International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

Novel Prediction of Diabetes Disease by Comparing K-Means with Logistic Regression with Improved Accuracy

R.Vinoth, Associate Professor Dr.S.R.Raja
Master of Computer Applications
Centre for Open and Digital Education, HITS, Chennai, Tamil Nadu

Abstract- Aim: This study aims to evaluate the effectiveness of the K-Means algorithm in comparison to Logistic Regression (LR) for analyzing a diabetes dataset. Diabetes is a critical and potentially fatal condition, and as it remains incurable, its prevention and management are vital public health concerns. Materials and Methods: For this research, a substantial dataset was sourced from the Kaggle Dataset - Diabetes Disease Analysis and Prediction, encompassing 13 clinical features pertinent to diabetes. The sample consisted of 10 instances, with additional control variables incorporated to account for possible confounding factors and enhance the accuracy of the findings. Both K-Means and Logistic Regression algorithms were employed for predictive analysis. Discussions: Two distinct analyses were conducted to assess the performance of the K-Means algorithm against the proposed LR algorithm. The outcomes indicated that the enhanced LR method yielded superior results. Result: The mean accuracy for the LR algorithm was recorded at 76.8%, while K-Means clustering achieved a mean accuracy of 46.2%, demonstrating that LR outperformed K-Means. The results suggest that machine learning techniques can effectively predict diabetes. The p-value obtained in this study was 0.001, which is less than the threshold of p=0.05, underscoring the importance of utilizing LR for diabetes prediction. Conclusion: The findings reveal that the extended LR algorithm achieved greater accuracy compared to the K-Means algorithm. Nonetheless, it is noted that Logistic Regression would benefit from a larger sample size to enhance the precision of the results.

Index Terms- Machine Learning, Diabetes, K-Means, Logistic Regression, Clustering, Accuracy, High Risk.

I. INTRODUCTION

In recent years, numerous studies have indicated that genetic testing can effectively identify individuals at high risk for developing diabetes, enabling the provision of tailored counseling and education. However, many of these studies were limited by small sample sizes and failed to consider confounding variables that could have affected the outcomes. Machine learning algorithms have become increasingly popular for predicting diabetes, yielding favorable results. This research has enhanced the accuracy of high-risk diabetes predictions (Sim et al. 2023), leading to improved disease management and patient outcomes. The significance of this research lies in its potential to advance diabetes prediction and the development of machine learning algorithms (Y. Zhang, Bartels, and Xiao 2022). Personalized healthcare is essential for patient treatment, and cost-effective healthcare solutions are necessary. This study explores the applications that have been utilized and improved for the early identification of high-risk diabetes. Patients are encouraged to monitor their diabetes treatment, utilizing early detection applications (Zou et al. 2022; Edoh, Pawar, and Mohammad 2018). Consequently, treatment is tailored to individual patients. In the context of disease prediction, population health management is enhanced, resulting in better patient outcomes and improved treatment options. This research employed two algorithmic approaches: K-Means and Logistic Regression (Zou et al. 2022; Edoh, Pawar, and Mohammad 2018; L. Zhang and Liu 2022). Among these, K-Means is an established method, and its accuracy was compared to that of an improved Logistic Regression algorithm, which demonstrated superior accuracy over the K-Means clustering approach. The objective of this study is to fill existing research gaps by enhancing the accuracy of the same dataset using an improved Logistic Regression model, thereby contributing to the accuracy and practical application of high-risk predictive models for diabetes (Miller and Shaw 2006). The research commenced with a thorough search of pertinent articles across reputable academic databases, including Google Scholar, PubMed, and ScienceDirect. Numerous research articles have been published regarding Machine Learning techniques applied to high-risk diabetes. Over the last five years, there have been 18,300 articles identified on Google Scholar, 462 on PubMed, and 95 on IEEE Xplore. Diabetes is a chronic metabolic condition affecting millions globally and is a significant contributor to morbidity and mortality (Ajabnoor et al. 2023).

International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

Timely identification and management of individuals at risk for diabetes are essential for preventing or delaying its onset and mitigating associated complications (Nabrdalik et al. 2023). Machine learning methods, including K-Means and Logistic Regression, have demonstrated potential in predicting diabetes (Ashok Kumar 2022). A thorough review and analysis of the literature is necessary to extract key findings, methodologies, and limitations from each study. The literature review should be organized by summarizing the main findings in a clear and structured format, utilizing headings and subheadings to categorize studies based on their methodologies or research questions (Fu et al. 2022). Nonetheless, prior research has encountered limitations related to dataset diversity, model transparency, and practical applicability (Damjanovska et al. 2022). Consequently, this study aims to evaluate the predictive accuracy of the K-Means clustering algorithm in comparison to an enhanced Logistic Regression model for identifying individuals at risk of diabetes, utilizing a large and diverse dataset (Zou et al. 2022). Additionally, the study will prioritize the development of a transparent and interpretable model suitable for clinical application to aid in diabetes prevention and management (Oliveira et al. 2023). The findings from this research could significantly influence strategies for diabetes prevention and management, ultimately alleviating the burden of this disease (Smokovski 2020).

II MATERIALS AND METHODS

Numerous research articles have been published regarding Machine Learning techniques applied to high-risk diabetes. Over the last five years, there have been 18,300 articles identified on Google Scholar, 462 on PubMed, and 95 on IEEE Xplore. Diabetes is a chronic metabolic condition affecting millions globally and is a significant contributor to morbidity and mortality (Ajabnoor et al. 2023). Timely identification and management of individuals at risk for diabetes are essential for preventing or delaying its onset and mitigating associated complications (Nabrdalik et al. 2023). Machine learning methods, including K-Means and Logistic Regression, have demonstrated potential in predicting diabetes (Ashok Kumar 2022). A thorough review and analysis of the literature is necessary to extract key findings, methodologies, and limitations from each study. The literature review should be organized by summarizing the main findings in a clear and structured format, utilizing headings and subheadings to categorize studies based on their methodologies or research questions (Fu et al. 2022). Nonetheless, prior research has encountered limitations related to dataset diversity, model transparency, and practical applicability (Damjanovska et al. 2022). Consequently, this study aims to evaluate the predictive accuracy of the K-Means clustering algorithm in comparison to an enhanced Logistic Regression model for identifying individuals at risk of diabetes, utilizing a large and diverse dataset (Zou et al. 2022). Additionally, the study will prioritize the development of a transparent and interpretable model suitable for clinical application to aid in diabetes prevention and management (Oliveira et al. 2023). The findings from this research could significantly influence strategies for diabetes prevention and management, ultimately alleviating the burden of this disease (Smokovski 2020).

The testing was performed on the Google Colab platform, a cloud-based environment designed for the development and evaluation of machine learning models (Kumar 2021). This platform supports widely-used programming languages like Python and offers access to advanced computing resources (McFarlane and Bakris 2012). The testing process involved creating and training a machine learning model utilizing data gathered from participants in both groups. A supervised learning methodology was employed, with input data comprising various demographic and clinical variables, including age, gender, body mass index, blood glucose levels, and other pertinent factors (Zou et al. 2022; Wang et al. 2021). The model's output variable indicated the presence or absence of diabetes in each participant. Data collection was conducted through multiple channels, such as medical records, physical assessments, and laboratory tests (Tarasewicz et al. 2023). Participants underwent a thorough health evaluation, which included an examination of their medical history, physical assessments, and laboratory tests to ascertain their diabetes status (C-Home Heuck et al. 2003). The data was gathered in a secure and confidential manner and subsequently stored in a database for further analysis (Zou et al. 2022).

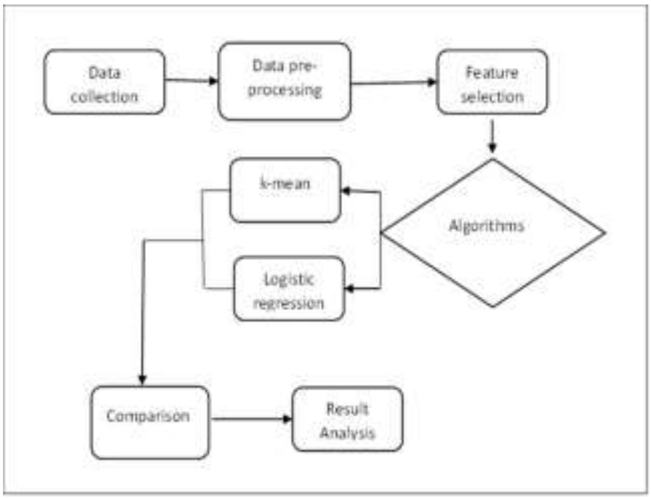

Figure1. Flow chart depicting the methodology adopted in the study

Pseudo Code
Group 1(K-Mean)

- Import libraries: pandas, sklearn.cluster.KMeans, sklearn.preprocessing, sklearn.metrics.accuracy_score.

- Load diabetes dataset using pd.read_csv() and store in 'diabetes_data' variable.
- Remove the 'Outcome' column from the dataset and assign remaining columns to 'X'. Assign the 'Outcome' column to 'y_true'.
- Scale the features in 'X' using Min-Max scaler by creating an object of MinMaxScaler() and using its fit_transform() method. Store the result in the 'X_scaled' variable.
- Create an empty list called 'accuracy_list' to store the accuracy scores.
- Iterate through the range of 10 to execute k-means clustering. In each iteration, instantiate a KMeans() object with n_clusters set to 2 and random_state assigned to i. Train the model using the scaled features by applying the fit() method. Use the labels_ attribute to predict the labels and store them in the 'y_pred' variable. Calculate the accuracy score with the accuracy_score() function, comparing the true labels 'y_true' with the predicted labels 'y_pred'. Convert the accuracy score to a percentage by multiplying it by 100 and add it to the 'accuracy_list'. 7. Display the accuracy scores by using the print() function along with the string 'Accuracy scores:' and the 'accuracy_list' variable. Group 2 (Logistic Regression) 1. Import the required libraries: pandas, LogisticRegression, preprocessing, train_test_split, and accuracy_score. 2. Load the diabetes dataset into the variable diabetes_data using the read_csv method from the pandas library. 3. Separate the dataset into features and labels, where X includes all columns except for the 'Outcome' column, and y consists solely of the 'Outcome' column. 4. Normalize the feature data using the Min-Max scaler from the preprocessing library, storing the result in the X_scaled variable. 5. Create an empty list to hold the accuracy scores. 6. For i in the range of 10, perform the following steps: a. Utilize the train_test_split function from the sklearn library to generate training and testing sets, specifying a test size of 0.1 and using random state "i." b. Instantiate a logistic regression object with random_state set to i, and apply the fit method to train the model. c. Use the predict method to forecast the labels for the testing data, and compute the accuracy score with the accuracy_score method from the sklearn library. d. Add the accuracy score to the accuracy_list. 7. Output the list of accuracy scores.
- Loop through range(10) to perform k-means clustering. In each iteration, create an object of KMeans() with n_clusters=2 and random_state=i. Train the model using the scaled features by employing the fit() method. Predict the labels using labels_ attribute and store them in the 'y_pred' variable. Compute the accuracy score using accuracy_score() function with true labels 'y_true' and predicted labels 'y_pred'. Multiply the accuracy score with 100 to convert it to percentage and append it to 'accuracy_list'.
- Print the list of accuracy scores using print() function with 'Accuracy scores:' string and 'accuracy_list' variable.

Group 2(Logistic Regression)

- Import necessary libraries: pandas, LogisticRegression, preprocessing, train_test_split, accuracy_score
- Load the diabetes dataset using the read_csv method from pandas library and store it in the diabetes_data variable.
- Split the dataset into features and labels, where X contains all columns except the 'Outcome' column, and y contains only the 'Outcome' column.
- Scale the feature data using Min-Max scaler from the preprocessing library, and store it in the X_scaled variable.
- Initialize an empty list to store the accuracy scores.
- For i in range 10, repeat the following steps:
- Apply the train_test_split function from the sklearn library to create training and testing sets, with a testing size of 0.1 and using the random state "i.".
- Create a logistic regression object with random_state=i, and use fit method to place the data.
- Predict the labels for the testing data using predict method, and calculate the accuracy score using accuracy_score method from sklearn library.
- Append the accuracy score to the accuracy_list.
- Print the list of accuracy scores.
- Statistical Analysis The analysis of diabetes disease prediction was conducted using Logistic Regression and K-Means clustering techniques. SPSS software version 29 was utilized to perform t-tests, enabling a comparison of the algorithms based on accuracy, mean values, standard deviation, and standard error. A t-test was executed at a 95% confidence interval to determine if there were significant differences between the algorithms. III. RESULTS Table 1 presents the classification accuracy for both K-Means clustering and Logistic Regression algorithms in predicting diabetes. The accuracy of K-Means is notably lower than that of Logistic Regression, recording only 47.52% compared to 84.41% for Logistic Regression. Table 2 details the independent samples comparison conducted to assess the performance of K-Means against the Logistic Regression algorithm for diabetes prediction. The mean accuracy for K-Means was found to be 46.2%, while the Logistic Regression algorithm achieved a mean accuracy of 76.8%. The standard deviation for K-Means was 17.04113, whereas the Logistic Regression algorithm had a standard deviation of 4.56557.

Table 1. Classification Accuracy for K-Mean and Logistic Regression

Algorithm	Accuracy_score	F1_score	Recall_score	Precision score
KMeans	46.2%	66%	55%	84%
Logistic Regression	76.8%	75%	69%	82%

Statistical Analysis The investigation into diabetes disease prediction employed Logistic Regression and K-Means clustering methodologies. The analysis was carried out using SPSS software version 29, which facilitated the execution of t-tests to compare the algorithms based on metrics such as accuracy, mean values, standard deviation, and standard error. A t-test was performed at a 95% confidence interval to ascertain whether significant differences existed between the algorithms.

III. RESULTS

Table 1 illustrates the classification accuracy of both K-Means clustering and Logistic Regression algorithms in the context of diabetes prediction. The accuracy of K-Means is significantly lower than that of Logistic Regression, achieving only 47.52% in contrast to 84.41% for Logistic Regression.

Accuracy	Algorithm	N	Mean	Std. Deviation	Std. Error Mean
	K-Means	10	46.2	17.04113	5.38888
	Logistic Regression	10	76.8	4.56557	1.44376

Table 2 provides a detailed comparison of independent samples to evaluate the performance of K-Means relative to the Logistic Regression algorithm for diabetes prediction. The mean accuracy for K-Means was determined to be 46.2%, while the Logistic Regression algorithm attained a mean accuracy of 76.8%. The standard deviation for K-Means was recorded at 17.04113, in comparison to a standard deviation of 4.56557 for the Logistic Regression algorithm

Table 3 illustrates the results of an independent samples comparison aimed at evaluating the efficacy of the K-Means algorithm in relation to a Logistic Regression model for predicting diabetes.

This analysis was conducted with a 95% confidence interval and a significance threshold of $p>0.05$. The findings indicated a statistically significant difference between the two methodologies, with a confidence interval of less than .001. Table 3. Statistical independent sample test comparing K-Means and Logistic Regression with a 95% confidence interval. The results demonstrate a statistically significant difference between the K-Means and Logistic

Regression, with $p=0.001$ $(p<0.05)$.

Accuracy		F	Sig	T	df	Significance One-sided p	Significance Two-sided p	Mean Difference	Std Error Difference	95% Confidence Interval Of the Difference Lower	95% Confidence Interval Of the Difference Upper
	Equal variances assumed	82.051	<001	-5.485	18	<.001	<001	-30.6000	35.5789	-42.32089	-18.87911
	Equal variances not assumed			-5.485	10.283	<001	<001	-30.6000	5.57893	-42.98403	-18.21597

Figure 2 presents a comparison of mean accuracy, demonstrating that the Logistic Regression algorithm surpassed the KMeans algorithm, achieving a mean accuracy of 76.80% in contrast to 46.20%.

Additionally, the standard deviation associated with the Logistic Regression algorithm was observed to be lower than that of the KMeans algorithm.

The results are illustrated in a graph where the X-axis represents the algorithms and the Y-axis indicates the mean accuracy. The error bars are represented as ± 2 standard deviations, corresponding to a 95% confidence interval.

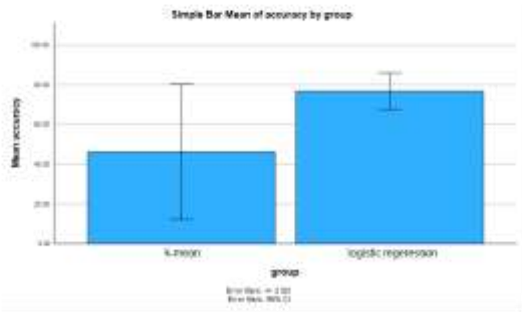

Figure 2. Comparison of mean accuracy

IV. DISCUSSION

The findings of this study demonstrate that Logistic Regression significantly surpassed K-Means in the prediction of diabetes, achieving a mean accuracy of 76.8% compared to K-Means' 46.2%. This suggests that advanced Logistic Regression methods may provide a more effective means of predicting diabetes than unsupervised learning methods like K-Means. One explanation for this disparity in accuracy is that Logistic Regression, as a supervised learning method, utilizes labeled data to enhance its predictive capabilities. In contrast, K-Means operates as an unsupervised learning technique, clustering data points based solely on their similarities without the benefit of labeled data for prediction. These results carry significant implications for the creation of predictive models for diabetes. There is potential to further enhance Logistic Regression by integrating additional features or methodologies to boost accuracy. Techniques such as recursive feature elimination or principal component analysis can be employed to pinpoint the most critical predictors of diabetes. Furthermore, regularization methods like Lasso or Ridge Regression can help mitigate overfitting and enhance the model's generalizability. It is essential to recognize that while Logistic Regression demonstrated superior performance over K-Means in this analysis, both methodologies possess unique advantages and limitations. Their effectiveness may vary based on the specific dataset and research objectives. K-Means can be beneficial for uncovering patterns and similarities within data, whereas Logistic Regression may be more adept at predicting outcomes based on a defined set of input features.

The accuracy score of K-Means is markedly inferior to that of Logistic Regression, achieving only 47.52% compared to Logistic Regression's 84.41%. This disparity highlights the superior predictive capability of Logistic Regression in diagnosing diabetes. Additionally, the f1_score for Logistic Regression surpasses that of K-Means, signifying a more favorable balance between precision and recall. K-Means records a recall score of merely 55%, whereas Logistic Regression achieves a recall score of 69%, demonstrating its enhanced ability to correctly identify true positive cases. Although K-Means exhibits a higher precision score of 84% in comparison to Logistic Regression's 82%, precision alone does not adequately assess the performance of a predictive model, as it may be affected by the prevalence of false positives. The findings presented in Table 1 indicate that Logistic Regression is a more effective algorithm than K-Means for diabetes prediction, as it consistently achieves superior accuracy, f1_score, and recall metrics. In summary, the outcomes of this study indicate that enhanced Logistic Regression is a more proficient method than K-Means for predicting diabetes, with notable improvements in accuracy. Future investigations should consider the implementation of additional feature selection and regularization techniques to further enhance the accuracy and generalizability of Logistic Regression models in diabetes prediction. Moreover, exploring other machine learning methodologies may yield further advancements in predictive accuracy for diabetes.

V. CONCLUSION

The evaluation of K-Means in conjunction with an improved Logistic Regression model for predicting diabetes demonstrated a notable enhancement in accuracy. The enhanced Logistic Regression achieved an accuracy rate of 76.80%, surpassing the K-Means model, which recorded an accuracy of 46.20%. Statistical analysis indicated a significant disparity between the mean accuracies of the two models, with the enhanced Logistic Regression exhibiting a higher mean accuracy. Furthermore, the effect size was considerable, underscoring the substantial difference between the performance of the two models.

REFERENCES

1. Ashok Kumar, M. 2022. Prediction of Type 2 Diabetes Mellitus Using Machine Learning Techniques.
2. Ajabnoor, G. M. A., Bahijri, S. M., Enani, S. M., Alsheikh, L., Ahmed, M., Alhozali, A., ... Al-Rubeaan, K. (2023). Exploring the Validity of Available Markers and Indices in the Diagnosis of Nonalcoholic Fatty Liver Disease (NAFLD) in People with Type 2 Diabetes in Saudi Arabia. Diseases (Basel, Switzerland), 11(1)..
3. C-Home Heuck, P-Kanagasabapathy A-Reinauer, Hans Reinauer, P-Kanagasabapathy Heuck (a-Reinauer C-Home), and World Health Organization. 2003. Laboratory Diagnosis and Monitoring of Diabetes Mellitus.
4. Damjanovska, Sofi, Daniel B. Karb, Alok Tripathi, Jessica Asirwatham, Sarah Delozier, Jaime A. Perez, Yngve Falck-Ytter, and Stanley Cohen. 2022. “Accuracy of Ultrasound Elastography and Fibrosis-4 Index (FIB-4) in Ruling Out Cirrhosis in Obese Non-Alcoholic Fatty

Liver Disease (NAFLD) Patients." Cureus 14 (9): e29445.
5. Edoh, Thierry, Pravin Pawar, and Sagar Mohammad, 2018. Pre-Screening Systems for Early Disease Prediction, Detection and Prevention. IGI Global.
6. Fu, Xiaomin, Yuhan Wang, Ryan S. Cates, Nan Li, Jing Liu, Dianshan Ke, Jinghua Liu, Hongzhou Liu, and Shuangtong Yan. 2022. "Implementation of Five Machine Learning Methods to Predict the 52-Week Blood Glucose Level in Patients with Type 2 Diabetes." *Frontiers in Endocrinology* 13: 1061507.
7. Kumar, Govind. 2021. Python Programming: With Google Colab Development Environment. Govind Kumar.
8. McFarlane, Samy I., and George L. Bakris. 2012. Diabetes and Hypertension: Evaluation and Management. Springer Science & Business Media.
9. Miller, D. Douglas, and Leslee J. Shaw. 2006. coronary artery disease: Diagnostic and Prognostic Models for Reducing Patient Risk. The Journal of Cardiovascular Nursing 21 (6 Suppl 1): S2–16; quiz S17–19.
10. Nabrdalik, Katarzyna, Hanna Kwiendacz, Karolina Drożdż, Krzysztof Irlik, Mirela Hendel, Agata M. Wijata, Jakub Nalepa, et al. 2023. "Machine Learning Predicts Cardiovascular Events in Patients with Diabetes: The Silesia Diabetes-Heart Project." *Current Problems in Cardiology*, March, 101694.
11. Oliveira, Bruno Alberto Soares, Giulia Zanon Castro, Giovanna Luiza Medina Ferreira, and Frederico Gadelha Guimarães. 2023. "CML-Cardio: A Cascade Machine Learning Model to Predict Cardiovascular Disease Risk as a Primary Prevention Strategy." *Medical & Biological Engineering & Computing*, January. https://doi.org/10.1007/s11517-022-02757-z.
12. Poretsky, Leonid. 2013. Principles of Diabetes Mellitus. Springer Science & Business Media.
13. Shrine, Nick, Abril G. Izquierdo, Jing Chen, Richard Packer, Robert J. Hall, Anna L. Guyatt, Chiara Batini, et al. 2023. "Multi-Ancestry Genome-Wide Association Analyses Improve Resolution of Genes and Pathways Influencing Lung Function and Chronic Obstructive Pulmonary Disease Risk." Nature Genetics 55 (3): 410–22.
14. Sim, Ruth, Chun Wie Chong, Navin Kumar Loganadan, Noor Lita Adam, Zanariah Hussein, and Shaun Wen Huey Lee. 2023. "Comparison of a Chronic Kidney Disease Predictive Model for Type 2 Diabetes Mellitus in Malaysia Using Cox Regression versus Machine Learning Approach." Clinical Kidney Journal 16 (3): 549–59.
15. Smokovski, Ivica. 2020. Managing Diabetes in Low Income Countries: Providing Sustainable Diabetes Care with Limited Resources. Springer Nature.
16. Tarasewicz, Dariusz, Andrew J. Karter, Noel Pimentel, Howard H. Moffet, Khanh K. Thai, David Schlessinger, Oleg Sofrygin, and Ronald B. Melles. 2023. "Development and Validation of a Diabetic Retinopathy Risk Stratification Algorithm." *Diabetes Care*, March. https://doi.org/10.2337/dc22-1168.
17. Tapia, German, Tommi Suvitaival, Linda Ahonen, Nicolai A. Lund-Blix, Pål R. Njølstad, Geir Joner, Torild Skrivarhaug, Cristina Legido-Quigley, Ketil Størdal, and Lars C. Stene. 2021. "Prediction of Type 1 Diabetes at Birth: Cord Blood Metabolites vs Genetic Risk Score in the Norwegian Mother, Father, and Child Cohort." The Journal of Clinical Endocrinology and Metabolism 106 (10): e4062–71.
18. Wang, Yingying, Lisha Yu, Yiying Wang, Jie Zhou, Yanli Wu, Tao Liu, Na Wang, and Chaowei Fu. 2021. "Postload Plasma Glucose but Not Fasting Plasma Glucose Had a Greater Predictive Value for Cardiovascular Disease in a Large Prospective Cohort Study in Southwest China." Frontiers in Cardiovascular Medicine 8: 815357.
19. Zhang, Lindong, and Min Liu. 2022. "Analysis of Diabetes Disease Risk Prediction and Diabetes Medication Pattern Based on Data Mining." Computational and Mathematical Methods in Medicine 2022 (October): 2665339.
20. Zhang, Yan, Matthew N. Bartels, and Han Xiao. 2022. Innovative Patterns and Technologies of Cardiac Rehabilitation in Patients With Coronary Artery Disease. Frontiers Media SA.
21. Zou, Yutong, Lijun Zhao, Junlin Zhang, Yiting Wang, Yucheng Wu, Honghong Ren, Tingli Wang, et al. 2022. "Development and Internal Validation of Machine Learning Algorithms for End-Stage Renal Disease Risk Prediction Model of People with Type 2 Diabetes Mellitus and Diabetic Kidney Disease." Renal Failure 44 (1): 562–70.
22. 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.
23. P. Prabhu, et al. Improving the performance of K-means clustering for high dimensional data set Int J Comput Sci Eng, 3 (6) (June 2011) ISSN: 0975-3397
24. Anjali Khandegar Khushbu Pawar diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm Int J Digital Appl Contemp Res, 5 (6) (2017)
25. J. Novakovic, S. Rankov Classification performance using principal component analysis and different value of the ratio R Int J Comput Commun Control, Vol. VI (2) (2011), pp. 317-327 ISSN 1841-9836, E-ISSN 1841-9844
26. Rakesh Motka, Viral Parmarl, Balbindra Kumar, A.R. Ver ma Diabetes mellitus forecast using different data mining techniques IEEE 4th international conference on computer and communication technology (ICCCT), IEEE (2013), pp. 99-103.