Disease Prediction System

Documentation

Project Overview

The Disease Prediction System is a machine learning-based web application tailored for diagnostic centers. Its primary objective is to assist healthcare professionals in the early detection of diseases such as Diabetes, Heart Disease, and Parkinson's Disease. The system uses patient data, machine learning models, and a web interface to provide reliable and efficient predictions.

Tasks:

- Data Preprocessing: Ensuring clean, consistent data for reliable model training.

- Model Training: Leveraging machine learning algorithms to achieve high accuracy.

- Deployment: Providing a real-time interactive interface via Streamlit.

Project Objectives

- Facilitate early diagnosis of critical diseases to improve treatment outcomes.

- Integrate machine learning models with an intuitive web-based interface.

- Enhance user experience with a simple, real-time prediction system for diagnostic use.

Project Significance

This system demonstrates the use of machine learning and web technologies to aid diagnostic centers in improving their services. The integration of models with a user-friendly application highlights the potential of technology in healthcare:

- Time Efficiency: Provides quick and accurate predictions.

- Scalability: Suitable for deployment in multiple diagnostic centers.

- User-Centric Design: Designed for non-technical healthcare professionals.

Project Scope

Inclusions

1. Diabetes Prediction: Evaluates on different features like:

- **Pregnancies** - Number of times the patient has been pregnant.
- **Glucose** - Plasma glucose concentration (measured after a glucose tolerance test).
- **Blood Pressure** - Diastolic blood pressure (mm Hg).
- **Skin Thickness** - Triceps skinfold thickness (mm).

- **Insulin** - 2-hour serum insulin (μU/ml).
- **BMI** - Body mass index (weight in kg/(height in m^2)).
- **Diabetes Pedigree Function** - A function that scores the likelihood of diabetes based on family history.
- **Age** - Age of the patient.

2. Heart Disease Prediction: Evaluates on different features like:

- **age** - Age of the individual.
- **sex** - Sex of the individual (1 = Male, 0 = Female).
- **cp** - Chest pain type (0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic).
- **trestbps** - Resting blood pressure (in mm Hg).
- **chol** - Serum cholesterol level (in mg/dl).
- **fbs** - Fasting blood sugar > 120 mg/dl (1 = True, 0 = False).
- **restecg** - Resting electrocardiographic results (0 = Normal, 1 = Having ST-T wave abnormality, 2 = Showing probable or definite left ventricular hypertrophy).
- **thalach** - Maximum heart rate achieved.
- **exang** - Exercise-induced angina (1 = Yes, 0 = No).
- **oldpeak** - ST depression induced by exercise relative to rest.
- **slope** - Slope of the peak exercise ST segment (0 = Upsloping, 1 = Flat, 2 = Downsloping).
- **ca** - Number of major vessels (0-4) colored by fluoroscopy.
- **thal** - Thalassemia (1 = Normal, 2 = Fixed Defect, 3 = Reversible Defect).

3. Parkinson's Disease Prediction: Evaluates on different features like:

- **MDVP:Fo(Hz)** - Average fundamental frequency.
- **MDVP:Fhi(Hz)** - Maximum fundamental frequency.
- **MDVP:Flo(Hz)** - Minimum fundamental frequency.
- **MDVP:Jitter(%)** - Variation in fundamental frequency.
- **MDVP:Jitter(Abs)** - Absolute jitter.
- **MDVP:RAP** - Relative amplitude perturbation.
- **MDVP:PPQ** - Five-point period perturbation quotient.
- **Jitter:DDP** - Average absolute difference of differences between cycles.
- **MDVP:Shimmer** - Variation in amplitude.
- **MDVP:Shimmer(dB)** - Amplitude variation in decibels.
- **Shimmer:APQ3** - Three-point amplitude perturbation quotient.
- **Shimmer:APQ5** - Five-point amplitude perturbation quotient.
- **MDVP:APQ** - Eleven-point amplitude perturbation quotient.
- **Shimmer:DDA** - Average absolute difference of differences between amplitudes.
- **NHR** - Noise-to-harmonics ratio.
- **HNR** - Harmonics-to-noise ratio.
- **RPDE** - Recurrence period density entropy.
- **DFA** - Signal fractal scaling exponent.
- **spread1** - Nonlinear measure of fundamental frequency variation.
- **spread2** - Nonlinear measure of variation in amplitude.
- **D2** - Signal nonlinear dynamic complexity measure.
- **PPE** - Pitch period entropy.

4. User Interface: Built using Streamlit for seamless interactions.

Implementation Details

Workflow

1. Data Preprocessing: Cleaning datasets and handling missing values.

2. Model Training:

  - Models: Random Forest Classifier.

  - Training-Testing Split: 80%-20%.

  - Metrics: Accuracy, precision, and recall.

3. Application Development:

  - Streamlit for building the web interface.

  - Joblib for saving and loading models.

Technical Stack

- Programming Language: Python.

- Libraries:

  - Scikit-learn

  - Pandas

  - Streamlit

  - Joblib

Limitations and Challenges

Limitations

1. Data Freshness: Models require periodic retraining for updated accuracy.

2. Scalability: Performance might degrade with a large number of simultaneous users.

Challenges

1. Feature Engineering: Ensuring relevant features for high prediction accuracy.

2. Response Times: Optimizing the Streamlit app for real-time performance.

3. User Experience: Interface for healthcare professionals as they understand the technical terms.

Results and Testing

1. Diabetes Prediction: Achieved 85% accuracy.

2. Heart Disease Prediction: Achieved 88% accuracy.

3. Parkinson's Disease Prediction: Achieved 90% accuracy.

Testing was conducted using both unit and integration tests to ensure reliability.

Conclusion

The Disease Prediction System demonstrates how technology can improve healthcare delivery through efficient and reliable disease detection. Future enhancements include deploying advanced models, enabling multi-language support, and optimizing for large-scale deployments.

Try it!

https://disease-prediction-using-ml.streamlit.app/