Heart Disease Prediction Using Machine Learning

Ethan Alexander

Project Overview

This project focuses on using machine learning techniques to predict heart disease through regression analysis. The workflow involved data preprocessing, model development, evaluation, and visualization to improve predictive accuracy and gain insights into the effectiveness of various models.

Data Exploration and Preprocessing

- **Dataset Analysis**:
  - Loaded a heart disease dataset and calculated feature correlations with the target variable to identify the four most informative features.
- **Data Cleaning and Transformation**:
  - Addressed missing values and performed encoding for categorical variables.
  - Applied min-max scaling and z-score normalization to ensure consistent feature scaling.
- **Data Splitting**:
  - Divided the dataset into training, validation, and testing sets using a 60:20:20 ratio, with a random seed of 42 for reproducibility.

Linear Regression to Predict Cholesterol Levels

- **Initial Model Performance**:
  - **Mean Squared Error (MSE)**: > 2000
  - **R^2 Score**: < 1%
- **Improvements**:
  - Applied normalization techniques and L1/L2 regularization.
  - Enhanced model performance, increasing R^2 to approximately 10%.
- **Visualization**:
  - Created scatter plots comparing actual vs. predicted cholesterol levels to evaluate model accuracy visually.

Logistic Regression for Heart Disease Prediction

- **Model Design**:
    - Utilized logistic regression with a polynomial degree of 2 to predict the likelihood of heart disease.
- **Key Observations**:
    - Increasing the polynomial degree (e.g., from 2 to 3) improved model fit.
    - However, degrees higher than 5 caused overfitting, which reduced generalizability on unseen data.

Polynomial Regression for Heart Disease Prediction

- **Experimentation**:
    - Conducted polynomial regression using a degree of 4 and set `max_iter` to 200 for optimization.
- **Comparison**:
    - Polynomial regression outperformed logistic regression due to its ability to model complex relationships between features and the target variable.
    - Achieved a higher proportion of correct predictions, showcasing its effectiveness for this dataset.

Key Insights and Takeaways

1. **Feature Selection**: Identifying and using the most relevant features significantly improved model performance.
2. **Normalization and Regularization**: These techniques reduced overfitting and enhanced the linear regression model's predictive power.
3. **Polynomial Regression**: By modeling complex patterns, polynomial regression provided the most accurate predictions but required careful tuning to avoid overfitting.

This project demonstrates the potential of machine learning in predictive health diagnostics, highlighting the importance of preprocessing and model selection to achieve meaningful results.