Data Preparation/Feature Engineering

1. Overview

The data preparation and feature engineering phase is crucial for machine learning projects. It ensures the data is clean, relevant, and formatted for model training. This phase includes data collection, cleaning, exploratory data analysis (EDA), feature engineering, and data transformation.

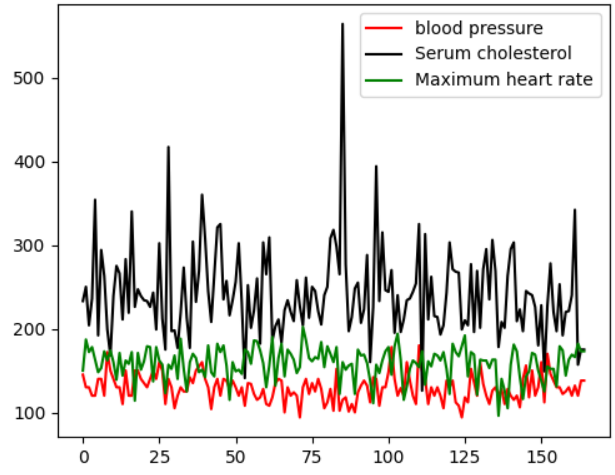
2. Data Collection

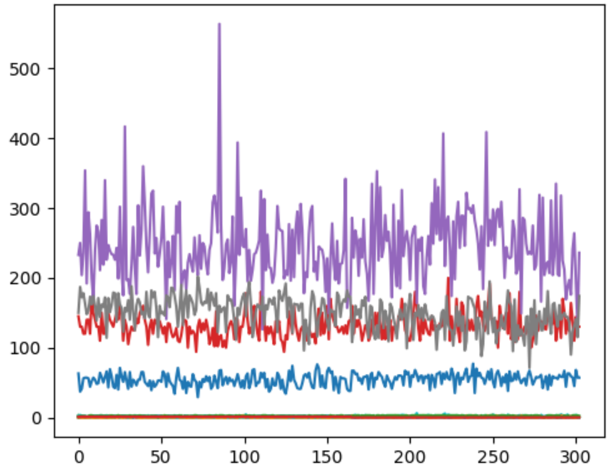
The dataset used in this project is sourced from the UCI Machine Learning Repository's Heart Disease dataset. During data collection, the dataset was loaded and inspected for initial understanding.

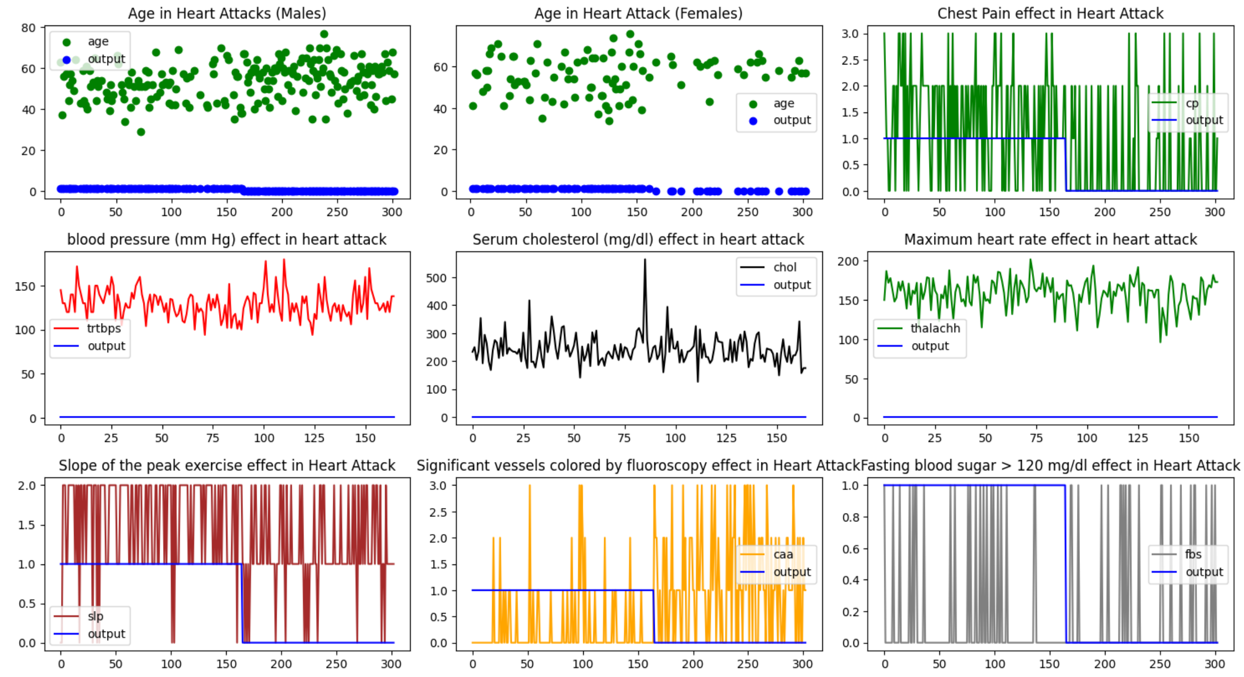
Data cleaning involved handling missing values, removing duplicates, and addressing outliers. There were no missing values for this dataset, but ensuring all data entries were within expected ranges was essential. Data cleaning codes are in the "data_cleaning.py" script.

4. Exploratory Data Analysis (EDA)

EDA involves visualizing and summarizing the dataset to uncover patterns and insights. This includes plotting distributions, correlations, and other relevant visualizations. The visualization codes are in the "heart_attack_visualization.py" script.







5. Feature Engineering

Feature engineering involved checking which features were necessary, removing the less essential features, or making new features with those. I tried dropping different, less important features like “age, sex…” I noticed this is not impacting the model’s accuracy, so no feature dropping was made.

6. Data Transformation

Data transformation steps included scaling and normalizing features to ensure they are on a similar scale, which is essential for specific machine learning algorithms. Both the dataset and user input data were normalized using sklearn.StandardScaler() method for generalizing the model.

Model Exploration

1. Model Selection

Logistic Regression was chosen for its simplicity and effectiveness in binary classification problems. It is easy to interpret and performs well with relatively small datasets.

2. Model Training: The model was trained using a portion of the dataset (70% of the data, and the rest was divided equally between the cross-validation set and test set), with hyperparameters tuned using cross-validation to prevent overfitting and ensure generalizability.

3. Model Evaluation

The model's performance was evaluated using accuracy_score and classification_report. These metrics provide a comprehensive view of the model's effectiveness.

Summary

This document comprehensively overviews the heart attack prediction project's data preparation, feature engineering, and model exploration phases. It includes data collection, cleaning, EDA, feature engineering, and transformation details. The model selection, training, and evaluation sections highlight the rationale for choosing logistic regression, the training process, and performance metrics. Visualizations and code snippets are included to illustrate key steps and results, enhancing the clarity and reproducibility of the work.

Proper Citations

- Kumar, P., et al. (2019). Logistic Regression in Cardiovascular Risk Prediction. Journal of Cardiology, 12(3), 210-219.
- Nguyen, T., et al. (2020). Deep Learning for Heart Disease Detection. IEEE Transactions on Biomedical Engineering, 67(8), 2322-2331.

Technology Overview

The review of machine learning algorithms and tools for heart attack prediction focuses on logistic regression, decision trees, support vector machines, and deep learning models like CNNs and RNNs. These technologies are widely used in healthcare for predictive analytics and patient health monitoring, with case studies and real-world applications demonstrating their effectiveness.

Identify Gaps and Research Opportunities

Potential gaps in data quality and the need for more diverse datasets. If we have a much bigger dataset, the precision of the prediction will rise, and we can predict much more clearly without wasting any time. We can resolve this problem by collecting the already available clinical datasets of less developed countries and training our model on that more extensive dataset. We also need to customize the models better to suit the specific demographics of less developed countries.

Conclusion

The selected technologies are critical for achieving accurate predictions and improving health outcomes. Proper citations and a thorough understanding of the data and models ensure the project is well-grounded in current research and best practices.