

T1421

by Turnitin Account

Submission date: 30-Mar-2024 01:52AM (UTC-0700)

Submission ID: 2335210112

File name: DOC-20240330-WA0002..pdf (861.87K)

Word count: 3996

Character count: 24286

CHAPTER-1

INTRODUCTION

Diabetes is a prevalent chronic metabolic disorder characterized by elevated blood sugar levels resulting from either insufficient insulin production or impaired insulin utilization by the body. With its incidence escalating globally, diabetes has emerged as a significant public health concern, imposing substantial burdens on healthcare systems and individuals alike. According to the World Health Organization (WHO), an estimated 422 million people worldwide were living with diabetes in 2014, with projections indicating a steady rise in prevalence.

The multifactorial nature of diabetes underscores the importance of early detection and intervention to mitigate its adverse health outcomes and complications. Timely diagnosis enables healthcare professionals to implement appropriate management strategies, including lifestyle modifications, pharmacotherapy, and regular monitoring, thereby improving patient outcomes and quality of life.

Machine learning, a subfield of artificial intelligence (AI) focused on developing algorithms capable of learning from data and making predictions or decisions, holds immense potential in revolutionizing healthcare delivery. In the context of diabetes, machine learning techniques offer a data-driven approach to predictive modeling, leveraging diverse patient characteristics and biomarkers to forecast disease risk or progression.

Attributes such as BMI, hobbies, lifestyles, and family history were not considered in the featured work. Incorporating these attributes into data collection can improve the accuracy of predictions and their applicability.

1.1 SCOPE OF THE CAPSTONE PROJECT

1.1.1 PROBLEM STATEMENT

- Diabetes is a chronic metabolic disorder characterized by high blood sugar levels over a prolonged period. It poses a significant global health burden, affecting millions of individuals worldwide. Early detection and management of diabetes are crucial in preventing its complications and improving patient outcomes.
- Machine learning (ML) techniques offer promising avenues for predicting diabetes risk based on various clinical and demographic features. However, developing accurate and reliable predictive models requires addressing several challenges and considerations.

1.1.2 EXISTING SYSTEM

- Researchers have explored data mining techniques like clustering and classification for diabetes detection in existing studies.
- Proposals have been made to leverage algorithms such as logistic regression for prediction, with comparisons conducted using a standard diabetic dataset.

1.1.3 PROPOSED SYSTEM

- This study proposes a binary classification system for diabetes detection using a local dataset.
- Machine learning algorithms, including Random Forest, Logistic Regression, and Support Vector Machines (SVM), will be employed for initial classification.
- The system aims to further enhance prediction accuracy through the integration of deep learning techniques.

1.1.4 OBJECTIVES

- To gather and preprocess diverse health data to establish a comprehensive foundation for model development.
- To develop the diabetes prediction model, rigorously validate its accuracy using various datasets, and refine it for optimal performance.
- To compare various machine learning models and analyze its results.

1.1.5 PURPOSE

- The project is to proactively identify individuals at risk of developing diabetes.
- This allows for early intervention and targeted preventive measures, leading to improved health outcomes, cost-effective resource allocation management.

CHAPTER-2

PROJECT PLANNING

This chapter describes the work breakdown structure, 2 Timeline Development Schedule, Cost Breakdown Structure, Cost Analysis, Risks Assessment.

2.1 Work Breakdown Structure diagram

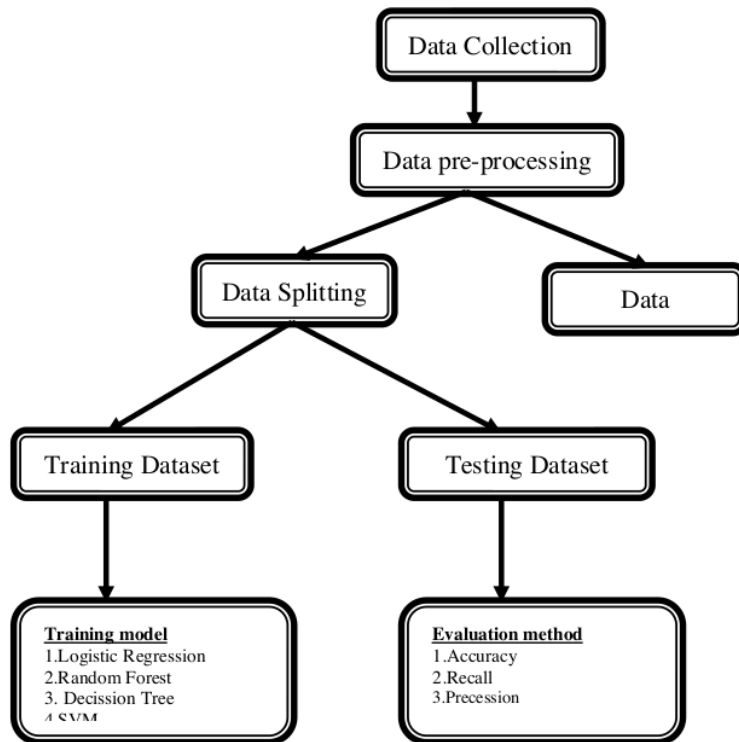


Fig 2.1 Work Breakdown Structure

2.2 Timeline Development – Schedule

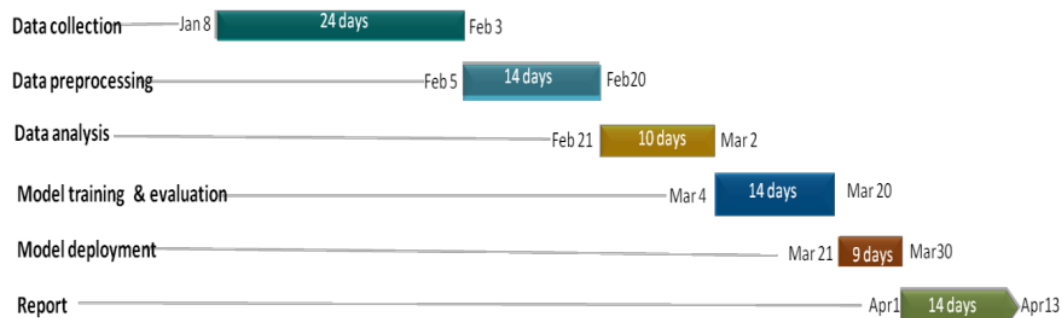


Fig2.2 Timeline Development – Schedule

2.3 Cost Analysis

Data Collection Costs	• Subscription fee for access to healthcare databases	₹2000
	• Costs associated with obtaining additional medical records	₹1000
Infrastructure Costs	• Purchase of high-performance computing hardware	₹50000
Personnel Costs	• Faculty advisor (supervision and guidance), Student researchers	₹0
Software Costs	• purchase of necessary software licenses	₹0
Training and Education Costs	• Workshops or seminars on machine learning techniques	₹2000
	• Machine learning training course for team members	₹2000
Testing and Validation Costs	• Printing costs for presentation materials	₹1000
	• Cost for acquiring additional labeled datasets for validation	₹2000
Deployment Costs	• Website hosting for project documentation or presentation	₹3000
Miscellaneous Costs	• Travel expenses	₹2000

8

2.4 Risks Assessment

A project risk assessment is a process that aims to gain a deeper understanding of which project tasks, deliverables, or events could influence its success.

1. Data Quality and Privacy

Risk Description	Likelihood	Impact	Mitigation Strategy
Incomplete or inaccurate data	High	High	Implement rigorous data cleaning and validation procedures.
Privacy concerns with health data	Medium	Medium	Anonymize and encrypt sensitive information, adhere to regulations.

2. Model Development and Validation

Risk Description	Likelihood	Impact	Mitigation Strategy
Model over fitting or under fitting	Medium	High	Regularize the model, use cross-validation, and monitor performance metrics.
Lack of interpretability of the model	High	Medium	Utilize interpretable models or provide model interpretation through explainability techniques

3. Ethical and Bias Considerations

Risk Description	Likelihood	Impact	Mitigation Strategy
Bias in the model predictions	Medium	High	Implement fairness-aware algorithms, conduct bias assessments regularly.
Lack of diversity in the training data	Low	Medium	Ensure diverse representation in the dataset and address data imbalance.

4. Model Deployment and Integration

Risk Description	Likelihood	Impact	Mitigation Strategy
Technical issues during model deployment	Medium		Conduct thorough testing in staging environments, and implement rollback plans.
Resistance or lack of user adoption	Low	Medium	Engage end-users early, provide training, and communicate the benefits clearly

5. Compliance and Regulatory

Risk Description	Likelihood	Impact	Mitigation Strategy
Failure to comply with healthcare regulations	Medium	High	Stay informed about regulations, involve legal experts, and adhere to standards.

2.5 Requirements Specification

2.5.1 Functional Requirements

Functional requirements in machine learning projects outline the specific functionalities, capabilities, and behaviors of the machine learning model and its associated components. These requirements focus on the tasks and processes related to data preparation, model training, evaluation, deployment, and maintenance. Here's a breakdown of functional requirements in machine learning.

1. Input Interface

- The system should provide a user interface for users to input their age, gender, fasting plasma glucose (FPG), cholesterol levels (Chol), triglycerides (Tri), ²⁴ high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, alanine aminotransferase (ALT) levels, and creatinine levels.
- Gender should be represented as binary values (0 for female, 1 for male).

2. Prediction Functionality

- The system should use machine learning algorithms ¹⁷ to predict whether the user has diabetes or not based on the input features provided.
- Predictions should be made in real-time and displayed to the user.

3. Output Display

- The system should display the prediction result to the user, indicating whether they are likely to have diabetes or not.
- Additionally, the system should provide some form of explanation or visualization to help users understand the prediction.

4. Accuracy Evaluation

- The system should include functionality ¹ to evaluate the accuracy of the prediction model using appropriate metrics such as accuracy, precision, recall, and F1-score.
- ⁶ This evaluation should be accessible to administrators or developers for model refinement purposes.

2.5.2 Non-Functional Requirements

Non-functional requirements in machine learning projects encompass aspects of system behavior and performance, emphasizing qualities like reliability, scalability, performance, security, and usability. These requirements are crucial for ensuring that deployed models meet desired standards and effectively address user and stakeholder needs. They contribute to the overall quality and behavior of the system, ensuring robustness,

trustworthiness, and sustainability. Here's a breakdown of non-functional requirements in machine learning.

1. Performance

- The system should respond to user inputs and provide predictions within a reasonable time frame, ideally within a few seconds.

2. Security

- User data, including sensitive health information, should be securely handled with data protection regulations.

3. Scalability

- It should be able to scale horizontally by adding more server instances if needed.

4. Reliability

- The system should be highly reliable, with minimal downtime or disruptions to service.

5. Usability

- The user interface should be intuitive and easy to use.

6. Compatibility

- The system should be compatible with web browsers to ensure accessibility for users.

7. Maintainability

- The codebase should be well-organized, documented, and modular to facilitate ease of maintenance and future updates.
- Version control should be utilized to track changes and collaborate on development efforts effectively

2.5.3 User Input Requirements

Age

- Users should input their age in years.
- Age should be a numeric value within a specified range (e.g., 18-100 years).

Gender

- Users should specify their gender.
- Gender should be represented as a binary value: 0 for female and 1 for male.

Fasting Plasma Glucose (FPG)

- Users should input their FPG level in milligrams per deciliter (mg/dL).
- FPG should be a numeric value within a specified range (e.g., 70-300 mg/dL).

Cholesterol Levels (Chol)

- Users should input their total cholesterol level in milligrams per deciliter (mg/dL).
- Cholesterol levels should be a numeric value within a specified range (e.g., 100-400 mg/dL).

Triglycerides (Tri)

- Users should input their triglyceride level in milligrams per deciliter (mg/dL).
- Triglyceride levels should be a numeric value within a specified range (e.g., 50-500 mg/dL).

High-Density Lipoprotein (HDL) Cholesterol

- Users should input their HDL cholesterol level in milligrams per deciliter (mg/dL).
- HDL cholesterol levels should be a numeric value within a specified range (e.g., 20-100 mg/dL).

Low-Density Lipoprotein (LDL) Cholesterol

- Users should input their LDL cholesterol level in milligrams per deciliter (mg/dL).
- LDL cholesterol levels should be a numeric value within a specified range (e.g., 50-200 mg/dL).

Alanine Aminotransferase (ALT) Levels

- Users should input their ALT level in international units per liter (IU/L).
- ALT levels should be a numeric value within a specified range (e.g., 5-100 IU/L).

Creatinine Levels

- Users should input their creatinine level in milligrams per deciliter (mg/dL).
- Creatinine levels should be a numeric value a specified range (e.g. 0.5-2.5 mg/dL).

2.5.4 Technical Constraints

Hardware Requirements

Processor : Any Processor above 500 MHz

RAM : 4GB

Hard Disk : 500GB

Input device : Standard keyboard and mouse.

Output device monitor : VGA with High Resolution

Software Requirements

Operating System : Windows 10

Programming : Python 3.12 and related libraries

IDE : Jupyter Notebook

2.6 Literature Survey

33	Title	Authors	Methods	Results	Dataset															
	Diabetes prediction using machine learning and explainable AI techniques	Isfauzzaman Tasin, Riasat Khan	XGBoost Classifier	Accuracy:- 88.5 Precision:- 0.81	Pima Indian dataset															
	Diabetes prediction using machine learning algorithms	M. Ramakrishna Murthy, M. Manoj Kumar,	38 Random Forest, Decision tree and Naïve Bayes,	Random forest Accuracy:-84.4 Precision:- 50.8 Decision tree Accuracy:-77.9 Precision:- 31.4 Naïve Bayes Accuracy:-76.3 Precision:- 34.2	Diabetes Health Indicators Dataset															
26	Diabetes prediction using Machine Learning algorithms and ontology	Hakim El Massari, Zineb Sabouri	49 VM, Logistic Regression, Naive Bayes, Decision Tree,	<table><tr><th>Model</th><th>ACC</th><th>PRE</th></tr><tr><td>SVM</td><td>77</td><td>78</td></tr><tr><td>LR</td><td>77</td><td>79</td></tr><tr><td>NB</td><td>76</td><td>80</td></tr><tr><td>DT</td><td>73</td><td>79</td></tr></table>	Model	ACC	PRE	SVM	77	78	LR	77	79	NB	76	80	DT	73	79	Pima Indian dataset
Model	ACC	PRE																		
SVM	77	78																		
LR	77	79																		
NB	76	80																		
DT	73	79																		
27	Machine Learning based Diabetes Prediction and Development of Smart Web Application	Nazin Ahmmeda , Rayhan Ahammeda	17 Random Forest (RF),Support Vector Machines (SVM),Logistic regression (LR),	<table><tr><th></th><th colspan="2">ACCURACY</th></tr><tr><th>Model</th><th>D1</th><th>D2</th></tr><tr><td>RF</td><td>80.26</td><td>96.18</td></tr><tr><td>SVM</td><td>80.26</td><td>91.49</td></tr><tr><td>LR</td><td>77.63</td><td>84.04</td></tr></table>		ACCURACY		Model	D1	D2	RF	80.26	96.18	SVM	80.26	91.49	LR	77.63	84.04	Pima Indian dataset
	ACCURACY																			
Model	D1	D2																		
RF	80.26	96.18																		
SVM	80.26	91.49																		
LR	77.63	84.04																		

2.7 System Design

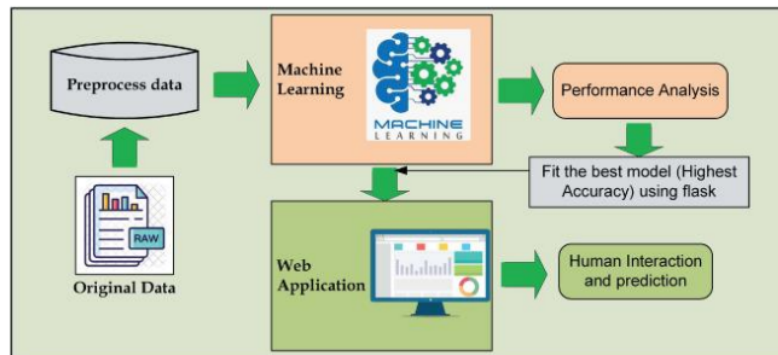


Fig2.7 System Design

1. Data Acquisition and Preprocessing

- **Data Source:** Utilize a credible medical dataset on diabetes, such as ³⁹ Pima Indians

Diabetes Dataset from UCI Machine Learning Repository.

32

- **Data Preprocessing:** Clean the data by handling missing values, outliers, and scaling numerical features for better model performance.

2. Feature Engineering

- Analyze the data to identify potential relationships between existing features and diabetes.
- Consider creating new features based on domain knowledge, such as body mass index (BMI) from weight and height.

3. Model Selection and Training

Choose a machine learning algorithm suitable for binary classification, such as:

- **Logistic Regression:** Effective baseline model for classification problems.
- **Random Forest:** Handles mixed-type data and avoids overfitting.
- **Support Vector Machines (SVM):** Powerful for high-dimensional data and clear class separation.
- Train the model on a portion of the preprocessed data, optimizing hyper parameters for best performance.

4. Model Evaluation

23

- Evaluate the trained model on a separate test set using metrics like accuracy, precision, recall, and F1 score.
- Consider using techniques like cross-validation to ensure generalizability of the model.

37

5. Deployment and User Interface (Optional):

- Develop a user-friendly interface where users can input relevant data points like age, weight, blood sugar levels, etc.
- The model predicts the likelihood of diabetes based on the input data.

6. Additional Considerations

- **Explain ability:** Explore Explainable AI (XAI) techniques to understand how the model arrives at its predictions, increasing user trust.
- **Privacy:** Ensure user data privacy by anonymizing sensitive information and following data security best practices.

2.8 Implementation

Random Forest Algorithm

Random Forest is a machine learning algorithm that belongs to the ensemble learning methods. It operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each decision tree in the forest is built using a subset of the training data and a subset of the features.

Here's how the algorithm works:

1. Random Sampling with Replacement (Bootstrap)

- Random Forest samples the training data with replacement, a process known as bootstrapping.
- This means that for each tree constructed, a subset of the original dataset is randomly chosen, and some samples may appear multiple times, while others may not appear at all.

2. Decision Tree Construction

For each tree in the forest:

- Randomly select a subset of features from the total features available.
- Choose the best split among the selected features to partition the data.
- Split the data into child nodes based on the chosen split.
- Repeat the process recursively for each child node until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf.

3. Voting (Classification) or Averaging (Regression)

After constructing multiple decision trees, when making predictions:

- For classification, each tree "votes" for the class label, and the mode of all votes is taken as the final prediction.
- For regression, each tree makes a prediction, and the average of all predictions is taken as the final prediction.

4. Hyper parameters:

- Random Forest has several hyperparameters that can be tuned for optimal performance, such as:
 - **n_estimators**: The number of trees in the forest.
 - **max_depth**: The maximum depth of each decision tree.
 - **min_samples_split**: The minimum number of samples required to split an internal

node.

- **max_features**: The number of features to consider when looking for the best split.

And many more, each affecting the behavior and performance of the model.

5. Out-of-Bag (OOB) Score

- Random Forest can optionally compute an out-of-bag score, which estimates the generalization performance of the model using the samples not included in the bootstrap sample used to train each tree.

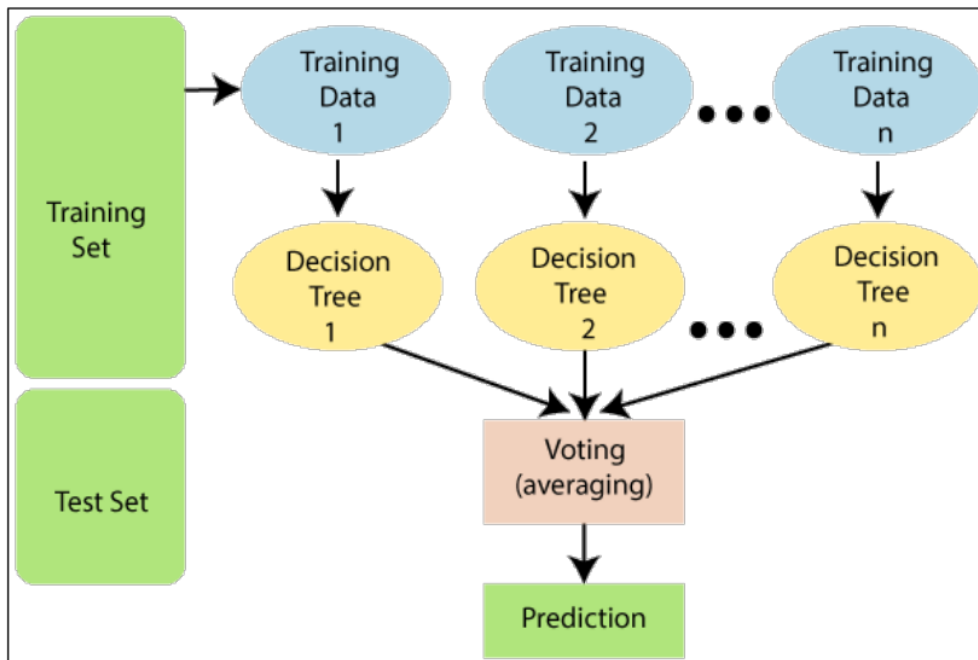


Fig2.8.1 The working of the Random Forest Algorithm

CHAPTER- 3

3.1 METHODOLOGIES

Machine learning methodologies are different approaches to train algorithms to learn from data. In short, they're the tools and techniques used to build intelligent systems. These methodologies can be broadly categorized by how data is provided.

- **Data Acquisition and Preprocessing**

- **Data Sources:** Electronic health records, population-based surveys, and wearable device data can provide valuable information on demographics, medical history, lifestyle habits, and lab test results.
- **Data Cleaning:** Incomplete or inaccurate data can hinder model performance. Techniques like missing value imputation and outlier removal ensure data quality.

- **Feature Selection and Engineering**

- **Feature Selection:** Identifying the most relevant factors influencing diabetes onset is vital. Statistical methods or correlation analysis can help pinpoint these features.
- **Feature Engineering:** Deriving new features from existing data can improve model accuracy. For instance, calculating body mass index (BMI) from weight and height.

- **Machine Learning Algorithms**

- **Logistic Regression:** This linear model estimates the probability of an individual developing diabetes based on various factors. It's interpretable, making it easy to understand the impact of each feature.
- **Support Vector Machines (SVM):** SVMs create a hyper plane that best separates data points belonging to different classes (diabetic and non-diabetic). They can handle complex relationships between features.
- **Decision Trees:** These tree-like structures classify data points based on a series of sequential decisions. They offer clear decision-making logic but can be prone to overfitting.
- **Random Forests:** Ensembles of decision trees, where each tree votes on the classification. This reduces over fitting and improves modelgeneralizability.
- **Artificial Neural Networks (ANNs):** Inspired by the human brain, ANNs learn complex patterns from data through interconnected layers. They excel at handling

non-linear relationships but can be computationally expensive to train.

- **Model Evaluation and Refinement**

- **Metrics:** Performance is assessed using metrics like accuracy, precision, recall, and F1-score. These metrics evaluate how well the model identifies individuals with and without diabetes.
- **Cross-validation:** The data is split into training and testing sets. The model is trained on the training data and evaluated on the unseen testing data to ensure generalizability.
- **Hyper parameter Tuning:** Adjusting the parameters of the chosen algorithm can significantly impact performance. Techniques like grid search or randomized search help optimize these parameters.

- **Ethical Considerations in Diabetes Prediction with ML**

- **Bias:** Datasets may harbor biases that can lead to discriminatory predictions. Mitigating bias through data selection and using fairness-aware algorithms is crucial.
- **Interpretability:** While some ML models are "black boxes," techniques like feature importance analysis can shed light on their decision-making process.
- **Privacy:** Protecting patient privacy is paramount. Secure data handling practices and anonymization techniques are essential.

3.1.1 Use cases

1. **Early Diagnosis and Prevention**

Machine learning models can analyze a patient's medical history, lifestyle factors, and genetic predispositions to identify individuals at high risk of developing diabetes. Early detection allows for timely interventions such as lifestyle modifications, diet changes, and medication to prevent or delay the onset of diabetes.

2. **Personalized Treatment Plans**

By analyzing patient data, including blood glucose levels, insulin sensitivity, and response to medication, machine learning models can assist healthcare providers in developing personalized treatment plans for diabetic patients. These plans can optimize medication dosage, timing, and types of treatment to better manage blood sugar levels and reduce the risk of complications.

3. **Remote Patient Monitoring**

Machine learning algorithms can be integrated into wearable devices or mobile applications to continuously monitor vital signs and glucose levels in diabetic patients.

Real-time data analysis can alert patients and healthcare providers to potential fluctuations in blood sugar levels, allowing for timely interventions and adjustments to treatment plans.

4. Clinical Decision Support Systems

Machine learning models can aid healthcare providers in clinical decision-making by analyzing patient data and providing recommendations for diagnosis, treatment, and management of diabetes. These decision support systems can assist healthcare providers in interpreting complex patient data and identifying the most effective interventions for individual patients.

5. Population Health Management

Machine learning algorithms can analyze large-scale healthcare data, including electronic health records, insurance claims, and demographic information, to identify high-risk populations and geographic areas for targeted diabetes prevention and management programs. By identifying populations at increased risk of diabetes, healthcare organizations can allocate resources more effectively and implement preventive interventions to reduce the burden of diabetes on the healthcare system.

6. Drug Discovery and Development

Machine learning models can analyze biological data, including genetic, proteomic, and metabolic information, to identify novel drug targets and biomarkers for diabetes. By uncovering underlying molecular mechanisms of diabetes and related complications, machine learning can accelerate the discovery and development of new therapeutic interventions for the treatment and prevention of diabetes.

7. Predictive Analytics for Complications

Machine learning models can predict the likelihood of diabetic complications such as diabetic retinopathy, nephropathy, neuropathy, and cardiovascular disease based on patient data, including medical history, lab results, and imaging studies.

CHAPTER- 4

TEST AND VALIDATION

4.1 Test plan

- **Objectives and Scope:** Define clear objectives, such as accurately predicting diabetes, and specify the scope, including features, evaluation metrics, and constraints.
- **Data Preparation:** Describe the dataset, preprocessing steps (e.g., normalization, handling missing values), and ensure its representativeness.
- **Model Evaluation Metrics:** Define relevant metrics like accuracy, precision, recall, F1-score, and choose based on project requirements.
- **Test Cases:** Develop diverse test cases, including edge cases and challenges like imbalanced classes or noisy data.
- **Cross-Validation Strategy:** Determine cross-validation strategy (e.g., k-fold), considering dataset size and resources.
- **Model Training and Validation:** Describe model training, hyper parameter tuning, validation strategy, and techniques to prevent overfitting.
- **Test Set Evaluation:** Define how the final model will be evaluated using the test set, ensuring consistency with validation set results.
- **Deployment Considerations:** Discuss considerations for deploying the model in production, including scalability and ongoing monitoring.
- **Documentation and Reporting:** Thoroughly document the test plan and generate clear reports summarizing results and insights.
- **Review and Iteration:** Review the test plan with stakeholders, gather feedback, and iterate as necessary to ensure alignment with project goals.

4.2 Test Case

S.no	Test Case	Description	Input	Expected Output	Actual Output	Result
01	Valid Input	Test the application with valid input	Age=18, Gender=1, FPG=140, Cholesterol=150, Triglycerides=120, HDL= 40, LDL=100, ALT=100, CCR= 0.70	Great! You DON'T have DIABETES	Great! You DON'T have diabetes	Pass
02	Valid Input	Test the application with valid input	Age=55, Gender=0, FPG=290, Cholesterol=190, Triglycerides=122, HDL= 47, LDL=120, ALT=105, CCR= 0.89	Oops! You have DIABETES	Oops! You have DIABETES	Pass
03	Incomplete input	Test the application's response with incomplete input	Age=60, Gender=0, FPG=290, Cholesterol=190, Triglycerides=122, HDL= 47, LDL=Empty, ALT=105, CCR= 0.89	Invalid input. Please fill in the form with appropriate values	Invalid input. Please fill in the form with appropriate values	Pass
04	Invalid Input for gender	Test the application's response with invalid input for gender	Gender =3	Invalid input. Please fill in the form with appropriate values	Invalid input. Please fill in the form with appropriate values	Pass

CHAPTER- 5

RESULT AND OUTPUT

5.1 Trained Models and its Accuracy

Sl No.	Model Name	Accuracy
1.	Random Forest Classifier/Gradient Boosting	90.77%
2.	Logistic Regression	90.00%
3.	Gaussian Process Classification (GPC)	89.23%
4.	Extra Trees	87.69%
5.	XGBoost	87.69%
6.	LightGBM	87.69%
7.	MLPClassifier	86.92%
8.	Support Vector Machine (SVM)	86.92%
9.	Decision Tree Classifier	81.53%

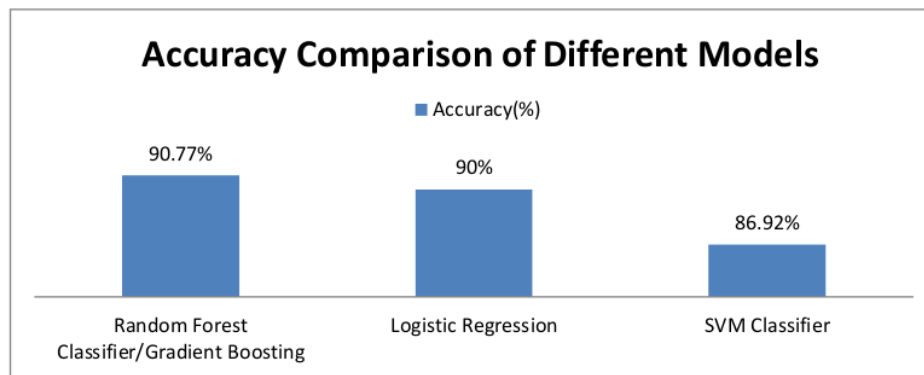


Fig5.1 Accuracy Comparison of Different Models

5.2 Random Forest Performance Analysis:-

Classification Report Graph:

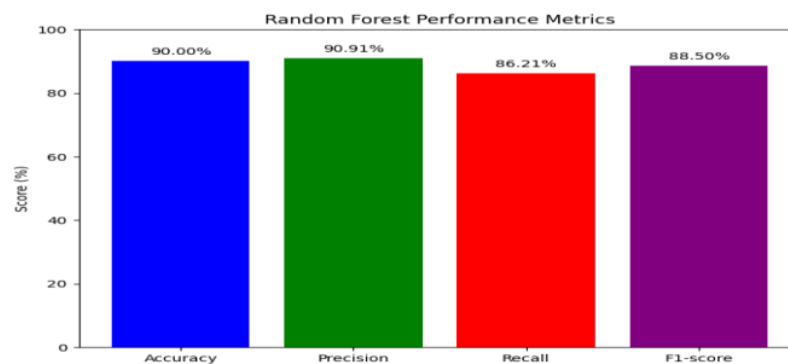


Fig 5.2 Classification Report Graph

5.3 Confusion matrix:

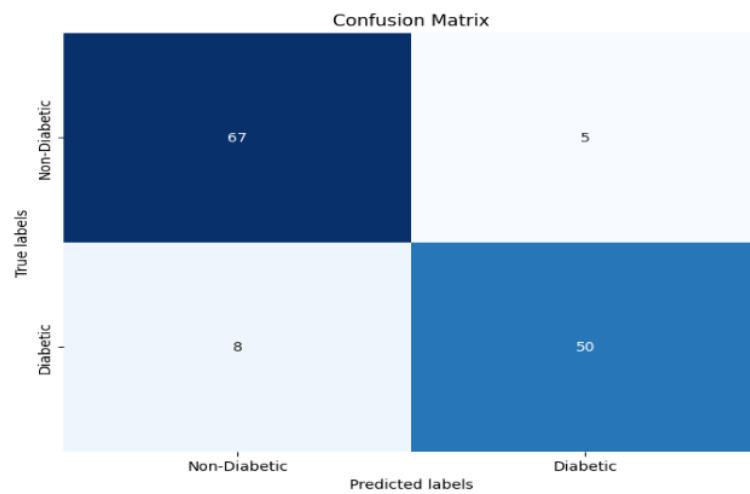


Fig5.3 Confusion matrix

5.4 Outputs

5.4.1 Home Page

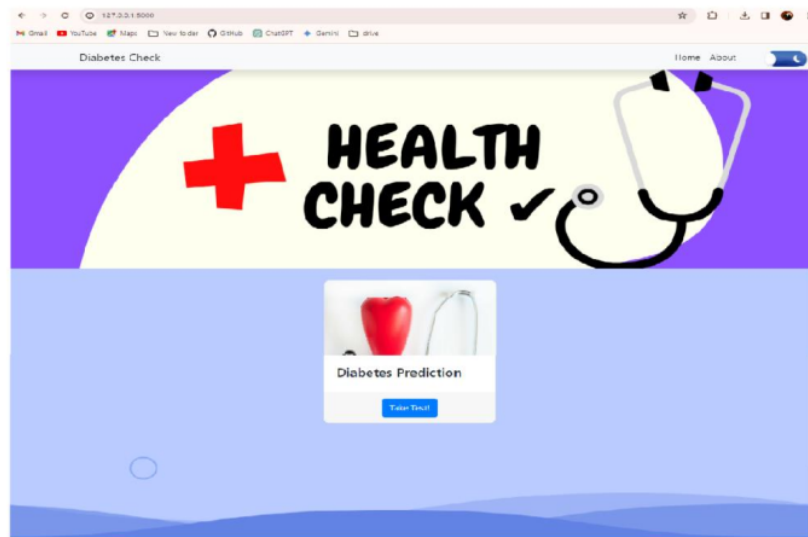


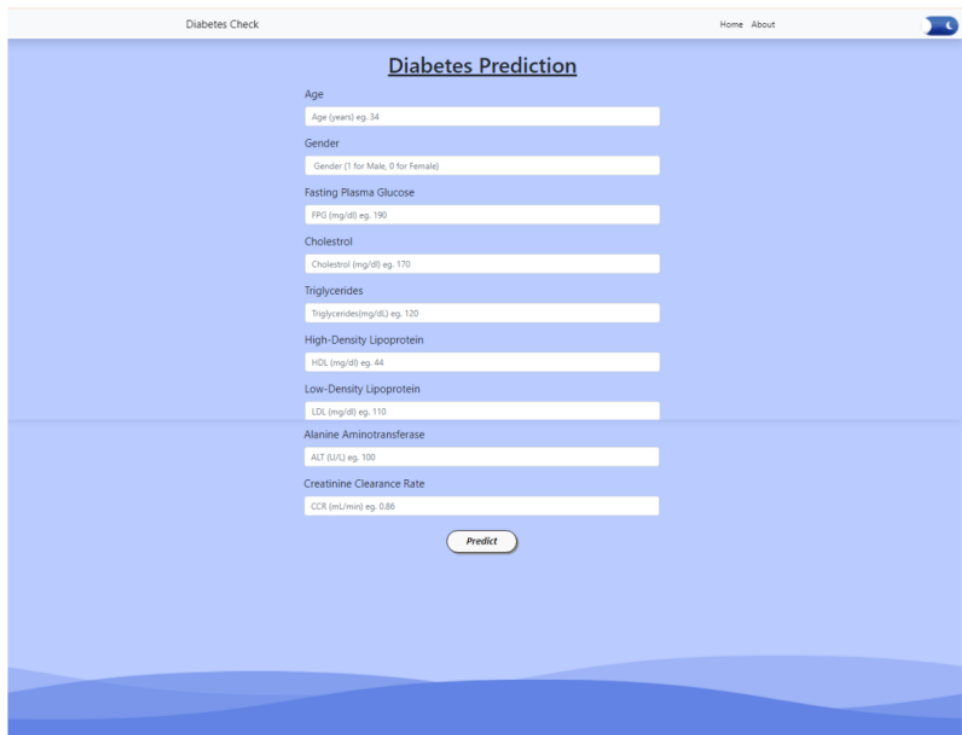
Fig5.4.1 Home Page

5.4.2 About Page



Fig5.4.2 About Page

5.4.3 Input Page



The screenshot shows a web application titled "Diabetes Check" with a navigation bar containing "Home" and "About" links. The main heading is "Diabetes Prediction". Below this, there are input fields for various medical parameters, each with a placeholder example value:

- Age: Age (years) eg. 34
- Gender: Gender (1 for Male, 0 for Female)
- Fasting Plasma Glucose: FPG (mg/dl) eg. 190
- Cholesterol: Cholesterol (mg/dl) eg. 170
- Triglycerides: Triglycerides(mg/dl) eg. 120
- High-Density Lipoprotein: HDL (mg/dl) eg. 44
- Low-Density Lipoprotein: LDL (mg/dl) eg. 110
- Alanine Aminotransferase: ALT (U/L) eg. 100
- Creatinine Clearance Rate: CCR (mL/min) eg. 0.86

A "Predict" button is located at the bottom of the input section. The background features a blue gradient with wavy lines at the bottom.

Fig5.4.3 Input Page

5.4.4 Negative Output

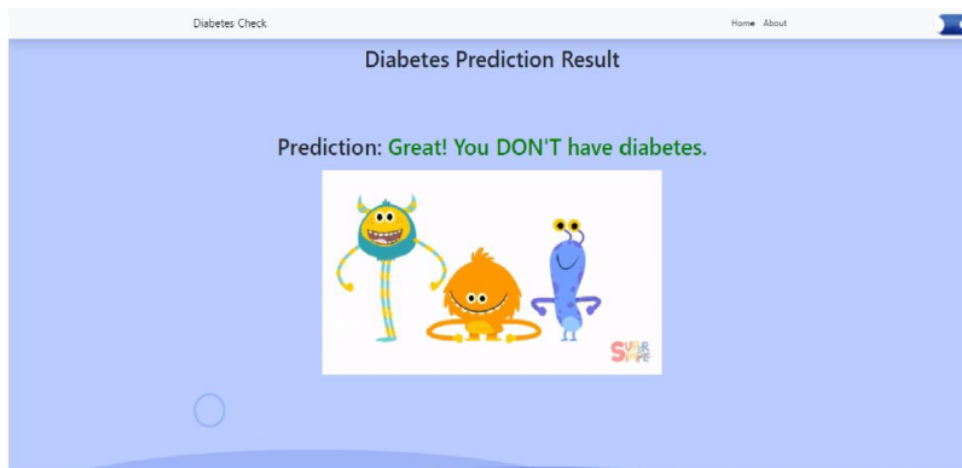


Fig5.4.4 Negative Output

5.4.5 Positive Output

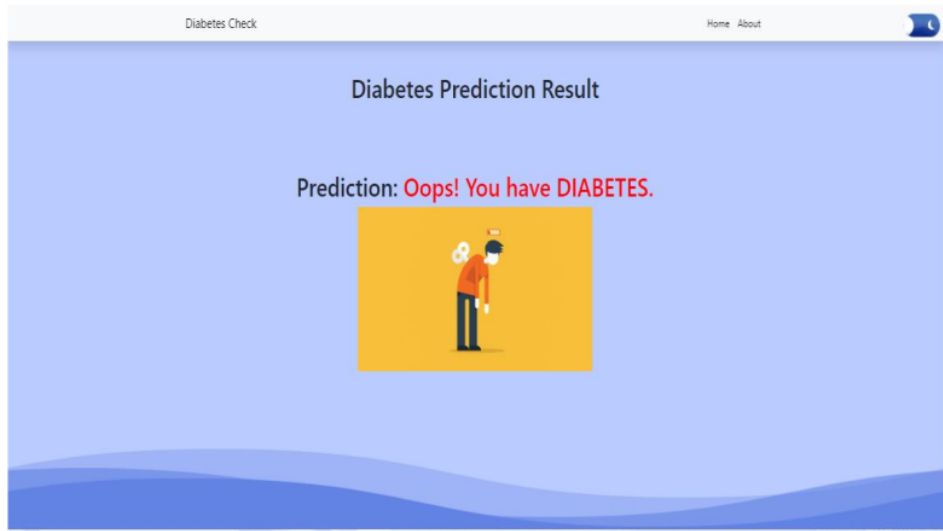


Fig5.4.5 Positive Output

5.5 Conclusion

In essence, this project investigated the effectiveness of machine learning models in diabetes classification. By applying random forest algorithms and logistic regression to a pre-processed dataset, the project achieved impressive accuracy of around 90% for both. This achievement indicates the promise of developing efficient and non-invasive diagnostic tools for diabetes prediction using a limited set of readily available medical attributes. This underscores the significant contribution machine learning can make to preventative healthcare by enabling the early identification of individuals at risk. Future efforts could explore leveraging deep learning techniques to uncover even more intricate patterns within the data. Additionally, incorporating a wider range of patient data, such as lifestyle habits, body mass index, and family history, has the potential to further refine diagnostic accuracy and personalize risk assessments. By pursuing these advancements, this project's findings lay the groundwork for the development of robust and informative clinical decision-making tools, empowering both patients and healthcare professionals in the fight against diabetes.

5.6 References

- <https://www.google.com/>
- <https://chat.openai.com>
- <https://bard.google.com/chat>
- <https://www.kaggle.com>
- <https://www.w3schools.com/>
- <https://github.com/>
- <https://sci-hub.se/>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/>
- https://www.researchgate.net/publication/343536932_A_Review_of_Diabetic_Prediction_Using_Machine_Learning_Techniques
- <https://sciencescholar.us/journal/index.php/ijhs/article/download/9617/6620>
- <https://www.amazon.in/Prediction-Diabetes-using-Vague-Theory/dp/3330074566>
- <https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction- using-machine-learning/>

ORIGINALITY REPORT

20%

SIMILARITY INDEX

16%

INTERNET SOURCES

9%

PUBLICATIONS

13%

STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	1 %
2	123dok.org Internet Source	1 %
3	Submitted to Intercollege Student Paper	1 %
4	Submitted to University of Wales Institute, Cardiff Student Paper	1 %
5	Submitted to Higher Education Commission Pakistan Student Paper	1 %
6	www.mdpi.com Internet Source	1 %
7	Submitted to Monmouth University Student Paper	1 %
8	Submitted to University of Central Lancashire Student Paper	1 %
9	Submitted to Flinders University	

1 %

10

Hassen Khelifi, Lefi Abdellaoui, Wassila Bouzid Sai. "Chapter 24 Prediction of Cutting Force and Surface Roughness in Turning Using Machine Learning", Springer Science and Business Media LLC, 2024

Publication

1 %

11

Submitted to Adtalem Global Education

Student Paper

1 %

12

Submitted to University of North Texas

Student Paper

1 %

13

Submitted to University of Surrey

Student Paper

1 %

14

robots.net

Internet Source

1 %

15

Submitted to University at Buffalo

Student Paper

<1 %

16

Submitted to Luton Sixth Form College

Student Paper

<1 %

17

Nazin Ahmed, Rayhan Ahammed, Md. Manowarul Islam, Md. Ashraf Uddin, Arnisha Akhter, Md. Al-Amin Talukder, Bikash Kumar Paul. "Machine Learning based Diabetes Prediction and Development of Smart Web

<1 %

Application", International Journal of Cognitive Computing in Engineering, 2021

Publication

18 Submitted to University of Sunderland <1 %
Student Paper

19 Submitted to Blue Mountain Hotel School <1 %
Student Paper

20 Submitted to Liverpool John Moores
University <1 %
Student Paper

21 ijritcc.org <1 %
Internet Source

22 www.devotion.greenvest.co.id <1 %
Internet Source

23 www.ijert.org <1 %
Internet Source

24 www.nature.com <1 %
Internet Source

25 Nadia Hamid, Jay M. Portnoy, Aarti Pandya.
"Computer-Assisted Clinical Diagnosis and
Treatment", Current Allergy and Asthma
Reports, 2023 <1 %
Publication

26 arxiv.org <1 %
Internet Source

27	www.scilit.net Internet Source	<1 %
28	www.scitechnol.com Internet Source	<1 %
29	Submitted to University of Alabama at Birmingham Student Paper	<1 %
30	Submitted to University of Strathclyde Student Paper	<1 %
31	assets.researchsquare.com Internet Source	<1 %
32	www.qlik.com Internet Source	<1 %
33	Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, Riasat Khan. "Diabetes prediction using machine learning and explainable AI techniques", Healthcare Technology Letters, 2022 Publication	<1 %
34	Submitted to Bede Polding Catholic College Student Paper	<1 %
35	Submitted to Birkbeck College Student Paper	<1 %
36	Submitted to University of Adelaide Student Paper	<1 %

37	Submitted to University of Hertfordshire Student Paper	<1 %
38	dokumen.pub Internet Source	<1 %
39	export.arxiv.org Internet Source	<1 %
40	www.researchsquare.com Internet Source	<1 %
41	Submitted to 2U Southern Methodist University Student Paper	<1 %
42	acikerisim.karabuk.edu.tr:8080 Internet Source	<1 %
43	cdr.lib.unc.edu Internet Source	<1 %
44	jamanetwork.com Internet Source	<1 %
45	listens.online Internet Source	<1 %
46	www.coursehero.com Internet Source	<1 %
47	www.geeksforgeeks.org Internet Source	<1 %
48	www.ncbi.nlm.nih.gov	

49

Hakim El Massari, Zineb Sabouri, Sajida Mhammedi, Noredine Gherabi. "Diabetes Prediction Using Machine Learning Algorithms and Ontology", Journal of ICT Standardization, 2022

Publication

<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography On