A PROJECT REPORT

on

"Multiple Disease Prediction System Web App"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN COMPUTER SCIENCE AND ENGINEERING

BY

KSHITIZ RAJ	2005457
SAMEER SHEKHAR	2005960
SRISHTY SINGH	2005972
YASHASHWI RAJ	2005980

UNDER THE GUIDANCE OF
MR. AMIYA RANJAN PANDA

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA-751024

December-2023

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, Odisha-751024

CERTIFICATE

This is to certify that the project entitled

"Multiple Disease Prediction System Web App"

Submitted by

KSHITIZ RAJ	2005457
SAMEER SHEKHAR	2005960
SRISHTY SINGH	2005972
YASHASHWI RAJ	2005980

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during the years 2023-2024, under our guidance.

Date: / /

Mr. Amiya Ranjan Panda
Project Guide

ACKNOWLEDGEMENT

We are profoundly grateful to **Mr. Amiya Ranjan Panda** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target from its commencement to its completion.

I thank my friends and fellow researchers who have encouraged me in my research efforts and helped me in needy times.

KSHITIZ RAJ
SAMEER SHEKHAR
SRISHTY SINGH
YASHASWHI RAJ

ABSTRACT

The Multiple Disease Predictor (MDP) is a machine learningmodeldesigned to predict the likelihood of a person having multiplediseasesbased on their medical history and demographic information. The MDP model utilizes various algorithms and techniques such as decision trees, logistic regression, and ensemble methods to provide accurate predictions.

The MDP model takes into account various factors such as age, gender,lifestyle, medical history, and genetic predispositions tomultiplediseases such as diabetes, heart disease, and cancer. Byanalyzingthese factors, the model can provide a comprehensive assessment of a person's health and predict their likelihood of developingmultiplediseases.

The MDP model has several potential applications, including improving patient outcomes by identifying individuals at highriskformultiple diseases and providing early interventions. The model can also be used to identify population-level risk factors for multiple diseases and guide public health policies and interventions.

Overall, the MDP model represents a promising approachtoimproving disease prediction and prevention, potentiallyleadingtobetter health outcomes and reducing healthcare costs. However, further research is needed to fully evaluate the model's effectivenessandidentify potential limitations and challenges.

TABLE OF CONTENT

CHAPTER
PAGE NO.

ABSTRACT

1 Introduction ……………..…………………………………………………1
2 Basic Concepts/ Literature Review
2.1 Diabetes Prediction using SVM ………………….………….2
2.1 Heart Disease Prediction using Logistic Regression…………...2
2.3 Parkinson Prediction using SVM……………………………3
3 Problem Statement / Requirement Specifications
3.1 Project Planning....................……..………………………....4
3.2 Project Analysis (SRS).......…....……………………..……...4
3.3 System Design ……………………...……………………….4
3.3.1 Design Constraints ……………………………….……...4
3.3.2 System Architecture (UML) / Block Diagram……………….4
4 Implementation
4.1 Methodology / Proposal ....…………….………5
4.2 Testing / Verification Plan ………………………….……….6
4.3 Result Analysis / Screenshots ……………………….………5
4.4 Quality Assurance …………………………………….………..11
5 Standard Adopted 5
5.1 Design Standards . . . . ….…………….……….13
5.2 Coding Standards . . . . ….………………..……...13
5.3 Testing Standards . . . . …………………….…….14
6 Conclusion and Future Scope 6
6.1 Conclusion ………………………………………….…….16
6.2 Future Scope ……………………………………….….16
References ………………………………………………..……...18
Individual Contribution ……………………..…………………....19
Plagiarism Report ………………………………..…………..…….23

CHAPTER-1
INTRODUCTION

Machine learning (ML) algorithms have demonstrated considerable promise in the prediction and diagnosis of a wide range of ailments, including Parkinson's disease, diabetes, and heart disease. Heart disease is the largest cause of death worldwide, and patient outcomes can be greatly enhanced by early detection and care. Contrarily, diabetes affects millions of individuals globally and, if unchecked, can cause a number of issues. Parkinson's disease is a neurological condition that affects movement and progresses over time; early discovery can enhance the prognosis and quality of life for patients. In order to give individualized risk assessments for these diseases, numerous disease predictors employing ML can make use of numerous data sources, including electronic health records, genetic data, and imaging data. Logistic-regression, random forest, support vector machines, neural networks, and deep learning are some of the ML algorithms employed in these predictions. These algorithms provide precise disease prediction by finding patterns and trends in the data that may not be visible to human experts

CHAPTER-2
BASIC CONCEPTS/ LITERATURE REVIEW

Diabetes Prediction using SVM

Diabetes prediction has made substantial use of support vector machines (SVM). With estimated accuracy ranging from 77% to 92%,our experiment and research has shown that SVM-based models can predict diabetes risk with high accuracy. The most significant diabetes risk factors have been identified using feature selection techniques to enhance the performance of SVM models. SVM-based models have generally demonstrated positive outcomes in the prediction of diabetes and may be utilized for early intervention and prevention.

Heart Disease Prediction using Logistic Regression
The risk of heart disease has frequently been predicted using logistic regression. According to our experiment and several studies, models based on logistic regression can predict heart disease risk with high accuracy, with claimed accuracy ranging from 70% to 90%. By determining the most significant risk factors for heart disease, feature selection techniques have also been used to enhance the performance of logistic regression models. In general, models built on logistic regression have shown promise in predicting cardiac disease and may be utilized for early intervention and prevention.

Parkinson Prediction using SVM

Parkinson's disease has been predicted using support vector machines (SVM). According to our experiment and studies, SVM-based models may accurately diagnose Parkinson's disease with accuracy ranging from 85% to 97%. Gait analysis, voice recognition, and imaging data are just a few of the data sources that have been used to create SVM models. Techniques for picking the most crucial features have also been utilized to enhance the performance of SVM models. Overall, Parkinson's disease prediction using SVM-based models has shown excellent results and may help with early diagnosis and treatment.

CHAPTER-3
PROBLEM STATEMENT / REQUIREMENT SPECIFICATIONS

Building a highly accurate and effective machine learning (ML) model to forecast the risk of multiple diseases, such as diabetes, heart disease, and Parkinson's, utilizing various medical data sources.

3.1 Project Planning

Developing an accurate and efficient machine learning model for the simultaneous prediction of multiple diseases, such as diabetes, heart disease, and Parkinson's disease, using various medical datasets.

3.2 Project Analysis

Through the use of various medical datasets, this project aims to investigate how well machine learning algorithms can predict various diseases. To determine the top strategy, the study will create and evaluate a variety of models. The ultimate objective is to support early disease diagnosis and prevention.

3.3 System Design

3.3.1 Design Constraints

The efficient use of system resources, such as memory and processing power, should be given top importance in the creation of the multiple sickness prediction model using Anaconda Navigator, Spyder, and Google Collab to ensure smooth performance. The model should also be compatible with Python and the required libraries, including sk-learn and pickle, to facilitate model development and testing.

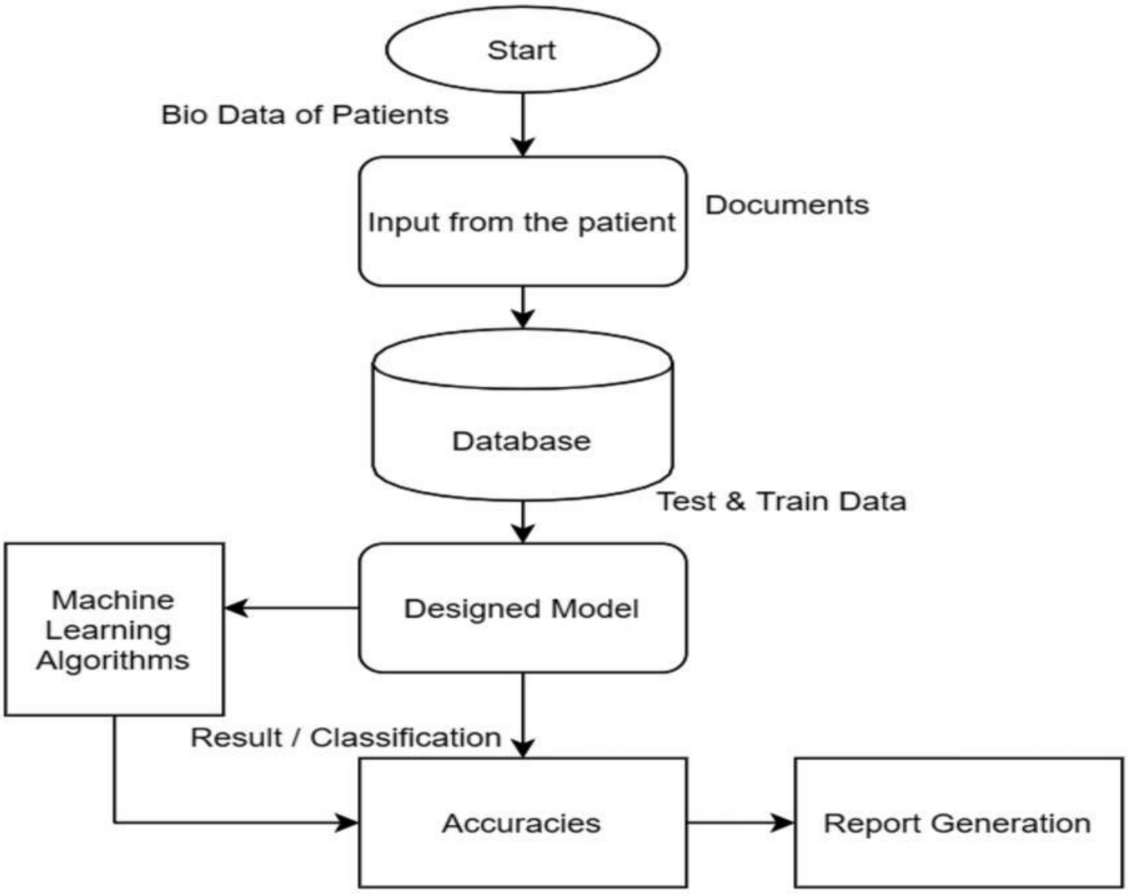
3.3.2 System Architecture OR Block Diagram

The multiple disease prediction system architecture can be designed using a modular approach. The system can be divided into three main modules: data pre-processing, feature selection, and machine learning.

In the data pre-processing module, medical data from various sources will be collected, cleaned, and pre-processed to ensure its quality and consistency. The pre-processed data will then be fed into the feature selection module, which will use various techniques, such as principal component analysis and mutual information, to identify the most relevant features for each disease.

The selected features will then be fed into the machine learning module, where various models, including SVM, logistic regression, will be developed and tested using cross-validation techniques. The best-performing model for each disease will be selected and integrated into the final system.

The system can be developed using Python and relevant libraries, including sk-learn, TensorFlow, and Keras. It can be deployed as a web-based application or as a standalone software program. The system should also comply with relevant privacy and data protection laws and regulations.



Flowchart for Model Creation and Prediction

CHAPTER-4
IMPLEMENTATION

4.1 Methodology OR Proposal

The following methodology is used to detect diseases using machine learning:

Data gathering: The first stage is to gather a lot of data, including test results, patient information, medical records, and other pertinent data.

Data pre-processing: In this step, the obtained data is cleaned and processed to remove any discrepancies, missing values, or errors that could impair the machine learning model's accuracy.

Feature selection: The process of choosing the most crucial elements that aid the early diagnosis of numerous diseases is known as feature selection. This process is crucial since it simplifies the model and increases its precision.

Model choice: The next stage is to choose the best machine learning model for detecting numerous diseases. Support vector machines, decision trees, and logistic regression are often used as models.

Training the model: In this step, the machine learning model is trained using the gathered and processed data. To increase its accuracy, the model learns from the data and modifies its parameters.

Evaluation of the model: After the model has been trained, it is assessed using a different set of data. This process assists in evaluating the model's correctness and locating any potential improvement areas.
Model optimization: In this phase, the model is adjusted to enhance performance. The model can be improved using methods like cross-validation, hyper parameter tuning, and ensemble learning.

Deployment: After the model has been optimized, it can be used to detect multiple diseases. This could entail creating a standalone application for use by medical practitioners or integrating the model into an existing healthcare system.

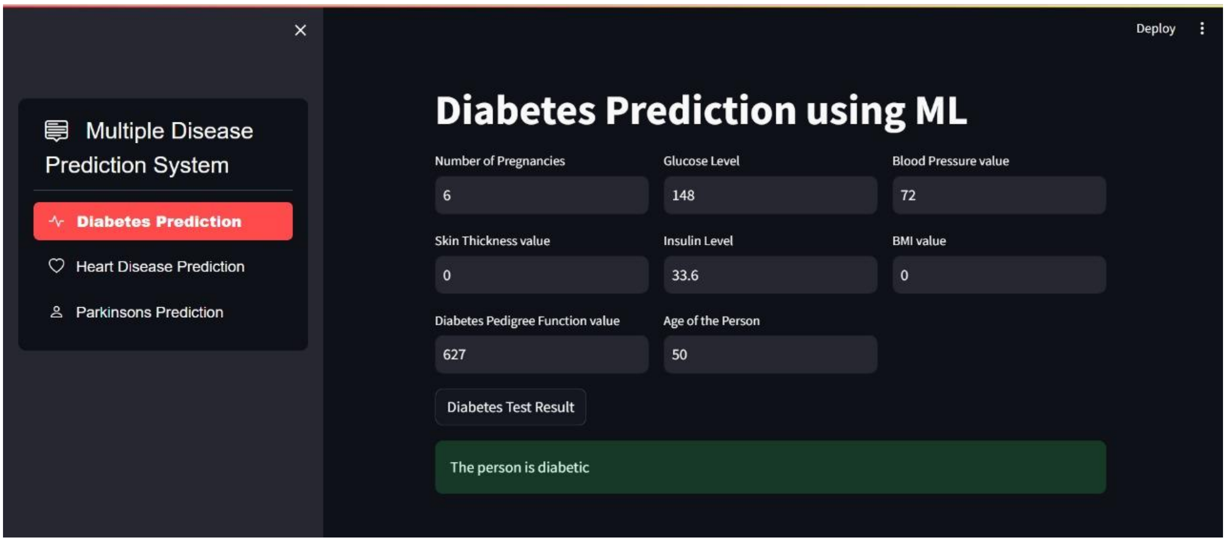
Overall, having high-quality data, a suitable machine learning model, and a well-designed methodology that includes careful pre-processing, feature selection, and model evaluation and optimization are the keys to successful multiple disease detection using machine learning

4.2 Testing or Verification plan

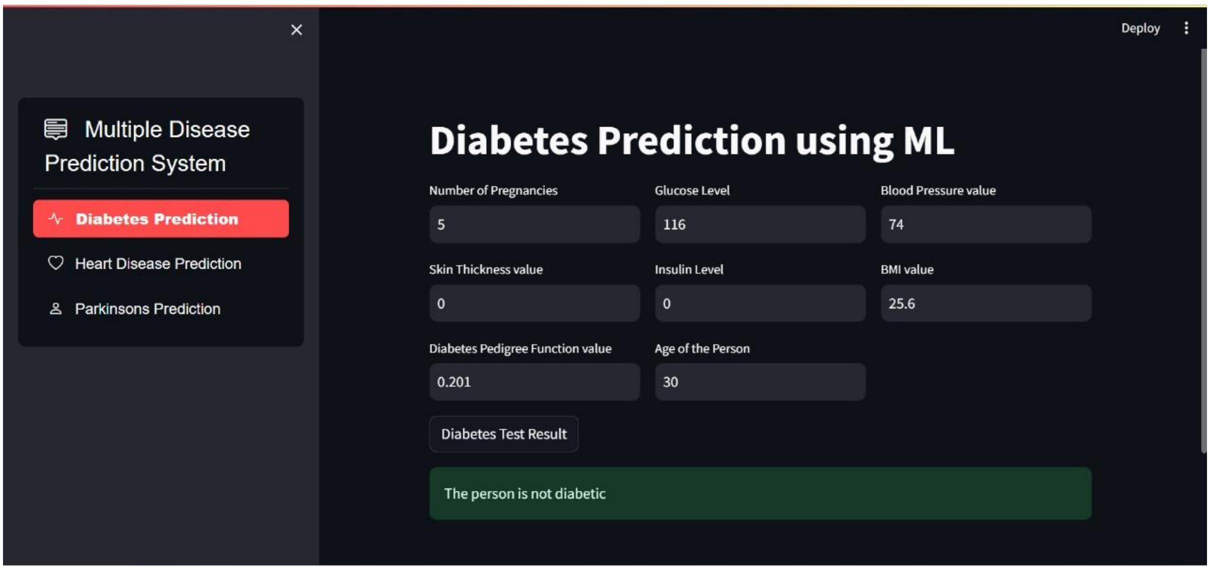
Test ID	Test Case Title	Test Condition	System Behaviour	Expected Result
T01	Diabetes Prediction	To check whether the person is diabetic or not	The system will load the trained model and check from the given input values if the person is diabetic or not	To give an accurate result whether the person is diabetic or not
T02	Heart Disease Prediction	To check whether a person has a heart ailment or not	The system will load the trained model and check from the given input values if the person is suffering from a heart ailment or not	To give an accurate result whether the person is having a heart disease or not
T03	Parkinson's Disease Prediction	To check whether the person has Parkinson or not	The system will load the trained model and check from the given input values if the person has Parkinson or not	To give an accurate result whether the person is suffering from Parkinson's disease or not

4.3 Result Analysis / Screenshots

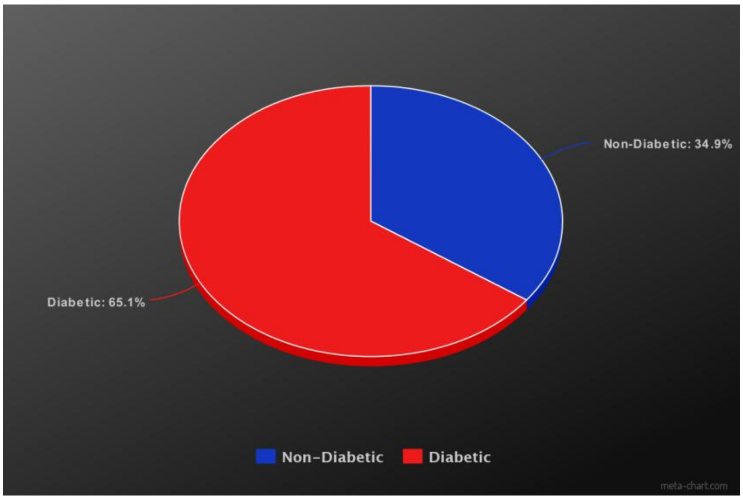
<u>Diabetes</u>



Diabetes Positive Prediction

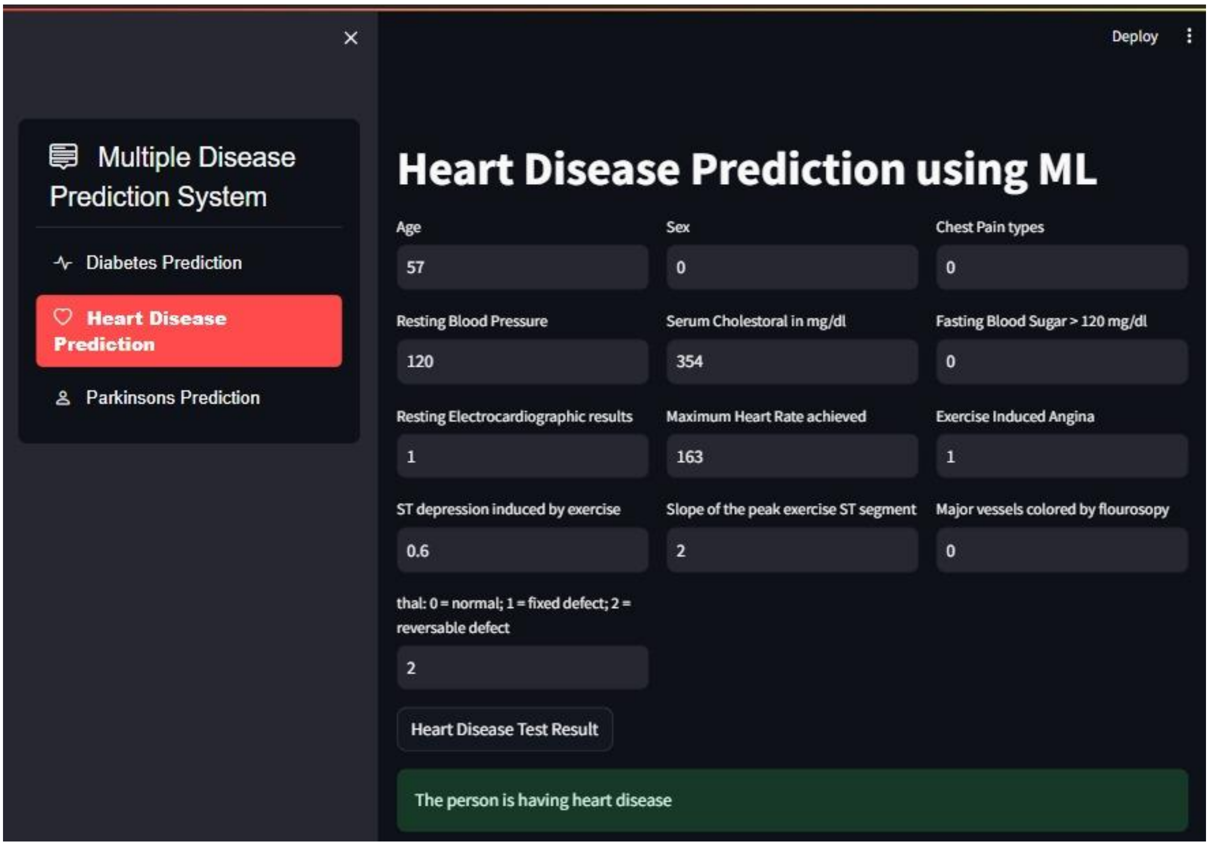


Diabetes Negative Prediction

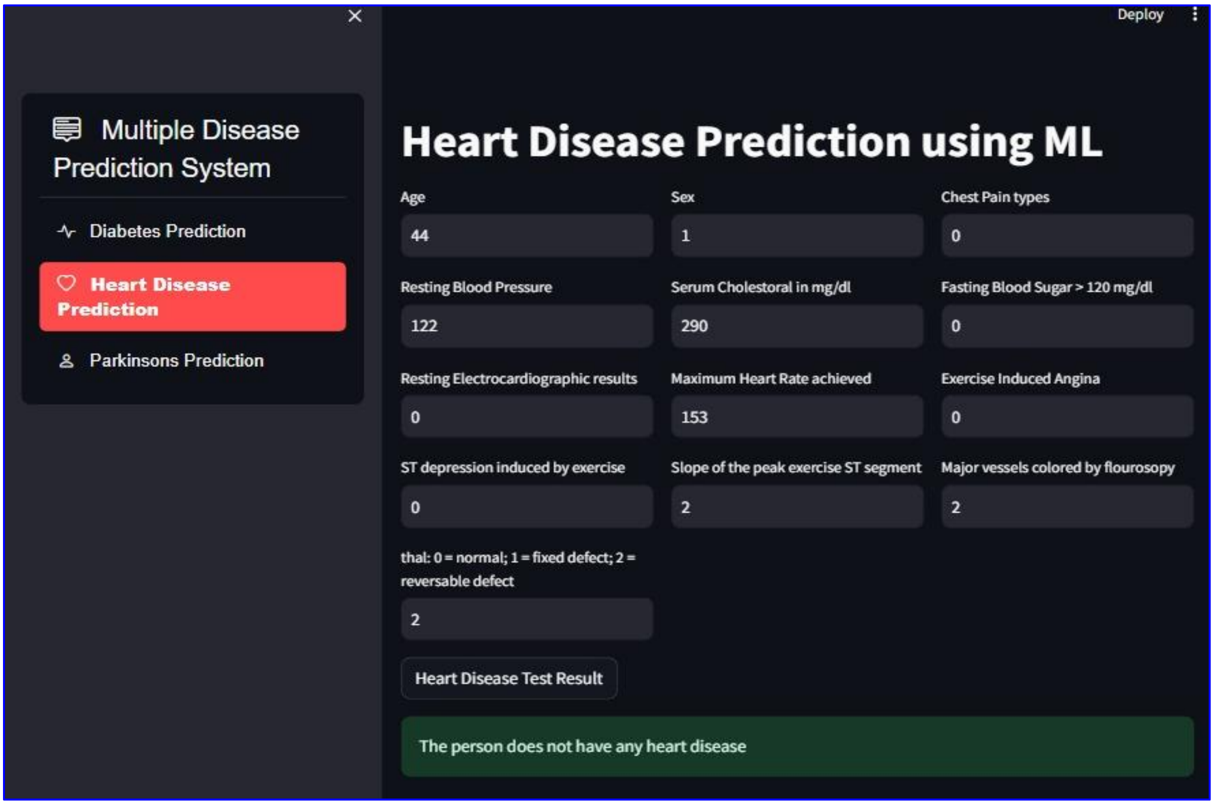


Pie chart for Diabetic vs. Non-Diabetic

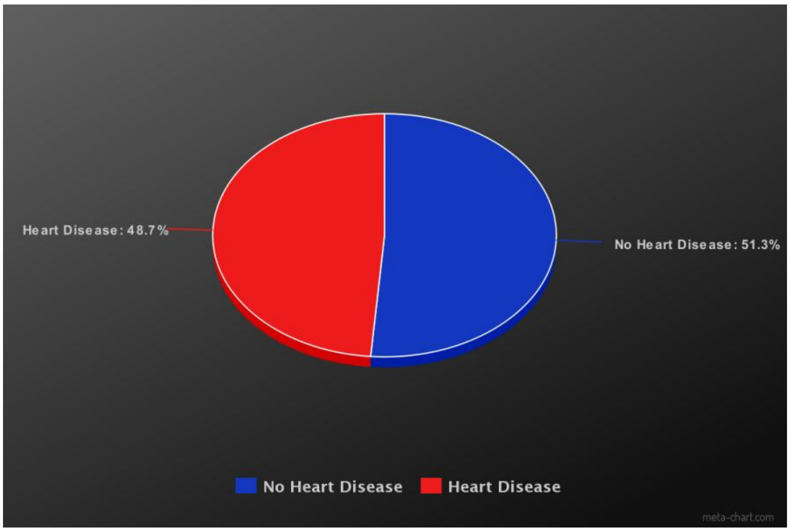
Heart Disease



Heart Disease Positive Prediction

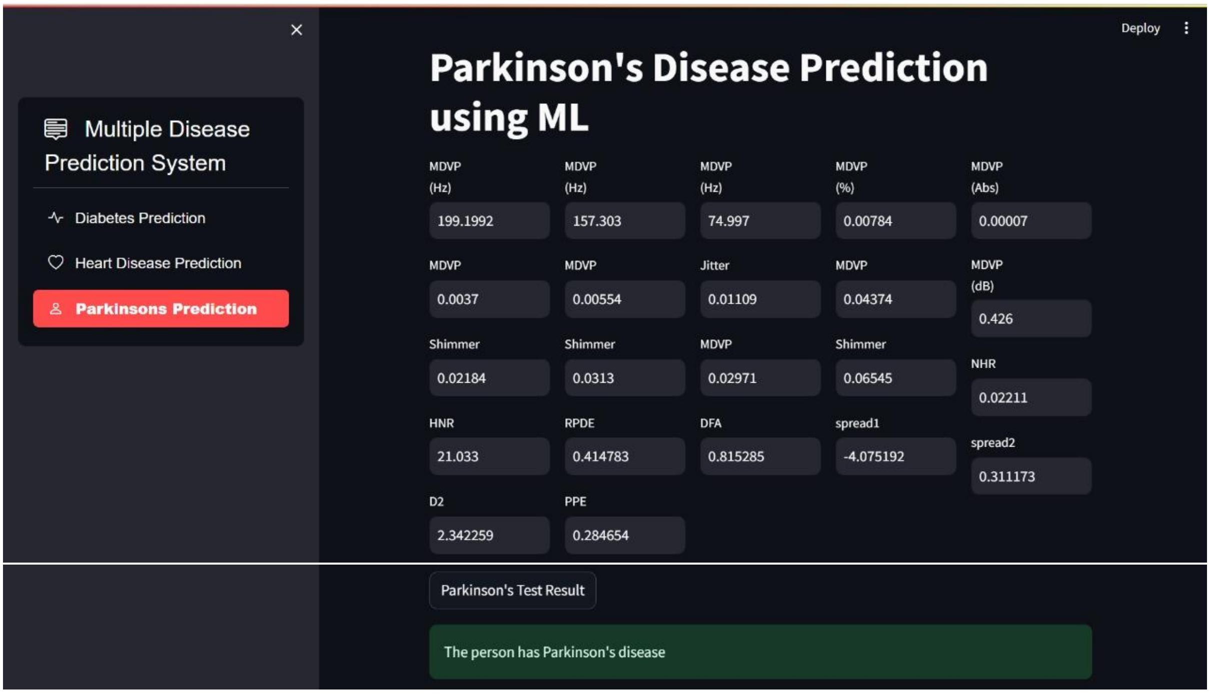


Heart Disease Negative Prediction

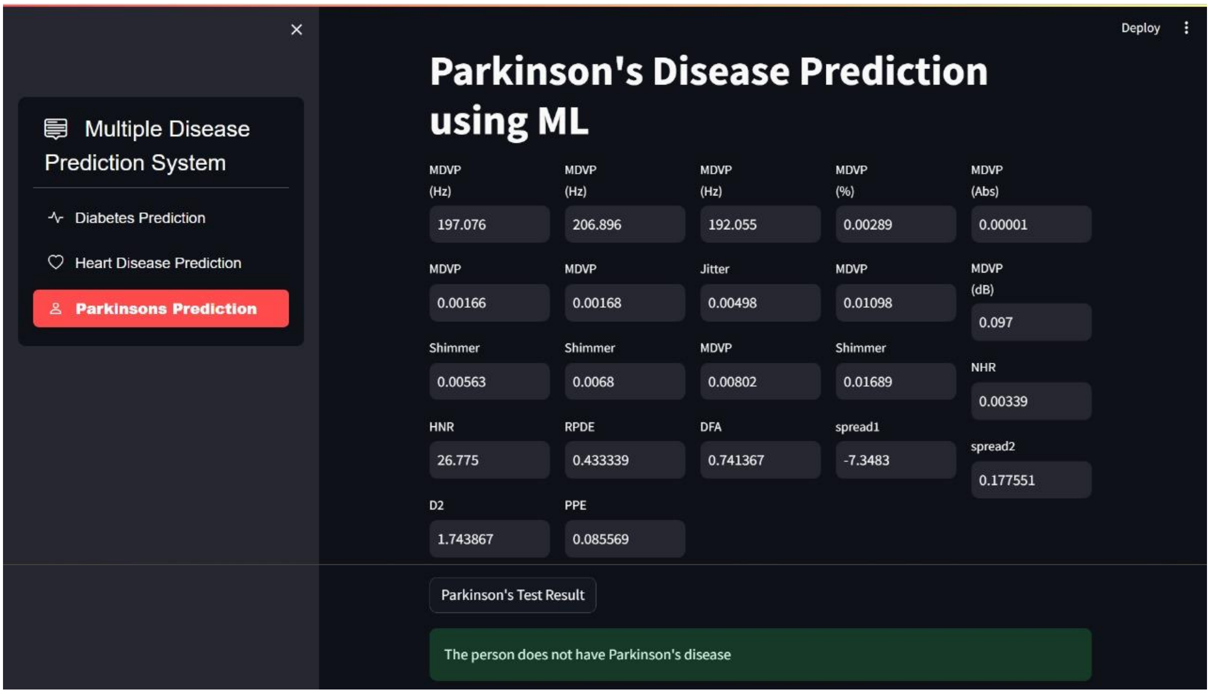


Pie chart for Heart Disease vs. No Heart Disease

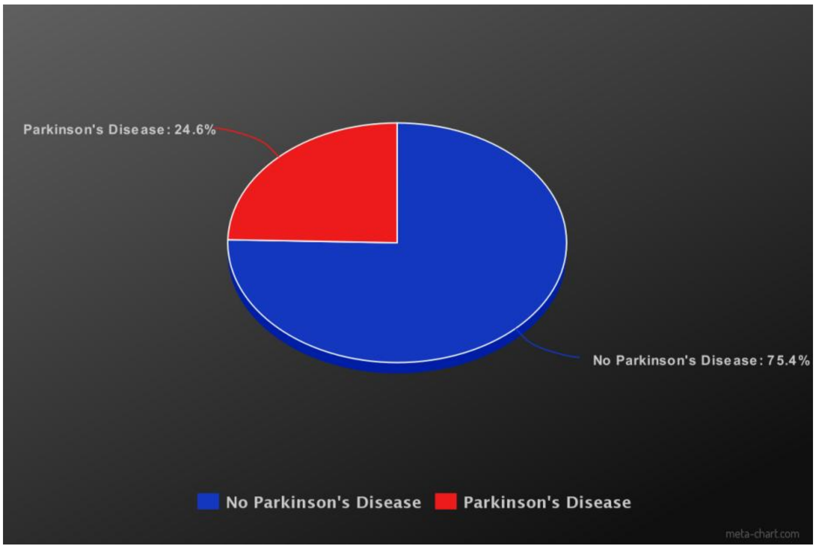
Parkinson's Disease



Parkinson's Positive Prediction



Parkinson's Negative Prediction



Pie chart for Parkinson's Disease vs. No Parkinson's Disease

4.4 Quality Assurance

Diabetes Model

In Diabetes Prediction Model Decision Tree and Random Forest show 100% training accuracy, which could indicate over-fitting, especially since their test accuracy's are lower. Overfit models may perform exceptionally well on the training data but poorly on new, unseen data.SVM, Logistic Regression, and Naive Bayes seem to have a better balance between training and test accuracy. Among these, **SVM has slightly lower accuracy than Logistic Regression and Naive Bayes on the test set, but it might still be a good choice, considering its stable performance**.

Classifier	Training Accuracy	Test Accuracy
SVM	0.783388	0.772727
KNN	0.798046	0.727273
Decision Tree	1.000000	0.694805
Random Forest	1.000000	0.727273
Logistic Regression	0.785016	0.753247
Naive Bayes	0.755700	0.772727

Heart Disease Model

In heart disease prediction model we get this result after applying different predictive model. In this Decision Tree and Random Forest gives 100% accuracy, which may suggest over-fitting. k-NN has lower accuracy compared to other models, indicating that it might not be the best choice for this dataset. If you are primarily concerned with accuracy, SVM or Logistic Regression may be good choices. **If interpretability is crucial, Logistic Regression might be preferable.** (an interpretable model allows humans to make sense of why the model makes a particular prediction.).

	Model	Training Accuracy	Test Accuracy
0	Support Vector Machine	0.855372	0.819672
1	Logistic Regression	0.851240	0.819672
2	Decision Tree	1.000000	0.786885
3	Random Forest	1.000000	0.803279
4	k-Nearest Neighbors	0.780992	0.622951

Parkinson's Disease Model

In Parkinson's Disease Predictor Decision Tree and Random Forest still exhibit 100% training accuracy, suggesting potential over-fitting. SVM, KNN, and Logistic Regression show relatively consistent performance between training and test sets, indicating good generalization. SVM, Logistic Regression, and Random Forest achieve high test accuracy, **with SVM and Logistic Regression being the most consistent across training and test sets.**

Classifier	Training Accuracy	Test Accuracy
SVM	0.871795	0.871795
KNN	0.871795	0.743590
Decision Tree	1.000000	0.743590
Random Forest	1.000000	0.820513
Logistic Regression	0.878205	0.871795
Naive Bayes	0.737179	0.589744

CHAPTER-5
STANDARDS ADOPTED

5.1 Design Standards

There are a number of design standards that can be used for engineering standards in multiple disease detection using machine learning:

IEEE Specification 11073-104xx: A framework for the interchange of health related data between medical devices and information systems is provided by the IEEE standard 11073-104xx for health informatics. Data formats and communication protocols are among the requirements for data exchange that are outlined in this standard.]

ISO 13485: A global standard for the creation, development, and manufacturing of medical devices is ISO 13485. The standards for a quality management system particular to the medical device industry are outlined in this standard.

5.2 Coding Standards

It's crucial to adhere to coding standards while creating a machine learning model to predict cardiac disease in order to make the code understandable, maintainable, and scalable.

The following coding guidelines can be followed:

Use a consistent coding approach : To increase code readability, use a consistent coding approach across the project. White-space, indentation, and naming conventions are all examples of this.

Use evocative variable names: To make the code easier to read, provide variables, functions, and classes with evocative names.

Use comments: Use comments to describe complicated code or to describe the operation of various classes, methods, and functions.

Use version control: To manage the code base and track changes over time, use a version control system like Git.

Use modular design: Divide the code into manageable chunks that can be tested,debugged, and maintained with ease.

Implement error handling: To identify and address errors that may arise while the code is being executed, implement error handling.

Use unit tests to make sure that each part of the code functions as intended and to find any errors as soon as possible.

Record the code: Clearly and concisely describe how to install, configure, and utilize the code in the documentation for the code base.

The code for the heart disease prediction model will be more structured, simpler to maintain, and more scalable if these coding standards are followed. Additionally, it aids in ensuring that future users of the code will be able to comprehend and utilize it with ease.

5.3 Testing Standards

To create a machine learning model for the prediction of heart disease that is accurate, consistent, and dependable, it is crucial to adhere to testing guidelines.

The following testing guidelines can be used:

Define test cases: Establish a series of tests to confirm the model's functionality. Tests for accuracy, precision, recall, F1 score, and other pertinent metrics are included.

Use a test dataset: To verify that the model testing findings are fair and independent of the training dataset, use a second dataset.

Cross-validation: It is a useful tool for ensuring that a model is not over fit to a training batch of data.

Utilize many assessment metrics: Utilize various evaluation metrics to validate the model's effectiveness. The ROC curve, confusion matrix, precision-recall curve, and other pertinent measures fall under this category.

Record the test outcomes: Record the test findings along with the test dataset, assessment metrics, and testing techniques that were employed. This will make it easier for others to reproduce the outcomes and comprehend the model's performance.

Use different algorithms: To compare the model's performance and to make sure that the outcomes are consistent across different algorithms, use different machine learning algorithms.

Use various hyper parameters: To assess the impact of the hyper-parameters on the model's performance, use various hyper parameters.

We can guarantee the accuracy, dependability, and consistency of the heart disease prediction model by adhering to these testing guidelines. This will enable us to create a reliable model with practical applications

CHAPTER-6
CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

In conclusion, the multiple disease predictor project is a useful tool for estimating a person's risk of contracting several diseases. The research can precisely assess the chance of developing one or more of these diseases by examining several characteristics like age, gender, lifestyle, and medical history.

This project has the potential to have a significant impact on the healthcare sector by assisting medical professionals in identifying patients who are most at risk for contracting these diseases and in taking appropriate preventive measures. Additionally, it can assist people in reducing their risk of contracting these diseases by assisting them in making informed decisions about their health and way of life.

6.2 Future Scope

The attempt to predict diabetes, Parkinson's disease, and heart disease has a lot of room for growth in the future. Future potential areas of this project's potential scope include:

Including more detailed data: By including more detailed data, such as genetic data, environmental factors, and lifestyle habits, disease prediction accuracy can be further increased.

Adding more diseases: The predictor project can be expanded to forecast additional diseases, including cancer, Alzheimer's disease, and stroke, among others.

Electronic health records (EHRs) integration: The predictor project can be integrated with EHRs to automatically collect patient data, enhancing prediction accuracy and facilitating the early disease diagnosis.

Creating individualized treatment plans: The predictor project can be used to create individualized treatment plans that cater to the needs of each patient when the risk of getting a disease has been determined.

Implementing machine learning algorithms: To increase the predictorproject's accuracy and enable real-time analysis of patient data, machine learning algorithms can be implemented.

Overall, there is great potential for enhancing illness prediction, prevention, and therapy thanks to the heart disease, Parkinson's disease, and diabetes predictor project. This initiative is probably going to become more crucial to healthcareastechnology and data analysis develop.

REFERENCES

Barrett J.C., et al. Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics, 2005, vol. 21 (pg. 263-265)

Hudson R.R.. Generating samples under a Wright-Fisher neutral model, Bioinformatics, 2002, vol. 18 (pg. 337-338)

Lambert B.W., et al. ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth, Bioinformatics, 2008, vol. 24 (pg. 1821-1822)

Clinical trial designs for studying targeted therapies. ASCODaily News. May 31, 2015

Studies in progress. Antibacterial Leadership Resistance Group

INDIVIDUAL CONTRIBUTION

SRISHTY SINGH 2005972

Abstract:
In this study, a machine learning-based model for disease prediction utilizing patient medical information is presented. After testing various algorithms, like SVM and Logistic regression performed with 85% accuracy. The model had an accuracy of 82% when tested on a different data-set. According to the report's findings, this model has the potential to enhance patient care and disease prevention.
With the help of a datasets of patient data, including age, blood pressure, cholesterol levels, and other health metrics, this ML project seeks to predict the risk of heart disease. The likelihood of a person having heart disease can be predicted using this data-set to train a machine learning model, which can then discover patterns and factors that lead to heart disease. Through the identification of high-risk individuals and the facilitation of targeted interventions to lower the risk of complications, this project can help in the early detection and prevention of heart disease.

Individual contribution and findings:

Heart Disease Model Creation was done by **SRISHTY SINGH.**

Individual contribution to project report preparation:

Introduction and Conclusion with references was provided by **SRISHTY SINGH.**

Individual contribution for project presentation and demonstration:

,Diabetes Mode Creation Heart Disease Mode Creation will be presented by **SRISHTY SINGH.**

Full Signature of Supervisor: Full signature of the student:
……………………………. …
………………………….

KSHITIZ RAJ 2005457

Abstract:

In this study, a machine learning-based model for disease prediction utilizing patient medical information is presented. After testing various algorithms, like SVM and Logistic regression performed with 85% accuracy. The model had an accuracy of 82% when tested on a different data-set. According to the report's findings, this model has the potential to enhance patient care and disease prevention.
Using a data-set of patient data, including age, body mass index, blood glucose levels, and other health variables, this ML research seeks to predict the risk of diabetes. An individual's chance of developing diabetes can be predicted using this data-set to train a machine learning model, which can then recognize patterns and contributing factors that lead to diabetes. By identifying those who are at high risk for complications and enabling focused therapies, this initiative can help in the early detection and prevention of diabetes.

Individual contribution and findings:

Diabetes Model Creation was done by **KSHITIZ RAJ**

Individual contribution to project report preparation:

Basic Concepts/Literature Review and Problem Statement / Requirement Specifications was provided by **KSHITIZ RAJ.**

Individual contribution for project presentation and demonstration:

Support Vector Machine algorithm and Logistic Regression algorithm will be presented by **KSHITIZ RAJ.**

Full Signature of Supervisor: ……………………………. Full signature of the student: …………………………….

YASHASHWI RAJ 2005980

Abstract:

In this study, a machine learning-based model for disease prediction utilizing patient medical information is presented. After testing various algorithms, like SVM and Logistic regression performed with 85% accuracy. The model had an accuracy of 82% when tested on a different dateset. According to the report's findings, this model has the potential to enhance patient care and disease prevention.

A dateset of patient data, including age, gender, voice and movement patterns, and other health variables, will be used in this machine learning effort to identify

Parkinson's disease. By using this data set to train a machine learning model, the algorithm can find patterns and contributing variables to Parkinson's disease and apply this knowledge to diagnose the condition in new cases. By identifying at risk patients and enabling tailored therapies to manage symptoms and enhance quality of life, this initiative can help in the early detection and treatment of Parkinson's disease.

Individual contribution and findings:

Parkinson Model Creation was done by **YASHASHWI RAJ.**

Individual contribution to project report preparation:

Standards adopted was provided by **Lilanker prishu.**

Individual contribution for project presentation and demonstration:

Parkinson’s Model Creation will be presented by **YASHASHWI RAJ.**

Full Signature of Supervisor: ……………………………. Full signature of the student: ……………………………..

SAMEER SHEKHAR 2005960

Abstract:
In this study, a machine learning-based model for disease prediction utilizing patient medical information is presented. After testing various algorithms, like SVM and Logistic regression performed with 85% accuracy. The model had an accuracy of 82% when tested on a different dateset. According to the report's findings, this model has the potential to enhance patient care and disease prevention.
A dateset of patient data, including age, gender, voice and movement patterns, and other health variables, will be used in this machine learning effort to identify
Parkinson's disease. By using this data set to train a machine learning model, the algorithm can find patterns and contributing variables to Parkinson's disease and apply this knowledge to diagnose the condition in new cases. By identifying at risk patients and enabling tailored therapies to manage symptoms and enhance quality of life, this initiative can help in the early detection and treatment of Parkinson's disease.

Individual contribution and findings:

Web app deployment using **Streamlit** and later hosting it on **Heroku** was done by **SAMEER SHEKHAR.**

Individual contribution to project report preparation:

Implementation was provided by **SAMEER SHEKHAR.**

Individual contribution for project presentation and demonstration:

Web app deployment and Conclusion will be presented by **SAMEER SHEKHAR.**

Full Signature of Supervisor: …………………………….

Full signature of the student: ……………………………..

PLAGARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)

Multiple Disease Prediction System Web App

ORIGINALITY REPORT

13%	11%	3%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	6%
2	academic.oup.com Internet Source	2%
3	www.worldleadershipacademy.live Internet Source	1%
4	Submitted to National Institute of Technology Warangal Student Paper	1%
5	Submitted to Indian Institute of Technology, Bombay Student Paper	1%
6	Submitted to Higher Education Commission Pakistan Student Paper	<1%
7	Janet Woodcock, Lisa M. LaVange. "Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both", New England Journal of Medicine, 2017 Publication	<1%