DYNAMIC USAGE OF MACHINE LEARNING FOR EARLY DETECTION OF DIABETES

Dr. G. Nalinipriya, A.H. Abdur Rahman Basil, H. Visalan

Abstract**—Diabetica is a web-based tool designed to predict the likelihood of diabetes using machine learning algorithms. By analyzing health data such as cholesterol levels, the platform offers convenient, personalized diabetes risk assessments from home. Diabetica employs logistic regression and other advanced techniques to provide accurate predictions, empowering users with valuable health insights. Built on the Django web framework, Diabetica ensures a seamless user experience with robust features like ORM for database management, built-in security measures, and a responsive user interface. Django's authentication system secures user data, while the admin interface facilitates efficient website management. Overall, Django is instrumental in creating a reliable, secure, and user-friendly web application for diabetes prediction.***

Index terms: Machine learning, Diabetes prediction, Web-based application, Health parameters, Logistic regression, Data analysis

I. INTRODUCTION

Diabetes, a chronic metabolic disorder characterized by elevated blood sugar levels, poses serious health risks and can lead to complications such as heart disease, stroke, and kidney failure. Managing diabetes requires regular blood glucose monitoring, a healthy lifestyle, and sometimes medication. Despite advancements, effective diabetes management remains challenging, impacting individuals physically, emotionally, and financially.

Early detection is crucial for diabetes management, enabling timely lifestyle changes and treatment. Current screening methods can be invasive, time-consuming, and costly. Diabetica, a web-based tool utilizing Django and logistic regression, aims to provide an accessible platform for assessing diabetes risk based on key health parameters like cholesterol levels. Leveraging machine learning, Diabetica offers personalized predictions and insights, empowering users to take proactive steps towards better health.

Diabetica , uses the Django web framework for a seamless user experience, robust database management, and security. The tool's frontend is developed with Django's templating system, ensuring dynamic content and responsiveness. The authentication system secures user data, while the admin interface allows efficient website management.

Machine learning, especially logistic regression, has revolutionized predictive modeling by analyzing complex data sets for accurate predictions. Logistic regression is ideal for binary classification, making it valuable in healthcare for tasks like diabetes prediction. By combining machine learning and logistic regression, Diabetica offers a powerful tool for early diabetes detection and management. Traditional diabetes management relies on regular glucose monitoring, dietary control, physical activity, and medication. Diagnosis involves tests like fasting blood sugar, oral glucose tolerance, and HbA1c. Treatment includes a balanced diet, regular exercise, and medications to control blood sugar levels.

In summary, Diabetica aims to improve diabetes management through early detection and personalized insights, making a significant impact on individuals' health and quality of life.

II. LITERATURE SURVEY

G. Parimala et al, in their study used Naive Bayes Classification is employed [1]. Naive Bayes is a widely used algorithm for classification tasks, especially in text classification, due to its simplicity and efficiency. It operates based on Bayes theorem, which calculates the probability of a hypothesis given the data. Diabetes prediction models utilize features such as glucose level, body mass index (BMI), age, insulin level, and external characteristics associated with diabetes (e.g., family history, diet, exercise habits), along with other standard components. These features collectively help in assessing the risk of developing diabetes. The model's strength lies in its inclusion of critical features like glucose level, BMI, age, and insulin level, which are known to be significant in diabetes prediction, thus enhancing its accuracy. However, the model's reliance solely on these features might limit its predictive capability by potentially overlooking other relevant factors not present in the dataset.

Md. Faisal et al, in their paper "Early Prediction of Diabetes Mellitus using Machine Learning Techniques," explore the application of four popular machine learning algorithms: Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), and C4.5 decision tree [2]. Their study focuses on utilizing a dataset comprising 16 attributes related to diabetes mellitus, including factors such as age, diet, hypertension, vision problems, and genetic factors. This project is situated within the domains of healthcare, particularly diabetes management and early prediction, as well as machine learning and data mining for medical applications. The research highlights the high prediction accuracy achieved, particularly with the C4.5 decision tree algorithm, emphasizing the importance of input feature quality and relevance in such predictive modeling efforts.

Nahla Barakat et al employed a range of machine learning techniques, including algorithms, feature engineering, data processing, and model evaluation, in their study [3]. The features considered in their research encompassed age, gender, BMI, glucose level, blood pressure, skin thickness, insulin level, family history, pregnancy status, physical activity, diet, and cholesterol levels. The study's focus was on using machine learning predictive models to identify students at risk of stress early on. However, the researchers noted the potential risk of bias in predictions based on historical data, emphasizing the need for continuous monitoring and adjustments to mitigate such biases.

Roshi Saxena et al utilize supervised machine learning algorithms like logistic regression, SVM, decision trees, random forest, and ensemble methods to predict diabetes based on health-related attributes [4]. Their study involves datasets containing glucose levels, BMI, and blood pressure to determine the likelihood of diabetes. Features such as age, BMI, glucose levels, insulin levels, blood pressure, and other health-related attributes are used to train machine learning models for predicting diabetes. While these models offer early detection and personalized treatment options, challenges include acquiring and processing large, high-quality datasets for training, and the need for specialized expertise to interpret results in medical contexts, which may raise ethical and privacy concerns.

Md. Maniruzzaman et al in thier study aims to develop a machine learning (ML)-based system for predicting diabetic patients, given the rising prevalence of diabetes and its associated complications [5]. Logistic regression (LR) was used to identify significant risk factors for diabetes, with seven factors including age, education, BMI, systolic BP, diastolic BP, direct cholesterol, and total cholesterol found to be significant. Four classifiers (naïve Bayes, decision tree, Adaboost, and random forest) were then used to predict diabetic patients, with the LR-based feature selection and random forest-based classifier combination showing the highest accuracy of 94.25% and an AUC of 0.95 for the K10 protocol. The study concludes that the combination of LR and random forest-based classifiers performs well and can be useful in predicting diabetic patients.

Parastoo Rahimloo et al int thier paper introduces a method for accurate diabetes prediction by combining neural networks and logistic regression [6]. It first presents the parameters for laboratory testing and their impact on diabetes prediction. Logistic regression is used to determine the impact of each parameter on the output, leading to the proposed model for accurate prediction. The results show that the combination of neural networks and logistic regression reduces errors in predicting diabetes, allowing for early detection and necessary actions. The study concludes that this combined approach is effective in predicting diabetes.

Ms.B.Sujani Reddy et al in thier methodology involves data collection from various sources, data preprocessing to clean and normalize the data, feature selection to identify important elements, model training using logistic regression, model evaluation using metrics like accuracy and ROC-AUC score, hyperparameter tuning for optimization, validation using techniques like cross-validation, and interpretation of the model coefficients [7] . The flowchart depicts the process from entering parameter values on the website to viewing the predicted outcome for diabetes. The conclusion highlights the benefits of using logistic regression for diabetes detection, including simplicity and interpretability, while acknowledging its limitations and suggesting areas for future research to enhance its capabilities. The author acknowledges CMR College of Engineering & Technology for providing facilities and support.

III. SYSTEM AND SOFTWARE REQUIREMENTS

The system requires a modern desktop or laptop computer, with at least an Intel Core i3 processor or equivalent. It should have a minimum of 4GB of RAM, although more is recommended for handling larger datasets. Sufficient storage space, typically a few gigabytes, is necessary to store the dataset and the model. The system is compatible with major operating systems such as Windows, macOS, and Linux.

For software requirements, the system relies on Python as the core programming language. It also requires the Django framework for web application development, and HTML/CSS for designing the user interface. Additionally, libraries like Pandas, NumPy, Matplotlib, and Scikit-learn are essential for data manipulation, analysis, and visualization.

IV. PROPOSED SYSTEM ARCHITECUTE

We have gone through many research paper which handles the same problem as our research.

The proposed system, Diabetica, consists of several main components:

1. **Scrape the Dataset Using Python:** The process involves acquiring a dataset relevant to diabetes prediction, including data on cholesterol levels, glucose levels, BMI, age, and other health indicators. Python data science libraries like Pandas, NumPy, and Scikit-Learn within Google Colab are used for

developing machine learning models, facilitating dataset exploration, preprocessing, and analysis.

2. **Data Processing:** The dataset undergoes preprocessing steps before model training. This includes cleaning the data by changing all commas to periods for numeric values, encoding the target variable into binary format, retaining non-null values, and removing unnecessary columns to streamline the dataset for training.

3. **Performance Metrics:** Critical performance metrics include the root mean square error (RMSE), Mean Absolute Error (MAE), R-squared (R2), Pearson correlation coefficient, accuracy, and confusion matrix. These metrics help evaluate the model's performance in predicting stress levels accurately.

The scope of the Diabetica project is to develop a web-based application that utilizes a simple logistic regression model to predict whether an individual has diabetes. The primary goal is to provide users with an easy-to-use interface to input their health data and receive an accurate prediction of their diabetes status. The application aims to be user-friendly, with features such as data validation, result visualization, and user notification. Data privacy and security measures will be in place to protect user information. Integration with wearables, personalized recommendations, and advanced analytics are considered for future enhancements. Overall, Diabetica aims to provide users with a valuable tool for predicting their diabetes risk and empowering them to make informed decisions about their health.

4.1 METHODOLOGY

The methodology of Diabetica follows a systematic flow chart, starting with user input of health data. The data then undergoes preprocessing, where it is cleaned, encoded, and prepared for analysis. Next, the preprocessed data is fed into a logistic regression model for predicting diabetes status. Once the model makes its predictions, a report is generated summarizing the results. Finally, the output, which indicates whether an individual is diabetic or not, is presented to the user in a clear and understandable format. This structured approach ensures that the prediction process is efficient, accurate, and user-friendly.

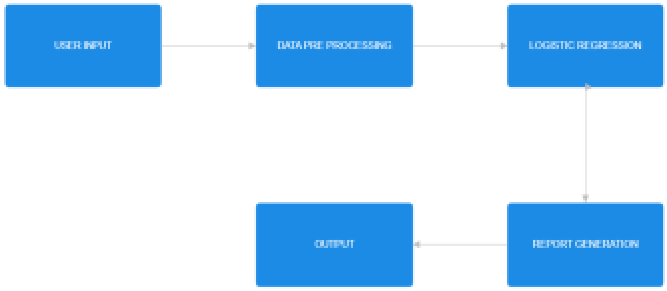


Fig.1 Flow Diagram

4.2 SYSTEM ARCHITECURE

The architecture of Diabetica is designed to be a web-based application, making it accessible to users through their browsers. The system architecture follows a client-server model, where the client, typically a web browser, interacts with the server hosting the Diabetica application.

On the client side, users interact with the application through a user-friendly interface. This interface allows users to input their health data, such as glucose levels, cholesterol levels, BMI, and other relevant indicators. The input data is then sent to the server for processing. The server side of the architecture is responsible for handling the data processing and prediction tasks. Upon receiving the user's input, the server preprocesses the data, which includes cleaning, encoding, and preparing it for analysis. The preprocessed data is then fed into a logistic regression model, which predicts the likelihood of the user having diabetes based on the input features.

Once the prediction is made, the server generates a report summarizing the results. This report includes the predicted diabetes status, along with any additional insights or recommendations based on the input data. Finally, the server sends the report back to the client, where it is presented to the user in a clear and understandable format.

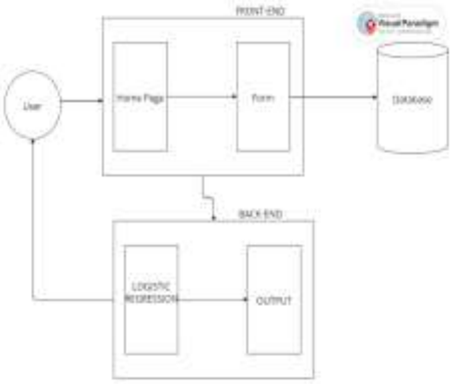


Fig.2 Architecure Diagram

4.3 MODULE DESCRIPTION

The Diabetica project consists of several main components, including data scraping using Python, data processing, training and evaluation of ML models. In the data scraping module, datasets relevant to diabetes prediction are acquired from external sources, encompassing various health indicators such as cholesterol levels, glucose levels, BMI, and age. Python libraries like Pandas, NumPy, and Scikit-Learn are leveraged within Google Colab to explore, preprocess, and analyze these datasets efficiently.

Subsequently, the data processing module involves cleaning the datasets by standardizing numeric values and encoding categorical variables. The target variable is prepared for binary classification, and the dataset is streamlined for model training.

The training and evaluation module focuses on loading the dataset into a Pandas DataFrame for manipulation and selecting relevant input features for modeling. Imputation techniques are applied to address missing values, and categorical data is encoded for machine learning. The DataFrame is then split into training and testing sets using a random train-test split stratified by target classes, ensuring a balanced distribution for model training and evaluation. This comprehensive approach enables the creation and assessment of data-driven models for predicting diabetes effectively.

V. IMPLEMENTATION

In the implementation of the Diabetica project, the Django development server is hosted to serve the web application. Upon accessing the server, users are directed to the home page where they are presented with a "Get Started" button. Clicking on this button navigates users to a new page where they are presented with two options:

1. Demo Button: Clicking on the "Demo" button triggers the application to randomly select a set of values from the Diabetes.csv dataset. The selected values are then used as input for the logistic regression model. The model predicts whether the individual has diabetes or not based on the input values, and the output is displayed to the user.

2. User Input Button: Clicking on the "User Input" button allows the user to input their own values for cholesterol level, glucose level, HDL cholesterol level, age, gender, height, weight, systolic blood pressure, diastolic blood pressure, waist circumference, and hip circumference. Once the user inputs these values and submits, the application performs logistic regression on the user-provided data to predict whether the individual has diabetes or not. The prediction output is then displayed to the user.

This implementation allows users to interact with the Diabetica web application, providing them with the ability to explore diabetes prediction using logistic regression both with demo data and their own input values.

5.1 PERFORMANCE MEASURES

In the Diabetica project, the Logistic Regression (LR) model is trained and evaluated using the dataset to predict diabetes. Logistic Regression is a linear model commonly used for binary classification tasks, making it suitable for predicting whether an individual has diabetes or not based on the input features.

The performance of the Logistic Regression model is assessed using the classification_report function from the sklearn.metrics module. This function provides metrics such as precision, recall, F1-score, and support for each class (diabetes and no diabetes), helping to evaluate the model's accuracy and effectiveness in classifying individuals.

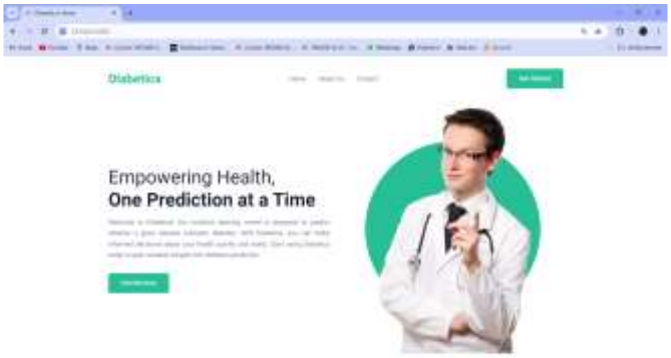
The LR model is trained using the input features and target variable from the dataset. The dataset is preprocessed to handle missing values, encode categorical variables, and prepare the data for training. After training, the model is tested using a separate test set to evaluate its performance on unseen data.The output of the LR model is used to predict whether an individual has diabetes or not. This prediction can be used to provide valuable insights for healthcare professionals and individuals to make informed decisions about diabetes management and prevention.

VI. RESULT AND DISCUSSION

In the project "Diabetica: Dynamic Usage of Machine Learning for Early Detection of Diabetes," we aimed to develop a robust predictive model for early detection of diabetes using machine learning techniques. The project's results provide valuable insights into diabetes prediction and proactive healthcare management. Leveraging machine learning methodologies, our project successfully developed a predictive model capable of categorizing individuals' likelihood of having diabetes. Through meticulous data collection and preprocessing, including tasks such as data cleaning, normalization, and feature extraction, our model was trained and evaluated on a comprehensive dataset to ensure its reliability and effectiveness in predicting diabetes. Evaluation metrics played a crucial role in assessing the performance of our diabetes prediction model. Metrics such as accuracy, precision, recall, and F1-score provided quantitative measures of the model's effectiveness in predicting diabetes. Analyzing the model's performance across these metrics helped us understand its strengths and limitations, enabling us to make informed decisions and refine the model further. Furthermore, feature importance analysis helped us identify the most influential variables in predicting diabetes. This analysis provided valuable insights into the key predictors or indicators of diabetes, enhancing the interpretability of our model and informing personalized healthcare strategies. The visualization of predicted diabetes outcomes further enriched our project's outcomes, providing stakeholders with intuitive insights into diabetes trends and patterns. Visual representations of diabetes predictions over time or across different demographic groups helped us identify critical trends and anomalies, guiding proactive healthcare interventions. The user interface and deployment aspect of our project ensured accessibility and usability for end-users. A user-friendly interface facilitated diabetes prediction and visualization of results, empowering individuals to monitor and manage their health effectively. Through validation testing and user feedback, we iteratively refined the user interface and model functionalities, ensuring alignment with user needs and preferences.

HOME PAGE:

The home page is designed using html and bootstrap, it has a "Get started" button which navigates to the next page. It also has links to contact and read the descrption.


Diabetica

Empowering Health,
One Prediction at a Time

Fig 3. Home Page of Diabetica

INPUT FOR CHOICES:

The main purpose of this page is to switch modes between a demo and the user input. On clicking the demo, the app will automatically choose a value from the dataset and demonstrate a sample of this project. On choosing the user input mode, a form will appear and user can input their own values and get the prediction for those data's.

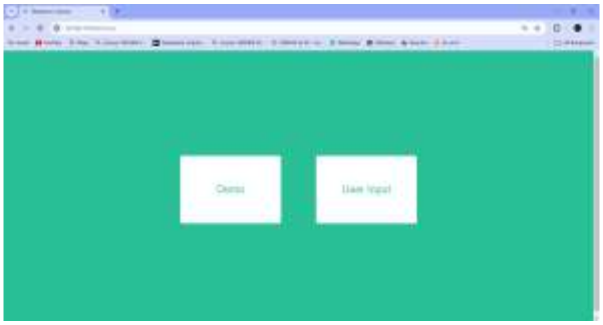
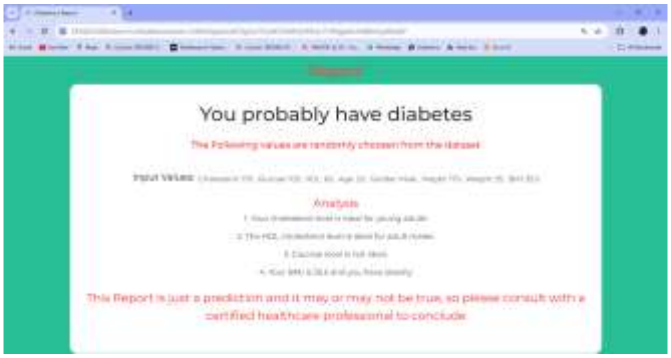


Fig 4. input for choices

REPORT PAGES:

The report page displays the report whether the patient has diabetes or not, apart from that it shows additional results of what levels are not ideal.


You probably have diabetes

The Following values are randomly chosen from the dataset

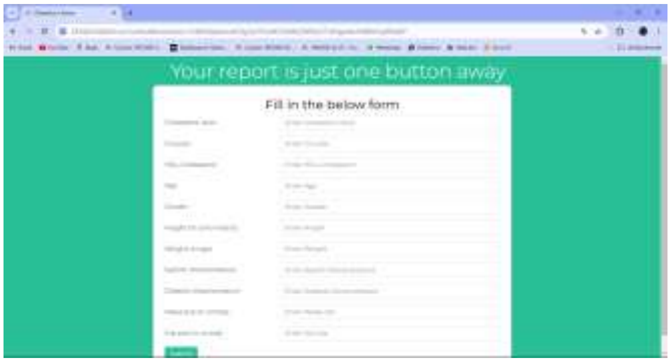
Input Values: [illegible]

Analysis

1. Your cholesterol level is ideal for young adults
2. The HDL cholesterol level is ideal for adult males
3. Glucose level is not ideal.
4. Your BMI is [illegible] and you have obesity

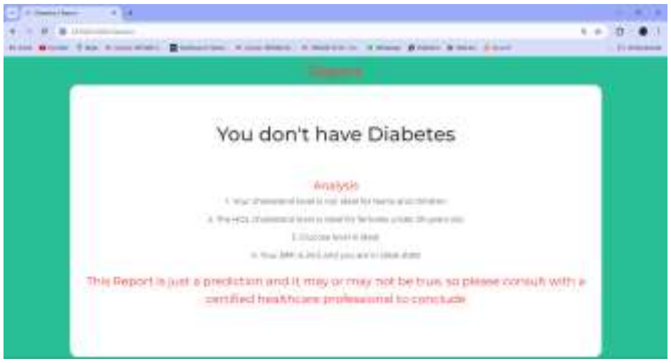
This Report is just a prediction and it may or may not be true, so please consult with a certified healthcare professional to conclude

Fig 5. Report of Demo page


Your report is just one button away

Fill in the below form

Fig 6. User Details Input Form


You don't have Diabetes

Analysis

1. Your cholesterol level is not ideal for teens and children
2. The HDL cholesterol level is ideal for females under 29 years old
3. Glucose level is ideal
4. Your BMI is [illegible] and you are in ideal state

This Report is just a prediction and it may or may not be true, so please consult with a certified healthcare professional to conclude

Fig 7. Report Page of User Input

The Fig 8 illustrates a comparison of three algorithms—Logistic Regression, SVM (Support Vector Machine), and Deep Learning—in terms of their performance. Each algorithm's performance is represented on a scale from 0 to 100, with Logistic Regression and SVM both achieving performance scores around 85, indicating strong predictive capabilities. Deep Learning, while still effective, shows a slightly lower performance score of around 80. This visual comparison highlights that while Deep Learning is powerful, traditional algorithms like Logistic Regression and SVM can sometimes offer competitive or even superior performance in specific contexts.

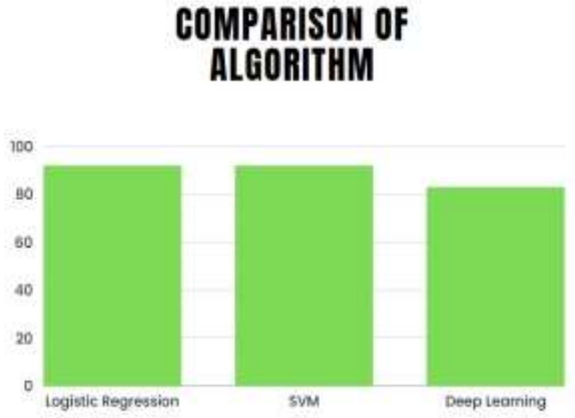


Fig 8. Comparison of various algorithms

VII. CONCLUSION

In the project "Diabetica", a web-based diabetes prediction model using logistic regression is developed. This model aims to predict the likelihood of an individual having diabetes based on various input features. The project involves several key components, including data preprocessing, model development, web application development, and deployment. The first step in the project is data preprocessing, which involves cleaning and preparing the dataset for training the model. This includes handling missing values, converting categorical variables into numerical representations, and scaling the features to ensure uniformity.

Next, a logistic regression model is trained using the preprocessed dataset. Logistic regression is chosen for its ability to predict binary outcomes, making it suitable for predicting diabetes (either the individual has diabetes or not).

Once the model is trained, a web application is developed to provide a user-friendly interface for users to input their data and obtain a prediction. The web application is designed to be intuitive and easy to use, allowing users to input relevant features such as glucose levels, BMI, age, and other health metrics. Upon submitting their data, the web application uses the trained logistic regression model to predict the likelihood of the user having diabetes. The prediction is then displayed to the user along with any additional insights or recommendations.

Finally, the web application is deployed to a server to make it accessible to users over the internet. This ensures that individuals can easily access the diabetes prediction tool and make informed decisions about their health. Overall, the Diabetica project demonstrates the use of logistic regression and web technologies to develop a practical and accessible tool for diabetes prediction, potentially aiding in early detection and management of diabetes.

REFERENCES

[1] G Parimala, R Kayalvizhi and S Nithiya, "Diabetes prediction using machine learning", In: International Conference on Computer Communication and informatics (ICCCI), 2023.

[2] Md. Faisal Farque, Asaduzzaman, Iqbal H. Sarker, "Machine learning techniques to predict diabetes mellitus". In: International Conference on Electrical, Computer and communication Engineering (ECCE) , 2019.

[3] Nahla Barakat, Andrew P. Bradely, Mohammed Nabil H barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus". In: IEEE Transactions on Information Technology in Biomedicine, 2012.

[4] Pragya Paudyal PhD, "Diabetes prediction using logistic regression in python". In: Medium.com, 2020.

[5] Zarinabegam Mundargi, Mayur Dabade, Yash Chindhe, Savani Bondre, Anannya Chaudhary, "Diabetes prediction using Logistic Regression". In: Renewable Energy, Green Computing and Sustainable Development, 2024.

[6] Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta and G. C. Sampada, "A Comprehensive review of various diabetic prediction models", In: Journal of Health care Engineering, 2022.

[7] Anushka Awasthi, Ishwar Gangwal, Mihir Jain, "Diabetes Prediction using Machine Learning: A Review". In: IJRASET Journal for Research in Applied Science and Engineering Technology, 2022.

[8] Turki Alghamdi, "Prediction of Diabetes Complications Using Computational Intelligence Techniques". In: Computation and Complex Data Processing Systems, 2023.
[9] Mohammed Mustafa, "Diabetes prediction dataset", In: Kaggle,2023.

[10] M. Ramakrishna Murthy, P. Ruchita, Ch. Bharat Teja, M. Manoj Kumar, T V S Lingeswararao. "Diabetes prediction using machine learning algorithms". In: International Research Journal of Modernization in Engineering Technology and Science ,2022.