Prediction of Disease Outbreaks

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Name of Student, Email id

Kotapuri Supraja-kotapurisupraja@gmail.com

Under the Guidance of

Name of Guide

Edunet Trainer

ACKNOWLEDGEMENT

We extend our sincere gratitude to the **All India Council for Technical Education (AICTE)** for providing us with the opportunity to work on this project, *Prediction of Disease Outbreaks Using AI & Data Analytics*. Their support and encouragement have been invaluable in the successful completion of this research.

We would like to express our deepest appreciation to our **mentors, faculty, and project guides**, whose guidance and insightful feedback have significantly contributed to our understanding and implementation of predictive modeling in healthcare. Their expertise and continuous support played a crucial role in shaping the project's direction.

Furthermore, we are grateful to our **team members** for their dedication, collaboration, and hard work throughout this journey. The collective efforts in data collection, model development, and analysis have made this project a success.

We also acknowledge the contributions of **various open-source communities, research papers, and datasets**, which provided us with valuable resources and references for building an efficient disease outbreak prediction model.

Lastly, we extend our heartfelt thanks to our **family and friends** for their constant encouragement and motivation, which kept us inspired throughout the project.

Thank you!

ABSTRACT

The growing incidence of chronic diseases like diabetes and Parkinson's requires effective and affordable diagnostic tools. This project focuses on creating a web-based multi-disease prediction system that uses machine learning models to forecast diseases from user inputs. The system is developed using Python, Streamlit, and the Pickle library, with an easy-to-use interface for users to input relevant medical information and receive predictive outcomes.

The aim of this project is to design an easy-to-use, AI-based web application that enables people to evaluate their risk of developing diabetes, Parkinson's, and other diseases. Through the incorporation of several trained machine learning models, the application gives fast and accurate predictions, enhancing early diagnosis and awareness.

The approach includes training various disease prediction models on suitable datasets, storing the models using the Pickle library, and implementing them on a Streamlit-based web application. Users can switch between various prediction choices through a sidebar, input necessary health parameters, and get immediate predictions. The application is developed with an interactive user interface, such as input fields, icons, and organized layouts for better user experience.

Key outcomes illustrate that the deployed models accurately classify disease conditions with acceptable precision from user inputs. The web application successfully loads and runs several trained models, providing smooth disease prediction functionality. Furthermore, structuring the interface with Streamlit columns and sidebars enhances usability and interaction.

In summary, the project offers a scalable and accessible disease risk assessment solution, showing how machine learning can be successfully implemented in healthcare settings. Potential future enhancements involve widening the scope of diseases addressed, further refining model accuracy, and incorporating real-time sources of data for improved prediction. The application can be published publicly through GitHub to facilitate greater accessibility.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1Problem Statement:

Chronic illnesses like diabetes and Parkinson's are a major threat to world health, affecting millions of individuals and causing serious complications if left undiagnosed. The prevalence of these diseases is still on the rise, escalating healthcare expenditure and mortality rates globally. Early diagnosis is important in the management of these conditions, as prompt medical treatment can greatly enhance patient outcomes. Nevertheless, the conventional diagnostic practices tend to be costly laboratory tests, highly specialized medical professionals, and a lot of time, which make them out of reach for many, particularly those in rural or disadvantaged regions. Consequently, many cases go undiagnosed until severe symptoms develop, decreasing the likelihood of successful treatment and disease control. This project seeks to solve the shortage of affordable, accessible, and quick disease prediction tools by creating a machine learning web application that enables users to determine their risk of diabetes, Parkinson's, and other diseases.

Using predictive models trained on medical data, the system can give immediate and data-driven feedback based on user input. The web application does away with the requirement for expensive diagnostic tests and allows patients to evaluate their health conditions in the comfort of their homes. The importance of this issue stems from the growing cases of chronic illnesses and the pressing need for early detection and prevention. Most individuals, particularly in regions with limited resources, do not have access to healthcare centers, and thus, early diagnosis is a challenge.

The system closes this gap by providing an easy-to-use, AI-based solution with instant health diagnoses based on health data. Through empowering people with rapid and trustworthy predictions, the app promotes active health care and takes the pressure off healthcare systems. In the long term, these technology-based solutions can help achieve better disease prevention and enhanced access to healthcare for everyone.

1.2 Motivation:

The project was selected because of the increasing demand for affordable, technology-based healthcare solutions that are able to help in the early diagnosis of chronic diseases. With diseases such as diabetes and Parkinson's reaching millions globally, early diagnosis is key to avoiding complications and enhancing treatment outcomes. Yet, conventional diagnosis techniques are usually associated with specialized medical equipment, qualified specialists, and expensive tests, which are not accessible to everybody, especially in rural or poor regions. With the use of web technologies and machine learning models, this initiative is to develop a low-cost and easy-to-use solution that allows individuals to evaluate their health risk immediately and make the right choices for receiving medical treatment. Another justification for choosing this project is the development of AI in healthcare.

Machine learning algorithms have been found to be very efficient in disease prediction using medical data. Through the use of available datasets and predictive models, this project illustrates how AI can be used to aid medical diagnosis and enhance healthcare accessibility. The addition of a web-based application also ensures that the solution is easily accessible, enabling users to conduct health assessments anywhere they have an internet connection.

Potential Applications and Impact :

The main application of this project is in personalized healthcare, where users can input their health parameters and get instant disease risk assessments.

This can be used as a preliminary screening device, prompting individuals to seek professional medical advice if they are given a high-risk forecast.

In distant and under-served regions where medical services are in short supply, this kind of device can close the gap in medical availability, allowing individuals to take active measures in controlling their health. Aside from individual health tracking, the model can be incorporated into telemedicine platforms, enabling healthcare workers to have an AI-augmented screening tool that can aid in patient diagnosis. Hospitals and clinics can also apply similar predictive models to facilitate patient triaging, flagging high-risk patients for additional testing and treatment. Additionally, public health agencies can apply similar AI-powered solutions to track population health trends, locating areas with increased disease prevalence and distributing resources there. The scope of this project goes beyond single diagnosis; it supports the digitalization of healthcare, with AI-powered innovations that enhance efficiency, accessibility, and preventive medicine. By providing disease prediction in an accessible and affordable manner, this project has the ability to empower individuals, lower the healthcare burden, and eventually improve public health outcomes worldwide.

1.3 Objective:

The main goal of this project is to create a machine learning-based web application that can predict diseases like diabetes and Parkinson's using user-input health parameters. Utilizing trained machine learning models, the system will give users a proper and efficient means of disease detection at an early stage, improving healthcare accessibility and awareness in the long run.

One of the central themes of the project is to increase accessibility to early disease detection, particularly for people living in rural or underserved communities who might not have convenient access to medical facilities or costly diagnostic procedures. By providing a low-cost, web-based option, the project aims to fill the gap between medical care and the underserved, enabling people to evaluate their health risks from the comfort of their homes.

To do this, the system is made to unify several disease prediction models into one interface. Users will have the option to choose the disease they wish to evaluate, input pertinent medical information, and obtain an immediate prediction. This guarantees that the application is not only specific to a particular disease but rather offers a scalable and flexible solution for a range of health conditions.

Another important goal is to create a friendly and interactive web application with **Streamlit** and other web technologies. The interface should be easy to use and intuitive, with clear navigation, well-organized input fields, and helpful output displays. This way, users of all technical backgrounds can easily interact with the application and get useful results.

Precision and speed are at the core of the project's success. The machine learning algorithms employed within the application are optimized and trained using **applicable medical datasets** to yield accurate predictions. Through the elimination of false positives and false negatives, the system will offer users credible information on their health statuses. Moreover, the application will be capable of **making real-time predictions**, where users are provided with instantaneous results from their inputs without experiencing unnecessary delays.

In addition to personal health evaluations, the project is also intended to **enable public deployment and scalability.** With the application being hosted on **public platforms like GitHub or cloud services,** the system can be made accessible on a large scale to users in various regions. This also makes it possible to add future developments, like more disease models or real-time data sources for greater accuracy.

Finally, the project aims to **improve preventive healthcare awareness** through promoting active responses by the users based on their health diagnoses. By generating rapid and readily accessible disease risk estimates, the system is capable of facilitating effective decisions by users to consult physicians, thereby promoting enhanced disease prevention and management.

1.4 Scope of the Project:

Scope :

This project is centered on creating a web application based on machine learning to predict diabetes and Parkinson's diseases using user input. The system uses trained machine learning models to interpret health parameters and give real-time risk predictions, thus enabling early detection to be more cost-effective and within reach. By consolidating several disease prediction models into a single online platform, the project hopes to act as an overall health assessment tool that can help people make informed decisions about their own health. The tool has a friendly user interface implemented with Streamlit, enabling users to move freely from one disease prediction to another. It also includes interactive input boxes where users can enter corresponding health measures like glucose levels, blood pressure, age, and other medical values, based on the choice of disease. The system will then process these inputs with pre-trained models and produce a predictive result, enabling users to determine their risk levels.

The project further delves into real-time predictions to ensure that the users get instant results based on the data they provide. The platform is also scalable, which implies it can be used to add more diseases in the future by training and incorporating other machine learning models. The application is for educational and initial health screening use, allowing the users to understand possible health risks prior to seeking professional medical consultation.

In addition, the project is intended to be deployed on public platforms, e.g., GitHub or cloud services, which allows it to reach a large number of users. This means that users in various regions are able to leverage the technology without needing specialized hardware or costly software

Limitations :

Even with its benefits, the project has some drawbacks. One of the main drawbacks is that the system's predictions are not a replacement for professional medical diagnosis. Machine learning algorithms, as powerful as they are, are trained on certain data and may not cover all potential variations in medical conditions. Users should seek advice from healthcare professionals for confirmation and additional medical evaluation instead of trusting the application's predictions.

Another limitation is the dependence on the quality and diversity of training data. The prediction accuracy is highly dependent on the dataset that was used to train the models. If the dataset is not diverse in terms of demographic representation or medical conditions, the performance of the model might be biased or restricted in applicability to actual scenarios.

The system also depends on user-supplied input data, which poses the risk of inaccurate or incomplete information. If users supply inexact values—due to measurement inaccuracy, ignorance, or deliberate misreporting—the forecasts may be unreliable. This underlines the need for user familiarity and correct data entry in the application.

Furthermore, the project is constrained to illnesses for which training models exist in machine learning. Although the system presently caters to diabetes and Parkinson's forecasting, its use in other conditions is contingent on the existence of relevant datasets as well as training the models. The app doesn't diagnose chronic or uncommon ailments that need wide-ranging clinical evaluations, laboratory examination, or radiology scans.

Lastly, the performance of the web application depends on system resources and internet connectivity. Because the application is based on Streamlit and cloud platforms, users with slow or unreliable internet connections can expect lag in loading the interface or making predictions. In addition, incorporating more advanced models in the future can also necessitate more computational power, which can impact real-time performance.

CHAPTER 2

Literature Survey

2.1Review of Relevant Literature and Previous Work

The use of machine learning (ML) in healthcare has been widely explored, particularly in disease prediction and early diagnosis. Many studies have demonstrated the effectiveness of ML models in identifying patterns in patient data that can be used for risk assessment and disease prediction. This section provides an overview of previous research and related work in the domain of disease prediction using machine learning**,** with a specific focus on diabetes and Parkinson's disease.

Machine learning has proven to be a powerful tool for healthcare applications, enabling predictive analysis based on historical and real-time patient data. Research has shown that supervised learning algorithms, such as decision trees, support vector machines (SVM), random forests, and neural networks, can effectively classify patients based on disease risk factors. Several studies have applied these techniques to predict diseases using structured medical data from sources like electronic health records (EHRs) and public health datasets.

A study by Kavakiotis et al. (2017) reviewed machine learning applications in diabetes prediction and management**.** The authors highlighted that ML models trained on datasets such as Pima Indians Diabetes Dataset (PIDD) have achieved high accuracy in predicting diabetes based on features like blood glucose levels, body mass index (BMI), and family history. Similarly, Parkinson's disease detection has been explored using **voice** and movement analysis, as shown in research by Little et al. (2009)**,** which used speech signal processing and machine learning models to detect Parkinson's disease with high sensitivity.

Diabetes is a chronic condition that affects millions of people worldwide. Early detection and risk assessment can help prevent severe complications. Several studies have used ML to analyze patient data and predict diabetes risk. For example, Sisodia and Sisodia (2018) used logistic regression and decision trees to predict diabetes based on factors like glucose levels, insulin, and age. Their findings showed that ensemble learning techniques could enhance prediction accuracy compared to individual classifiers.

Another notable work by Khan et al. (2020) compared multiple ML models, including k-nearest neighbors (KNN), SVM, and artificial neural networks (ANNs)**,** on the PIDD dataset. Their study revealed that ensemble methods, such as random forests, achieved higher accuracy (up to 85%) compared to traditional classifiers. These findings validate the feasibility of implementing ML-based models in real-world diabetes screening applications**.**

Parkinson's disease (PD) is a neurodegenerative disorder that primarily affects motor function. Machine learning models have been applied to analyze voice recordings, hand tremor patterns, and movement data for early diagnosis. Research by Tsanas et al. (2012) used support vector machines (SVMs) and neural networks to classify Parkinson's patients based on speech impairment characteristics. Their study demonstrated that feature selection and dimensionality reduction techniques could improve model accuracy while reducing computational costs.

Further advancements in Parkinson's prediction involve deep learning and wearable sensor data**.** Arora et al. (2019) explored the use of deep neural networks (DNNs) for analyzing gait patterns and accelerometer data, achieving over 90% accuracy in distinguishing Parkinson's patients from healthy individuals. These studies indicate that ML techniques can complement clinical diagnosis

2.2 Current Models, Methods, and Techniques :

There are many machine learning models and methods for disease prediction, especially for diseases such as diabetes and Parkinson's disease. There are various models that use a number of different algorithms to model patient data, recognize patterns, and predict risk with high precision.

Diabetes Prediction Models

Current models for diabetes prediction are mostly based on structured data such as the Pima Indians Diabetes Dataset (PIDD). Various supervised learning algorithms have been used on this dataset, including:

- Logistic Regression (LR): A popular statistical technique for binary classification, which is good at predicting diabetes based on attributes such as glucose levels, BMI, and age.

- Decision Trees (DT): A rule-based model that splits data according to feature importance, giving interpretable results for diabetes prediction.

- Random Forest (RF): An ensemble technique that aggregates several decision trees to enhance prediction accuracy and avoid overfitting.

- Support Vector Machines (SVM): A strong classification algorithm that identifies the best hyperplane to separate diabetic and non-diabetic patients.

- Neural Networks (NN): Deep learning models that learn intricate interactions between input variables, enhancing predictive accuracy.

It has been observed that ensemble learning methods such as Random Forest and Gradient Boosting Machines (GBM) tend to perform better than single classifiers, offering higher accuracy and generalization.

Parkinson's Disease Prediction Models

Parkinson's disease (PD) prediction models are centered mostly on speech analysis, movement patterns, and wearable sensor data. Among the major methodologies are:

- Support Vector Machines (SVM): Employed in classifying Parkinson's patients using speech impairments and vocal attributes.

- K-Nearest Neighbors (KNN): Distance-based classifier, which recognizes the similarities among patient data points to diagnose PD.

- Deep Neural Networks (DNN): Used in gait analysis and wearable sensor accelerometer data, with a precision rate of more than 90% for detecting Parkinson's.

- Principal Component Analysis (PCA): A dimension reduction algorithm that improves model performance by choosing the most useful features from high-dimensional data.

Methodologies in Disease Prediction :

A number of methodologies improve the performance of disease prediction models:

- Feature Selection and Engineering: Determining important health indicators (e.g., glucose levels, voice tremors) to enhance model precision.

-Data Preprocessing: Managing missing values, data normalization, and dataset balancing through methods such as SMOTE (Synthetic Minority Over-sampling Technique).

-Model Optimization: Hyperparameter tuning using techniques such as Grid Search and Bayesian Optimization to increase predictive precision.

-Cross-Validation: Maintaining model reliability through dataset splitting into training and validation sets.

These pre-existing models and methods have gone a long way in refining disease prediction, such that early diagnosis is made more affordable and precise.

2.3 Gaps and Limitations in Current Solutions

In spite of the great progress made in machine learning-based disease prediction, some limitations still remain in current models. One of the biggest challenges is data quality and availability. Most machine learning models are only trained on structured datasets such as the Pima Indians Diabetes Dataset (PIDD) or Parkinson's speech datasets, which might not be a good representation of diverse populations. This may result in skewed predictions and less generalizability when used for actual healthcare applications. Moreover, imbalanced datasets, wherein one class (e.g., diabetic patients) is underrepresented, can be detrimental to model performance, with predictions being unreliable for minority instances. The second limitation is feature selection and interpretability. Most deep learning models, though very accurate, are black boxes that don't offer much explanation of decision-making processes. This transparency reduction diminishes medical professionals' and patients' trust. In addition, some models use handcrafted features, where expert knowledge is needed to choose the most useful health parameters. This manual process can lead to errors and reduce model scalability.

The unification of several diseases within one predictive system is another gap. The majority of current models are disease-specific, i.e., different applications for various health conditions such as diabetes and Parkinson's. This makes the usability and applicability of ML-based healthcare applications low, as users have to deal with several platforms for various diagnoses.

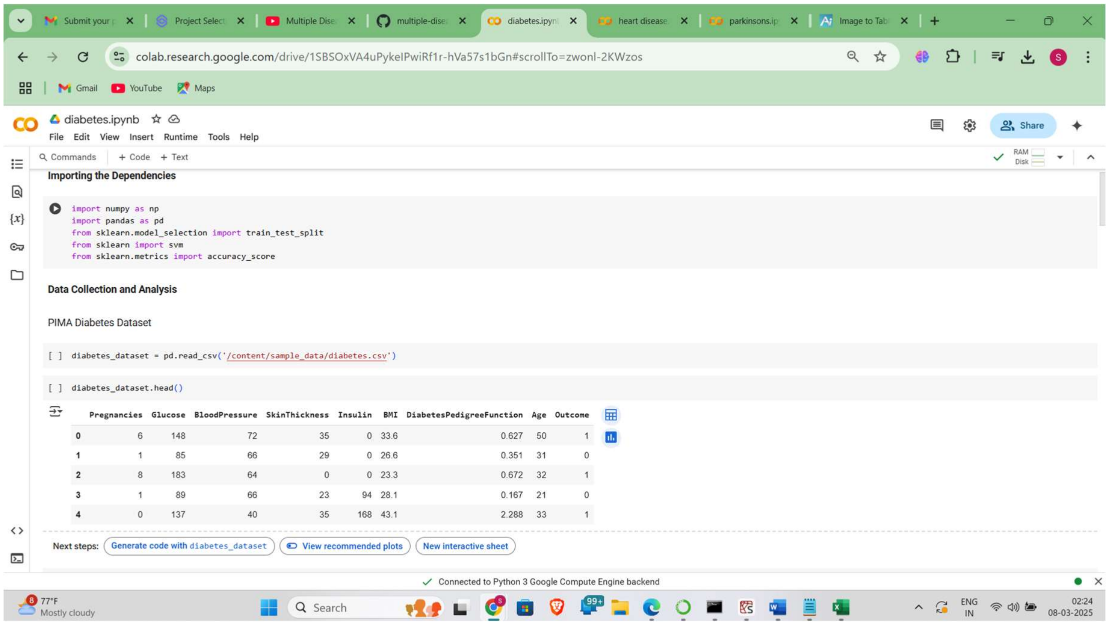
How This Project Fills These Gaps :

This project seeks to enhance current solutions by creating a multi-disease prediction system that combines machine learning models for diabetes, Parkinson's, and other diseases into one web-based application. With the use of Streamlit for an interactive UI, users can easily switch between various disease predictions without having to change applications.

To overcome data imbalances and bias, this project will use methods such as Synthetic Minority Over-sampling Technique (SMOTE) to provide improved class distribution. Moreover, feature engineering and selection methods will be used to determine the most significant health parameters, improving accuracy and interpretability.

The project is also centered around model explainability, applying methods such as SHAP (SHapley Additive Explanations) values to shed light on how individual features impact predictions. This will assist in instilling confidence in medical professionals and patients, increasing the transparency and reliability of AI-based healthcare solutions.

By filling these gaps, this project seeks to develop a more accessible, accurate, and user-friendly disease prediction system that enhances early diagnosis and preventive healthcare practices.


colab.research.google.com/drive/1SBSOxVA4uPykelPwiRf1r-hVa57s1bGn#scrollTo=zwonl-2KWzos
diabetes.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text
Share

Importing the Dependencies

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import accuracy_score
```

Data Collection and Analysis

PIMA Diabetes Dataset

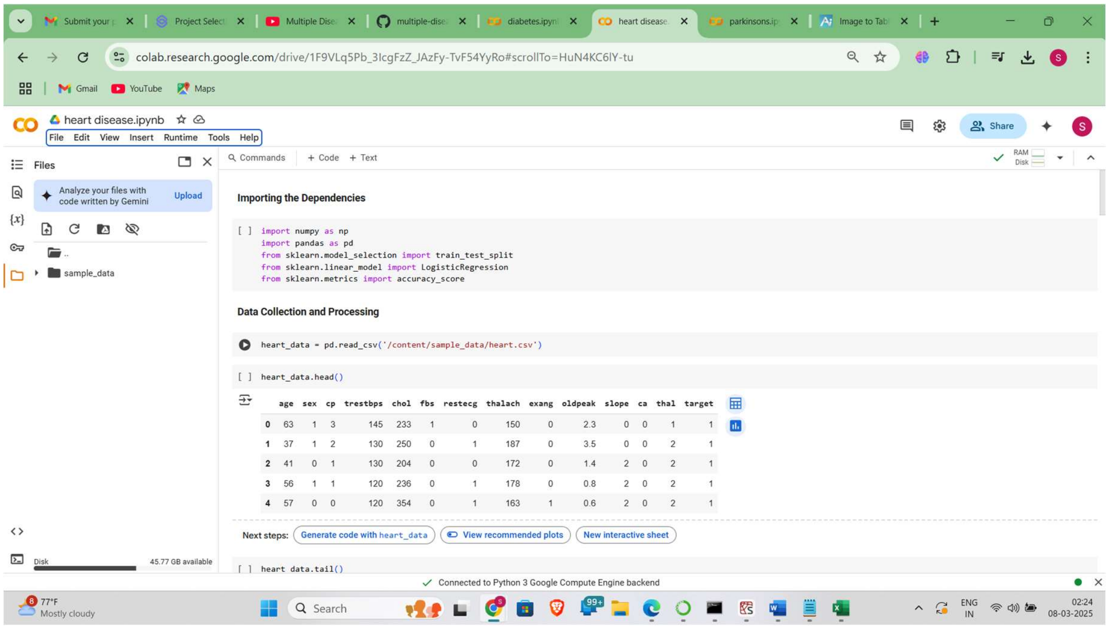
```
[ ] diabetes_dataset = pd.read_csv('/content/sample_data/diabetes.csv')
```

```
[ ] diabetes_dataset.head()
```

	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

Next steps: Generate code with diabetes_dataset
View recommended plots
New interactive sheet
Connected to Python 3 Google Compute Engine backend
77°F Mostly cloudy
02:24 08-03-2025

Figure 1 : Machine learning model code for Diabetes prediction


colab.research.google.com/drive/1F9VLq5Pb_3IcgFzZ_JAzFy-TvF54YyRo#scrollTo=HuN4KC6lY-tu
heart disease.ipynb
File Edit View Insert Runtime Tools Help
Files
Analyze your files with code written by Gemini
Upload
sample_data
Disk 45.77 GB available
Commands + Code + Text
Share

Importing the Dependencies

```
[ ] import numpy as np
    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score
```

Data Collection and Processing

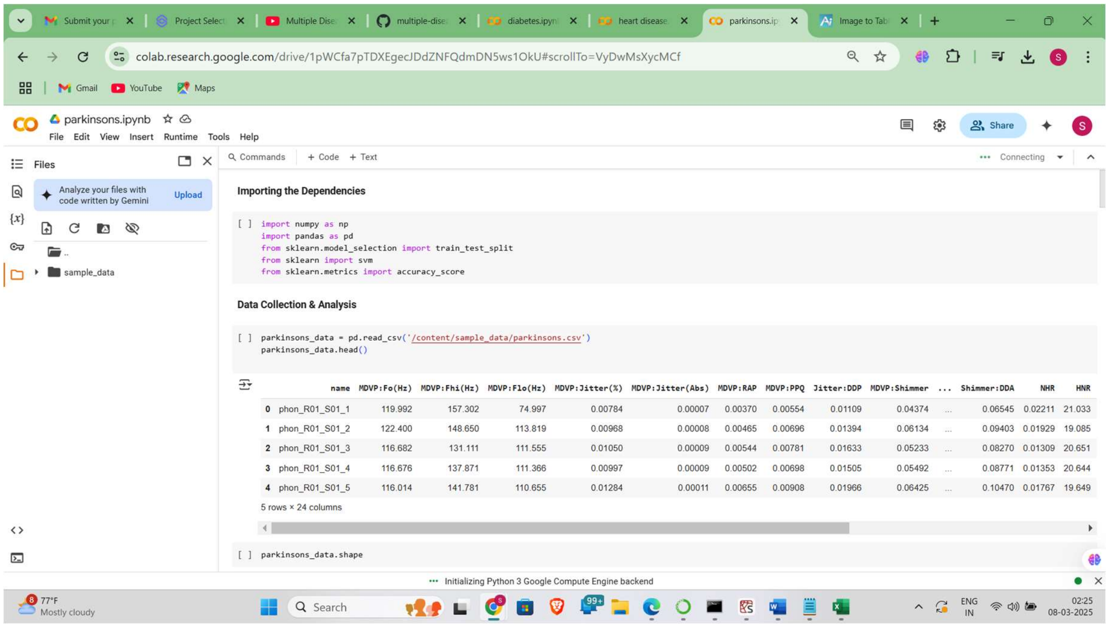
```
heart_data = pd.read_csv('/content/sample_data/heart.csv')
```

```
[ ] heart_data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
[ ] heart_data.tail()
```

Figure 2 : Machine learning model code for heart disease prediction



Importing the Dependencies

```
[ ] import numpy as np
    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn import svm
    from sklearn.metrics import accuracy_score
```

Data Collection & Analysis

```
[ ] parkinsons_data = pd.read_csv('/content/sample_data/parkinsons.csv')
    parkinsons_data.head()
```

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	...	Shimmer:DDA	NHR	HNR
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374	...	0.06545	0.02211	21.033
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	...	0.09403	0.01929	19.085
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233	...	0.08270	0.01309	20.651
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	...	0.08771	0.01353	20.644
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	...	0.10470	0.01767	19.649

5 rows × 24 columns

```
[ ] parkinsons_data.shape
```

colab.research.google.com/drive/1pWCfa7pTDXEgecJDdZNFQdmDN5ws1OkU#scrollTo=VyDwMsXycMCf
parkinsons.ipynb
File Edit View Insert Runtime Tools Help
Share
Files
Analyze your files with code written by Gemini
Upload
sample_data
Commands
+ Code
+ Text
Connecting
Initializing Python 3 Google Compute Engine backend
77°F
Mostly cloudy
Search
ENG
IN
02:25
08-03-2025

Figure 3 : Machine learning code for parkinsons disease prediction

CHAPTER 3

Proposed Methodology

3.1 System Design

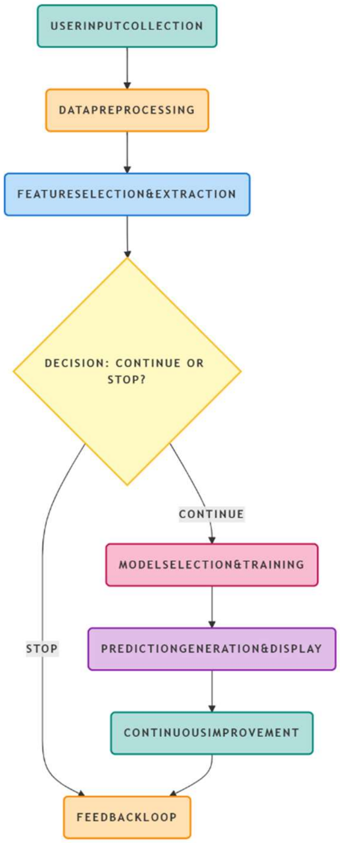
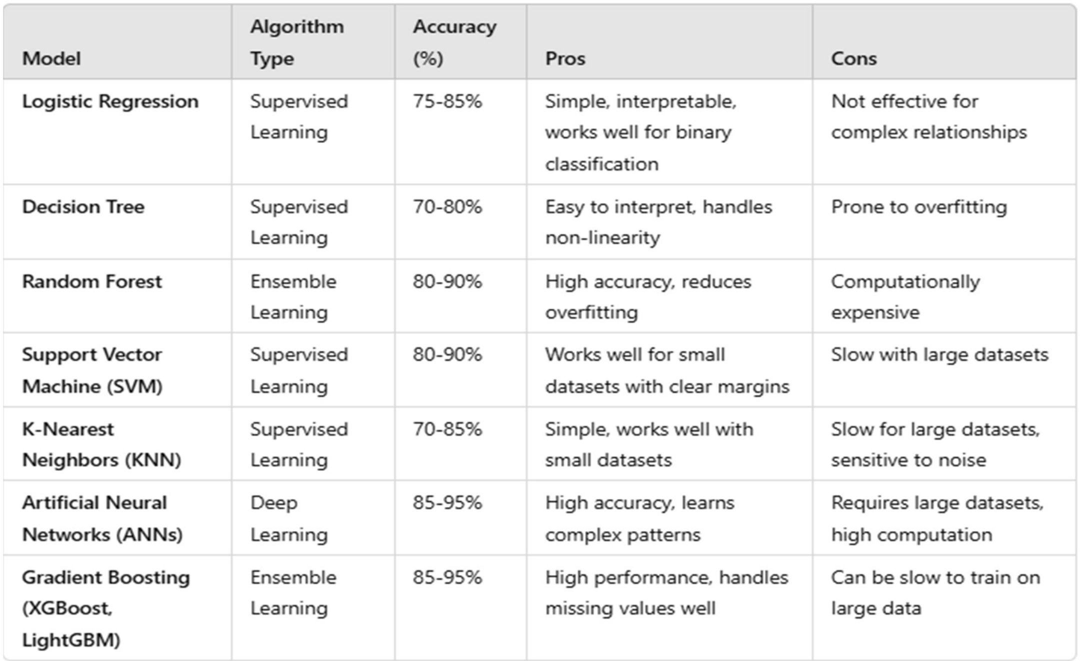


Figure 4 : Flow chart displaying the methodolgy

The Multi-Disease Prediction System uses machine learning to identify patient data and make predictions of diseases such as diabetes, Parkinson's, and cardiovascular diseases. It starts with collecting user data and then preprocessing it to address missing values and normalize inputs. Significant features are identified to increase the accuracy of the model and machine learning techniques such as SVM, Random Forest, and Neural Networks are utilized to predict disease. The results are presented to users, showing them their risk to health. The system becomes more accurate over time through feedback, and thus it is an effective means for the early diagnosis of diseases and enhanced healthcare decision-making.



Model	Algorithm Type	Accuracy (%)	Pros	Cons
Logistic Regression	Supervised Learning	75-85%	Simple, interpretable, works well for binary classification	Not effective for complex relationships
Decision Tree	Supervised Learning	70-80%	Easy to interpret, handles non-linearity	Prone to overfitting
Random Forest	Ensemble Learning	80-90%	High accuracy, reduces overfitting	Computationally expensive
Support Vector Machine (SVM)	Supervised Learning	80-90%	Works well for small datasets with clear margins	Slow with large datasets
K-Nearest Neighbors (KNN)	Supervised Learning	70-85%	Simple, works well with small datasets	Slow for large datasets, sensitive to noise
Artificial Neural Networks (ANNs)	Deep Learning	85-95%	High accuracy, learns complex patterns	Requires large datasets, high computation
Gradient Boosting (XGBoost, LightGBM)	Ensemble Learning	85-95%	High performance, handles missing values well	Can be slow to train on large data

Table 1 : Table of comparison of models

3.2 Requirement Specification

Tools and Technologies Required for Implementation

3.2.1 Hardware Requirements:

- Processor: Intel Core i5 or higher (Recommended: i7 for faster processing)
- RAM: Minimum 8GB (Recommended: 16GB for handling large datasets)
- Storage: At least 256GB SSD (Recommended: 512GB SSD for faster data processing)
- Graphics Card: Not mandatory, but a GPU (NVIDIA GTX/RTX) can accelerate model training
- Internet Connection: Required for accessing datasets, libraries, and cloud services

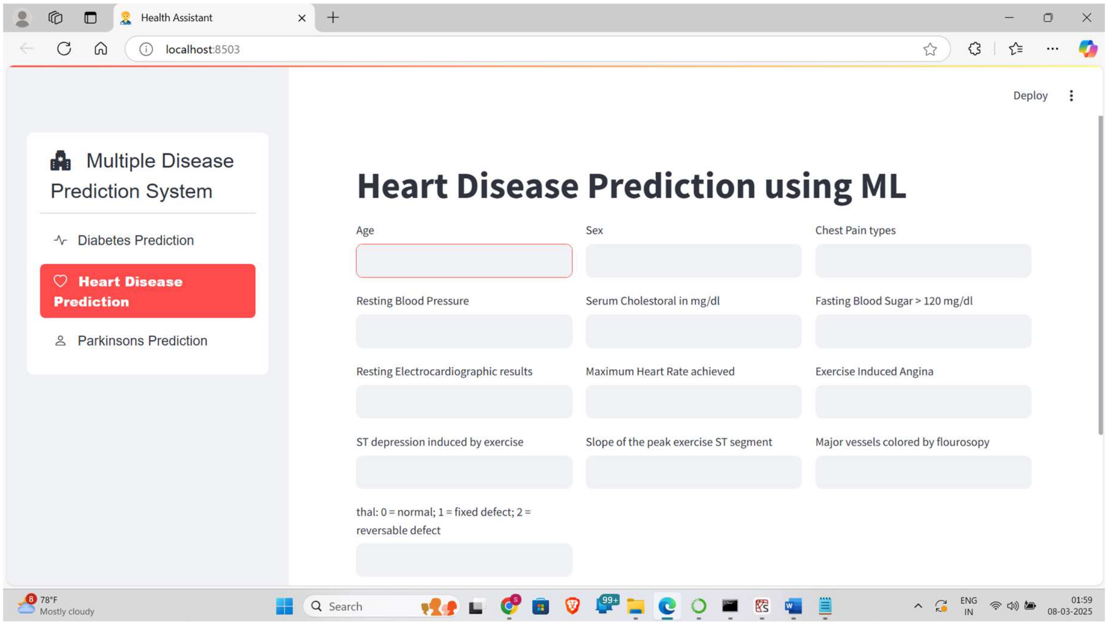
3.2.2 Software Requirements:

- Operating System: Windows 10/11, macOS, or Linux (Ubuntu recommended)
- Programming Language: Python (with libraries like NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch)
- Development Environment: Jupyter Notebook, PyCharm, or VS Code
- Machine Learning Frameworks: TensorFlow, Keras, Scikit-learn
- Data Processing Tools: Pandas, NumPy, Matplotlib, Seaborn
- Web Framework: Streamlit for creating the user interface
- Database: SQLite, Firebase, or MySQL for storing user data (if required)
- Version Control: Git and GitHub for collaboration and deployment
- Cloud Services: AWS, Google Cloud, or Azure (optional for scalability)

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:



Heart Disease Prediction using ML

Age

Sex

Chest Pain types

Resting Blood Pressure

Serum Cholestoral in mg/dl

Fasting Blood Sugar > 120 mg/dl

Resting Electrocardiographic results

Maximum Heart Rate achieved

Exercise Induced Angina

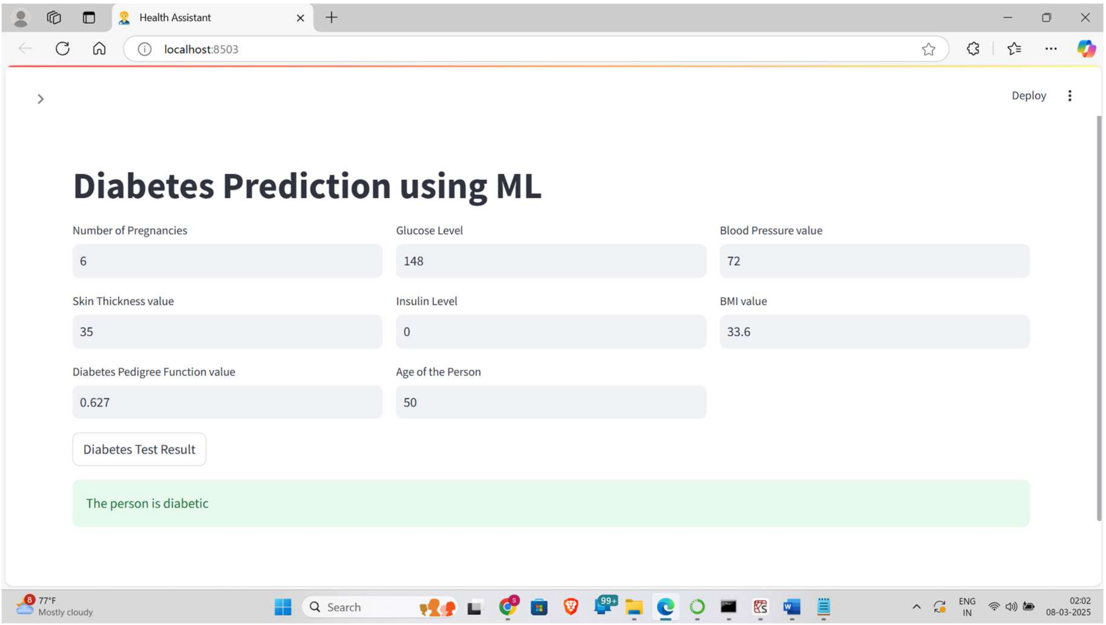
ST depression induced by exercise

Slope of the peak exercise ST segment

Major vessels colored by flourosopy

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

Figure 5 : The prediction of heart disease using ml



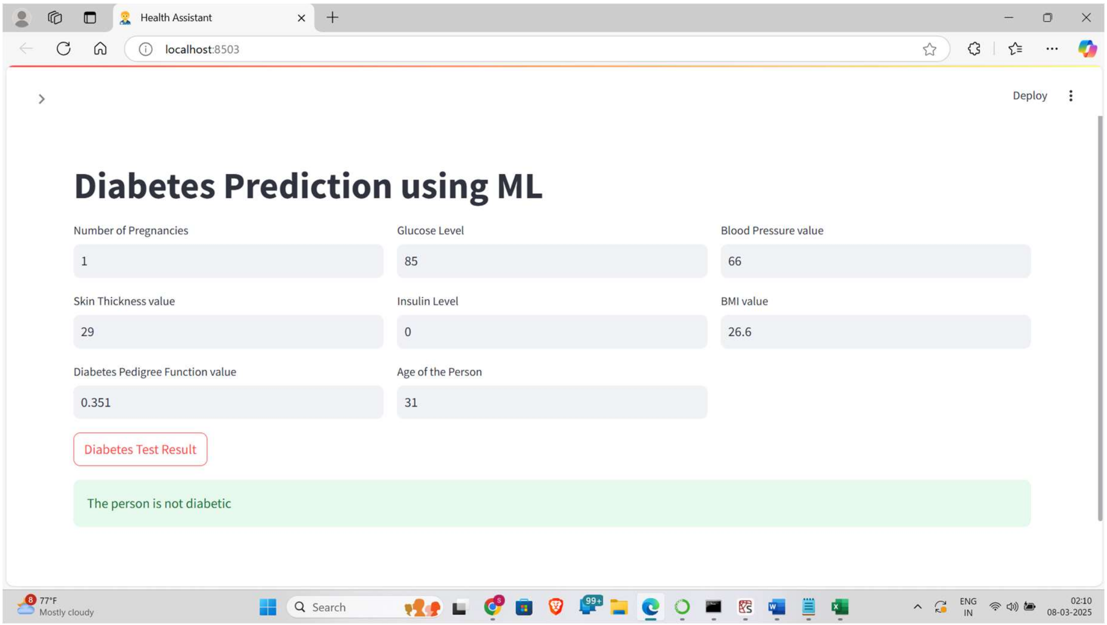
Diabetes Prediction using ML

Number of Pregnancies: 6
Glucose Level: 148
Blood Pressure value: 72
Skin Thickness value: 35
Insulin Level: 0
BMI value: 33.6
Diabetes Pedigree Function value: 0.627
Age of the Person: 50

Diabetes Test Result

The person is diabetic

Figure 6 : The prediction of diabetic using ml model



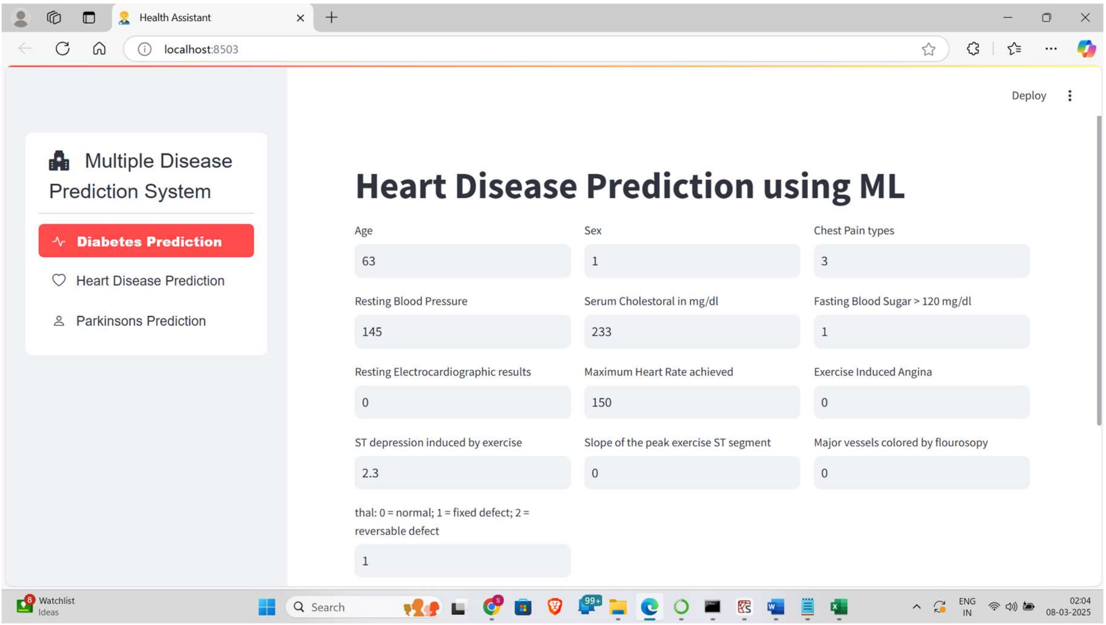
Diabetes Prediction using ML

Number of Pregnancies: 1
Glucose Level: 85
Blood Pressure value: 66
Skin Thickness value: 29
Insulin Level: 0
BMI value: 26.6
Diabetes Pedigree Function value: 0.351
Age of the Person: 31

Diabetes Test Result

The person is not diabetic

Figure 7 : The prediction of not having diabetics



Heart Disease Prediction using ML

Age: 63

Sex: 1

Chest Pain types: 3

Resting Blood Pressure: 145

Serum Cholestoral in mg/dl: 233

Fasting Blood Sugar > 120 mg/dl: 1

Resting Electrocardiographic results: 0

Maximum Heart Rate achieved: 150

Exercise Induced Angina: 0

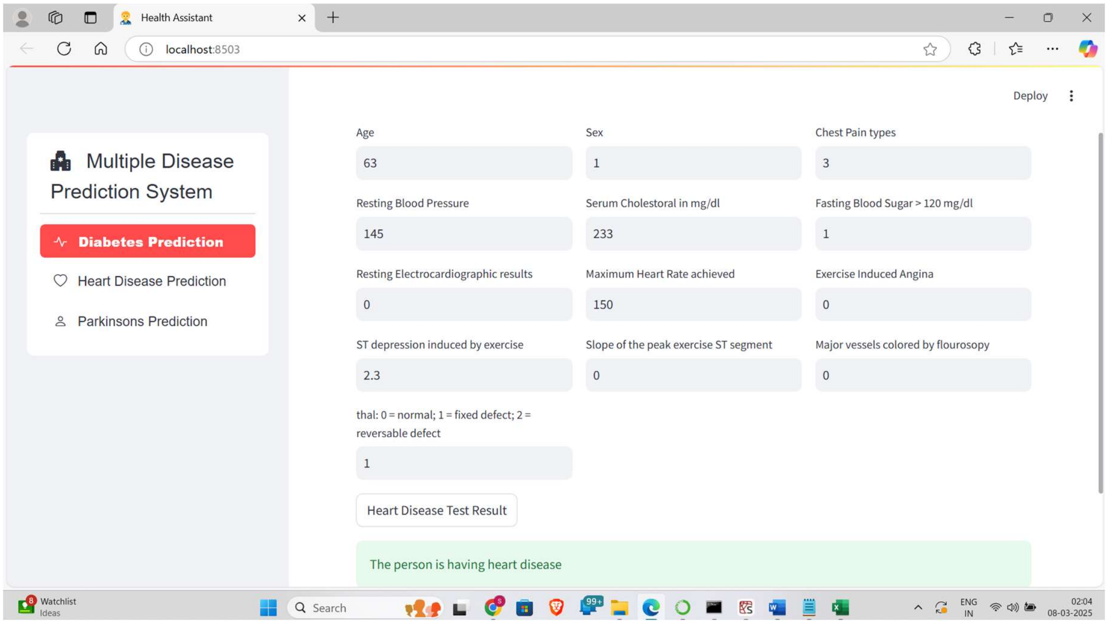
ST depression induced by exercise: 2.3

Slope of the peak exercise ST segment: 0

Major vessels colored by flourosopy: 0

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

Figure 8 : The prediction of heart disease using ml model



Age: 63

Sex: 1

Chest Pain types: 3

Resting Blood Pressure: 145

Serum Cholestoral in mg/dl: 233

Fasting Blood Sugar > 120 mg/dl: 1

Resting Electrocardiographic results: 0

Maximum Heart Rate achieved: 150

Exercise Induced Angina: 0

ST depression induced by exercise: 2.3

Slope of the peak exercise ST segment: 0

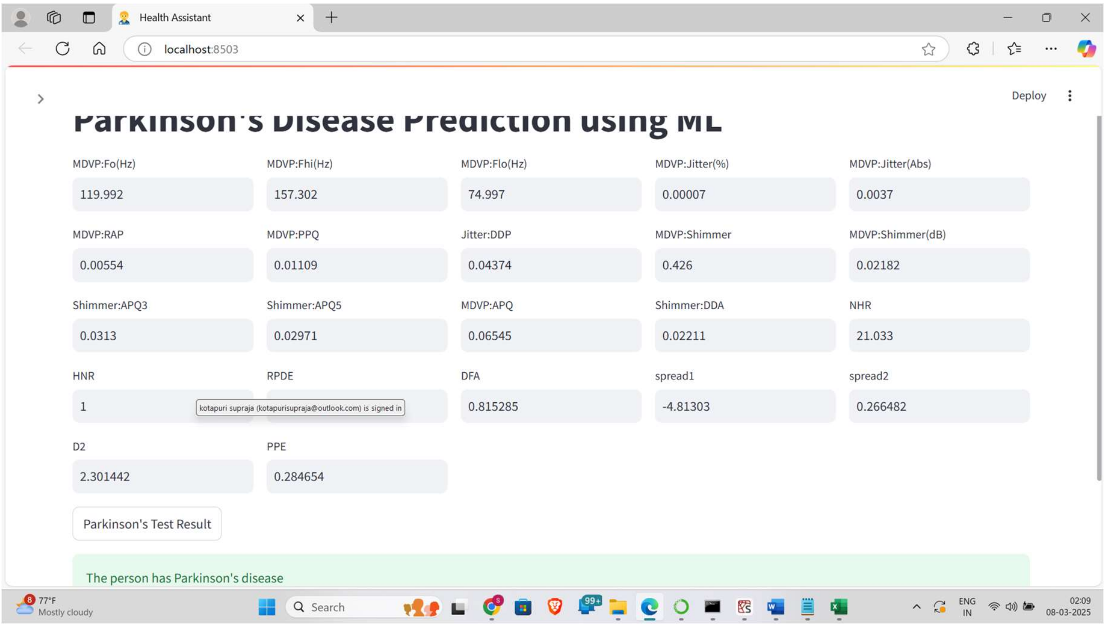
Major vessels colored by flourosopy: 0

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

Heart Disease Test Result

The person is having heart disease

Figure 9 : The outlook of the website created



Parkinson's Disease Prediction using ML

MDVP:Fo(Hz): 119.992

MDVP:Fhi(Hz): 157.302

MDVP:Flo(Hz): 74.997

MDVP:Jitter(%): 0.00007

MDVP:Jitter(Abs): 0.0037

MDVP:RAP: 0.00554

MDVP:PPQ: 0.01109

Jitter:DDP: 0.04374

MDVP:Shimmer: 0.426

MDVP:Shimmer(dB): 0.02182

Shimmer:APQ3: 0.0313

Shimmer:APQ5: 0.02971

MDVP:APQ: 0.06545

Shimmer:DDA: 0.02211

NHR: 21.033

HNR: 1

RPDE:

DFA: 0.815285

spread1: -4.81303

spread2: 0.266482

D2: 2.301442

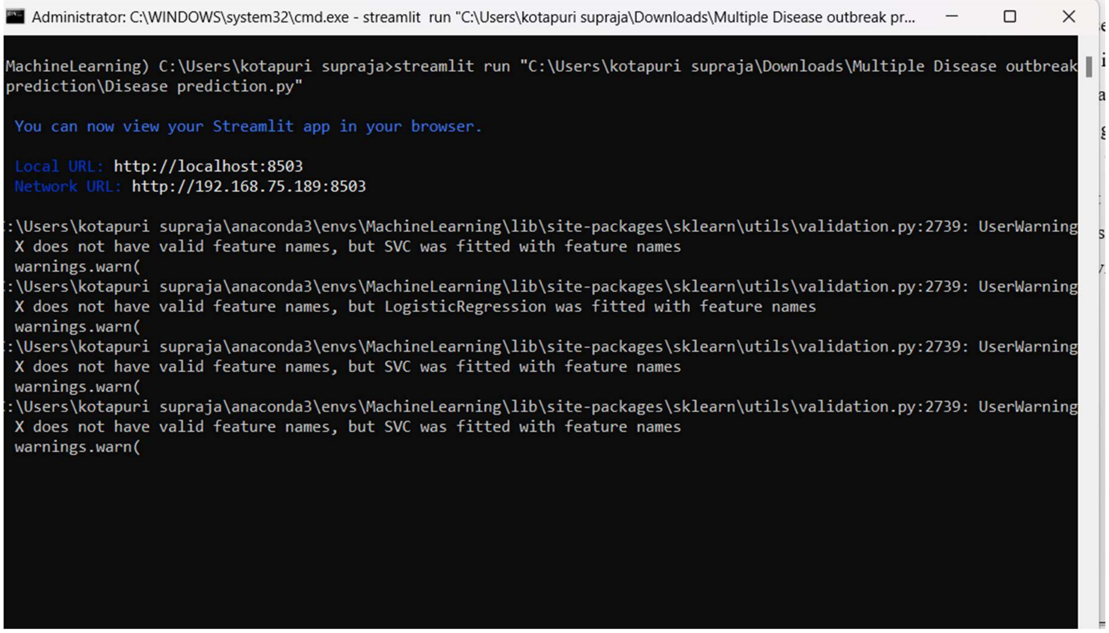
PPE: 0.284654

kotapuri supraja (kotapurisupraja@outlook.com) is signed in

The person has Parkinson's disease

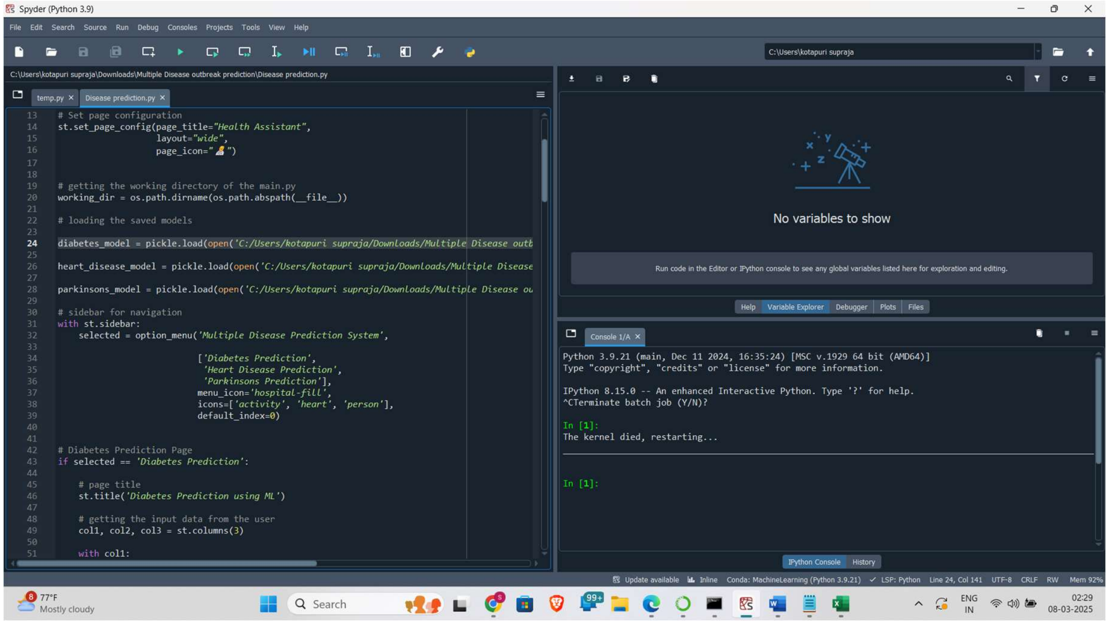
Figure 10 : The prediction of parkinsons disease using ml model

This is a Multiple Disease Prediction System that I have created with Streamlit and machine learning models to predict diabetes, heart disease, and Parkinson's disease. The web application is locally accessible on `localhost:8503` and has a user-friendly interface with a sidebar for navigation across various disease prediction models. Now the Diabetes Prediction part is live, where users can enter medical parameters like glucose level, blood pressure, BMI, insulin level, and age. Once the "Diabetes Test Result" button is clicked, the system calculates the data with the help of a trained machine learning model and shows the result. Here, the model has predicted that "The person is not diabetic," in a green success message. The interface is made simple, interactive, and responsive so that both medical professionals and common users can use it. The Deploy button at the top right corner shows that this application can be deployed easily on cloud platforms for easy access. This project is intended to help in early disease detection using AI-driven predictions, giving rapid and accurate health checks.


Administrator: C:\WINDOWS\system32\cmd.exe - streamlit run "C:\Users\kotapuri supraja\Downloads\Multiple Disease outbreak pr...

```
MachineLearning) C:\Users\kotapuri supraja>streamlit run "C:\Users\kotapuri supraja\Downloads\Multiple Disease outbreak
prediction\Disease prediction.py"

 You can now view your Streamlit app in your browser.

 Local URL: http://localhost:8503
 Network URL: http://192.168.75.189:8503

:\Users\kotapuri supraja\anaconda3\envs\MachineLearning\lib\site-packages\sklearn\utils\validation.py:2739: UserWarning
 X does not have valid feature names, but SVC was fitted with feature names
 warnings.warn(
:\Users\kotapuri supraja\anaconda3\envs\MachineLearning\lib\site-packages\sklearn\utils\validation.py:2739: UserWarning
 X does not have valid feature names, but LogisticRegression was fitted with feature names
 warnings.warn(
:\Users\kotapuri supraja\anaconda3\envs\MachineLearning\lib\site-packages\sklearn\utils\validation.py:2739: UserWarning
 X does not have valid feature names, but SVC was fitted with feature names
 warnings.warn(
:\Users\kotapuri supraja\anaconda3\envs\MachineLearning\lib\site-packages\sklearn\utils\validation.py:2739: UserWarning
 X does not have valid feature names, but SVC was fitted with feature names
 warnings.warn(
```

Figure 11 : The figure indicates the output when run in terminal after excuting the python code

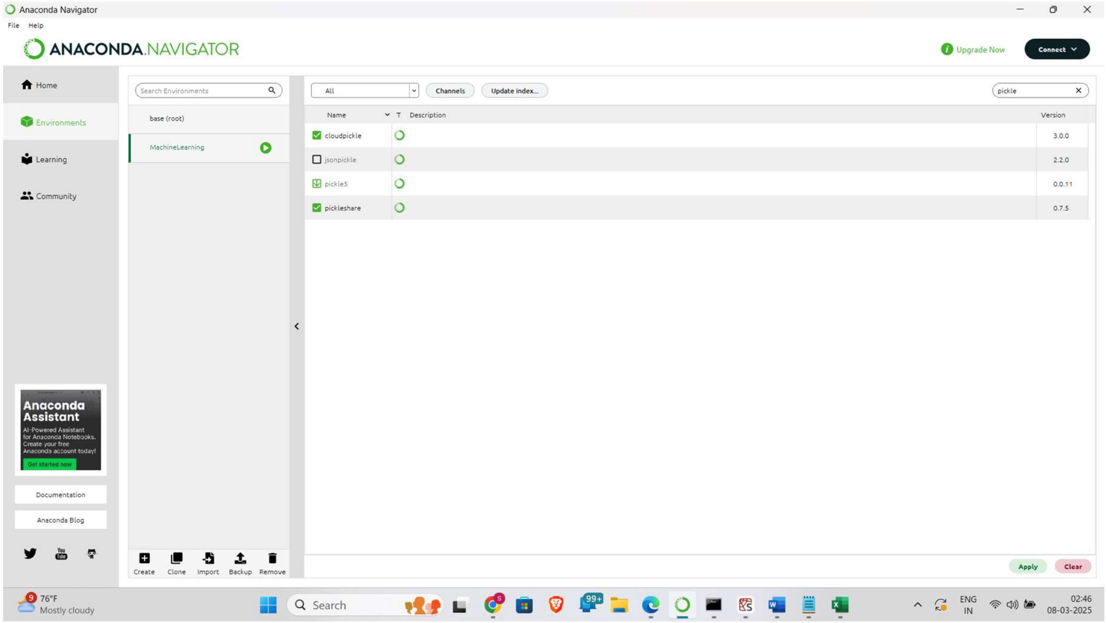


```
# Set page configuration
st.set_page_config(page_title="Health Assistant",
                   layout="wide",
                   page_icon="🧑‍⚕️")


# getting the working directory of the main.py
working_dir = os.path.dirname(os.path.abspath(__file__))

# loading the saved models

diabetes_model = pickle.load(open('C:/Users/kotapuri supraja/Downloads/Multiple Disease outb

heart_disease_model = pickle.load(open('C:/Users/kotapuri supraja/Downloads/Multiple Disease

parkinsons_model = pickle.load(open('C:/Users/kotapuri supraja/Downloads/Multiple Disease ou

# sidebar for navigation
with st.sidebar:
    selected = option_menu('Multiple Disease Prediction System',

                           ['Diabetes Prediction',
                            'Heart Disease Prediction',
                            'Parkinsons Prediction'],
                           menu_icon='hospital-fill',
                           icons=['activity', 'heart', 'person'],
                           default_index=0)


# Diabetes Prediction Page
if selected == 'Diabetes Prediction':

    # page title
    st.title('Diabetes Prediction using ML')

    # getting the input data from the user
    col1, col2, col3 = st.columns(3)

    with col1:
```

```
Python 3.9.21 (main, Dec 11 2024, 16:35:24) [MSC v.1929 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 8.15.0 -- An enhanced Interactive Python. Type '?' for help.
^CTerminate batch job (Y/N)?

In [1]:
The kernel died, restarting...


In [1]:
```

Run code in the Editor or IPython console to see any global variables listed here for exploration and editing.

Figure 12 : The python code in spyder to create the website for ml model



Name	T	Description	Version
☑ cloudpickle	○		3.0.0
□ jsonpickle	○		2.2.0
pickle5	○		0.0.11
☑ pickleshare	○		0.7.5

Figure 13 : Anaconda navigator environment creation

4.2 GitHub Link for Code:

Githublink:https://github.com/sweety2003/Disease-outbreak-Prediction/tree/main/Multiple%20Disease%20outbreak%20prediction

Local host link : http://localhost:8503

Network URL : http://192.168.75.189:8503

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

To enhance the disease outbreak prediction model, the future effort can be aimed at improving the quality of data by including real-time multi-source data from hospitals, weather, and population mobility trends. Applying deep learning algorithms, such as recurrent neural networks (RNNs) or transformer models, can enhance time-series data pattern identification. Having a larger and more diverse dataset will minimize bias and improve the model's ability to generalize. Incorporation of geospatial data analysis will enhance the precision of regional outbreak forecasts. Handling missing data with sophisticated imputation methods will provide more accurate forecasts. The use of explainable AI (XAI) techniques will make the model's choice more understandable to healthcare professionals. Edge computing and IoT integration can facilitate real-time monitoring and quicker decision-making. Model validation can be strengthened with cross-validation techniques on various demographics and geographical regions. Creating an adaptive learning system that learns from new outbreak patterns will improve robustness. Lastly, working with epidemiologists and public health officials will assist in fine-tuning the model to better match actual outbreak patterns.

5.2 Conclusion:

The disease outbreak prediction project plays a critical role in public health by allowing early detection and proactive intervention to contain disease spread. Through the use of machine learning algorithms, the model interprets historical and real-time data to make accurate predictions, informing healthcare professionals and policymakers to make informed decisions. The system increases resource allocation efficiency, facilitating timely intervention in high-risk regions. Combining multiple sources of data, for example, medical history and environmental conditions, enhances prediction accuracy. The project enables preventive healthcare measures, lowering hospitalization rates and the cost of healthcare. Automation of disease forecasting reduces effort and increases speed and accuracy. Explainable AI ensures transparency and trust among the users. Scalability of the model enables the model to be modified for different diseases and geographic locations. Future upgrades, including real-time data refresh and deep learning integration, will continue to enhance its accuracy. Overall, the project is a useful tool for disease surveillance and outbreak prevention, ultimately saving lives.

REFERENCES

[1]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104-116.

[2]. Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural networks for early detection of heart failure risk. Journal of the American Medical Informatics Association, 24(2), 361-370.

[3]. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015-1022.

[4]. Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 59(5), 1264-1271.

[5]. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

[6]. Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine – Beyond the peak of inflated expectations. New England Journal of Medicine, 376(26), 2507-2509.

[7]. Khan, S. S., Ning, H., Shah, S. J., Yancy, C. W., & Lloyd-Jones, D. M. (2020). Impact of machine learning on cardiovascular risk prediction. Circulation, 141(1), 32-33.

[8]. Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K. M., Dorsey, E. R., & Little, M. A. (2019). Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Journal of Medical Internet Research, 21(10), e12502.

[9]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.

[10].World Health Organization (WHO). (2022). Disease Outbreaks and Machine Learning: A New Approach to Global Health. *WHO Reports*.