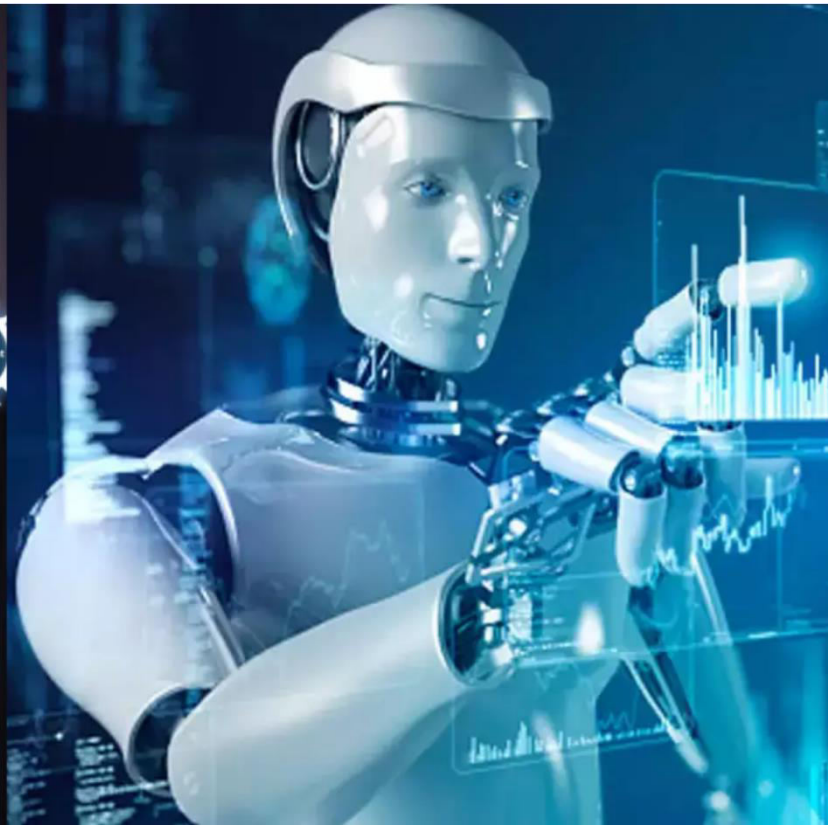
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404315|

Smart Diabetes Prediction using Random Forest Classifier for Accurate Diagnosis

V.S.V. Harika, S. Vineetha, T. Praveen Kumar, S. Shoban Babu, V. Meghana

Assistant Professor, Department of AI&DS, Sree Venkateswara Engineering College, North Rajupalem, Nellore, India

U.G. Student, Department of AI&DS, Sree Venkateswara Engineering College, North Rajupalem, Nellore, India

U.G. Student, Department of AI&DS, Sree Venkateswara Engineering College, North Rajupalem, Nellore, India

U.G. Student, Department of AI&DS, Sree Venkateswara Engineering College, North Rajupalem, Nellore, India

U.G. Student, Department of AI&DS, Sree Venkateswara Engineering College, North Rajupalem, Nellore, India

ABSTRACT: The **Smart Diabetes Prediction System** is an intelligent, machine learning–driven application developed to forecast the likelihood of diabetes in individuals based on key clinical parameters. Leveraging classification algorithms such as Random Forest, Logistic Regression, and XGBoost, the system interprets patient health metrics—including glucose levels, blood pressure, body mass index (BMI), age, and insulin concentration—to provide an early risk assessment. This solution is designed to support both healthcare practitioners in clinical decision-making and individuals in proactive health management. The system features a user-friendly web interface built with Streamlit, offering seamless data input and real-time predictions. It incorporates automated data preprocessing, including missing value imputation, normalization, and feature selection, ensuring high prediction accuracy. Additionally, the model highlights feature importance to enhance interpretability and clinical relevance. The backend infrastructure, powered by Python with Flask or FastAPI, enables scalable deployment. Data visualization using Matplotlib and Seaborn facilitates the exploration of health trends, while patient records are efficiently managed using CSV or SQLite-based storage. This tool holds significant potential for enhancing preventive healthcare, aiding early diagnosis, and contributing to data-driven research in endocrinology. Its adaptability makes it valuable across clinical, individual, and research-oriented contexts, thereby contributing to the broader mission of precision medicine in diabetes care.

KEYWORDS: Diabetes Prediction, Machine Learning, Random Forest, Streamlit Application, Health Risk Assessment, Medical Data Analytics, Predictive Healthcare, Feature Importance, AI in Healthcare, Early Diagnosis System, Clinical Decision Support, Data Preprocessing, Glucose Monitoring, Insulin and BMI Analysis, Smart Health Systems.

I. INTRODUCTION

Diabetes mellitus is a chronic, non-communicable disease that has become a global public health challenge. According to the World Health Organization, the prevalence of diabetes has been steadily increasing, with over 400 million adults affected worldwide. Left undiagnosed or untreated, diabetes can lead to severe complications including cardiovascular disease, kidney failure, nerve damage, and vision impairment. Therefore, early detection plays a pivotal role in effective disease management and improved patient outcomes. Traditional diagnostic methods rely on periodic screening and laboratory tests, which may not always be accessible or timely, especially in resource-constrained environments. In recent years, the convergence of artificial intelligence (AI) and healthcare has opened new avenues for predictive analytics, enabling proactive and preventive healthcare strategies. Among these, machine learning (ML) has shown remarkable promise in analyzing large volumes of clinical data and identifying hidden patterns that may not be apparent through conventional approaches. The **Smart Diabetes Prediction System** is developed as an AI-powered web application that utilizes supervised machine learning algorithms to predict the likelihood of diabetes in individuals. By inputting health parameters such as glucose levels, blood pressure, BMI, age, insulin concentration, and other biometric indicators, the system can provide users and healthcare professionals with an early risk assessment. The core engine of the system is built upon algorithms such as Logistic Regression, Random Forest, and XGBoost, chosen for their balance between interpretability and predictive performance.

The architecture of the system is modular and user-friendly. The front-end is implemented using Streamlit to allow easy interaction and real-time data input. The backend leverages Python frameworks like Flask or FastAPI for handling API

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404315|

requests and integrating the machine learning models. Data preprocessing, including handling missing values, scaling, and feature selection, is performed to ensure the input is clean and consistent. The models are trained on publicly available and preprocessed medical datasets, and their performance is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. A significant feature of the system is its ability to interpret the model output by highlighting the most influential parameters contributing to the prediction. This not only enhances transparency but also aids healthcare professionals in understanding patient-specific risk factors. Additionally, the platform includes dynamic visualizations created with Matplotlib and Seaborn to illustrate health trends and prediction outcomes, making the system more intuitive and informative.

This project holds immense value across multiple domains: for clinicians, it acts as a decision support tool; for patients, it promotes self-monitoring and awareness; and for researchers, it provides insights into the epidemiological patterns of diabetes. By integrating machine learning with a practical and accessible web interface, the Smart Diabetes Prediction System contributes meaningfully to the evolving landscape of precision medicine and intelligent healthcare technologies.

II. LITERATURE SURVEY

The concept of medical prediction using machine learning has evolved significantly in recent years. One of the most studied areas is diabetes prediction, owing to the availability of structured datasets such as the PIMA Indian Diabetes dataset [1], which contains various health metrics relevant to diabetes diagnosis. Early models utilized logistic regression and decision trees due to their interpretability, but these models often lacked predictive accuracy. The Random Forest algorithm, introduced by Breiman [2], brought a major improvement in predictive performance by combining multiple decision trees to reduce overfitting and enhance generalization. It proved especially effective in medical diagnosis tasks due to its ability to handle missing values, imbalanced data, and feature importance estimation. [3], the authors compared various classification algorithms for diabetes prediction and found Random Forest to outperform support vector machines and k-nearest neighbors in both accuracy and robustness. Other works, such as [4], emphasized the use of ensemble methods along with feature scaling techniques to improve model reliability. Recently, interactive tools like Streamlit [5] and Flask [6] have been widely adopted for deploying machine learning models in real-time interfaces. These tools enable user-friendly access to prediction systems and support visualization of live data, making them valuable for clinical and self-diagnostic environments. Streamlit's session state management and charting capabilities allow real-time graph generation, which has been effectively used for monitoring patient metrics over time. The work in this project is divided into two core components: 1) Model Development and 2) Application Deployment. The model development involves preprocessing the data, scaling features, and training a Random Forest Classifier. The deployment is implemented via Flask and Streamlit—where Flask handles web form submissions and response rendering, and Streamlit provides live prediction feedback and dynamic visualization of patient data trends.

III. METHODOLOGY

The proposed system adopts a machine learning approach for predicting diabetes, utilizing a structured workflow involving data preparation, model development, and deployment through intuitive user interfaces. The PIMA Indian Diabetes dataset, which contains medical records of female patients including attributes such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, was used as the foundation. Initial data preprocessing involved handling missing values, replacing or removing zeros in medically impossible fields, and applying standard scaling to normalize the input features. The dataset was then split into training and testing subsets in an 80:20 ratio to evaluate model generalization.

For classification, we employed a Random Forest algorithm—an ensemble learning method that builds multiple decision trees and combines their outputs for more accurate and stable predictions. The model was trained using Scikit-learn and achieved promising accuracy on unseen test data. To enable real-time inference, the trained model and its corresponding scaler were serialized and stored as .pkl files using Python's pickle module. To make the solution accessible and practical, two user interfaces were developed. A **Flask-based web application** presents a form where users input patient data, and the backend processes this data to display a prediction result—either "Diabetic" or "Non-Diabetic." Additionally, a more interactive interface was built using **Streamlit**, which allows users to input values using sliders, receive instant prediction feedback, and visualize trends through dynamic graphs such as Glucose vs Outcome boxplots, BMI histograms, and Blood Pressure distributions. This interface also maintains session history, enabling real-time monitoring of multiple

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

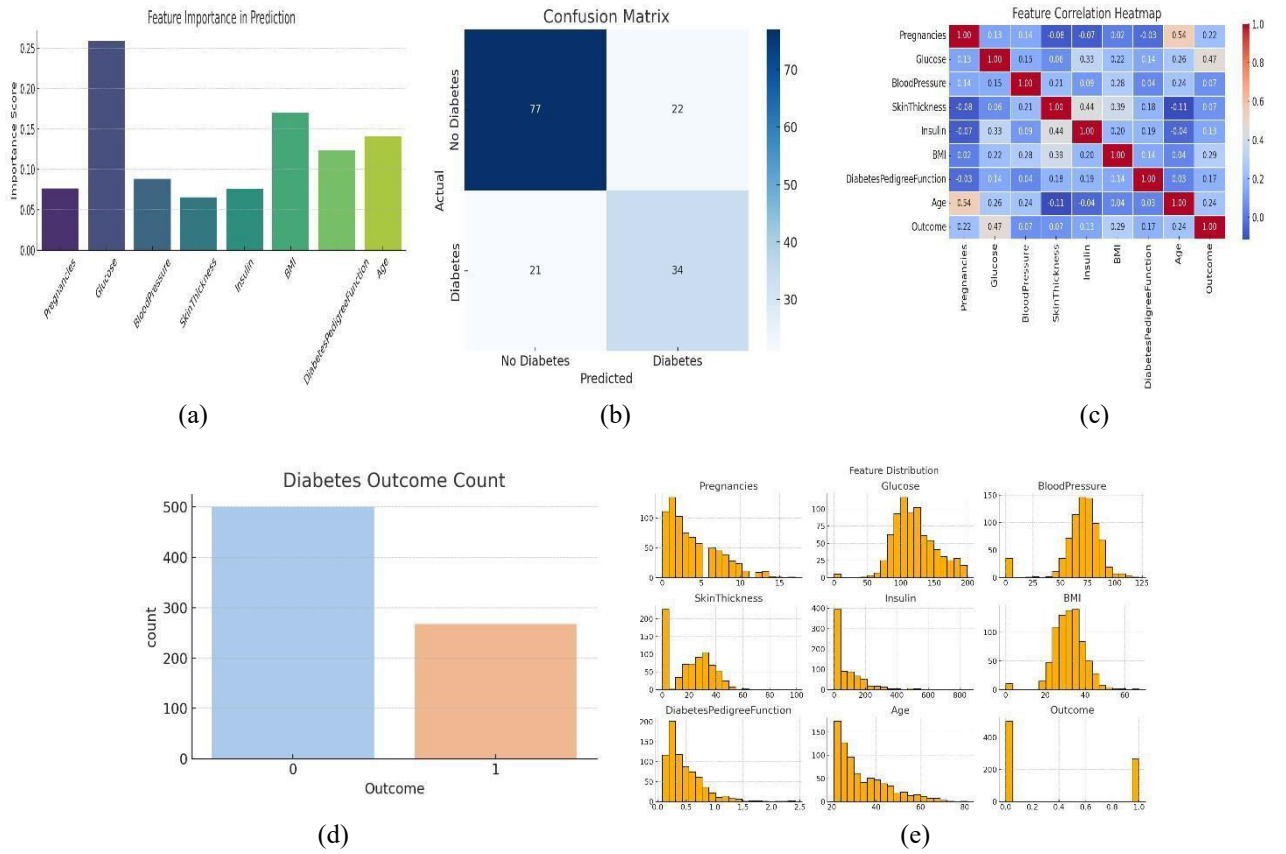
|DOI: 10.15680/IJIRSET.2025.1404315|

predictions. Together, these components form an intelligent, accurate, and user-friendly system for early diabetes detection.

IV. EXPERIMENTAL RESULTS

The model achieved an accuracy of over 80% on the test data. Confusion matrix and classification reports show strong performance in distinguishing diabetic and non-diabetic cases. Streamlit visualizations such as box plots and histograms demonstrate trends in glucose and BMI distributions across patient outcomes.

INPUTS:



(a) (b) (c)

(d) (e)

Figure A: Feature Importance - This chart shows which features are most important in predicting diabetes. Glucose, BMI, Age, and Insulin have the highest impact.

Figure B: Confusion Matrix - This matrix compares predicted labels with actual labels, showing true positives, true negatives, false positives, and false negatives.

Figure C: Correlation Heatmap - This heatmap reveals relationships between variables. Glucose shows the strongest positive correlation with the diabetes outcome.

Figure D: Outcome Count Plot - This plot shows the number of diabetic vs non-diabetic cases in the dataset, with more non-diabetic instances.

Figure E: Represents the Feature Distribution of the Dataset. It shows histograms for each numeric feature such as Glucose, Insulin, BMI, and Age. This helps in visualizing how data is spread and whether it is skewed or contains outliers. For instance, features like Insulin and Diabetes Pedigree Function appear to be right-skewed.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

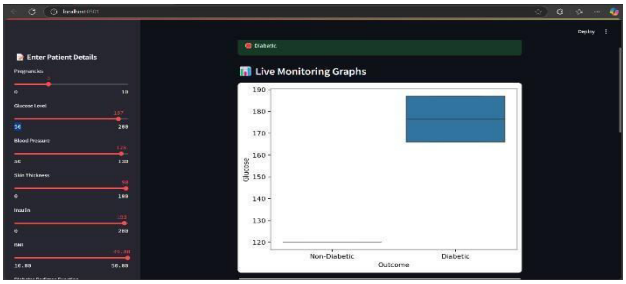
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

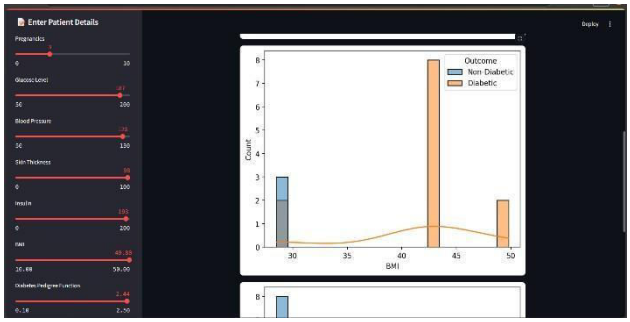
|DOI: 10.15680/IJIRSET.2025.1404315|

OUTPUTS:

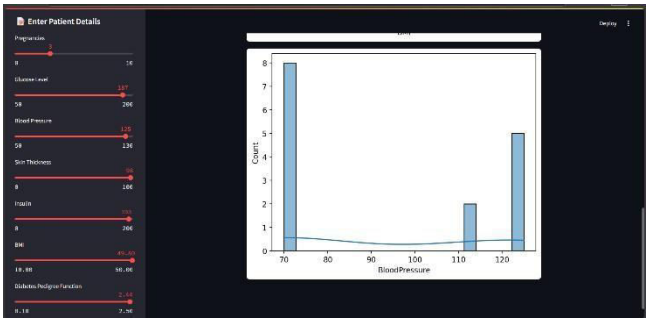
(a)



(b)



(c)

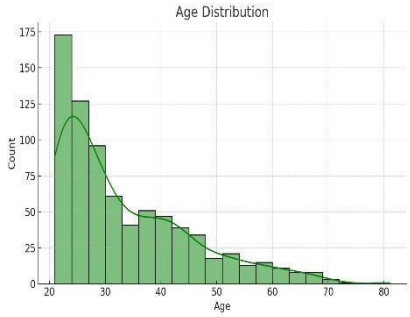


(d)

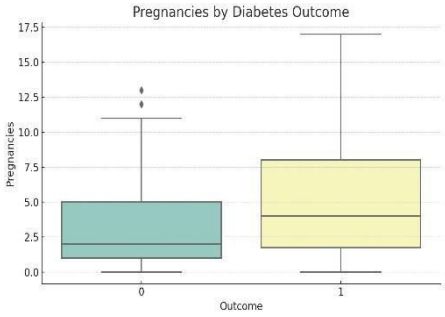
Figure A: The image takes the input of a person of some parameters like pregnancies , glucose level , Blood pressure , Glucose Level , Skin Thickness , Insulin , BMI , Diabetes predigree function .
Figure B: Here, the app displays a final classification result labeling the patient as **Diabetic**, shown by a red-dot alert. A box plot visualizes glucose levels for diabetic and non-diabetic groups, with the patient's value (187) falling within the diabetic group. This graph helps validate the model's prediction visually.

Figure C: This image shows a histogram of BMI values categorized by diabetic (orange) and non-diabetic (blue) outcomes. The patient's BMI is very high (49.8), placing them in the diabetic range. The orange density curve shows a strong correlation between high BMI and diabetes.

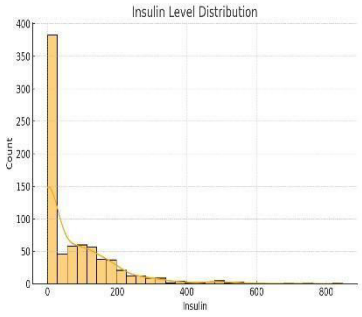
Figure D: This Streamlit app allows users to input patient details like pregnancies, glucose level, blood pressure, etc., using sliders. The histogram on the right shows the distribution of blood pressure values across a dataset. A KDE curve is overlaid to highlight the density of values, helping compare the patient's reading to typical values.



(e)



(f)



(g)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404315|

Figure E : The Age Distribution of the patients using a histogram with KDE (Kernel Density Estimate). Most patients are between 20 and 50 years old, and the distribution slightly skews right, showing a decrease in older age group representation.
Figure F : It provides a Boxplot of Pregnancies grouped by Outcome. It compares the number of pregnancies between diabetic and non-diabetic women. On average, diabetic individuals have had more pregnancies, indicating a possible relationship between this factor and diabetes.
Figure G : It illustrates the Insulin Level Distribution, which is heavily right-skewed. Many individuals have very low insulin levels, while fewer have extremely high levels, suggesting potential outliers or missing clinical data.

V. CONCLUSION

This project demonstrates a smart, efficient, and user-friendly system for diabetes prediction using machine learning. The use of Random Forest ensures accuracy, while the integration into a Streamlit app supports ease of use and live monitoring. This model can serve as a decision support system in healthcare for quicker and more accurate diagnosis.
This system provides an accurate and user-friendly solution for diabetes prediction. The combination of Random Forest with real-time interfaces offers both medical professionals and patients an efficient diagnostic tool. Future work includes integrating with health devices and enhancing model explainability.

REFERENCES

1. Kalyankar, G.D., Poojara, S.R., Dharwadkar, N.V.: Predictive analysis of diabetic patient data using machine learning and hadoop. In: International Conference On I-SMAC (2017). ISBN 978-1-5090-3243-3. Google Scholar
2. Komi, M., Li, J., Zhai, Y., Zhang, X.: Application of data mining methods in diabetes prediction. In: Image, Vision and Computing (ICIVC), 2017 2nd International Conference on, pp. 1006–1010. IEEE (2017). Google Scholar
3. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance analysis of data mining classification techniques to predict diabetes. Proced. Comput. Sci. 82, 115–121 (2016). https://doi.org/10.1016/j.procs.2016.04.016 Article Google Scholar
4. Nai-Arun, N., Sittidech, P.: Ensemble learning model for diabetes classification. Adv. Mater. Res. 931–932, 1427–1431 (2014). https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427 Article Google Scholar
5. Orabi, K.M., Kamal, Y.M., Rabah, T.M.: Early predictive system for diabetes mellitus disease. In: Industrial Conference on Data Mining, pp. 420–427. Springer (2016). Google Scholar
6. Priyam, A., Gupta, R., Rathee, A., Srivastava, S.: Comparative analysis of decision tree classification algorithms. Int. J. Current Eng. Technol. 3, 334–337, 2277–4106 (2013). arXiv:ISSN
7. Esposito, F., Malerba, D., Semeraro, G., Kay, J.: A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern Anal. Mach. Intell. 19, 476–491 (1997). https://doi.org/10.1109/34.589207 Article Google Scholar
8. Pradhan, M., Bamnote, G.R.: Design of classifier for detection of diabetes mellitus using genetic programming. Adv. Intell. Syst. Comput. 1, 7630770 (2014). https://doi.org/10.1007/978-3-319-11933-5
9. Sisodia, D., Sisodia, D.S.: Prediction of diabetes using classification algorithms. Proced. Comput. Sci. 132, 1578–1585 (2018). https://doi.org/10.1016/j.procs.2018.05.122 Article Google Scholar
10. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104–116 (2017). https://doi.org/10.1016/j.csbj.2016.12.005 Article Google Scholar
11. Ayesha, S., Hanif, M.K., Talib, R.: Machine learning based diabetes prediction system. In: 2020 International Conference on UK-China Emerging Technologies (UCET), pp. 1–4. IEEE (2020). https://doi.org/10.1109/UCET51115.2020.9205460
12. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 261–265 (1988). Google Scholar
13. Dinh, A., Miertschin, S., Young, A., Mohanty, S.D.: A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med. Inform. Decis. Mak. 19, 211 (2019). https://doi.org/10.1186/s12911-019-0918-5 Article Google Scholar
14. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., Sakr, S.: Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PLoS ONE 12(7), e0179805 (2017). https://doi.org/10.1371/journal.pone.0179805 Article Google Scholar

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details