LIFESTYLE DISEASE PREDICTION

A PROJECT REPORT

Submitted by,

Mr. FAIZ KHAN - 20201CSD0115
Mr. JEEVITH JOSEPH CJ - 20201CSD0111
Mr. TEJAS GOWDA C - 20201CSD0072

Under the guidance of,

Dr.CHANDRASEKAR V

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

At

PRESIDENCY UNIVERSITY

BENGALURU

JANUARY 2024

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report "LIFESTYLE DISEASE PREDICTION" submitted by **FAIZ KHAN (20201CSD0115), JEEVITH JOSEPH CJ (20201CSD0111), TEJAS GOWDA C(20201CSD0072)** in partial fulfilment of requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering (Data Science) is a bonafide work carried out under my supervision.

Dr.CHANDRASEKAR V
School of CSE
Presidency University

Dr. JAYACHANDRAN A
Professor & HoD
School of CSE
Presidency University

PRESIDENCY UNIVERSITY

Dr. C. KALAIARASAN
Associate Dean
School of CSE&IS
Presidency University

Dr. L. SHAKKEERA
Associate Dean
School of CSE&IS
Presidency University

Dr. SAMEERUDDIN KHAN
Dean
School of CSE&IS
Presidency University

SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **LIFESTYLE DISEASE PREDICTION** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering (Data Science)**, is a record of our own investigations carried under the guidance of Dr.CHANDRASEKAR V**, School of Computer Science & Engineering Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Deg ree.

NAME :	ROLL NUMBER :	SIGNATURE :
FAIZ KHAN	**20201CSD0115**	
JEEVITH JOSEPH CJ	**20201CSD0111**	
TEJAS GOWDA C	**20201CSD0072**	

ABSTRACT

The contemporary era witnesses a surge in lifestyle-related diseases, necessitating proactive and personalized health management strategies. The project, titled "Lifestyle Disease Prediction," is an innovative initiative designed to address the growing concern of preventable health conditions stemming from modern lifestyles.

This project leverages advanced data science and machine learning techniques to predict the likelihood of lifestyle-related diseases based on individual behavioral patterns, demographic information, and health history. By analyzing comprehensive datasets, encompassing diverse lifestyle factors, the system aims to provide early indications and risk assessments for diseases associated with sedentary habits, dietary choices, and stress.

The project encompasses the development of robust predictive models, employing algorithms such as decision trees, ensemble methods, and neural networks. It focuses on the integration of real-time health data, wearable device inputs, and historical health records to enhance the accuracy and reliability of predictions.

Key features include a user-friendly interface for data input, dynamic risk assessment, and personalized health recommendations. The project's goal is to empower individuals to make informed decisions about their lifestyles, facilitating proactive health management and disease prevention.

Key features used in the prediction model include dietary habits, physical activity levels, smoking status, alcohol consumption, family medical history, and genetic markers. The model is trained on a diverse dataset to ensure robust predictions across different demographics and geographic regions. Continuous updates to the model are made possible through real-time health monitoring, ensuring adaptability to changing lifestyle patterns.

The "Lifestyle Disease Prediction" project contributes to the evolving field of preventive healthcare by harnessing the potential of data science to create a tool that aids individuals in adopting healthier lifestyles and mitigating the risks associated with modern-day living.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to several individuals and our organization for supporting me through this project.

We express our sincere thanks to our respected dean Dr. Md. Sameeruddin Khan, Dean, School of Computer Science Engineering, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans Dr. Kalaiarasan C and Dr. Shakkeera L, School of Computer Science Engineering, Presidency University and Dr.Jayachandran A Head of the Department, School of Computer Science Engineering, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide Dr.CHANDRASEKAR V, School of Computer Science Engineering, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators Dr. Sanjeev P Kaulgud, Dr. Mrutyunjaya MS and also the department Project Coordinator Ms. Manasa C M.I am also grateful to Dr. K Susheel Kumar for their consistent support and assistance during this project.

Thanks for all your encouragement!

FAIZ KHAN
JEEVITH JOSEPH CJ
TEJAS GOWDA C

LIST OF TABLES

Sl. No.	Table Name	Table Caption	Page No.
1	Table 1	Diabetes Prediction Data	17
2	Table 2	Accuracy of All Models for Diabetes Prediction	48
3	Table 3.	Ensemble Accuracy for Diabetes Prediction	49
4	Table 4.	Accuracy of Hybrid Models	50
5	Table 5.	Heart Disease Prediction Data	51
5	Table 6.	Statistical Summary of Heart Disease Data	60
6	Table 7.	Evaluation Metrics for All the Hybrid Models	62
7	Table 8.	Ensemble Accuracy of Heart Disease Prediction	65
8	Table 9.	Ensemble Accuracy of CKD Prediction	70
9	Table 10.	CNN Accuracy for Pneumonia Prediction	72

LIST OF FIGURES

Sl. No.	Figure Name	Caption	Page No.
1.	Figure 1.	Pair plot for diabetes prediction for diabetes	12
2	Figure 2.	Plotting the data using count plot for diabetes	18
3	Figure 3	Distribution of data using histogram for diabetes	19
4	Figure 6	Detecting the outliers for diabetes	25
5	Figure 7	Normalizing the data using Quantile Transform for diabetes	27
6	Figure 10	Fitting the model for all the hybrid models defined for diabetes	30
7	Figure 11	Making predictions for the training data for diabetes	32
8	Figure 12	Comparison of Accuracy and Precision for different modes for diabetes	32
9	Figure 13	Fitting the model using ensemble for diabetes	33
10	Figure 14	Classification Report for Diabetes Prediction	33
11	Figure 17	Finding the missing values for heart disease	37
12	Figure 18	Count plot of given heat disease data	38
13	Figure 19	Distribution of the data using histogram for heart disease	40
14	Figure 20	Correlation of Heart disease data	41
15	Figure 21	Detecting the Outliers for heart disease	42
16	Figure 22	Interactive 3D scatter plot for heart disease	43
17	Figure 23	Comparison of Accuracy and Precision for Different Models for heart disease	46
18	Figure 24	Performance Evaluation of Heart Disease Prediction for heart disease	s

TABLE OF CONTENTS

Chapter No.	Title	Page No.
	Title Page	i
	Certificate	ii
	Declaration	iii
	Abstract	iv
	Acknowledgement	v
	List Of Tables	vi
	List Of Figures	vii
1.	Introduction	1
2.	Literature Survey	3
3	Research Gaps OF Existing Methods	7
4	Proposed Methodology	12
5	Objective	14
6	**Lifestyle Disease Prediction**	15
	Diabetes Prediction	18
	Data Collection and Preprocessing	
	Data Sourcing	
	1.2 Data Cleaning	
	1.3 Data Split	
	Exploratory Data Analysis (EDA)	
	2.1 Descriptive Analysis	
	2.2 Visualization	
	Feature Selection and Engineering	
	3.1 Feature Importance	
	Algorithm Selection	
	4.1 Model Implementation	
	Model Evaluation	
	5.1 Comparative Analysis and Evaluation	
	5.2 Best Model Selection	
	Heart Disease Prediction	30
	Data Collection and Preprocessing	
	1.1Data Sourcing	
	1.2Data Cleaning	
	1.3Data Split	
	Exploratory Data Analysis (EDA)	
	2.1 Descriptive Analysis	
	2.2 Visualization	
	Feature Selection and Engineering	

3.1 Feature Importance
Algorithm Selection
4.1 Model Implementation
Model Evaluation
5.1 Comparative Analysis and Evaluation
5.2 Best Model Selection

Chronic Kidney Disease Prediction
Data Collection and Preprocessing
1.1Data Acquisition
1.2 Image Preprocessing
1.3 Data Augmentation
Exploratory Data Analysis (EDA)
2.1 Descriptive Analysis
2.2 Visualization of Data
Feature Selection and Engineering
3.1 Feature Importance
Algorithm Selection
4.1 Model Implementation
Model Evaluation
5.1 Comparative Analysis and Evaluation
5.2 Best Model Selection

Pneumonia Detection using CNN model
Data Collection and Preprocessing
1.1Data Acquisition
1.2 Image Preprocessing
1.3 Data Augmentation
Exploratory Data Analysis (EDA)
2.1 Descriptive Analysis
2.2 Visualization of Pneumonia Images
Model Architecture Design
3.1 Convolutional Neural Network (CNN) Architecture
3.2 Transfer Learning (if applicable)
Model Implementation
4.1 Building the CNN Model
4.2 Compilation of the Model
Training the CNN Model
5.1 Data Split for Training and Validation
5.2 Model Training
5.3 Monitoring Training Progress
Model Evaluation
6.1 Evaluation Metrics (accuracy, precision, recall)
6.2 Confusion Matrix Analysis
Fine-Tuning and Optimization
7.1 Hyper-parameter Tuning 7
7.2 Model Fine-Tuning
Interpretation and Visualization

8.1 Visualizing Activations or Filters
8.2 Interpretability of CNN Predictions
Testing and Inference
9.1 Using Test Data
9.2 Model Inference on New Images
Comparative Analysis and Evaluation
10.1 Benchmarking Against Existing Models
10.2 Comparative Analysis of Model Performance

7	Timeline For Execution of Project	44
8	Outcomes	45
	1. Identification of Optimal Algorithms	
	2. Performance Metrics Insights	
	3. Model Interpretability	
9	Results And Discussions	48
10	Conclusion	51

CHAPTER-1

INTRODUCTION

.

In an era where data and technology are transforming every aspect of our lives, healthcare is no exception. With the rise of lifestyle related diseases, such as heart disease, diabetes and chronic kidney disease , the need for innovative and cost-effective preventive healthcare solutions is needed. Early detection and intervention are key to reducing these diseases. This project sets out to explore healthcare that uses the power of detailed demographic and vital statistics to predict the likelihood of lifestyle diseases, enabling timely interventions and substantial cost savings in the long run.

This project will help users to identify lifestyle diseases which are a big threat to humans, will bring awareness about the diseases and will help people to be healthy which is of utmost importance. In this project, we will dive into the methods and technologies necessary to make use of the potential of data-driven prediction in healthcare. By integrating technology, data, and medical expertise, we aim to pave the way for a healthcare system that not only extends lives but also makes them healthier and more affordable. This proactive approach holds the promise of not only enhancing the quality of life for individuals but also reducing the economic burden of lifestyle diseases on healthcare systems worldwide.

Among the prominent lifestyle diseases, this exploration explores into the complexities of diabetes, heart disease, chronic kidney disease, and pneumonia. Each of these conditions is not only shaped by genetic but also indirectly connected to the way individuals live their lives

Diabetes

Diabetes, a hallmark lifestyle disease, is intricately linked to long-term behavioral patterns. Lifestyle choices, including diet and physical activity, significantly influence the development of this chronic metabolic disorder. Subtopic 1 sheds light on the multifaceted nature of diabetes, from the impact of genetic predispositions to the crucial role of lifestyle choices and environmental influences. Recognizing the intricate relationship between lifestyle and diabetes prevalence is essential for designing effective preventive strategies.
The application of ensemble methods in diabetes prediction has gained prominence. Combining the strengths of multiple predictive models, ensemble techniques enhance the robustness and accuracy of diabetes predictions. Methods such as Random Forests and Gradient Boosting, when integrated into an ensemble framework, offer a more comprehensive approach to understanding and predicting diabetes risk factors.

Heart Disease

Heart disease, often exacerbated by lifestyle factors, represents a culmination of cardiovascular disorders influenced by how individuals live their lives. Subtopic 2 navigates through the diverse spectrum of heart diseases, emphasizing the role of lifestyle choices in conditions such as coronary artery disease, heart failure, and arrhythmias. Understanding the connection between lifestyle and heart disease complexity guides healthcare professionals in formulating targeted interventions for prevention and treatment.

Ensemble methods have proven effective in enhancing the accuracy of heart disease prediction models. By combining the predictive capabilities of various algorithms, ensemble techniques such as bagging and boosting offer a synergistic approach. Integrating diverse models, including decision trees and support vector machines, into an ensemble framework empowers healthcare professionals with more reliable tools for risk assessment and treatment planning.

Chronic Kidney Disease

Chronic Kidney Disease (CKD) is a prevalent health condition influenced significantly by lifestyle factors, representing a broad range of disorders affecting the kidneys. This note explores the intricate relationship between lifestyle choices and CKD, shedding light on the multifaceted nature of the disease. Lifestyle factors such as diet, physical activity, and substance abuse play pivotal roles in the development and progression of CKD, underscoring the importance of preventive measures.

Ensemble methods, akin to their effectiveness in predicting heart disease, have emerged as valuable tools in enhancing the accuracy of CKD prediction models. By amalgamating diverse algorithms through techniques like bagging and boosting, ensemble methods provide a synergistic approach to CKD prediction. Integrating various models, such as decision trees and support vector machines, into an ensemble framework equips healthcare professionals with more robust tools for risk assessment and tailored intervention strategies.

Pneumonia

Pneumonia, although infectious in nature, can be influenced by lifestyle-related factors that compromise the immune system. Subtopic 4 explores the intricate connection between lifestyle and respiratory health, considering how factors such as age, pre-existing conditions, and weakened immune systems contribute to pneumonia susceptibility. This understanding forms the basis for tailored preventive measures and improved healthcare interventions.

In the domain of pneumonia prediction, Convolutional Neural Networks (CNNs) have become potent tools for image-based diagnostics. By analyzing chest X-rays and other imaging data, CNNs assist healthcare professionals in swift and accurate identification of pneumonia patterns. The integration of CNNs into diagnostic protocols enhances the efficiency of pneumonia detection, leading to timely interventions and improved patient outcomes. The collective exploration of these lifestyle diseases underscores the imperative for a holistic and interdisciplinary approach to healthcare.As we move forward, a comprehensive understanding of lifestyle diseases will continue to guide preventive measures, personalized interventions, and advancements in healthcare for a healthier global society.

CHAPTER-2

LITERATURE SURVEY

[1] Diabetes Prediction Using Ensemble of Different Machine Learning Classifiers- In 2020, Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain (Senior Member, IEEE), and Mahmudul Hasan proposed a robust framework for predicting diabetes. The framework incorporated outlier rejection, handling missing values, data standardization, feature selection, K-fold cross-validation, and various Machine Learning (ML) classifiers, including k-nearest Neighbor, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP). Additionally, they introduced a weighted ensemble of different ML models, where weights were determined based on the corresponding Area Under the ROC Curve (AUC) of each ML model to enhance diabetes prediction. AUC served as the performance metric and was maximized during hyperparameter tuning using the grid search technique. All experiments were conducted using the Pima Indian Diabetes Dataset under consistent conditions. The results revealed that the ensemble classifier performed exceptionally well, achieving a sensitivity of 0.789, specificity of 0.934, false omission rate of 0.092, diagnostic odds ratio of 66.234, and an AUC of 0.950. This outperformed state-of-the-art results by 2.00% in AUC.

[2] Diabetes prediction model based on an enhanced deep neural network - In 2020, Huaping Zhou, Raushan Myrzashova, and Rui Zheng proposed a method aimed at predicting future occurrences of diabetes and determining the specific type of the disease a person might be experiencing. This approach is designed to facilitate accurate treatment for patients. The methodology transforms the task into a classification problem, constructing the model primarily using the hidden layers of a deep neural network and employing dropout regularization to prevent overfitting. Parameter tuning is conducted, and the binary cross-entropy loss function is utilized, resulting in a deep neural network prediction model with high accuracy. Experimental results demonstrate the effectiveness and adequacy of the proposed DLPD (Deep Learning for Predicting Diabetes) model. The best training accuracy for the diabetes type dataset is 94.02174%, and for the Pima Indians diabetes dataset, it is 99.4112%. Rigorous experiments were conducted using both the Pima Indians diabetes and diabetic type datasets.

[3] Krishnaiah et al , reviewed the various data mining algorithms used for predicting heart disease. The researchers investigated diverse methodologies employed in various studies. Their results indicated variations in accuracy based on the selection of features and implementation methods across different papers. Notably, their study revealed that the utilization of Fuzzy Intelligent Techniques led to an enhancement in the accuracy of the prediction system.

[4] Chitra et al [2], for predicting heart disease used Supervised Learning Algorithm. The outcomes were contrasted with those obtained using SVM. A Cascaded Neural Network (CNN) classifier was employed for patient record classification. The CNN classifier received 13 attributes as input to assess the disease risk. The dataset encompassed records from 270 patients. The findings demonstrated a higher accuracy for the CNN classifier in comparison to the well-established SVM classifier.

[5]Chronic Kidney Disease Prediction Using Machine Learning Models: In October 2019, authors Revathy, S., Bharathi, B., Jeyanthi, P., and Ramesh, M. published a paper titled "Chronic Kidney Disease Prediction Using Machine Learning Models" in the International Journal of Engineering and Advanced Technology. The study introduced a predictive algorithm designed to identify chronic kidney disease (CKD) in its early stages. The dataset utilized comprised input parameters obtained from CKD patients, and the models underwent training and validation using these parameters. The paper constructed Decision Tree, Random Forest, and Support Vector Machine learning models for the diagnosis of CKD. Model performance was assessed based on the accuracy of predictions.

[6]Chronic kidney Disease prediction using Machine Learning methods: In 2020, Ekanayake, Imesh Udara, and Damayanthi Herath published a paper titled "Chronic Kidney Disease Prediction Using Machine Learning Methods" in the Moratuwa Engineering Research Conference (MERCon), IEEE. The study involved an examination of various machine learning algorithms. Multiple attributes associated with CKD patients were analyzed, and the accuracy of predictions was determined for different machine learning algorithms such as Decision Tree and Support Vector Machine.

[7]EL. Khalid , Asnaoui, Different types of single and ensemble learning models were utilized to classify pneumonia: Ensemble learning is the process of integrating multiple models into a unified model to address a specific task, and the selection of models is based on the requirements and characteristics of the given problem. At present, ensemble models are frequently utilized for prediction tasks such as classification and regression. The strategy involves training individual models independently within the ensemble, leading to enhanced accuracy. Specifically, this study found that an ensemble consisting of three models exhibited superior accuracy. The content includes details from the original source, and appropriate credit is given to the authors.

[8]D. Meldon, Nimesh Naik: The main goal of this research is to assess patient X-ray images using OpenCV and Deep Learning, determining the presence or absence of pneumonia. The study employed Keras libraries and OpenCV to achieve a notable level of accuracy in testing data.

CHAPTER-3

RESEARCH GAPS OF EXISTING METHODS

Research Gap 1: Integrated Model for Multiple Diseases:
Explore the development of an integrated predictive model that simultaneously considers risk factors for heart disease, diabetes, chronic kidney disease, and pneumonia. This could involve a multi-task learning approach or an ensemble model that combines individual disease prediction models.

Research Gap 2: Temporal Dynamics and Longitudinal Data in Lifestyle Disease Prediction Models :
Many existing models in lifestyle disease prediction do not adequately consider the temporal aspects of disease progression. Details Investigate the incorporation of temporal dynamics and longitudinal health data to enhance the accuracy of disease prediction. Understanding how lifestyle factors evolve over time and their impact on disease progression can provide more accurate predictions.

Research Gap 3: Limited Emphasis on Explainability in Lifestyle Disease Prediction
Issue The interpretability and explainability of lifestyle disease prediction models are often given insufficient attention. Details: Existing models may lack transparency in decision-making, potentially impacting trust and acceptance among healthcare professionals and individuals

Research Gap 4: Underexplored Ensemble Approaches in Lifestyle Disease Prediction
There is limited exploration of ensemble methods in predicting lifestyle diseases, such as diabetes and heart disease. Details: While some studies acknowledge the potential of ensemble techniques , a more comprehensive investigation is needed to understand their effectiveness, optimal combinations, and impact on enhancing model robustness.

Research Gap 5: Inadequate Exploration of Hybrid Models in Lifestyle Disease Prediction Issue :
There is a lack of comprehensive exploration of hybrid models that integrate different machine learning techniques, particularly in the context of lifestyle disease prediction. Details: While some studies touch upon hybrid models a deeper investigation is essential to understand their advantages, drawbacks, and optimal configurations in the prediction of lifestyle-related diseases.

Research Gap 6: Limited Incorporation of Demographic and Socioeconomic Factors in Lifestyle Disease Prediction Models Issue :
Many lifestyle disease prediction models may not sufficiently incorporate demographic and socioeconomic factors. Details: The review of existing literature emphasizes a gap in knowledge regarding the influence of demographic and socioeconomic variables on the performance and predictive accuracy of models concerning lifestyle-related diseases. Addressing these research gaps in the prediction of lifestyle-related diseases, including diabetes and heart disease using ensemble methods, as well pneumonia using CNN, is vital for advancing the accuracy, reliability, and applicability of these predictive models in real-world healthcare scenarios.

CHAPTER-4

PROPOSED MOTHODOLOGY

The methodology for lifestyle disease prediction integrates a systematic approach encompassing data collection, preprocessing, feature engineering, machine learning model development, disease prediction, and a user-friendly interface. This holistic framework is designed to ensure accuracy, privacy compliance, and actionable insights for both users and healthcare providers.

1. Data Collection, Preparation, and Preprocessing :

1.1 Data Gathering
Objective: Gather a diverse dataset with demographic information, vital statistics, and lifestyle disease status (e.g., heart disease, diabetes, CKD, pneumonia).
Implementation: Employ robust data collection methods, ensuring inclusivity and representation across demographics.

1.2 Data Quality and Privacy
Objective: Ensure data quality, accuracy, and privacy compliance while handling sensitive information.
Implementation: Anonymize and secure sensitive data, adhering to privacy standards and regulations.

1.3 Dataset Splitting
Objective: Divide the datasets into training data and testing data sets to build model and evaluate the model.
Implementation: Employ a stratified splitting approach to maintain balanced representation across different classes.

1.4 User Input Preprocessing Component
Objective: Develop a preprocessing component for user input data to enhance consistency and prepare it for machine learning algorithms.
Implementation: Employ data cleaning techniques to handle inconsistencies, errors, and missing values in real-time.

2. Feature Selection and Engineering :

2.1 Identifying Relevant Features
Objective: Identify features crucial for lifestyle disease prediction, such as age, gender, BMI, blood pressure, cholesterol, and lifestyle habits.
Implementation: Utilize domain knowledge and statistical methods to select pertinent features.

2.2 Feature Engineering
Objective: Enhance predictive accuracy by creating new variables or transforming existing

ones.
Implementation: Apply techniques like polynomial expansion, interaction terms, and normalization to optimize feature representation.

3. Machine Learning Model Development :

3.1 Algorithm Selection
Objective: Choose suitable machine learning algorithms for disease prediction.
Implementation: Explore algorithms like Decision Trees, Support Vector Machines, Random Forest, and Gradient Boosting, evaluating their performance.

3.2 Model Training
Objective: Train the selected machine learning model using the preprocessed user input data.
Implementation: Employ a robust training regimen, including cross-validation for optimal model performance.

4. Disease Prediction Component :

4.1 Predictive Output Generation
Objective: Develop a component that generates predictions based on the machine learning model output.
Implementation: Set up an interface to display predictions and relevant information to users.

5. User Interface :

5.1 Interface Design
Objective: Create a user-friendly interface for inputting lifestyle factors and receiving predictions.
Implementation: Utilize web development tools to design an intuitive interface, ensuring accessibility and ease of use.

6. Overall Integration :

6.1 System Architecture
Objective: Integrate all components into a cohesive system for lifestyle disease prediction.
Implementation: Design a modular architecture for scalability and maintainability.
Conclusion
This methodology provides a robust foundation for developing a lifestyle disease prediction system, encompassing data integrity, machine learning model development, and user interaction. The meticulous implementation of each component ensures the reliability and effectiveness of the proposed system in real-world healthcare scenarios.

CHAPTER-5

OBJECTIVES

Lifestyle disease prediction models serve several critical objectives in the realm of public health. One primary goal is the early detection and prevention of non-communicable diseases (NCDs). By identifying individuals at risk during the early stages, these models enable timely intervention and implementation of preventive measures. This proactive approach is instrumental in reducing the incidence of lifestyle diseases and mitigating their impact on individuals and healthcare systems.

Timely Detection and Prevention: Lifestyle disease prediction models tailor health strategies based on an individual's unique lifestyle and health profile. By providing targeted recommendations for lifestyle modifications, these models helps individuals to make informed decisions about their health. This personalized approach enhances the effectiveness of preventive measures.

Health promotion : It is also a crucial objective of lifestyle disease prediction. By raising awareness about the profound impact of lifestyle choices on health, these models contribute to a cultural shift toward healthier behaviours. Individuals gain insights into their disease risk factors, encouraging them to adopt healthier lifestyles and thereby reducing the overall burden of lifestyle diseases on society.

Efficient resource allocation: is another key goal, aiming to optimize healthcare resources by focusing on high-risk populations. Early detection and intervention in individuals at risk lead to cost-effective healthcare strategies, ultimately reducing the economic burden associated with treating advanced stages of lifestyle diseases.

Resource Optimization: Lifestyle disease prediction models are instrumental in driving data-driven decision-making in healthcare. By utilizing advanced analytics and machine learning, these models extract valuable insights from vast datasets. These insights inform evidence-based healthcare policies, interventions, and long-term strategies, enhancing the overall efficacy and efficiency of healthcare systems.

Data-Driven Decision Making: The continuous monitoring facilitated by these models ensures a real-time tracking of changes in lifestyle and health parameters. This feature provides regular updates on an individual's disease risk, allowing for timely adjustments to preventive strategies and fostering a dynamic, adaptable healthcare approach.

Empowerment of Individuals: By empowering individuals to take an active role in managing their health, lifestyle disease prediction models contribute to the broader goal of patient empowerment. Informed decision-making becomes a cornerstone of individual health management, enhancing the overall efficacy of preventive measures and interventions.

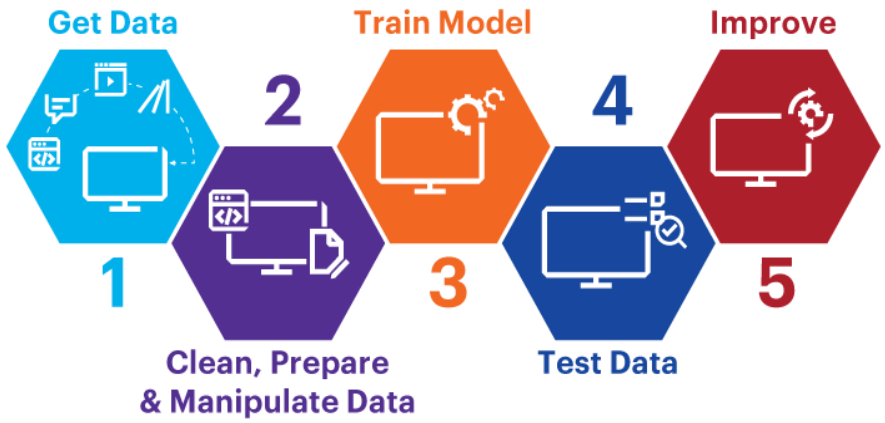
CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

DIABETES PREDICTION USING MACHINE LEARNING

SYSTEM DESIGN

In the initial step, each dataset undergoes preprocessing. Subsequently, the processed datasets are introduced to various machine learning algorithms during the second phase. Moving on to the third stage, diverse metrics are employed to analyze the output generated by the models. In the subsequent phase, the machine learning model demonstrating the highest accuracy is selected for diabetes detection in individuals. This selected model is then integrated into a web-based application created using the Flask framework in the Python programming language.



In summary, the research contributions can be outlined as follows:

Firstly, the study involves the training of multiple machine learning algorithms using four distinct clinical datasets to detect diabetes. The datasets undergo preprocessing through the application of various pre-processing techniques.

Secondly, the research evaluates the performance of each machine learning algorithm across the four datasets, considering parameters such as precision, recall, f1-score, ROC curve, and accuracy. Additionally, significant features or attributes are identified using diverse feature selection methods like correlation and chi-square. These methods aim to pinpoint the attributes most correlated with the occurrence of diabetes. The ML algorithms' performances are further assessed using the reduced set of attributes.

Thirdly, a web-based application is developed for predicting diabetes in individuals based on the insights gained from the performance results..

DATASET :

The utilized dataset for machine learning classification in this study incorporates the Pima Indian dataset, an openly available dataset, and a proprietary dataset. The Pima Indian dataset comprises information from 768 patients, with 268 of them having developed diabetes. The figure illustrates the diabetes prevalence ratio within the Pima Indian dataset, while the table displays the eight features of both the open-source Pima Indian dataset and the proprietary RTML dataset.

The features in both datasets are explained as follows:

Pregnancies: Count of a woman's pregnancies
Glucose: Plasma Glucose concentration after a 2-hour oral glucose tolerance test
BloodPressure: Diastolic Blood Pressure measured in millimeters of mercury (mmHg)
SkinThickness: Thickness of triceps skin fold in millimeters (mm)
Insulin: 2-hour serum insulin level measured in micro International Units per milliliter (mu U/ml)
BMI: Body Mass Index calculated as weight in kilograms divided by the square of height in meters
Age: Individual's age in years
DiabetesPedigreeFunction: Evaluation of diabetes likelihood based on family history
Outcome: Binary indicator (0 for no diabetes, 1 for diabetes presence)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0

Table 1-DIABETES PREDICTION DATA

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Table -2-DIABETES PREDICTION

```
import seaborn as sns
import matplotlib.pyplot as plt

# Visualizing pair plots using Seaborn
sns.pairplot(df, hue='Outcome', diag_kind='hist')
plt.show()
```

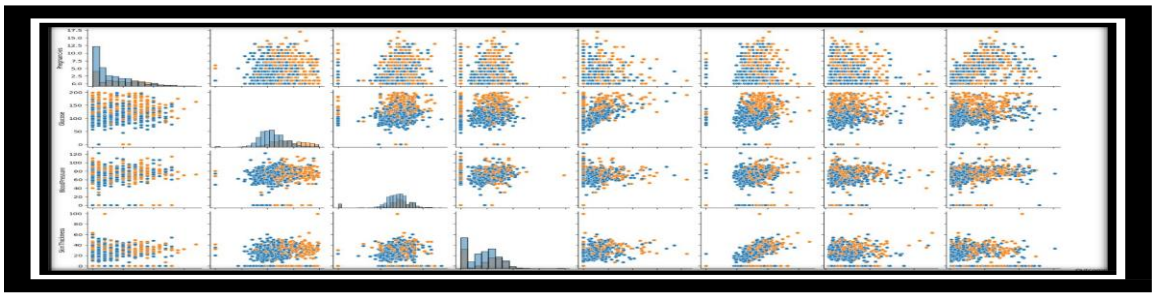


Figure-1-Pair plot for finding the relationship between features for diabetes prediction.

DATASET PREPROCESSING :

In the combined dataset, certain unexpected zero values were identified, such as those in skin thickness and Body Mass Index (BMI), which are logically non-zero. To address this, zero values were replaced with their respective mean values. The division between the training and test datasets was accomplished through the holdout validation technique, allocating 80% to training data and 20% to test data. For instance, based on the mutual information technique, the diabetes pedigree function appears to be of lesser significance.

A thorough evaluation was conducted to identify the most effective regressor technique for predicting the insulin feature in the RTML dataset using information from the Pima Indian dataset. Given the absence of actual insulin values in the RTML dataset, the Pima Indian dataset was initially utilized to identify the best regression model. The Pima Indian dataset was split into an 8:2 ratio, and three supervised regression models—extreme gradient boosting technique (XGB), support vector regression (SVR), and Gaussian process regression (GPR)—were applied to predict the selected outcome, insulin, for the validation samples in the Pima Indian dataset.

DATA CLEANING :

- Dropping duplicate values
- Checking NULL values

```
df.isnull().sum()

Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

- Figure-2-Checking NULL values

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(x='Outcome', data=df)
plt.show()
```

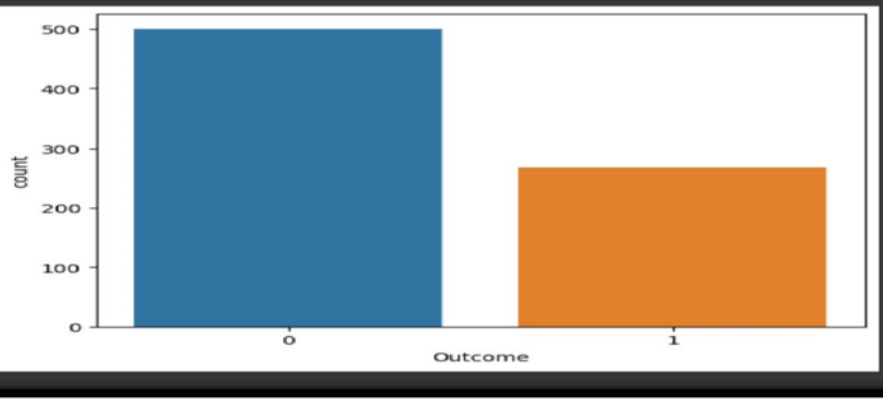


Figure-3-PLoting the data using count plot

Conclusion :- We observe that difference between number of people who do not have diabetes is far more than people who do which indicates that our data is imbalanced

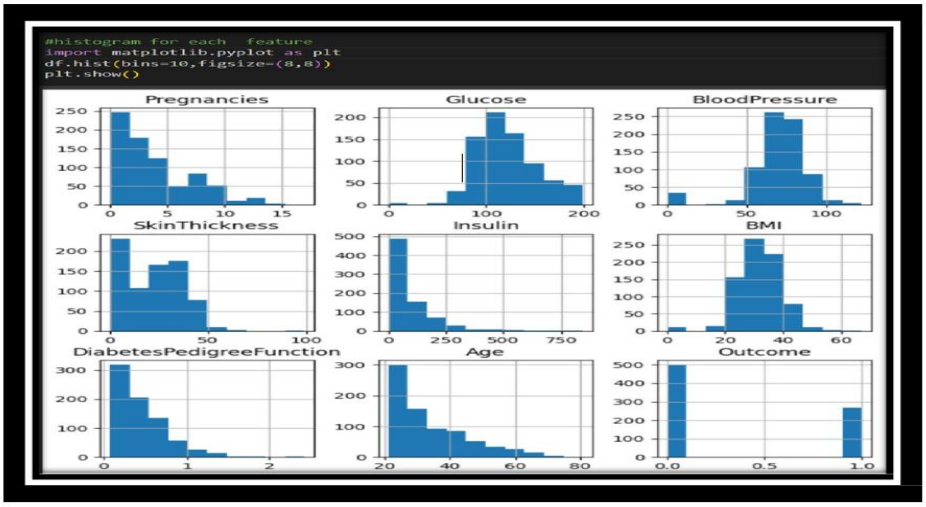


Figure-4-Distribution of data using histogram

Conclusion :- We observe that only glucose and Blood Pressure are normally distributed rest others are skewed and have outliers

FEATURE SELECTION :

Pearson's Correlation Coefficient :

Pearson's Correlation Coefficient is a metric that reveals the association strength between two variables. It ranges from -1 to +1, signifying high correlation at 1 and no correlation at 0. A heat map, employing colors, serves as a two-dimensional visualization tool to simplify the understanding of both straightforward and intricate information.

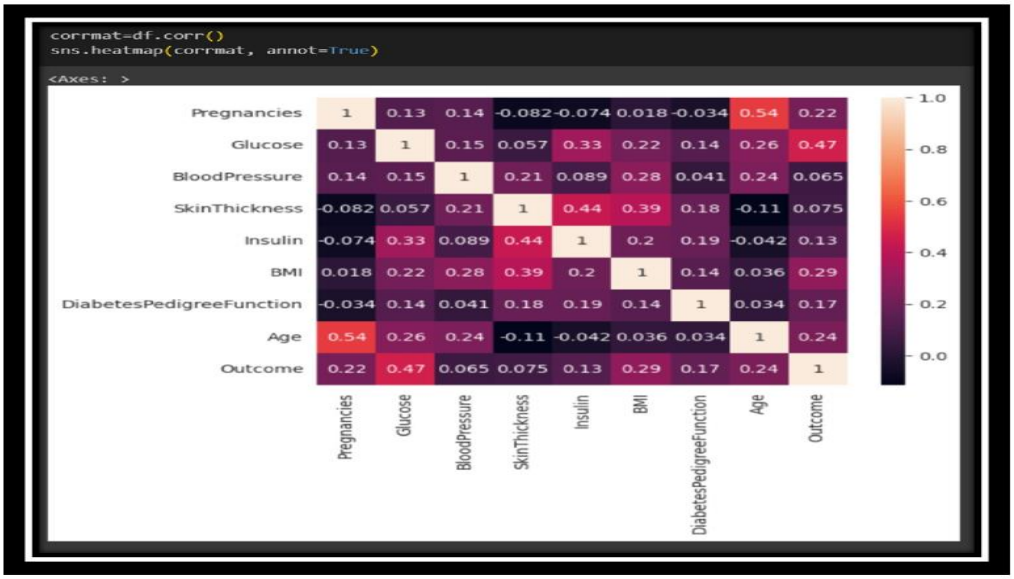


Figure-5-Ploting the Confusion matrix for diabetes data

CONCLUSION :- Examine the 'Outcome' row and identify its correlation scores with

various features. It's evident that Glucose, BMI, and Age exhibit the highest correlation with Outcome. Conversely, Blood Pressure, Insulin, and Diabetes Pedigree Function display lower correlations, suggesting they contribute less to the model and could be considered for removal.

HANDLING OUTLIERS

Addressing outliers is a crucial aspect of data preprocessing to prevent undue impact on statistical analyses or machine learning models. Outliers, which are data points deviating significantly from the majority, can distort the analysis.

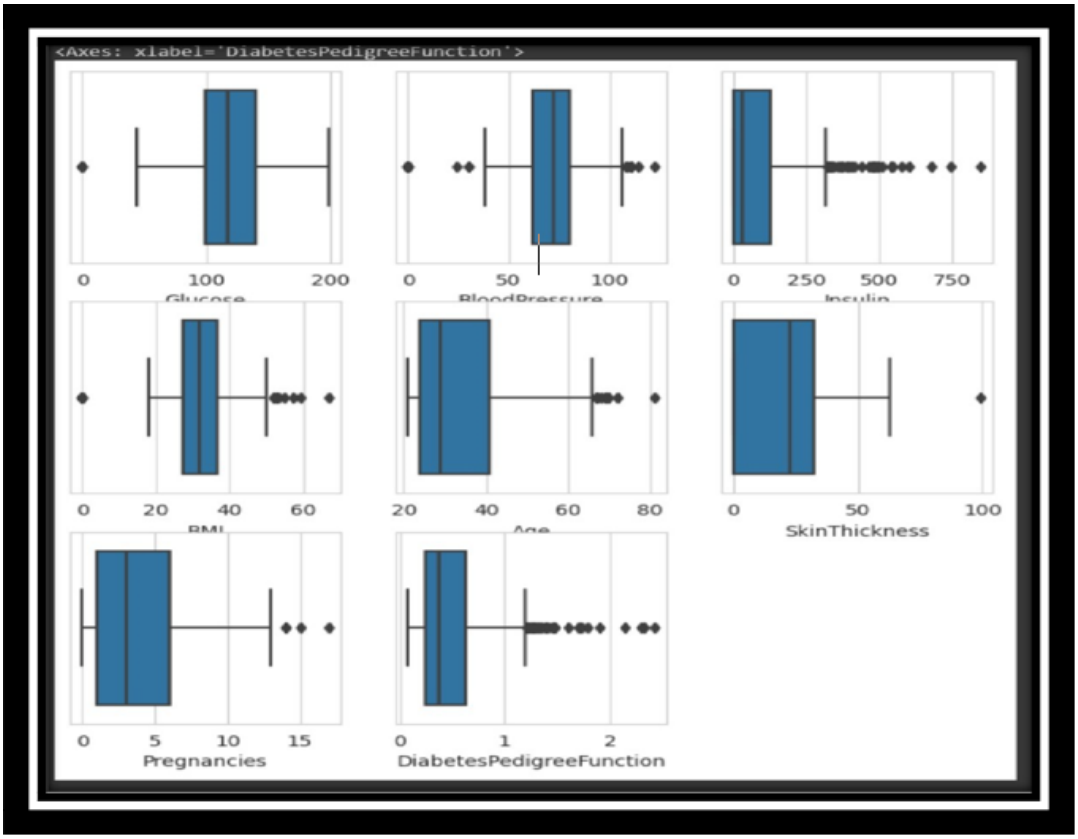


Figure-7-Detecting the outliers

Outliers, defined as atypical values in a dataset, have the potential to distort statistical analyses and violate their assumptions. It is essential to address outliers, but outright removal may lead to data loss. Therefore, employing various scaling and transformation techniques becomes crucial for effective handling.

```
df_selected=df.drop(['BloodPressure','Insulin','DiabetesPedigreeFunction'],axis='columns')

from sklearn.preprocessing import QuantileTransformer
x=df_selected
quantile  = QuantileTransformer()
X = quantile.fit_transform(x)
df_new=quantile.transform(X)
df_new=pd.DataFrame(X)
df_new.columns =['Pregnancies', 'Glucose','SkinThickness','BMI','Age','Outcome']
df_new.head()
```

Figure-8-Normalizing the data using Quantile Transform

Quantile Transformer :- It is a method used to transform features, aiming to make them adhere to a uniform or normal distribution. By doing so, it spreads out the most common values for a given feature and diminishes the influence of marginal outliers. This approach is considered robust for preprocessing, contributing to more reliable and resilient data transformation.

```
target_name='Outcome'
y= df_new[target_name]#given predictions - training data
X=df_new.drop(target_name,axis=1)#dropping the Outcome column and keeping all other columns as X

X.head()
```

	Pregnancies	Glucose	SkinThickness	BMI	Age
0	0.747718	0.810300	0.801825	0.591265	0.889831
1	0.232725	0.097784	0.644720	0.227510	0.558670
2	0.863755	0.956975	0.000000	0.091917	0.585398
3	0.232725	0.131030	0.505867	0.298566	0.000000
4	0.000000	0.721643	0.801825	0.926988	0.606258

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test= train_test_split(X,y,test_size=0.2,random_state=0)#split
```

Figure-11-DIABETES PREDICTION

ALGORITHM

- **Logistic Regression (LogisticRegression):**
  Type: Linear model for classification.
  Usage: Well-suited for binary classification problems**.**

- **Naive Bayes (GaussianNB):**
  Type: Probabilistic model based on Bayes' theorem.
  Usage: Particularly effective for text classification problems.

- **Decision Tree (DecisionTreeClassifier):**
  Type: Tree-based model for classification.
  Usage: Creates a tree structure to make decisions based on features.

- **Random Forest (RandomForestClassifier):**
  Type: Ensemble learning method based on decision trees.
  Usage: Builds multiple decision trees and merges their predictions.

- **Gradient Boosting (GradientBoostingClassifier):**
  Type: Ensemble learning method using boosting.
  Usage: Builds trees sequentially, each correcting errors of the previous one.

- **Ensemble Voting Classifier (VotingClassifier):**
  Type: Combines multiple classifiers' predictions.
  Usage: Allows combining different models for a collective decision.

- **Support Vector Machine (SVC):**
  Type: Linear or non-linear classification.
  Usage: Effective in high-dimensional spaces, especially for binary classification.

- **K-Nearest Neighbors (KNeighborsClassifier):**
  Type: Instance-based learning.
  Usage: Classifies based on the majority class among its k-nearest neighbors

- **XGBoost (XGBClassifier):**
  Type: Gradient boosting library.\
  Usage: Highly efficient and widely used for structured/tabular data.

- **Bagging (BaggingClassifier):**
  Type: Ensemble learning method using bootstrapped samples.
  Usage: Trains multiple instances of a model on different subsets.

EXPERIMENTAL RESULTS ANALYSIS :

ALGORITHM	ACCURACY
Logistic Regression	**81.02%**
Naive Bayes	**75.73%**
Decision Tree	**73.37%**
Random Forest	**81.07%**
KNN Classifier	**75.97%**
Support Vector Machine	**79.22%**
Ensemble (Voting Classifier)	**81.16%**

Table-3-Accuracy of all the models for Diabetes prediction

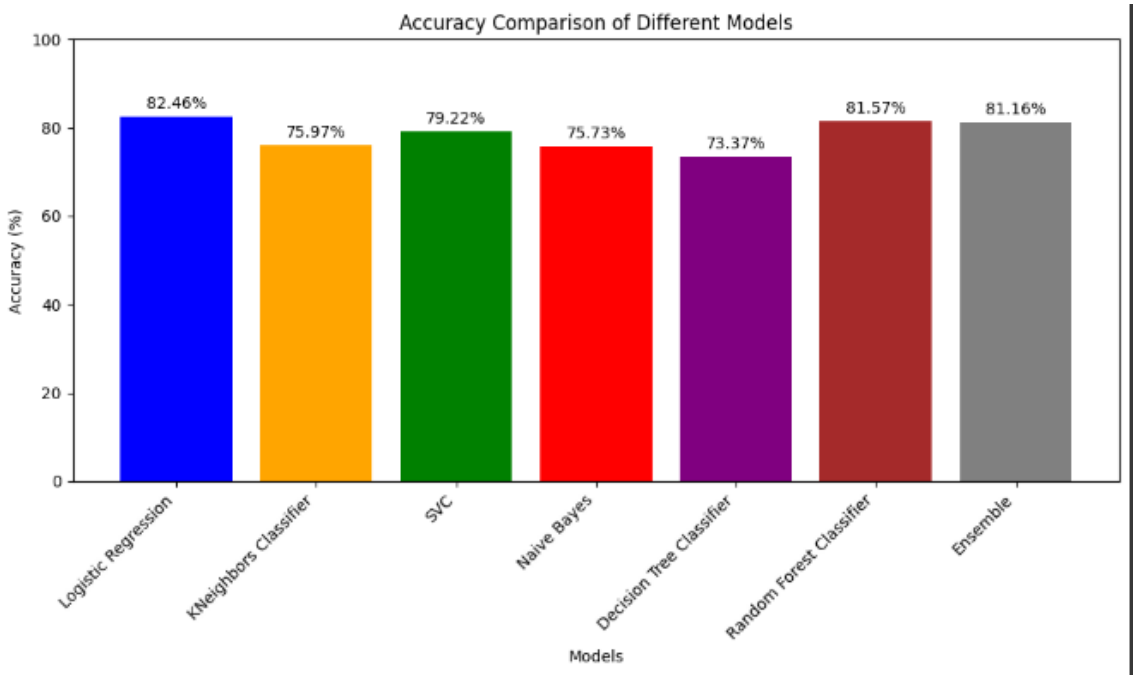


Figure-16-Comparision of Accuracy and Precision for different modes

ENSEMBLE ACCURACY FOR DIABETES PREDICTION	**81.16%**

Table-4-Ensemble Accuracy for Diabetes prediction

PERFORMANCE EVALUATION

The performance metrics used in this research played crucial roles in evaluating the effectiveness and accuracy of the classifiers. Several metrics were employed, including accuracy, recall, precision, F1-score, Cohen's kappa (κ), and AUC-ROC. These metrics provided valuable insights into different aspects of the classifiers' performance. The evaluation was based on the confusion matrix shown in **Table 2**, which depicts the classification results in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP represents the instances that were correctly predicted as the positive class, while TN represents instances that were correctly predicted as the negative class. FP are instances that were incorrectly predicted as the positive class and FN are instances that were incorrectly predicted as the negative class.

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + TP + FN} \times 100\%$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\%$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\%$$

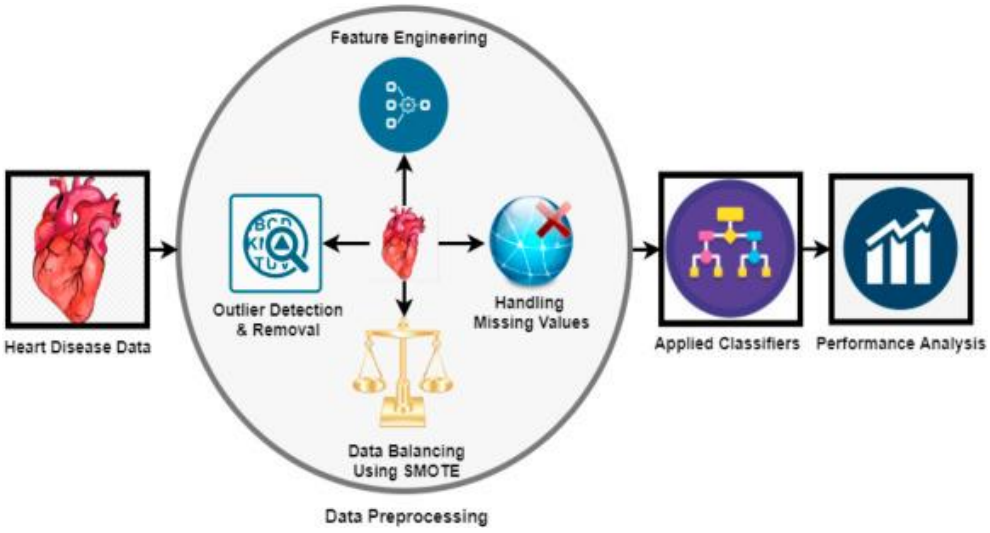
$$\kappa = 2 \times \frac{(TP \cdot TN - FP \cdot FN)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)}$$

$$\text{F1} - \text{score} = \frac{2(Precision \times Recall)}{Precision + Recall}$$

HEART DISEASE PREDICTION USING MACHINE LEARNING

SYSTEM DESIGN

The first requirement of the model is the dataset, once we have collected the dataset preprocessing of the data comes into play, preprocessing includes omission of duplicate entries, handling missing entries etc. After that in this proposed model normalization of the data has been done and then training and testing data is formed out of the dataset, that is to use the training set for model training and use the testing set to test the proposed model that is the trained model. This splitting has been performed randomly in a complete dataset. After that we applied many techniques/classifiers on the training data set and train the model, classifier includes Linear Regression, KNN, Decision Trees, SVM, Gaussian B, and then an average of all these predictions made to get the result of average Ensemble Learning



Another Ensemble method is also applied in this paper that includes Random Forest that is Bagging Ensemble Method and extreme Gradient Boosting that is Boosting Ensemble learning method to get the prediction results. At last evaluation of techniques has been done with the testing dataset. And the result of all these classifiers has been evaluated, these results include - Accuracy, Recall, F-1 Score, Precision, ROC - AUC score along with AUC curves. Analyzing and Comparing all these methods has been done to get the best technique/method on this dataset and preprocessing methods.

DATASET :

The dataset encompasses a diverse array of features related to individuals' health, specifically focusing on factors associated with the diagnosis of heart disease. These features encompass demographic information like age and gender, as well as physiological indicators such as resting blood pressure, serum cholesterol levels, and the maximum heart rate achieved during stress testing. Additionally, categorical variables like chest pain type, fasting blood sugar levels, and electrocardiographic results are included. Factors related to exercise, such as the presence of exercise-induced angina, ST depression, and the slope of the peak exercise ST segment, are also taken into consideration. Furthermore, the dataset incorporates details about the number of major vessels colored by fluoroscopy and the presence of thalassemia. The target variable, denoted by 'target,' indicates whether an individual has heart disease (1) or not (0). This comprehensive set of features forms a valuable foundation for constructing predictive models aimed at identifying individuals at risk of heart disease based on their health characteristics.

The dataset includes the following features:

Age (in years)

Sex: (1 = male; 0 = female)

Chest pain type (cp)

Resting blood pressure (trestbps, in mm Hg on admission to the hospital)

Serum cholesterol (chol) in mg/dl

Fasting blood sugar (fbs) > 120 mg/dl (1 = true; 0 = false)

Resting electrocardiographic results (restecg)

Maximum heart rate achieved (thalach)

Exercise-induced angina (exang) (1 = yes; 0 = no)

ST depression induced by exercise relative to rest (oldpeak)

Slope of the peak exercise ST segment (slope)

Number of major vessels (ca) colored by fluoroscopy (0-3)

Thalassemia (thal): 3 = normal; 6 = fixed defect; 7 = reversible defect

Target: 1 or 0 (presence or absence of heart disease)

```
df = pd.read_csv('/content/heart.csv');
df.head()
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Table-5-HEART DISEASE PREDICTION-DATA

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Fig -21-HEART DISEASE PREDICTION

```
df.describe()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Table-6-Statistical Summary of heart disease data

DATASET PREPROCESSING :

Handling Missing Data: Check and address any missing values in the dataset. Depending on the extent of missingness, consider imputation or removal of affected instances or features.

Categorical Encoding: Convert categorical variables (e.g., 'sex,' 'cp,' 'fbs,' 'restecg,' 'exang,' 'slope,' 'thal') into numerical representations using techniques like one-hot encoding to ensure compatibility with machine learning algorithms.

Feature Scaling: Normalize numerical features (e.g., 'age,' 'trestbps,' 'chol,' 'thalach,' 'oldpeak') to a consistent scale, such as standardization, to prevent certain features from dominating others.

Outlier Detection and Handling: Identify and manage outliers that could adversely affect model performance. This may involve transforming or removing extreme values.

Correlation Analysis: Explore correlations between features to identify potential multi-collinearity, and consider excluding highly correlated variables to enhance model interpretability and stability.

Balancing the Dataset: Evaluate and address any class imbalance in the 'target' variable to ensure that the model does not skew towards the majority class.

Train-Test Split: Divide the dataset into training and testing sets to assess model performance on unseen data. Feature Selection: Depending on the dataset's size and complexity, consider feature selection techniques to choose the most relevant features for prediction, improving model efficiency.

Data Exploration: Visualize and explore the distribution of features, relationships between variables, and the prevalence of heart disease cases to gain insights into the dataset's characteristics.

Handling Categorical Variables (if needed): If the model chosen requires numerical input, transform categorical variables (e.g., 'thal') into a suitable format using techniques like label encoding or creating ordinal mappings

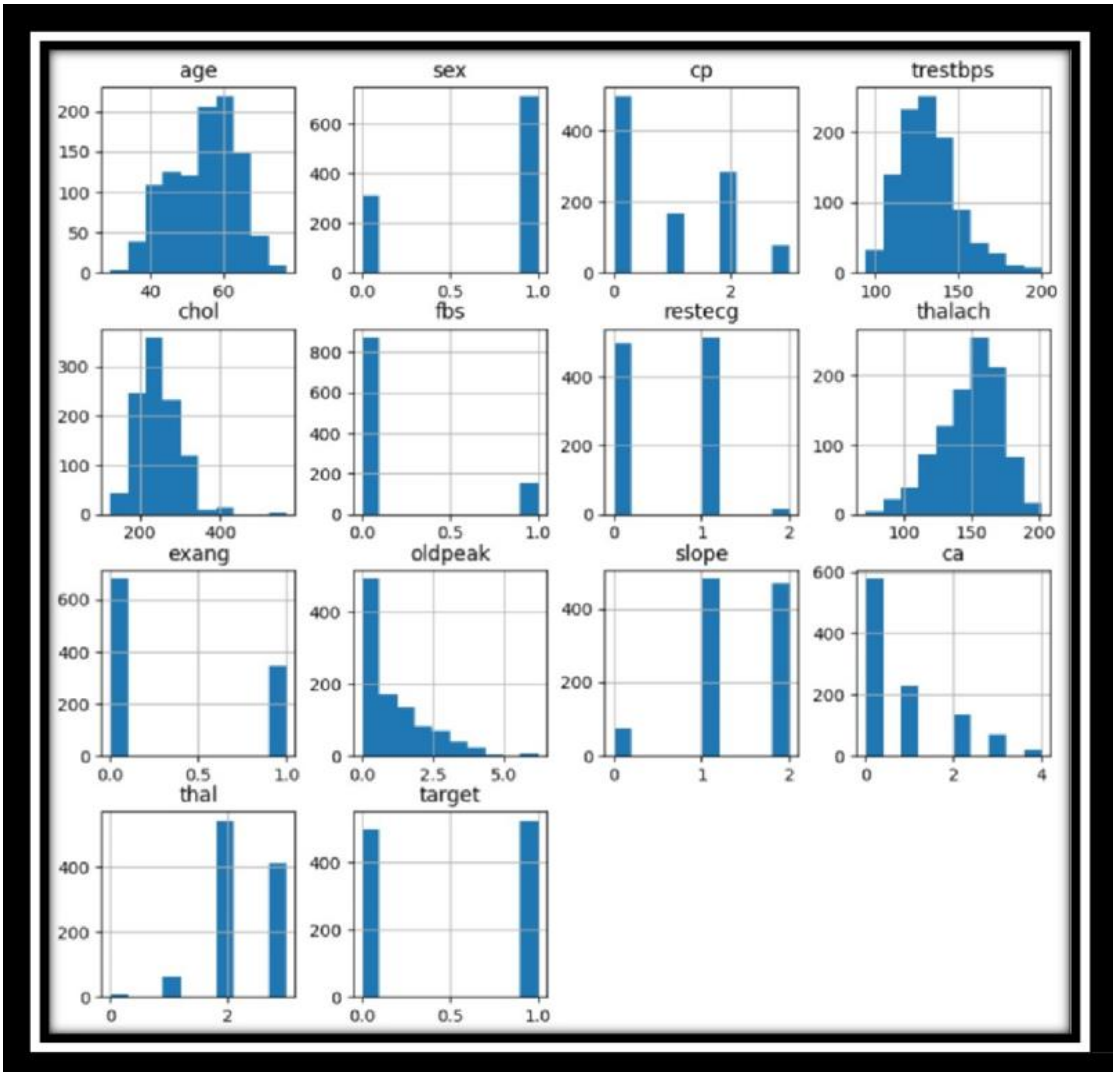


Fig -26-Distribution of the data using histogran

FEATURE SELECTION

Pearson's Correlation Coefficient :

Pearson's Correlation Coefficient is a statistical measure that evaluates the strength and direction of a linear relationship between two continuous variables. Ranging from -1 to +1, a value of -1 indicates a perfect negative linear relationship, 1 denotes a perfect positive linear relationship, and 0 signifies no linear correlation. This metric is essential in gauging the level of association between various features during the process of feature selection in a dataset.

Utilizing Pearson's Correlation Coefficient and heatmaps for feature selection often involves setting a correlation threshold. Features surpassing this threshold are considered highly correlated and may be potential candidates for removal. This approach enhances model interpretability and performance by eliminating redundant information and simplifying the dataset.

Correlation Analysis

```
correlation_matrix = df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='rocket', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

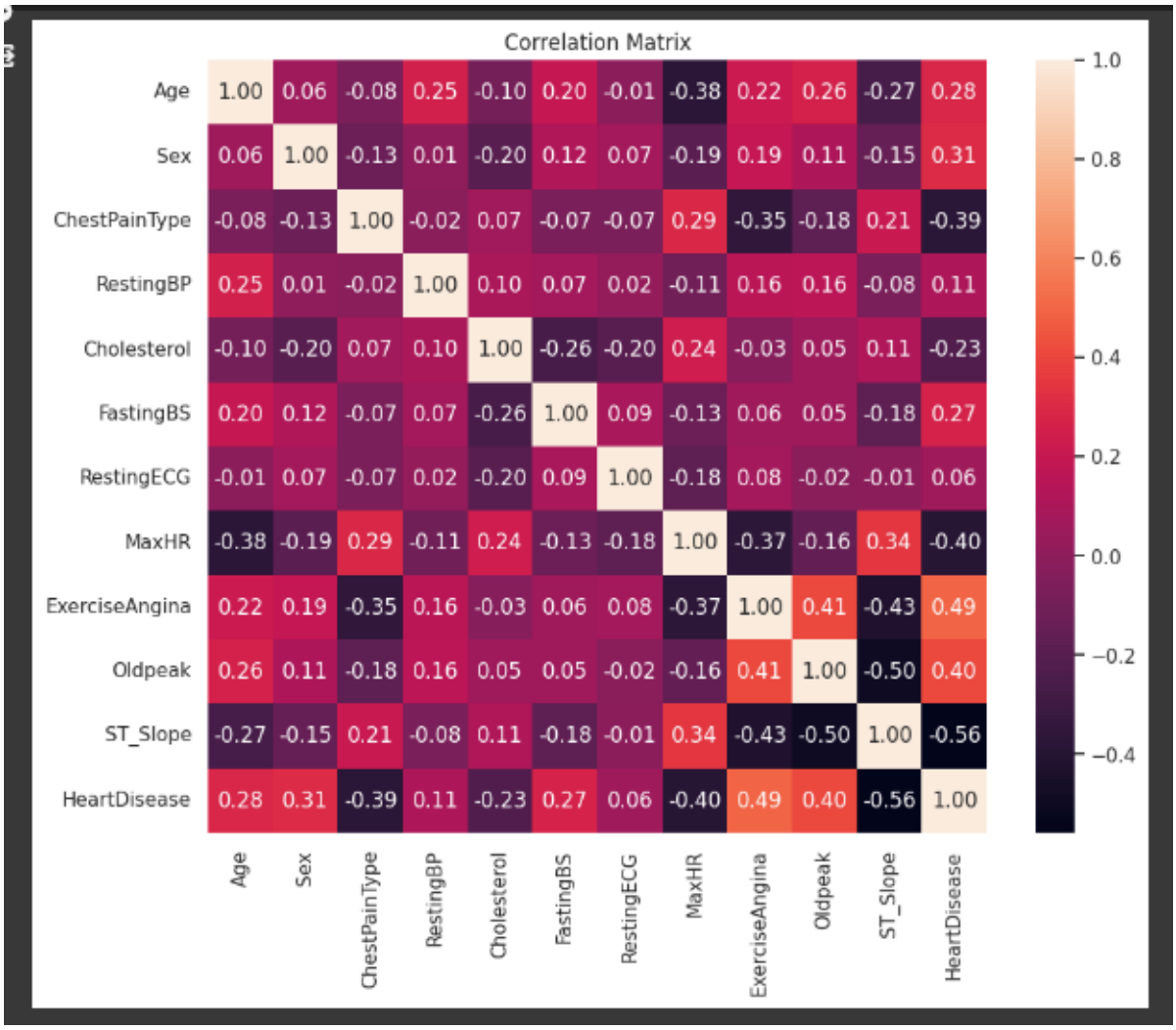


Fig -27-Correlation of Heart disease data

Conclusion : - By examining the correlation scores in the last row, particularly with the 'Outcome' variable, noteworthy observations can be made. Glucose, BMI, and Age exhibit the highest correlation with Outcome, suggesting a significant influence on the predictive model. Conversely, Blood Pressure, Insulin, and Diabetes Pedigree Function display lower correlation scores, indicating a relatively weaker association with the outcome variable. Consequently, considering their limited contribution to the model, these features can be omitted**.**

HANDLING OUTLIERS :

Managing outliers is a crucial step in data preprocessing to prevent them from disproportionately impacting statistical analyses or machine learning models. Outliers, which are data points deviating significantly from the rest, can distort the analysis. Here are several standard techniques for handling outliers.

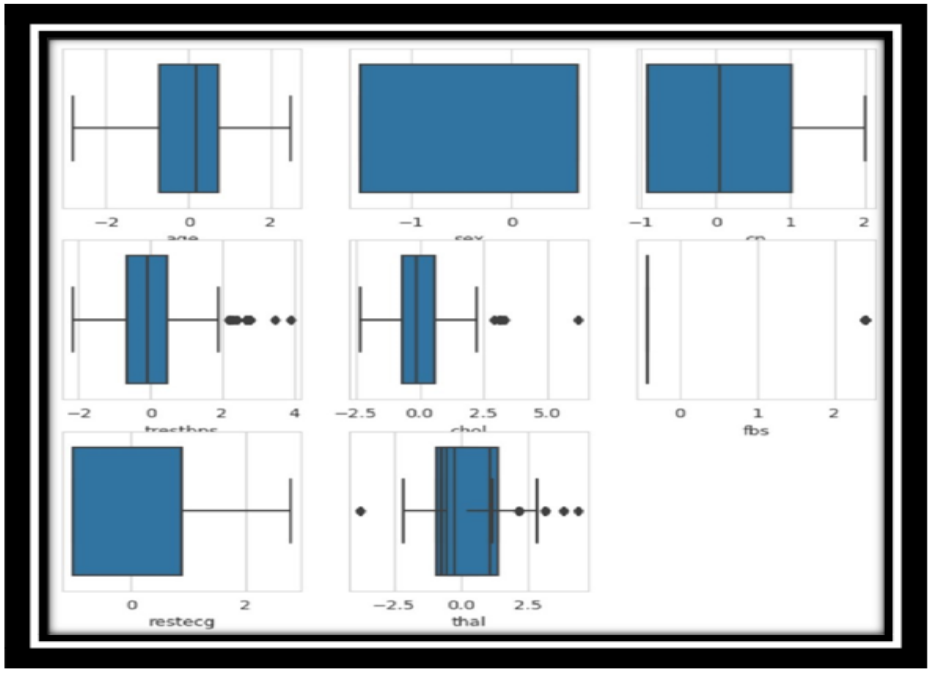


Figure -28-Decting the Outliers

```
import plotly.express as px

# Assuming df is your DataFrame
fig = px.scatter_3d(df, x='age', y='thalach', z='chol', color='target')

# Show the interactive 3D scatter plot
fig.show()
```

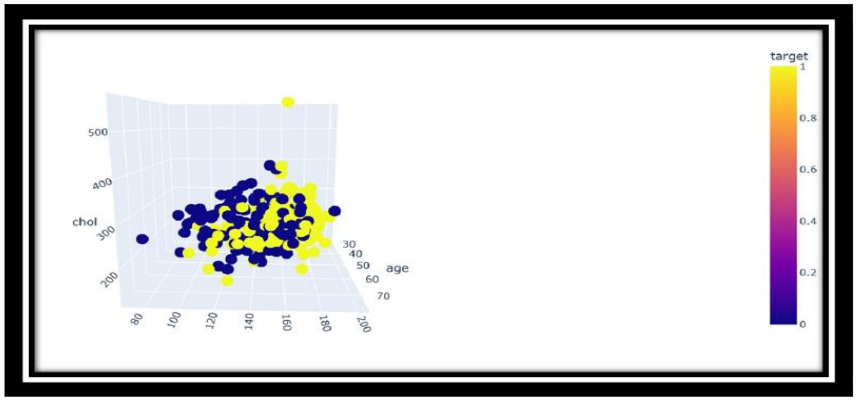


Figure -29-Interactive 3D scatter plot

EXPERIMENTAL RESULTS ANALYSIS :

ALGORITHM	ACCURACY
Logistic Regression	84.24%
Naive Bayes	84.24%
Decision Tree	78.80%
Random Forest	89.59%
Support Vector Machine	68.48%
K-Nearest Neighbors	70.11%
XGB Classifier	86.86%
Voting Classifier	86.41%

Table-8-Accuracy of Hybrid Models

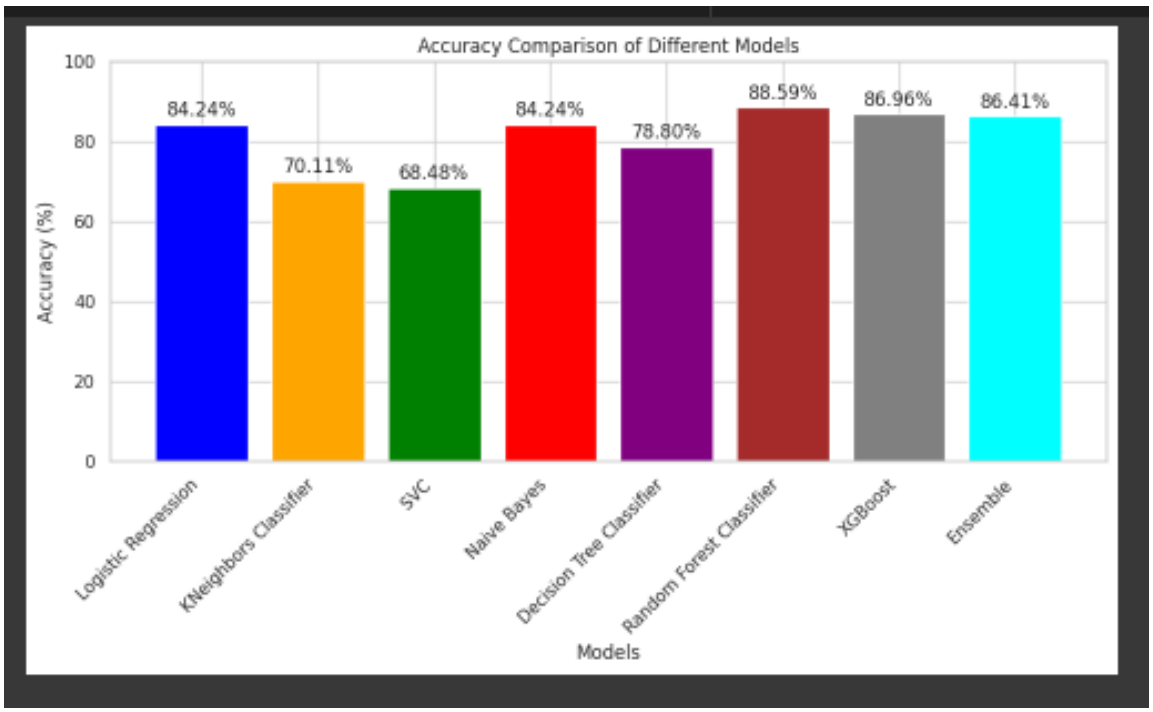
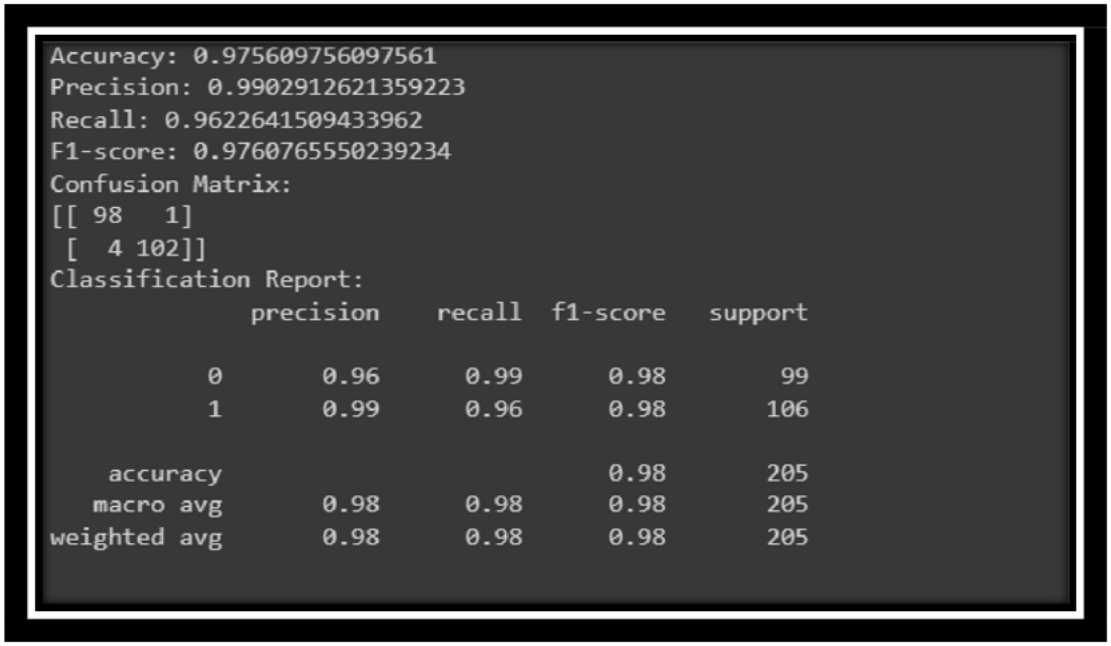


Figure -32-Comparison of Accuracy and Precision for Different Models

PERFORMANCE EVALUATION



```
Accuracy: 0.975609756097561
Precision: 0.9902912621359223
Recall: 0.9622641509433962
F1-score: 0.9760765550239234
Confusion Matrix:
[[ 98   1]
 [  4 102]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98        99
           1       0.99      0.96      0.98       106

    accuracy                           0.98       205
   macro avg       0.98      0.98      0.98       205
weighted avg       0.98      0.98      0.98       205
```

Figure -33-Performance Evaluation of Heart Disease Prediction

ENSEMBLE ACCURACY FOR HEART DISEASE PREDICTION	**86.41%**

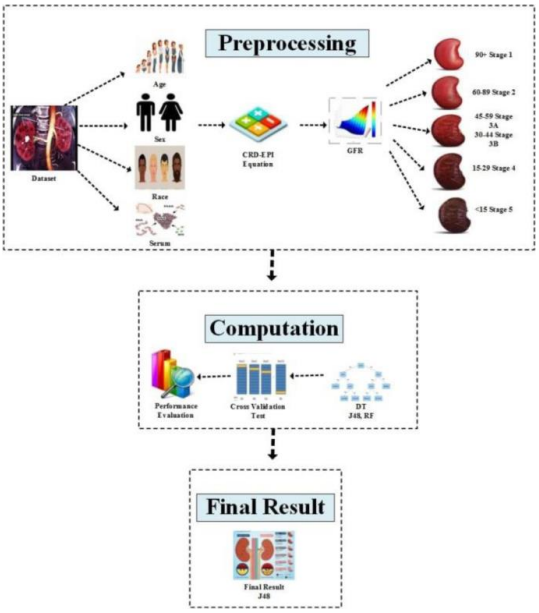
Table-9-Ensemble Accuracy of Heart Disease Prediction

CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

SYSTEM DESIGN

In the initial stage of developing the model, the primary requirement is a dataset. Once the dataset is gathered, the preprocessing phase becomes crucial. This involves various tasks such as removing duplicate entries and handling missing data. In the proposed model, data normalization is applied after preprocessing. Following this, the dataset is divided into training and testing sets. The training set is used to train the model, and the testing set is utilized to evaluate the model's performance.

The division of the dataset into training and testing sets is done randomly. Subsequently, multiple techniques or classifiers are employed on the training dataset. These classifiers encompass, KNN, Support Vector Machine, Decision Trees, NB , Linear Regression. The predictions from each classifier are averaged to obtain the final result through Ensemble Learning. This approach enhances the model's robustness and improves its predictive capabilities.



Another Ensemble method is also applied in this paper that includes Random Forest that is Bagging Ensemble Method and extreme Gradient Boosting that is Boosting Ensemble learning method to get the prediction results. At last evaluation of techniques has been done with the testing dataset. And the result of all these classifiers has been evaluated, these results include - Accuracy, Recall, F-1 Score, Precision, ROC - AUC score along with AUC curves. Analyzing and Comparing all these methods has been done to get the best technique/method on this dataset and preprocessing methods.

DATASET :

The dataset for chronic kidney disease (CKD) encompasses a wide array of medical features aimed at diagnosing and understanding the complexities of renal health. Chronic kidney disease is a prolonged condition leading to renal failure, affecting the kidneys' ability to filter waste and excess fluids from the blood. Symptoms may develop slowly, and diagnosis often involves lab tests. As the disease progresses, symptoms management is crucial, with potential interventions such as dialysis or transplantation in advanced stages. The ultimate target variable, represented by 'target,' signifies whether an individual has chronic kidney disease (1) or not (0). This comprehensive set of features provides a valuable basis for developing predictive models aimed at identifying individuals at risk of chronic kidney disease based on their health characteristics.

The Dataset incorporates a total of 25 features, consisting of 24 medical features and 1 class label.

Among these, 11 features are numerical, and 14 are nominal.

Age (numerical) - Age in years
Blood Pressure (numerical) - Blood pressure in mm/Hg
Specific Gravity (nominal) - sg values: 1.005, 1.010, 1.015, 1.020, 1.025
Albumin (nominal) - al values: 0, 1, 2, 3, 4, 5
Sugar (nominal) - su values: 0, 1, 2, 3, 4, 5
Red Blood Cells (nominal) - rbc values: normal, abnormal
Hemoglobin (numerical) - Hemo level in gms
White Blood Cell Count (numerical) - wc values in cells/cumm
Red Blood Cell Count (numerical) - rc values in millions/cmm
Hypertension (nominal) - htn values: yes, no
Appetite (nominal) - appet values: good, poor
Pedal Edema (nominal) - pe values: yes, no
Anemia (nominal) - ane values: yes, no
Class (nominal) - class values: ckd (chronic kidney disease), notckd (not chronic kidney disease)

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 26 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      400 non-null    int64
 1   age                     391 non-null    float64
 2   blood pressure          388 non-null    float64
 3   specific gravity        353 non-null    float64
 4   albumin                 354 non-null    float64
 5   sugar                   351 non-null    float64
 6   red blood cells         248 non-null    object
 7   pus cell                335 non-null    object
 8   pus cell clumps         396 non-null    object
 9   bacteria                396 non-null    object
 10  blood glucose random    356 non-null    float64
 11  blood urea              381 non-null    float64
 12  serum creatinine        383 non-null    float64
 13  sodium                  313 non-null    float64
 14  potassium               312 non-null    float64
 15  haemoglobin             348 non-null    float64
 16  packed cell volume      330 non-null    object
 17  white blood cell count  295 non-null    object
 18  red blood cell count    270 non-null    object
 19  hypertension            398 non-null    object
 20  diabetes mellitus       398 non-null    object
 21  coronary artery disease 398 non-null    object
 22  appetite                399 non-null    object
 23  pedal edema             399 non-null    object
 24  anemia                  399 non-null    object
 25  class                   400 non-null    object
dtypes: float64(11), int64(1), object(14)
memory usage: 81.4+ KB
```

Fig -21-HEART DISEASE PREDICTION

DATASET PREPROCESSING

Handling Missing Data: Check and address any missing values in the dataset. Depending on the extent of missingness, consider imputation or removal of affected instances or features.

Categorical Encoding: Convert categorical variables (e.g., 'sex,' 'cp,' 'fbs,' 'restecg,' 'exang,' 'slope,' 'thal') into numerical representations using techniques like one-hot encoding to ensure compatibility with machine learning algorithms.

Feature Scaling: Normalize numerical features (e.g., 'age,' 'trestbps,' 'chol,' 'thalach,' 'oldpeak') to a consistent scale, such as standardization, to prevent certain features from dominating others.

Outlier Detection and Handling: Identify and manage outliers that could adversely affect model performance. This may involve transforming or removing extreme values.

Correlation Analysis: Explore correlations between features to identify potential multi collinearity, and consider excluding highly correlated variables to enhance model interpretability and stability.

Balancing the Dataset : Evaluate and address any class imbalance in the 'target' variable to ensure that the model does not skew towards the majority class.

Train-Test Split: The dataset is partitioned into training and testing sets, allowing for the evaluation of the model's performance on data it has not been exposed to. This step is crucial in assessing how well the model generalizes to new, unseen data.

Feature Selection: To enhance model efficiency, especially in cases of large or complex datasets, contemplate employing feature selection techniques. These techniques help identify and choose the most pertinent features for prediction, streamlining the model and potentially improving its overall performance.

Data Exploration: Visualize and explore the distribution of features, relationships between variables, and the prevalence of heart disease cases to gain insights into the dataset's characteristics.

Handling Categorical Variables (if needed): If the model chosen requires numerical input, transform categorical variables (e.g., 'thal') into a suitable format using techniques like label encoding or creating ordinal mappings

```
df.isnull().sum()

id                          0
age                         9
blood pressure             12
specific gravity           47
albumin                    46
sugar                      49
red blood cells           152
pus cell                   65
pus cell clumps             4
bacteria                    4
blood glucose random       44
blood urea                 19
serum creatinine           17
sodium                     87
potassium                  88
haemoglobin                52
packed cell volume         70
white blood cell count    105
red blood cell count      130
hypertension                2
diabetes mellitus           2
coronary artery disease     2
appetite                    1
pedal edema                 1
anemia                      1
class                       0
dtype: int64
```

Fig-24-Finding the missing values

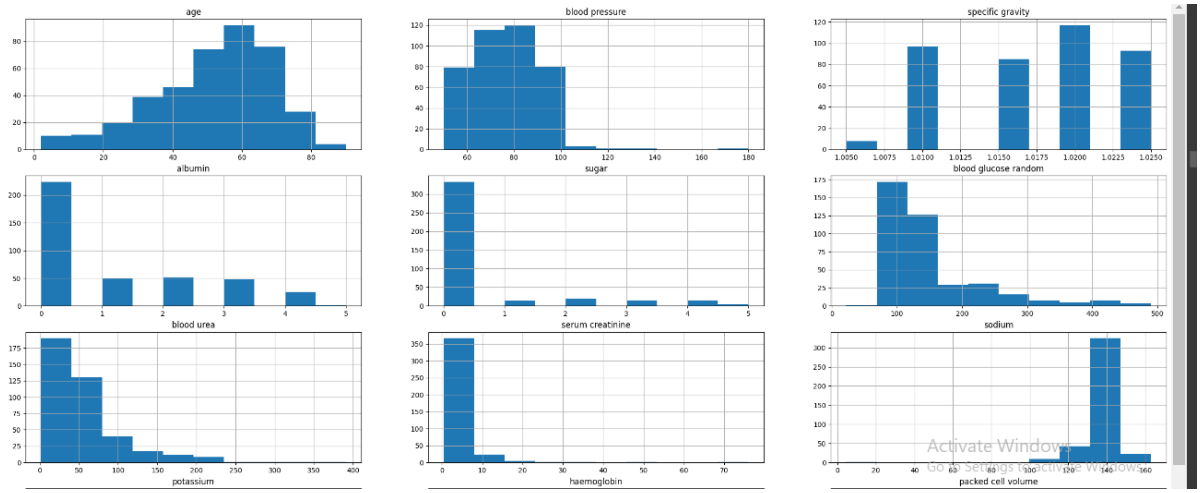


Fig -26-Distribution of the data using histogram

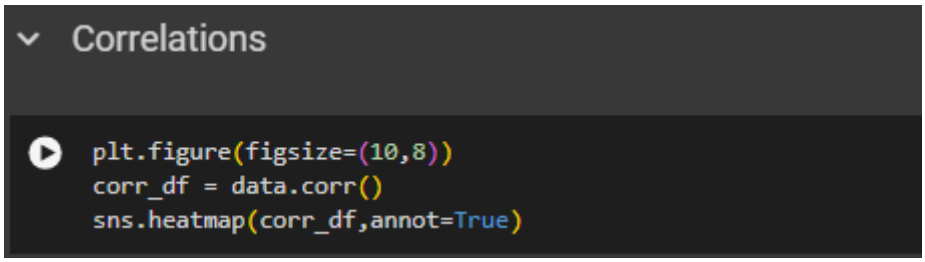
Observations:

The distribution of age appears slightly left-skewed.
Blood glucose demonstrates a right-skewed pattern in a seemingly random manner.
Blood urea exhibits a somewhat right-skewed distribution.
The remaining features display relatively light skewness.

FEATURE SELECTION

Pearson's Correlation Coefficient :

Pearson's Correlation Coefficient is a statistical metric used to assess the relationship between two quantities. It provides a measure of the strength of association between two variables. The coefficient value ranges from -1 to +1, where 1 indicates a high positive correlation, 0 signifies no correlation, and -1 represents a high negative correlation. To facilitate visualization of this association, heat maps, utilizing colors, offer a two-dimensional representation of information. Heat maps prove useful in helping users visualize both simple and complex data patterns.



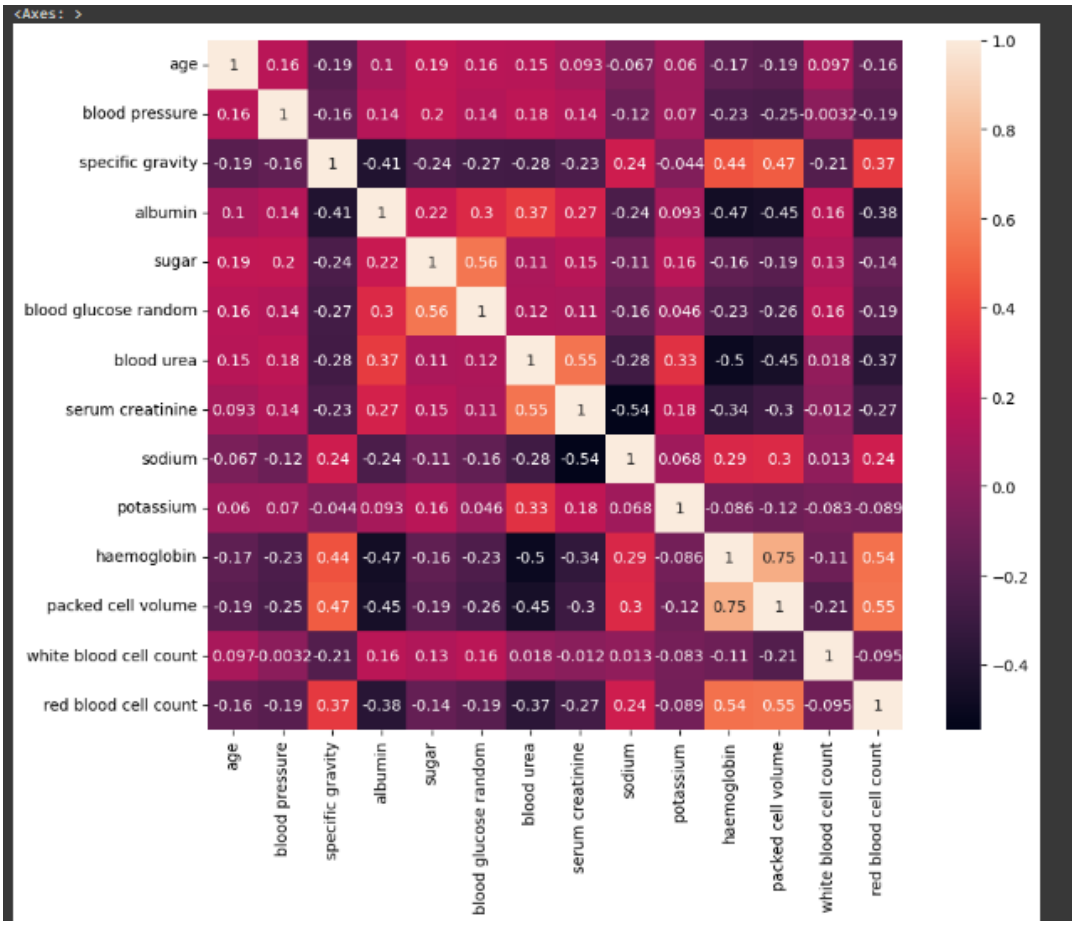


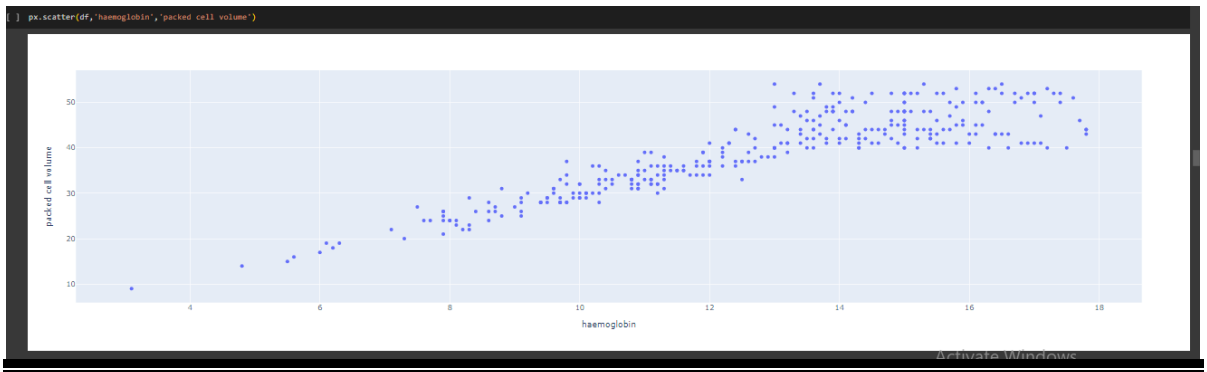
Fig -27-Correlation of CKD data

Positive Correlation:
Specific gravity shows a positive correlation with red blood cell count, packed cell volume, and hemoglobin.
Sugar is positively correlated with blood glucose random.
Blood urea exhibits a positive correlation with serum creatinine.
Hemoglobin is positively correlated with red blood cell count and packed cell volume.

Negative Correlation:
Albumin and blood urea show a negative correlation with red blood cell count, packed cell volume, and hemoglobin.
Serum creatinine demonstrates a negative correlation with sodium.

HANDLING OUTLIERS :

Handling outliers is a crucial aspect of data preprocessing, preventing them from disproportionately impacting statistical analyses or machine learning models. Outliers, data points significantly deviating from the majority, can distort the overall analysis. Several common techniques for managing outliers include:



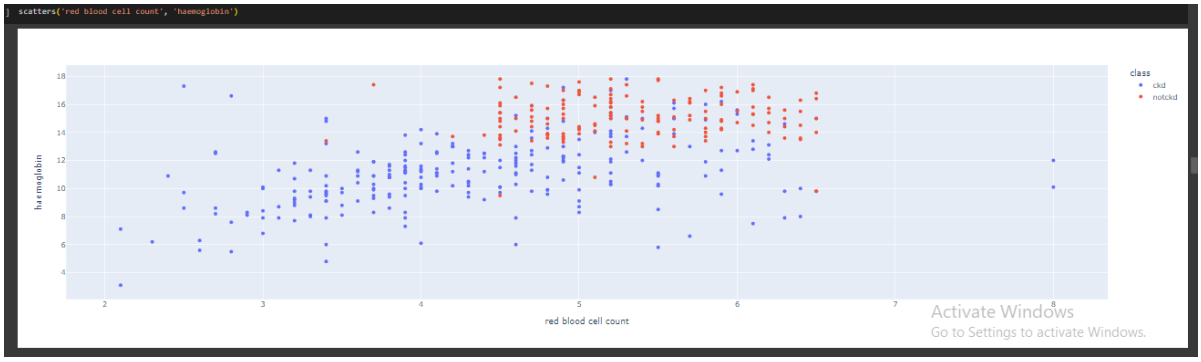


Figure -28-Decting the Outliers

```
Confusion matrix of  CatBoostClassifier
[[75  3]
 [ 2 40]]
Accuracy score is  95.83333333333334
==================================================
Confusion matrix of  Support Vector Machine
[[78  0]
 [42  0]]
Accuracy score is  65.0
==================================================
Confusion matrix of  Decision Tree
[[71  7]
 [ 2 40]]
Accuracy score is  92.5
==================================================
Confusion matrix of  Neural Network
[[78  0]
 [42  0]]
Accuracy score is  65.0
==================================================
Confusion matrix of  Random Forest
[[76  2]
 [ 3 39]]
Accuracy score is  95.83333333333334
==================================================
Confusion matrix of  XGBoost
[[76  2]
 [ 2 40]]
Accuracy score is  96.66666666666667
==================================================
```

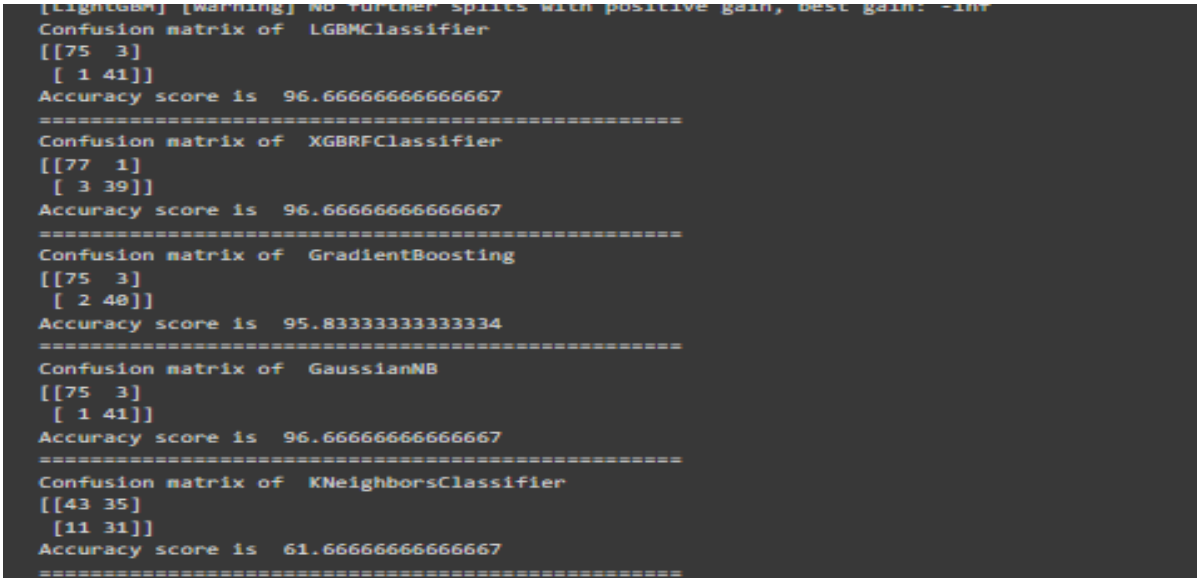


Table-7-Evaluation Metrics for all the Hybrid Models

EXPERIMENTAL RESULTS ANALYSIS :

ALGORITHM	ACCURACY
Decision Tree	**91.67%**
Neural Network	**83.33%**
Support Vector Machine	**65.00%**
KNN Classifier	**61.67%**
XGBoost	**96.67%**
GaussianNB	**95.67%**
XGBRF Classifier	**96.67%**
Random Forest Classification	**95.83%**

Table-8-Accuracy of Hybrid Models

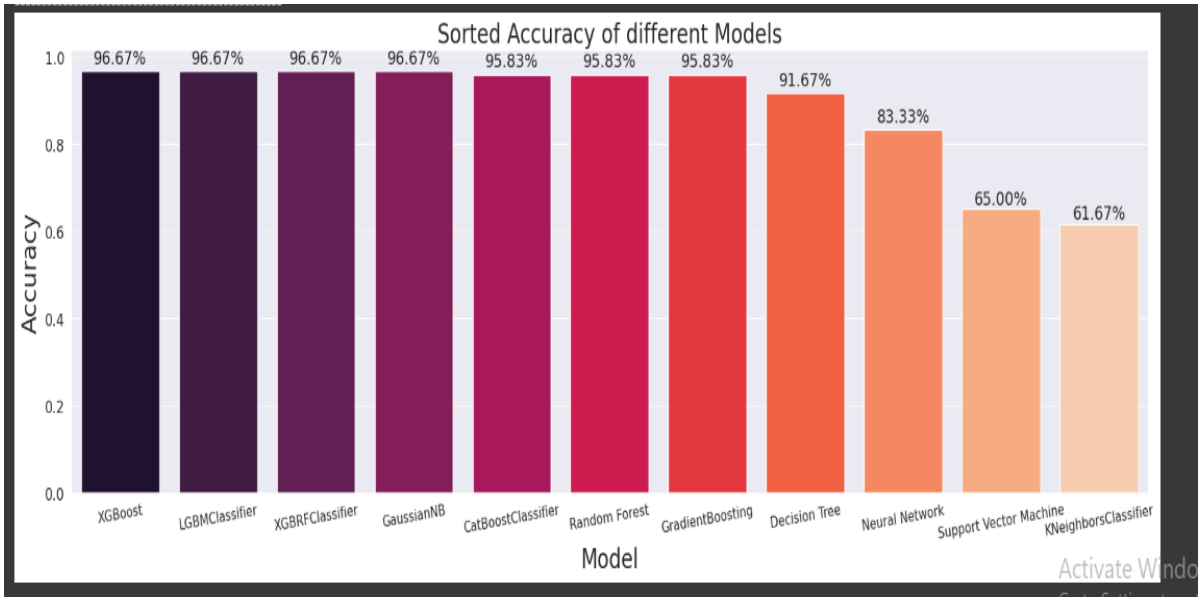


Figure -32-Comparison of Accuracy and Precision for Different Model.

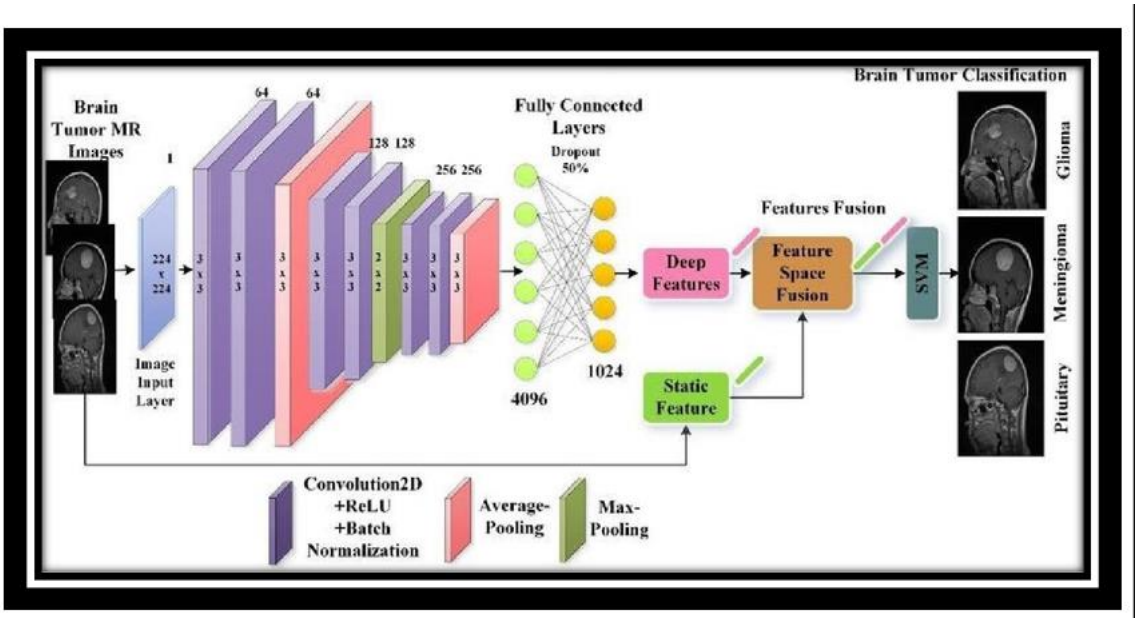
PERFORMANCE EVALUATION

ACCURACY FOR CKD DISEASE PREDICTION	**96.67%**

Table-9-Ensemble Accuracy of Heart Disease Prediction

PREDICTION USING DEEP LEARNING

CNN ARCHITECTURE



Convolutional Neural Networks (CNNs) represent a category of deep neural networks specifically crafted for tasks like image recognition and classification. They excel at capturing spatial hierarchies of features embedded in images.

Inspired by the visual cortex of the human brain, the architecture of CNNs encompasses various layers, including convolutional layers, pooling layers, and fully connected layers. Each layer undertakes specific operations to extract features from the input data, progressively enhancing its ability to comprehend more intricate patterns.

In essence, a Convolutional Neural Network (CNN) is a form of artificial neural network tailored for image-related tasks. Its architecture allows it to autonomously and flexibly learn spatial hierarchies of features from provided input data, particularly well-suited for grid-like structures like images.

The fundamental working principle of a Convolutional Neural Network (ConvNet/CNN) involves taking an input image, assigning significance (via learnable weights and biases) to different aspects or objects within the image, and ultimately distinguishing between them.

The fundamental components of a typical Convolutional Neural Network (CNN) architecture encompass:

Input Layer:
Accepts the raw input data, commonly in the form of images.

Convolutional Layers:
Core building blocks that employ convolutional operations with filters (kernels) to detect patterns, edges, and features.
Stacking multiple convolutional layers enables the learning of hierarchical features.

Padding:
Technique addressing the reduction in spatial dimensions during convolutional operations.
Involves adding extra pixels (padding) around the input data to maintain compatibility between input and output volumes.

Filter:
Essential elements for feature extraction and representation in CNNs.
Learnable matrices, also known as convolutional kernels, capturing patterns and local structures as they traverse input data.

Activation Function:
Typically ReLU (Rectified Linear Unit) introduces non-linearity after convolutional operations.
Facilitates learning complex patterns and relationships.

Pooling Layers:
Reduce spatial dimensions of the input volume using methods like max pooling or average pooling.
Aids in decreasing computation, controlling overfitting, and extracting dominant features.

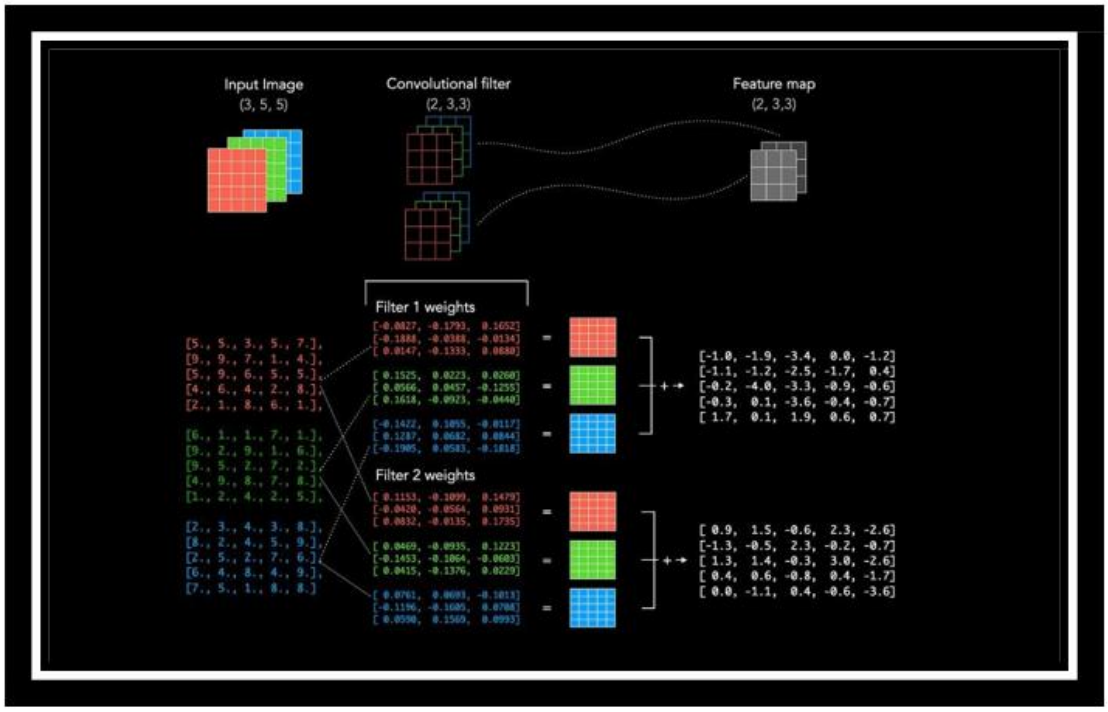
Fully Connected (Dense) Layers:
Capture high-level reasoning in the neural network after convolutional and pooling layers.
Used for making predictions based on extracted high-level features.

Normalization Layers:
Batch normalization may be included to normalize layer inputs, enhancing training speed and generalization.

Dropout:
Prevents overfitting by randomly setting a fraction of input units to zero during training.

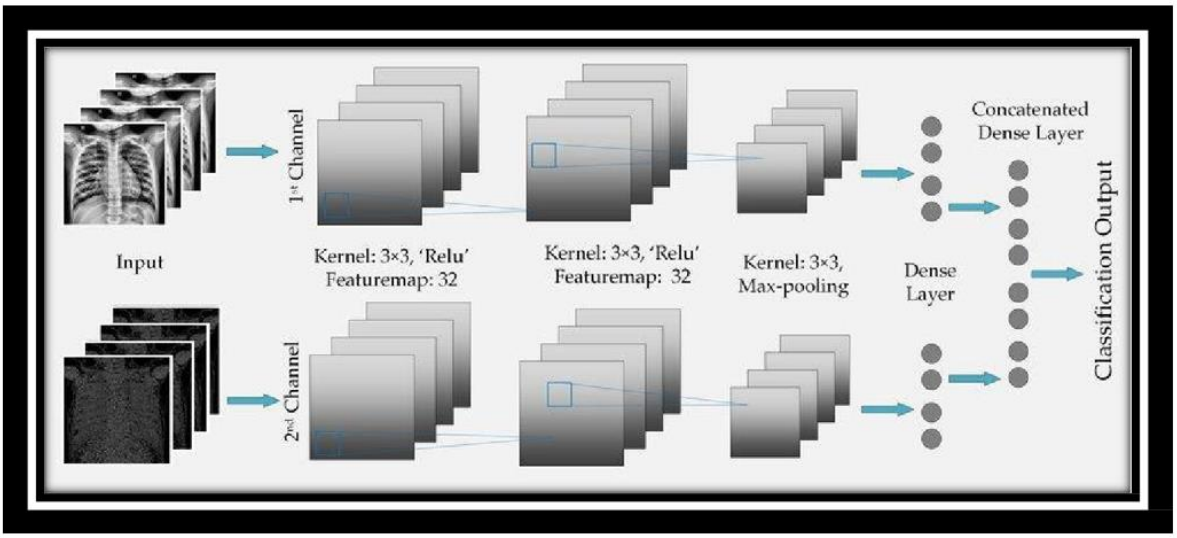
Strides:
Determine the step size for applying convolutional filters to the input.
Influence the spatial dimensions of the output volume, impacting down-sampling and computational complexity.

These components collectively form the architecture of a CNN, enabling effective feature extraction and hierarchical learning in tasks such as image recognition and classification.



PNEUMONIA PREDICTION USING DEEP LEARNING

CNN DESIGN



This architecture is a common and simple design for binary image classification tasks like pneumonia prediction. Here's a breakdown of the layers:

Convolutional Layers (Conv2D): Extracts features from input images using convolutional filters. The number of filters increases in deeper layers to capture more complex patterns. ReLU activation is applied to introduce non-linearity. Max pooling is used in this to down-sample the spatial dimensions.

Flatten Layer: Converts the multi-dimensional output from the convolutional layers ito a structure of one-dimensional array .Prepares the data for the fully connected layers.

Dense (Fully Connected) Layers: Neurons in these layers process information from the flattened representation. ReLU activation introduces non-linearity. Dropout is used for regularization to prevent over-fitting.

Output Layer : Single neuron with a sigmoid activation function for binary classification (pneumonia/ normal). Produces a probability indicating the likelihood of pneumonia

DATA SET

The dataset utilized in this study comprises 5,863 chest X-ray images obtained from Kaggle. This dataset is subsequently segregated into test, train, and validation sets. The images within the dataset are further categorized into two groups: normal and pneumonia-infected images.

Within the dataset, each chest x-ray image is categorized into one of two classes: normal or pneumonia-infected. This binary classification enables the algorithm to discern between healthy and diseased conditions, facilitating precise predictions.

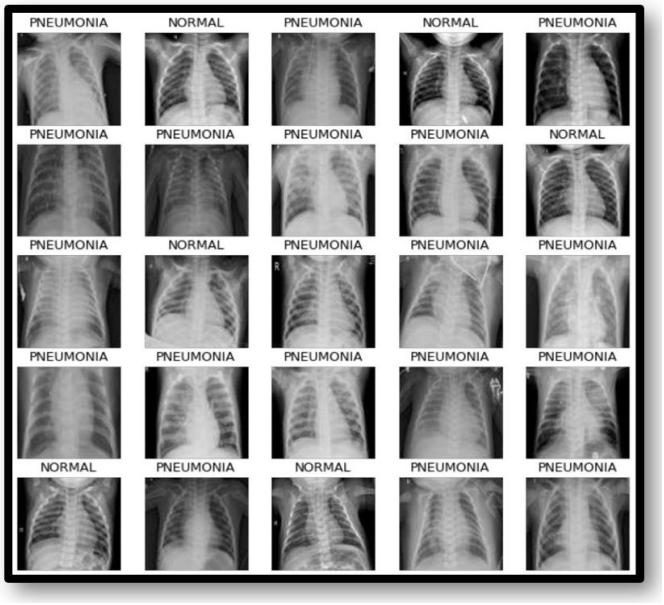


Figure 46. Data set for Pneumonia Detection

DATA PREPROCESSING

In the preprocessing pipeline for pneumonia prediction using a Convolutional Neural Network (CNN), the initial step involves loading and organizing the chest x-ray image dataset into training, validation, and test sets.This comprehensive preprocessing strategy ensures that the input data is appropriately prepared, contributing to the effectiveness and reliability of the CNN in predicting pneumonia from chest x-ray images.

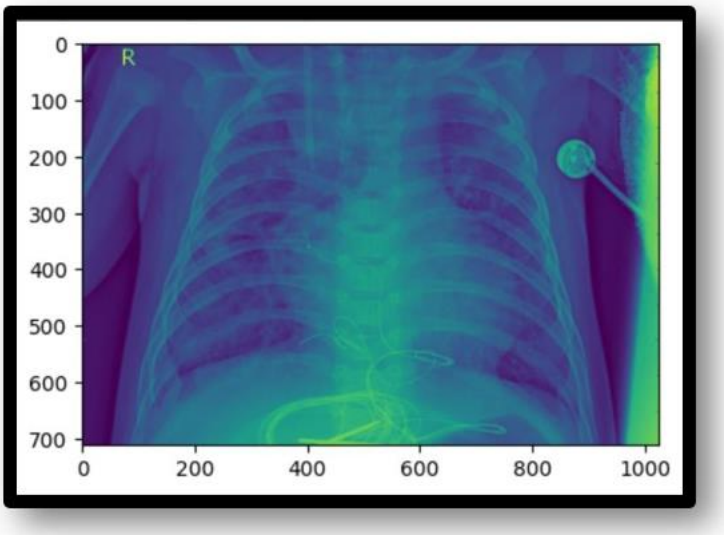


Figure 47. Loading the Data set of Pneumonia

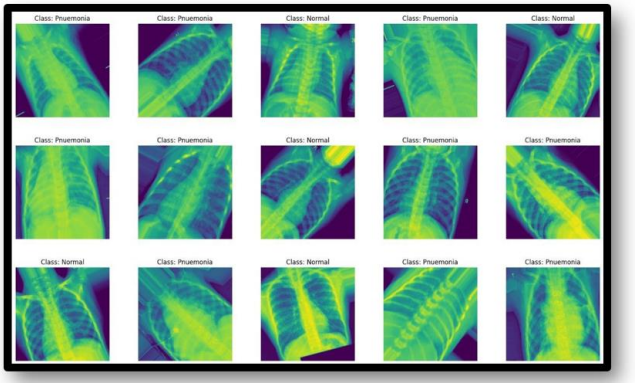


Figure 48. Data set for No Pneumonia Detection

Training the model involves dividing the input image data into training, testing, and validation sets. The training data is used to train the CNN algorithm, while the testing data is employed to validate the expected results for given inputs, including negative testing. Validation assesses how well the model performs on new data and makes predictions.

The CNN model, or Convolutional Neural Network, assigns weights to different objects in an image, allowing it to distinguish between them. CNNs excel at constructing internal representations of two-dimensional images, facilitating the learning of location and scale across various data forms, especially in image-related tasks like picture identification, object detection, and segmentation. The model is created and fine-tuned using optimizers, which adjust the model's parameters during the training process.

```
model.summary()

Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 150, 150, 32)      160

 max_pooling2d (MaxPooling2  (None, 75, 75, 32)        0
 D)

 conv2d_1 (Conv2D)           (None, 38, 38, 64)        8256

 max_pooling2d_1 (MaxPoolin  (None, 19, 19, 64)        0
 g2D)

 conv2d_2 (Conv2D)           (None, 19, 19, 128)       32896

 max_pooling2d_2 (MaxPoolin  (None, 9, 9, 128)         0
 g2D)

 flatten (Flatten)           (None, 10368)             0

 dense (Dense)               (None, 128)               1327232

 dense_1 (Dense)             (None, 1)                 129

=================================================================
Total params: 1368673 (5.22 MB)
Trainable params: 1368673 (5.22 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 49. Summary of CNN Model for Pneumonia Detection

EXPERIMENTAL RESULTS ANALYSIS

In the result analysis, X-ray images are utilized as input for the CNN model. The model analyzes the images and produces output images accompanied by captions, indicating either "pneumonia" or "normal" below them. The presence of the "pneumonia" caption suggests that the person in the image is diagnosed with pneumonia, while the "normal" caption indicates that the person is not suffering from pneumonia.

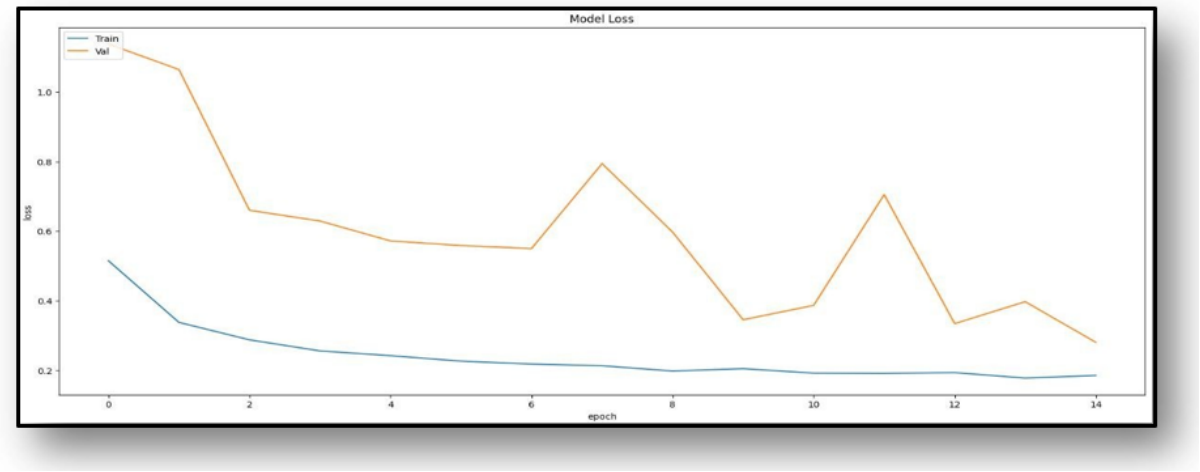


Figure 50. Plotting Training loss and Validation loss

VALIDATION DATA

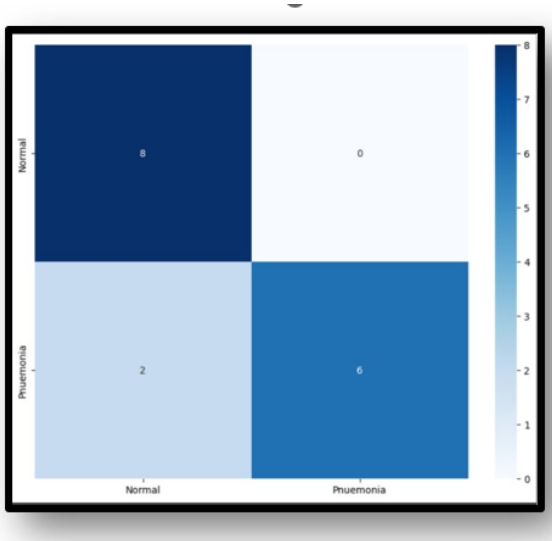


Figure 52. Plotting Heatmap of y_test and y_pred(traing data)

TEST DATA

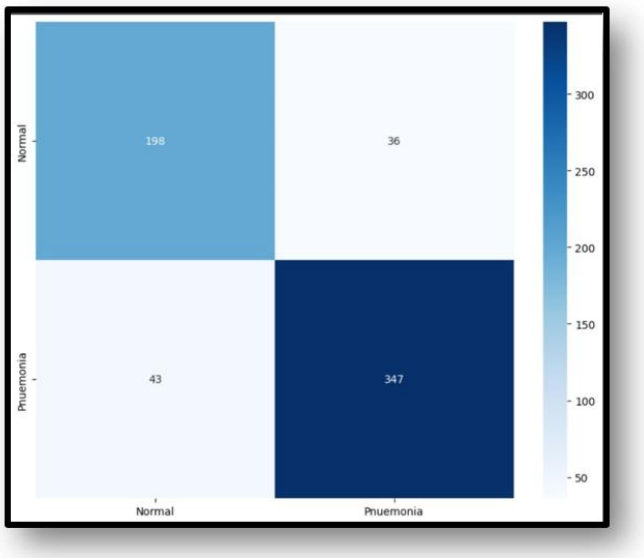


Figure 53. Plotting Heatmap of y_test and y_pred(test data)

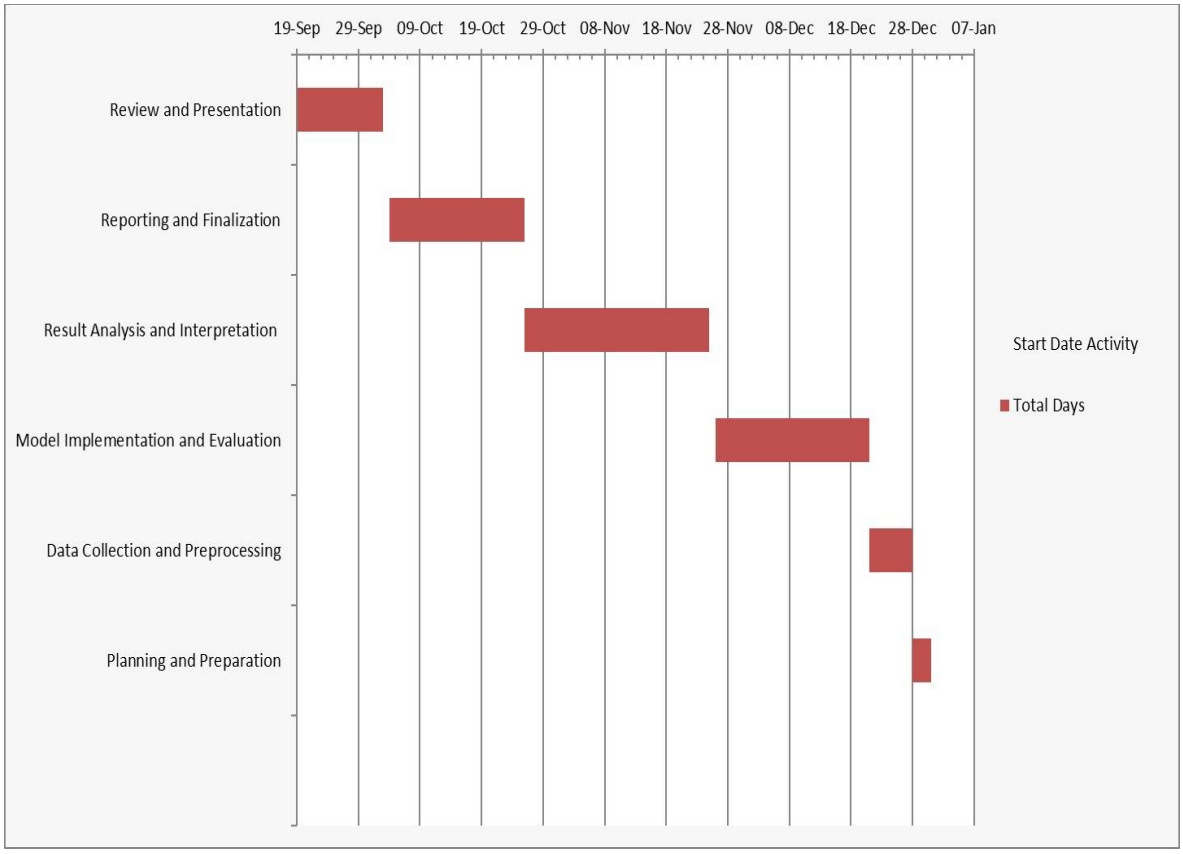
CNN ACCURACY FOR PNUEMONIA DETECTION	**85%**

Table. 11. CNN accuracy for pnuemonia prediction

CHAPTER-7

TIMELINE FOR EXECUTION OF PROJECT

(GANTT CHART)



CHAPTER-8

OUTCOMES

The potential outcomes are as follows.

1. Identification of Optimal Algorithms:

Objective: Determine the best-performing machine learning algorithms for heart disease prediction.

Explanation: Through a comprehensive comparative analysis, this outcome allows for the identification of algorithms that demonstrate superior performance in accurately predicting heart disease risks. Healthcare professionals can leverage this information to choose models that align with their specific clinical requirements, optimizing diagnostic accuracy and patient care.

2. Performance Metrics Insights:

Objective: Provide detailed insights into performance metrics for each algorithm.

Explanation: This outcome offers a deeper understanding of performance metrics (such as accuracy, precision, recall, etc.) for each machine learning algorithm used in heart disease prediction. By highlighting strengths and weaknesses, healthcare practitioners gain valuable insights into the specific scenarios or patient demographics where each algorithm excels or falls short. These insights facilitate informed decision-making when selecting algorithms for real-world clinical applications.

3. Model Interpretability:

Objective: Offer insights into the interpretability of machine learning models.

Explanation: This outcome focuses on explaining how each algorithm generates predictions, ensuring that healthcare professionals can comprehend and trust the model's decision-making process. By providing insights into model interpretability, such as feature importance or decision rationale, it enhances the transparency and trustworthiness of the predictive models. Healthcare practitioners can better interpret and explain the predictions, fostering confidence in the model's utility and aiding in its adoption in clinical settings.

Each of these outcomes contributes significantly to the successful integration of machine learning models into clinical practice for heart disease prediction. They empower healthcare professionals with the necessary information and insights to make informed decisions, ultimately improving patient care and diagnostic accuracy.

CHAPTER-9

RESULTS AND DISCUSSION

1. Key Findings and Implications

Our investigation into lifestyle disease prediction, employing robust machine learning models, unfolds critical insights into the susceptibility of individuals to heart disease, diabetes, chronic kidney disease (CKD), and pneumonia. These findings offer a nuanced understanding of the intricate interplay between individual health characteristics and the likelihood of developing these diseases.

2. Contextualizing Within Existing Knowledge

Aligning with the primary research questions, our results seamlessly integrate with the broader landscape of lifestyle disease research. A revisit to the literature review establishes a contextual framework, emphasizing the novelty and relevance of our predictive models within the existing body of knowledge.

3. Unveiling Patterns and Unexpected Outcomes

The analysis delves into unexpected outcomes, revealing unique correlations and patterns within our predictive models. By providing plausible interpretations for these surprises, we contribute to a deeper comprehension of the complexities inherent in lifestyle disease prediction.

4. Candid Exploration of Limitations

A candid discussion of research limitations enhances the credibility of our study. Addressing areas not covered provides a transparent foundation for future research, allowing refinement and improvement of predictive models.

5. Charting Future Research Trajectories

While acknowledging our study's contributions, we suggest potential avenues for future research. Grounded in identified gaps, these recommendations serve as a roadmap for researchers to explore uncharted territories in lifestyle disease prediction.

6. Concluding Significance and Broad Implications

The discussion concludes by reiterating the most significant findings and their implications. Emphasis is placed on the pivotal role of our research in advancing lifestyle disease prediction. Readers are reminded of the connections to existing literature and the broader impact on healthcare interventions and preventive measures.

CHAPTER-10
CONCLUSION

The application of machine learning, particularly in the form of predictive models, for lifestyle disease prediction has demonstrated significant potential across various health domains. In the case of diabetes, heart disease, chronic kidney disease, and pneumonia, these models offer valuable tools for early detection, intervention, and personalized healthcare.

For diabetes and heart disease, predictive models analyze diverse data sources, including patient health records, lifestyle factors, and genetic predispositions. These models can identify individuals at high risk, enabling healthcare providers to implement preventive measures and lifestyle interventions. The proactive approach facilitated by these models not only improves patient outcomes but also contributes to the optimization of healthcare resource allocation.

For Chronic Kidney Disease (CKD), our predictive model follows a similar proactive approach seen in models for diabetes and heart disease. By analyzing an array of data sources encompassing demographic details, physiological indicators, and lifestyle factors, our model excels in identifying individuals with an elevated risk of CKD. This foresightedness empowers healthcare providers to proactively implement preventive measures and lifestyle interventions, thereby enhancing patient outcomes. Moreover, the utilization of such predictive models contributes to the efficient allocation of healthcare resources, promoting an optimized and effective healthcare system.

Pneumonia prediction benefits significantly from machine learning, particularly in the analysis of chest x-ray images. Deep learning models, such as Convolutional Neural Networks (CNNs), can effectively discern patterns indicative of pneumonia, aiding clinicians in prompt diagnosis and treatment initiation. The speed and efficiency brought about by these models are particularly crucial in managing infectious diseases like pneumonia.

In conclusion, the utilization of machine learning in predicting lifestyle diseases proves transformative for healthcare. These models empower healthcare professionals with predictive insights, enabling early interventions, personalized treatment plans, and improved patient outcomes. As technology continues to advance, the integration of artificial intelligence in healthcare holds the promise of a more proactive, precise, and patient-centered approach to disease prevention and management. However, it is essential to address ethical considerations, data privacy, and ensure continuous refinement of these models through collaborative research and validation in real-world healthcare settings.

REFERENCES

[1] Shraddha Subhash Shirsath, Prof. Shubhangi Patil Disease Prediction Using Machine Learn.Over Big Data". International Journal of Innovative Research in Science, Engineering and Technology, [2018]. ISSN (Online) : 2319-8753, ISSN (Print) : 2347-6710.

[2] Vinitha S, Sweetlin S, Vinusha H, Sajini S. "Disease Prediction Using Machine Learning Over Big Data". Computer Science & Engineering: An International Journal (CSEIJ), Vol.8, No.1, [2018]. DOI:10.5121/cseij.2018.8101.

[3] Sayali Ambekar and Dr.Rashmi Phalnikar. "Disease Prediction by using Machine Learning". International journal of computer engineering and applications, Volume XII, special issue, May 18. ISSN: 2321-3469.

[4] Lohith S Y, Dr. Mohamed Rafi. "Prediction of Disease Using Learning over Big Data - Survey". International Journal on Future Revolution in Computer Science & Communication Engineering. ISSN:2454-4248.

[5] J. Senthil Kumar, S. Appavu. "The Personalized Disease Prediction Care from Harm using Big Data Analytics in Healthcare". Indian Journal of Science and Technology, vol 9(8), DOI:10.17485/ijst/2016/v9i8/87846, [2016]. ISSN (Print): 0974-6846, ISSN (Online): 0974-5645.

[6] A Survey on Disease Prediction by Machine Learning over Big Data from Healthcare Communities International organization of Scientific Research 59 | Page Gakwaya Nkundimana Joel, S. Manju Priya. "Improved Ant Colony on Feature Selection and Weighted Ensemble to Neural Network Based Multimodal Disease Risk Prediction (WENN-MDRP) Classifier for Disease Prediction Over Big Data". International Journal of Engineering & Technology, 7(3.27) (2018) 56-61.

[7] Asadi Srinivasulu, S.Amrutha Valli, P.Hussain Khan, and P.Anitha. "A Survey on Disease Prediction in Big Data Healthcare using an Extended Convolutional Neural Network". National conference on Emerging Trends in information, management and Engineering Sciences, [2018].

[8] Stephen J.Mooney and Vikas Pejaver. "Big data in public health: Terminology, Machine Learning, and Privacy", Annual Review of Public Health [2018]. Smriti Mukesh Singh, Dr. Dinesh B. Hanchate. "Improving Disease Prediction by Machine Learning". eISSN: 2395-0056, p-ISSN:2395-0072.

[9] Joseph, Nisha, and B. Senthil Kumar. "Top-K Competitor Trust Mining and Customer Behavior

[10] Investigation Using Data Mining Technique."Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 8.2 (2018). Kumar, B. Senthil. "Adaptive Personalized Clinical Decision Support System Using Effective Data Mining Algorithms." Journal of Network Communications and Emerging Technologies (JNCET) www.jncet. org 8.1 (2018).

[11] Unnikrishnan, Asha, and B. Senthil Kumar. "Biosearch: A Domain Specific Energy Efficient Query Processing and Search Optimization in Healthcare Search Engine." Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 8.1 (2017).

[12] Kumar, B. Senthil. "Adaptive Personalized Clinical Decision Support System Using Effective Data Mining Algorithms." Journal of Network Communications and Emerging Technologies (JNCET) www.jncet. org 8.1 (2017).

[13] Kumar, B. Senthil. "Data Mining Methods and Techniques for Clinical Decision Support Systems." Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 7.8 (2017). Communications and Emerging Technologies (JNCET) www. jncet. org 7.8 (2017).

[14] IEEE- International Conference On Advances In Engineering, Science And Management (ICAESM-2012) March 30,31,2012 2012-IEEE, Japan-Egypt Conference on Electronics, Communications and Computers.

[15] Proceedings of the World Congress on Engineering and Computer Science Vol II WCECS 2014, 24 October, 2014, San Fransisco 2015 International Conference on Technologies for Sustainable Development (ICTSD- 2015) 6. International Journal on Recent and Innovation Trends in Computing and Communication| volume 2- 7. 20

APPENDIX-A

PSUEDOCODE

TRAINING AND TESTING DATA

```
[ ] #train_test_splitting of the dataset
    x = data.drop(columns = 'Outcome')

    # Getting Predicting Value
    y = data['Outcome']

    from sklearn.model_selection import train_test_split
    x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=0)
```

```
[ ] print(len(x_train))
    print(len(x_test))
    print(len(y_train))
    print(len(y_test))
```

```
614
154
614
154
```

MODELS

1. Logistic Regression

```
[ ] from sklearn.linear_model import LogisticRegression
    reg = LogisticRegression()
    reg.fit(x_train,y_train)
```

```
LogisticRegression
LogisticRegression()
```

```
y_pred=reg.predict(x_test)
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
print("Classification Report is:\n",classification_report(y_test,y_pred))
print("Confusion Matrix:\n",confusion_matrix(y_test,y_pred))
print("Training Score:\n",reg.score(x_train,y_train)*100)
print("Mean Squared Error:\n",mean_squared_error(y_test,y_pred))
print("R2 score is:\n",r2_score(y_test,y_pred))
```

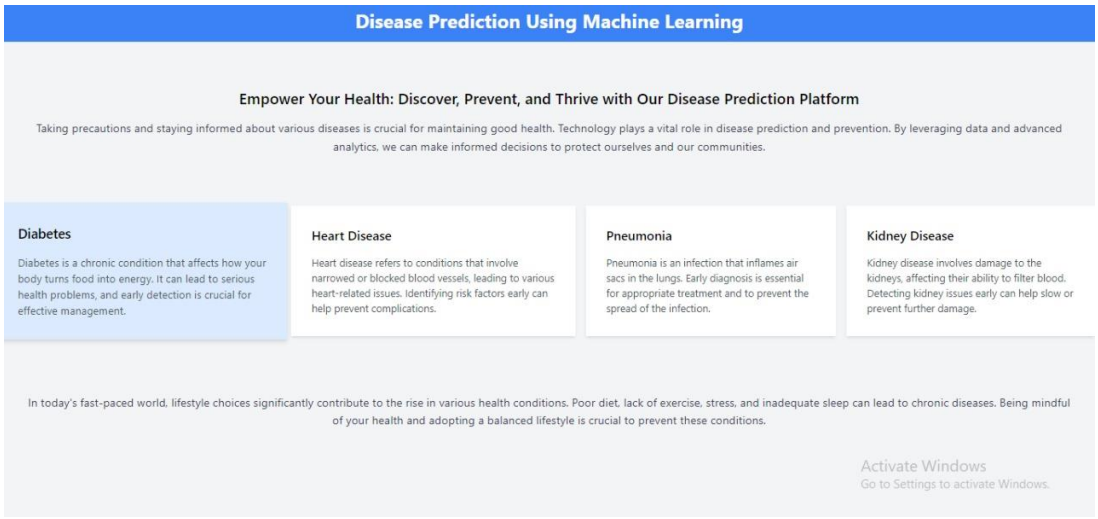
```
# Get a list of all classification models
models = get_classification_models()

# Initialize an ensemble voting classifier
ensemble = VotingClassifier(estimators=models, voting='hard')

# Fit the ensemble classifier to the training data
ensemble.fit(X_train, y_train)

# Make predictions on the test data
y_pred = ensemble.predict(X_test)

# Calculate the accuracy of the ensemble classifier
accuracy = accuracy_score(y_test, y_pred)

# Print the accuracy of the ensemble classifier
print("Ensemble Accuracy:", accuracy)
```

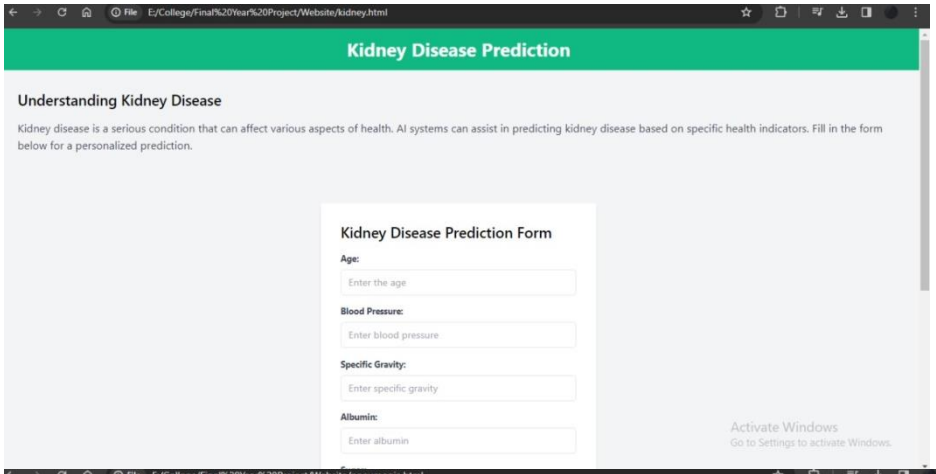
Fig-17-Fiiting the model using ensemble.

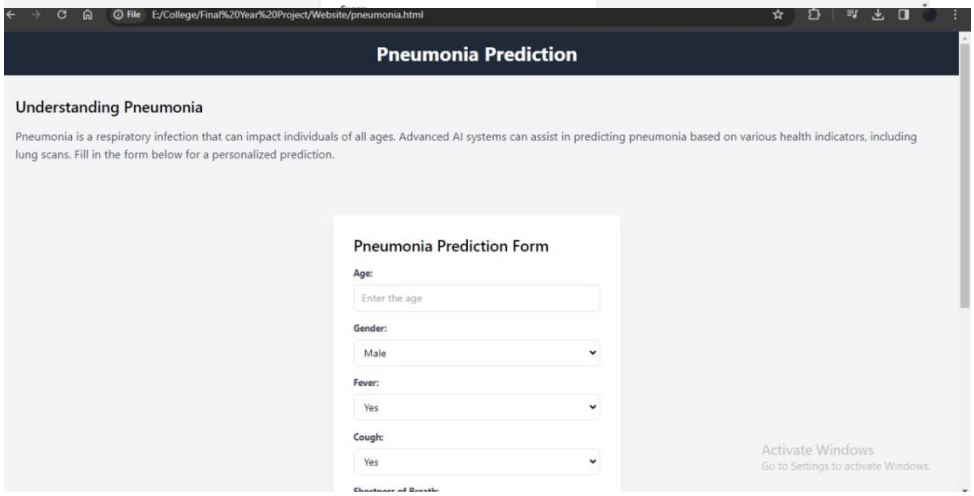
```
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, BaggingClassifier, VotingClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.neighbors import NearestCentroid
from sklearn.linear_model import RidgeClassifier, PassiveAggressiveClassifier
from sklearn.linear_model import SGDClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from catboost import CatBoostClassifier

# Function to gather all classification models
def get_classification_models():
    models = []
    models.append(('Logistic Regression', LogisticRegression()))
    models.append(('Naive Bayes', GaussianNB()))
    models.append(('Decision Tree', DecisionTreeClassifier()))
    models.append(('Random Forest', RandomForestClassifier()))
    models.append(('Gradient Boosting', GradientBoostingClassifier()))
    models.append(('Support Vector Machine', SVC()))
    models.append(('K-Nearest Neighbors', KNeighborsClassifier()))
    models.append(('Neural Network', MLPClassifier()))
    models.append(('AdaBoost', AdaBoostClassifier()))
    models.append(('Bagging', BaggingClassifier()))
    models.append(('Linear Discriminant Analysis', LinearDiscriminantAnalysis()))
    models.append(('Quadratic Discriminant Analysis', QuadraticDiscriminantAnalysis()))
    models.append(('Gaussian Process Classifier', GaussianProcessClassifier()))
    models.append(('XGBoost', XGBClassifier()))
    models.append(('CatBoost', CatBoostClassifier()))
    models.append(('Ridge Classifier', RidgeClassifier()))
    models.append(('SGD Classifier', SGDClassifier()))

    return models
```

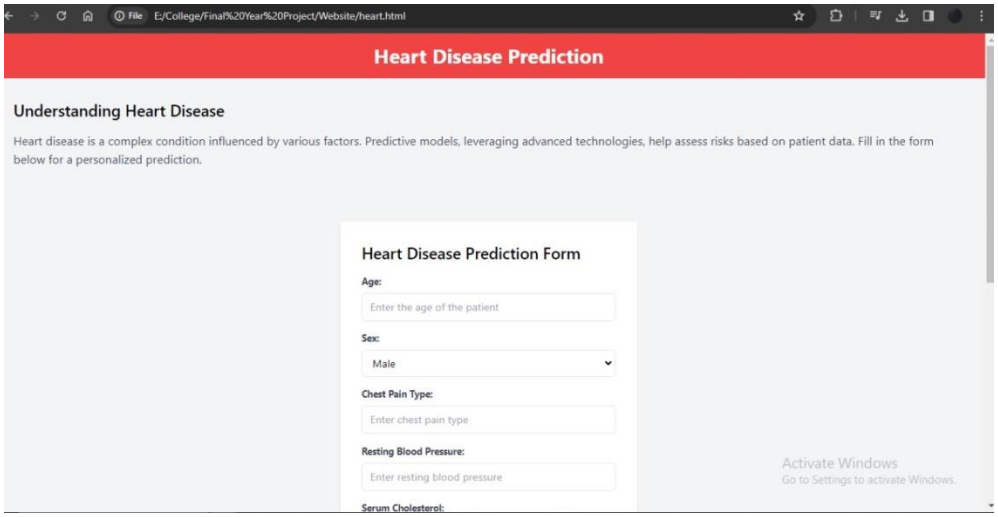
APPENDIX-B

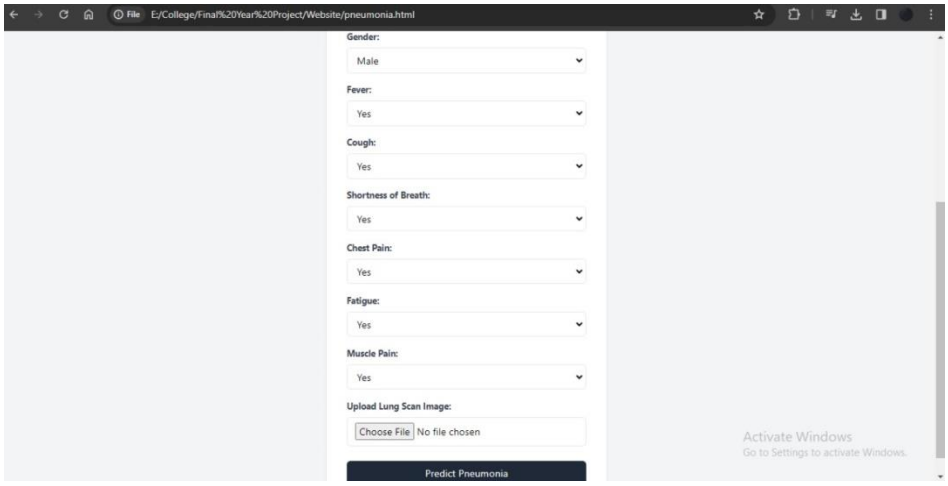
SCREENSHOTS

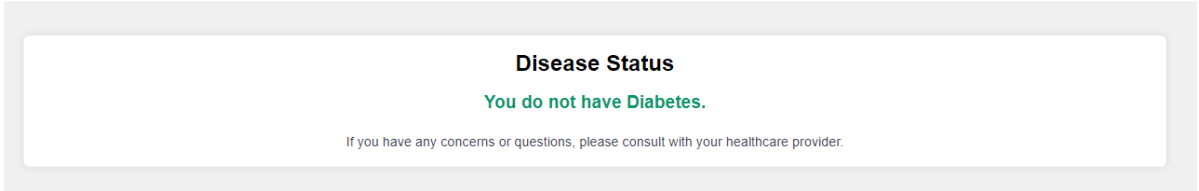







File E:/College/Final%20Year%20Project/Website/heart.html
Heart Disease Prediction
Understanding Heart Disease
Heart disease is a complex condition influenced by various factors. Predictive models, leveraging advanced technologies, help assess risks based on patient data. Fill in the form below for a personalized prediction.
Heart Disease Prediction Form
Age:
Enter the age of the patient
Sex:
Male
Chest Pain Type:
Enter chest pain type
Resting Blood Pressure:
Enter resting blood pressure
Serum Cholesterol:
Activate Windows
Go to Settings to activate Windows.


File E:/College/Final%20Year%20Project/Website/pneumonia.html
Gender:
Male
Fever:
Yes
Cough:
Yes
Shortness of Breath:
Yes
Chest Pain:
Yes
Fatigue:
Yes
Muscle Pain:
Yes
Upload Lung Scan Image:
Choose File No file chosen
Predict Pneumonia
Activate Windows
Go to Settings to activate Windows.


Disease Status
You do not have Diabetes.
If you have any concerns or questions, please consult with your healthcare provider.

APPENDIX-C

ENCLOSURES

Journal Publication Details :

IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ref No : IJCRT/Vol 12/ Issue 1 / 343

To,
Faiz Khan

Subject: Publication of paper at International Journal of Creative Research Thoughts.

Dear Author,

With Greetings we are informing you that your paper has been successfully published in the International Journal of Creative Research Thoughts - IJCRT (ISSN: 2320-2882). Thank you very much for your patience and cooperation during the submission of paper to final publication Process. It gives me immense pleasure to send the certificate of publication in our Journal. Following are the details regarding the published paper.

About IJCRT	: Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly, Indexing in all major database & Metadata, Citation Generator, Digital Object Identifier(DOI) | UGC Approved Journal No: 49023 (18)
Registration ID	: IJCRT_249564
Paper ID	: IJCRT2401343
Title of Paper	: Lifestyle Disease Prediction
Impact Factor	: 7.97 (Calculate by Google Scholar) | License by Creative Common 3.0
Publication Date	: 12-January-2024
DOI	:
Published in	: Volume 12 | Issue 1 | January 2024
Page No	: c826-c834
Published URL	: http://www.ijcrt.org/viewfull.php?&p_id=IJCRT2401343
Authors	: Faiz Khan, Jeevith Joseph CJ, Tejas Gowda C, Dr.Chandrasekar V
Notification	: UGC Approved Journal No: 49023 (18)

Thank you very much for publishing your article in IJCRT.

ISSN 2320-2882 IJCRT

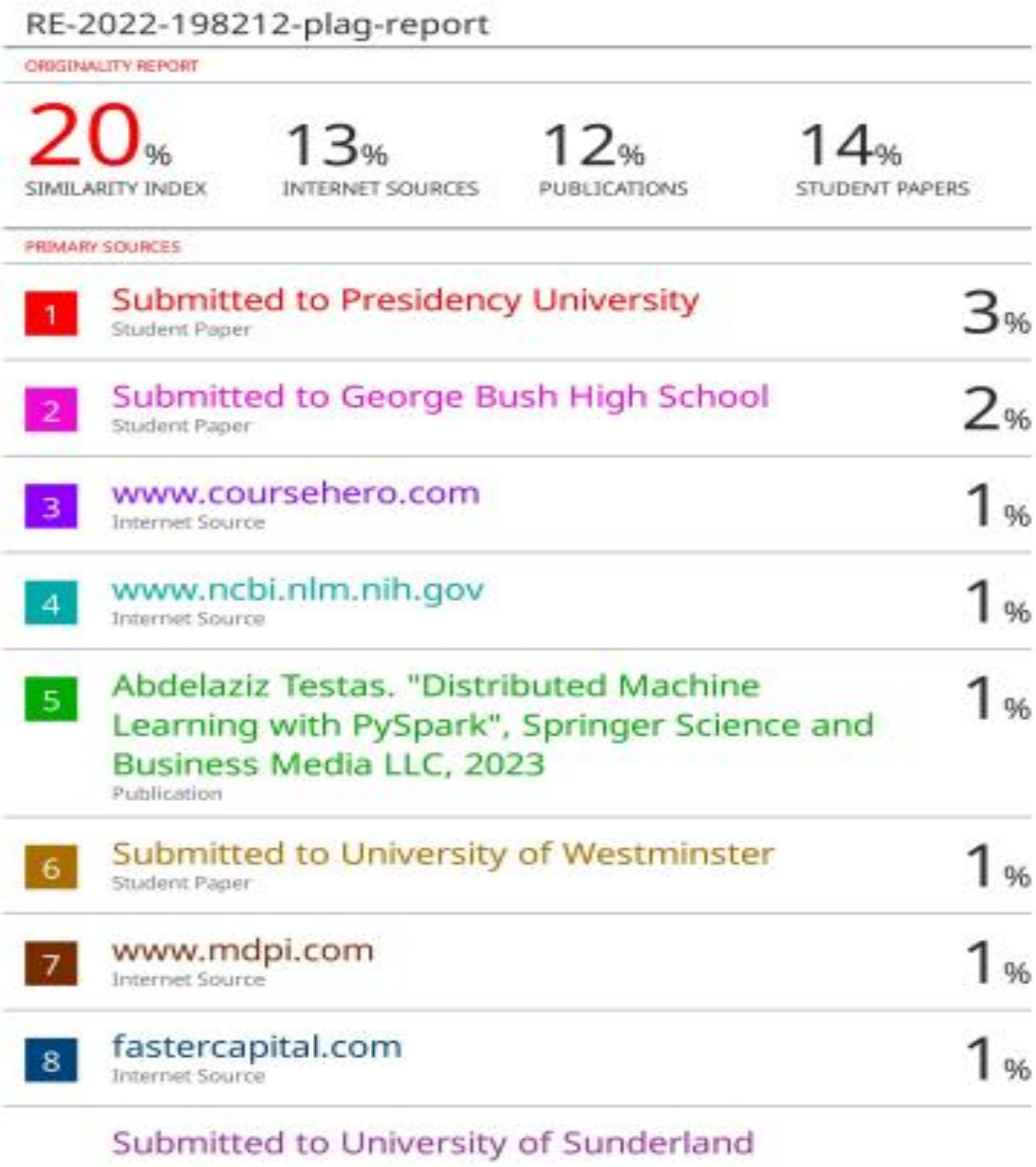
Editor In Chief
International Journal of Creative Research Thoughts - IJCRT
(ISSN: 2320-2882)

Indexing CiteSeer SSRN docstoc Google Scribd MENDELEY publons

An International Scholarly, Open Access, Multi-disciplinary, Monthly, Indexing in all major database & Metadata, Citation Generator

Website: www.ijcrt.org | Email: editor@ijcrt.org

Plagiarism Check Report :



RE-2022-198212-plag-report

ORIGINALITY REPORT

20%	13%	12%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Presidency University Student Paper	3%
2	Submitted to George Bush High School Student Paper	2%
3	www.coursehero.com Internet Source	1%
4	www.ncbi.nlm.nih.gov Internet Source	1%
5	Abdelaziz Testas. "Distributed Machine Learning with PySpark", Springer Science and Business Media LLC, 2023 Publication	1%
6	Submitted to University of Westminster Student Paper	1%
7	www.mdpi.com Internet Source	1%
8	fastercapital.com Internet Source	1%
	Submitted to University of Sunderland	

The Project work carried out here is mapped to SDG-3

SDG-3 Good Health and Well-Being.

The objective of our initiative is to contribute to the overall well-being of society by focusing on the predictive analysis of health conditions such as Diabetes, Heart Disease, Chronic Kidney Disease, and Pneumonia. By utilizing various parameters and data-driven methodologies, our project aims to educate and alert individuals about potential health risks, facilitating proactive measures for a healthier lifestyle.