A
Mini Project Report on

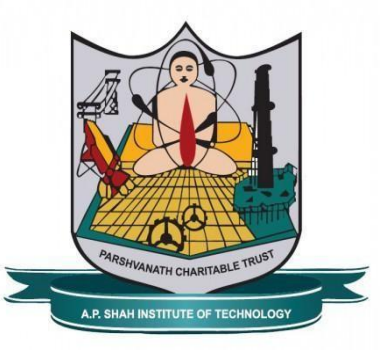
Diabetes Prediction using Machine Learning

Submitted in partial fulfillment of the requirements
for the degree of
BACHELOR OF ENGINEERING
IN
Computer Science & Engineering
Artificial Intelligence & Machine Learning

by

Kartik Parmar (22106101)
Pratham Patange (22106045)
Niraj Patel (22106050)
Jay Patil (22106020)
Under the guidance of

Prof. Kanchan Wankhede



Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai
2024-2025

A. P. SHAH INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled "**Diabetes Prediction using Machine Learning"** is a bonafide work of Kartik Parmar (22106101), Pratham Patange (22106045), Niraj Patel (22106050), Jay Patil (22106020) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Kanchan Wankhede	Dr. Jaya Gupta
Mini Project Guide	Head of Department

A. P. SHAH INSTITUTE OF TECHNOLOGY

Project Report Approval

This Mini project report entitled "**Diabetes Prediction using Machine Learning"** by **Kartik Parmar, Pratham Patange, Niraj Patel and Jay Patil** is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AI&ML) ***2024-25***.

External Examiner: _____

Internal Examiner: _____

Place: APSIT, Thane

Date:

Declaration

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission hasnot been taken when needed.

Kartik Parmar	Pratham Patange	Niraj Patel	Jay Patil
(22106101)	(22106045)	(22106050)	(22106020)

ABSTRACT

The Diabetes prediction is a critical area in healthcare, focusing on using machine learning techniques to forecast the likelihood of an individual developing diabetes based on various risk factors and health indicators. Diabetes is a chronic condition characterized by high blood sugar levels, which, if left unmanaged, can lead to severe complications such as heart disease, kidney failure, and vision loss. Early prediction and diagnosis are essential for managing the disease effectively and reducing its long-term impact. Machine learning models, such as Support Vector Machine (SVM) are increasingly being used to predict diabetes due to their ability to handle large datasets and detect complex patterns that may not be obvious to traditional statistical methods. These models are trained on datasets that include a variety of features like age, body mass index (BMI), blood pressure, glucose levels, cholesterol levels, lifestyle factors (such as diet and physical activity), and genetic predisposition. By analyzing the relationships between these variables and the occurrence of diabetes in the training data, the model learns to identify patterns and markers indicative of diabetes risk. For example, elevated glucose levels combined with high BMI and family history may strongly suggest a higher risk. The accuracy and reliability of these predictions depend on several factors, including the quality and comprehensiveness of the data, the appropriateness of the selected features, and the effectiveness of the machine learning algorithm used. The ultimate goal is to enhance patient outcomes through personalized healthcare strategies, reduce healthcare costs associated with diabetes management, and improve overall public health by preventing the onset of diabetes in high-risk populations.

Keywords: Diabetes, Algorithm, Healthcare

Index

Table Of Figure

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Diabetes is a growing global health issue, affecting millions of people and leading to serious complications such as heart disease, kidney failure, nerve damage, and vision loss. Traditionally, diabetes is diagnosed through biochemical tests, such as fasting blood sugar or HbA1c levels, and by recognizing clinical symptoms, which often only become apparent after significant damage has already occurred. This delayed diagnosis can result in missed opportunities for early intervention and prevention, which are crucial for minimizing the impact of diabetes and improving long-term health outcomes. To address this challenge, there is a need for predictive tools that can identify individuals at risk of developing diabetes before symptoms or irreversible damage occur. A diabetes predictor based on machine learning models offers an innovative solution for early detection. Machine learning algorithms, such as Support Vector Machine (SVM), can analyze complex datasets to identify patterns and correlations between various health indicators like age, body mass index (BMI), blood pressure, glucose levels, cholesterol, lifestyle factors, and family history and the likelihood of developing diabetes.

By learning from past data, these models can predict the risk of diabetes with high accuracy, even when clinical symptoms are not yet evident. This predictive capability enables healthcare providers to identify at-risk individuals early, allowing for timely intervention strategies such as recommending lifestyle changes, dietary modifications, regular monitoring, and preventive medication. Such proactive measures can significantly delay or even prevent the onset of diabetes, thereby reducing the disease's long-term complications and the associated healthcare costs.

The objective of this project is to develop a comprehensive diabetes prediction model using advanced machine learning techniques. By leveraging data-driven approaches, the project aims to provide a powerful tool for healthcare professionals to predict diabetes risk and implement preventive strategies more effectively. This model not only has the potential to improve individual patient outcomes but also contributes to broader public health efforts in managing the rising diabetes epidemic.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1-HISTORY

The history of diabetes prediction dates back to the early efforts in medical research to understand the risk factors and early indicators of diabetes. Initially, diabetes prediction relied heavily on statistical methods and clinical observations, where healthcare professionals identified patterns based on individual risk factors such as age, family history, obesity, and lifestyle habits. These early approaches were limited in scope and accuracy, often only identifying diabetes once symptoms were already present or the disease had significantly progressed. As computing power and data collection methods improved, researchers began to explore more sophisticated statistical models, like logistic regression, to predict diabetes risk based on a combination of demographic and clinical variables.

The advent of machine learning in the late 20th and early 21st centuries marked a significant turning point in diabetes prediction. Machine learning algorithms, such as decision trees, neural networks, and support vector machines (SVM), offered a more dynamic approach by analyzing large datasets to uncover complex, non-linear relationships between various risk factors and diabetes outcomes. These models could learn from historical data, improve over time, and make more accurate predictions about who might develop diabetes. This shift enabled the creation of more comprehensive prediction models that incorporated a broader range of data, including genetic information, laboratory test results, and even real-time data from wearable devices.

In recent years, advancements in artificial intelligence (AI) and big data have further transformed diabetes prediction. Deep learning, a subset of machine learning, has enabled the development of models that can process and learn from enormous datasets, including complex data from electronic health records (EHRs), wearable devices, and continuous glucose monitors. These deep learning models can identify subtle, non-linear relationships and interactions between variables, providing highly accurate predictions and enabling earlier detection of diabetes risk.

Overall, the history of diabetes prediction reflects a progression from simple risk factor analysis to sophisticated, data-driven models that are helping to transform diabetes care by enabling earlier detection, more precise interventions, and ultimately, better patient outcomes.

2.2-LITERATURE REVIEW

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem.[1]

Changes in body condition are caused by consuming a diet containing sugar, high fat, protein, and lack of water consumption, plus, more and more people are obese due to lack of exercise. Obesity is one of the few health problems in the world that still becomes a research topic. Many studies have stated that obesity is one of the risk factors that cause type 2 Diabetes Mellitus. World Population Data Sheet 2017 revealed that there were 7.5 billion people in the world. In 2017, the International Diabetes Federation (IDF) explained that diabetes was experienced by 425 million people aged 20-79 years. The highest number of people with diabetes was between the ages of 40 and 59 years. Diabetes caused a total of 4 million people to die in 2017.[2]

The Pima Indian diabetic database at the UCI machine learning laboratory has become a standard for testing data mining algorithms to see their prediction accuracy in diabetes data classification. The proposed method uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machine can be successfully used for diagnosing diabetes disease. SVM is a set of related supervised learning method used in medical diagnosis for classification and regression. [3]

The problems begin when diabetes is not detected at an early phase and diagnosed properly at the appropriate time. Different machine learning techniques, as well as ontology-based ML techniques, have recently played an important role in medical science by developing a system that can detect diabetes patients. This paper provides a comparative study and review of the most popular machine learning techniques. Various types of classification algorithms were considered namely: SVM, KNN, ANN, Naive Bayes, Logistic regression, and Decision Tree. The experimental results showed that the best accuracy goes for ontology classifiers and SVM.[4]

Exploiting Machine Learning Algorithms (MLA) is essential if healthcare professionals are able to identify diseases more effectively. In order to improve the medical diagnosis of diabetes this research explored and contrasts various MLA that can identify diabetes risk early. Machine learning can use past outcomes to make smart decisions on present cases that were before unknown. To evaluate how well the algorithms worked in Pima Indian data collection was used. The accuracy that was achieved using 10-fold cross-validation was 77.86 percent across the five algorithms (neural network, SVM, DTM, LR, and Naive Bayes) that were used.[5]

CHAPTER 3

PROBLEM STATEMENT

3. Problem Statement

The increasing prevalence of diabetes worldwide presents a significant public health challenge, with millions of individuals at risk of developing this chronic disease and its associated complications. Traditional methods for diagnosing diabetes often detect the condition only after it has progressed to a more severe stage, missing crucial opportunities for early intervention and prevention. This delay in diagnosis can lead to long-term health issues, increased healthcare costs, and reduced quality of life for patients. There is a critical need for an effective and accurate predictive tool that can identify individuals at high risk of developing diabetes well before symptoms appear. The problem is further compounded by the complexity of diabetes, which is influenced by a wide range of factors, including genetic predisposition, lifestyle choices, and other health conditions. Developing a reliable diabetes prediction model using machine learning techniques, such as Support Vector Machine (SVM), can help address this gap by analyzing diverse datasets to uncover patterns that indicate diabetes risk. Such a model could provide healthcare providers with a powerful tool for early diagnosis and personalized intervention strategies, ultimately improving patient outcomes. Addressing this problem not only has the potential to improve individual patient outcomes by preventing or delaying the onset of diabetes but also to alleviate the overall burden on healthcare systems, reduce costs associated with diabetes management, and enhance public health efforts globally.

CHAPTER 4

EXPERIMENTAL SETUP

4. Experimental Setup

4.1 Hardware Setup

1. **CPU:** core i5 or higher version
2. **RAM:** recommended 4GB and More
3. **STORAGE:** 256GB Disk Space or More
4. **OS:** Microsoft Windows 7, Microsoft Windows 8, Microsoft Windows 10 or later

4.2 Software Setup

Programming Languages:

- Python
- Streamlit
- Mongo DB

Libraries and Frameworks:

- NumPy
- Pandas
- scikit-learn (sklearn)

CHAPTER 5

PROPOSED SYSTEM & IMPLEMENTATION

5. Proposed system & Implementation

5.1 Block diagram of proposed system

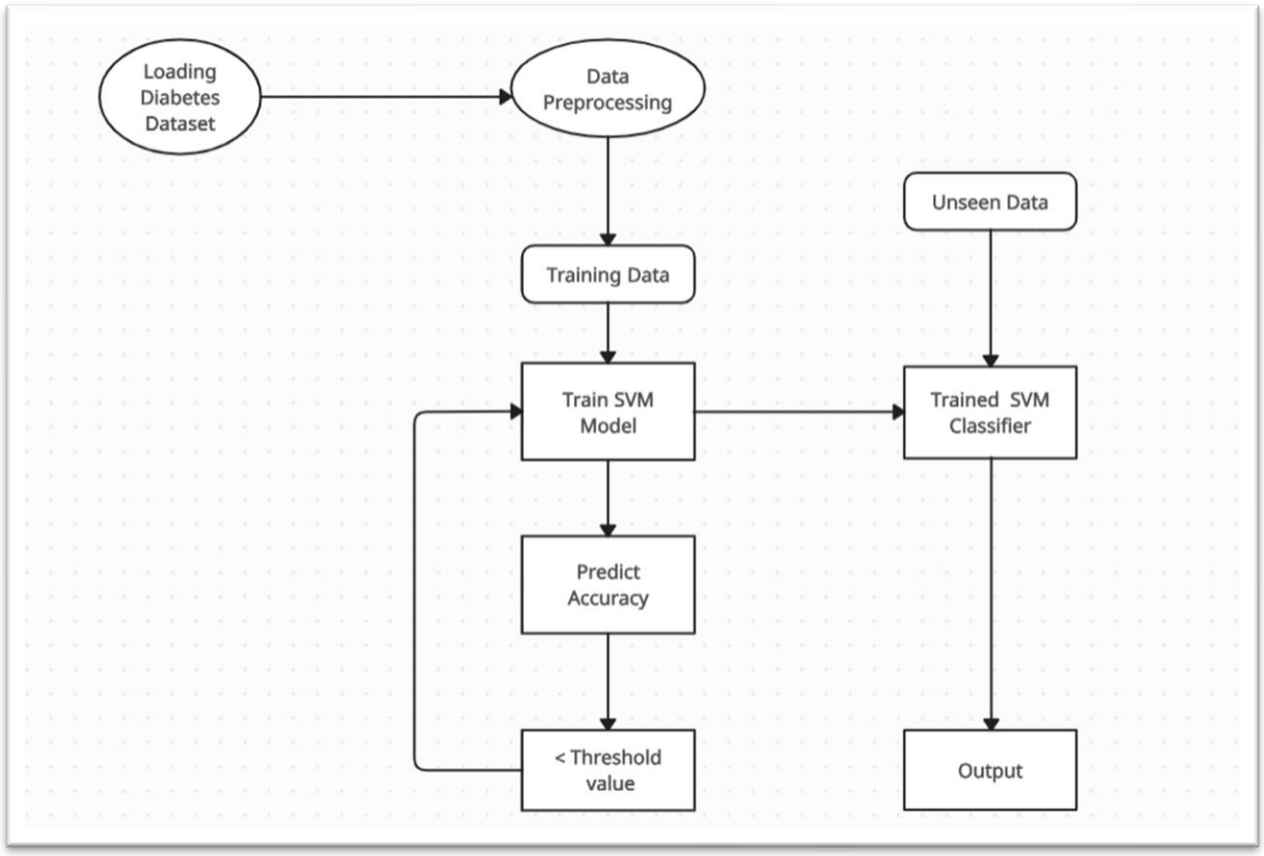


Fig 5.1.1 Block Diagram

5.2 Description of block diagram

1. Diabetes Dataset: This represents the entire dataset containing information about diabetes patients. It's likely split into two parts: training data and test data.

2. Training Data: This subset of the dataset is used to train the SVM model. It contains features (e.g., age, blood sugar level, BMI) and corresponding labels (e.g., diabetic or non-diabetic).

3. Train SVM Model: The SVM model is built using the training data. This involves finding the optimal hyperplane that separates the data points into different classes based on their features.

4. Predict Training Data: The trained SVM model is used to predict the labels of the training data points. This helps assess the model's performance on data it has already seen.

5. Select Badly Predicted Training Data: The training data points that were misclassified by the SVM model are identified. These are the points that the model struggled to predict correctly.

6. SVM Classifier: This refers to the trained SVM model itself. It's the machine learning algorithm that will be used to make predictions on new, unseen data.

7. Output: This is the final result of the process. It could be the predicted labels for the test data or any other relevant metrics that evaluate the model's performance, such as accuracy, precision, recall, or F1-score.

5.3 Implementation

Fig 5.3.1 Register Page

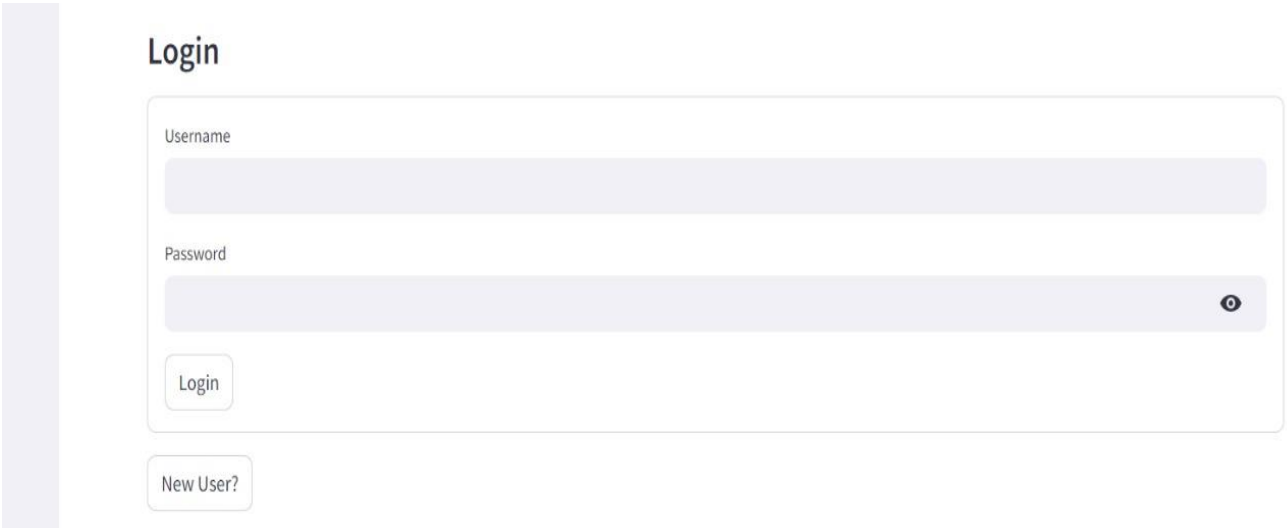


Fig 5.3.2 Login Page

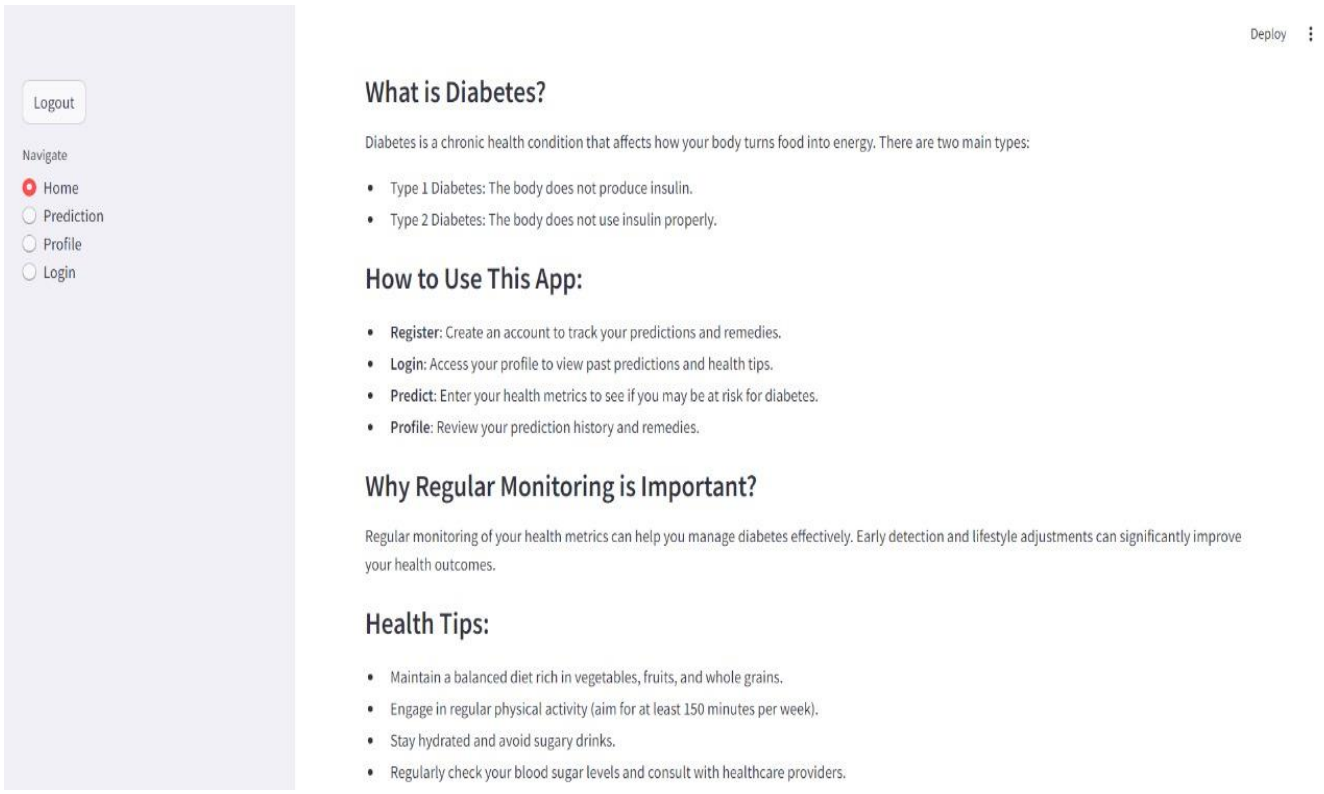


Fig 5.3.3 Home Page

Fig 5.3.4 Prediction Page

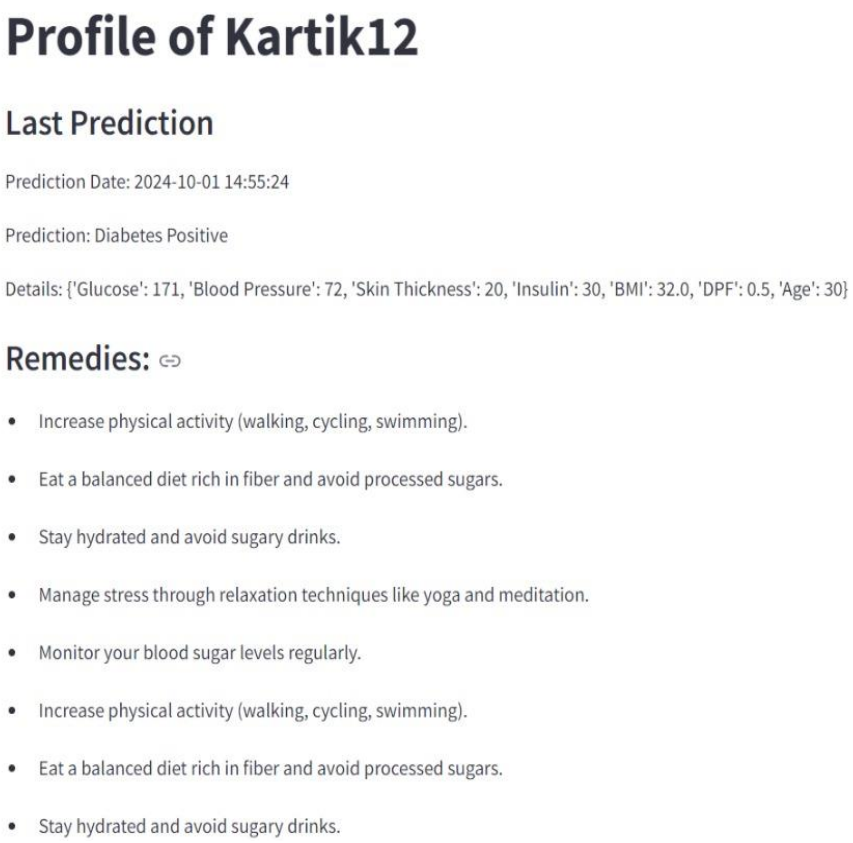


Fig 5.3.5 Profile Page

CHAPTER 6

CONCLUSION

6. Conclusion

In conclusion, the development of a diabetes predictor using machine learning techniques represents a significant advancement in the early detection and management of diabetes. By leveraging the power of algorithms such as Support Vector Machine (SVM), this predictive model can analyze complex datasets to accurately identify individuals at high risk of developing diabetes, often before clinical symptoms become apparent. This approach enables healthcare providers to implement early interventions, such as lifestyle modifications, preventive treatments, and continuous monitoring, which can prevent or delay the onset of diabetes and its associated complications. The successful implementation of such a model could transform diabetes care, leading to better patient outcomes, reduced healthcare costs, and a decrease in the overall burden of the disease. As data availability and machine learning techniques continue to evolve, there is great potential for further refining these models, making them more accurate, accessible, and integral to personalized healthcare strategies. The diabetes predictor is not just a tool for risk assessment but a crucial component in the shift towards preventive medicine and improved public health outcomes.

References

[1] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.

[2] Anggoro, Dimas Aryo, and Dian Novitaningrum. "Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease." ICIC Express Letters 15.1 (2021): 9-18.

[3] Classification Of Diabetes Disease Using Support Vector Machine V. Anuja Kumari, R.Chitra

[4] H. E. Massari, Z. Sabouri, S. Mhammedi and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," in Journal of ICT Standardization, vol. 10, no. 2, pp. 319-337, 2022, doi: 10.13052/jicts2245-800X.10212.

[5] Bhat, Salliah Shafi, Venkatesan Selvam, Gufran Ahmad Ansari, Mohd Dilshad Ansari, and Md Habibur Rahman. "Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district bandipora." *Computational Intelligence and Neuroscience* 2022, no. 1 (2022): 2789760.

URL

[1] https://www.sciencedirect.com/science/article/pii/S1877050918308548
[2] http://www.icicel.org/ell/contents/2021/1/el-15-01-02.pdf
[3] https://www.researchgate.net/profile/Chitra-Jegan/publication/320395340_Classification_Of_Diabetes_Disease_Using_Support_Vector_Machine/links/59e1c7a80f7e9b97fbe72c89/Classification-Of-Diabetes-Disease-Using-Support-Vector-Machine.pdf
[4] https://ieeexplore.ieee.org/document/10255389
[5] **https://doi.org/10.1155/2022/2789760**