"Disease Prediction System using Machine Learning”

Jayanth.R

Github link

15.08.2024

1. Introduction

Project Overview:

The "Disease Prediction System using Machine Learning" project aims to develop an intelligent system that predicts the likelihood of a person having a particular disease based on various health-related features. The system utilizes machine learning algorithms to analyze historical health data and provide predictions, contributing to early disease detection and proactive healthcare management.

Objectives:

1. **Data Collection:**
   - Gather a diverse dataset containing relevant health features such as age, gender, BMI, blood pressure, cholesterol levels, and family medical history.
2. **Data Preprocessing:**
   - Perform data cleaning and preprocessing to handle missing values and outliers.
   - Normalize or standardize features to ensure a consistent scale.
3. **Feature Selection:**
   - Employ feature selection techniques to identify the most influential variables for disease prediction.
   - Ensure that selected features contribute significantly to the accuracy of the machine learning models.
4. **Model Development:**
   - Explore and implement machine learning algorithms including Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM).
   - Evaluate and compare the performance of different models using metrics like accuracy, precision, recall, and F1-score.

5. **Cross-Validation:**

    - Implement cross-validation techniques to assess the generalization performance of the models and mitigate overfitting.

6. **Hyperparameter Tuning:**

    - Fine-tune hyperparameters of selected models to optimize their performance.

2. Data Sources

Dataset Description

The dataset used in this project consists of various health-related features and the target variable for predicting heart disease. The dataset contains 303 samples and 14 features. The features are:

1. **age**: Age of the patient (int64)
2. **sex**: Gender of the patient (1 = male, 0 = female) (int64)
3. **cp**: Chest pain type (1-4) (int64)
4. **trestbps**: Resting blood pressure (mm Hg) (int64)
5. **chol**: Serum cholesterol level (mg/dl) (int64)
6. **fbs**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false) (int64)
7. **restecg**: Resting electrocardiographic results (0-2) (int64)
8. **thalach**: Maximum heart rate achieved (int64)
9. **exang**: Exercise induced angina (1 = yes, 0 = no) (int64)
10. **oldpeak**: Depression induced by exercise relative to rest (float64)
11. **slope**: Slope of the peak exercise ST segment (1-3) (int64)
12. **ca**: Number of major vessels (0-3) colored by fluoroscopy (int64)
13. **thal**: Thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect) (int64)
14. **target**: Presence or absence of heart disease (1 = defective heart, 0 = healthy heart) (int64)

Data Summary

- **Total Records**: 303
- **Features**: 14
- **Non-Null Count**: All features have 303 non-null values, indicating no missing data in the dataset.

Feature Description

- **Age**: Numeric feature representing the patient's age.
- **Sex**: Binary feature representing the patient's gender.
- **Chest Pain Type (cp)**: Categorical feature describing the type of chest pain experienced.
- **Resting Blood Pressure (trestbps)**: Numeric feature representing the resting blood pressure in mm Hg.
- **Cholesterol (chol)**: Numeric feature indicating the serum cholesterol level.
- **Fasting Blood Sugar (fbs)**: Binary feature indicating whether the fasting blood sugar level is greater than 120 mg/dl.
- **Resting Electrocardiographic Results (restecg)**: Categorical feature describing the results of the resting electrocardiogram.
- **Maximum Heart Rate Achieved (thalach)**: Numeric feature showing the highest heart rate achieved.
- **Exercise Induced Angina (exang)**: Binary feature indicating the presence of exercise-induced angina.
- **Oldpeak**: Numeric feature showing the depression induced by exercise relative to rest.
- **Slope of Peak Exercise ST Segment (slope)**: Categorical feature describing the slope of the peak exercise ST segment.
- **Number of Major Vessels (ca)**: Numeric feature indicating the number of major vessels colored by fluoroscopy.
- **Thalassemia (thal)**: Categorical feature describing the type of thalassemia present.
- **Target**: Binary feature indicating the presence (1) or absence (0) of heart disease.

3. Methodology

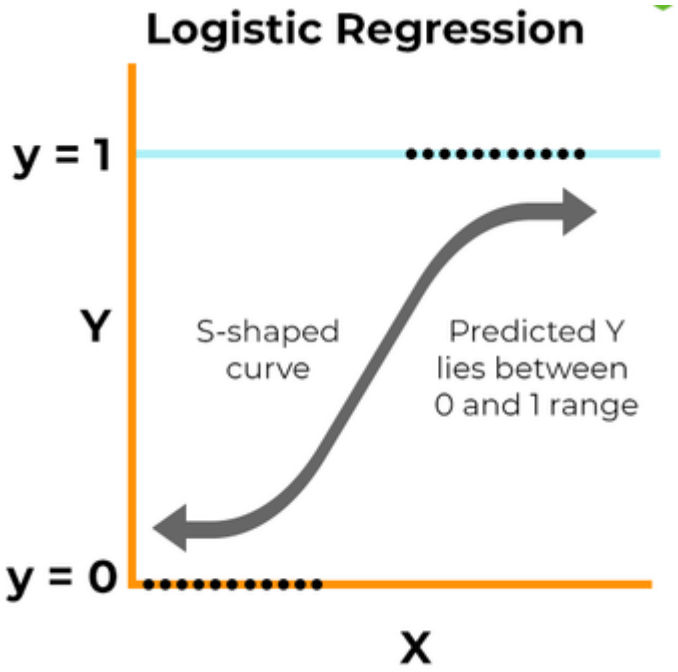
Data Processing

1. **Loading Data**: The dataset is loaded into a Pandas DataFrame for analysis.
2. **Exploratory Data Analysis (EDA)**: Initial exploration includes checking the dataset's shape, data types, and statistical measures. Distribution of the target variable is also examined.
3. **Handling Missing Values**: The dataset is checked for missing values, though in this case, there are none.
4. **Feature and Target Separation**: The dataset is divided into features (independent variables) and target (dependent variable).
5. **Model Training**: The divided dataset is used to train the models
6. **Model Evaluation**: Evaluate the model using metrics like Accuracy score and Cross Validation.
7. **Hyperparameter tuning**: Find the optimal parameters for the model to obtain optimal results/
8. **Building a Predictive System**: Build a Predictive System to allow users to enter values into the model for Disease predictions.

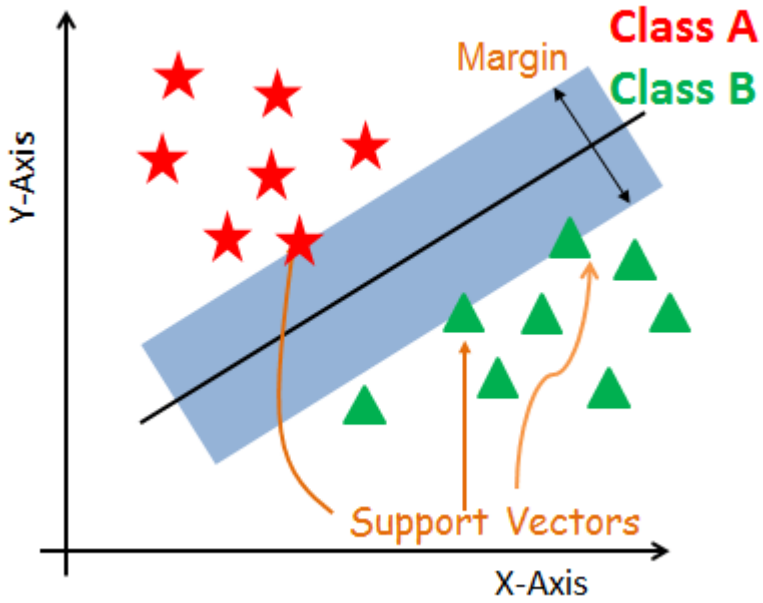
4. Model Architecture

The disease prediction system uses several machine learning models for classification:

1. **Logistic Regression**: A statistical model used for binary classification based on the relationship between independent variables and the binary outcome.



2. **Support Vector Machine (SVM)**: A classification algorithm that finds the hyperplane that best separates the classes.



3. **Decision Tree Classifier**: A model that uses a tree-like graph of decisions to classify the data.
4. **Random Forest Classifier**: An ensemble learning method that combines multiple decision trees to improve classification accuracy.

5. Model Training and Evaluation

1. **Splitting Data**: The dataset is split into training and testing sets, with 80% of data used for training and 20% for testing.
2. **Training Models**: Each model (Logistic Regression, SVM, Decision Tree, Random Forest) is trained on the training set.
3. **Evaluation Metrics**: Models are evaluated based on accuracy scores on both training and testing data. Cross-validation is used to assess model performance across multiple folds.
4. **Hyperparameter Tuning**: Grid search is used to find the best parameters for the Random Forest, Decision Tree, and SVM models.

6. Performance Metrics

- **Logistic Regression**: Provides a balance between training and testing accuracy.
- **SVM**: Shows strong generalization with high accuracy on the test set.
- **Decision Tree**: Tends to overfit the training data, resulting in lower performance on the test set.
- **Random Forest**: Achieves the highest accuracy among all models and shows the best performance overall.

7. Best Model Selection

Based on accuracy and cross-validation results, the Random Forest model is selected as the best performing model for this dataset.

```
Random Forest Accuracy on Test Data: 0.819672131147541
Decision Tree Accuracy on Test Data: 0.7377049180327869
SVM Accuracy on Test Data: 0.8032786885245902
```

8. Instructions for Using the Prediction System

1. **Data Input**: Ensure that the input data is in the same format as the training data (with the same features and data types).
2. **Model Loading**: Load the trained model (e.g., Random Forest) for predictions.
3. **Prediction**: Use the model to predict the likelihood of heart disease based on new patient data.
4. **Interpret Results**: The output will indicate whether the patient is predicted to have a defective heart (1) or a healthy heart (0).

9. Conclusion

The **Disease Prediction System** developed in this project effectively leverages machine learning techniques to predict heart disease based on patient data. Through a systematic approach involving data preprocessing, model training, evaluation, and hyperparameter tuning, the Random Forest model emerged as the most accurate and reliable classifier for this task.

Key takeaways include:

- The system can assist healthcare professionals by providing quick and accurate predictions, potentially aiding in early diagnosis and treatment planning.
- The Random Forest model's high accuracy and robustness make it the preferred choice for deployment in a clinical setting.
- The model's performance can be further improved with more data, feature engineering, and advanced tuning techniques.

This project demonstrates the power of machine learning in solving complex medical problems and opens the door for future enhancements, such as integrating more sophisticated models or expanding the system to predict other types of diseases. With continued development, the **Disease Prediction System** could become a valuable tool in the healthcare industry, contributing to better patient outcomes and more efficient healthcare delivery.