Mini Project Report on

GlucoSage – Diabetes Risk Predector

Submitted in partial fulfillment of the requirements of the degree of Bachelor in Engineering

By

Dhwani Darji (21UF16744CM134)

Anuja Agare (21UF16441CM126)

Aniket Jadhao (21UF16149CM144)

Ojas Hedau (21UF16824CM143)

Under the guidance of

Prof. Pallavi Khodke

DEPARTMENT OF COMPUTER ENGINEERING

SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE

CHEMBUR, MUMBAI – 400088.

University of Mumbai

(AY 2023-24)

Mahavir Education Trust's
SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE
Chembur, Mumbai - 400 088
Department of Computer Engineering

UG Program in Computer Engineering is re-accredited by N. B. A. New Delhi from AY 2022-23 for 3 years up to 30.06.2025. Awarded 'A' Grade (3.16 CGPA) by N. A. A. C. w. e. f. 20.10.2021

Certificate

This is to certify that the report of the Mini Project entitled **GlucoSage – Diabetes Risk Predictor** is a bonafide work of

Name of Student	Class	Roll No
Dhwani Darji	TE9	26
Anuja Agare	TE4	01
Aniket Jadhao	TE3	24
Ojas Hedau	TE9	08

submitted to the
UNIVERSITY OF MUMBAI

during Semester VI

in

COMPUTER ENGINEERING

(Ms. Pallavi Khodke)
Guide

(Prof. Uday Bhave)
I/c Head of Department

(Dr. Bhavesh Patel)
Principal

Mini Project Approval

This Mini Project entitled "**GlucoSage – Diabetes Risk Predictor"** by **Dhwani Darji, Anuja Agare, Aniket Jadhao, Ojas Hedau** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1……
(Internal Examiner Name & Sign)

2……
(External Examiner name & Sign)

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We 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 has not been taken when needed.

Name of student	Class	Roll No.	Signature
Dhwani Darji	TE9	26	
Anuja Agare	TE4	1	
Aniket Jadhao	TE3	24	
Ojas Hedau	TE9	8	

Date:

Place:

Contents

Abstract

The involvement of technology in the medical sector was one of the most significant milestones which we have yet achieved. Machine learning models can be used to predict various serious diseases like diabetes in humans at an early and curable stage. Every year, 1.6 million individuals lose their lives due to the same illness. Despite advancements in medicine and our body's resilience, the prevalence of diabetes has increased in recent years. By analysing the medical data of diabetic patients, we can identify common patterns and potentially predict the onset of diabetes at an early stage. This proactive approach aids in combating the disease and promoting well-being. Thus, we used computer algorithms like logistic regression and random forest to build and test our prediction model.

Our findings showed that age, high weight, lack of physical activity, high blood sugar levels, and family history of diabetes were important factors in predicting diabetes risk. The model worked well across different groups of people based on age, gender, and ethnicity.

This research provides a useful tool for healthcare providers to identify people at risk of diabetes early. This could lead to better personalized care and strategies to prevent diabetes. However, there were some limitations, such as relying on self-reported lifestyle information.

In conclusion, using advanced computer analysis can help create accurate tools to predict diabetes risk. Future work could involve adding more details like blood markers and using new technologies like wearable devices to improve predictions.

Keywords: Diabetes predictor, random forest classifier, support vector machine

Acknowledgement

We would like to express our sincere gratitude to all those who contributed to the successful completion of the diabetes risk predictor project. First and foremost, we extend our heartfelt appreciation to Dr. Bhavesh Patel, our esteemed Principal, for supporting us and believing in our project. We couldn't have done it without their encouragement. We also want to thank Prof. Uday Bhave, our Head of the Department, for guiding us throughout the project. Their advice was really helpful.

Special thanks to our project guide Prof. Pallavi Khodke, Asst. Prof in Department of Computer Engineering. She only not encouraged us to pursue our work but also initiated selfbelief and confidences. Furthermore, we extend our thanks to the faculty members of the Department of Computer Engineering, whose expertise and feedback enriched our understanding and facilitated the progress of our work. Our sincere appreciation also goes to our classmates who contributed to various aspects of this project.

In conclusion, we are thankful to all those who played a role, however big or small, in bringing this project to fruition. Your support and encouragement has been invaluable.

1. Introduction

Diabetes is a serious and long-lasting health condition caused by problems with insulin, a hormone that controls blood sugar levels. According to the International Diabetes Federation (IDF), around 463 million adults currently have diabetes, and this number could reach 700 million by 2045. Most cases are of Type 2 diabetes, which accounts for 90% of all diabetes cases. Diabetes is responsible for about 4.2 million deaths worldwide and is a major risk factor for kidney failure, heart attacks, and strokes.

The impact of diabetes extends beyond high blood sugar levels. It significantly increases the risk of various health complications, including heart disease, stroke, kidney failure, nerve damage, and vision impairment. Diabetes-related complications contribute to reduced quality of life and increased healthcare costs.

Recently, there's been a lot of interest in using computers to predict who might get diabetes in the future. These computers use something called "machine learning," which is a fancy way of saying they can look at a lot of information and find patterns. Things like age, weight, family history, and lifestyle habits can all give clues about who might be at risk for diabetes. Early diagnosis of diabetes is crucial because it allows for timely treatment and can help prevent complications. Medical experts use various methods to identify people at risk of developing diabetes before they show symptoms. Recently, there's been a lot of interest in using machine learning—a type of computer technology—to predict who might get diabetes in the future. In the face of this burgeoning epidemic, early intervention and preventive strategies are paramount. Timely identification of individuals at risk before symptoms manifest is pivotal in mitigating the deleterious effects of diabetes. To this end, healthcare professionals are increasingly turning to innovative approaches, including the burgeoning field of machine learning, to enhance predictive capabilities

Studies have shown that machine learning algorithms like decision trees, random forests, and others can effectively predict diabetes risk. These algorithms are good at handling large amounts of data and can identify patterns that indicate someone's likelihood of developing diabetes. To figure out which algorithm works best, researchers use different measures like accuracy, ROC values, F-score, and how long it takes to make predictions. While accuracy is important, it's not the only thing to consider when choosing the best method for predicting diabetes risk.

The goal is to develop and evaluate a machine learning-based diabetes risk prediction model that can accurately identify individuals at risk of developing diabetes. By incorporating multiple risk factors into the predictive model and to create a reliable prediction model that uses important factors like age, weight, family history of diabetes, and lifestyle habits. By using advanced computer techniques, we hope to improve early detection of diabetes and ultimately help people manage their health better. Considerations of computational efficiency, interpretability, and scalability are paramount, ensuring practical applicability in real-world healthcare settings. As such, our effort seeks to bridge the gap between theoretical promise and practical utility by developing and evaluating a machine learning-based diabetes risk prediction model.

We aim to contribute to the advancement of diabetes prevention and management and to enhance early detection efforts and facilitate targeted preventive interventions. Early identification of individuals at risk of diabetes can lead to timely interventions, such as lifestyle modifications and preventive measures, ultimately reducing the incidence of diabetes-related complications and improving overall health outcomes. Our goal is to make a computer program that can predict who might get diabetes based on all this information. By doing this, we hope to help people make healthier choices and prevent diabetes-related problems before they happen.

Our project wants to make it easier for doctors to spot people who might be at risk for diabetes early. This way, they can get the help they need to stay healthy and avoid serious health issues. Ultimately, we want to use technology to improve healthcare and make sure everyone has the chance to live a healthy life.

2. Literature Survey

Author(s)	Year	Title	Key Finding
Farida Mohsen, Hamada R. H. AlAbsi, & Zubair Shah	2023	A scoping review of artificial intelligence-based methods for diabetes risk prediction	Promising Performance of AI Models Variety of Data Modalities and Techniques
Nikos Fazakis Otilia Kocsis Elias Dritsas Sotiris Alexiou	2021	Machine Learning tools for Long Term Type 2 Diabetes Risk Prediction	Apply different ensemble algorithms to a dataset, combining different families of ML models to predict the risk of T2DM(Type 2 Diabetes Mellitus)
Apratim Sadhu Abhimanyu Jadli	2021	Early stage diabetes risk prediction: A Comparative Analysis of Classification Algorithms	Comparative study of classification algorithm to find the best and accurately suitable algorithm
Gary S Collins, Susan Mallett, Omar and Ly-Mee Yu	2011	Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting	Impact of inadequate Methods Usage Impact on Reliability

3. Problem Statement

We need better ways to predict who might develop diabetes before they show symptoms. Current methods often miss early signs, leading to delayed treatment and higher healthcare expenses. This research aims to create a smart computer system using machine learning to predict diabetes risk accurately. By considering different factors like age, lifestyle, and family history, we hope to catch diabetes early and prevent complications effectively.

4. Objectives

- **Develop a Diabetes Risk Prediction Model:** Create a computer model using machine learning that can accurately predict the risk of developing diabetes based on factors like age, lifestyle, and family history.

- **Integrate Diverse Risk Factors:** Incorporate various important factors into the prediction model, including demographics, clinical data, and lifestyle habits, to enhance the accuracy of diabetes risk assessment.

- **Enable Early Detection:** Enable early identification of individuals at high risk of diabetes before symptoms appear, allowing for timely interventions and preventive measures

5. Methodology

5.1 The Dataset

The dataset that we have used is taken from the UCI repository. It contains 520 instances and 16 attributes with a few missing values which have been pre-processed by ignoring the tuples with incomplete values. The dataset is summarised in Table 1

TABLE I. ATTRIBUTE DESCRIPTION

Attributes	Description
Age	20 years - 65 years
Sex	1.Male, 2.Female
Polyuria (Excessive Urination)	1.Yes, 2.No.
Polydipsia(Excessive Thirst)	1.Yes, 2.No.
Sudden Weight loss	1.Yes, 2.No.
Weakness	1.Yes, 2.No.
Polyphagia(Excessive Hunger)	1.Yes, 2.No.
Genital thrush	1.Yes, 2.No.
Visual blurring	1.Yes, 2.No.
Itching	1.Yes, 2.No.
Irritability	1.Yes, 2.No.
Delayed Healing	1.Yes, 2.No.
Partial Paresis	1.Yes, 2.No.
Muscle Stiffness	1.Yes, 2.No.
Alopecia	1.Yes, 2.No.
Obesity	1.Yes, 2.No.
Class	1.Yes, 2.No.

After pre-processing the dataset a total of 520 instances remained. Out of these 520 instances, 320 are positive and 200 are negative values The two class variables(positive or negative) are used to find whether a patient has a risk of diabetes or not.

5.2 Experimental Procedure

The experimental procedure is carried out in the steps described below:
The dataset is partitioned into training and test set in a ratio of 80 : 20 respectively using and the two most suited classification algorithms is applied to classify the dataset into positive and negative classes. Four evaluation metrics: classification accuracy, F-score, ROC value and computation time is calculated to compare the performance of the specified algorithms. This was done to find the best classification algorithm for our problem statement.

5.3 Algorithms

1) Support Vector Machines :

Support Vector Machines(SVM) was first introduced by Vapnik. SVM works by selecting critical samples from all classes known as support vectors and separating the classes by generating a function that divides them as broadly as possible using these support vectors. Therefore, it can be said that a mapping between an input vector to a high dimensionality space is made using SVM that aims to find the most suitable hyperplane that divides the data set into classes . This linear classifier aims to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance, by finding the best-suited hyperplane .

Upon comparing various kernel, Radial Basis function(RBF) kernel was used to classify the data, also known as the Gaussian kernel. When training an SVM with the Radial Basis Function (RBF) kernel, two hyperparameters must be considered: C and gamma. The hyperparameter C, common to all SVM kernels, trades off misclassification of training examples against the simplicity of the decision surface. A low value of C smoothens the decision surface, while a high value of C aims at classifying all training examples correctly. The extent of influence of a single training example is defined by gamma. The larger the value of gamma is, the closer other examples must be to be affected.

The distance between data points is measured by the Gaussian kernel:

$$K_{rbf}(x_i,x_i)=\exp(-\gamma\|x_i-x_j\|^2)$$

Here, x_i and x_i are data points, $\| x_i - x_j \|$ denotes Euclidean distance. The choice of kernel functions is dependent on the respective data and specific domain problem. The various values of C and gamma is checked over the validation set. The appropriate value of C for best accuracy on the validation set is selected based on its accuracy. The best validation score is obtained for gamma=0.1 and C=1.7.

2) Random Forests:

A random forest is essentially an ensemble of a number of decision trees. The logic that sticks with random forest is to combine different sets of values from training sets to form decision trees thus reducing the chances of overfitting and misclassification by averaging the results of various decision trees.

In the model described in the paper, we have used max-depth of 13 and 100 estimators to classify the data points. On increasing the max-depth, we experienced a decrease in the classification accuracy.

5.4 Analysis and comparison

After analysing and comparing the two algorithms, we were able to clearly make a concise comparison between the two models based on various aspects as summarized in table II

TABLE II. COMPARISON BETWEEN THE MODELS

Aspect	Random Forest Classifier	Support Vector Machine
Model Complexity	Ensemble method with multiple decision trees	Uses hyperplanes to separate data points
Performance	Robust and resistant to overfitting	Powerful with strong theoretical guarantees
Interpretability	Provides feature importances	Produces less interpretable models
Scalability	Relatively scalable, can handle large datasets	Less scalable, especially with non-linear kernels
Handling Imbalanced Data	Naturally robust to class imbalance	May require additional techniques for imbalanced data
Flexibility	Can handle both numerical and categorical data	Effective for binary classification tasks

5.5 Results And Observations

In this section, the classification accuracy, F-score and ROC value of both the algorithms used for the diabetes dataset is presented which clearly depicts the best fitted algorithm for this particular problem statement. All the evaluation metrics are summarised in table III.

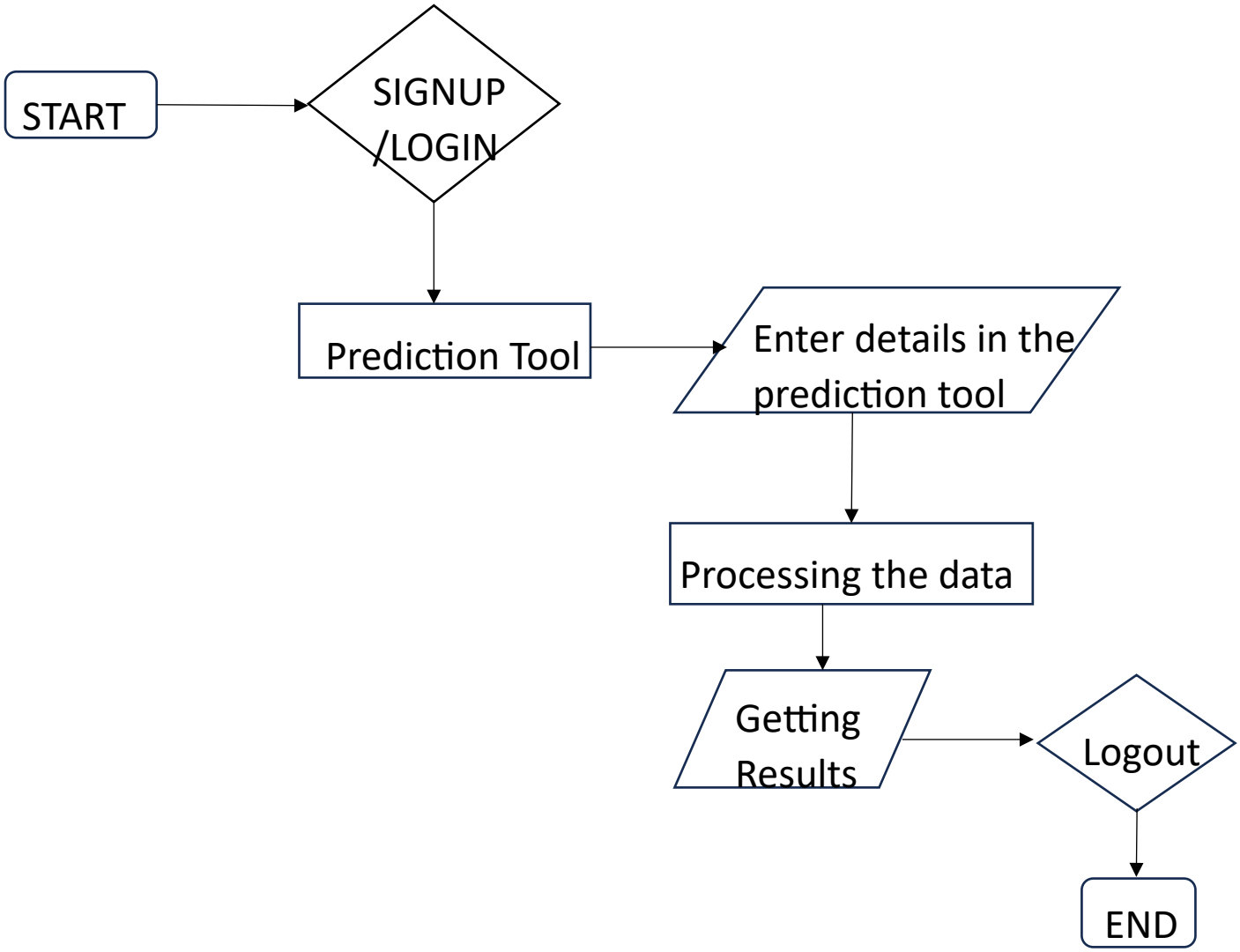
TABLE III. COMPARISON OF ACCURACY, F-SCORE, ROC VALUES OF 2 MODELS

Model	Accuracy	F-Score	ROC
Random Forests	98.0778%	0.9790	0.9979
Support Vector Machine	61.53%	0.6029	0.6566

The results drawn from training the machine learning models clearly indicate that Random Forest Classifier proved to be the best model among the models for the concerned dataset with an accuracy score of 98.0778%, ROC score of 0.9979 and F-score of 0.9790.

Hence, as per the results obtained we can firmly believe that Random Forest Classifier is one of the most effective algorithms against binary-based classification datasets.

6. Flowchart


START
SIGNUP /LOGIN
Prediction Tool
Enter details in the prediction tool
Processing the data
Getting Results
Logout
END

7. Software Requirements

1. Operating systems like windows.
2. Visual Studio Code code-editor.
3. Web browsers like chrome, safari, firefox, opera, etc.
4. Coding languages :

    Front end: HTML, CSS, Javascript

    Back end: Node.js, Express.js

    Database: MongoDB

    Model language: Python

8. Conclusion

Our project aimed to create a computer model that can predict the risk of developing diabetes before symptoms appear. We focused on using factors like age, lifestyle habits, family history, and clinical data to build a more accurate prediction tool. Our model plays a crucial role in identifying individuals at higher risk of diabetes. This early detection allows healthcare providers to intervene early with tailored preventive strategies.

The significance of our work lies in leveraging technology, specifically machine learning, to improve healthcare outcomes. By harnessing the power of data analysis and predictive algorithms, we can empower healthcare professionals to take proactive steps in managing diabetes.

Our project underscores the importance of early intervention in diabetes management. By identifying high-risk individuals sooner, we can initiate appropriate interventions such as lifestyle modifications, dietary changes, and regular monitoring to prevent or delay the onset of diabetes-related complications.

While our model represents a step forward in diabetes risk prediction, it is part of a broader effort to advance preventive healthcare strategies. Continued collaboration between researchers, healthcare providers, and technology experts is crucial in refining and deploying these tools effectively.

Ultimately, our goal is to contribute to better health outcomes for individuals at risk of diabetes worldwide. Early detection facilitated by our predictive model has the potential to reduce the burden of diabetes-related complications and improve quality of life for many.

9. Future Scope

- **Preventive Recommendations:** Expanding our project to include preventive recommendations based on risk profiles can empower individuals to take proactive steps towards preventing diabetes. This may involve integrating lifestyle suggestions, such as diet and exercise plans, into our predictive model outputs.

- **Long-Term Follow-Up Studies:** Conducting long-term follow-up studies to track outcomes and monitor the progression of diabetes among individuals identified as high-risk by our predictive model. This can provide valuable insights into the effectiveness of early interventions and preventive strategies.

- **Expansion to Other Health Conditions:** We can explore using our predictive model for other health conditions like heart disease or obesity. By adapting our model to assess risks for these related conditions, we can provide a more comprehensive approach to preventive healthcare.

10. References

[1] Apratim Sadhu, Abhimanyu Jadli ‘Early stage diabetes risk prediction: A Comparative Analysis of Classification Algorithms’ In 2021 International Advanced Research Journal in Science, Engineering and Technology

[2] Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, Konstantinos Moustakas ‘Machine Learning tools for Long Term Type 2 Diabetes Risk Prediction’ , July 2021

[3] Saxena, K., Khan, Z., & Singh, S. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm.

[4] Mller, Andreas C., and Sarah Guido. Introduction to Machine Learning with Python: a Guide for Data Scientists. O’Reilly, 2017.