Respondent Cardiovascular Classification

Danilo Babić SV42/2020

1. Motivation

Cardiovascular diseases are a group of disorders characterized by damage to the heart and blood vessels. Understanding and accurately classifying cardiovascular diseases is crucial for medical diagnostics and treatment planning. Early detection and precise classification can lead to more effective treatments and improved patient outcomes. This work addresses the problem of classifying cardiovascular diseases by categorizing subjects from the dataset into those who have any type of cardiovascular disease and those who do not, essentially tackling a binary classification problem.

2. Research questions

The specific problem addressed in this work is the classification of cardiovascular diseases using a dataset consisting of 70,000 samples and 13 features. Each sample represents an individual participant and includes attributes such as age, height, gender, weight, systolic and diastolic blood pressure, cholesterol and glucose levels, along with indicators of alcohol and cigarette consumption, and physical activity. Out of these features, 6 are numeric.

The dataset underwent preprocessing to address various issues. The 'ID' feature, deemed irrelevant for model training, was removed, as was the 'presence of cardiovascular diseases' feature, which was reserved for model evaluation purposes. Invalid values were also handled: heights below 120 cm were excluded, while weights were constrained using a calculated average BMI threshold for individuals of average height (165 cm). Negative and extreme values of systolic and diastolic blood pressure were adjusted: negative values were converted to their absolute counterparts, and extreme values (above 370 mmHg for systolic and below 50 mmHg, above 360 mmHg for diastolic and below 20 mmHg) were removed from the dataset.

After preprocessing, the dataset now contains 67,776 samples and 11 features, ready for further analysis and classification tasks aimed at improving the understanding and management of cardiovascular diseases through machine learning techniques.

3. Related work

Machine learning algorithms play a pivotal role in analysing large-scale datasets, such as those containing medical records of individuals susceptible to cardiovascular diseases. One of the foundational methods used is **logistic regression**, which models the probability of a binary outcome based on input features. Its simplicity and interpretability make it an attractive choice for initial exploratory analyses and as a benchmark against more complex models.

Decision trees and their ensemble counterpart, **random forests**, are also widely applied in cardiovascular diseases classification tasks. Decision trees partition the dataset into subsets based on feature values, making them adept at capturing nonlinear relationships. Random forests aggregate multiple decision trees, reducing variance and improving generalization performance. They excel in

handling high-dimensional data and are robust to noise and outliers commonly found in healthcare datasets.

Support vector machines (SVM) offer another powerful approach for binary classification. SVMs find the optimal hyperplane that separates classes by maximizing the margin between them. They are effective in scenarios where clear boundaries exist between classes, leveraging kernel functions to handle nonlinear relationships in data.

In recent years, **deep learning** techniques, particularly **neural networks**, have gained prominence in healthcare analytics.

Practical implementation of these algorithms requires careful consideration of data preprocessing, feature selection, and model evaluation. Preprocessing steps often involve handling missing values, normalizing data, and addressing outliers, as seen in the earlier discussion of cleaning the CVD dataset. Feature engineering plays a crucial role in extracting informative signals from raw data, enhancing the predictive power of models.

Model evaluation is critical to ensuring the reliability and generalizability of classification results. Techniques such as cross-validation and performance metrics like accuracy, precision and recall provide insights into model effectiveness and help optimize parameters.

4. Methodology

In addressing the complex task of classifying cardiovascular diseases (CVDs) through machine learning, a systematic and methodical approach was adopted to leverage data preprocessing, feature engineering, dimensionality reduction, and model selection techniques. This essay will provide a detailed explanation of how Logistic Regression and Random Forest algorithms were utilized to achieve accurate and insightful predictions.

The initial step involved loading and thoroughly inspecting the dataset (cardio_train.csv). Through exploratory data analysis (EDA), anomalies such as negative values in blood pressure (ap_hi and ap_lo) and unrealistically low heights were identified and systematically removed to ensure data quality and reliability in subsequent analyses.

To prepare the dataset for modeling, numerical features underwent standardization using StandardScaler. This preprocessing step is crucial as it normalizes the data, mitigating the influence of varying scales and improving the convergence and performance of machine learning algorithms.

A critical phase of the process involved exploring correlations among features using Seaborn's heatmap visualization. This analysis provided valuable insights into the relationships between variables, guiding decisions on feature selection and helping to identify potential multicollinearity issues.

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining essential information. By plotting the cumulative explained variance ratio, an optimal number of principal components (n_components=7) was determined to effectively capture significant variance in the data.

Random Forest was selected for its robustness and ability to capture complex interactions and nonlinear relationships within the dataset. This ensemble learning method aggregates multiple decision trees, each trained on different subsets of the data with randomly selected features. By combining predictions from multiple trees, Random Forest reduces overfitting and enhances prediction accuracy. It proved particularly effective in identifying intricate patterns in the data that might be overlooked by simpler model like Logistic Regression.

5. Discussion

Both models underwent rigorous evaluation using metrics such as accuracy, precision and recall. Cross-validation techniques ensured the reliability and generalizability of the models by assessing performance across multiple folds of the dataset. Hyperparameter tuning, especially prominent in Random Forest through parameters like n_estimators and max_depth, optimized model performance to achieve the best possible outcomes.

Logistic Regression showed moderate performance across key metrics: sensitivity (recall) at 0.651, precision at 0.724, accuracy at 0.703, and an F1 score of 0.686. On the other hand, Random Forest Tree demonstrated slightly superior results: sensitivity of 0.701, precision of 0.714, accuracy of 0.711, and an F1 score of 0.707. These metrics indicate Random Forest's effectiveness in classification tasks, particularly with improved recall while maintaining comparable precision. Overall, Random Forest Tree's marginally higher F1 score (0.707) compared to Logistic Regression (0.686) underscores its superior balance between precision and recall, crucial for accurate classification tasks.

6. References

- https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data

Note: ChatGPT helped format some of the sentances above, I chose everything I wanted to write, he just wrote it better. ☺