“ML-Driven Early Detection for Optimal Health – Empowering You with Accurate Predictive Health Analytics”

SUMMER INTERNSHIP SOPHOMORES PROJECT

SPJ 2001

AP20232410000130

Summer Id: 20240007

Zeroth Review

UNDER THE GUIDANCE OF

Prof. Anil Vitthalrao Turukmane

DEPARTMENT OF SCOPE (Speci. Networking And Security)

SUBMITTED BY

Madhu Alapaka - 21MIS7022.

Kothamasu Kaarthikeya - 21MIS7039.

Pulluru Nikhilesh - 21MIS7087.

DEPARTMENT OF SCOPE(Integ M.TechSpec.in Software Engineering)

VIT-AP UNIVERSITY

Title of the Project	**ML-Driven Early Detection for Optimal Health – Empowering You with Accurate Predictive Health Analytics"**
Software Requirements	**Operating System**: Windows **10/11** (64-bit) **Programming Language**: Python **ML Learning Framework:** ML Modules, Scikit – Learn **IDEs Tool**: Visual Studio Code, Jupyter Notebook. **Version Control System**: Git **Data Manipulation:** NumPy, Pandas **Model Evaluation:** Streamlit **Package Management and Build tools:** Suppose used OpenCv
Hardware Requirements	**Processor**: Intel i5/i7(8th generation or newer recommended for better performance) **Memory (RAM)**: At least 4 GB or More **Storage**: SSD with at least 250 GB free space for software and project files. **Hardware:** Laptop **GPU:** NVIDIA RTX 3080 with 2GB VRAM **Cooling:** High-performance CPU cooler and adequate case fans or an external cooling pad **Ports and Connectivity: Wi-Fi 6** for faster and more reliable wireless connectivity **Keyboard and Touchpad**: Backlit keyboard, Precision touchpad, or external mouse for accurate control with High display(HD)

ABSTRACT

In this project, we develop a comprehensive Multiple Disease Prediction System using Machine Learning techniques in Python, with a deployment interface built on Streamlit. The system aims to provide early detection and risk assessment for multiple diseases, specifically focusing on Diabetes, Heart Disease, Parkinson's Disease etc. Leveraging a diverse array of machine learning algorithms, the system analyzes patient data to predict the likelihood of these conditions, enabling timely medical intervention and personalized healthcare management.

Our approach involves several key stages: data collection and preprocessing, feature engineering, model training and evaluation, and deployment. We gather and clean data from reputable medical datasets, ensuring high-quality input for our predictive models. Through rigorous feature engineering, we identify and transform critical predictors that enhance model accuracy.

We employ various machine learning models tailored to each disease, optimizing them through cross-validation and performance metrics to ensure robust predictions. The final models are then integrated into a user-friendly web application using Streamlit, providing an interactive platform for users to input their health data and receive instant predictions. this project aims to assist individuals and healthcare providers in early disease detection, ultimately contributing to better health outcomes and proactive disease management.

Keywords: Disease Prediction, Machine Learning, Lab report data or Clinical data, Multi-class classification, data/image preprocessing, data augmentation, user-friendly(UI) web application, Support Vector Machine (SVM).

Modules Involved Are:

- **Data Ingestion and Preprocessing**
- **Model Evaluation and Validation**
- **User Interface with Streamlit.**
- **Reporting and Visualization.**
- **Machine Learning Model Building Module.**
- **DiseaseSpecific Predictive Models.**

1. Data Ingestion and Preprocessing:

Data Acquisition Module: Collect raw data from various source

Data Refinement Module: Clean and preprocess data to ensure quality.

Feature Creation Module: Develop new features to enhance model performance.

2. Model Evaluation and Validation:

Validation Engine: Conduct thorough validation of predictive models.

Performance Metrics Module: Calculate and display model performance metrics.

3. User Interface with Streamlit:

Interactive Dashboard: Provide a user-friendly interface for data input and results visualization.

User Input Interface: Facilitate user data entry and parameter adjustments.

Prediction Display: Show prediction outcomes and risk assessments interactively.

4. Reporting and Visualization

The **Reporting and Visualization** module involves generating comprehensive reports on model performance metrics like accuracy, precision, and recall, as well as visualizing data distributions, prediction results, and trends over time. This module also includes audit logs to track user activities and system operations, ensuring transparency and compliance.

5. Machine Learning Model Building Module:

This is the core module where you train the model to predict diseases based on the prepared data. Key steps include:

- **Model Selection:** Choosing appropriate machine learning algorithms like:
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Random Forest
  - Deep Learning models (for complex relationships between features and diseases)
- **Model Training:** Training the chosen model on a portion of the data to learn the underlying patterns between features and disease outcomes.
- **Hyperparameter Tuning:** Adjusting the model's internal parameters to optimize its performance on a separate validation dataset.

6. DiseaseSpecific Predictive Models

- **Heart/Cardio Disease Prediction Module:**

Logistic Regression, Decision Trees, Random Forest, SVM, etc.

Specific risk factors: cholesterol levels, blood pressure, age, smoking status, etc.

- **Diabetes/Gluco Prediction Module:**

Logistic Regression, KNearest Neighbors, Random Forest, etc.

Specific risk factors: BMI, glucose levels, insulin levels, age, etc.

- **Kidney/Nephro Disease Prediction Module:**

Logistic Regression, SVM, Random Forest, etc.

Specific risk factors: serum creatinine, glomerular filtration rate (GFR), blood pressure, etc.