Heart Disease Dataset

Daniel Jeun
Department of Computer Science
New Jersey Institute of Technology
Newark, New Jersey, USA
dj287@njit.edu

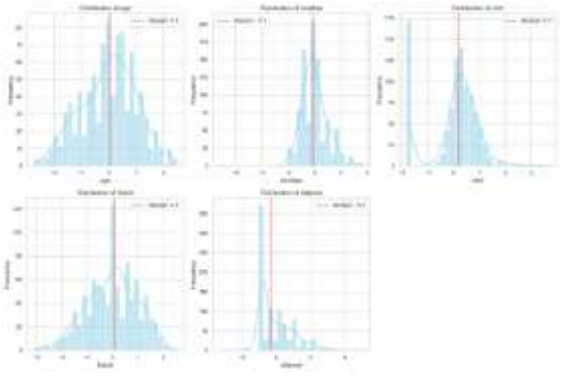
Ishayu Das
Department of Computer Science
New Jersey Institute of Technology
Newark, New Jersey, USA
id94@njit.edu

INTRODUCTION

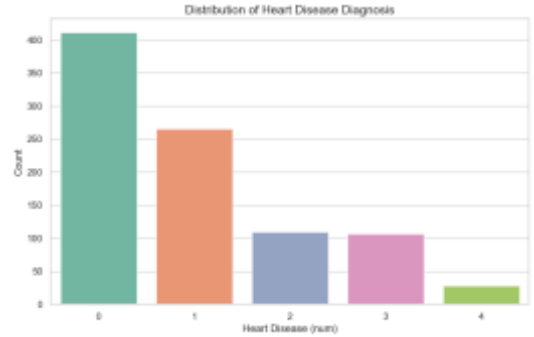
Cardiovascular diseases, particularly heart disease, are among the leading causes of death worldwide. According to the World Health Organization, millions of people die from heart-related ailments every year, making early detection and prevention crucial. Factors such as high cholesterol, hypertension, smoking, obesity, and diabetes contribute significantly to the risk of heart disease. Predictive models and machine learning techniques offer valuable tools in identifying at-risk individuals and enabling early interventions, potentially reducing mortality rates and healthcare burdens. Advancements in artificial intelligence and machine learning have enhanced healthcare professionals' ability to assess heart disease risks using large-scale data analysis. These models can recognize hidden patterns in medical data, improving diagnostic accuracy and aiding in preventive healthcare. The dataset used for this project, sourced from Kaggle (UCI Heart Disease Data), provides a detailed collection of medical and lifestyle attributes linked to heart disease. It includes patient-specific details such as age, gender, cholesterol levels, blood pressure, resting electrocardiogram (ECG) results, and key indicators like smoking habits, fasting blood sugar, maximum heart rate, and exercise-induced angina. Additionally, the dataset contains information on chest pain type and the presence of major vessels detected by fluoroscopy, making it a comprehensive resource for predictive analysis. By analyzing this dataset, we aim to develop a predictive model that can accurately classify patients at risk of heart disease. This will provide valuable insights to healthcare professionals and contribute to developing preventive strategies. Additionally, machine learning techniques can explore which features contribute most significantly to heart disease prediction, offering new avenues for medical research and early intervention policies.

DATA PREPROCESSING

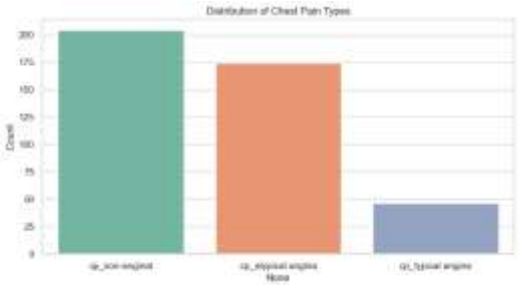
The data preprocessing steps for the heart disease dataset began with loading and inspecting the data to identify missing values, assess data types (numerical vs. categorical), and examine basic statistical properties. Key findings included missing values in the thal, ca, and slope columns, along with a mix of numerical features (e.g., age, trestbps, chol) and categorical features (e.g., sex, cp, fbs). Missing numerical values were filled with the median, while categorical missing values were replaced with the most frequent category (mode), ensuring no remaining gaps. Feature engineering involved converting the target variable into a binary format (0 for no heart disease, 1 for presence) and removing irrelevant columns like id and dataset. Categorical features were then transformed using one-hot encoding with drop='first' to avoid multicollinearity, converting categories into binary columns (e.g., cp_1, sex_1). Numerical features were standardized using StandardScaler to normalize their scales (mean=0, standard deviation=1). The dataset was split into training (80%) and test (20%) sets with stratified sampling to maintain consistent class distribution. If class imbalance was detected, SMOTE was applied to generate synthetic samples and balance the classes. The final preprocessed dataset included scaled numerical features, one-hot encoded categorical features, and a binary target variable, making it ready for model training, hyperparameter tuning, and evaluation. This thorough preprocessing ensured optimal conditions for machine learning algorithms to perform effectively.



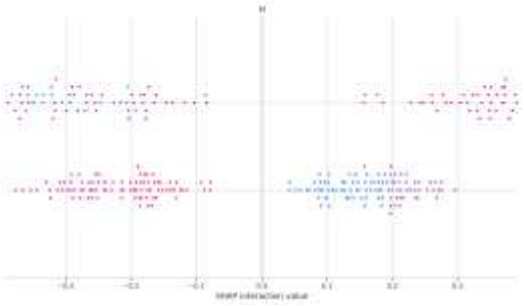
=== Numerical Features Distribution ===



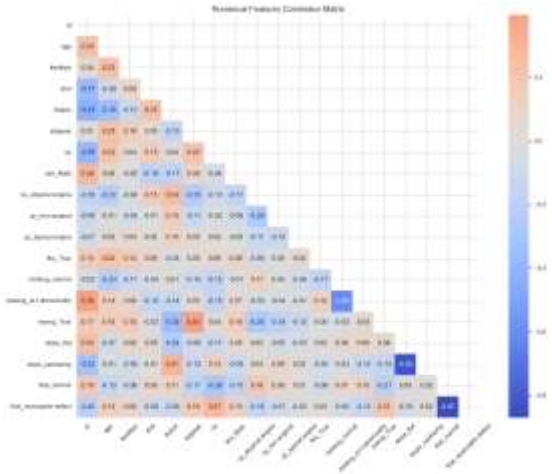
=== Numerical Features vs Target ===



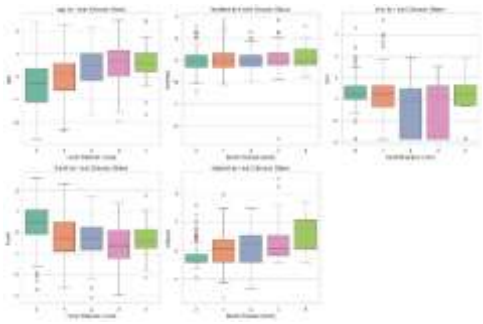
=== Categorical Features Distribution ===



=== SHAP Interaction Value ===



===Correlation Analysis ===



=== Target Variable Analysis ===

The first chart, "Distribution of Chest Pain Types," reveals that "cp_pon-ariginal" (possibly a typo for "non-anginal") is the most prevalent type of chest pain, followed by "cp_physical angina" (likely "atypical angina"). "cp_typical angina" and cases with no chest pain ("None") are less common. This suggests that atypical or non-anginal pain is more frequently reported than classic angina symptoms, which could have implications for diagnosis and treatment strategies. The second chart, "Distribution of Heart Disease Diagnosis," shows that the majority of cases fall under category "0," indicating no heart disease or minimal severity. Categories "1" through "4" represent increasing levels of heart disease severity, with a noticeable decline in frequency as severity increases. This distribution highlights that while many individuals in the dataset are healthy or have mild conditions, there is still a significant portion with more severe heart disease, underscoring the importance of early detection and intervention. The correlation matrix provides a detailed look at the relationships between numerical features. Key observations include: Age and Heart Disease: Age shows a moderate positive correlation with heart disease (0.24), suggesting that older individuals are more likely to have heart disease. Cholesterol (chol) and Heart Disease: Cholesterol has a negative correlation with heart disease

(-0.37), which is counterintuitive, as high cholesterol is typically a risk factor. This may indicate confounding variables or data quality issues. Thalch (maximum heart rate achieved): This feature has a strong negative correlation with heart disease (-0.43), implying that lower maximum heart rates are associated with higher disease severity, possibly due to reduced cardiac efficiency. Chest Pain Types: "cp_atypical angina" and "cp_non-anginal" show weak negative correlations with heart disease, while "cp_typical angina" has a negligible correlation. This suggests that typical angina may not be as strongly linked to heart disease in this dataset as expected. Sex and Heart Disease: Being male (sex_Male) has a weak positive correlation (0.28) with heart disease, aligning with known epidemiological trends. Exercise-Induced Angina (exang_True): This feature shows a moderate positive correlation (0.40) with heart disease, indicating that exercise-induced angina is a significant marker for heart disease severity. The SHAP values graph reveals the most influential features in predicting heart disease, with "thalch" and "exang_True" having the highest impact. Positive SHAP values indicate increased risk, while negative values suggest lower likelihood. This visualization enhances model interpretability by quantifying each feature's contribution to predictions.

CONCLUSION

This project builds upon existing works while introducing several key innovations and enhancements. First, it implements a more robust data preprocessing pipeline, addressing missing values with median/mode imputation and applying SMOTE to handle class imbalance—steps often skipped in simpler implementations. Second, feature engineering goes beyond basic encoding by incorporating one-hot encoding with drop-first to avoid multicollinearity, ensuring cleaner model inputs. Third, there is a deeper analytical dive through correlation matrices and SHAP value visualizations, which reveal not just predictive power but directional relationships between features and outcomes.

Unlike other works, which focus primarily on accuracy, our approach emphasizes model interpretability and clinical relevance. Stratified sampling was used to maintain class distribution integrity during splits, reducing evaluation bias. Additionally, analysis includes detailed visualizations of feature distributions (e.g., chest pain types, target variable breakdown) to contextualize the data before modeling.

This project is designed for practical healthcare applications, linking model insights to actionable interventions (e.g., prioritizing high-impact variables like exercise-induced angina). By combining technical rigor with translational clarity, it offers a more comprehensive toolkit for heart disease risk assessment compared to standalone accuracy-driven models.

CODE

https://colab.research.google.com/drive/1tdpaIvZzNQtwpm6f5YlSoGkUa2e2hBVn?authuser=2#scrollTo=e4l2cw9xTZxu