech. J. 13 (2015) 8-17.

[10] Song Y, Cook NR, Albert CM, Van Denburgh M, Manson JE: Effect of homocysteine-lowering treatment with

folic acid and B vitamins on risk of type 2 diabetes in women: a randomized, controlled trial. Diabetes 2009; 58: 1921– 1928.

[11] Sathar, G., Naveen, S., Varma, D.V., Reshma, M., & Nayak, J. (2020). COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DIABETIC PREDICTION.

[12] Ibrahim, N.H., Mustapha, A., Rosli, R., & Helmee, N.H. (2013). A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients.

[13] Chandrakala, & Madhuri, S. (2020). Analysis of Eye Retina for Diabetic Detection using PCA & SVM Methods.

[14] Li, T., Jia, Y., Wang, S., Wang, A., Gao, L., Yang, C., & Zou, H. (2019). Retinal Microvascular Abnormalities in Children with Type 1 Diabetes Mellitus Without Visual Impairment or Diabetic Retinopathy. Investigative ophthalmology & visual science, 60 4, 990-998 .