A1: Proposal of a PRML Problem - Description, Motivation, Characterization, Dataset

Description:

Our project focuses on developing a Predictive Risk Model for Heart Disease (PRMHD) using Pattern Recognition and Machine Learning (PRML). The model aims to analyze a range of medical and lifestyle factors to predict the likelihood of an individual developing heart disease.

Motivation:

Heart disease is a leading cause of death worldwide. Early diagnosis can lead to effective treatment, but traditional diagnostic methods are often slow and expensive. A computational model can provide quick, accurate, and cost-effective risk assessments.

Characterization:

The problem is multi-dimensional, involving variables like age, cholesterol levels, blood pressure, etc. It is a classification problem where the outcome is binary: either the individual is at risk of heart disease or not.

Dataset:

The dataset for our Predictive Risk Model for Heart Disease (PRMHD) is a comprehensive collection of medical and lifestyle variables. It includes demographic information such as age and sex, along with medical metrics like exercise-induced angina, number of major vessels, types of chest pain, resting blood pressure, cholesterol levels, fasting blood sugar, and electrocardiographic results. Additionally, it captures the maximum heart rate achieved during stress tests. The dataset is designed to be binary in its outcome, categorizing individuals as having either a higher or lower chance of experiencing a heart attack.

The Cleveland Heart Disease dataset from the UCI Machine Learning Repository was selected for this project. Although the database contains a total of 76 attributes, our focus is on a subset of 14 key attributes, as these are the ones most commonly cited in published research. Specifically, we are utilizing only the Cleveland database for this endeavor. The initial steps involve data cleaning to address any missing values, followed by Exploratory Data Analysis (EDA) to gain insights into the dataset.

A2: Goals, Questions to be Investigated, Decision-Making

The primary goal of this project is to identify patterns that signify a high risk of developing heart disease. Building upon this, we aim to develop a machine learning model capable of predicting the likelihood of heart disease with high accuracy. Ultimately, the objective is to seamlessly integrate this predictive model into existing healthcare systems, thereby aiding medical professionals in their decision-making processes.

Several key questions guide this project. First, we seek to identify the most significant factors contributing to heart disease, which will inform the feature selection for our machine learning model. Second, we aim to evaluate the model's accuracy in predicting heart disease, as this will be crucial for its clinical applicability. Lastly, we are interested in exploring how the model can adapt to new data for continually improved predictions over time.

Model Validation:

The model will be validated using a variety of techniques:

- Cross-Validation: To ensure that the model performs well on unseen data.
- Precision, Recall, and F1 Score: To measure the model's performance in terms of both false positives and false negatives.
- ROC Curve and AUC: To evaluate the model's ability to distinguish between positive and negative classes.

The model will be trained on a large dataset and validated rigorously. Once validated, it can be used to predict the likelihood of heart disease in new patients, thereby aiding in early diagnosis and treatment planning.

The model is designed to serve as a streamlined decision-support tool for healthcare providers. It will classify patients into either a high-risk or low-risk category based on various health and lifestyle factors. This binary classification will enable medical professionals to make quick, informed decisions regarding the need for further diagnostic tests or immediate intervention.

A3: PRML Design Steps, Pattern Recognition, Machine Learning Algorithms

The project aims to predict heart disease, which inherently is a Pattern Recognition and Machine Learning (PRML) problem. The design steps for a PRML solution typically involve data collection, feature extraction, model training, and evaluation. In our case, the dataset comprises various features like age, sex, blood pressure, and cholesterol levels, among others, which are essential indicators of heart health. The objective is to recognize patterns in these features that are indicative of the likelihood of heart disease.

The project involves identifying intricate patterns in medical data that are indicative of heart disease. These patterns could be linear or non-linear relationships between variables like age, cholesterol levels, and blood pressure. The ability to recognize these patterns is crucial for the successful prediction of heart disease, making it a pattern recognition problem.

Machine learning algorithms will be employed to learn these patterns from the data automatically. The model will be trained to generalize from the training data to unseen cases, adapting and improving its predictive power over time. This iterative learning process qualifies it as a machine learning problem.

The project is not just about identifying patterns (Pattern Recognition) or making predictions (Machine Learning); it's about doing both in a synergistic manner. The pattern recognition techniques will feed into the machine learning algorithms, providing a comprehensive solution to a complex problem.

The model will use pattern recognition to identify complex relationships among variables like age, blood pressure, and cholesterol levels, thereby recognizing patterns indicative of heart disease.

Machine Learning Algorithms:

- Logistic Regression: Logistic Regression is chosen for its simplicity and effectiveness in binary classification. It serves as a computationally inexpensive and interpretable baseline model.
- Random Forest: Random Forest is used for its ability to handle multiple features and provide feature importance. It offers robustness against overfitting and is useful for understanding key predictors.
- Support Vector Machines (SVM): SVM is selected for its strength in high-dimensional spaces. It is robust to outliers and can capture complex relationships in the data.

Why These Algorithms Are Suited:

These algorithms are well-suited for classification problems and have been proven effective in medical diagnosis scenarios. They offer a good mix of simplicity and complexity, allowing for both interpretability and high accuracy.

References:

Ahsan, M. and Siddique, Z., 2021. "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review". Artificial Intelligence in Medicine, 2022, Article ID 102289. DOI: 10.1016/j.artmed.2022.102289. PDF.

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S.D. and Singh, P., 2021. "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning". Computational Intelligence and Neuroscience, 2021, Article ID 8387680. DOI: 10.1155/2021/8387680. PDF.

Fatima, M. and Pasha, M., 2017. "Survey of Machine Learning Algorithms for Disease Diagnostic". Journal of Intelligent Learning Systems and Applications, 9(1), pp.1-12. DOI: 10.4236/JILSA.2017.91001. PDF.

Rahmanpritom, R., 2021. "Heart Attack Analysis & Prediction Dataset". Kaggle. Available at: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.