Heart Disease Prediction: A Comparative Study of Naive Bayes and Linear Regression Models

Problem Definition

The core problem tackled in this assignment is binary classification: determining whether a patient has heart disease based on various medical attributes and test results. This prediction task is particularly valuable in clinical settings where early identification can lead to timely interventions and improved patient outcomes. The challenge lies in effectively processing mixed data types (continuous and categorical variables) in the model inputs and also handling real-world data quality issues such as missing values and variability in readings.

Dataset Description

The analysis utilizes the UCI Heart Disease Dataset, specifically the Cleveland data, which contains 303 patient records with 14 clinical attributes. The dataset includes both demographic information (age, sex) and clinical measurements such as chest pain type, resting blood pressure, serum cholesterol levels, maximum heart rate achieved, and results from various cardiac tests.

The target variable represents the presence of heart disease on a scale of 0-4, where 0 indicates no disease and values 1-4 represent increasing severity levels. For this binary classification task, the target was transformed to distinguish between no disease (0) and disease present (1-4). The dataset presents typical real-world challenges including missing values in the 'ca' and 'thal' columns, along with varying scales across numerical features that require preprocessing attention.

Data Cleaning and Preprocessing

Handling Missing Values

The initial data exploration revealed missing values in two categorical features: 'ca' and 'thal', originally encoded as '?' symbols in the dataset. An approach was taken to address these gaps, prioritizing data retention over removal given the relatively small dataset size of 303 samples. (usual for medical data)

Approach and Reasoning:
The missing values were handled using feature-wise mode imputation. For each affected

column, missing entries were replaced with the most frequently occurring value within that specific feature. This method was chosen over row-wise imputation or deletion as this helped us retain:

1. Statistical Validity: Column-wise mode imputation maintains the natural distribution of each feature, ensuring that the most common category within each variable serves as a reasonable substitute for missing data.
2. Data Preservation: With only 6 rows containing missing values, deletion would have resulted in minimal data loss. However, imputation was preferred to maintain the full sample size, which is applied in medical datasets where each patient record represents valuable information.

- Before imputation: 'ca' column contained 5 missing values, 'thal' column contained 2 missing values
- After imputation: Both 'ca' and 'thal' columns showed zero missing values, with mode values (0.0 for 'ca' and 3.0 for 'thal') filling the gaps

Addressing Data Quality Issues

The boxplot analysis revealed the presence of outliers in several continuous variables, particularly in cholesterol levels, resting blood pressure, and maximum heart rate. However, the number of extreme values was relatively small (max in a single feature was 6/303) and appeared to be legitimate medical measurements rather than data entry errors.

Outlier Treatment Decision:
Rather than removing outliers, which could eliminate clinically relevant cases of patients with genuinely extreme values, a conservative approach was adopted. The decision was made to retain all data points as they likely represent real medical conditions that could provide valuable information for model training. This approach acknowledges that in medical data, extreme values hold diagnostic importance.

Feature Engineering

Standardization and Scaling

The continuous numerical features (age, trestbps, chol, thalach, oldpeak) were standardized using StandardScaler to achieve zero mean and unit variance. This transformation was essential for several reasons:

Scale Differences: The original features had vastly different ranges - age (29-77), cholesterol (126-564), and resting blood pressure (94-200) - which could lead to features with larger scales dominating the model learning process.

1. Algorithm Requirements: Both Ridge and Lasso regression require standardized features to ensure uniform penalty across all variables. Without standardization,

features with larger scales would receive disproportionately smaller penalty coefficients.

2. Model Performance: Standardization improves convergence for gradient-based algorithms and ensures that distance-based calculations used in classification techniques treat all features equally.

Categorical Variable Encoding

The dataset contained several categorical variables that required applicable encoding strategies based on their characteristics:

Binary Variables: Features like 'sex', 'fbs' (fasting blood sugar), and 'exang' (exercise-induced angina) were already optimally encoded as 0/1 and required no additional transformation.

Ordinal Variables: Variables such as 'cp' (chest pain type), 'restecg' (resting ECG results), and 'slope' (ST segment slope) maintained their existing integer encoding as they represent meaningful ordered categories reflecting disease severity or physiological states.

Nominal Variables: The 'thal' (thalassemia type) variable was identified as nominal since its values (3, 6, 7) represent different types of thalassemia without inherent ordering. This variable was processed using one-hot encoding to ensure order and to prevent the model from learning ordinal relationships which don’t exist.

The dataset is small (303 samples) and adding many new features risks overfitting. Existing features are already clinically meaningful. So, I avoided creation of new features from existing data.

Model Training and Comparison

Training Setup

The modeling approach compared two fundamentally different machine learning paradigms: generative (Naive Bayes) versus discriminative (Linear Regression) methods. A train-test split (80-20) ensured balanced representation of both classes across training and testing sets, with a fixed random state guaranteeing reproducible results across all model comparisons. (This was stratified)

Naive Bayes Implementation

Model Selection: GaussianNB was chosen over BernoulliNB due to the mixed nature of the feature set, which includes both continuous measurements and categorical variables. GaussianNB handles continuous features naturally by and assumes normal distributions for each class. Hence, this choice

Smoothing Parameter Exploration: Two alpha values were systematically tested:

- $\alpha = 1.0$
- $\alpha = 0.01$ (minimal laplace smoothing)

Results Analysis:
The comparison revealed interesting patterns in smoothing effectiveness:

- Naive Bayes with $\alpha = 0.01$: Accuracy = 0.869, ROC-AUC = 0.954
- Naive Bayes with $\alpha = 1.0$: Accuracy = 0.853, ROC-AUC = 0.930

The lower smoothing parameter ($\alpha = 0.01$) showed better performance, suggesting that the training data provides reliable probability estimates and that minimal smoothing allows the model to better capture the underlying feature patterns.

Linear Regression Models

Implementation Strategy: Linear regression was adapted for binary classification using a 0.5 probability threshold. Three variants were implemented to explore regularization effects:

1. Standard Linear Regression: Base model without regularization
2. Ridge Regression: L2 regularization testing α values [0.1, 1.0, 10.0] (custom grid created)
3. Lasso Regression: L1 regularization testing α values [0.01, 0.1, 1.0](custom grid created)

Performance Results:

- Linear Regression: Accuracy = 0.869, ROC-AUC = 0.938
- Ridge Regression best values ($\alpha = 0.1$): Accuracy = 0.869, ROC-AUC = 0.938
- Lasso Regression ($\alpha = 0.01$): Accuracy = 0.853, ROC-AUC = 0.936

Evaluation Results

The evaluation employed multiple metrics to assess model performance from different perspectives:

Performance Summary:
All models achieved strong performance, with accuracy scores ranging from 0.853 to 0.869 and ROC-AUC scores between 0.930 and 0.954. The best performing model was Naive Bayes with $\alpha = 0.01$, achieving 86.9% accuracy and showed good recall (96.4%) for detecting heart disease cases.

From a clinical perspective the best model showed,

- High recall (96.4%): Successfully identifies nearly all heart disease cases, minimizing dangerous false negatives

- Moderate specificity (78.8%): Correctly identifies most healthy patients, though some false positives occur
- Only 1 false negative: Good performance for medical screening where missing disease cases has negative consequences
- 7 false positives: Manageable number of healthy patients requiring additional screening.

Discussion

Model Performance Interpretation

The results demonstrate that both Naive Bayes and Linear Regression approaches achieve comparable performance on this heart disease prediction task. There is a narrow performance gap showing that the dataset's properties are well-suited to both generative and discriminative approaches.

Smoothing Effects in Naive Bayes:
The superior performance of minimal smoothing ($\alpha = 0.01$) indicates that the dataset provides sufficient samples for reliable likelihood estimation. The 1.6% accuracy improvement over standard smoothing suggests that the training data is representative and that excessive regularization actually lowers model performance by over-smoothing genuine data patterns.

Regularization Impact in Linear Models:
Ridge regression method showed no improvement over standard linear regression, suggesting that correlation among feature values is not a significant issue in this dataset. Lasso regression's feature selection limiting to 12 from 16 provides model explainability while maintaining competitive performance, even though it has a slightly lower accuracy than the linear model.

Limitations and Considerations

Dataset Size: The relatively small dataset (303 samples) limits the complexity of models that can be effectively trained and may affect general assumptions to broader populations.

Feature Engineering: While only basic preprocessing was performed, more complex feature engineering (interaction terms, domain-specific transformations) might improve performance.

Clinical Validation: The models would require extensive clinical validation before deployment in real healthcare settings such as testing on diverse patient samples.

Trade-offs seen: While Lasso regression provides feature selection benefits, the reduction in model complexity comes at the cost of slightly decreased predictive performance.

Practical Implications

The high recall achieved by the best-performing model makes it particularly suitable for screening applications where the primary goal is to avoid missing potential heart disease cases. The moderate specificity level suggests that the model would be most effective as part of a two-stage assessment process, where positive predictions indicate further diagnostic testing rather than immediate treatment.

AI Tool Usage Disclosure

Tools Utilized

This analysis benefited from several AI-powered tools to enhance development efficiency and code quality:

1. Claude (Anthropic): Primary AI tool for coding and debugging
2. ChatGPT (OpenAI): Code review and reduction.

AI Contributions

- Visualization Creation: Assisted in generating matplotlib and seaborn code for creating comprehensive evaluation visualizations including confusion matrices, ROC curves, and performance heatmaps as shown.
- Debugging Support: Helped identify and resolve issues with data type conversions, missing value handling, and sklearn parameter configurations.

Original Contributions

- Exploratory Data Analysis : Interpretation of data quality issues, feature relationships, and preprocessing decisions. Went through the dataset manually and ran boxplots, missing value analysis.
- Model Selection: Reasoning for choosing GaussianNB over BernoulliNB and choice of regularization parameter selection in each case.
- Context: Analysis of results within medical context, including recall/true negative trade-offs and deployment method in a practical scenario.
- Critical Evaluation: Assessment of model limitations, potential improvements, and applicability.