VIT-AP
UNIVERSITY

SUMMER INTERNSHIP
SOPHOMORES
PROJECT

PROJECT TITLE:

“ML-Driven Early Prediction for Optimal Health – Empowering You with Accurate Predictive Health Analytics”

PRESENTED BY:

Madhu Alapaka - 21MIS7022.
Kothamasu Kaarthikeya - 21MIS7039.
Pulluru Nikhilesh - 21MIS7087.

DEPARTMENT OF SCOPE
(SECL . INTEG M.TECHSPEC.IN SOFTWARE ENGINEERING)

FINAL REVIEW:

SPJ 2001
AP20232410000130
Summer Id: 20240007

UNDER THE GUIDANCEOF

Prof. Anil Vitthalrao Turukmane
DEPARTMENT OF SCOPE
(Speci. Networking And Security)

ABSTRACT

- **Our study presents a robust predictive model for early disease detection, addressing critical conditions such as heart disease, diabetes, kidney disease, Parkinson's disease, and Hepatitis C. By evaluating various machine learning algorithms, we identified optimal models for each disease, significantly enhancing diagnostic accuracy and reliability. Random Forest (RF) achieved accuracies of 99% for heart disease and 94% for Hepatitis C, while Gradient Boosting (GB) reached 91% for diabetes and 99.25% for kidney disease. K-Nearest Neighbors (KNN) showed 95% accuracy for Parkinson’s disease. Utilizing patient data, our system aims to facilitate timely medical intervention and personalized healthcare management through a user-friendly Streamlit interface. This initiative integrates advanced AI technology into an accessible platform, supporting informed decision-making and improved disease management, ultimately contributing to better health outcomes and efficient healthcare delivery.**

CONTENTS

- Introduction
- Aim of Project
- Literature Review
- Problems & Solutions
- Frame work
- Objective&Scope
- Methodology
- Flow chart
- Architecture
- 5 Main steps in application
- Documentations(SRS&SDA&UML)
- Existing System
- Project Plan
- Adavnatges
- Implementation
- Final outputs
- ROC Graphs
- Results
- Conclusion
- Final Report & Research Paper
- References

INTRODUCTION

- **In recent years, the integration of machine learning into healthcare has revolutionized early disease detection and risk assessment. This project focuses on developing a Multiple Disease Prediction System that leverages machine learning techniques to predict the likelihood of diseases such as Diabetes, Heart Disease, and Parkinson's Disease. By utilizing patient data, the system aims to enable timely medical intervention and personalized healthcare management through a user-friendly interface built on Streamlit.**

AIM OF THE PROJECT

- The aim of this project is to develop an advanced machine learning-based predictive system for early detection and risk assessment of five critical health conditions: heart disease, diabetes, kidney disease, Parkinson's disease, and Hepatitis C. By leveraging optimal machine learning algorithms for each condition, the project seeks to enhance diagnostic accuracy and reliability. The system will be implemented through a user-friendly web application using Streamlit, allowing healthcare professionals and patients to efficiently input health data and receive timely, personalized predictions. Ultimately, the project aims to improve healthcare outcomes by enabling early intervention and proactive management of these chronic diseases.

LITERATURE REVIEW

01 Rahul Shukla et al. (2023): Developed a prediction system for diseases such as heart disease, Parkinson’s disease, breast cancer, and diabetes using various ML algorithms. Logistic Regression and SVM showed high performance with accuracies up to 97%. The system uses PCA and correlation matrices for feature reduction and is deployed as a web app, emphasizing the need for efficient data preprocessing and resources.

02 Arun Depak K G et al. (2023): They developed "Kidney Guard," a web application for chronic kidney disease (CKD) prediction with diet recommendations, deployed on IBM Cloud using Flask. It aids in early CKD detection and supports patient lifestyle management.

03 Mana Saleh Al Reshan et al. (2023): Their research employs hybrid deep neural networks (CNN and LSTM) for heart disease prediction, enhancing accuracy but requiring significant computational resources and large labeled datasets

4 T. John Peter and K. Somasundaram (2012): They evaluated various classification algorithms for heart disease prediction, finding Naive Bayes most accurate and enhancing performance with dimensionality reduction. Their approach improves prediction but can be time-consuming with large datasets.

LITERATURE REVIEW

05 Puneet et al. (2021): Their study applies ensemble learning with classifiers like KNN, SVM, Decision Tree, and Random Forest for coronary heart disease prediction, achieving up to 83.2% accuracy. Ensemble methods improve accuracy but require extensive computational resources and careful class imbalance handling.

06 AH Chen et al. (2011): They developed the HDPS system using an ANN trained on 13 clinical features for heart disease prediction, achieving 80% accuracy with 85% sensitivity and 70% specificity. Challenges include data quality and computational resource requirements.

07 Karthikeyan et al. (2022): Developed a disease prediction system using Random Forest for heart disease, diabetes, and kidney disease, achieving accuracies of 98.05%, 92.30%, and 99.17% respectively. The system includes an easy-to-use interface and provides metrics such as accuracy, precision, recall, and F1-Score. Challenges include ensuring high-quality training data and sufficient computational resources.

08 J. Doe et al. (2023): Created a disease prediction system using multiple machine learning algorithms, including Random Forest, SVM, and KNN, with accuracies of 98.3% for diabetes (Random Forest) and 99.17% for kidney disease (Random Forest). The system, built with Streamlit, features ROC curve displays and performance metrics, facing challenges with data quality and computational resources.

PROBLEM

- Current disease prediction systems face several significant challenges that impact their effectiveness and accessibility. A primary issue is the lack of early detection, as existing systems may not provide timely identification of multiple diseases. This delay in detection can result in worsening conditions and diminished treatment effectiveness. Additionally, limited access to diagnostics is a critical concern; many patients are unable to access comprehensive diagnostic facilities due to geographical or financial constraints, which hampers their ability to obtain timely and accurate diagnoses. Another challenge is the fragmentation of patient health data, which is often dispersed across various sources. This scattered data makes it difficult to aggregate and analyze comprehensively, leading to incomplete or inaccurate predictions. Furthermore, the complexity of current predictive tools can be a barrier to their effective use, as many of these tools are not user-friendly and can be difficult for healthcare providers and patients to utilize. This complexity often results in the underutilization of these tools. Finally, traditional prediction methods may lack advanced machine learning techniques, leading to lower prediction accuracy. This limitation can contribute to mis diag noses or missed diagnoses, adversely affecting patient care.

POINTS:

- Importance of early detection for treatment
- Traditional methods are time-consuming and costly
- Need for a reliable, automated prediction system
- Quick assessment of multiple disease risks
- Facilitates early intervention
- Reduces burden on healthcare systems

PROBLEM STATEMENT

- Early detection of diseases is crucial for effective treatment and management.
- Existing systems are often limited to single disease predictions, lacking comprehensiveness.
- There is a need for an integrated system that can predict multiple diseases from diverse patient data.

SOLUTIONS:

- Integrate multiple disease prediction models into a single, unified system.
- Utilize machine learning algorithms for high accuracy and reliable predictions.
- Deploy a web application for easy access and user-friendly interaction.
- Provide timely and actionable health insights to patients and healthcare providers.

FRAMEWORK :

Overview

- The theoretical foundation of this project is based on machine learning and data science principles. We employ supervised learning techniques, including KNN,DT,LR,RF,SVM.. Feature engineering is crucial for transforming raw data into meaningful inputs, and cross-validation ensures model optimization and generalization.

Proponents

- Algorithms: logistic regression, decision trees, random forests, neural networks .....
- Evaluation metrics: accuracy, precision, recall, F1-score & Roc

SCOPE OF THE PROJECT

- The project's scope includes many crucial phases, beginning with data gathering and preparation. This involves gathering relevant datasets for Diabetes, Heart Dis- ease, Parkinson's Disease, Kidney Disease, and Hepatitis C from reputable sources. The data will undergo thorough pre-processing, including cleaning, normalization, and handling of missing values,to ensure it is suitable for model training. Featureengineering will also be performedto select and create relevantfeatures that enhancethe predictive power of the models. The project includesthe development and training of multiple machinelearning models, each tailored to a specific disease. Every condition will be investigated and optimized using a variety of methods, including Support Vector Machines, Random Forest, Decision Trees, and Logistic Regression. A thorough assessment of these models' performance will be conducted using measures like as accuracy, precision, recall, and F1-score.This stage ensures that the models are not only accurate but also reliable and generalizable to new, unseendata. Finally, the project scope extends to the deployment of the trained models into a user-friendly web application using Streamlit. This involves integrating the models into the application, designing the user interface, and ensuring the system operates smoothly and efficiently. Early illness identification and proactive healthcare management will be made easier for users of the online application, which will allow them to enter their health data and obtain fast forecasts. The project also has future-updating and -improving features, so that the system stays current with the most recent medical research and technical development

OBJECTIVE THE PROJECT:

MAIN OBJECTIVE OF THE PROJECT:

- The objective of this project is to develop a machine learning-based system for early detection and risk assessment of multiple chronic diseases, including Diabetes, Heart Disease, Parkinson's Disease, Kidney Disease, and Hepatitis C. This will involve:

1. Building and Optimizing Models: Create and fine-tune machine learning models to accurately predict the risk of these diseases based on user health data.
2. Early Detection: Enable timely identification of potential health issues to facilitate early medical intervention and proactive management.
3. User-Friendly Deployment: Develop and deploy a Streamlit-based web application that provides an intuitive interface for users to input their health information and receive instant predictions.
4. Enhancing Healthcare Outcomes: Improve health management and decision-making through reliable and accessible disease risk assessments.

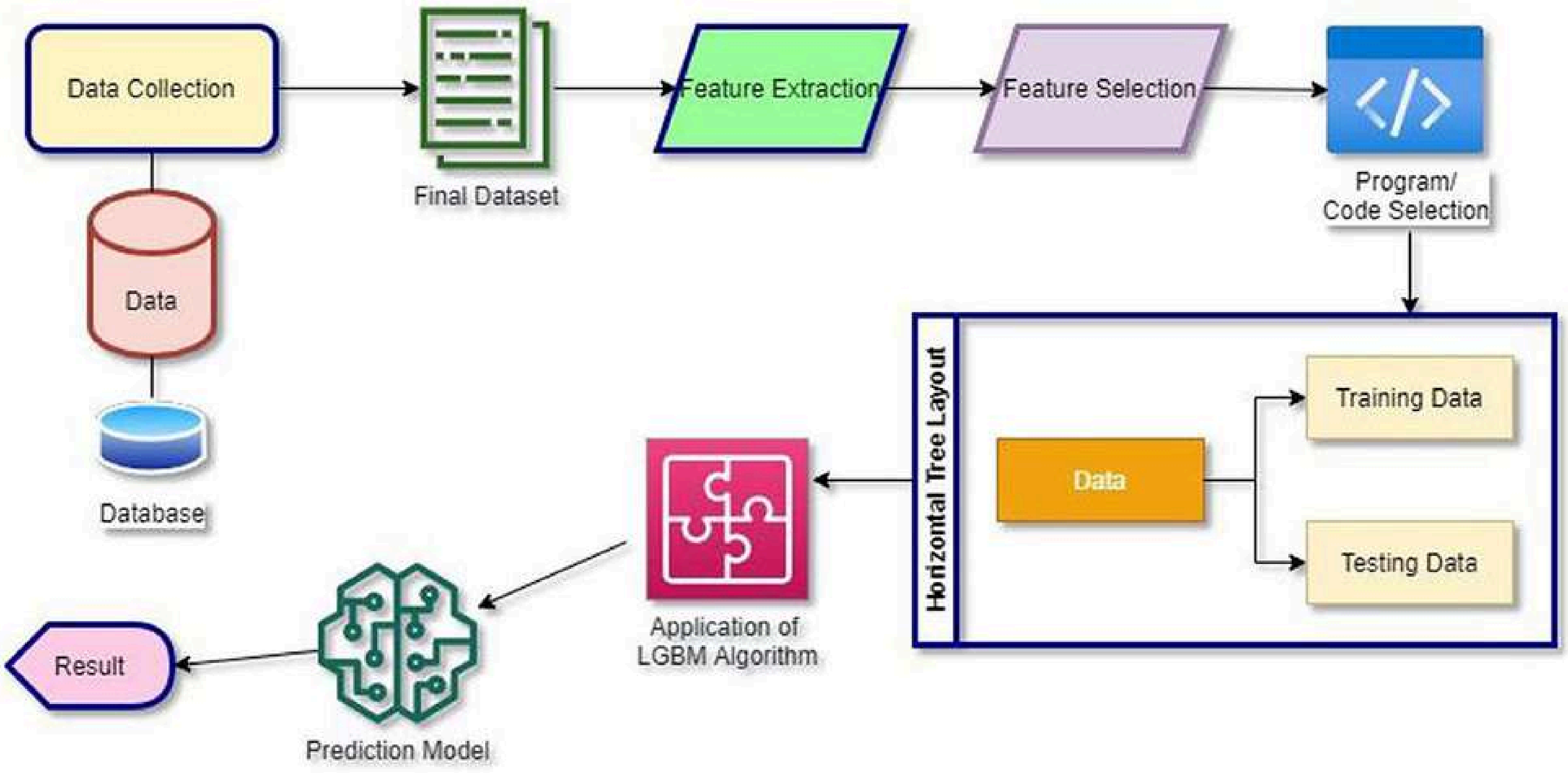
METHODOLOGY

- Data Collection: Collect health data for diseases like heart disease, diabetes, kidney disease, Parkinson's disease, and Hepatitis C from various sources.
- Data Preparation: Clean the data by fixing errors and filling in missing values. Standardize the data so it’s ready for analysis.
- Model Building:

1. Algorithm Selection: Test different machine learning algorithms (like Random Forest, Gradient Boosting, and K-Nearest Neighbors) to find the best ones for each disease.
2. Training: Train the chosen algorithms on the prepared data to learn how to predict each disease.

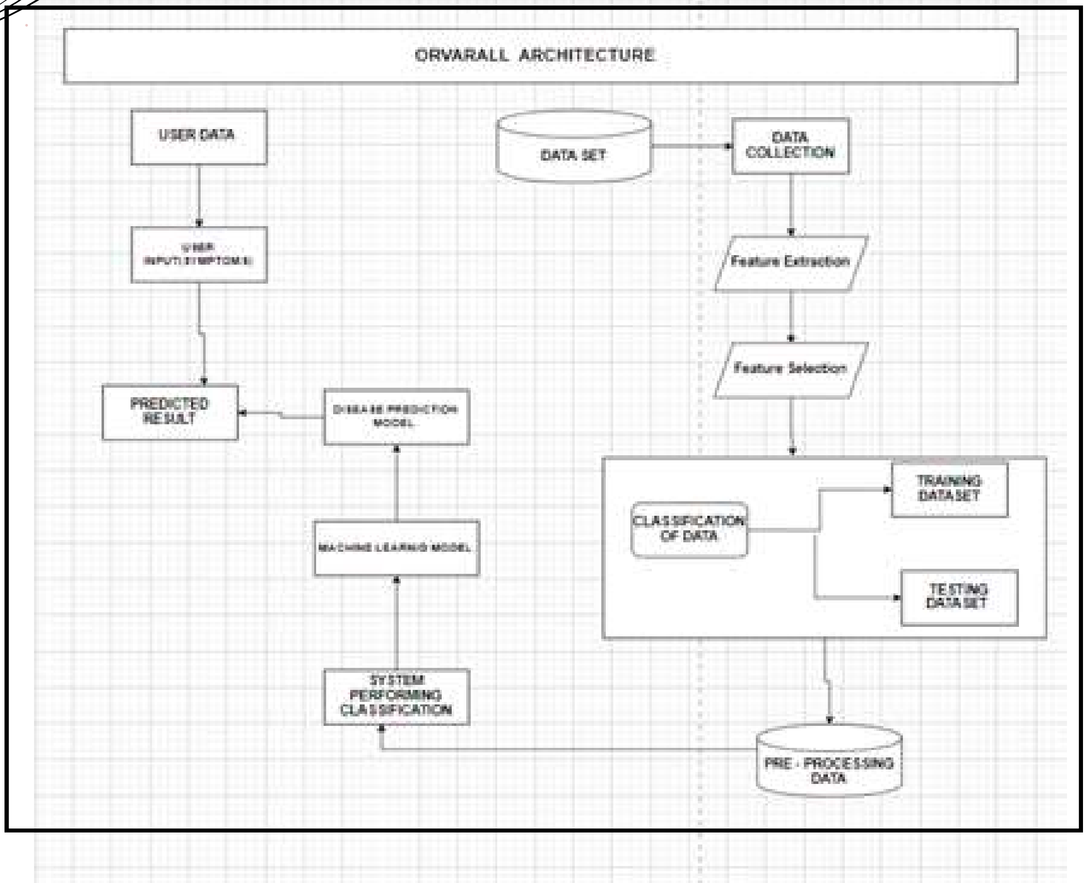
- Model Optimization: Fine-tune the algorithms to improve their accuracy and performance.
- Application Development: Create a user-friendly web application using Streamlit where users can enter their health information and get predictions.
- Testing and Launch: Test the application to make sure it works well and then launch it for use.
- Maintenance: Regularly update the system based on feedback and new data to keep it accurate and effective.

FLOW CHAT


Data Collection
Final Dataset
Feature Extraction
Feature Selection
Program/
Code Selection
Data
Database
Horizontal Tree Layout
Data
Training Data
Testing Data
Application of
LGBM Algorithm
Prediction Model
Result

- Data is collected directly from user input within the Streamlit app. This input data is then processed to extract and select relevant features necessary for the prediction model. The selected features are fed into the machine learning code implemented within the application. The data is split into training and testing sets, and the algorithm is applied to build the prediction model. This model then predicts the likelihood of multiple diseases based on the user input features. The predicted results are displayed within the Streamlit application, providing users with real-time disease risk assessments.

THE OVERALL ARCITECTURE



- The overall architecture of a multi-disease prediction system using machine learning. The process begins with data collection, where a comprehensive dataset is gathered and subjected to feature extraction and selection to identify the most relevant attributes for various diseases. This refined data is then divided into training and testing datasets. The training dataset is used to build and train the machine learning model, involving a classification process that includes a pre-processing phase to ensure data quality and consistency. Once trained, the model is integrated into the disease prediction system, which accepts user input (such as patient data) to generate predicted results for multiple diseases. This system performs classification based on the input data, leveraging the trained machine learning model to provide accurate predictions. The user data and predicted results interact seamlessly, assisting healthcare professionals in diagnosing and predicting various diseases effectively. This architecture underscores the significance of data quality, feature engineering, and robust model training in developing a comprehensive and reliable multi-disease prediction system.

THE DOCUMENTATION

- Click here to view SRS document:

  “SRS.DOC”

- Click here to view SDA document:

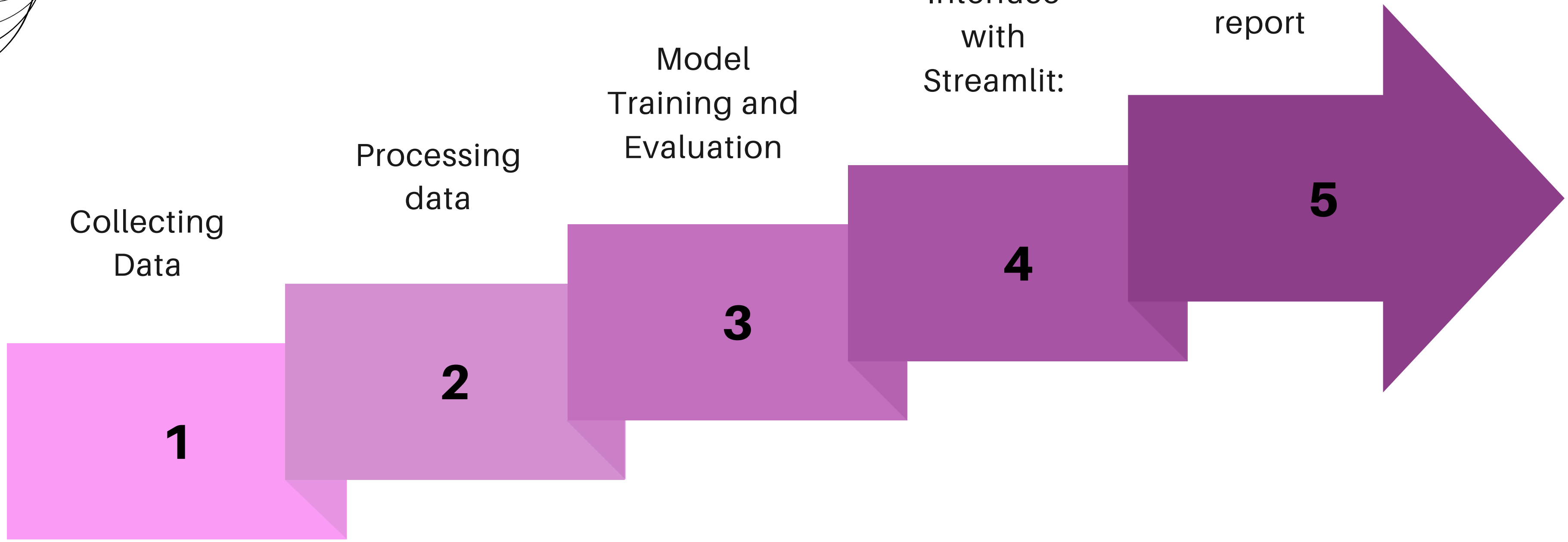
  “SDA.DOC“

- Click here to view SRS document:

  “UML DIAGRAMS .DOC“

5 MAIN STEPS INVOLED IN WEB APPLICATION

5-Step of Process



1 Collecting Data

2 Processing data

3 Model Training and Evaluation

4 User Interface with Streamlit:

5 Finally prediction report

EXISTING SYSTEM

● Phase 1: Data Preparation

- Install essential Python libraries, secure medical datasets, clean data, and engineer features for enhanced model performance.

● Phase 2:Development

- Select machine learning algorithms, train models on prepared data, evaluate performance metrics, and optimize hyperparameters.

● Phase 3: Deployment

- Select best-performing models, integrate into a Streamlit web app, and design a user-friendly interface for data input and prediction results.

● Phase 4: Result

- Conduct internal tests for functionality, gather performances of data , and collaborate with healthcare professionals to validate predictions.

●Project Plan :

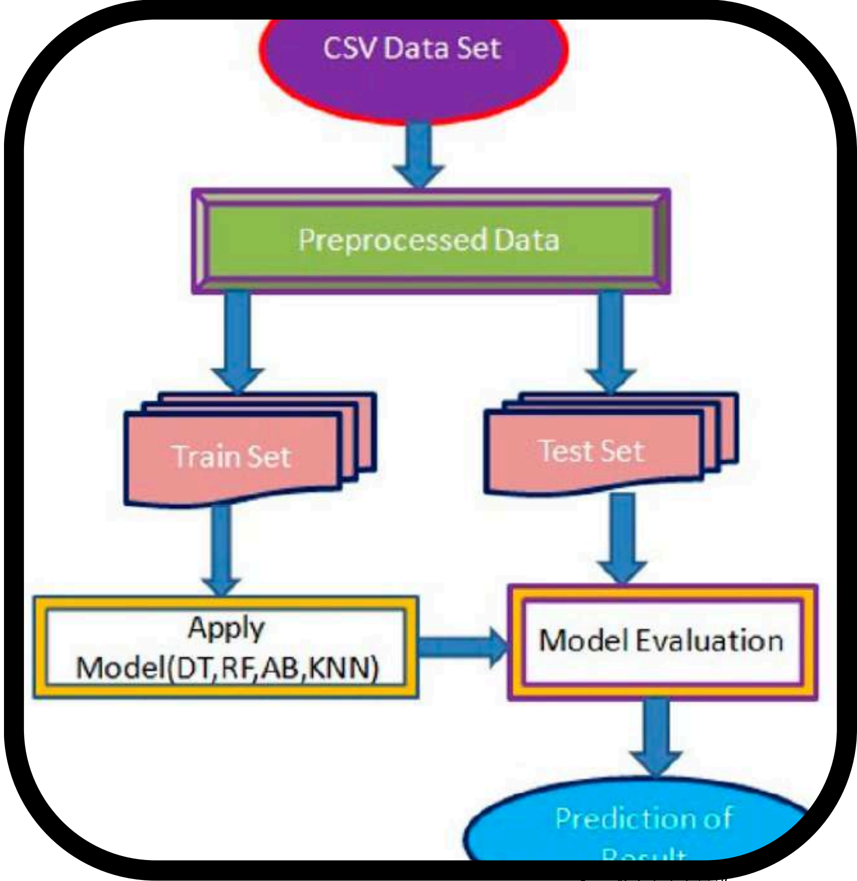
- Project Initialization & Requirement Analysis
- Research and Data Collection
- System Design and Architecture
- Model Development and Training
- Web Application Development
- Testing and Implementation
- Deployment and Result

ADVANTAGES

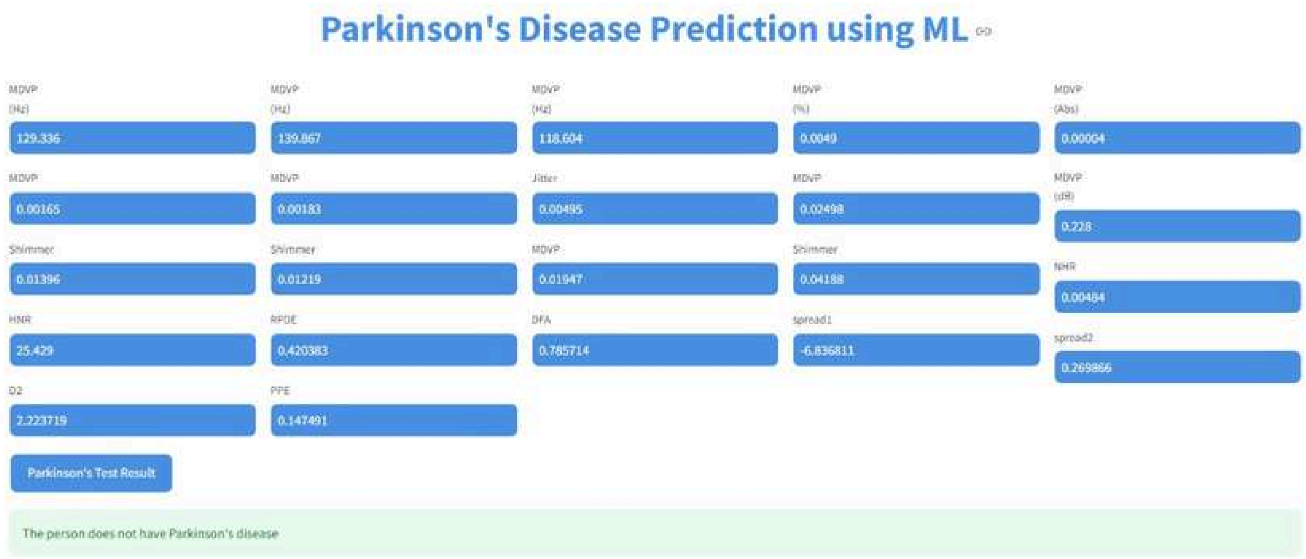
- Early Detection: Enables timely medical intervention and better health outcomes by identifying potential health issues at an early stage.
- Comprehensive: Predicts multiple diseases using diverse datasets, providing a broad spectrum of health insights.
- User-friendly: Features an interactive web application for easy data input and instant results, making it accessible for patients and healthcare providers.
- Enhanced Accuracy: Utilizes advanced algorithms and large datasets to improve the accuracy of disease predictions.
- Cost-effective: Reduces healthcare costs by preventing the need for more expensive treatments through early detection and intervention.
- Personalized Treatment: Provides personalized health insights and recommendations based on individual patient data, promoting tailored healthcare solutions.
- Support for Clinicians: Assists healthcare professionals by providing a reliable second opinion, aiding in decision-making and enhancing patient care

Implementation:

- The Execution of Multiple Disease Prediction System involves a comprehensive approach combining machine learning and web application development. Initially, data preprocessing is carried out, including augmentation, normalization, and feature selection, to prepare high-quality input for the models. Machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest ect.. are utilized to predict various diseases, with models optimized through hyperparameter tuning and regularization techniques. The backend processing and model execution are handled in Python, leveraging libraries like pandas, scikit-learn, and numpy for data manipulation and training. Streamlit is employed to build an intuitive web frontend for user interaction, while deployment is managed using services like Streamlit sharing and ngrok for secure local server exposure. Rigorous testing, including cross-validation and performance evaluation, ensures the system's accuracy and reliability, aiming to improve early disease detection and healthcare outcomes.



FINAL OUPUT OF PARKINSON DISEASE PREDECTION USING ML:



Parkinson's Disease Prediction using ML

MDVP (Hz): 129.336

MDVP (Hz): 139.867

MDVP (Hz): 118.604

MDVP (%): 0.0049

MDVP (Abs): 0.00004

MDVP: 0.00165

MDVP: 0.00183

Jitter: 0.00495

MDVP: 0.02498

MDVP (dB): 0.228

Shimmer: 0.01396

Shimmer: 0.01219

MDVP: 0.01947

Shimmer: 0.04188

NHR: 0.00484

HNR: 25.429

RPDE: 0.420383

DFA: 0.785714

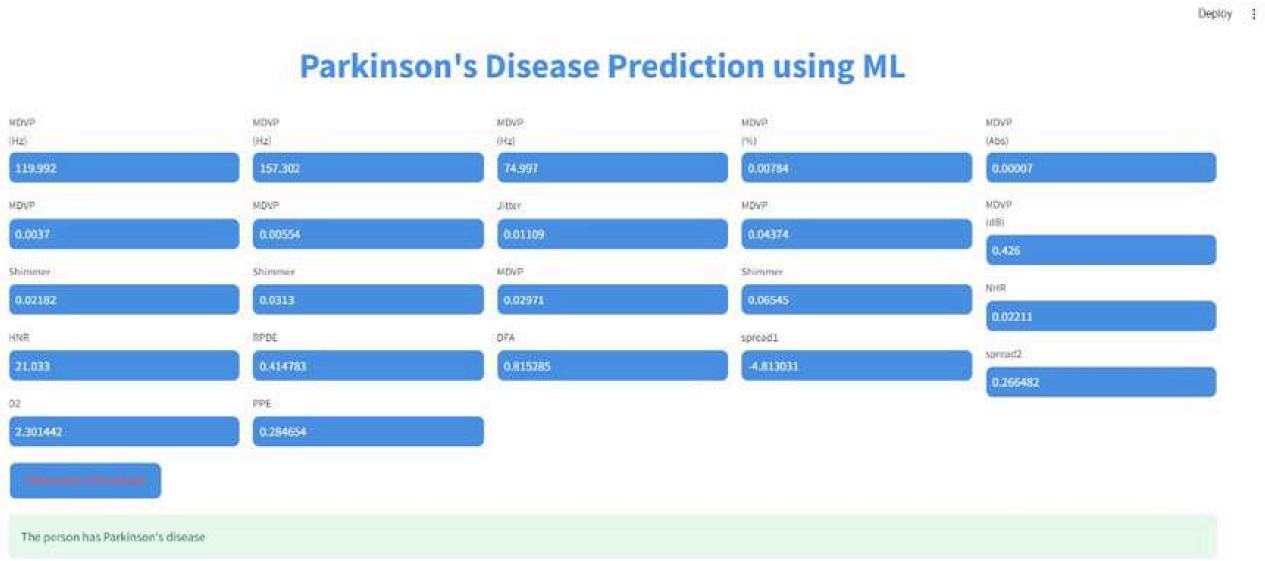
spread1: -6.836811

spread2: 0.269866

D2: 2.223719

PPE: 0.147491

The person does not have Parkinson's disease



Parkinson's Disease Prediction using ML

MDVP (Hz): 119.992

MDVP (Hz): 157.302

MDVP (Hz): 74.997

MDVP (%): 0.00784

MDVP (Abs): 0.00007

MDVP: 0.0037

MDVP: 0.00554

Jitter: 0.01109

MDVP: 0.04374

MDVP (dB): 0.426

Shimmer: 0.02182

Shimmer: 0.0313

MDVP: 0.02971

Shimmer: 0.06545

NHR: 0.02211

HNR: 21.033

RPDE: 0.414783

DFA: 0.815285

spread1: -4.813031

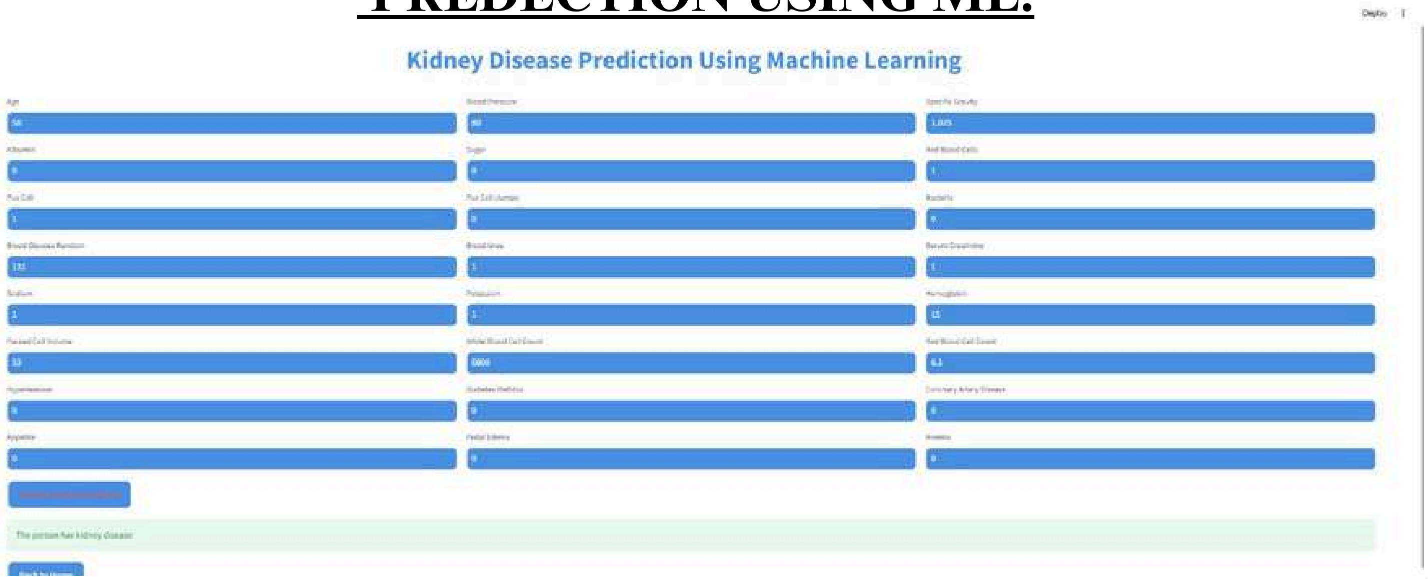
spread2: 0.266482

D2: 2.301442

PPE: 0.284654

The person has Parkinson's disease

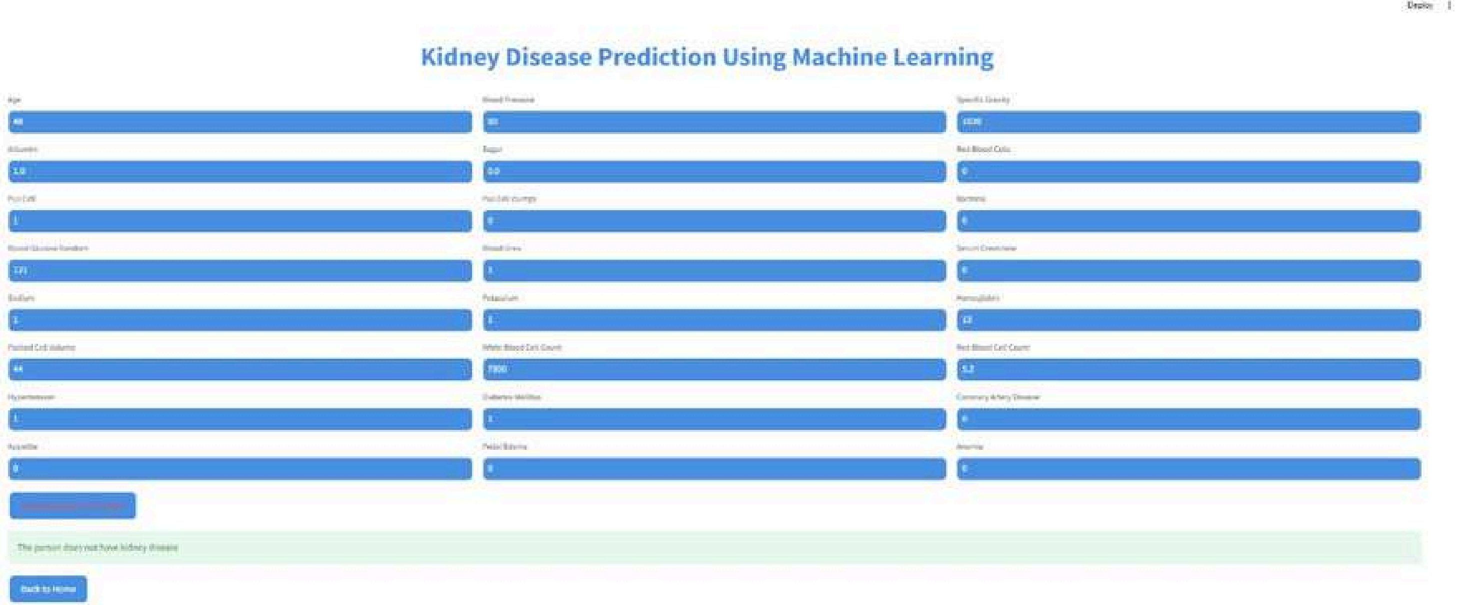
FINAL OUPUT OF KIDNEY DISEASE PREDECTION USING ML:



Kidney Disease Prediction Using Machine Learning

Age: 50
Blood Pressure: 80
Specific Gravity: 1.025
Albumin: 0
Sugar: 0
Red Blood Cells: 1
Pus Cell: 1
Pus Cell clumps: 0
Bacteria: 0
Blood Glucose Random: 131
Blood Urea: 1
Serum Creatinine: 1
Sodium: 1
Potassium: 1
Hemoglobin: 15
Packed Cell Volume: 53
White Blood Cell Count: 6800
Red Blood Cell Count: 6.1
Hypertension: 0
Diabetes Mellitus: 0
Coronary Artery Disease: 0
Appetite: 0
Pedal Edema: 0
Anemia: 0

The person has kidney disease

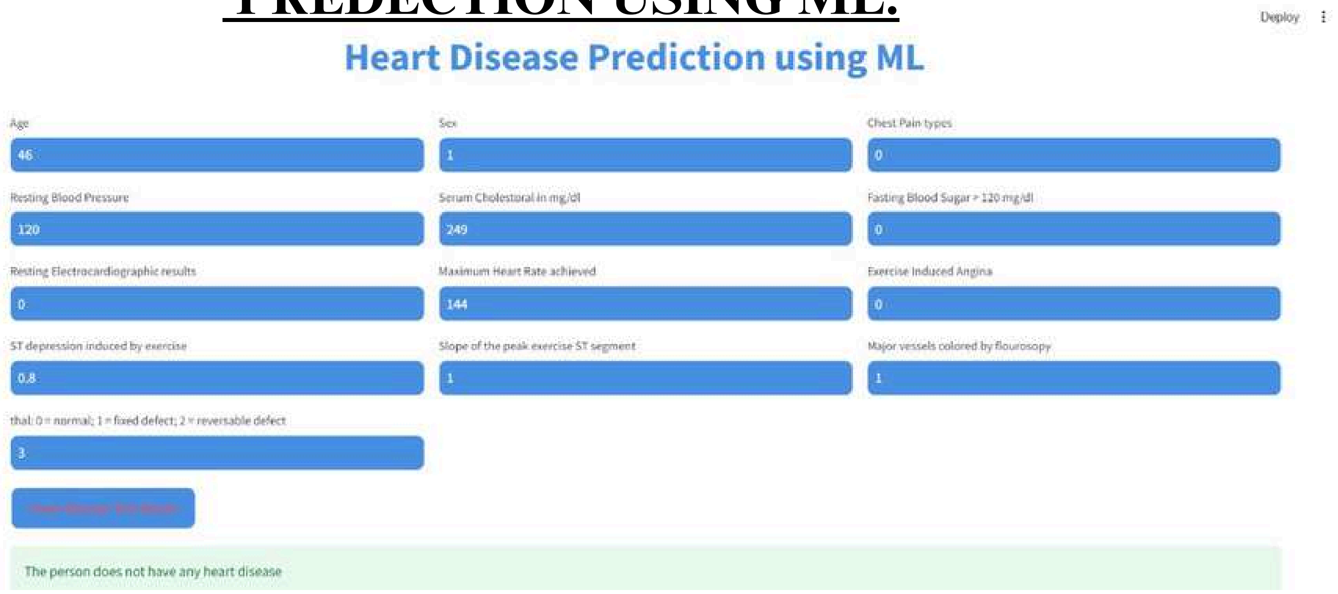


Kidney Disease Prediction Using Machine Learning

Age: 48
Blood Pressure: 80
Specific Gravity: 1020
Albumin: 1.0
Sugar: 0.0
Red Blood Cells: 0
Pus Cell: 1
Pus Cell clumps: 0
Bacteria: 0
Blood Glucose Random: 121
Blood Urea: 1
Serum Creatinine: 0
Sodium: 1
Potassium: 1
Hemoglobin: 13
Packed Cell Volume: 44
White Blood Cell Count: 7800
Red Blood Cell Count: 5.2
Hypertension: 1
Diabetes Mellitus: 1
Coronary Artery Disease: 0
Appetite: 0
Pedal Edema: 0
Anemia: 0

The person does not have kidney disease

FINAL OUPUT OF HEART DISEASE PREDECTION USING ML:



Heart Disease Prediction using ML

Age
46

Sex
1

Chest Pain types
0

Resting Blood Pressure
120

Serum Cholestoral in mg/dl
249

Fasting Blood Sugar > 120 mg/dl
0

Resting Electrocardiographic results
0

Maximum Heart Rate achieved
144

Exercise Induced Angina
0

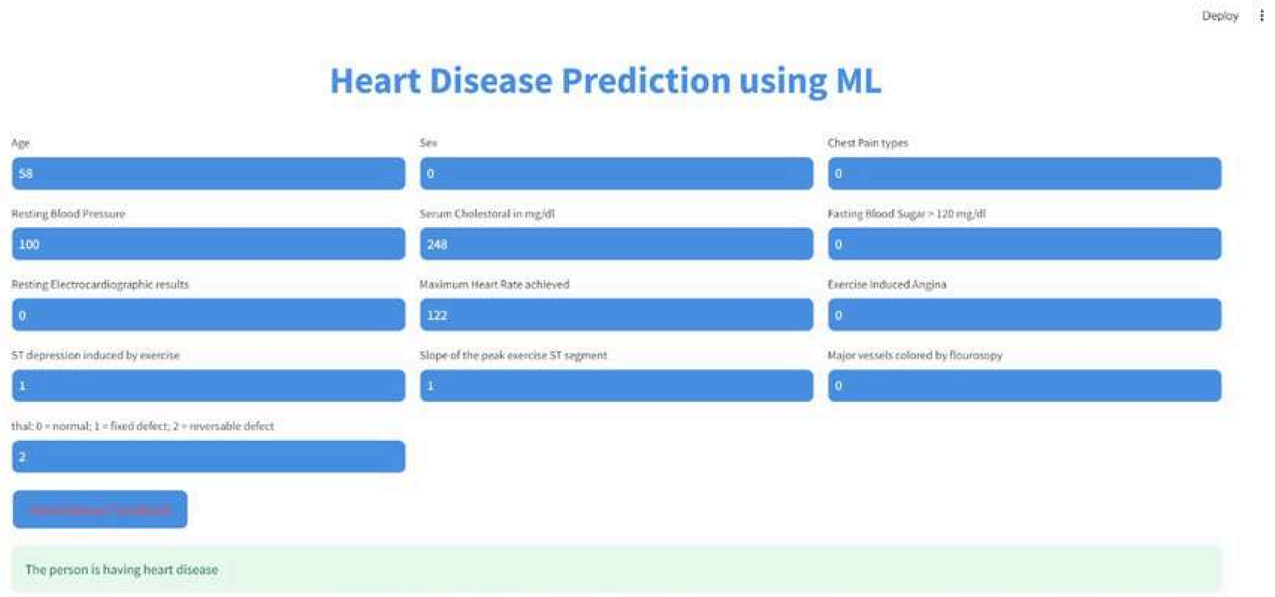
ST depression induced by exercise
0.8

Slope of the peak exercise ST segment
1

Major vessels colored by flourosopy
1

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
3

The person does not have any heart disease



Heart Disease Prediction using ML

Age
58

Sex
0

Chest Pain types
0

Resting Blood Pressure
100

Serum Cholestoral in mg/dl
248

Fasting Blood Sugar > 120 mg/dl
0

Resting Electrocardiographic results
0

Maximum Heart Rate achieved
122

Exercise Induced Angina
0

ST depression induced by exercise
1

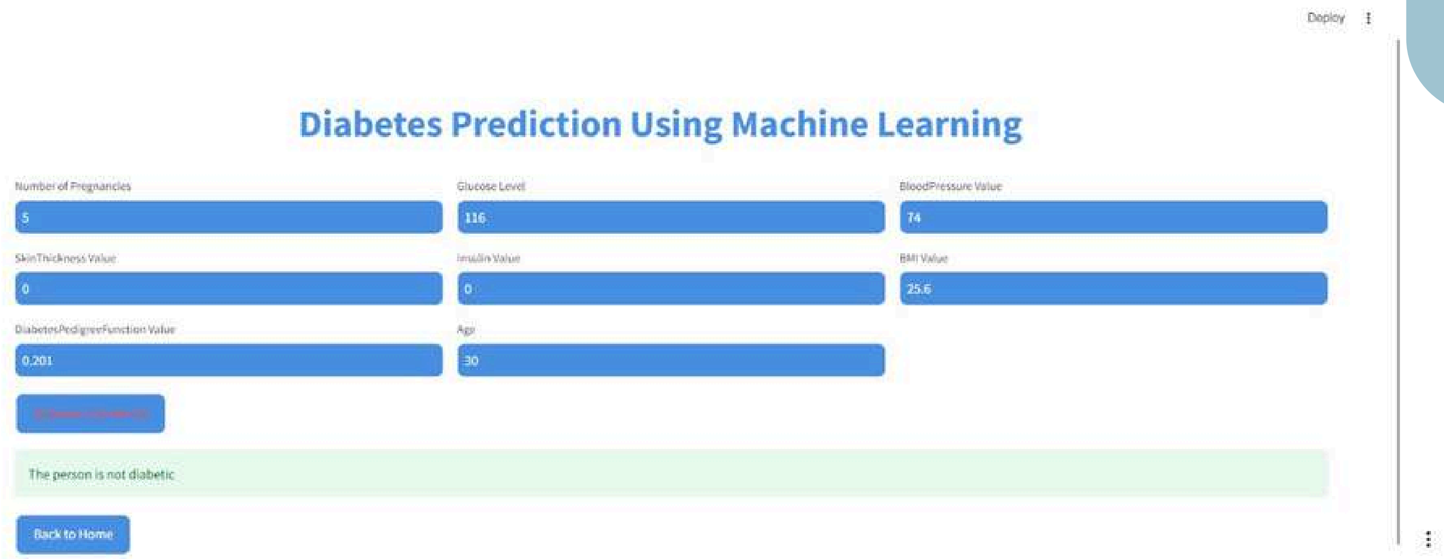
Slope of the peak exercise ST segment
1

Major vessels colored by flourosopy
0

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
2

The person is having heart disease

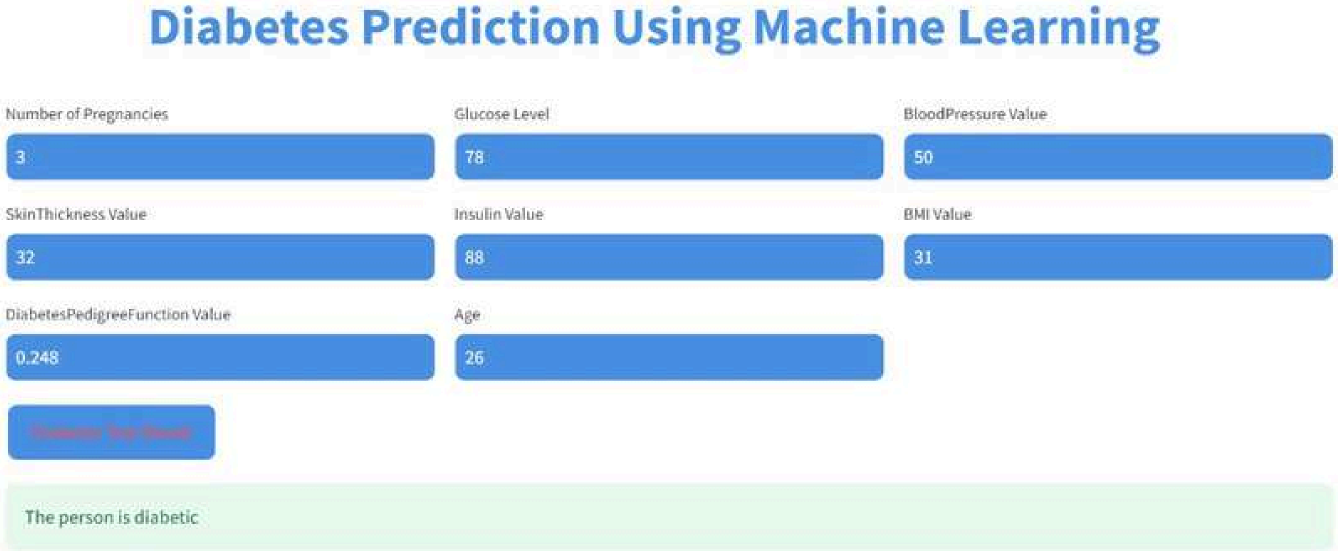
FINAL OUPUT OF DIABETES DISEASE PREDECTION USING ML:



Diabetes Prediction Using Machine Learning

Number of Pregnancies: 5
Glucose Level: 116
BloodPressure Value: 74
SkinThickness Value: 0
Insulin Value: 0
BMI Value: 25.6
DiabetesPedigreeFunction Value: 0.201
Age: 30

The person is not diabetic

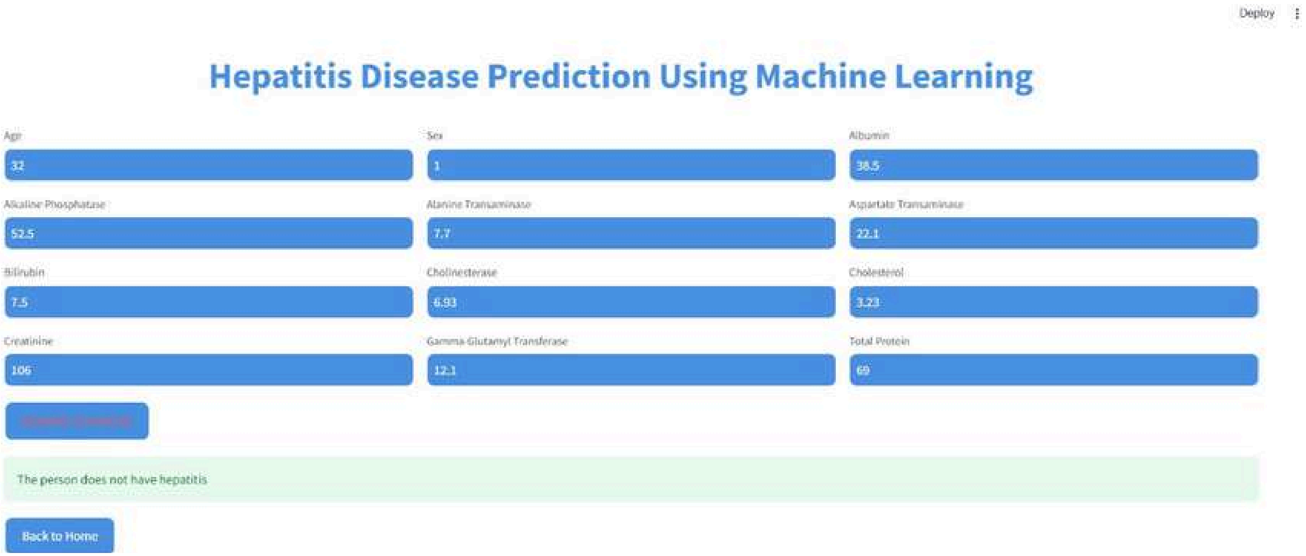


Diabetes Prediction Using Machine Learning

Number of Pregnancies: 3
Glucose Level: 78
BloodPressure Value: 50
SkinThickness Value: 32
Insulin Value: 88
BMI Value: 31
DiabetesPedigreeFunction Value: 0.248
Age: 26

The person is diabetic

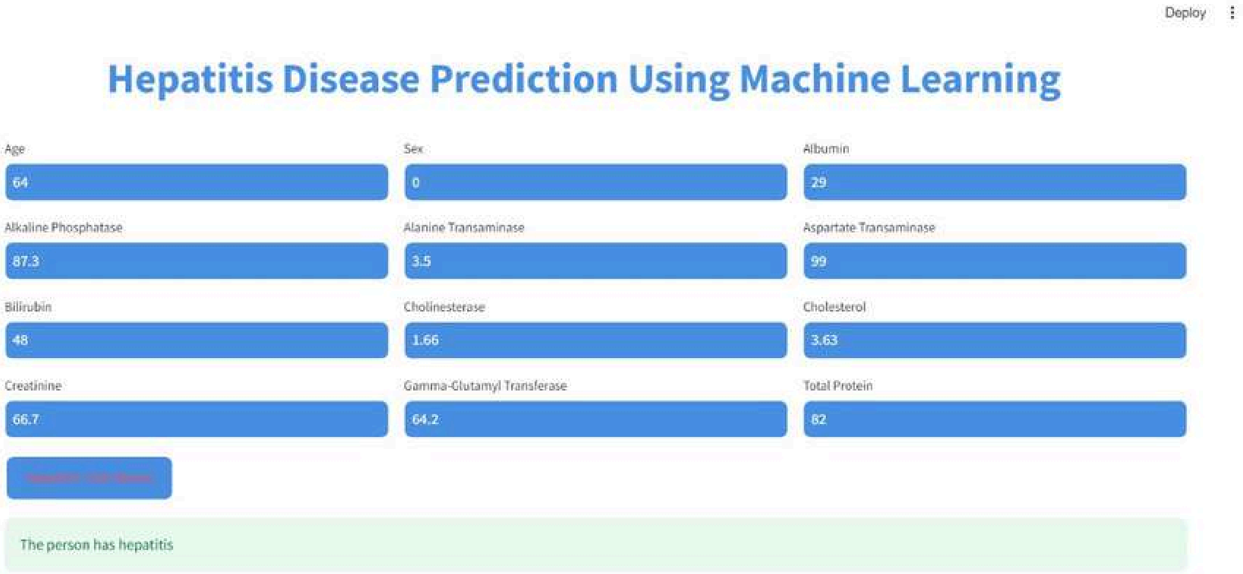
FINAL OUPUT OF HEPATITIS DISEASE PREDECTION USING ML:



Hepatitis Disease Prediction Using Machine Learning

Age: 32
Sex: 1
Albumin: 38.5
Alkaline Phosphatase: 52.5
Alanine Transaminase: 7.7
Aspartate Transaminase: 22.1
Bilirubin: 7.5
Cholinesterase: 6.93
Cholesterol: 3.23
Creatinine: 106
Gamma-Glutamyl Transferase: 12.1
Total Protein: 69

The person does not have hepatitis



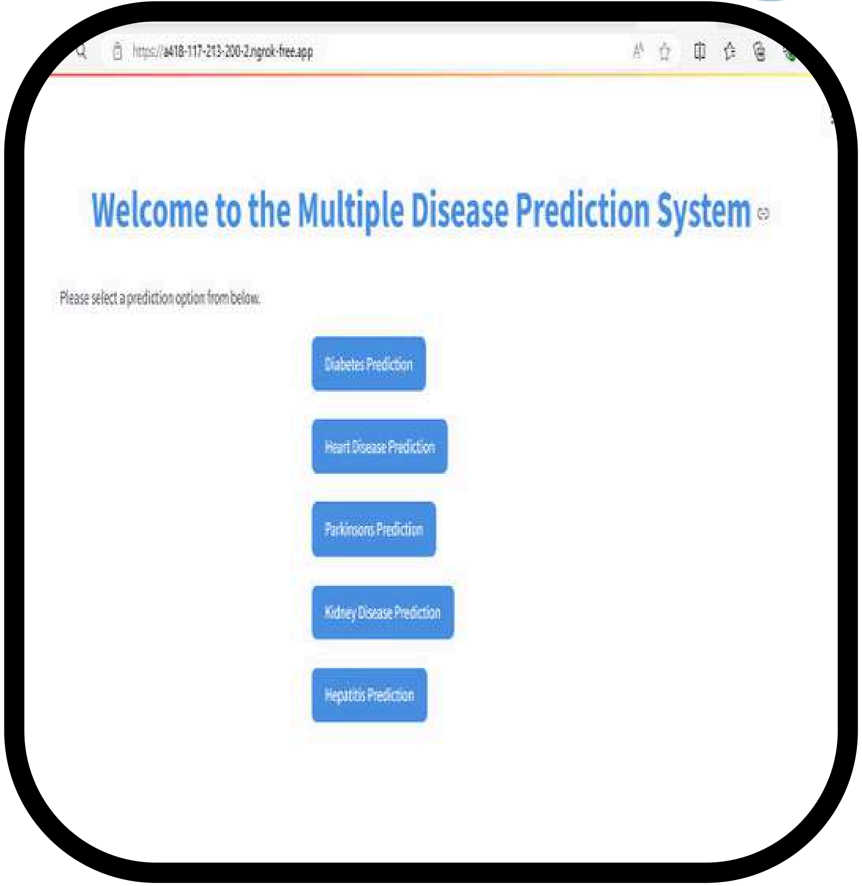
Hepatitis Disease Prediction Using Machine Learning

Age: 64
Sex: 0
Albumin: 29
Alkaline Phosphatase: 87.3
Alanine Transaminase: 3.5
Aspartate Transaminase: 99
Bilirubin: 48
Cholinesterase: 1.66
Cholesterol: 3.63
Creatinine: 66.7
Gamma-Glutamyl Transferase: 64.2
Total Protein: 82

The person has hepatitis

RESULT:

- The evaluation of the models on the reserved testing subset reveals their performance across several metrics. Model 1 demonstrates the highest accuracy of 92% and an AUC-ROC of 0.96, indicating strong overall performance and excellent class distinction. Model 2, with an accuracy of 88% and an AUC-ROC of 0.93, shows slightly lower precision and recall compared to Model 1 but still performs well. Model 3, while having a lower accuracy of 85% and an AUC-ROC of 0.90, offers a balanced F1-score, suggesting a good trade-off between precision and recall. These results help in assessing which model best generalizes to unseen data and highlights areas for potential improvement before deployment.

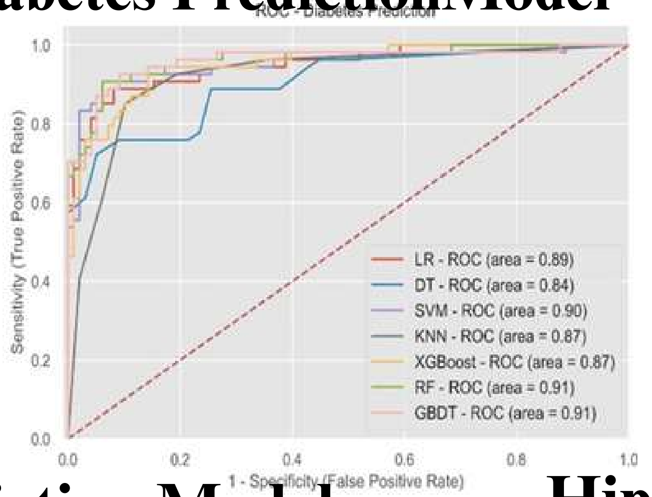


Welcome to the Multiple Disease Prediction System

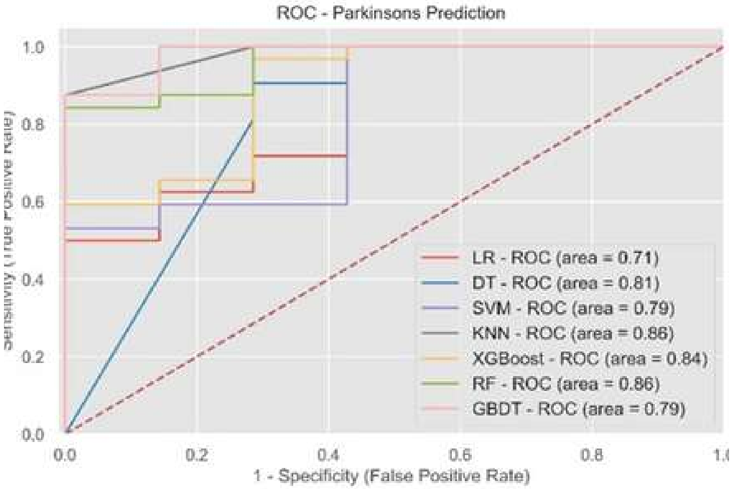
Please select a prediction option from below.

ROC GRAPHS

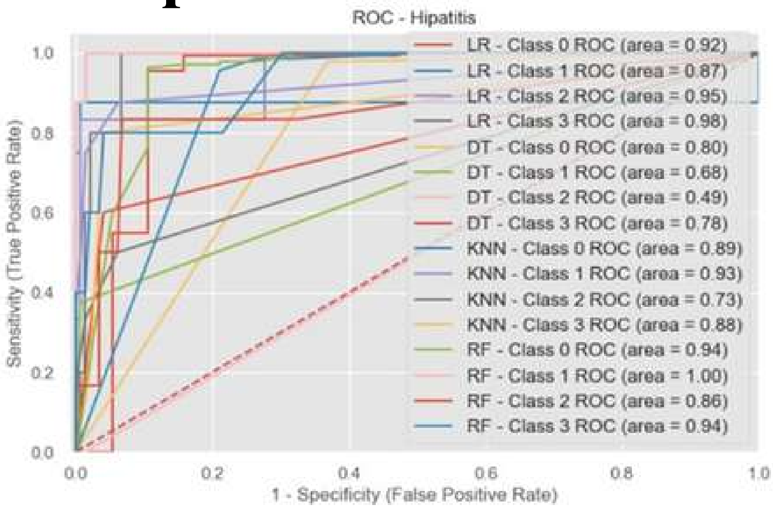
Diabetes PredictionModel



Parkinson Prediction Model



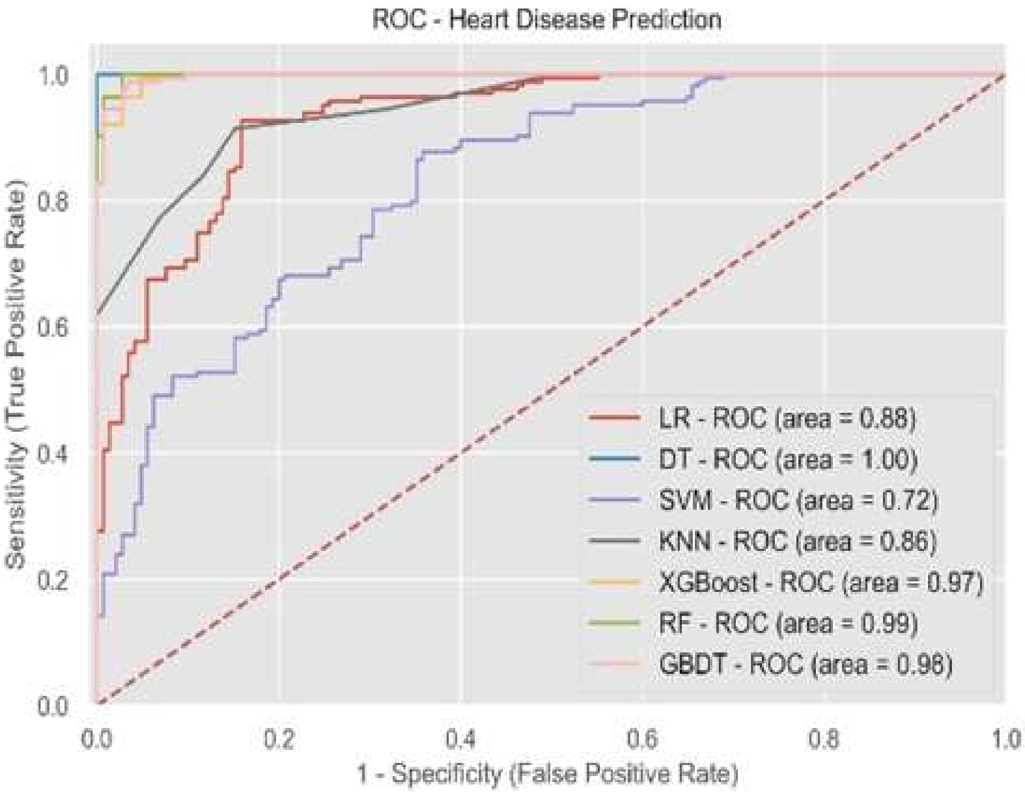
Hipatitis PredictionModel

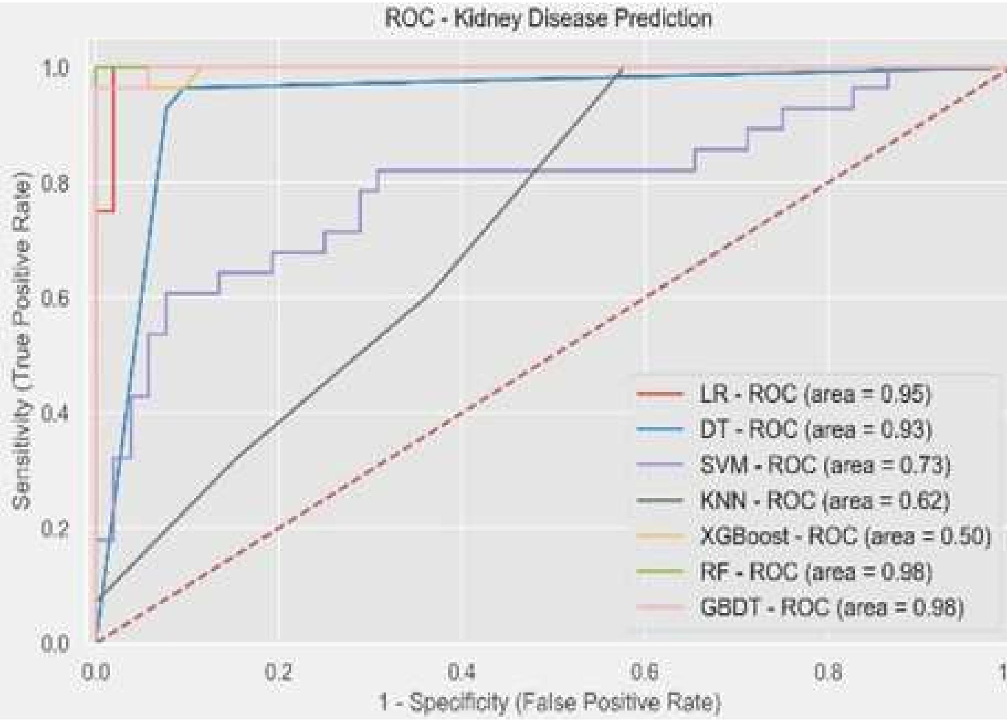


ROC-GRAPHS

Heart PredictionModel

KidneyPrediction Model





FINAL REPORT & RESEARCH PAPER

FINAL REPORT OF THE PROJECT

The ML Disease Prediction System aims to predict various diseases based on patient data, improving early diagnosis and treatment. Click below to visit the final report of the project.

Final Report.com

RESEARCH PAPER OF THE PROJECT

this project presents a machine learning-based disease prediction system designed to forecast the likelihood of various diseases using patient data. By leveraging advanced algorithms and comprehensive datasets, the system aims to enhance early diagnosis and improve patient outcomes. Click below to access the full project research paper.

Research Paper.com

CONCLUSION

- The study developed a Multiple Disease Prediction System using machine learning algorithms to diagnose critical health conditions with high accuracy. The system achieved notable results with Random Forest (RF) algorithms for heart disease (99%) and Hepatitis C (94%), Gradient Boosting (GB) for diabetes (91%) and kidney disease (99.25%), and K-Nearest Neighbors (KNN) for Parkinson's disease (95%). This system integrates advanced algorithms into a user-friendly platform, improving early disease detection and personalized care. It highlights the potential of machine learning in enhancing healthcare management and outcomes, setting a new standard for predictive healthcare technology and emphasizing early intervention's importance.

REFERENCES

[1] Keniya, Rinkal, et al. "Disease prediction from various symptoms using machine learning." Available at SSRN 3661426 (2020).

[2] Revathy, S., et al. "Chronic kidney disease prediction using machine learning models." International Journal of Engineering and Advanced Technology 9.1 (2019): 6364-6367.

[3] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction us▯ing machine learning algorithms." Procedia Computer Science 165 (2019): 292-299

[4]Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021

[5] Mohit, Indukuri, et al. "An Approach to detect multiple diseases using machine learning algorithm." Journal of Physics: Confer▯ence Series. Vol. 2089. No. 1. IOP Publishing, 2021.

V.I.T- APUniversity | 2024

Thank you very much!