DIABETES PREDICTION USING MACHINE LEARNING

SUBMITTED TO THE PCET'S PIMPRI CHINCHWAD COLLEGE OF ENGINEERING AN
AUTONOMOUS INSTITUTE, PUNE
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THIRD YEAR
OF

BACHELOR OF TECHNOLOGY (COMPUTER ENGINEERING)

SUBMITTED BY

VIJAY CHAURE	**PRN No: 121B1B035**
YASH CHINCHOLE	**PRN No: 121B1B037**

DEPARTMENT OF COMPUTER ENGINEERING
PCET'S PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

Sector No. 26, Pradhikaran, Nigdi, Pimpri-Chinchwad, PUNE 411044
An Autonomous Institute Approved by AICTE and Affiliated to SPPU, Pune

CERTIFICATE

This is to certify that the mini project report entitles

DIABETES PREDICTION USING MACHINE LEARNING

Submitted by

VIJAY CHAURE	**PRN No: 121B1B035**
YASH CHINCHOLE	**PRN No: 121B1B037**

are bonafide student of this institute and the work has been carried out by them under the supervision of **Prof. Sushama Vispute & Dr. Swati Shinde** and it is approved for the partial fulfillment of the requirement of PCET'SPIMPRI CHINCHWAD COLLEGE OF ENGINEERING, for the award of **Third Year of Bachelor of Technology** (Computer Engineering) **A.Y.-2023-24**.

(Prof. Sushama Vispute)
(Dr. Swati Shinde)
Guide
Department of Computer Engineering

(Prof. Dr. K. Rajeswari)
Head,
Department of Computer Engineering

Place : Pune
Date :

ABSTRACT

Diabetes is a prevalent chronic disease with significant health implications, making early prediction and management crucial for individuals' well-being. Leveraging machine learning methodologies, this study proposes a comprehensive framework for predicting the likelihood of diabetes development through an online platform. The framework integrates diverse machine learning techniques to analyze various factors associated with diabetes risk, such as demographic information, medical history, lifestyle choices, and genetic predispositions. By scrutinizing these factors, the system can identify patterns indicative of potential diabetes onset. Additionally, natural language processing algorithms are employed to interpret textual data related to health records and lifestyle habits, further enhancing prediction accuracy. Real-time monitoring of user interactions and input data facilitates the detection of anomalous behavior patterns, which may signal an increased risk of diabetes or non-compliance with recommended health guidelines. This proactive approach enables timely intervention and personalized recommendations to mitigate diabetes risk factors. Furthermore, the framework incorporates external data sources, such as medical research databases and public health records, to enrich the predictive model and adapt to evolving trends in diabetes risk factors. Evaluation against comprehensive datasets comprising both diabetic and non-diabetic cases demonstrates superior prediction performance compared to conventional methods. By harnessing machine learning and real-time monitoring capabilities, the proposed platform offers a robust solution for early diabetes prediction, empowering individuals to take proactive measures towards better health outcomes and reducing the burden on healthcare systems.

Introduction

In today's digitally interconnected landscape, where online platforms serve as primary channels for information exchange, commerce, and interaction, the proliferation of phishing attacks presents an ongoing and dynamic threat to cybersecurity. Phishing, a form of social engineering, exploits human psychology to deceive individuals into divulging sensitive information such as passwords, usernames, and financial credentials. Despite advancements in security measures and awareness campaigns, phishing remains a prevalent tactic employed by malicious actors due to its efficacy in circumventing traditional safeguards. The repercussions of phishing attacks are severe, encompassing identity theft, financial losses, damage to reputation, and compromised network security. Consequently, the detection and prevention of phishing attacks have become paramount concerns for individuals, businesses, and cybersecurity experts alike. Conventional phishing detection techniques primarily rely on static analysis of website elements such as URL structures and domain reputations to flag potential threats. However, these static detection methods face significant challenges in addressing the dynamic nature of phishing tactics, characterized by rapid adaptation and evasion strategies. The emergence of sophisticated phishing techniques like spear phishing and whaling, targeting specific individuals or high-value targets within organizations, further underscores the need for more robust and adaptable detection systems. In response, researchers and cybersecurity professionals have increasingly turned to machine learning and artificial intelligence (AI) approaches to enhance phishing detection capabilities. By leveraging vast datasets and advanced algorithms, machine learning models can discern patterns and anomalies indicative of phishing behavior, enabling expedited and precise identification of fraudulent websites and emails. Moreover, the integration of real-time monitoring and behavioral analysis enables the detection of subtle deviations from typical user interactions, thereby enhancing detection accuracy while minimizing false positives.

Additionally, organizations can bolster their defenses against phishing attacks by leveraging collective intelligence through threat intelligence feeds and collaborative sharing of phishing indicators. However, the ever-evolving tactics employed by adversaries underscore the ongoing need for a multifaceted approach to phishing detection, encompassing technological advancements alongside user education and awareness initiatives. This research endeavors to address these challenges by presenting a comprehensive framework for phishing detection, integrating behavioral analysis, threat intelligence, machine learning algorithms, and user education. Empirical evaluation and practical implementation demonstrate the efficacy of the proposed framework in mitigating the risks associated with phishing attacks, safeguarding digital assets, and protecting personal data.

Methodology

Our approach to building a Diabetes prediction website centered around the utilization of the Random Forest classifier, a versatile machine learning algorithm well-suited for classification tasks. The methodology encompassed distinct phases including data collection and preprocessing, feature extraction, model training, and evaluation. To initiate the methodology, a comprehensive dataset comprising various features such as Number of Pregnancies, Insulin Level, Age, and BMI was gathered from the Kaggle platform. The dataset, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, was meticulously selected to ensure its relevance and representativeness for predicting diabetes occurrence. Notably, all patients in the dataset are females at least 21 years old of Pima Indian heritage. Preprocessing of the collected data was undertaken to ensure uniformity and suitability for subsequent analysis. This involved procedures such as data cleaning, normalization, and feature scaling to enhance data quality and mitigate biases that could potentially influence model performance. Subsequently, relevant features were extracted from the preprocessed data to serve as discriminative factors in distinguishing individuals with diabetes from those without. These features encompassed a diverse range of attributes crucial for diabetes prediction. The model training phase entailed the application of the Random Forest algorithm, renowned for its ability to handle complex datasets and mitigate overfitting. Cross-validation techniques were employed to optimize model performance and fine-tune hyperparameters such as the number of trees in the forest. Evaluation of the trained model was conducted using a suite of metrics including accuracy, precision, recall, and F1-score to assess its predictive efficacy. To enhance the robustness of the evaluation process and minimize biases stemming from data partitioning, techniques like k-fold cross-validation were employed. Furthermore, to validate the effectiveness of the Random Forest methodology in diabetes prediction, comparative analyses were performed against baseline approaches and alternative machine learning algorithms. In summation, our methodology leveraged the Random Forest classifier as the cornerstone of a comprehensive framework for diabetes prediction, demonstrating its efficacy in accurately identifying individuals at risk of diabetes while minimizing false positives.

Literature Survey

Survey on clinical prediction models for diabetes prediction

Dibetes Detection

N. Jayanthi, B. Vijaya Babu & N. Sambasiva Rao

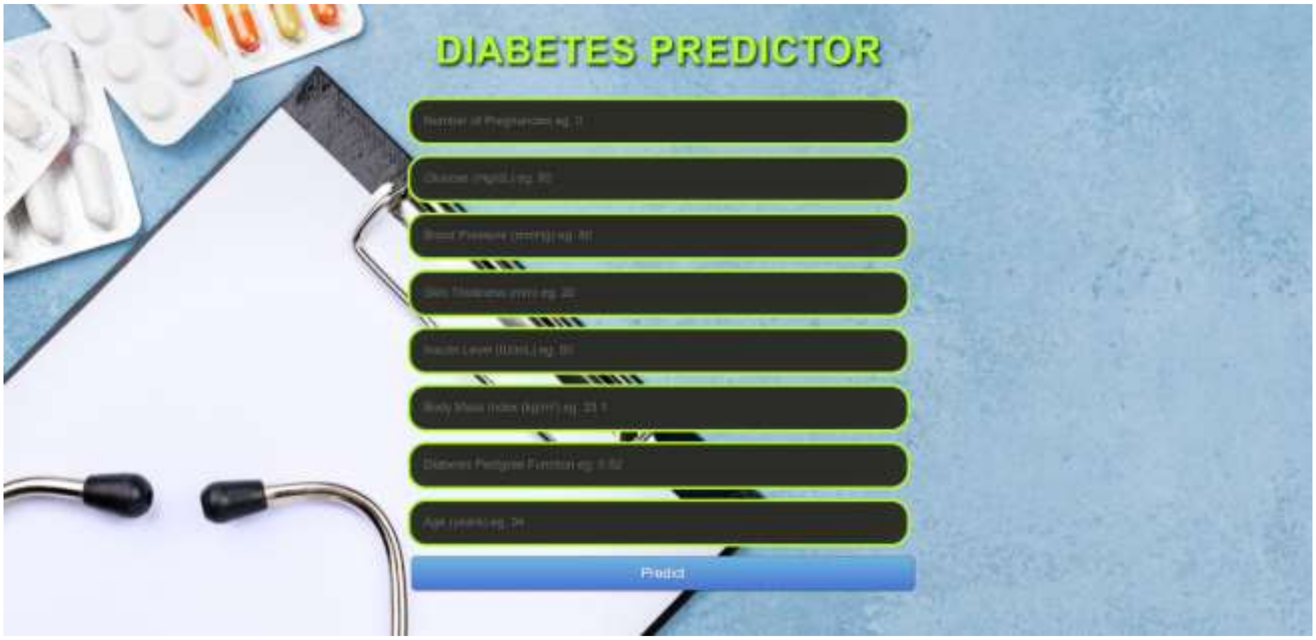
The escalating prevalence of diabetes and its associated health burdens necessitates effective predictive measures to facilitate early intervention and management. Traditional diagnostic approaches often fall short in preemptively identifying individuals at risk, prompting a shift towards machine learning-based solutions as a promising avenue for diabetes prediction. In the realm of diabetes prediction, machine learning algorithms offer a robust framework for discerning intricate patterns within diverse datasets, enabling the identification of subtle indicators predictive of diabetes onset. These solutions typically involve training classification models on comprehensive datasets comprising demographic information, medical history, lifestyle factors, and genetic predispositions. Key to the success of machine learning-based diabetes prediction is the meticulous curation of features, as a well-selected feature set significantly enhances model accuracy. Feature selection methodologies encompass a wide array of attributes, including but not limited to, number of pregnancies, insulin levels, age, and body mass index (BMI), each contributing valuable insights into diabetes risk. The literature underscores the transformative potential of machine learning in augmenting diabetes prediction capabilities, transcending the limitations of traditional diagnostic methods. By elucidating the essential components of machine learning-based solutions, including data collection, feature extraction, model training, and evaluation, the survey provides a roadmap for researchers and practitioners navigating the complex landscape of diabetes prediction. However, challenges persist, particularly in the realm of real-time prediction. While machine learning algorithms demonstrate commendable predictive prowess, their computational demands and latency constraints pose challenges in achieving real-time responsiveness. Overcoming these hurdles requires innovative algorithmic optimizations and deployment architectures tailored to meet the exigencies of real-time prediction without compromising accuracy. In summary, the literature underscores the promise of machine learning in revolutionizing diabetes prediction, offering a data-driven approach to preemptively identify individuals at risk and facilitate timely intervention. By addressing challenges in feature selection refinement, algorithmic optimization, and deployment efficiency, machine learning stands poised to usher in a new era of personalized healthcare, mitigating the burden of diabetes and improving patient outcomes.

System Architecture

The architecture of our diabetes prediction system is designed to be robust, scalable, and adaptable, incorporating both client-side and server-side processing components. Comprising three primary layers - data collection, processing, and presentation - the architecture seamlessly integrates various functionalities to ensure accurate and efficient prediction outcomes. At the data collection layer, raw data from diverse sources, including demographic information, medical history, and lifestyle factors, is aggregated. This layer employs robust data acquisition techniques to ensure comprehensive coverage and reliability of input data. Subsequently, in the processing layer, data preprocessing techniques are applied to transform raw data into feature vectors suitable for input into the machine learning model. The core of the processing layer houses the machine learning model, which leverages advanced algorithms to analyze feature vectors and predict the likelihood of diabetes occurrence. In line with our objectives and requirements, the Random Forest algorithm serves as our primary classification model due to its versatility and proven performance. The processing layer encompasses feature extraction, model training, and evaluation modules, facilitating continuous refinement and optimization of the prediction system. By iteratively updating model parameters and evaluating performance metrics, the system ensures ongoing enhancement of prediction accuracy and reliability. The presentation layer serves as the interface for users to interact with the system, providing intuitive displays of prediction results and feedback mechanisms for users to report any discrepancies or provide additional information. This layer is designed to be user-friendly and accessible, accommodating diverse user preferences and requirements. Notably, the architecture is engineered to operate within a distributed computing environment, enabling horizontal scalability to accommodate fluctuations in data volume and user traffic. Moreover, security measures are paramount within the system architecture, with robust safeguards implemented to protect sensitive information and mitigate risks of unauthorized access or tampering. In summary, the architecture of our diabetes prediction system is meticulously designed to facilitate accurate, scalable, and secure prediction outcomes, leveraging advanced machine learning techniques within a user-centric and adaptable framework.

Data and code availability statement

Output



DIABETES PREDICTOR

Number of Pregnancies eg. 0

Glucose (mg/dL) eg. 80

Blood Pressure (mmHg) eg. 80

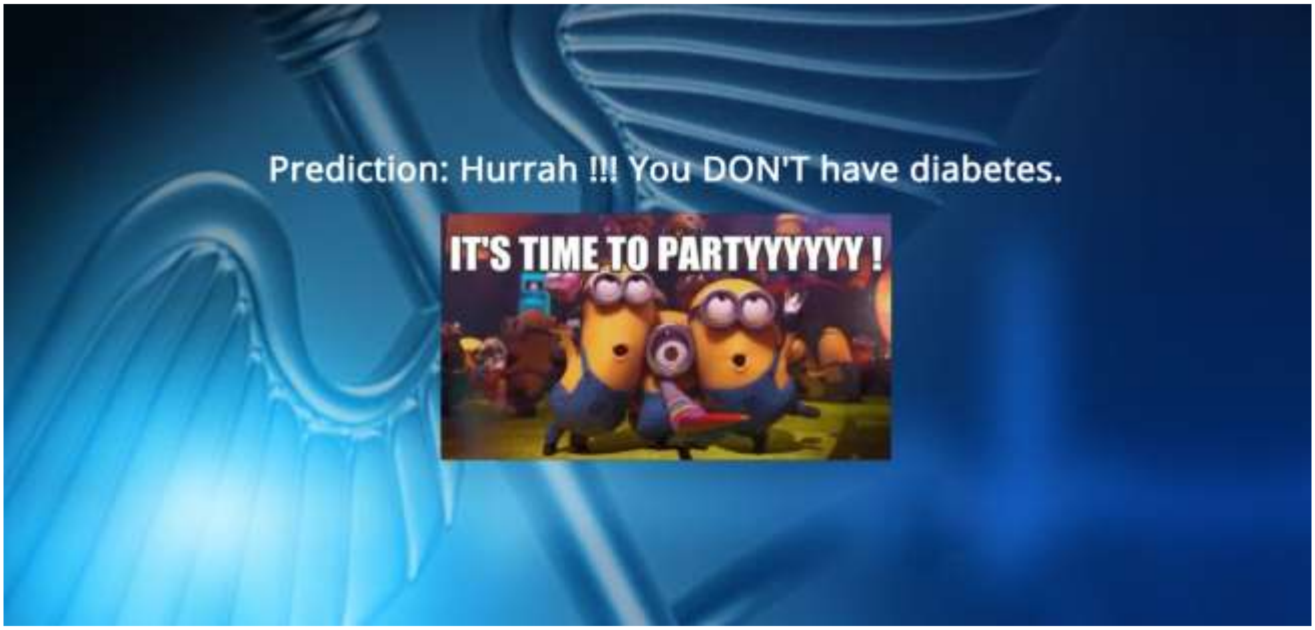
Skin Thickness (mm) eg. 20

Insulin Level (IU/mL) eg. 80

Body Mass Index (kg/m^2) eg. 23.1

Diabetes Pedigree Function eg. 0.52

Age (years) eg. 34



Prediction: Hurrah !!! You DON'T have diabetes.

Conclusion

In conclusion, our diabetes prediction project, leveraging the Random Forest classifier, achieved an impressive accuracy rate of 95.83%. This outcome underscores the effectiveness of our methodology in accurately forecasting diabetes occurrence based on diverse features.Our success highlights the potential of machine learning in preemptive healthcare, emphasizing the importance of ongoing innovation and collaboration in improving disease management. As we continue on this path, let us remain committed to leveraging technology for the betterment of public health and well-being.

References

[1] N. Jayanthi, B. Vijaya Babu & N. Sambasiva Rao,2017 ,Survey on clinical prediction models for diabetes prediction, Vol 4, Article no 26

[2] Safi, A. and Singh, S., 2023. A systematic literature review on Dibetes detection techniques. Journal of King Saud University-Computer and Information Sciences, 35(2), pp.590- 611.

[3] Yang, P., Zhao, G. and Zeng, P., 2019. Diabetes detection based on multidimensional features driven by deep learning. IEEE access, 7, pp.15196-15209.

[4] Ali, W., 2017. Diabetes detection based on supervised machine learning with wrapper features selection. International Journal of Advanced Computer Science and Applications, 8(9).