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

Veda Bhavishya Gudivaka

23b01a4245@svecw.edu.in

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

I would like to take this opportunity to express my heartfelt gratitude to everyone who contributed, directly or indirectly, to the successful completion of this project.

Firstly, I would like to extend my sincere thanks to my supervisor, **Mr. Jay Rathod Sir**, for his valuable guidance and constant support throughout this project. His expertise not only helped me generate innovative ideas but also motivated me to approach challenges with confidence and clarity. His valuable advice, continuous support, and constructive feedback have been a great source of motivation and inspiration. His trust in our abilities gave us the confidence to complete this project successfully.

It has been an honour to work under his guidance during this project. His support not only helped us in the project but also contributed to our personal and professional growth. We truly appreciate his time, guidance, and encouragement in every aspect of this work.

Veda Bhavishya Gudivaka

ABSTRACT

This project is about building a **Disease Outbreak Prediction System** using **Machine Learning (ML)** to help detect chronic diseases like **Diabetes, Heart Disease, and Parkinson's Disease** at an early stage. Detecting these diseases early is important because it allows for better treatment and improves the chances of recovery. Traditional methods of diagnosis can take a lot of time and resources, so this project focuses on creating a faster, data-driven solution.

We used ML models to predict the risk of diseases based on patient information like health parameters. First, we cleaned and processed the data to make it suitable for training the models. To handle imbalanced data, we used a technique called **Tomek Links**. We then trained four ML models: **Logistic Regression, Decision Tree, Random Forest,** and **Support Vector Classifier (SVC)**. These models were evaluated based on **accuracy, precision, recall,** and the **confusion matrix**. The best-performing model was saved as a .sav file for deployment.

To make this system easy to use, we created a simple **web application** using **Streamlit** in Python. This app allows users to enter their health details and get real-time predictions about their disease risk. It can help both individuals and healthcare professionals make informed health decisions.

In the future, we plan to improve the system by adding real-time data and optimizing the models to make predictions even more accurate and reliable.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement

Chronic diseases like **Diabetes, Heart Disease, and Parkinson's Disease** are major causes of death and long-term health issues worldwide. These diseases often develop silently, with symptoms appearing only when the condition has already progressed, making early detection difficult. If these diseases are diagnosed late, it can lead to serious health problems, a lower quality of life, and higher medical expenses.

Traditional diagnostic methods rely on manual check-ups, lab tests, and consultations with specialists. These methods are often **time-consuming, expensive,** and not easily available in rural or underdeveloped areas.

Why This Problem is Important:

- **Early Detection:** Helps catch diseases early, reducing health risks.
- **Better Health Outcomes:** Improves disease management and quality of life.
- **Cost-Effective:** Reduces medical costs by avoiding expensive treatments for advanced stages.
- **Fast & Efficient:** Provides quick, data-driven predictions.
- **Accessible:** Can be useful in remote areas with fewer healthcare facilities.
- **Decision Support:** Assists doctors in making accurate diagnoses.

1.2 Motivation

The motivation behind this project is to solve the critical problem of **early detection** of chronic diseases like **Diabetes, Heart Disease,** and **Parkinson's Disease.** These diseases are becoming more common, and traditional diagnostic methods are often **slow, expensive,** and **inaccessible** to many people.

Machine Learning (ML) can analyze large amounts of medical data, find hidden patterns, and make accurate predictions. This makes ML an ideal tool for improving early diagnosis and helping doctors make better decisions.

Where This Can Be Used:

- **Healthcare Diagnostic Tools:** Helps doctors detect diseases quickly.
- **Personal Health Apps:** People can monitor their health at home.
- **Telemedicine Support:** Assists in online medical consultations.
- **Public Health Risk Assessment:** Identifies disease risks in communities.
- **Insurance Risk Evaluation:** Helps insurance companies assess health risks.

Impact of This Project:

- **Early Disease Detection:** Leads to better patient outcomes.
- **Accessible Healthcare:** Helps people in remote areas get medical support.
- **Lower Healthcare Costs:** Saves money by reducing expensive treatments.
- **Data-Driven Decisions:** Supports doctors with accurate data insights.

1.3 Objective

The goal of this project is to develop a **Machine Learning-based system** that can detect chronic diseases like **Diabetes, Heart Disease,** and **Parkinson's Disease** at an early stage.

Key Objectives:

1. **Predict Disease Risks:** Identify the chances of having a disease based on patient data.
2. **Enable Early Diagnosis:** Help doctors catch diseases early for timely treatment.
3. **Balance Data:** Use the **Tomek Links** technique to handle imbalanced data for better model performance.
4. **Develop a Web App:** Create an easy-to-use app with **Streamlit** to provide real-time health predictions.

5. **Support Medical Decisions:** Help both individuals and healthcare professionals make informed health decisions.

1.4 Scope of the Project

What the Project Covers:

- **Disease Prediction:** Focuses on early detection of **Diabetes, Heart Disease,** and **Parkinson's Disease** using ML models.
- **ML Models Used:** Includes **Logistic Regression, Decision Tree, Random Forest,** and **Support Vector Classifier (SVC).**
- **Data Processing:** Uses **Tomek Links** to handle imbalanced data, improving model accuracy.
- **Web App Deployment:** Offers a real-time, user-friendly web interface using **Streamlit.**
- **Decision Support:** Helps doctors and individuals make data-driven health decisions.

Limitations of the Project:

- **Limited Disease Coverage:** Currently focused only on three diseases.
- **Data Dependency:** The model's accuracy depends on the quality of the data provided.
- **No Real-Time Data:** Doesn't connect to live medical data or electronic health records.
- **Generalization Issues:** May not work perfectly for all populations due to differences in datasets.
- **Basic User Input:** The model predicts based only on the provided data, without considering full medical history.

This project is a step towards making healthcare **faster, more accessible,** and **data-driven,** with room for future improvements.

CHAPTER 2

Literature Survey

2.1 Review of Related Work

Machine Learning (ML) has been widely used to help detect chronic diseases like Diabetes, Heart Disease, and Parkinson's Disease at an early stage. Researchers have shown that ML models such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVMs) are effective in predicting the risk of these diseases.

To improve accuracy, many studies focus on fixing problems with imbalanced data (where some disease cases are much fewer than others) using techniques like Tomek Links. Additionally, the performance of these models is measured using metrics like accuracy, precision, recall, and confusion matrix.

In recent years, there's been a growing trend of creating web-based applications that use these ML models. These apps allow people to check their health risks in real-time, making healthcare more accessible and engaging for users.

2.2 Existing Models and Techniques

Many ML-based systems have already been developed for predicting different diseases:

- Diabetes Prediction Models:
  - Use models like Logistic Regression, Decision Tree, Random Forest, and Support Vector Classifier (SVC).
  - These models analyze features such as glucose levels, BMI, insulin levels, and other health indicators.
- Heart Disease Prediction Models:
    - Use Ensemble Methods like Random Forest along with Decision Trees, Logistic Regression, and SVC.

- Focus on factors like blood pressure, cholesterol levels, and ECG readings to predict heart disease risks.

- Parkinson's Disease Prediction Models:
    - Apply ML models like SVC and Random Forest.
    - These models analyse data related to voice changes, tremors, and motor functions for early detection.

Besides these models, many healthcare apps and mobile platform have started using ML to allow people to check their health risks on their own before consulting a doctor

2.3 Gaps in Existing Solutions and How This Project Solves Them

Problems with Current Systems:

Class Imbalance Issues:

Many models struggle with datasets where there are far fewer cases of a disease compared to healthy cases. This causes the models to be biased toward predicting the majority class (healthy) correctly while missing disease cases.

Limited Real-Time Accessibility:

Some systems don't have user-friendly apps for people to check their health status in real-time.

Generalization Challenges:

Models trained on specific datasets may not work well for people from different regions or backgrounds due to dataset biases.

Complex Deployment:

Many ML models are hard to deploy, making them difficult for non-technical users to access and use.

How This Project Solves These Problems:

Fixing Class Imbalance:

Uses a data balancing technique called Tomek Links to remove overlapping data points, helping the model make more accurate predictions.

Real-Time Web App:

Builds an interactive and easy-to-use web application using Streamlit, allowing people to get instant health risk predictions.

Robust Model Selection:

Compares multiple ML models like Logistic Regression, Decision Tree, Random Forest, and SVC to choose the one that performs best with the data.

Simple Deployment:

Uses model serialization (saving models as .sav files) to make deployment easy. This allows even people without technical skills to use the app effortlessly.

This project not only improves disease detection but also makes it accessible, user-friendly, and reliable for everyone.

CHAPTER 3

Proposed Methodology

3.1 System Design

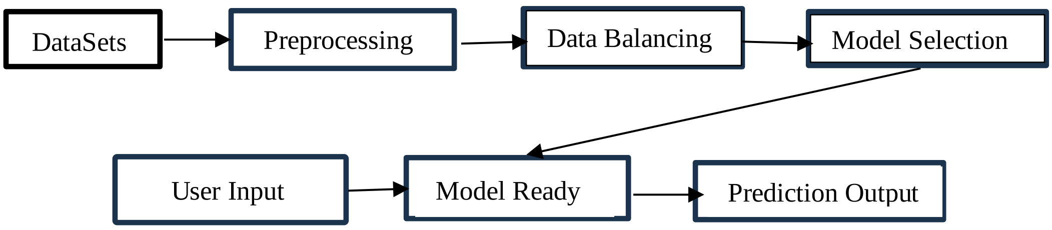

Fig.1: Disease Prediction System Architecture

Explanation of the Diagram:

1. **Dataset:**
   - Contains medical records with patient health parameters and symptoms, forming the base for model training.
2. **Preprocessing:**
   - Cleans and normalizes data, handles missing values, and prepares it for effective model learning.
3. **Data Balancing (Tomek Links):**
   - Balances the dataset by removing overlapping samples, improving model performance on imbalanced data.
4. **Model Selection (ML Algorithms):**
   - Trains models like **Logistic Regression, Decision Tree, Random Forest, and SVC**, selecting the best based on performance metrics.
5. **Model Ready:**
   - The optimized model is saved for deployment in the web application.
6. **User Input:**
   - Users enter health data into the web app, serving as input for real-time predictions.
7. **Prediction Output:**

- The model processes the input and predicts disease risk, aiding early diagnosis and timely medical consultation.

This flow ensures efficient disease prediction using ML techniques for better healthcare outcomes.

3.2 Requirement Specification

To implement the disease prediction system, the following hardware and software requirements are necessary:

3.2.1 Hardware Requirements:

- Processor: Intel Core i5 or higher (for faster computation and model training)
- RAM: 8 GB (minimum) for smooth data processing and web deployment
- Storage: 256 GB SSD or higher for efficient data handling and model storage
- Internet Connection: Stable connection for deploying and accessing the web application

3.2.2 Software Requirements:

- Programming Language: Python (for data preprocessing, model development, and deployment)
- Libraries & Frameworks:
    - Pandas, NumPy: For data manipulation and analysis
    - Scikit-learn: For implementing machine learning models (Logistic Regression, Decision Tree, Random Forest, SVC)
    - Imbalanced-learn: For applying Tomek Links for data balancing
    - Matplotlib, Seaborn: For data visualization and analysis
    - Streamlit: For building and deploying the interactive web application
    - Pickle: For model serialization and deployment
- Development Environment: Jupyter Notebook, VS Code (for coding and model development)
- Operating System: Windows 10 or higher / Linux / macOS (compatible with Python and required libraries)

These tools and technologies ensure efficient development, deployment, and performance of the disease prediction system.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

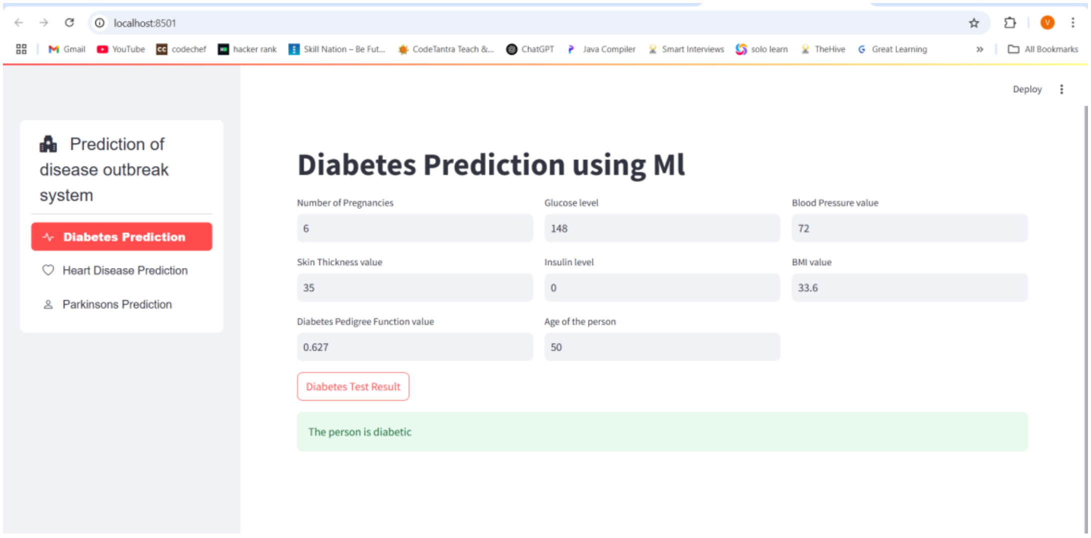

Fig. 2: Person with Diabetes

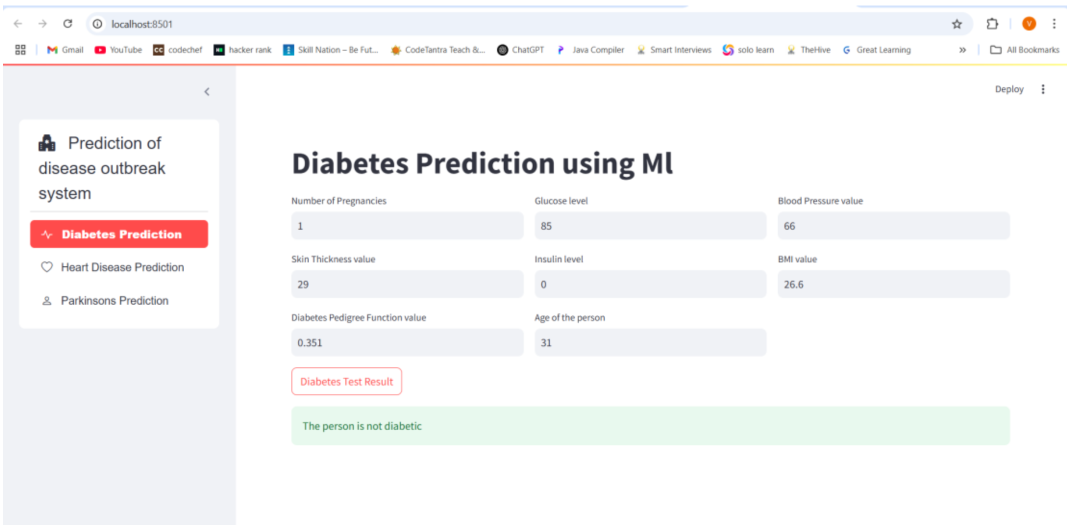

Fig. 3: Person without Diabetes

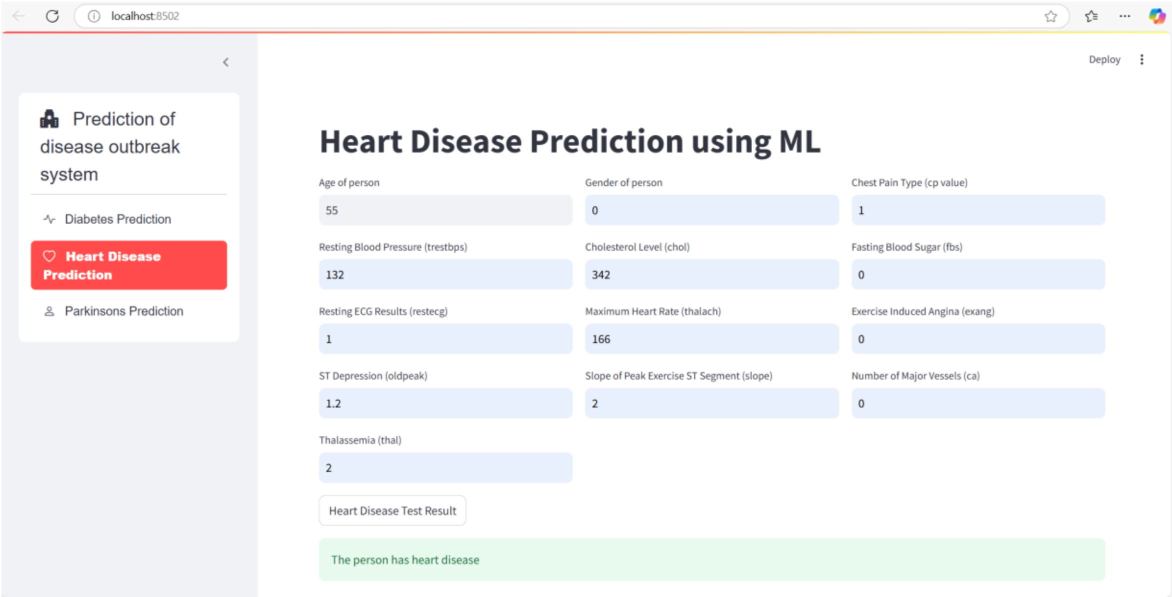

Fig. 4: Person With Heart Disease

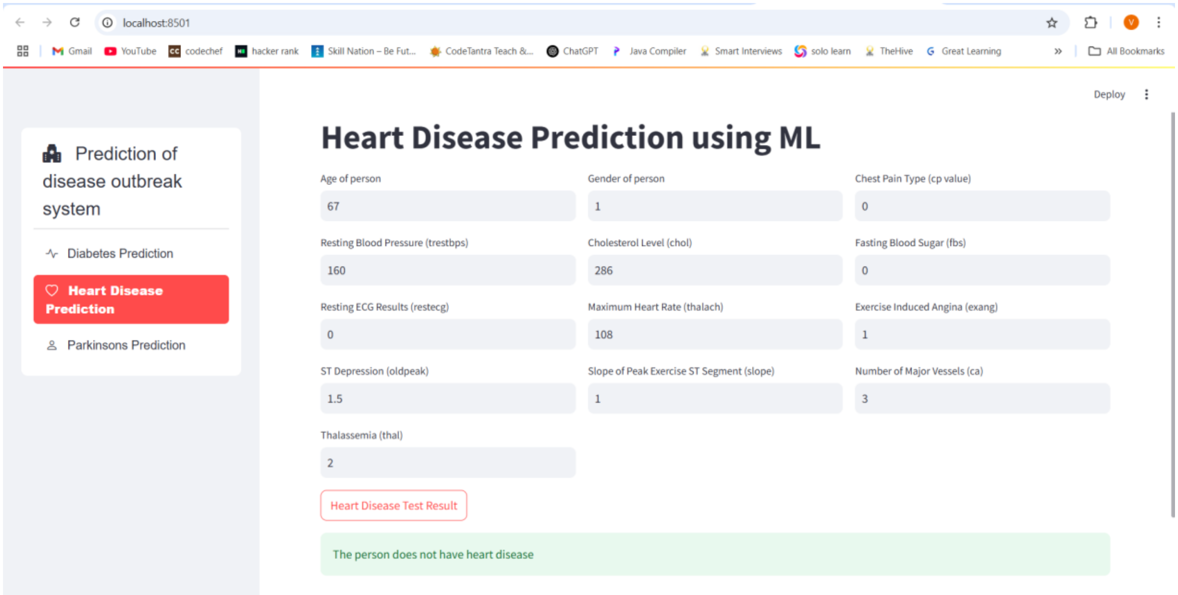

Fig. 5: Person Without Heart Disease

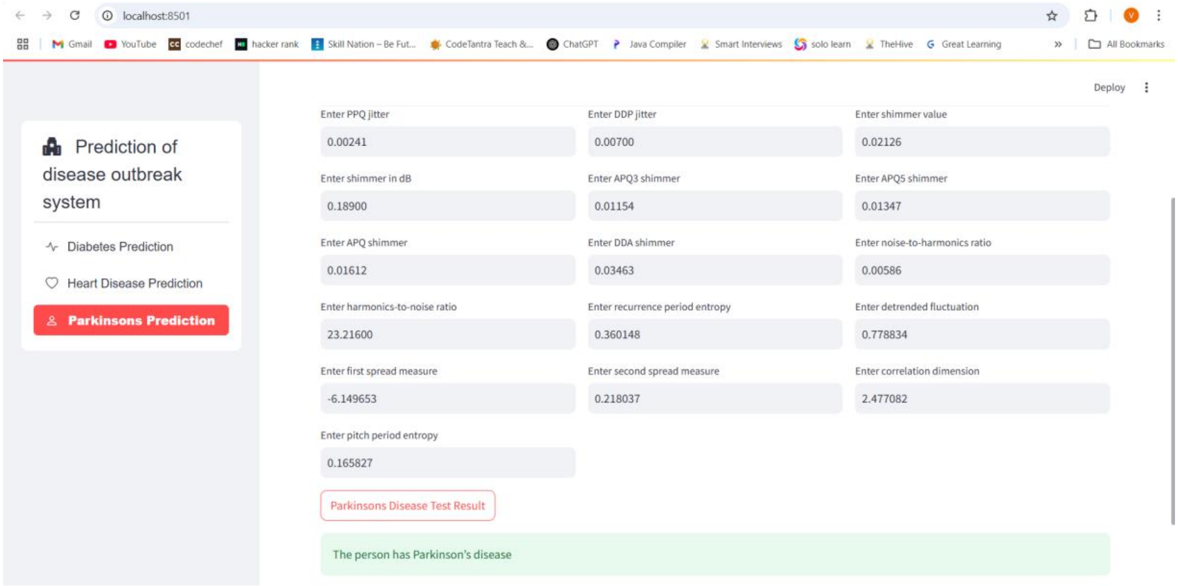

Fig. 6: Person with Parkinson's Disease

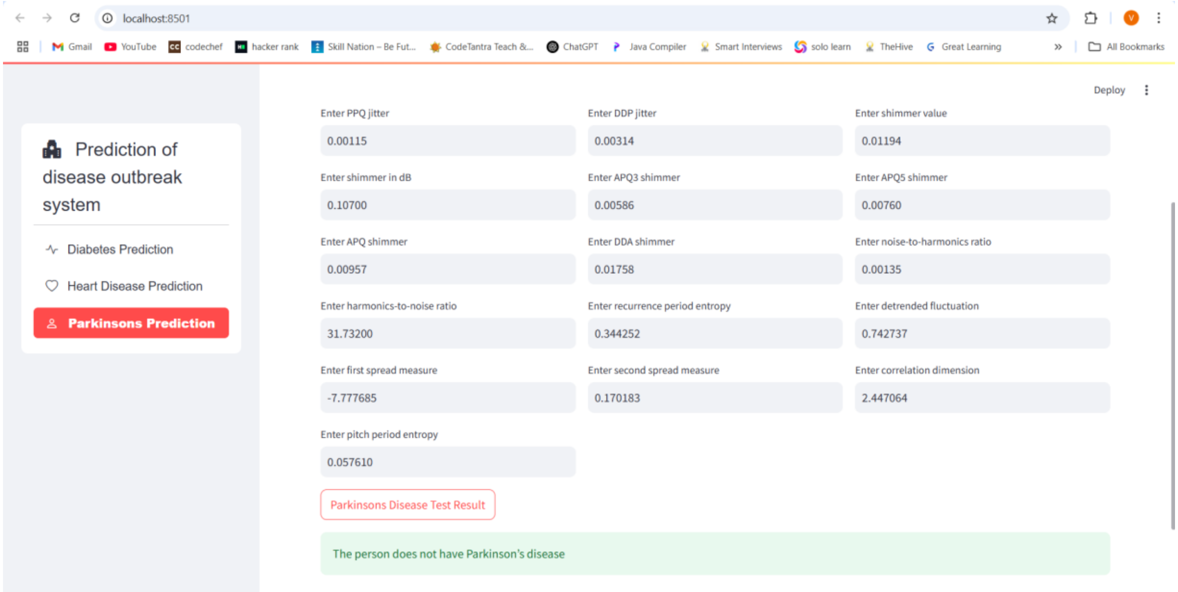

Fig. 7: Person Without Parkinson's Disease

4.2 GitHub Link for Code:

https://github.com/Vedabhavishya/Prediction-of-Disease

CHAPTER 5

Discussion and Conclusion

5.1 Future Work

Advanced Model Implementation:

- In the future, we can improve the system's accuracy by adding advanced models like XGBoost or Neural Networks.
- We can also use hyperparameter tuning techniques to fine-tune the models and get better results.

Real-Time Data & Scalability:

- The system can be upgraded to work with real-time data from health monitoring devices like fitness trackers or smartwatches.
- To make the system more accessible to a larger number of users, it can be deployed on cloud platforms, improving scalability and performance.

5.2 Conclusion

This project successfully created a Disease Prediction System using machine learning to help project successfully created a disease Prediction System using machine learning to stage. We used models like Logistic Regression, Decision Tree, Random Forest, and Support Vector Classifier to make accurate predictions based on patient data.

To improve performance, we handled data imbalances using the Tomek Links technique. Additionally, the system was developed using Streamlit, making it easy for users to cjheck their health risks in real-time with a simple interface.

Overall, this project shows how machine learning can play an important role in healthcare by offering fast, reliable, and accessible disease prediction support.

REFERENCES

[1]. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.

[2]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[3]. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

[4]. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[5]. Tomek, I. (1976). *Two Modifications of CNN*. IEEE Transactions on Systems, Man, and Cybernetics, 6(11), 769–772.

[6]. Microsoft & SAP TechSaksham Initiative. (2023). Al and Machine Learning in Healthcare: Transformative Learning Materials

[7]. Streamlit Documentation. (2024). *Streamlit: The Fastest Way to Build and Share Data Apps*. Retrieved from https://docs.streamlit.io