IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction Using Machine Learning and Chat-Bot

[1]Mr. Arun P., [2]Yunish Sapkota, [3]Muthineni Shashank, [4]Kolli Venkata Abhishek

[1]Assistant Professor, [2]Student, [3]Student, [4]Student

[1]Computer Science and Engineering,

[1]Cambridge Institute of Technology, Bengaluru , India

Abstract: Diabetes, a widespread chronic condition globally, has spurred interest in using machine learning for prognosis and management. This effort aims to develop a predictive framework using machine learning tools, analyzing various data sources to identify at-risk individuals. By employing XGBoost and other machine learning algorithms, it processes large datasets efficiently. Additionally, it integrates a diabetes forecasting chatbot, offering personalized guidance based on individual risk factors. This system shows promise in enhancing diabetes management and health outcomes by accurately identifying high-risk individuals and providing tailored support. Overall, it underscores the potential of machine learning and chatbot technologies in improving chronic disease management.

Index Terms **– Diabetes, Prediction, Machine Learning, Chat Bot.**

I. INTRODUCTION

The healthcare sector houses extensive databases containing structured, semi-structured, and unstructured data. Utilizing big data analytics, these vast datasets are analyzed to uncover hidden information and patterns, extracting valuable insights. In developing nations such as India, Diabetes Mellitus (DM) has emerged as a significant health concern, classified as a Non-Communicable Disease (NCD) affecting a large population. Statistics from 2017 show that approximately 425 million people are affected by diabetes, with an estimated 2-5 million deaths attributed to the condition annually. Projections suggest a rise to 629 million by 2045.

Diabetes Mellitus (DM) comprises various types, including Type-1, also known as Insulin-Dependent Diabetes Mellitus (IDDM), requiring insulin injections due to insufficient insulin production. Type-2, or Non-Insulin-Dependent Diabetes Mellitus (NIDDM), occurs when cells struggle to utilize insulin effectively. Gestational Diabetes, classified as Type-3, develops during pregnancy due to elevated blood sugar levels. DM leads to long-term complications and significantly increases health risks for those affected.

Predictive analysis integrates machine learning algorithms, data mining methods, and statistical techniques to extract insights and forecast future occurrences based on historical and current data. Its application in healthcare facilitates informed decision-making and anticipates forthcoming events. Utilizing predictive analysis in healthcare involves machine learning and

regression methods, aiming to accurately diagnose diseases, improve patient care, optimize resources, and enhance clinical outcomes. Machine learning plays a crucial role in this process, allowing computer systems to learn from previous experiences without explicit programming for each instance, thus reducing human intervention and enabling automated processes with minimal errors. Presently, conventional diabetes detection relies on laboratory tests such as fasting blood glucose and oral glucose tolerance, which, while effective, are time-consuming.

II. LITERATURE SURVEY

Numerous endeavors have been dedicated to developing algorithms and programs aimed at early diabetes detection. One such effort is detailed in "Diabetes Prediction using Machine Learning Algorithms" [1], which underscores the gravity of diabetes mellitus as a significant health concern influenced by factors like age, obesity, lifestyle choices, and genetic predisposition. The study emphasizes the heightened risk of complications such as heart disease, stroke, and nerve damage associated with diabetes. While conventional hospital practices rely on diagnostic tests followed by tailored treatments, the potential of Big Data Analytics in healthcare lies in offering predictive insights from extensive datasets. However, existing classification and prediction methods often fall short in terms of accuracy. This literature review introduces a novel diabetes prediction model that incorporates additional factors, thereby enhancing classification accuracy. Furthermore, a pipeline model is proposed to further refine the diabetes prediction process.

Similarly, "Diabetes Prediction using Machine Learning Techniques" [2] underscores the significant health risks posed by diabetes, including heart-related issues, kidney problems, hypertension, and eye damage. Early detection is deemed pivotal for effective management. This project endeavors to predict diabetes onset employing various Machine Learning Techniques, leveraging patient data. Techniques such as K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF) are harnessed for classification, with Random Forest demonstrating superior predictive capability. The study accentuates the effectiveness of machine learning in diabetes prediction, with Random Forest emerging as the most accurate model among those evaluated.

Moreover, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers" [3] presents a framework for diabetes prediction that integrates outlier rejection, missing value imputation, data standardization, feature selection, and ensembling of various machine learning classifiers, including k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP). The introduction of weighted ensembling based on Area Under ROC Curve (AUC), optimized through grid search during hyperparameter tuning, is highlighted. Experiments conducted on the Pima Indian Diabetes Dataset demonstrate superior performance, with the proposed ensembling classifier achieving notable metrics. The framework outperforms state-of-the-art methods, underscoring its efficacy in diabetes prediction. The public availability of the source code ensures reproducibility and facilitates further research.

Furthermore, "Prediction of Diabetes Empowered with Fused Machine Learning" [4] emphasizes the criticality of early disease prediction in the medical field, especially for diseases like diabetes with significant global health risks. With modern dietary habits heightening risk factors, symptom understanding becomes paramount for prediction. Machine learning (ML) algorithms emerge as valuable tools for disease detection. This article introduces a fused machine learning approach for diabetes prediction, incorporating Support Vector Machine (SVM) and Artificial Neural Network (ANN) models to analyze datasets and determine diabetes diagnosis. The dataset is partitioned into training and testing data, and the output from these models serves as input for a fuzzy logic system, ultimately determining diabetes diagnosis. The proposed fused ML model achieves commendable prediction accuracy, surpassing previous methods.

III. DATASET

The dataset utilized in this project is the Pima Indian Diabetes Dataset, a well-known resource in the fields of machine learning and data science. It comprises information about 768 women of Pima Indian descent aged 21 or older, residing near Phoenix, Arizona, USA. Collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in the late 1970s and early 1980s, the dataset consists of 8 attributes, including pregnancy frequency, age, body mass index (BMI), blood pressure, and glucose concentration, among others. The target variable of the dataset is binary, indicating whether the woman developed diabetes within five years of the initial examination. This dataset holds significant importance in research and education, particularly in the development and evaluation of machine learning algorithms for diabetes prediction. Moreover, it has played a crucial role in studying the relationships between various risk factors and the onset of diabetes.

In recent years, there has been a notable effort to expand the dataset's scope and gather more comprehensive health-related data concerning Pima Indians and other indigenous communities. The primary objective is to enhance understanding of the complex factors influencing diabetes and other health conditions within these populations. By doing so, researchers aim to develop more effective interventions for prevention and treatment tailored to the specific needs of these communities. This expansion seeks to provide deeper insights into health dynamics, enabling targeted and efficient healthcare interventions for indigenous populations.

IV. METHODOLOGY

4.1 User Interface (UI):

The user interface (UI) offers an interactive platform for users to engage with the chat-bot by prompting input. Moreover, users have the option to bypass the chat-bot and utilize a separate interface equipped with sliders and buttons for selecting the desired input and visualizing data. The UI or front end is developed in Python, leveraging Streamlit for creating a user-friendly web application experience.

4.2 Preprocessing & Feature Extraction:

During this crucial step, the dataset undergoes a meticulous process of division into training and testing subsets, ensuring that the predictive model is robust and reliable. Careful consideration is given to feature selection, where attributes deemed pertinent to the prediction task are identified and incorporated into the analysis. This entails a comprehensive evaluation of each attribute's relevance and contribution to the predictive accuracy.

In the training phase, the algorithm is primed to discern patterns and relationships within the data, with the objective of accurately predicting the likelihood of diabetes onset. To achieve this, all attributes of the dataset, except for the outcome variable, are harnessed to furnish the algorithm with ample information for learning and inference.

Simultaneously, the dataset's attributes are scrutinized from another perspective to facilitate personalized suggestions for individuals. Attributes that are immutable or cannot be readily altered, such as age, heredity, and pregnancy status, serve as anchors for identifying analogous cases without diabetes. By leveraging these fixed attributes, the algorithm seeks to uncover similar instances within the dataset and extrapolate personalized recommendations tailored to individual circumstances.

In essence, this preparatory phase lays the groundwork for both predictive modeling and personalized intervention strategies, ensuring that the algorithm is equipped with the requisite data and insights to make informed predictions and offer relevant suggestions for mitigating the risk of diabetes.

4.3 Prediction Module:

In the diabetes prediction system, a diverse array of machine learning algorithms is employed, including XG-Boost, Random Forest, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree. This variety of algorithms offers users the flexibility to select the most suitable one for their specific input scenario. Each algorithm has its unique strengths and weaknesses, and by providing users with multiple options, the system ensures that they can choose the most effective one based on their data characteristics and requirements.

During the testing and training phases, each algorithm is rigorously evaluated to assess its performance and accuracy in predicting diabetes. This evaluation process yields an accuracy score, which quantifies the algorithm's ability to make correct predictions. By providing users with these accuracy scores, the system empowers them to make informed decisions about which algorithm to use, enhancing the effectiveness and utility of the prediction process.

Overall, the inclusion of multiple machine learning algorithms and the evaluation of their performance contribute to the robustness and reliability of the diabetes prediction system, enabling users to leverage the most effective algorithm for their specific needs and circumstances

4.4 Suggestion Module:

The suggestion module is an integral part of the system, activated when diabetes is detected in a user. It generates personalized recommendations by identifying similar entries in the dataset where the outcome is negative (indicating no diabetes) and then computing the differences in attributes between the user's input and these similar entries. Machine learning algorithms such as K-Nearest Neighbors (KNN) are employed for this task, considering attributes like Age, Diabetes Pedigree Function (DPF), Pregnancy, and others.

To illustrate, let's consider a scenario:

User Input:-
- Pregnancies = 2
- Glucose = 90
- BloodPressure = 60
- SkinThickness = 25
- Insulin = 94
- BMI = 35
- DiabetesPedigreeFunction = 0.18
- Age = 24

Similar Entry from Dataset:
- Pregnancies = 1
- Glucose = 71
- BloodPressure = 66
- SkinThickness = 23
- Insulin = 94
- BMI = 28.1
- DiabetesPedigreeFunction = 0.167
- Age = 21

Calculating the differences between the dataset entry and the user's input:
- BMI difference = -6.9

- Glucose difference = -19
- BloodPressure difference = 6

Based on these differences, the user would be suggested to decrease their weight and lower their glucose level. However, it's essential to emphasize that these suggestions are not a substitute for medical diagnosis. Users are strongly recommended to seek professional medical examination for accurate diagnosis and appropriate treatment. By leveraging machine learning techniques and comparing user input with similar dataset entries, the system provides tailored suggestions aimed at promoting healthier lifestyle choices and managing diabetes risk factors.

4.5 Chat-Bot Module:

The chat-bot operates on the foundation of OpenAI's Chat-GPT 3.5 turbo, leveraging its advanced language processing capabilities to interact with users. This module seamlessly integrates with the machine learning algorithms responsible for prediction and suggestion generation. By combining the power of these algorithms with the natural language generation abilities of Chat-GPT, the chat-bot delivers intuitive and contextually relevant responses based on the results obtained from the ML algorithms.

One notable aspect is the separation between the language model and the prediction/suggestion module. This design allows for flexibility and scalability, as updates or upgrades can be applied independently to each component. For instance, if there are advancements in language processing technology or newer versions of GPT become available, the chat-bot can be easily upgraded to incorporate these improvements without disrupting the underlying prediction and suggestion functionalities.

Furthermore, this modular architecture enables the possibility of transitioning to entirely new language models or incorporating different AI models in the future. This adaptability ensures that the chat-bot remains agile and capable of evolving alongside advancements in AI technology, ultimately enhancing the user experience and performance of the system.

V. RESULT

The program was executed on a laptop featuring robust hardware specifications, including a quad-core Intel i5-10300H CPU, 16 GB of dual-channel DDR4 system memory, and a 500 GB NVME storage drive. These specifications provide ample computing power and memory capacity to efficiently handle the dataset and execute the machine learning algorithms. To ensurer reliable performance evaluation, the dataset was meticulously divided into training and testing subsets, maintaining a balanced ratio of 1:4 between them. This approach enables thorough testing of the algorithms' predictive capabilities while minimizing the risk of overfitting or bias.

Throughout the study, we conducted numerous tests using various input scenarios to comprehensively assess the performance of each algorithm. These tests spanned a diverse range of inputs, capturing different demographic profiles associated with diabetes. By systematically varying the inputs, we aimed to evaluate how effectively each algorithm could adapt and generalize across a spectrum of real-world scenarios.

Each algorithm underwent rigorous testing, with multiple iterations conducted to ensure robustness and reliability of the results. The accuracy of each algorithm was meticulously recorded for each test scenario, and the average accuracy across all tests was calculated to provide a comprehensive overview of their performance. The culmination of these efforts is summarized in the table below, which showcases the accuracy of the evaluated algorithms. These accuracy scores serve as valuable metrics for comparing the performance of each algorithm and guiding informed decision-making in algorithm selection for our diabetes prediction system.

Table 5.1: Accuracy Comparison

Algorithm	Accuracy (%)
XB-Boost	82
Random Forest	84
Naive Bayes	75
Logistic Regression	80
SVM	74
KNN	72
Decision Tree	64

Among the evaluated algorithms, the Random Forest classifier emerged as the most accurate, boasting an impressive average accuracy of 84%. On the other end of the spectrum, the Decision Tree algorithm exhibited the lowest accuracy, achieving a score of 64%. Interestingly, in addition to its high accuracy, the Random Forest algorithm also demonstrated remarkable efficiency during the training phase, along with the XG-Boost algorithm. This efficiency suggests that these algorithms not only excel in predictive performance but also in computational speed, making them favorable choices for real-time or resource-constrained environments.

These findings highlight the importance of not only accuracy but also computational efficiency when selecting an algorithm for deployment in practical applications. By prioritizing both aspects, we can ensure not only reliable predictions but also swift processing, enhancing the overall effectiveness and usability of our diabetes prediction system.

VI. CONCLUSION

In our research paper, we conducted an extensive evaluation of seven machine learning algorithms, namely K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree, and XGBoost, for predicting diabetes using our model. Our analysis revealed that XGBoost consistently outperformed the other algorithms in terms of prediction accuracy for diabetes.

Looking ahead, our future scope involves training the algorithms on even larger datasets, a task that necessitates leveraging cloud platforms for computing resources. By harnessing the scalability and flexibility offered by cloud infrastructure, we aim to further enhance the predictive capabilities of our model and accommodate the growing demands of healthcare data analysis.

Moreover, our experimental results underscore the effectiveness of our Diabetes ChatBot in both accurately comprehending users' queries and facilitating their understanding of the predictions made by ML models. This highlights the potential of our Chatbot as a valuable tool for domain experts, such as healthcare workers, in interpreting and utilizing ML models effectively, particularly in the context of disease diagnosis.

In the future, it would be beneficial to explore the real-world applications of our Diabetes ChatBot in settings such as doctors' offices, laboratories, or professional environments. Understanding how stakeholders interact with the system in practical settings can provide valuable insights into its usability and impact, guiding further enhancements and refinements.

Additionally, we aim to investigate the integration of our program into existing scientific and professional workflows to maximize its utility and promote widespread adoption. By seamlessly integrating with established work streams, our Diabetes Chatbot has the potential to streamline processes and empower stakeholders to make informed decisions based on ML predictions effectively.

VII. ACKNOWLEDGEMENT

We extend our heartfelt gratitude to Mr. Arun P, Assistant Professor in the Department of Computer Science and Engineering at CITech, for his invaluable guidance and technical insights that greatly contributed to the success of our project. We also appreciate the support and assistance of our friends, teachers, and parents, whose contributions enriched our work and motivated us throughout this journey.

REFERENCES

[1] Aishwarya Mujumdar, V Vaidehi Dr., "Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science Volume 165, 2019, Pages 292-299.

[2] Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learning Techniques" International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181 Vol. 9 Issue 09, September-2020.

[3] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[4] U. Ahmed et al., "Prediction of Diabetes Empowered with Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.