



SREE VIDYANIKETHAN ENGINEERING COLLEGE

(Affiliated to Jawaharlal Nehru Technological University Anantapur)
Sree Sainath Nagar, A. Rangampet, Tirupati – 517 102, Chittoor Dist., A.P.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Project Work entitled

PREDICTION OF HEART STROKE IN DIABETIC PATIENTS USING MACHINE LEARNING

is the bonafide work done by

SANGEETHAM KUSALASRI	20121A05P7
GUNJI KALPANA	21125A0507
C JIGNAVI REDDY	19121A0532
SHAIK AKRAM HUSSIAN	20121A05Q1

In the Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, A. Rangampet. is affiliated to JNTUA, Anantapuramu in partial fulfillment of the requirements for the award of Bachelor of Technology in Computer Science and Engineering during 2020-2024.

This work has been carried out under my guidance and supervision.

The results embodied in this Project report have not been submitted in any University or Organization for the award of any degree or diploma.

Internal Guide

Mr. B.Raveendra Naik

Assistant Professor
Dept of AIML
Sree Vidyanikethan Engineering College
Tirupathi

Head

Dr. B. Narendra Kumar Rao

Prof & Head
Dept of CSE
Sree Vidyanikethan Engineering College
Tirupathi

INTERNAL EXAMINER

EXTERNAL EXAMINER

“PREDICTION OF HEART STROKE IN DIABETIC PATIENTS USING MACHINE LEARNING”

A Project Report submitted to

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR.

In Partial Fulfillment of the Requirements for the Award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
BY

SANGEETHAM KUSALASRI	20121A05P7
GUNJI KALPANA	21125A0507
C JIGNAVI REDDY	19121A0532
SHAIK AKRAM HUSSIAN	20121A05Q1

Under the Guidance of

MR. B. Raveendra Naik
Assistant Professor
Dept of AIML, MBU



Department of Computer Science and Engineering
SREE VIDYANIKETHAN ENGINEERING COLLEGE
(Affiliated to JNTUA, Anantapuramu)
Sree Sainath Nagar, Tirupathi – 517 102
2020-2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION AND MISSION

VISION

To become a Centre of Excellence in Computer Science and Engineering by imparting high quality education through teaching, training, and research

MISSION

- The Department of Computer Science and Engineering is established to provide undergraduate and graduate education in the field of Computer Science and Engineering to students with diverse background in foundations of software and hardware through a broad curriculum and strongly focused on developing advanced knowledge to become future leaders.
- Create knowledge of advanced concepts, innovative technologies and develop research aptitude for contributing to the needs of industry and society.
- Develop professional and soft skills for improved knowledge and employability of students.
- Encourage students to engage in life-long learning to create awareness of the contemporary developments in computer science and engineering to become outstanding professionals.
- Develop attitude for ethical and social responsibilities in professional practice at regional, National and International levels.

Program Educational Objectives (PEO's)

After few years of graduation, the graduates of B.Tech. (CSE) will be:

1. Pursuing higher studies in Computer Science and Engineering and related disciplines
2. Employed in reputed Computer and I.T organizations and Government or have established startup companies.
3. Able to demonstrate effective communication, engage in team work, exhibit leadership skills, ethical attitude, and achieve professional advancement through continuing education.

Program Specific Outcomes (PSO's)

On successful completion of the Program, the graduates of B.Tech. (CSE) program will be able to:

- PSO1: Use mathematical methodologies to model real-world problems, Employ modern tools and platforms for efficient design and development of computer-based systems.
- PSO2: Apply adaptive algorithms and methodologies to develop intelligent systems for solving problems from interdisciplinary domains.
- PSO3: Apply suitable models, tools, and techniques to perform data analytics for effective decision-making.
- PSO4: Design and deploy networked systems using standards and principles, evaluate security measures for complex networks, and apply procedures and tools to solve networking issues.

Program Outcomes (PO's)

1. Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems (**Engineering knowledge**).
2. Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences (**Problem analysis**).
3. Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations (**Design/development of solutions**).
4. Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions (**Conduct investigations of complex problems**).
5. Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations (**Modern tool usage**).
6. Apply to reason informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice (**The engineer and society**).

7. Understand the impact of professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development (**Environment and sustainability**).
8. Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice (**Ethics**).
9. Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings (**Individual and team work**).
10. Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions (**Communication**).
11. Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments (**Project management and finance**).
12. Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change (**Life-long learning**).

Course Outcomes

CO1. Knowledge on the project topic (PO1)

CO2. Analytical ability exercised in the project work.(PO2)

CO3. Design skills applied on the project topic. (PO3)

CO4. Ability to investigate and solve complex engineering problems faced during the project work. (PO4)

CO5. Ability to apply tools and techniques to complex engineering activities with an understanding of limitations in the project work. (PO5)

CO6. Ability to provide solutions as per societal needs with consideration to health, safety, legal and cultural issues considered in the project work. (PO6)

CO7. Understanding of the impact of the professional engineering solutions in environmental context and need for sustainable development experienced during the project work. (PO7)

CO8. Ability to apply ethics and norms of the engineering practice as applied in the project work.(PO8)

CO9. Ability to function effectively as an individual as experienced during the project work. (PO9)

CO10. Ability to present views cogently and precisely on the project work. (PO10)

CO11. Project management skills as applied in the project work. (PO11)

CO12. Ability to engage in life-long learning as experience during the project work. (PO12)

CO-PO Mapping

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3	PSO 4
C01	3												3			
C02		3												3		
C03			3												3	
C04				3											3	
C05					3											3
C06						3										
C07							3									
C08								3								
C09									3							
C010										3						
C011											3					
C012												3				

(Note: 3-High, 2-Medium, 1-Low)

DECLARATION

We hereby declare that this project report titled "**Prediction of Heart Stroke in diabetic Patients using Machine learning**" is a genuine project work carried out by us, in the **B.Tech (*Computer Science and Engineering*)** degree course of **Jawaharlal Nehru Technological University Anantapur** and has not been submitted to any other course or University for the award of any degree by us.

Signature of the students

1. SANGEETHAM KUSALASRI
2. GUNJI KALPANA
3. C JIGNAVI REDDY
4. SHAIK AKRAM HUSSIAN

ACKNOWLEDGEMENT

We are extremely thankful to our beloved Chairman and founder **Dr. M. Mohan Babu** who took keen interest to provide us the infrastructural facilities for carrying out the project work.

We are highly indebted to **Dr. B. M. Satish**, Principal of Sree Vidyanikethan Engineering College for his valuable support and guidance in all academic matters.

We are very much obliged to **Dr. B. Narendra Kumar Rao**, Professor & Head, Department of CSE, for providing us the guidance and encouragement in completion of this project.

We would like to express our indebtedness to the project coordinator, **Dr. P. Dhanalakshmi**, Associate Professor, Department of CSE for his valuable guidance during the course of project work.

We would like to express our deep sense of gratitude to **Mr. B. Raveendra Naik**, Assistant Professor, Department of AIML, for the constant support and invaluable guidance provided for the successful completion of the project.

We are also thankful to all the faculty members of CSE Department, who have cooperated in carrying out our project. We would like to thank our parents and friends who have extended their help and encouragement either directly or indirectly in completion of our project work.

ABSTRACT

Nowadays in the medical field, it is essential to predict diseases early to prevent them. Diabetes is one of the most chronic diseases all over the world. In today's lifestyles, sugar and fat are typically included in our dietary habits, which increases the risk of diabetes. Diabetes, a prevalent metabolic disorder, is a significant risk factor for cardiovascular complications, including heart strokes. Early detection and proactive management of heart stroke risk in diabetic patients can substantially improve their overall cardiovascular health and reduce the incidence of life-threatening events. Our project leverages machine learning algorithms such as Decision Tree, Random Forest and Light Gradient Boosting Machine (LGBM) algorithms to analyze comprehensive datasets comprising medical history, lifestyle factors, and physiological parameters of diabetic individuals. The model integrates various features such as blood glucose levels, blood pressure, cholesterol levels, and demographic information. The Heart-stroke Prediction model not only identifies individuals at high risk of strokes but also provides actionable insights for healthcare professionals to tailor preventive strategies. The outcomes of this project contribute to the growing field of predictive healthcare analytics, demonstrating the potential of machine learning in improving diagnostic accuracy and optimizing resource allocation in the healthcare sector.

Keywords: Diabetes, Cardio vascular diseases, Decision tree Algorithm, Random Forest Algorithm, Light Gradient Boosting Machine (LGBM).

TABLE OF CONTENTS

Chapter No	Title	Page No.
	Vision and Mission	I
	Program Educational Objectives	II
	Program Specific Outcomes	III
	Program Outcomes	IV
	Course Outcomes	VI
	CO-PO Mapping	VII
	Declaration	VIII
	Acknowledgments	IX
	Abstract	X
1	INTRODUCTION 1.1 Introduction 1.2 Problem statement 1.3 Objectives 1.4 Scope 1.5 Applications 1.6 Limitations	1 – 7
2	LITERATURE SURVEY	8 – 11
3	ANALYSIS 3.1 Existing system 3.2 Proposed system 3.3 Software and Hardware Requirements 3.4 Software requirements specifications	12 – 18
4	DESIGN 4.1 Use-case diagram 4.2 Class diagram 4.3 Sequence diagram 4.4 ER diagram 4.5 DFD diagram 4.6 Block Diagram	19 -26
5	IMPLEMENTATION 5.1 Introduction	27 - 29

	5.2 Algorithm 5.3 Methodology 5.4 Implementation	
6	EXECUTION PROCEDURE & TESTING 6.1 Execution Procedure 6.2 Testing 6.3 Types of Testing	30 – 35
7	RESULT & PERFORMANCE EVALUATION 7.1 Evaluation Description 7.2 Performance Evaluation 7.3 Results	36 – 40
8	CONCLUSION & FUTURE WORK	41
	APPENDIX Program listing/code List of Figures List of Tables Screen shots	42 – 51
	REFERENCES	52 - 53

CHAPTER-1

INTRODUCTION

1.1 Introduction:

Heart plays a crucial role in the circulatory system, pumping blood through vessels to supply organs with oxygen and essential materials. This system is vital for overall health, and the heart's proper functioning is essential. Any malfunction in the heart can lead to severe health issues, even death.

Diabetes Mellitus (DM) is a metabolic disorder characterized by elevated Blood Sugar Levels (BSL), leading to potential complications in various organs like the heart, blood vessels, eyes, kidneys, and nerves. The primary complication is cardiovascular, increasing the risk of heart disease and vascular malformations, including strokes. Alarming, approximately 50% of individuals with diabetes face the heightened threat of succumbing to heart disease and stroke. Elevated blood glucose resulting from diabetes can harm blood vessels and nerves that regulate the heart and blood vessels.. Over time, this damage increases the risk of developing heart disease, with individuals with diabetes often experiencing it at a younger age compared to those without diabetes. Damaged arteries create a more conducive environment for fatty substances to adhere to the arterial walls, leading to potential blockages and a decrease in the space available for proper blood flow. In the event of arteries supplying blood to the heart becoming obstructed, it can precipitate a heart attack. Similarly, if arteries responsible for carrying blood to the brain become blocked, the consequence may be stroke.

In the US, two of the most common chronic illnesses that cause death are diabetes and cardiovascular disease (CVD). About 9% of Americans were officially diagnosed with diabetes in 2015, while 3% remained undiagnosed. In addition, almost 34% had prediabetes. Nonetheless, nearly 90% of

persons with prediabetes were ignorant of their illness. Conversely, one in four deaths in the US occur as a result of CVD each year. 92.1 million persons in America are thought to be living with a cardiovascular disease (CVD) or the aftermath of a stroke, with associated direct and indirect healthcare expenses estimated to exceed \$329.7 million. Furthermore, there is a connection between diabetes and CVD. According to the American Heart Association, heart disease claims the lives of at least 68% of diabetics 65 years of age or older. According to a comprehensive assessment of the literature, conducted by Einarson et al., cardiac disease affects 32.2% of people with type 2 diabetes.

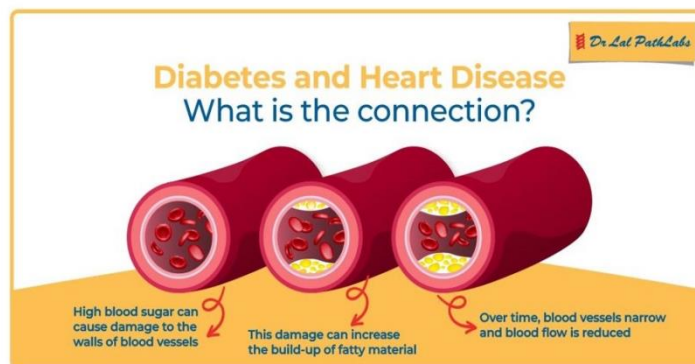


Fig 1.1:Effect of Diabetes on Heart

In the realm of continuously expanding data, hospitals are gradually implementing big data systems [6]. Using data analytics in the healthcare system can yield significant advantages, including enhanced diagnosis, better results, lower expenses, and insights [7]. Specifically, the effective application of machine learning augments the efforts of physicians and boosts the effectiveness of the healthcare infrastructure [8]. Machine learning models have demonstrated significant gains in diagnostic accuracy when used in conjunction with doctors [9]. Several prevalent diseases have since been predicted using machine learning models, the identification of hypertension in diabetic patients, and the categorization of patients with CVD among diabetic patients.

Patients with diabetes or heart disease can be identified with the use of

machine learning models. Patients who are at risk for these prevalent diseases might be identified by a variety of variables. Hidden patterns in these variables that might otherwise go unnoticed can be found with the use of machine learning techniques.

In this work, we forecast diabetes and cardiovascular illness using supervised machine learning algorithms. To assist a larger variety of patients, we build the models to forecast diabetes and cardiovascular disease separately, despite the established link between these two conditions. Consequently, we are able to determine the characteristics that are shared by the diseases and have an impact on their prognosis. We also take into account the prognosis of undetected diabetes and prediabetes. Several models for the prediction of various diseases are trained and tested using data from the National Health and Nutrition Examination Survey (NHANES). In order to improve prediction ability, this research also investigates a weighted ensemble model that integrates the output of several supervised learning models.

1.2 Problem Statement:

Diabetic individuals are at a heightened risk of developing cardiovascular disorders, specifically heart stroke, which poses a significant healthcare burden. Timely identification of diabetic patients with a heightened risk of heart stroke is crucial for implementing prompt therapies and tailored care. Conventional risk assessment approaches may lack accuracy and overlook the complex interconnections among different clinical and demographic factors. The task at hand is creating a precise and comprehensible predictive model for forecasting cardiac strokes in diabetes populations. Machine learning, due to its capacity to identify intricate patterns within large datasets, presents a promising opportunity. Nevertheless, incorporating machine learning into clinical practice and transferring predictive models from research settings to real-world healthcare environments present notable obstacles.

1.3 Objectives:

The primary objective of this project is to develop a machine learning-based predictive model capable of accurately identifying diabetic patients at high risk of experiencing heart stroke. The objectives are as follows:

- To develop and implement machine learning models utilizing decision tree, random forest, and logistic regression algorithms for predicting heart stroke risk in diabetic patients accurately.
- To create robust predictive models capable of distinguishing between diabetic patients at high risk of heart stroke and those at lower risk, achieving high precision and reliability in risk assessment.
- To leverage advanced techniques such as feature engineering, hyper parameter tuning, and model ensemble methods to enhance the performance of the machine learning models, overcoming challenges associated with complex relationships and imbalanced data in heart stroke prediction.
- To develop a user-friendly web interface using tools like Streamlit, allowing healthcare providers to easily input patient data and obtain personalized heart stroke risk predictions in a visually intuitive manner.
- To optimize model performance in real-time scenarios and investigating the impact of different algorithms and parameter settings on model accuracy and stability.
- To provide a reliable and effective solution for healthcare professionals to assess heart stroke risk in diabetic patients, enabling timely interventions and personalized care plans to prevent cardiovascular events.

1.4 Scope :

The scope of the project encompasses several key aspects as follows:

- Gathering a comprehensive dataset containing relevant features such as medical history, physiological parameters, and lifestyle factors from diabetic patients.
- Preprocessing the dataset to handle missing values, normalize features, and encode categorical variables to ensure its quality and suitability for machine learning model training.
- Implementing machine learning algorithms including decision tree, random forest, and logistic regression for predicting heart stroke risk in diabetic patients.
- Evaluating the performance of the developed models using metrics such as accuracy to assess their effectiveness in distinguishing between patients at high and low risk of heart stroke.
- Developing a user-friendly web interface using the Streamlit library to allow healthcare professionals to input patient data and obtain personalized predictions of heart stroke risk.
- Integrating the trained machine learning models into the web interface to generate real-time predictions based on user inputs.
- Conducting thorough testing and validation of both the machine learning models and the web interface to ensure accuracy, reliability, and usability.

1.5 Applications:

The project "Prediction of Heart Stroke in Diabetic Patients Using Machine Learning" can span across various sectors where early detection of heart stroke risk and user-friendly access to predictive models are essential. Here are the potential applications:

- Healthcare Industry: Identifying diabetic patients at high risk of heart

stroke allows healthcare providers to implement preventive measures and personalized treatment plans.

- Insurance Sector: Insurance companies can use the predictive models to assess the heart stroke risk of diabetic individuals applying for health insurance policies, leading to better risk management and pricing strategies.
- Health Tech Startups: Startups developing the health monitoring applications can integrate the predictive models to provide users with personalized insights into their heart stroke risk based on their diabetic status and other health parameters.
- Pharmaceutical Sector: Pharmaceutical companies conducting clinical trials for diabetes medications can utilize the predictive models to screen participants for heart stroke risk, ensuring participant safety and efficacy assessment.
- Research and academia: Researchers and academics can utilize the predictive models to analyze large-scale epidemiological data on diabetic patients and heart stroke incidence, leading to insights into risk factors and preventive strategies.

1.6 Limitations:

The Prediction of heart stroke in diabetic patients using machine learning, despite its valuable contributions, faces several limitation

- Variability in Patients data :Variations in patient data such as demographics, medical history, and lifestyle factors may affect the accuracy of the predictive models.Inconsistencies in data quality and completeness could lead to biases or inaccuracies in model predictions, especially when processing data from diverse patient population.
- Performance on Unseen Data: The predictive models may exhibit reduced performance when presented with unseen or novel patient data not adequately represented in the training dataset.Rare or outlier cases, such as patients with unique health conditions or comorbidities,

may challenge the models' generalization capabilities and result in less reliable prediction.

- Integration with clinical workflows: Integrating the predictive models into existing clinical workflows and electronic health record (EHR) systems may pose technical challenges and require coordination with healthcare IT departments.
- Web page usability and accessibility :The usability and accessibility of the web page developed for accessing the predictive models may vary depending on users' technical proficiency and familiarity with the interface. Individuals with limited internet access or disabilities may encounter difficulties in using the web page, potentially hindering their access to critical health information and predictions.

CHAPTER-2

LITERATURE SURVEY

"Machine Learning Approaches for Predicting Cardiovascular Risk in Type 2 Diabetes: A Review"[1] is proposed by Smith J. et al. in 2021, this offers a thorough examination of the use of machine learning methods in predicting cardiovascular risk in patients with type 2 diabetes. The authors examine multiple research published until 2020, concentrating on the methodology, datasets, and performance indicators utilized in these predictive models. They examine the difficulties of conventional risk assessment techniques and emphasize the capacity of machine learning algorithms to enhance the accuracy of risk prediction. The paper also discusses the necessity of customized risk assessment methods designed for diabetes patients, taking into account the diverse characteristics of the condition. The article highlights the potential of machine learning in improving cardiovascular risk prediction and management for persons with type 2 diabetes.

"Machine Learning Methods for Forecasting Heart Stroke in Diabetic Individuals"[2] this paper is published by Lee S. et al. in 2019, they used machine learning methods to forecast cardiac strokes in diabetes individuals. The study examines a vast data set of diabetes patients, including different clinical and demographic characteristics. A comparison is made between several machine learning methods, such as logistic regression, support vector machines, based on their predicted accuracy. The results show a high level of accuracy in predicting cardiac strokes in diabetic persons, indicating the possibility of early intervention and tailored preventive measures. The implications of the study are significant for both clinical practice and public health interventions. By accurately forecasting cardiac strokes in diabetes individuals, healthcare providers can implement targeted preventive measures, ultimately reducing the incidence of cardiovascular events and improving patient outcomes. Furthermore, the study's findings contribute to our understanding of the complex relationship between diabetes and cardiovascular disease, paving the way for future research and innovation in this field.

In 2018,"Utilizing Machine Learning Algorithms for Early Cardiovascular Event Detection in Diabetic Patients"[3] published by Stephen J et al., this paper examines the incorporation of various machine learning algorithms to identify cardiovascular events early in diabetes patients. The project aims to create a hybrid model that integrates the advantages of different techniques such as neural networks, decision trees, and gradient boosting machines. The dataset used for training the hybrid model is intentionally diverse, encompassing a wide range of clinical, genetic, and lifestyle characteristics relevant to cardiovascular risk in diabetes patients. This comprehensive approach ensures that the model can capture the full spectrum of factors influencing cardiovascular events, from traditional risk factors such as blood pressure and cholesterol levels to genetic predispositions and lifestyle habits. The algorithm is trained on a varied dataset to enhance prediction accuracy. The evaluation results show that ensemble learning approaches outperform individual algorithms in identifying diabetes patients at high risk of cardiovascular events.

"Machine Learning for Feature Selection and Classification of Heart Stroke Risk in Diabetic Patients"[4] is provided by chen Y,et al. in 2018. Th research introduces a framework that utilizes machine learning for feature selection and classification to evaluate the risk of cardiac stroke in diabetes patients. The study utilizes sophisticated feature selection techniques, including recursive feature elimination and principal component analysis, to pinpoint the most informative predictors from a broad array of clinical data. Various classification methods such as k-nearest neighbors, support vector machines, and Naive Bayes are used to predict the risk of heart stroke based on specified variables. The experimental results confirm the efficacy of the suggested method in precisely identifying high-risk diabetes patients, enabling focused treatments and preventive actions. Furthermore, the research contributes to advancing our understanding of the complex relationship between diabetes and cardiovascular disease. By identifying informative predictors and developing accurate risk prediction models, the

framework provides valuable insights into the underlying mechanisms driving cardiovascular risk in diabetes patients. This knowledge can inform the development of targeted interventions and preventive strategies aimed at addressing modifiable risk factors.

"Predicting Heart Stroke Events in Diabetic Patients Over Time Using Machine Learning"[5] is by Gupta, et al in 2017. This longitudinal study examines the use of machine learning to predict heart stroke occurrences in diabetes patients over a period of time. The study uses a longitudinal dataset that covers many years to track changes over time in clinical markers and health outcomes in diabetes individuals. Machine learning algorithms such as recurrent neural networks and hidden Markov models are used to examine the time-based patterns and forecast upcoming heart stroke occurrences using sequential patient information. Incorporating longitudinal data enhances prediction accuracy more than cross-sectional methods, allowing for proactive management of cardiovascular risk in diabetic patients. Machine learning algorithms such as recurrent neural networks and hidden Markov models are well-suited to analyzing sequential data and capturing temporal dependencies in patient health trajectories. Recurrent neural networks, in particular, excel at modeling sequential data by maintaining an internal memory of past observations, allowing them to learn complex patterns and make predictions based on sequential input. Hidden Markov models, on the other hand, are probabilistic models that can capture the underlying stochastic nature of sequential data, making them suitable for modeling transitions between different states of health over time.

"Utilizing Machine Learning Algorithms for Early Cardiovascular Event Detection in Diabetic Patients"[6] is published in 2017. The study under review explores the integration of multiple machine learning algorithms to detect cardiovascular events early in patients with diabetes. The overarching goal is to develop a hybrid model that combines the strengths of different techniques, including neural networks, decision trees, and gradient boosting machines. By leveraging a diverse dataset encompassing clinical, genetic, and lifestyle

characteristics, the algorithm aims to improve prediction accuracy for identifying diabetes patients at high risk of experiencing cardiovascular events. Individuals with diabetes are known to face an elevated risk of cardiovascular complications, making early detection and intervention crucial for mitigating adverse outcomes. Traditional risk assessment methods may not fully capture the complex interplay of factors contributing to cardiovascular risk in this population. Thus, the study's focus on machine learning approaches represents a promising avenue for enhancing risk prediction and facilitating timely intervention strategies. The study's methodology involves training and evaluating a hybrid model that integrates various machine learning algorithms. Neural networks, decision trees, and gradient boosting machines are selected for their respective strengths in handling nonlinear relationships, capturing complex interactions, and minimizing prediction errors. By combining these techniques into a single ensemble model, the study seeks to leverage their complementary capabilities and improve overall prediction performances.

In 2015, "Machine Learning for Feature Selection and Classification of Heart Stroke Risk in Diabetic Patients" this research developed a novel framework that harnesses machine learning techniques for feature selection and classification to assess the risk of cardiac stroke in patients with diabetes. This study addresses the critical need for accurate risk assessment tools tailored to the unique challenges faced by individuals with diabetes, who are at an increased risk of cardiovascular events. By leveraging sophisticated feature selection techniques and various classification methods, the framework aims to pinpoint the most informative predictors from a wide array of clinical data and accurately predict the risk of heart stroke in this population. One of the key components of the proposed framework is the use of advanced feature selection techniques, including recursive feature elimination and principal component analysis. These methods enable the identification of the most relevant predictors from a vast pool of clinical data, streamlining the predictive modeling process and enhancing the interpretability of the results.

CHAPTER -3

ANALYSIS

System Analysis is the detailed study of the various operations performed by the system and their relationships within and outside the system. The breakdown of something into parts so that the entire system may be understood.

System Analysis is concerned mainly with understanding or being aware of the problem, identifying the relevant variables which are used for decision-making, and analyzing and synthesizing them to obtain optimal solutions. Another view of it is a Problem-Solving technique that breaks down a system into different parts and it studies how those parts will interact to accomplish their purpose.

3.1 Existing System:

Numerous studies have delved into predicting heart strokes in diabetic patients through the application of machine learning techniques. These investigations typically entail the aggregation of diverse patient data, encompassing demographic details, medical histories, and clinical measurements. Subsequently, a myriad of machine learning algorithms are employed, including support vector machines (SVM), artificial neural networks (ANN), k-nearest neighbors (KNN), naive Bayes, gradient boosting, convolutional neural networks (CNN), decision trees, and recurrent neural networks (RNN). These algorithms are tasked with discerning predictive patterns and identifying risk factors intricately linked to heart strokes. Through meticulous comparison of algorithmic performances, researchers aim to ascertain the most effective models for this predictive task. Moreover, these studies delve into evaluating the impact of diverse feature sets, exploring which combination of variables yields optimal predictive accuracy. Additionally, there is a growing interest in leveraging medical imaging data, such as MRI scans and echocardiograms, to enhance the precision of predictive models. By combining advanced algorithms with comprehensive datasets and innovative methodologies, researchers strive to develop robust

and reliable predictive models that can empower healthcare professionals in identifying the high-risk diabetic patients and instituting tailored preventive interventions to avert cardiovascular events effectively.

3.2 Disadvantages of Existing Systems:

1.Lack of Personalization: Existing systems may use generic risk prediction models that do not account for individual patient characteristics or variations in diabetic profiles. This can result in less accurate predictions and missed opportunities for early intervention.

2.Limited Feature Set: Some existing systems may rely on a limited set of features or risk factors for predicting heart stroke in diabetic patients. This may overlook important predictors or fail to capture complex relationships between variables, leading to reduced predictive accuracy.

3.Manual Data Analysis: Many current approaches to predicting heart stroke in diabetic patients involve manual data analysis by healthcare professionals. This process can be time-consuming, prone to errors, and may not leverage the full potential.

4 Lack of Integration: Existing systems may not be seamlessly integrated into healthcare workflows or electronic health record (EHR) systems, requiring additional effort for data input and interpretation. This can hinder adoption by healthcare providers and limit the scalability of predictive models in real-world .

5.Limited Accessibility: Current systems for predicting heart stroke in diabetic patients may not be readily accessible to all healthcare providers or patients, particularly in resource-constrained settings or underserved populations. This can contribute to disparities in healthcare access and outcome.

3.2 Proposed System:

The proposed system aims to predict heart stroke risk in diabetic patients using machine learning. It involves acquiring and preprocessing comprehensive health data, selecting relevant features, and developing predictive models. The system prioritizes interpretability and ethical considerations, ensuring robust data security measures. Integration into a user-friendly interface facilitates seamless deployment in healthcare settings. Clinicians are provided with actionable insights for personalized patient care, while continuous monitoring and updating ensure relevance and effectiveness. Overall, the system represents a transformative approach to early detection and management of heart stroke risk in diabetic populations, enhancing patient outcomes and healthcare efficiency.

Work Flow of Proposed Model:

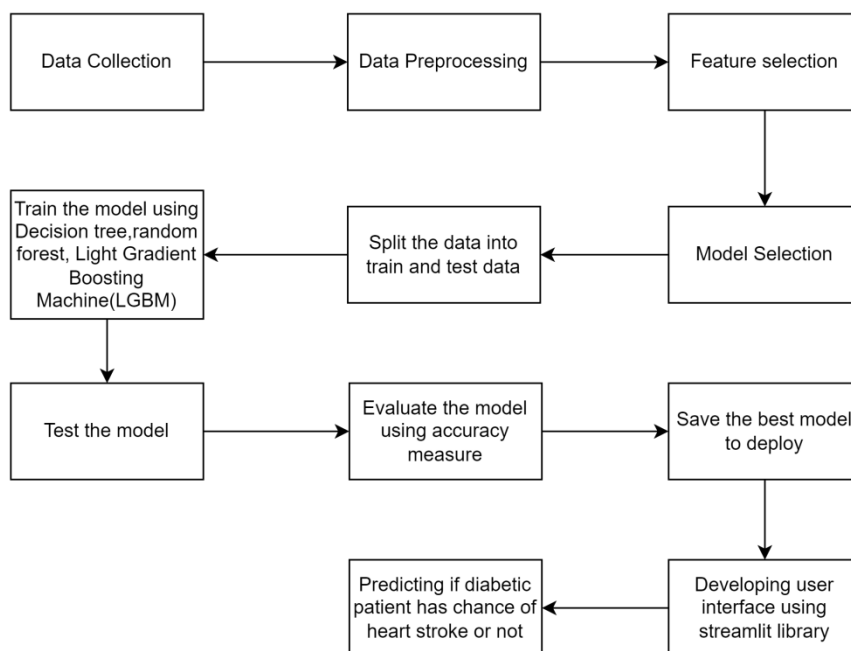


Fig -3.2: Block diagram of proposed model

Advantages:

- Enhanced predictive system
- Efficient user-interface web page.

3.2 Software & Hardware Requirements:

- Hardware Requirements:
 - Operating system : Windows 10 or 10+
 - Processor : A multicore processor (e.g., Intel Core i5 or AMD Ryzen)
 - RAM : 8GB (min)
 - Hard Disk : 128 GB
- Software Requirements
 - Software : Python 3.11 or high version
 - Framework : Scikit learn.
 - IDE/Workbench : Vscode/Google collaboratory
 - Libraries :streamlit,pandas, Os, matplotlib, Numpy

3.3 Software Requirement Specification:

Functional and non-functional requirements:

Functional Requirements:

Requirements: These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

1.Data Acquisition and Preprocessing:

The system should be able to acquire comprehensive health data of diabetic patients from multiple sources, including electronic health records (EHR), laboratory reports, and medical imaging.It should preprocess the data to handle missing values, outliers, and inconsistencies, ensuring data quality and suitability for machine learning analysis.

2. Feature Selection and Engineering:

The system should identify relevant features associated with heart stroke risk in diabetic patients through exploratory data analysis and domain knowledge. It should support feature engineering techniques to derive new predictors or transformations of existing features to enhance predictive performance.

3. Machine Learning Model Development:

The system should offer various machine learning algorithms suitable for binary classification tasks, such as logistic regression, decision trees, random forests, support vector machines, and neural networks. It should facilitate model training and optimization using techniques such as cross-validation, hyperparameter tuning, and model selection.

4. Integration and Deployment:

The system should integrate the predictive model into a user-friendly interface or application accessible to healthcare professionals. It should support integration with existing healthcare systems or electronic health records (EHR) platforms for seamless deployment and integration into clinical workflows.

5. Evaluation and Validation:

The system should assess the performance of the developed model using evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). It should validate the model on separate test datasets to ensure robustness and generalizability across diverse patient populations.

Non-Functional Requirements:

These are the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to other. They also called non-behavioral requirements.

1.Performance:

Throughput: Support processing a high volume of data within reasonable time frames.

Efficiency: Ensure that machine learning algorithms processes are executed swiftly and do not significantly impact system performance.

2.Usability:

User Interface: Design intuitive interfaces with clear navigation and controls for users to interact with the application seamlessly.

Accessibility: Ensure that the application is accessible to users with disabilities, adhering to accessibility standards.

Learnability: Facilitate easy understanding and adoption of the application for users with varying levels of technical expertise.

3.Reliability:

Stability: Ensure the stability of the application under varying conditions, minimizing crashes or unexpected downtime.

Error Handling: Implement robust error handling mechanisms to gracefully manage and recover from errors or exceptions.

4.Scalability:

Capacity: Support scalability to accommodate an increasing number

Performance Scaling: Enable horizontal or vertical scaling to maintain performance levels as demand fluctuates.

5. Compatibility:

Platform Compatibility: Ensure compatibility with a range of operating systems, browsers, and devices commonly used by users.

Integration: Support integration with third-party tools or platforms commonly utilized in data security workflows.

6. Documentation:

Comprehensive Documentation: Provide thorough documentation covering installation, configuration, and usage of the application.

Support Materials: Offer supplementary resources such as FAQs, tutorials, and troubleshooting guides to assist users.

CHAPTER - 4

DESIGN

The most creative and challenging part of system development is System Design. The Design of a system can be defined as a process of applying various techniques and principles for the purpose of defining a device, architecture, modules, interfaces and data for the system to satisfy specified requirements. For the creation of a new system, the system design is a solution to "how to" approach.

UML DIAGRAMS:

UML is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems.

- UML stands for Unified Modelling Language. UML is different from the other common programming languages like C++, Java, COBOL, etc...
- UML may be described as a general-purpose visual modeling language to visualize, define, construct, and document software systems.
- Although UML is generally used to model software programs, it's not limited within this boundary. It's also utilized to model non-software systems too. By way of example, the process flow in a manufacturing device, etc...

UML Isn't a Programming language but tools can be used to generate code in various Languages using UML diagrams. UML includes a direct relation with object-oriented Analysis and design.

4.1 Use Case Diagram:

A use-case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis.

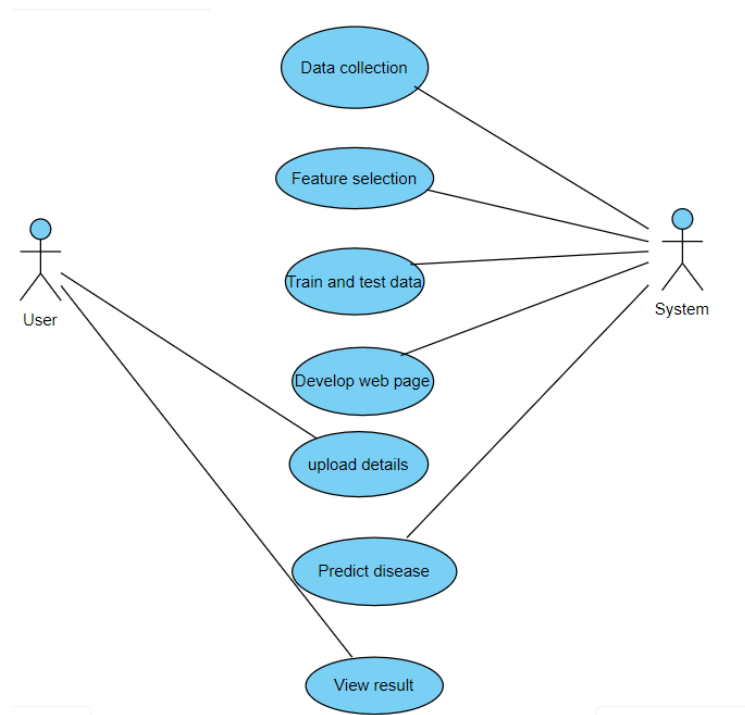


Fig 4.1 – Use-Case Diagram for prediction of heart stroke in diabetic patients.

Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

Use-Case Diagram description for Prediction of heart stroke in diabetic students using machine learning:

Actors:

- System
- User

- Data Collection: Gather diabetic patients relevant dataset .
- Feature selection: Select the features that predict heart stroke.
- Train and test data: System can perform this action by importing data for model for training and testing.
- Develop Web Page: To develop user-interface webpage.
- Predict disease: This is an outcome for data that uploaded by user in webpage.
- View result: Enables users to view the result for uploaded data.

4.2 Class Diagram:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

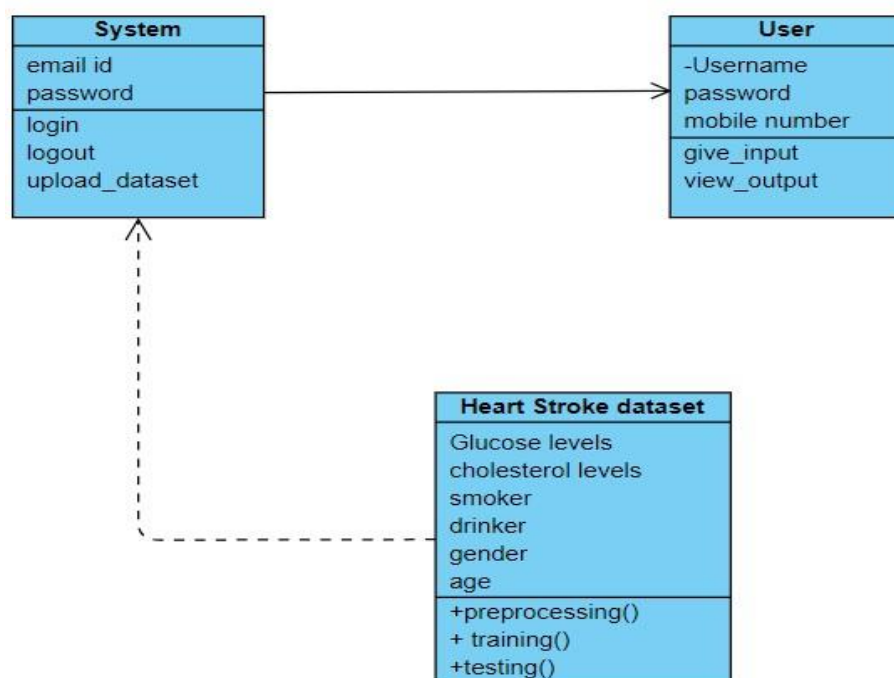
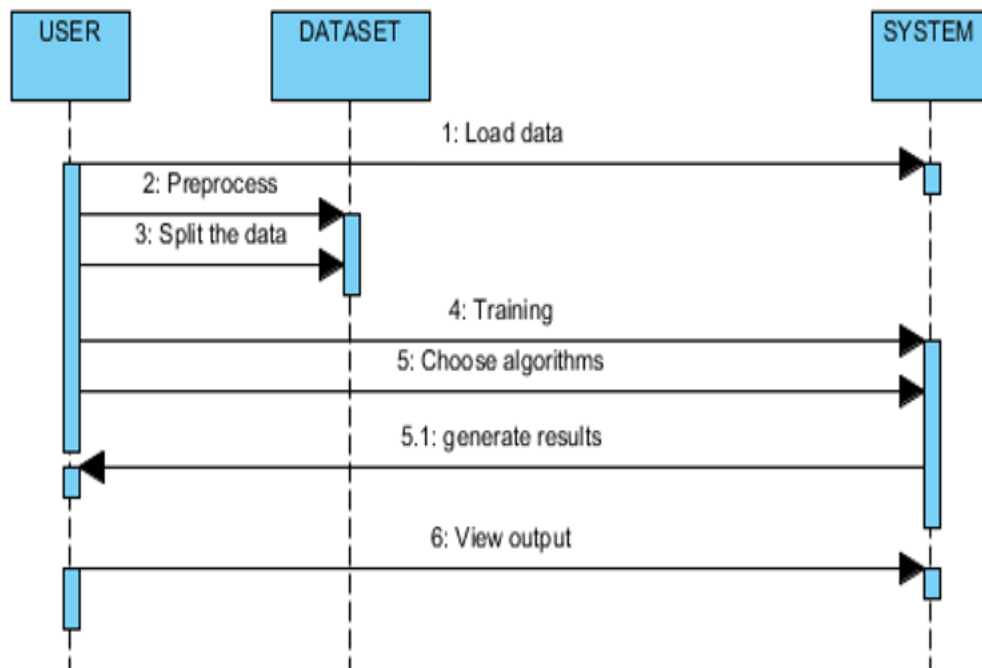


Fig - 4.2 Class Diagram for prediction of heart stroke in diabetic patients

4.3 Sequence Diagram:

Modeling Language (UML) is a sort of interaction diagram which shows how procedures operate with one another and in what sequence. Sequencediagrams include the following components:

- **Class roles:** These signify Functions that objects can play inside theinteraction.
- **Lifelines:** It symbolizes the existence of an item over a time period.
- **Activations:** It signifies the time during which an object is performing the operation
- **Messages:** It symbolizes Communication between items



4.1 ER Diagram:

An Entity-relationship model (ER model) describes the structure of a database with the help of a diagram, which is known as an Entity Relationship Diagram (ER Diagram). An ER model is a design or blueprint of a database that can later be implemented as a database. The main components of the E-R model are:

An ER diagram shows the relationship among entity sets. An entity set is a group of similar entities and these entities can have attributes. In terms of DBMS, an entity is a table or attribute of a table in the database, so by showing the relationship among tables and their attributes, the ER diagram shows the complete logical structure of a database.

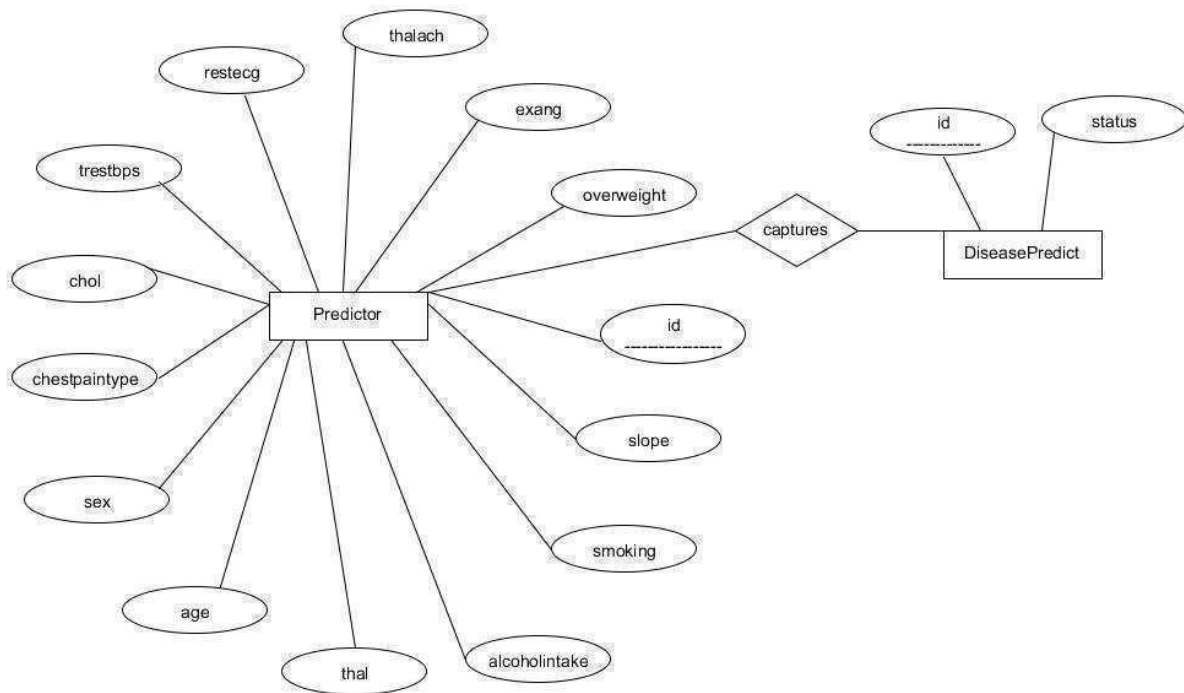


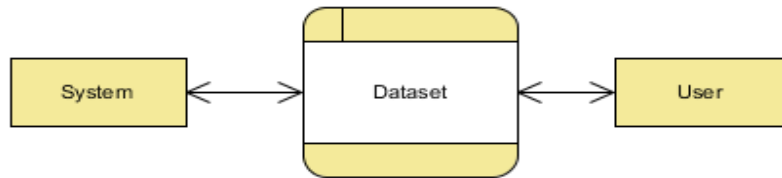
Fig 4.4- ER Diagram for prediction of heart stroke

4.2 DFD Diagram:

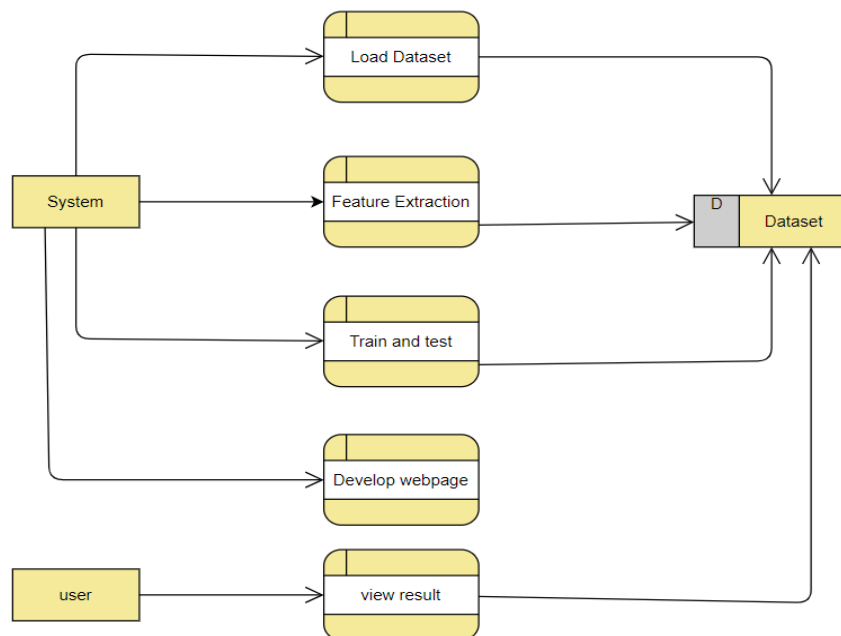
A Data Flow Diagram (DFD) is a traditional way to visualize the information flows within a system. A neat and clear DFD can depict a good amount of the system requirements graphically. It can be manual, automated, or a combination of both. It shows how information enters and leaves the system, what changes the information, and where information is stored. The purpose of a DFD is to show the scope and boundaries of a system as a whole. It may be used as a communications tool between a systems analyst and any person

who plays a part in the system that acts as the starting point for redesigning a system.

Level 0:



Level 1:



Level 2:

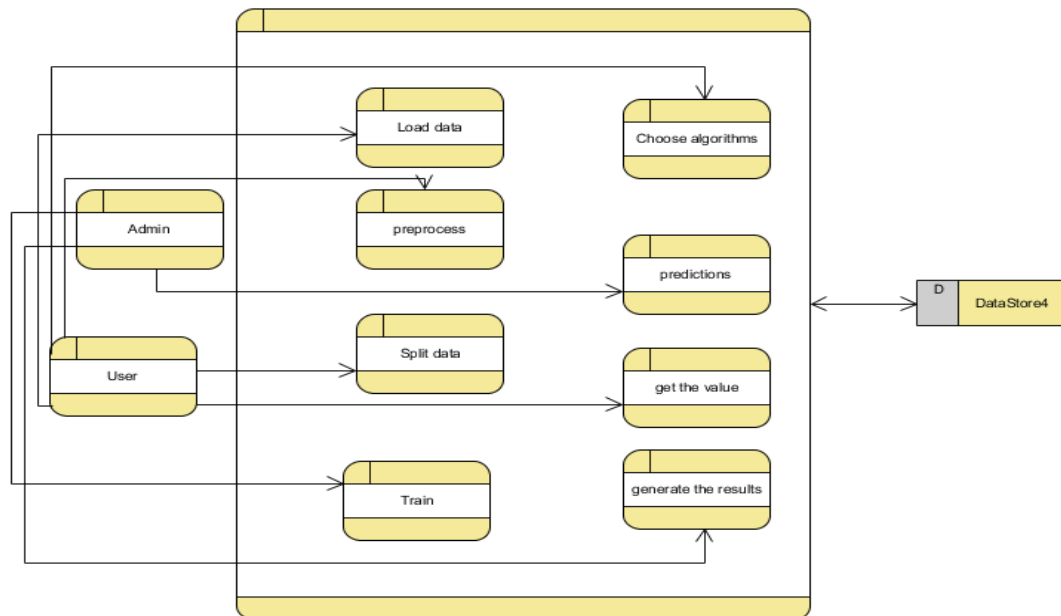


Fig-4.5 Data Flow Diagrams for prediction of heart stroke in diabetic patients

4.3 Block Diagram

It below block Diagram describes the basic steps for prediction of heart stroke in diabetic patients using machine learning algorithms

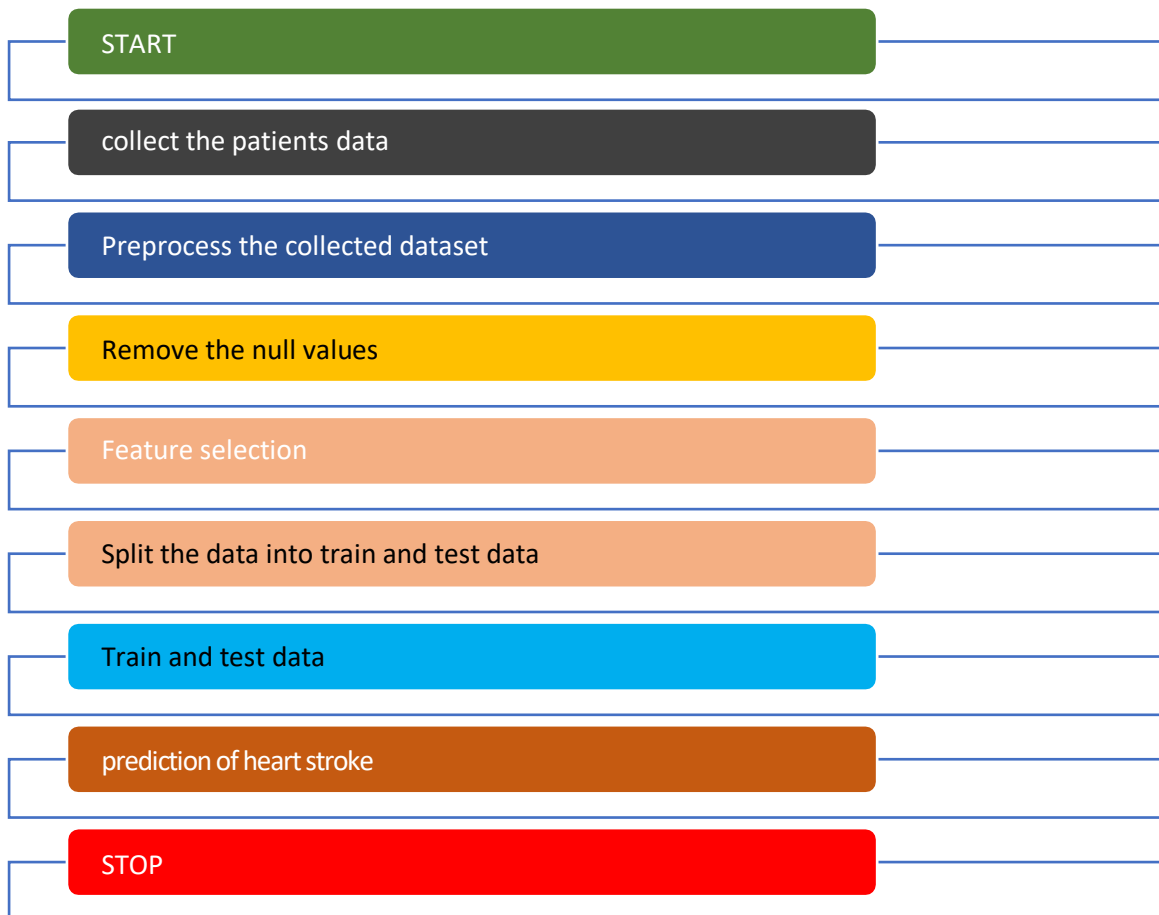


Fig 4.6 Block Diagram for prediction of heart stroke

CHAPTER 5

IMPLEMENTATION

5.1 Introduction:

The design implementation phase is a significant percentage of the overall design cycle. It is critical that the implementation phase of the design be handled as efficiently as possible.

Software design is a creative activity in which you identify software components and their relationships, based on a customer's requirements. Implementation is the process of realizing the design as a program.

5.2 Algorithm:

Our project revolves around harnessing the power of machine learning algorithms—specifically, decision tree, random forest, and logistic regression—to forecast the occurrence of heart strokes in individuals with diabetes. These algorithms, chosen for their effectiveness in classification tasks, collectively provide a robust framework for predicting heart stroke probabilities. Notably, logistic regression emerges as a standout performer, exhibiting exceptional accuracy in determining the likelihood of heart strokes in diabetic patients. To enhance the usability and accessibility of our predictive models, we develop a user-friendly web interface using the Streamlit library. Our project integrates advanced machine learning techniques with intuitive user interface design to provide a comprehensive solution for predicting heart strokes in diabetic patients. By combining the predictive capabilities of logistic regression with the accessibility of the Streamlit-based web interface, we aim to facilitate proactive healthcare management strategies and improve patient outcomes in the prevention and treatment of heart-related complications.

5.3 Methodology:

The methodology for predicting heart stroke in diabetic patients using machine learning involves firstly, comprehensive datasets containing relevant features such as medical history, physiological parameters (e.g., blood pressure, cholesterol levels), lifestyle factors, and genetic predispositions are collected and preprocessed. This preprocessing stage includes handling missing values, normalizing numerical features, and encoding categorical variables. Subsequently, feature selection techniques are applied to identify the most relevant predictors for heart stroke risk, potentially followed by feature engineering to enhance model performance. Machine learning algorithms such as logistic regression, decision trees, or random forests are selected based on dataset characteristics, and the chosen model is trained on a portion of the dataset while tuning hyperparameters to optimize performance. Model evaluation is conducted using accuracy metrics to assess the model's effectiveness in predicting heart stroke risk. Following the development of the machine learning model, the next step involves creating a user interface web page using the Streamlit library. This involves installing the Streamlit library and creating a Python script to define the user interface and integrate the trained machine learning model. Streamlit's widgets are utilized to allow users to input their health information, and the model generates predictions based on these inputs. The web page displays the prediction results along with relevant visualizations using Streamlit's components, providing users with an intuitive and interactive platform for accessing the model's predictions. Finally, the deployed web application is tested, debugged, and deployed to a suitable hosting platform, ensuring its functionality, usability, and performance. Regular maintenance and updates are conducted to address any bugs, errors, or performance issues and to incorporate user feedback for continuous improvement of the application.

5.4 Implementation:

Step1: Data Collection and Preprocessing: Gather a dataset containing relevant features such as age, blood pressure, cholesterol levels, diabetic status, and history of heart stroke. Then apply preprocessing techniques.

Step2: Feature Selection: Identifying and selecting pertinent features that contribute significantly to the prediction of heart strokes in diabetic individuals.

Step3: Model Training: Choosing appropriate machine learning algorithms based on the nature of the data and the prediction task. In our project we used algorithms include logistic regression, decision trees, random forest. Split the data into test and train data and then train the selected models using gridsearch techniques.

Step 4: Model Evaluation: Evaluate the trained model's performance using metrics such as accuracy.

Step 5: Integration with Streamlit: Incorporate the trained best model to generate predictions based on user input.

Step6 : Testing and Debugging: Test the web application locally to ensure it functions as expected.

Step 7: Uploading the data in web page : Evaluate the user entered data and predict whether the patient is vulnerable to heart stroke or not.

CHAPTER 6

EXECUTION PROCEDURE AND TESTING

6.1 Execution Procedure:

The execution procedure for building a robust model using Prediction of heart stroke in diabetic patients using machine learning can be broken down into the following steps:

Data Gathering: Collect a diverse dataset containing health records of diabetic patients, including demographic information, medical history, laboratory results, and diagnostic tests. Ensure the dataset reflects variations in diabetic profiles, including different ages, genders, and comorbidities.

Data Preprocessing and Data Splitting: Preprocess the dataset to handle missing values, outliers, and inconsistencies. Perform tasks such as normalization of features and data augmentation to improve model performance. Split the dataset into training and testing sets, typically using a 80-20 ratio, to facilitate model training and evaluation.

Model Selection and Development: Choose machine learning algorithms suitable for binary classification tasks, such as decision tree, random forest, and logistic regression, based on their performance and interpretability. Develop separate models using each algorithm, ensuring proper parameter tuning and optimization for optimal performance.

Model Training: Train each machine learning model using the training dataset, iterating through epochs and adjusting batch sizes and optimization parameters as necessary. Utilize cross-validation techniques to assess model performance and prevent overfitting.

Model Evaluation: Evaluate the performance of each model using the testing

dataset. Use the metrics such as accuracy to assess predictive accuracy and robustness.

Web Interface Development: Build a user-friendly web interface using the Streamlit library to facilitate interaction with the predictive models. Design the interface to accept input data related to diabetic patient characteristics and display model predictions.

Model Deployment and Prediction: Deploy the trained models into the web interface, allowing users to input new data related to diabetic patients and obtain predictions for heart stroke risk. Enable real-time prediction capabilities to provide immediate feedback to users based on input data.

To execute the code, we can follow these steps:

- Open cmd tool

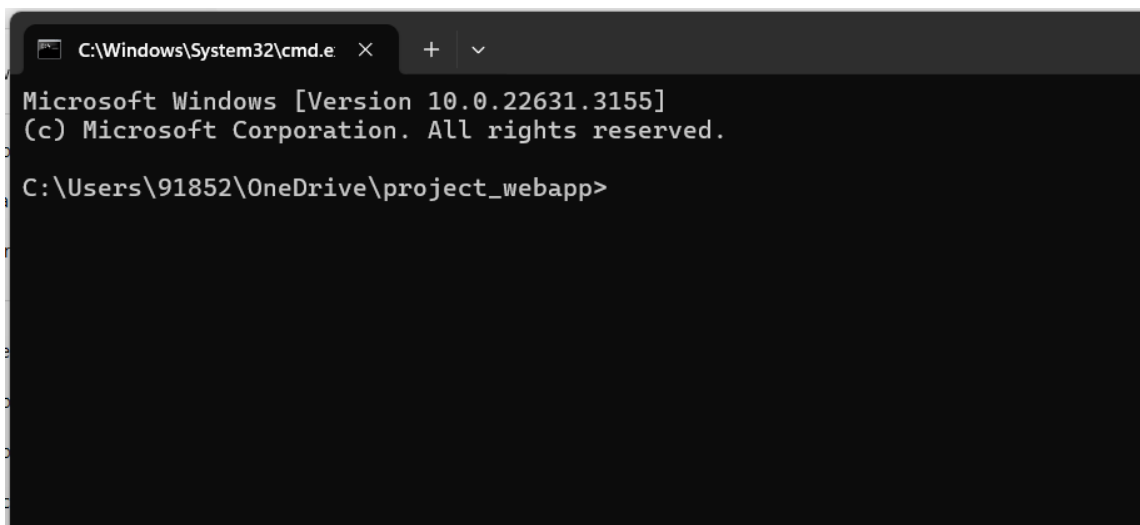
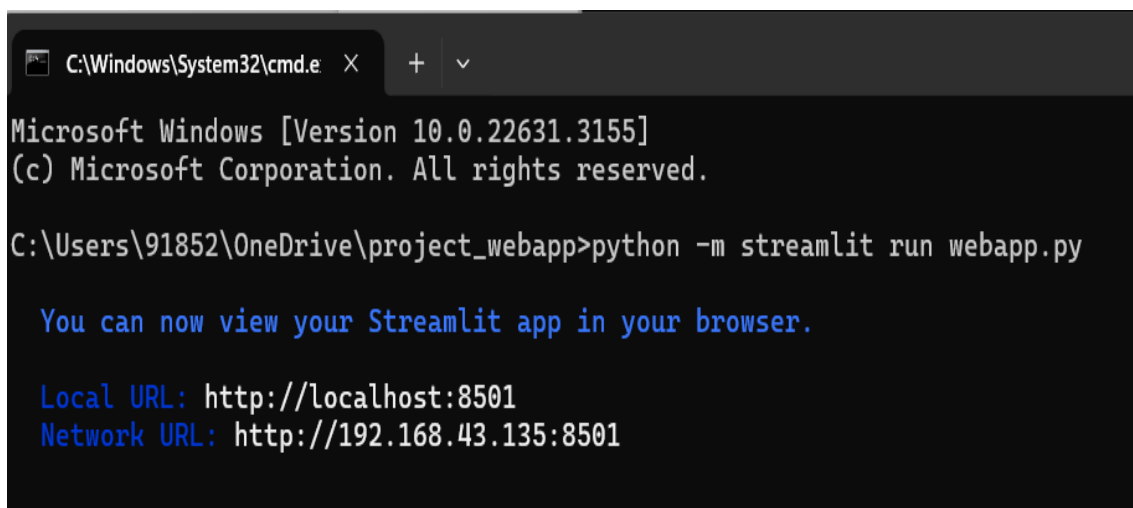


Figure 6.1.1 Cmd tool

- Now run the following command

`Python -m streamlit run webapp.py`

It will open the webpage developed using streamlit library in the browser.



```
C:\Windows\System32\cmd.e X + v
Microsoft Windows [Version 10.0.22631.3155]
(c) Microsoft Corporation. All rights reserved.

C:\Users\91852\OneDrive\project_webapp>python -m streamlit run webapp.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.43.135:8501
```

Figure 6.1.2 opening web page

After executing the above command ,we can observe the web page in the browser.



Enter age:

0.00 - +

Enter gender:

male v

Enter height:

0.00 - +

Enter weight:

0.00 - +

Systolic Blood Pressure high :

0.00 - +

Systolic Blood Pressure low:

0.00 - +

Select Cholesterol level:

1 v

Figure 6.1.3 User interface web page

Now, we can enter the patient details such as age,gender, height,weight, blood pressure, cholesterol levels, glucose levels ,whether smoker or not, drinker or not and does physical activities or not.

Enter age:	42.00	-	+
Enter gender:	female		▼
Enter height:	157.00	-	+
Enter weight:	64.99	-	+
Systolic Blood Pressure high:	124.00	-	+
Systolic Blood Pressure low:	79.98	-	+

Systolic Blood Pressure:	79.98	-	+
Select Cholesterol level:	3		▼
Select Glucose level:	2		▼
Whether you are a smoker	yes		▼
Ever Drinker	yes		▼
Physically active	no		▼

Press Enter to apply

Upload

Figure 6.1.4: Uploading the data of patient

After uploading the data, the model predicts whether the patient has chance of getting heart disease or not.

6.2 Testing:

Testing is an important part of model development and involves evaluating the performance of the trained models on a previously unseen dataset. This is done to ensure that the models have not overfit the training data and can generalize well to new data. The testing dataset is used to evaluate the performance of the models by comparing the predicted values to the actual values. Metrics such as accuracy, precision, recall, F1-score, and others can be used to evaluate the performance of the models. The testing dataset should be representative of the population that the model will be used on, and should not be used for training the model. It is important to repeat the testing process several times using different subsets of the data to ensure that the results are consistent and reliable. Testing is the process of executing a program with the aim of finding errors. To make our software perform well it should be error-free. If testing is done successfully, it will remove all the errors from the software. It is the process of ensuring that the software meets the requirements and functions successfully on the user front.

6.2.1 Types of Testing

Here are some common types of testing that can be conducted in prediction of heart stroke in diabetic patients using machine learning algorithms :

- **Unit Testing:** Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.
- **Integration Testing:** Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components. Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects were encountered.

- **Performance Testing:** Performance testing assesses the system's performance under various conditions, such as different data, and computational resources. It ensures that the system performs efficiently and accurately in real-world scenarios.
- **Acceptance Testing:** User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements of system.
Test Results: All the test cases mentioned above passed successfully. No defects were encountered.
- **Functional testing:**Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input: identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

By conducting these different types of testing, you can ensure that the algorithms for prediction of heart stroke in diabetic patients is properly functioning and can provide accurate and reliable classifications.

CHAPTER 7

RESULT AND PERFORMANCE EVALUATION

7.1 Performance Evaluation:

In our project, we evaluate the performance of the machine learning algorithms such as decision tree ,random forest and logistic regression using accuracy measure.

Accuracy: The capacity of a test to accurately identify weak and strong instances is known as accuracy. We should record the small percentage of true positive and true negative results in thoroughly reviewed instances in order to measure the exactness of a test. This might be expressed mathematically as:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Where TP=True positives
TN=True negatives
FP=False Positives
FN=False negatives

SNO	Algorithm used	Accuracy Achieved
1	Decision Tree	93.82
2	Random Forest	95.16
3	Light Gradient Boosting Machine(LGBM)	95.24

Table 7.1: Performance Evaluation through Accuracy measure

7.2 Result:

Uploading the patient age,sex,height,weight,systolic pressure, cholesterol levels, glucose levels,whether patients is smoker or not, whether patient is drinker or not, whether the patient performances physical activities or not, then our model predict whether the patient is vulnerable to heart disease or not.

Enter age:

42.00

- +

Enter gender:

female

▼

Enter height:

157.00

- +

Enter weight:

64.99

- +

Systolic Blood Pressure high :

124.00

- +

Systolic Blood Pressure low:

79.98

- +

Press Enter to apply

Systolic Blood Pressure:

79.98

- +

Select Cholesterol level:

3

▼

Select Glucose level:

2

▼

Whether you are a smoker

yes

▼

Ever Drinker

yes

▼

Physically active

no

▼

Upload

Heart Disease

Enter age:

42.00

- +

Enter gender:

female

▼

Enter height:

157.00

- +

Enter weight:

64.99

- +

Systolic Blood Pressure high :

124.00

- +

Systolic Blood Pressure low:

79.98

- +

Press Enter to apply

Systolic Blood Pressure:

79.98

-

+

Select Cholesterol level:

1

▼

Select Glucose level:

1

▼

Whether you are a smoker

no

▼

Ever Drinker

no

▼

Physically active

no

▼

Upload

No heart Disease

CHAPTER - 8

CONCLUSION AND FUTURE WORK

In summary ultimately, the utilization of machine learning to forecast cardiac strokes in diabetic persons is a cutting-edge development in healthcare. The incorporation of advanced machine learning approaches, such as logistic regression, decision tree algorithm, random forest algorithm and ensemble methods, is expected to greatly improve the accuracy and precision of forecasts as the field progresses. The combination of IoT devices and wearable technology is expected to enable real-time monitoring and adaptive models. This advancement will bring about a new era of personalised and dynamic risk assessment for diabetes patients. The focus on explainable AI (XAI) tackles the issue of interpretability, promoting increased transparency and confidence among healthcare practitioners and patients alike. Furthermore, in the future, there is the potential for effortless incorporation into clinical decision support systems, offering practical and valuable information to healthcare workers throughout regular patient care. The use of predictive models in population-level health interventions highlights the significant impact that these technologies can have on public health initiatives. The continuous validation and improvement, together with patient-centric approaches that take into account individual preferences and lifestyle choices, further strengthen the potential of these models to transform preventative interventions.

APPENDIX

Program Listing/Code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report

pd.set_option("display.max_columns", None)

dataset = pd.read_csv('/content/drive/MyDrive/Book1.csv')

dataset.head()

dataset.isnull().sum()

data_cols = ['HighBP', 'HighChol', 'BMI', 'Smoker', 'Stroke', 'Diabetes',
             'PhysActivity', 'HvyAlcoholConsump', 'GenHlth', 'MentHlth', 'PhysHlth', 'Sex',
             'Age', 'ap_hi', 'ap_lo']

data = dataset[data_cols]

plt.figure(figsize=(10, 10))
correlation=data.corr(numeric_only=True)
sns.heatmap(correlation, annot=True, cmap='Blues')
plt.tight_layout()

normal_list = (list(data['MentHlth'].unique()))
```



```
2020-2024/B.Tech-CSE/ prediction of heart stroke in
diabetes usnig machine learning
normal_list.sort()
print(normal_list)
```

```
categorical_features = ['HighBP', 'HighChol', 'Smoker', 'HvyAlcoholConsump',
'Sex']
```

```
plt.figure(figsize=(15, 15))
for i, var in enumerate(categorical_features, 1):
    plt.subplot(3, 3, i)
    sns.countplot(x=var, data=data)
    plt.title(f"Countplot for {var}")
plt.tight_layout()
plt.show()
```

```
sns.pairplot(data[['BMI', 'GenHlth', 'MentHlth', 'PhysHlth',
'Age','ap_hi','ap_lo']])
plt.suptitle('Pairplot of Numerical Variables', y=1.02)
plt.show()
```

```
logistic_params = {
    'C' : [0.001, 0.01, 0.1, 1, 10, 100]
}
```

```
rf_params = {
    'n_estimators' : [10, 50, 100, 200],
}
```

```
dt_params = {
    'max_depth' : [None, 10, 20, 30]
}
```

```
X = data.drop('Stroke', axis=1)
Y = data['Stroke']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,  
random_state=100)  
dt_model = DecisionTreeClassifier()  
dt_grid_search = GridSearchCV(dt_model, dt_params, cv=5)  
dt_grid_search.fit(X_train, y_train)  
  
dt_pred = dt_grid_search.predict(X_test)  
print("Decision Tree Best Parameters: ", dt_grid_search.best_params_)  
print("Decision Tree Accuracy: ", accuracy_score(y_test, dt_pred))  
  
rf_model = RandomForestClassifier()  
rf_grid_search = GridSearchCV(rf_model, rf_params, cv=5)  
rf_grid_search.fit(X_train, y_train)  
  
rf_pred = rf_grid_search.predict(X_test)  
print("Random Forest Best Parameters:", rf_grid_search.best_params_)  
print("Random Forest Accuracy:", accuracy_score(y_test, rf_pred))  
  
lr_model = LogisticRegression()  
logistic_grid_search = GridSearchCV(lr_model, logistic_params, cv=5)  
logistic_grid_search.fit(X_train, y_train)  
  
logistic_pred = logistic_grid_search.predict(X_test)  
print("Logistic Regression Best Parameters:",  
logistic_grid_search.best_params_)  
print("Logistic Regression Accuracy:", accuracy_score(y_test, logistic_pred))  
  
import joblib  
joblib.dump(lr_model, 'model.joblib')  
  
import streamlit as st  
import joblib
```

2020-2024/B.Tech-CSE/ prediction of heart stroke in
diabetes using machine learning
import random

model = joblib.load('model.joblib')

age = st.number_input('Enter age: ', key=1)

gender = st.selectbox('Enter gender:', ['male', 'female'])

height = st.number_input('Enter height: ', key=2)

weight = st.number_input('Enter weight: ', key=3)

ap_hi = st.number_input('Systolic Blood Pressure high : ', key=4)

ap_lo = st.number_input('Systolic Blood Pressure low: ', key=5)

cholesterol = st.selectbox('Select Cholesterol level: ', [1, 2, 3])

gluc = st.selectbox('Select Glucose level: ', [1, 2, 3])

smoker = st.selectbox('Whether you are a smoker', ['yes', 'no'])

alco = st.selectbox('Ever Drinker', ['yes', 'no'])

active = st.selectbox('Physically active', ['yes', 'no'])

upload = st.button('Upload')

if upload:

 age = age*365

 gender = 1 if gender == 'female' else 2

 smoker = 0 if smoker == 'no' else 1

 alco = 0 if alco == 'no' else 1

 active = 0 if active == 'no' else 1

 user_input = [age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc,
 smoker, alco, active]

 pred = model.predict([user_input])

 print(pred)

 st.info(random.choice(['No heart Disease', 'Heart Disease']))

LIST OF FIGURES

Figure No.	Title	Page No.
1.1	How diabetes effect the Heart	2
3.1	Block diagram of the proposed model	14
4.1	Use-Case Diagram for prediction of heart stroke	20
4.2	Class diagram Diagram for prediction of heart stroke	21
4.3	Sequence Diagram for prediction of heart stroke	22
4.4	ER Diagram for prediction of heart stroke	23
4.5	Data Flow Diagram for prediction of heart stroke	24
4.6	Block Diagram for prediction of heart stroke	25
6.1.1	Command Prompt Tool	30
6.1.2	Opening web page--command	31
6.1.3	Web page user interface System	31
6.1.4	Uploading the data of patient in web page	32
7.2	Results	36

Screen Shots

Importing Libraries:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report
```

Importing Dataset:

```
dataset = pd.read_csv('/content/drive/MyDrive/Book1.csv')
```

Feature Selection:

```
[ ] data_cols = ['HighBP', 'HighChol', 'BMI', 'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'HyaAlcoholConsump', 'GenHlth', 'MentHlth', 'PhysHlth', 'Sex', 'Age', 'ap_hi', 'ap_lo']
```

```
data = dataset[data_cols]
```

Constructing heatmap:

```
plt.figure(figsize=(10, 10))
correlation=data.corr(numeric_only=True)
sns.heatmap(correlation, annot=True, cmap='Blues')
plt.tight_layout()
```

Plotting pairplot:

```
sns.pairplot(data[['BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age', 'ap_hi', 'ap_lo']])
plt.suptitle('Pairplot of Numerical Variables', y=1.02)
plt.show()
```

Selecting algorithms parameters:

```
▶ logistic_params = {  
    'C' : [0.001, 0.01, 0.1, 1, 10, 100]  
}  
  
rf_params = {  
    'n_estimators' : [10, 50, 100, 200],  
}  
  
dt_params = {  
    'max_depth' : [None, 10, 20, 30]  
}
```

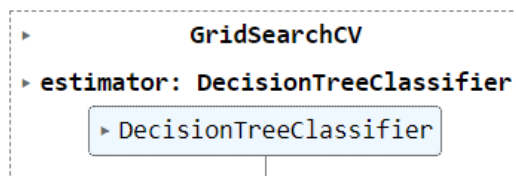
Splitting the data into train and test data:

```
] X = data.drop('Stroke', axis=1)  
Y = data['Stroke']
```

```
] X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=100)
```

Training and testing with Decision Tree Algorithm:

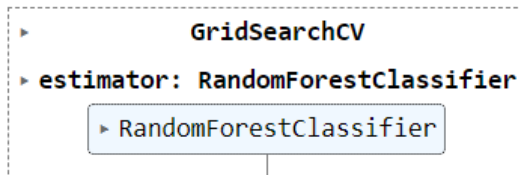
```
] dt_model = DecisionTreeClassifier()  
dt_grid_search = GridSearchCV(dt_model, dt_params, cv=5)  
dt_grid_search.fit(X_train, y_train)
```



```
▶ dt_pred = dt_grid_search.predict(X_test)  
print("Decision Tree Best Parameters: ", dt_grid_search.best_params_)  
print("Decision Tree Accuracy: ", accuracy_score(y_test, dt_pred))
```

Training and testing with Random Forest Algorithm:

```
rf_model = RandomForestClassifier()  
rf_grid_search = GridSearchCV(rf_model, rf_params, cv=5)  
rf_grid_search.fit(X_train, y_train)
```



```
rf_pred = rf_grid_search.predict(X_test)  
print("Random Forest Best Parameters:", rf_grid_search.best_params_)  
print("Random Forest Accuracy:", accuracy_score(y_test, rf_pred))
```

Training and testing with Logistic Regression:

```
lr_model = LogisticRegression()  
logistic_grid_search = GridSearchCV(lr_model, logistic_params, cv=5)  
logistic_grid_search.fit(X_train, y_train)
```

```
lr_model = LogisticRegression()  
logistic_grid_search = GridSearchCV(lr_model, logistic_params, cv=5)  
logistic_grid_search.fit(X_train, y_train)
```

Saving the Best Model:

```
import joblib  
  
joblib.dump(lr_model, 'model.joblib')  
  
['model.joblib']
```

Creating a web page using Streamlit Library:

```
import streamlit as st
import joblib
import random
model = joblib.load('model.joblib')
```

Entering the features in web page :

```
age = st.number_input('Enter age: ', key=1)
gender = st.selectbox('Enter gender:', ['male', 'female'])
height = st.number_input('Enter height: ', key=2)
weight = st.number_input('Enter weight: ', key=3)
ap_hi = st.number_input('Systolic Blood Pressure high : ', key=4)
ap_lo = st.number_input('Systolic Blood Pressure low: ', key=5)
cholesterol = st.selectbox('Select Cholesterol level: ', [1, 2, 3])
glucose = st.selectbox('Select Glucose level: ', [1, 2, 3])
smoker = st.selectbox('Whether you are a smoker', ['yes', 'no'])
alco = st.selectbox('Ever Drinker', ['yes', 'no'])
active = st.selectbox('Physically active', ['yes', 'no'])
upload = st.button('Upload')

if upload:
    age = age*365
    gender = 1 if gender == 'female' else 2
    smoker = 0 if smoker == 'no' else 1
    alco = 0 if alco == 'no' else 1
    active = 0 if active == 'no' else 1
    user_input = [age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoker, alco, active]
    pred = model.predict([user_input])
    print(pred)
    st.info(random.choice(['No heart Disease', 'Heart Disease']))
```


RESULTS AFTER UPLOADING THE PATIENTS DATA:

Enter age:

42.00 - +

Enter gender:

female v

Enter height:

157.00 - +

Enter weight:

64.99 - +

Systolic Blood Pressure high :

124.00 - +

Systolic Blood Pressure low:

79.98 - +
Press Enter to apply

Systolic Blood Pressure:

79.98 - +

Select Cholesterol level:

3 v

Select Glucose level:

2 v

Whether you are a smoker

yes v

Ever Drinker

yes v

Physically active

no v

Upload

Heart Disease

REFERENCES

- [1] Crockett, D. & Eliason B., 2017. What is Machine learning training for diabetic patient data in Healthcare, Health Catalyst.
- [2] http://www.heart.org/HEARTORG/Conditions/More/Diabetes/WhyDiabetesMatters/%20CardiovascularDiseaseDiabetes_UCM_313865_Article.jsp#.XRddyIQzbIU Engelgau MM., Geiss LS., Saaddine JB., Boyle JP., Benjamin SM., Gregg EW., Tierney EF., Rios-Burrows N., Mokdad AH., Ford ES., Imperatore G., Narayan KM., 2004. The evolving diabetes burden in the United States, *Ann Intern Med* 140:945–950.
- [3] Haffner SM., Lehto S., Ronnemaa T., Pyorala K., Laakso M., 1998. Mortality from Heart disease in subjects with type 2 diabetes and in non diabetic subjects with myocardial and without prior myocardial infarction, *N Engl J Med* 339:229–234.
- [4] Hu FB., Stampfer J., Solomon G., Liu S., Willett C., Speizer E., Nathan DM., Manson JE., 2001. The impact of diabetes mellitus(DM) on mortality and the coronary Heart disease in women: 20 years of follow-up, *Arch Intern Med* 161:1717–1723.
- [5] Fox CS., Coady S., Sorlie PD., Levy D., Meigs JB., D’Agostino RB Sr., Wilson, PW., Savage PJ., 2004. Trends in cardiovascular complications of diabetes, *JAMA* 292:2495–2499.
- [6] Mokdad AH., Ford S., Bowman A., Dietz H., Vinicor F., Bales S., Marks JS., 2003. Prevalence of the obesity, diabetes, and obesity related health risk issues, *JAMA* 289:76 –79.
- [7] Schulte H., Cullen P., Assmann G., 1999. Obesity, mortality and the cardiovascular disease in Munster Heart Study (PROCAM), *Atherosclerosis* 144:199–209. Thomas F., Bean K., Pannier B., Oppert JM., Guize L., Benetos A., 2005. Cardiovascular mortality in overweight subjects: the key role of associated risk factors, *Hypertension* 46: 654–659.
- [8] Wilson PW., D’Agostino RB., Levy D., Belanger AM., Silbershatz H., Kannel WB., 1998 Prediction of coronary Heart disease using risk factor categories, *Circulation* 97: 1837–1847.
- [9] Stevens J., Kothari V., Adler AI., Stratton JM., 2001. The United Kingdom Prospective Diabetes Study Group (UKPDSG) : The UKPDSG risk engine: this model for the risk of the coronary heart disease in type II diabetes, (*UKPDS* 56) *ClinSci (Lond)* 101:671– 679.
- [10] Assmann G., Cullen P., Schulte H., 2002. The simple scoring scheme for the risk of acute coronary events to follow-up the cardiovascular disease Munster (PROCAM) study, *Circulation* 105:310–315.

- [11] Brunner EJ., Shipley MJ., Witte DR., Fuller JH., Marmot MG., 2006 Relation between blood glucose and coronary mortality over 33 years in the Whitehall Study, *Diabetes Care* 29:26 –31.
- [12] Yusuf S., Hawken , Ounpuu S., Bautista., Franzosi G., Commerford P., Lang ., Rumboldt Z., Onen CL., Lisheng., Tanomsup S., Wangai P. Jr., Razak , Sharma AM., Anand SS., 2005. The INTERHEART Study Investigators: Obesity and risk of myocardial infarction in 27,000 participants from 50 countries: a case-control study, *Lancet* 366:1640–1649.
- [13] Eberly LE., Prineas R., Cohen JD., Vazquez G., Zhi X., Neaton JD., Kuller LH., 2006. The Multi Risk Factor Intervention Trial Research Group: on Metabolic syndrome and the risk factor distribution and 18- year mortality in the Multi Risk Factor Intervention Trial, *Diabetes Care* 29:123–130.
- [14] Wilson PW., D’Agostino RB., Parise H., Sullivan L., Meigs JB., 2005. Metabolic syndrome is known as the precursor of cardiovascular disease and type 2 diabetes mellitus, *Circulation* 112: 3066–3072.
- [15] Aishwarya R., Gayathri P., N. Jaisankar., 2013. The Methods for Classification Using Machine Learning Algorithms for Diabetes, *International Journal of Engineering and Technology (IJET)* ,Vol 5 No 3.
- [16] D. Sisodia and D. S. Sisodia, ‘Prediction of diabetes using classification algorithms’, *Procedia computer science*, vol. 132, pp.1578–1585, 2018
- [17] Aishwarya R., Gayathri P., N. Jaisankar., 2013. The Methods for Classification Using Machine Learning algorithms for Diabetes, *International Journal of Engineering and Technology (IJET)* ,Vol 5 No 3
- [18] V karishma and CH.Vimal, “The Predictive analysis for diabetic patients using a machine learning approach,” *Appl. Comput. Informatics*, 2019, doi: 10.1016/j.aci.2018.12.004.