

DIABETES PREDICTION USING ML

A PROJECT REPORT

Submitted to

Visvesvaraya Technological University

BELAGAVI - 590 018

by

Rohan Ponnanna KK

4SU19CS076

Seetaram Naik

4SU19CS088

Srinivas S

4SU19CS101

Sulekha PB

4SU19CS103

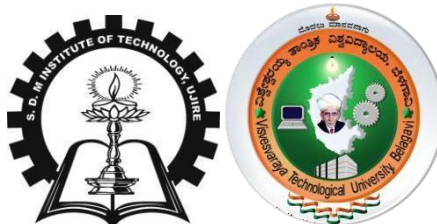
Under the guidance of

Mr. Chaithanya D

Assistant Professor

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering



Department of Computer Science & Engineering

SDM INSTITUTE OF TECHNOLOGY

UJIRE - 574 240

2022-2023

SDM INSTITUTE OF TECHNOLOGY

(Affiliated to Visvesvaraya Technological University, Belagavi)

UJIRE – 574 240

Department of Computer Science and Engineering

CERTIFICATE

Certified that the Project Work titled '**Diabetes Prediction Using ML**' is carried out by **Mr. Seetaram Naik**, USN: **4SU19CS088** bonafide student of SDM Institute of Technology, Ujire, in partial fulfillment for the award of the degree of **Bachelor of Engineering** in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi during the year 2022-2023. It is certified that all the corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said Degree.

Mr. Chaithanya D
Asst. Professor and Guide

Dr. Thyagaraju G S
Professor and Head

Dr. Ashok Kumar T
Principal

Signature with date and seal:

External Viva

Name of the Examiners:

Signature with Date

1.

2.

Acknowledgement

It is our pleasure to express our heartfelt thanks to Mr. Chaithanya D, Assistant Professor, Department of Computer Science and Engineering, for his supervision and guidance which enabled us to understand and develop this project.

We are indebted to Dr. Thyagaraju G S, Head of the Department and Dr. Ashok Kumar T, Principal, for their advice and suggestions at various stages of the work. We also extend our heartfelt gratitude to Dr. H Manoj T Gadiyar, Project Coordinator for his assistance and the management of SDM Institute of Technology, Ujire, for providing us with a good learning environment, library and laboratory facilities.

Lastly, we take this opportunity to offer our regards to all of those who have supported us directly or indirectly in the successful completion of this project work.

Rohan Ponnanna KK

Seetaram Naik

Srinivas S

Sulekha PB

Abstract

Diabetes is a habitual complaint with the eventuality to beget a worldwide health care extremity. The International Diabetes Federation reports that currently there are 382 million individuals worldwide who have diabetes. In the next decade, by 2035, this number is projected to increase to 592 million, which is double the current figure. Diabetes mellitus or indicate diabetes is a complaint caused due to the increase position of blood glucose. colorful traditional styles, grounded on physical and chemical tests, are available for diagnosing diabetes. still, early vaticination of diabetes is relatively grueling task for medical interpreters due to complex interdependence on colorful factors as diabetes affects mortal organs similar as order, eye, heart, jitters, bottom etc. Data wisdom styles have the eventuality to profit other scientific fields by slipping new light on common questions. One similar task is to help make prognostications on medical data. Machine literacy is an arising scientific field in data wisdom dealing with the ways in which machines learn from experience. The end of this design is to develop a system which can perform early vaticination of diabetes for a case with a advanced delicacy by combining the results of different machine literacy ways. This design aims to prognosticate diabetes via three different supervised machine literacy styles including SVM, Logistic regression, KNN. This design also aims to propose an effective fashion for earlier discovery of the diabetes complaint using Machine literacy algorithms and end to end deployment using streamlit.

Table of Contents

	Page No.
Abstract	i
Table of Contents	ii
List of Figures	iv
List of Tables	v
Chapter 1 Introduction	1
1.1 Project Introduction	1
1.2 Problem Description	1
Chapter 2 Literature Review	2
2.1 Literature Survey	2
2.2 Comparative Analysis of the Related Work	4
2.3 Summary	4
Chapter 3 Problem Formulation	5
3.1 Problem Statement	5
3.2 Objectives of the Present Study	5
3.3 Summary	5
Chapter 4 Requirements and Methodology	6
4.1 Hardware Requirements	6
4.2 Software Requirements	6
4.3 Methodology Used	6
Chapter 5 System Design	7
5.1 Architecture of the Proposed System	7
5.2 System Flowchart	8
Chapter 6 Implementation	9
6.1 Pseudocode	9
Chapter 7 System Testing, Results and Discussion	10
7.1 System Testing	10
7.2 Result Analysis	10
7.3 Summary	14

Chapter 8	Conclusion and Scope for Future Work	15
8.1	Conclusion	15
8.2	Scope for Future work	15
References		
Personal Profile		

List of Figures

	Page No.
Figure 5.1 Architecture of the Proposed Diabetes Prediction System	7
Figure 5.2 System Flowchart of the Proposed System	8
Figure 7.1 Feature Importance for Logistic Regression model	11
Figure 7.2 Feature Importance for RandomForest model	11
Figure 7.3 Login Page	12
Figure 7.4 Home Page	12
Figure 7.5 Input Page	13
Figure 7.6 Diet Recommendation Page	13
Figure 7.7 Exercise Information Page	14

List of Tables

	Page No.
Table 2.1 Comparative Analysis	4
Table 4.1 Hardware Requirement	6
Table 4.2 Software Requirement	6
Table 7.1 Unit Test Cases	10
Table 7.2 Analysis of the Algorithms	10

Introduction

1.1 Project Introduction

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and management of diabetes can help prevent serious complications and improve the quality of life of those affected. Machine Learning (ML) has emerged as a promising tool for early detection and prediction of diabetes. In this project, we aim to develop an ML model that can accurately predict the onset of diabetes based on a set of patient features, such as common diabetic symptoms. To achieve this, we will use a dataset of historical patient data, including both diabetic and non-diabetic patients.

We will use various ML algorithms such as logistic regression, decision trees, random forests, and support vector machines to develop the predictive model. We will also evaluate the performance of each algorithm and select the one that performs best in terms of accuracy, precision, recall, and F1 score. The ultimate goal of this project is to develop a reliable and accurate predictive model that can assist medical professionals in diagnosing and managing diabetes, thereby improving patient outcomes and quality of life and giving them more information such as diet planning and exercise information to control the diabetes.

1.2 Problem Description

Doctors often depend on conventional knowledge and research summaries to guide their treatment decisions, but these approaches can be time-consuming and may not always uncover important patterns in patient data. In contrast, machine learning can help identify significant patterns more quickly, but this technique requires large amounts of data. Unfortunately, for many diseases, the amount of available data is limited. Additionally, the number of individuals without the disease far exceeds those with the disease, leading to an imbalance in the dataset.

Some prediction system doesn't meet the recommended accuracy level for medical purpose. Common peoples cannot access advanced prediction system at free of cost. Most of the prediction system don't have user friendly interface. Existing model doesn't have proper diet recommendation system to maintain a healthy lifestyle.

Literature Review

2.1 Literature Survey

In the paper titled “A data mining approach for the diagnosis of diabetes mellitus. Intelligent Systems and Control (ISCO)” [1], the authors Kumari, Sonu, and Archana Singh aims to predict diabetes using ensemble voting classifiers on the Pima Indian diabetes dataset. The dataset is evaluated using various classification algorithms, including Decision Tree, K-Nearest Neighbors, Random Forest, and Support Vector Machines (SVM). Through extensive experimentation and analysis, the results reveal that Support Vector Machines achieved the highest accuracy of 80% for the dataset. Furthermore, employing 10-fold cross-validation yielded an accuracy of 81%. These findings suggest that SVM outperformed other algorithms in predicting diabetes on the Pima Indian diabetes dataset.

In the paper titled “Performance Analysis of Classifier Models to Predict Diabetes Mellitus” [2], the authors Pradeep Kandhasamy S. Balamurali studied the performance of various data mining algorithms for predicting diabetes is examined. The study focuses on comparing the accuracy of classifiers, including Decision Tree, K-Nearest Neighbors, Random Forest, and Support Vector Machines (SVM), using 10-fold cross-validation. The goal is to classify patients with diabetes accurately. Through rigorous experimentation and analysis, the authors evaluate the performance of these algorithms and assess their effectiveness in predicting diabetes. The results of the study provide insights into the comparative performance of these data mining techniques for diabetes prediction and offer valuable guidance for future research in this domain.

In the paper titled “ Predictive analysis of diabetes using J48 algorithm of classification techniques in Contemporary Computing and Informatics” [3] by authors Pradeep and Dr. Naveen, the performance of various machine learning techniques is compared and measured based on their accuracy. The study specifically focuses on Decision Tree, Support Vector Machines (SVM), Logistic Regression, and Random Forest algorithms. The authors investigate how the accuracy of these techniques varies before and after pre-processing the data. By conducting a thorough analysis, they identify the impact of pre-processing on the performance of each algorithm. The results of this study shed light on the comparative effectiveness of these machine learning techniques.

In the paper titled “Review of Predictive Analysis Techniques for Analysis Diabetes Risk” [4] by authors Sonali Vyas, Rajeev Ranjan, Navdeep Singh, and Arohan Mathur presents a comparative performance evaluation of the Gradient Boosting Machine (GBM) model and the Logistic Regression model for data classification. The study measures the Area Under the Receiver Operating Characteristic curve (AROC) and sensitivity for both models using 10-fold cross-validation. The results indicate that the proposed GBM model achieved an AROC of 84.7% with a sensitivity of 71.6% for the dataset. On the other hand, the proposed Logistic Regression model achieved an AROC of 84.0% with a sensitivity of 73.4% for the same dataset. These findings provide insights into the performance and comparative effectiveness of the GBM and Logistic Regression models for data classification. The study contributes to the field of machine learning by showcasing the capabilities of these models in handling classification tasks. It can serve as a reference for researchers and practitioners in selecting suitable models for similar classification scenarios

In the paper titled “Prevalence and Early Prediction of Diabetes Using Machine Learning” [5] by authors Salliah Shafi and Gufran Ahmad Ansari, various machine learning algorithms are utilized and their accuracy is assessed. To enhance the precision of the results, a preprocessing technique is employed. Comparisons are made between the performance of different machine learning techniques and the effectiveness of the preprocessing approach. Among the evaluated machine learning techniques, the Naive Bayes method exhibits greater overall precision, particularly when coupled with the preprocessing technique. Additionally, the most commonly used type of Bayesian network, the Naive Bayesian network, demonstrates the highest accuracy values for the classification of diabetes and cardiovascular disease (CVD) at 99.51% and 97.92% respectively. These findings highlight the significance of preprocessing in improving precision and showcase the superiority of Naive Bayes and Naive Bayesian network in accurately classifying diabetes and CVD. The research serves as a valuable reference for researchers and practitioners seeking to employ machine learning algorithms and preprocessing techniques in healthcare applications.

2.2 Comparative Analysis of the Related Work

The table 2.1 discusses the comparative analysis of the current systems in light of the suggested proposal.

Table 2.1 Comparative Analysis

Sl. No	Author(s)	Algorithms/Techniques	Performance Measures
1.	Sonu Kumari and Archan Singh	Ensemble voting classifiers.	Accuracy
2.	Pradeep Kandhasamy, S. Balamurali	Data mining techniques, Decision Tree, K-Nearest Neighbors, Random Forest, SVM.	Accuracy
3.	Pradeep and Dr.Naveen	Decision tree SVM and Logistic Regression, Random forest.	Accuracy
4.	Sonali Vyas, Rajeev Ranjan, Navdeep Singh , Arohan Mathur	GBM model, Logistic Regression	Accuracy
5.	Salliah Shafi and Gufran Ahmad Ansari	Navies Bayes method.	Accuracy

2.3 Summary

The point of this project is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The sheer volume of data and various sources of noise and variability make analysis complex. Human mistakes or various laboratory test scan entangle the procedure of identification of the disease. By utilizing predictive algorithms, this model has the ability to accurately predict whether a patient is likely to have diabetes, enabling healthcare professionals to provide timely clinical intervention and potentially prevent loss of life. This technology has the potential to revolutionize the way we diagnose and manage diabetes, as it allows for early identification of at-risk individuals, leading to earlier interventions and improved outcomes.

Problem Formulation

3.1 Problem Statement

Doctors rely on common knowledge for treatment. When conventional knowledge falls short, medical studies are conducted after a certain number of cases have been reviewed. This traditional process can be time-consuming, whereas the use of machine learning can expedite the identification of patterns in medical data. However, machine learning techniques require a large amount of data, which may be limited depending on the disease. Additionally, the number of samples without the disease can be much greater than those with the disease, creating an imbalance in the data set.

Some prediction system doesn't meet the recommended accuracy level for medical purpose. Common peoples cannot access advanced prediction system at free of cost. Most of the prediction system don't have user friendly interface. Existing model doesn't have proper diet recommendation system to maintain a healthy lifestyle.

3.2 Objectives of the Present Study

The objectives of the proposed project are as follows:

1. To analyse and choosing the suitable algorithm for the early prediction of diabetes to the user.
2. To improve the overall accuracy of the model to get the accurate result.
3. To give a diet recommendation for to maintain a healthy lifestyle.
4. Hosting the web platform in which user can view and interact with the model.

3.3 Summary

The best solution for early detection of diabetes is using machine learning techniques. The classification algorithms are more accurate when compared to traditional imaging techniques. Developing an early diabetes prediction system can be useful for many doctors as well as patients by helping them be alert and take the required medications to prevent the diabetes.

Requirements and Methodology

4.1 Hardware Requirements

The hardware requirements for the proposed project are depicted in Table 4.1.

Table 4.1: Hardware Requirements

Sl. No	Hardware/Equipment	Specification
1.	Graphics Card	Intel 621 Graphics card or 2GB
2.	RAM	4GB or above

4.2 Software Requirements

The software requirements for the proposed project are depicted in Table 4.2.

Table 4.2: Software Requirements

Sl. No	Software	Specification
1.	Code Editor	VS Code 64 bit
2.	Python	Python 3 and above
3.	Framework	Streamlit

4.3 Methodology Used

The proposed Diabetes prediction system is implemented using the following steps:

Step 1 : Import the required libraries and diabetes dataset.

Step 2 : Preprocess the data to remove and split the data into training data and test data.

Step 3 : Try different machine learning algorithm to train the model.

Step 4 : Evaluate the model based on the accuracy of the result and choose the best model.

Step 5 : Implementing the user friendly interface along with diet recommendation for the user.

System Design

5.1 Architecture of the Proposed System

Figure 5.1 shows the architecture of the proposed system.

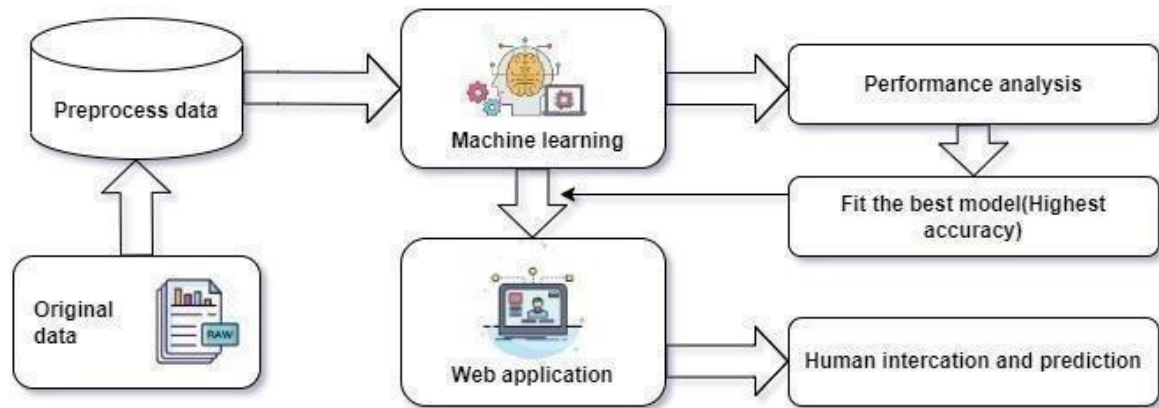


Figure 5.1: Architecture of the Proposed System

The architecture of diabetes prediction using machine learning is illustrated above. The original data is preprocessed and used for machine learning model. The performance is analyzed and the model with the best accuracy score is selected. After the evaluation of the model the web application is created using Streamlit framework for the same through which the user can interact and can get the result. Diet recommendation is also provided in this web application.

5.2 System Flowchart

A system flowchart is a way of depicting how data flows in a system and how decisions are made to control events. Figure 5.2 depicts the system flowchart.

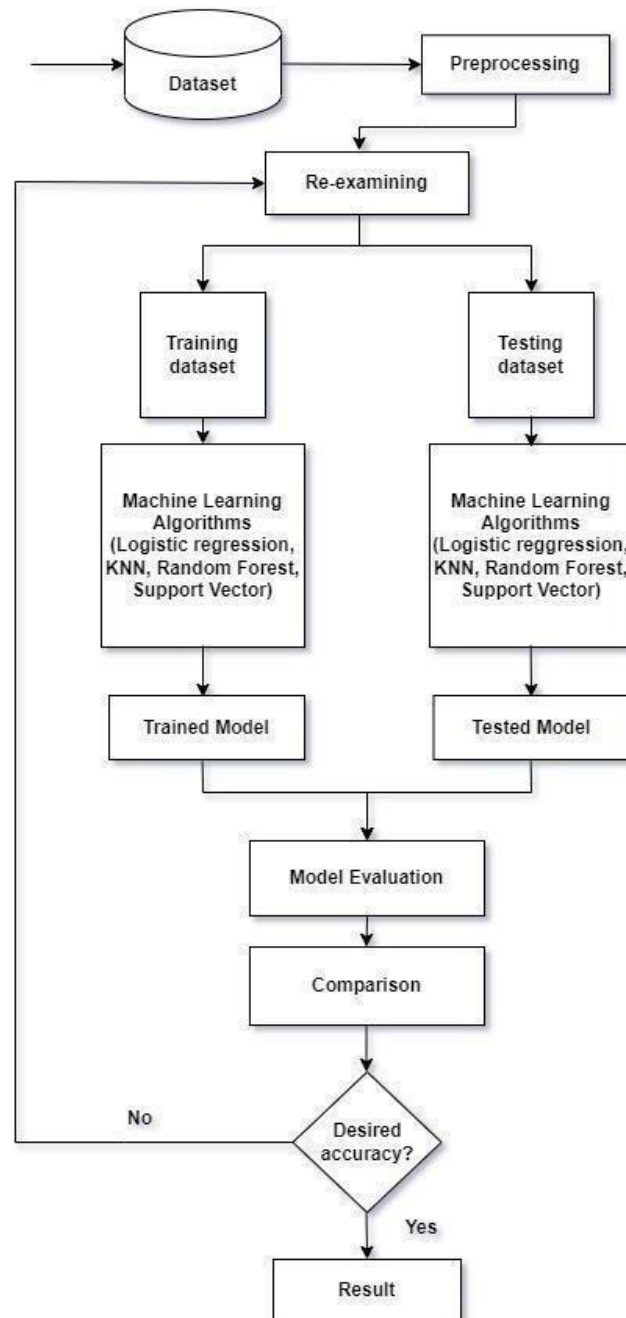


Figure 5.2: System Flowchart of Proposed System

At the initial stage the dataset is preprocessed, examined and splitted for training and testing purpose. Different algorithms are used for testing purpose and the model is evaluated. After evaluation process, if the accuracy is met then the result is predicted. Otherwise it is again re-examined.

Implementation

6.1 Pseudocode

Input: Training data

1. For loop: Iterate from 1 to k (a specified number of iterations or epochs). For each training data instance d_i .
2. For each training data instance d_i : This indicates that the following steps will be performed on each training data instance.
3. Initialize the weight of instance d_j : Initialize the weight of instance d_j to $P(1 | d_j) * (1 - P(1 | d_j))$. This step assigns an initial weight to the instance based on the probability of its class label being 1.
4. Finalize $f(j)$ for the data: This step is not clear from the provided code snippet. It mentions finalizing a function $f(j)$ using the class value (z_j) and weights (w_j). It's likely that this step involves creating a regression function or model using the data and its corresponding weights.
5. Classification Label Decision: Assign a class label to the instance based on the probability $P(1 | d_j)$. If $P(1 | d_j)$ is greater than 0.5, assign class label 1; otherwise, assign class label 2.
6. In summary, the code snippet combines regression and classification techniques to estimate a target value (z_i) based on the probability of the class label being 1 ($P(1 | d_j)$), and subsequently assigns class labels (1 or 2) based on a threshold of 0.5 for the probability.

System Testing, Results and Discussion

7.1 System Testing

Table 7.1: Unit Test Cases

Test case number	Input	Stage	Expected behavior	Observed behavior	Status P=Pass F=Fail
1	Login credentials	Login page	If details matched, the input page is shown	As expected	P
2	Registration of user	Input page	Storing user data in a file with hashed password	As expected	P
3	Enter input values from the test set	Input page	The result should appear either non diabetic and diabetic	As expected	P

7.2 Result Analysis

The main aim of the project was to predict the patient's Diabetic Status using machine learning algorithms. Table 7.2 shows the analysis that was performed on the four algorithms with different training and testing sizes. It was found that Logistic Regression was the most accurate in all the cases.

Table 7.2: Analysis of the Algorithms

Training Size	Testing Size	Accuracy (%)			
		Decision Tree	Random Forest	Logistic Regression	SVM
80%	20%	93	94	96	92
70%	30%	93	94	95	92

Figure 7.1 shows the bar graph for the feature importance for Logistic Regression model where the train set size was 80% and the test set size was 20%.

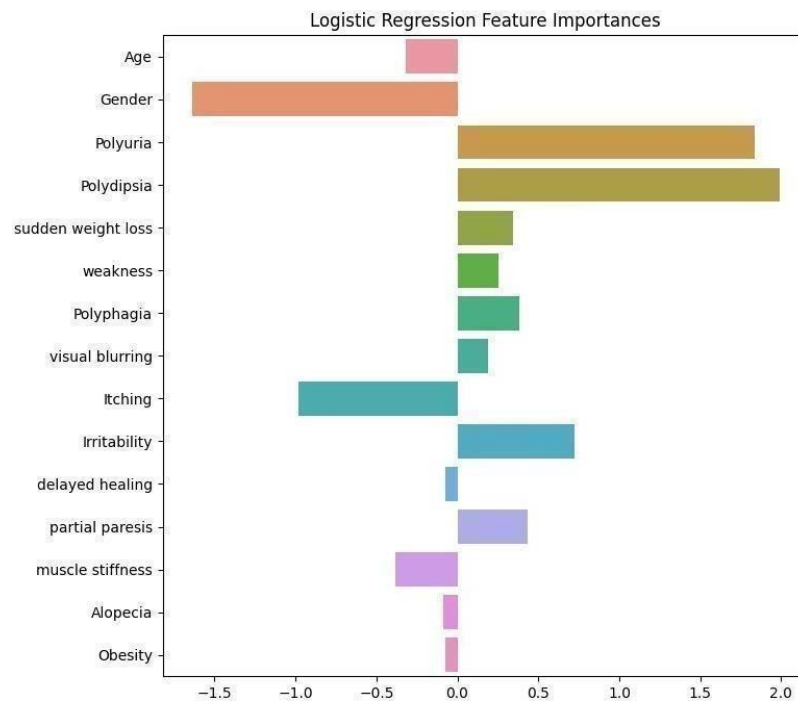


Figure 7.1: Feature Importance for Logistic Regression model

Figure 7.2 shows the bar graph for the feature importance for Random Forest model where the train set size was 70% and the test set size was 30%.

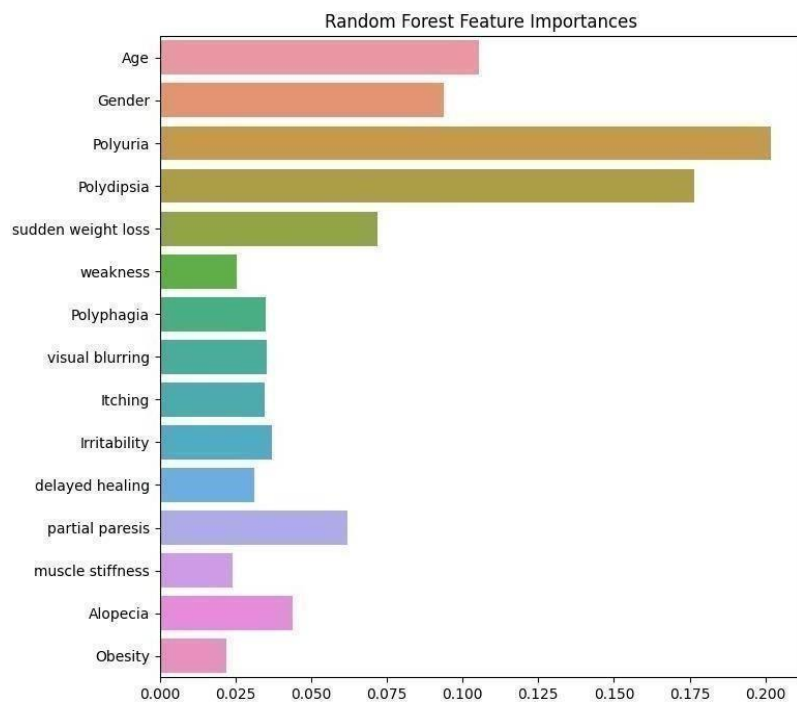


Figure 7.2: Feature Importance for Random Forest model

Figure 7.3 is the login page for the users who use this application.

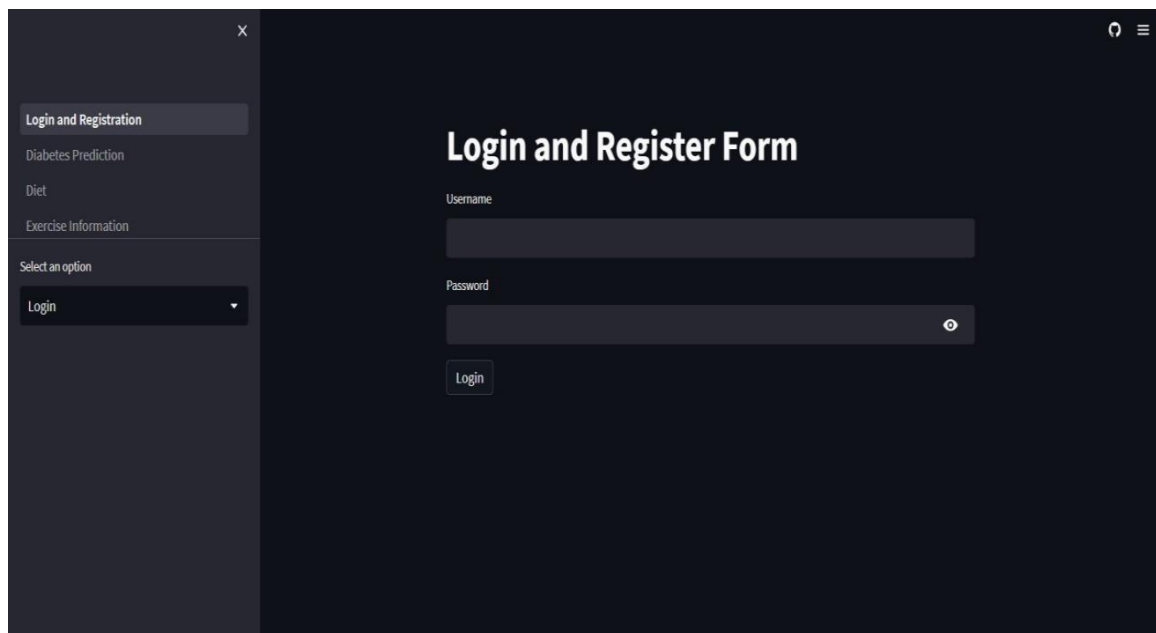
The screenshot shows a web application interface with a dark theme. On the left is a sidebar menu with a close button (X) at the top. The menu items are 'Login and Registration' (highlighted), 'Diabetes Prediction', 'Diet', and 'Exercise Information'. Below these is a section 'Select an option' with a dropdown menu currently showing 'Login'. The main content area has the title 'Login and Register Form' in white. Below the title are two input fields: 'Username' and 'Password'. The 'Password' field has an eye icon to toggle visibility. A 'Login' button is positioned below the password field. In the top right corner of the main area, there are icons for a search/magnifying glass and a hamburger menu.

Figure 7.3: Login Page

Figure 7.4 is the Home page of our application where we are displaying some basic information about Diabetes.

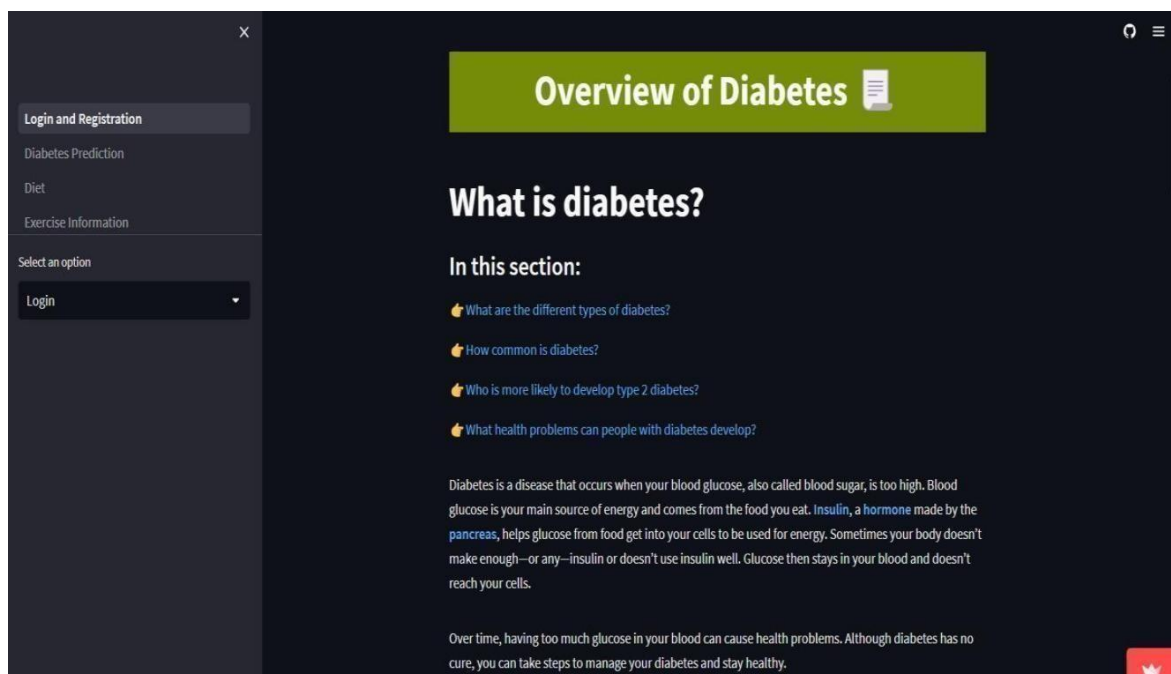
The screenshot shows the home page of the application. The sidebar menu on the left is identical to Figure 7.3, with 'Login and Registration' highlighted. The main content area features a green header bar with the text 'Overview of Diabetes' and a document icon. Below this is the section title 'What is diabetes?'. Underneath, it says 'In this section:' followed by a list of four topics, each preceded by a thumbs-up icon: 'What are the different types of diabetes?', 'How common is diabetes?', 'Who is more likely to develop type 2 diabetes?', and 'What health problems can people with diabetes develop?'. A paragraph of text follows, explaining that diabetes is a disease where blood glucose is too high, and mentioning insulin and the pancreas. At the bottom, another paragraph states that although there is no cure, steps can be taken to manage the condition. A small red heart icon is visible in the bottom right corner of the main content area.

Figure 7.4: Home Page

Figure 7.5 is the page where we take input from user to predict diabetes and the result is displayed.

Figure 7.5: Input Page

Figure 7.6 is the diet recommendation page for giving meal plan to the user.

MEAL TYPE	FOOD ITEM	QUANTITY	kcal
Breakfast	• Tea/Coffee (No Sugar)	1 cup	35
	• Idli with 1 cup sambhar OR	1 nos	121
	1 Masala dosa with ¼ cup sambhar/ 2		
	Idlis with ½ cup sambhar/ Rice cereal		
	flakes with 1 cup milk/ 1 cup Semolina	½ katori	135
	porridge/ 1 cup Wheat porridge/ ½ cup	1 cup	35
	Upma + 1 vada		
Mid Morning	• Any 1 fruits:		
	Musk melon /	100 gm	17
	Water melon /	100 gm	16
	Orange juice	200 gm	18

Figure 7.6: Diet Recommendation Page

Figure 7.7 is the Exercise information page to keep the diabetes in level.



Figure 7.7: Exercise Information Page

7.3 Summary

The application was developed using the Streamlit framework. The programming languages that were used were Python, HTML and CSS. The figures in the previous section showed the snapshots of various pages of the application. Since Random Forest was found to be the most accurate among the four algorithms, the prediction model was created using it.

Conclusion and Scope for Future Work

8.1 Conclusion

The project proposes a way of constructing an easier and simpler model for the prediction of stroke for an individual just based on some physical entities such as age, gender, work type, and many other. Proposed system tends out to be much more economical and less time consuming in predicting out whether an individual will be having stroke or not. The system developed will be simple and easy to use which makes it much more versatile. It forbids the complication of other projects and has higher accuracy comparatively, making it suitable for medical purpose.

8.2 Scope for Future Work

The project can be further improvised by considering real time datasets taken from the hospital using doctor's consent. This method further provides much scope for easier usage and much more economical when compared to the traditional medical methods. The project can be further utilized by providing a front-end window with a device separately for its prediction.

References

- [1] Kumari, S., & Singh, A. (2019). "A data mining approach for the diagnosis of diabetes mellitus". 2019 7th International Conference on Intelligent Systems and Control (ISCO). doi:10.1109/isco.2019.6481182.
- [2] Kandhasamy, J. P., & Balamurali, S. (2019). "Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*", 47, 45–51. doi:10.1016/j.procs.2019.03.182
- [3] Pradeep, K. R., & Naveen, N. C. (2020). "Predictive analysis of diabetes using J48 algorithm of classification techniques". 2020 2nd International Conference on Contemporary Computing and Informatics (IC3I).doi:10.1109/ic3i.2020.7917987
- [4] Vyas, S., Ranjan, R., Singh, N., & Mathur, A. (2020). "Review of Predictive Analysis Techniques for Analysis Diabetes Risk". 2020 Amity International Conference on Artificial Intelligence (AICAI). doi:10.1109/aicai.2020.8701236 .
- [5] Salliah Shafi and Gufran Ahmad Ansari, (2022). "Analysis of Diabetes mellitus using Machine Learning Techniques". INSPEC 22572071, DOI: 10.1109/IMPACT55510.2022.10029058.
- [6] R.Y.Toledo and L.Martinez, "A food recommender system considering nutritionalinformation and user preferences",2022.
- [7] M. Bernardini and E. Frontoni, "Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine" , 2023.

Personal Profile



Mr. Chaithanya D
Asst. Prof.
Project Guide

Mr. Chaithanya D received M.Tech degree from Shri Jayachamrajendra College of Engineering in 2013 and completed B.E in Information Science and Engineering from Vivekananda College Of Engineering and Technology in the year 2010.

He is currently working as Assistant Professor in the Department of Computer Science & Engineering at SDM Institute of Technology.

Area of Interest : Data Mining , Data Analytics.

Email ID : chaitanya@sdmit.in



Name: Rohan Ponnanna KK

USN: 4SU19CS076

Address: Akshaya nilaya, Koppa, Periyapatna.

E-mail ID: 19c03@sdmit.in

Contact Phone No: 7899062415



Name: Seetaram Naik

USN: 4SU19CS088

Address: Ankola, Uttara Kannada - 581314

E-mail ID: 19c16@sdmit.in

Contact Phone No: 7619645353



Name: Srinivas S

USN: 4SU19CS101

Address: P.O Peradala, Badiadka, Kasaragod.

E-mail ID : 19c27@sdmit.in

Contact Phone No : 6238219056



Name: Sulekha PB

USN: 4SU19CS103

Address : Chintamani Math, Amaravati,
Hospete.

E-mail ID : 4su19cs103@sdmit.in

Contact Phone No: 9390444592
