MACHINE LEARNING

OPEN ENDED LAB

MULTIPLE DISEASE PREDICTION SYSTEM

GROUP MEMBERS	**Khansa Asif(CS-21045)** **Syeda Hafsa(CS-21046)** **Muhammad Bilal (CS-21091)**
COURSE CODE	**CS-324**
SECTION	**A**
SUBMITTED TO	**Miss Mahnoor Malik**

Abstract:

This report presents the development of a Multiple Disease Prediction System using machine learning (ML) techniques. The project involved data collection, preprocessing, feature extraction, model training, evaluation, and deployment. The system was designed to predict the likelihood of diabetes and heart disease in patients, providing a tool for individuals to monitor their health based on their inputs. Using datasets from Kaggle, we implemented Logistic Regression and Support Vector Machines (SVM) in a Google Colab environment. The front-end interface was developed using Streamlit to ensure a user-friendly experience. The results demonstrated that both models could predict diseases effectively, with Logistic Regression showing high accuracy. Future advancements include incorporating additional diseases, using more sophisticated ML algorithms, and enhancing the user interface.

Introduction:

In recent years, the integration of machine learning into healthcare has shown great promise in predictive diagnostics, allowing for earlier and more accurate disease detection. This project, the Multiple Disease Prediction System, aims to leverage machine learning algorithms to predict the presence of diabetes and heart disease. By analyzing patient data inputs, the system provides valuable insights into an individual's health status, potentially aiding in early diagnosis and treatment. This report details the process of developing such a system, from initial data exploration to model deployment, highlighting the technical and practical aspects of the project.

Objective:

The primary objective of this project is to develop a reliable and user-friendly application that predicts the risk of diabetes and heart disease based on patient-provided data.

Development Tools:

To achieve the project objectives, the following tools and libraries were used:

- **Programming Language**: Python
- **Development Environment**: Google Colab, Spyder
- **Libraries**:
    - **Data Processing**: numpy, pandas
    - **Visualization**: matplotlib, seaborn
    - **Machine Learning**: scikit-learn
    - **Utilities**: itertools, pickle, os
    - **Web Interface**: streamlit, streamlit-option-menu
- **Data Source**: Kaggle datasets for diabetes and heart disease

Development Steps:

The development process for the Multiple Disease Prediction System involved several key steps, detailed below:

Data Collection:

Datasets were sourced from Kaggle, specifically targeting diabetes and heart disease. These datasets provided comprehensive features necessary for training robust predictive models.

Data Preprocessing:

Data preprocessing was critical to ensure the quality and usability of the data. Steps included:

- **Cleaning**: Removing duplicates and handling missing values.
- **Encoding**: Converting categorical variables into numerical values suitable for model training.
- **Normalization**: Scaling numerical features to standardize the input data.

Exploratory Data Analysis (EDA):

EDA was performed to uncover insights from the data. Visualizations such as histograms, box plots, and scatter plots were used to explore the distribution of variables and their relationships. This step helped in understanding the underlying patterns and informed feature selection and engineering.

Feature Engineering:

- **Feature Selection**:

Diabetes: We retained key features like Glucose, BMI, Age, etc., for their direct relevance to diabetes prediction.
Heart Disease: We selected significant attributes such as Age, ChestPainType, and MaxHeartRate, based on their medical importance.

- **Feature Transformation**:

Standardization was applied to all features in both datasets to ensure uniform scaling.
For heart disease data, categorical variables like Sex and ChestPainType were numerically encoded to integrate seamlessly with our models

Model Building:

Two machine learning algorithms were chosen for model building:

- **Logistic Regression**: Effective for binary classification with linear decision boundaries and regularization to prevent overfitting.
- **Support Vector Machine (SVM)**: Suitable for high-dimensional spaces and capable of handling non-linear boundaries using kernel functions.

Model Training:

The models were trained using the preprocessed datasets. The training process involved:

- Splitting the data into training and testing sets to evaluate model performance.
- Using cross-validation to ensure the robustness of the models.
- Tuning hyper-parameters to optimize model performance.

Model Evaluation:

Models were evaluated based on their accuracy, precision, recall, and F1 score. Logistic Regression and SVM were compared to determine the best-performing model for each disease prediction task.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression(with python package)	88.52%	87.87%	90.62%	89.23%
Logistic Regression(without python package)	77.00%			
Support Vector Machine(with python package)	77.27%	75.67%	51.85%	61.53%
Support Vector Machine(without python package)	18.18%			

Integration:

The trained models for diabetes and heart disease prediction were integrated into a user-friendly Streamlit application. This web-based interface allows users to easily input their health data and receive immediate predictions.

- **Navigation**: Users can switch between different disease prediction models using a sidebar menu.
- **Input Forms**: Tailored input forms guide users including help option to enter relevant health data within specified ranges, making the process straightforward and user-friendly.
- **Real-Time Predictions**: Users receive instant feedback on their disease risk after submitting their data.
- **Error Handling**: The application provides prompts and validations to ensure data is entered correctly.

Validation:

The application was tested to ensure that it correctly predicts the likelihood of diabetes and heart disease based on user inputs. The system's predictions were validated against known test data to confirm its accuracy and reliability.

Results:

The project yielded the following results:

- **Logistic Regression**: Demonstrated high accuracy in diabetes prediction system. Its ability to handle high-dimensional data and provide clear decision boundaries made it effective for this application.

- **Support Vector Machine (SVM)**: Performed well, particularly in cases with non-linear relationships. Its robustness to overfitting in high-dimensional spaces contributed to its success in heart disease prediction.
- **User Interface**: The Streamlit application provided a user-friendly platform for health monitoring, allowing users to input their data and receive predictions seamlessly.

Future Advancements:

To further enhance the Multiple Disease Prediction System, the following improvements are proposed:

- **Incorporate Additional Diseases**: Expand the system to include predictions for more diseases, broadening its applicability.
- **Use Advanced ML Algorithms**: Implement more sophisticated algorithms such as Gradient Boosting, Neural Networks, or Ensemble methods to potentially improve predictive performance.
- **Enhance User Interface**: Improve the interface to make it more intuitive and accessible, possibly adding more interactive features and visualizations.
- **Real-Time Data Integration**: Allow real-time data input and processing to provide instant predictions and feedback.

Conclusion:

The Multiple Disease Prediction System successfully integrated machine learning algorithms to predict diabetes and heart disease. The use of Logistic Regression and SVM provided effective models for binary classification tasks, with Logistic Regression showing particularly high accuracy. The project demonstrated the potential of ML in healthcare applications, offering a tool for early disease detection and health monitoring.

Snapshots of the Website:

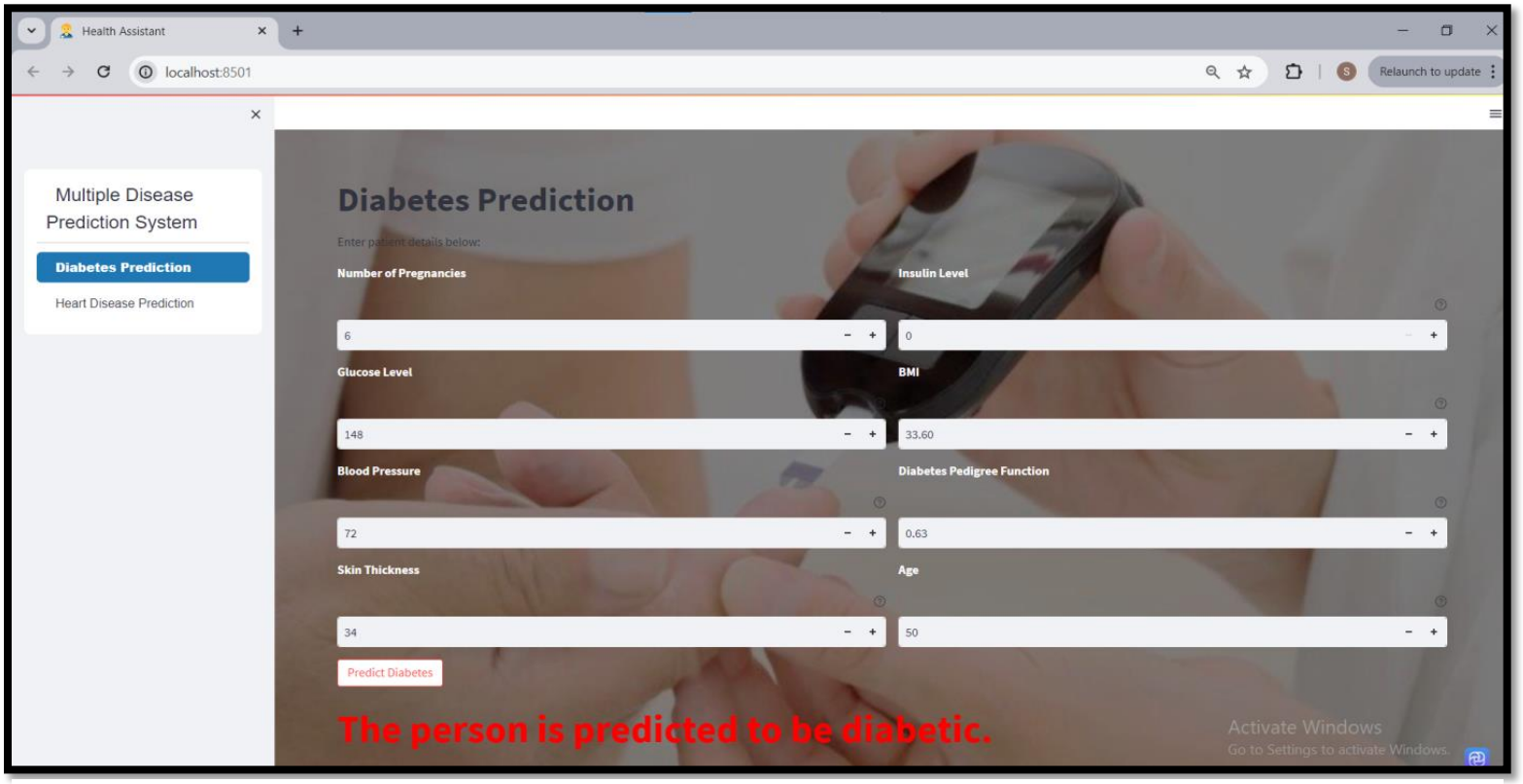

Figure 1

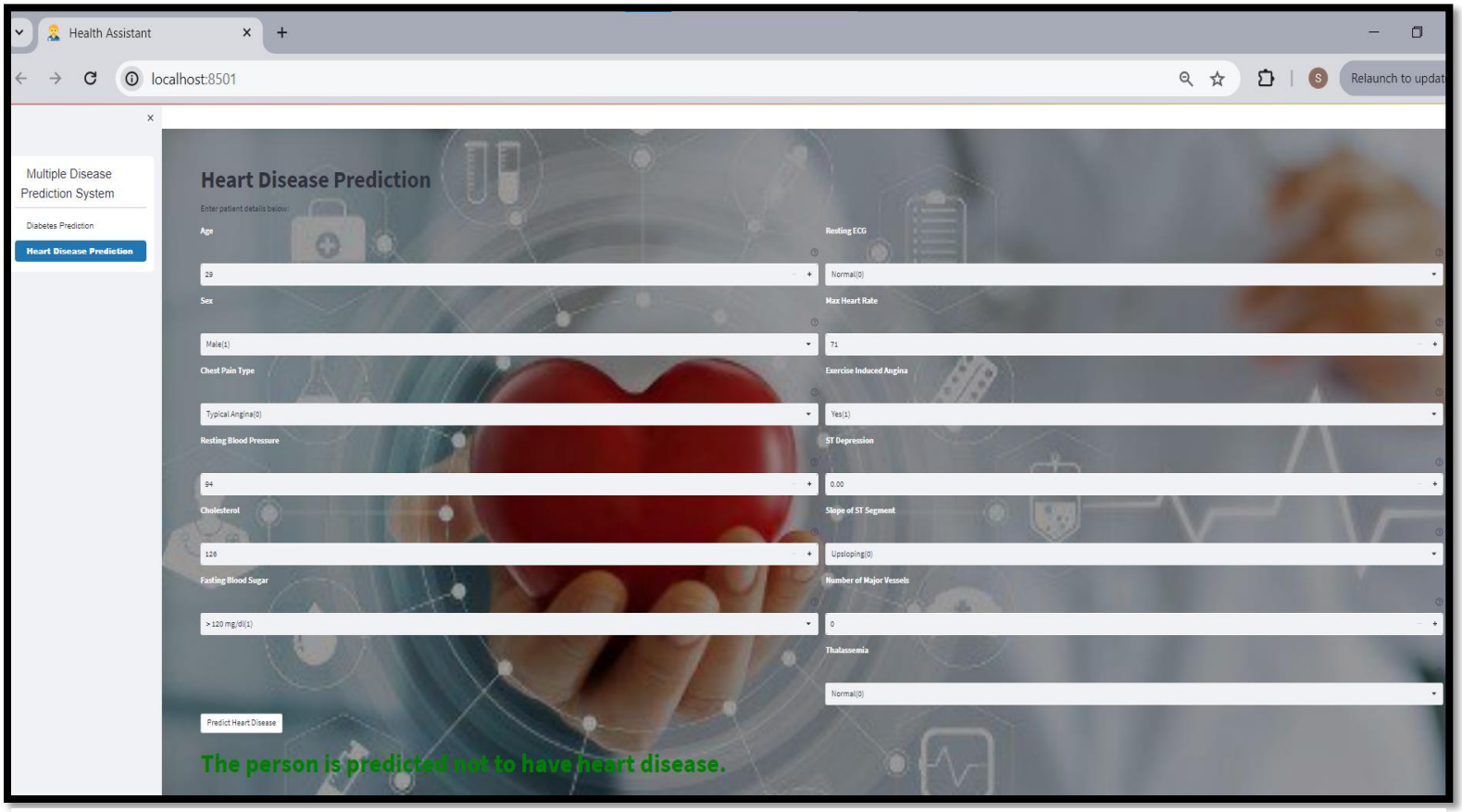

Figure 2