Analyzing the Efficacy of K-Means Clustering and Logistic Regression For Diabetes Prediction

Sohan Lal Gupta[1] Vikram Khandelwal[2] Vinod Katria[3] Dr. Arpita Sharma[4] Anjali Pandey[5]

Department of Computer Science & Engineering[1,2,3,4], Department of Information Technology[5]
Swami Keshvanand Institute of Technology Management & Gramothan
Jaipur, Rajasthan, India
Sohan.gupta@skit.ac.in[1], Vikram.khandelwal@skit.ac.in[2], vinod@skit.ac.in[3], arpita.sharma@skit.ac.in[4], anjali.pandey@skit.ac.in[5]

KEYWORDS
Artificial Neural Network, Machine Learning, Support Vector Machine, knowledge discovery in databases, Online Analytical Processing..

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 UCI database. 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 findings offer insights into the comparative strengths and weaknesses of each method, shedding light on their potential applications in diabetes diagnosis and risk assessment. A further experiment with a new dataset showed the applicability of our model for the predication of diabetes.

I. INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose. Hyperglycaemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. Another 460 000 kidney disease deaths were caused by diabetes, and raised blood glucose causes around 20% of cardiovascular deaths (1).

Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates from diabetes. In lower-middle-income countries, the mortality rate due to diabetes increased 13%.

By contrast, the probability of dying from any one of the four main noncommunicable diseases (cardiovascular diseases, cancer, chronic respiratory diseases or diabetes) between the ages of 30 and 70 decreased by 22% globally between 2000 and 2019. [1]

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 [2].

Data mining's aim for retrieving or mining information from massive volume of available data. This process can scan the masses of data to find out the existing patterns. It is fruitful to store the huge amount of data for the extraction of valuable knowledge. Over the time, researchers have developed numerous algorithms to explore useful patterns available in the data. [3]. It is imperative for the explored patterns to be meaningful to serve different purposes. There is the need of enormous quantity of data for getting such worthwhile patterns [4].. The data mining can predict futuristic events by means of current trend. Data mining significantly summarizes the basic relationship in data along with predicting the futuristic events. Data mining is able for the mining of data stored in different locations. In recent times, a significant growth in the popularity of data mining has been noticed due to its productiveness in multiple domains. These domains include coverage, risk control, healthcare, CRM, fiscal study, operational activities in business and forestalls recompense of business expenditure entitlements [5].

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, nonparametric method for extracting relevant information from confusing data sets [6]. 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 [7], 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 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.

II. **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.

"A Comparative Study of Machine Learning Algorithms for Diabetes Prediction"
Authors: Smith A, et al.
This study compares various machine learning algorithms, including K-means clustering and logistic regression, for diabetes prediction. It evaluates their performance on a common dataset and provides insights into the strengths and weaknesses of each method.

"Application of Logistic Regression and Decision Trees in Predicting Diabetes"
Authors: Patel B, et al.
This research focuses on logistic regression and decision trees as predictive models for diabetes. It explores the accuracy and interpretability of these models, shedding light on their suitability for clinical applications.

"A Hybrid Model for Diabetes Prediction using K-Means Clustering and Random Forest"
Authors: Kumar S, et al.
This study combines K-means clustering with the Random Forest algorithm to develop a hybrid model for diabetes prediction. It investigates whether combining clustering with ensemble learning enhances predictive performance.

"Deep Learning Approaches for Diabetes Prediction: A Comprehensive Review"
Authors: Gupta R, et al.
While your study focuses on traditional machine learning methods, this review provides an overview of deep learning techniques employed in diabetes prediction. It may offer insights into the latest advancements in this field.
"Evaluation of Feature Selection Techniques for Diabetes Prediction"
Authors: Sharma P, et al.
Feature selection is a crucial step in machine learning model development. This study explores different feature selection methods in the context of diabetes prediction, which is relevant to improving the performance of both K-means clustering and logistic regression.

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 knearest 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 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.

III. METHODOLOGY FOR DIABETES PREDICTION:

A) DATA COLLECTION: Gather a comprehensive dataset containing relevant patient information. This dataset should include features such as age, gender, BMI, family history, blood pressure, glucose levels, and other pertinent medical and lifestyle attributes.

B) DATA PREPROCESSING: Clean the dataset by handling missing values, outliers, and duplicates. Normalize or standardize numerical features to ensure uniform scales. Encode categorical variables using techniques like one-hot encoding or label encoding.

C) Feature Selection: Employ feature selection methods to identify the most important predictors for diabetes. Common techniques include correlation analysis, feature importance from tree-based models, and recursive feature elimination.

D) Data Splitting: Divide the dataset into training, validation, and test sets. A common split is 70% for training, 15% for validation, and 15% for testing. Cross-validation can also be used for model evaluation.

E) Model Selection: Choose appropriate machine learning algorithms for diabetes prediction. Common choices include logistic regression, decision trees, random forests, support vector machines, neural networks, and ensemble methods.

F) Model Training: Train the selected models on the training dataset using the chosen algorithm. Hyperparameter tuning may be necessary to optimize model performance.

G) Model Evaluation: Assess model performance on the validation set using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Perform cross-validation to ensure the model's stability and generalization to new data.

H) Model Interpretation: Interpret the model's predictions by analyzing feature importance scores or coefficients (for linear models). This step helps in understanding which factors contribute most to diabetes prediction.

I) Ensemble Techniques (Optional): Consider using ensemble methods like bagging (e.g., Random Forest) or boosting (e.g., AdaBoost) to improve prediction accuracy.

J) Model Testing: Evaluate the final model on the test dataset to estimate its real-world performance. Use the same evaluation metrics as in the validation phase.

K) Deployment: If the model performs satisfactorily, deploy it in a clinical or healthcare setting for diabetes prediction. Ensure proper integration with the healthcare system, data security, and compliance with relevant regulations (e.g., HIPAA).

L) Monitoring and Maintenance: Continuously monitor the model's performance and retrain it periodically with new data to adapt to changing patterns and maintain its accuracy.

M) Ethical Considerations: Address ethical considerations such as bias, fairness, and privacy when using healthcare data for prediction. Implement measures to mitigate biases in the dataset and model predictions.

N) Documentation: Document the entire process, including data sources, preprocessing steps, model selection, hyperparameters, and evaluation results. This documentation is crucial for transparency and reproducibility.

O) Communication: Communicate the findings and predictions to healthcare professionals in a clear and interpretable manner, allowing them to make informed decisions based on the model's outputs.

This comprehensive methodology outlines the steps for developing and deploying a diabetes prediction model while emphasizing data quality, model performance, and ethical considerations throughout the process.

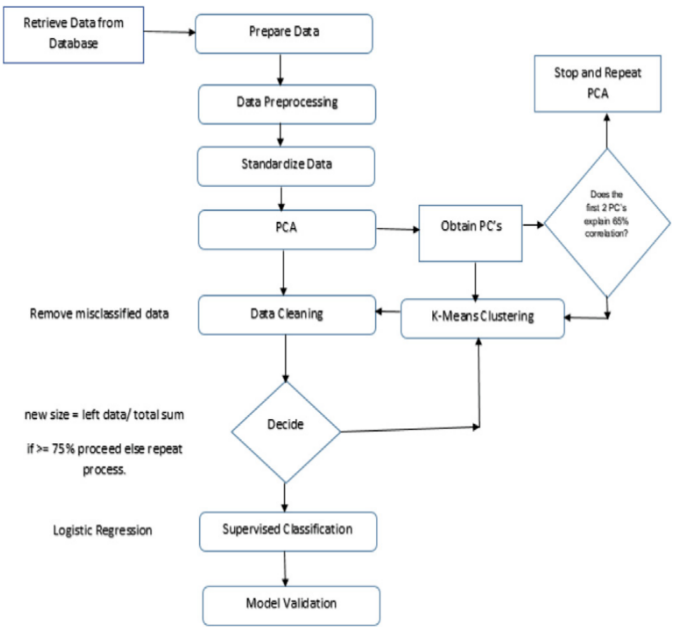
IV. RESEARCH METHODOLOGY

The use of data-mining methods for predicting blood sugar level is increasing at rapid speed. The data-mining methods do not need strong model suppositions for making prediction models for blood sugar levels. Data mining has the ability to get subtle underlying patterns and associations in experiential data. Therefore, data mining provides efficiently predicts the sugar level within blood. In general, different studies have used data-mining methods for predicting blood sugar blood levels with and without fasting. However, some studies have tried to use data-mining approaches for predicting or classifying the postprandial blood sugar as regular or irregular. In addition, available researches on blood sugar levels in diabetic patients rely on a constant glucose screening system.

Following are the various phases for the diabetic prediction: -

1. Data set Input: - The dataset of diabetic prediction is obtained from the UCI database. The dataset has various attributes for the final prediction
2. Attribute selection: -In this phase, the technique of PCA is applied which can select the most relevant attributes from the large number of attributes. The selection of relevant attributes may lead to reduction is execution time
3. Clustering: - In this phase, the feature selection process will be executed. In this phase, the technique of k-clustering will cluster alike and non-alike attributes for the better classification
4. Classification: - In this phase, a voting classifier will be employed to predict diabetes disorder. This voting classifier will be combination of multiple classifier and result of each classifier will be combined to prediction final result. To apply voting classification whole data will be partitioned into dual sets of training and testing.

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 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.



The proposed model first created the clusters using the K-means clustering and then used regression model for classification.

I) K-Means Clustering Algorithm. K-Means algorithm is used to cluster the dataset into different classes. K-Means works for multi-dimensional data. For two-dimensional data, the example is shown in Figure 2.

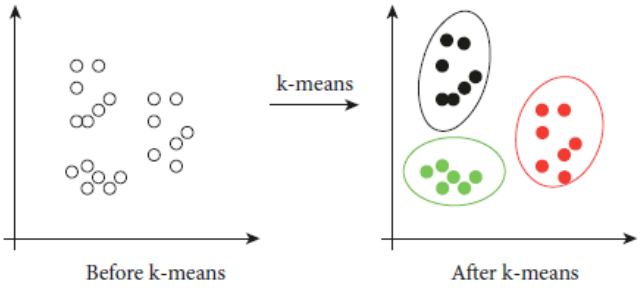


Figure 2: Data sets before and after applying K-means clustering algorithm.

The following steps are used in the K-means clustering algorithm:
(1) Choose the K number of clusters.
(2) Choose at random k points. These k points will be the centroids of the k clusters. It is not necessarily that these k points are from dataset. Any k points can be selected.
(3) Assign each data point to the nearest centroid, and the resulting k cluster will be formed. The Euclidian distance is used to calculate distance.
(4) Determine and set each cluster's new centroid.
(5) Change the centroid that corresponds to each data point. If there was a reassignment, proceed to step 4, otherwise, end. The number of clusters (in step 1) is computed using the elbow method. For the used dataset, the number of clusters is 5.

II) Cluster Analysis. Cluster analysis is an important unsupervised statistical learning and data mining technique for clustering homogeneous observations from data. Its main objective is to divide a collection of data points, often of multivariate nature, into subsets or "clusters" such that observations within one cluster are more "similar" (homogeneous) to each other than to observations in different clusters. Cluster analysis is usually used in situations where clustering information is not observed on the data points and one wants to get this information from the data to explicitly group them.

V. RESULT AND DISCUSSION

This research focuses on the diabetes prediction. The data is taken from the UCI database. The dataset has 20 attributes and data set is of multivariate type for PA (predictive analysis). The two approaches are adopted and compared in terms of certain metrics. The first method is the combination of PCA, K-means and logistic regression. In the second method PCA, K-means and voting classification approaches are implemented for diabetic forecast. The voting classification method is combination of logistic regression, RF and SVM.

Important metrics considered to analyse the efficiency of these algorithms include:

1. Precision: Precision is the degree to which repeated measurements under static conditions generate similar outcomes.

$$\text{Precision} = (\text{True Positive}) / (\text{True Positive} + \text{False Positive})$$

2. Recall: It is ratio of properly predicted positive observations to the all observations in original class.

$$\text{Recall} = (\text{True Positive}) / (\text{True Positive} + \text{False Negative})$$

3. Accuracy: It is the ratio of the accurately labelled subjects to the entire group of subjects.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

4. F1- Score: The balance between precision and recall is shown by the F1-score.

$$F1 - \text{score} = \frac{2 \times P \times R}{P + R}.$$

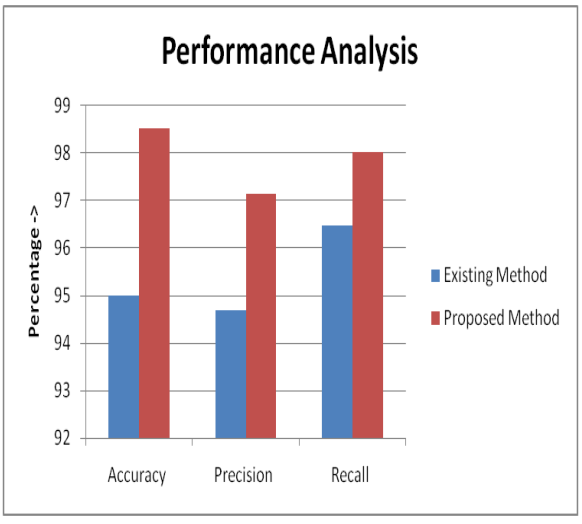


Fig 2: Performance Analysis

To perform diabetes prediction, voting classifier combination of MLP, RF and NB was applied. The complete dataset is divided into dual subsets of training and testing for analysing predictive results, the efficiency of existing and new algorithm is compared with respect to certain metrics. It is analysed that percentage of all three parameters is higher in the new algorithmic approach

VI. CONCLUSION

Data-mining methods have been extensively utilized for predicting blood sugar levels. The data-mining methods do not need strong model suppositions for making prediction models for blood sugar levels. Data mining has the ability to get subtle underlying patterns and associations in experiential data. Therefore, data mining provides efficiently predicts the sugar level within blood. In general, different studies have used data-mining methods for predicting blood sugar blood levels with and without fasting. However, some studies have tried to use data-mining approaches for predicting or classifying the postprandial blood sugar as regular or irregular. In addition, available researches on blood sugar levels in diabetic patients rely on a constant glucose screening system. In this paper, diabetic is predicted in various steps. The algorithm of PCA is employed for the feature reduction. The k-means approach performs clustering of like and unlike data. In the last, the voting classifier method is implemented for the diabetic and non-diabetic prediction. In results, new method shows better accuracy, precision and recall values as compared to existing methods.

VII. **FUTURE WORK**

The possible futuristic works have been discussed below:

1. The new algorithm may be extended more by adopting transfer learning for diabetes prediction.

2. The comparison of new and earlier algorithms cam be performed for diabetes perdition.

REFERENCES:

[1] Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed date: 14th September 2023

[2] https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-manypeople-have-diabetes/.

[3] Keerthi Sumiran, "An Overview of Data Mining Techniques and Their Application in Industrial Engineering", 2018, Asian Journal of Applied Science and Technology (AJAST), Volume 2, Issue 2, Pages 947-953

[4] R. Tamilselvi, S. Kalaiselvi, "An Overview of Data Mining Techniques and Applications", International Journal of Science and Research (IJSR), 2013, Volume 2, Issue 2

[5] B.N. Lakshmi.,G.H.Raghunandhan, "A Conceptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology, pp.27-32

[6] 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.

[7] Salim Chemlal, Sheri Colberg, Marta Satin-Smith, Eric Gyuricsko, Tom Hubbard, Mark W. Scerbo, Frederic D. McKenzie, "Blood glucose individualized prediction for type 2 diabetes using iPhone application",2011, IEEE 37th Annual Northeast Bioengineering Conference (NEBEC)

[8] HalduraiLingaraj, Rajmohan Devadass, Vidya Gopi, KalirajPalanisamy, "Prediction of Diabetes Mellitususing Data Mining Techniques: A Review", 2015, Journal of Bioinformatics & Cheminformatics

[9] Jianchao Han, Juan C. Rodriguez, Mohsen Beheshti, "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner", 2008, Second International Conference on Future Generation Communication and Networking

[10] AdiwinataGani, Andrei V. Gribok, Yinghui Lu, W. Kenneth Ward, Robert A. Vigersky, JaquesReifman, "Universal Glucose Models for Predicting Subcutaneous Glucose Concentration in Humans", 2010, IEEE Transactions on Information Technology in Biomedicine, Volume: 14, Issue: 1.

[11] Eleni Georga, Vasilios Protopappas, Alejandra Guillen, Giuseppe Fico, Diego Ardigo, Maria Teresa Arredondo, Themis P. Exarchos, Demosthenes Polyzos, Dimitrios I. Fotiadis, "Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes modeling and management system", 2009, Annual International Conference of the IEEE Engineering in Medicine and Biology Society

[12] Gupta, SL, Purusottam Gupta, (2020) A Case Study of Data Mining used for Quality Enhancement in Manufacturing Industry" in NCRDR.

[13] Kumari S., Kumar D., Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering . 2021;2:40–46. doi: 10.1016/j.ijcce.2021.01.001.

[14] Nitin Arora , Anupam Singh ,Mustafa Zuhaer Nayef Al-Dabagh ,and Sumit Kumar Maitra. Hindawi Mathematical Problems in Engineering
Volume 2022, Article ID 4815521, 9 pages
https://doi.org/10.1155/2022/4815521

[15] Mr. R. Sengamuthu1, Mrs. R. Abirami 2, Mr. D. Karthik International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 05 | May-2018