Singular Value Decomposition from Scratch and Its Applications

Abstract

This project explores machine learning techniques for predicting heart disease using the Cleveland dataset from the UCI repository. The dataset was cleaned and preprocessed to handle missing values, categorical variables, and skewed distributions. A variety of models including Logistic Regression, Random Forest, XGBoost, SVM (linear and RBF), and LightGBM were trained and evaluated. Logistic Regression emerged as one of the most effective models with an accuracy of 88.52

1 Introduction

Heart disease remains one of the leading causes of death globally. Early detection can help in timely intervention and treatment. This project aims to develop a machine learning model to predict the presence of heart disease using patient data from the Cleveland Heart Disease dataset available on the UCI Machine Learning Repository.

2 Dataset Description

The Cleveland dataset contains 14 attributes including demographic, clinical, and diagnostic features. The original target variable ranges from 0 to 4, representing increasing severity of disease. For binary classification, it was transformed into a binary target where:

- `0` – Absence of heart disease
- `1` – Presence of heart disease (i.e., `target > 0`)

2.1 Features:

- age, sex, cp (chest pain type), trestbps (resting blood pressure), chol (cholesterol), fbs (fasting blood sugar), restecg (resting ECG), thalach (max heart rate), exang (exercise induced angina), oldpeak (ST depression), slope, ca (number of vessels colored), thal (thalassemia)

3 Exploratory Data Analysis (EDA)

- Missing values were found in the 'ca' and 'thal' columns, filled using mode.
- Target class distribution: ˜54.1% with no disease, ˜45.9% with disease.
- Histograms showed mostly normal distributions for continuous variables except for 'oldpeak', which was right-skewed.
- IQR based outlier analysis revealed few outliers: 23 in 'cp', 9 in 'trestbps', 5 in 'chol', 45 in 'fbs', 1 in 'thalach', and 5 in 'oldpeak'; others had none.

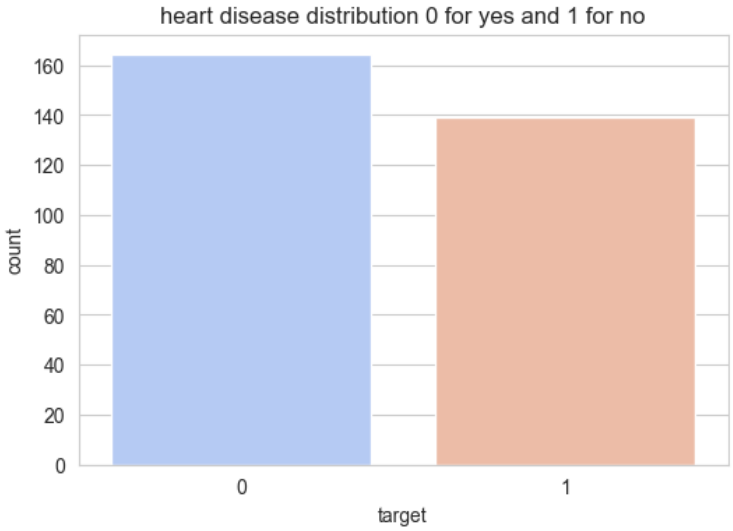

Figure 1: Target Class Distribution

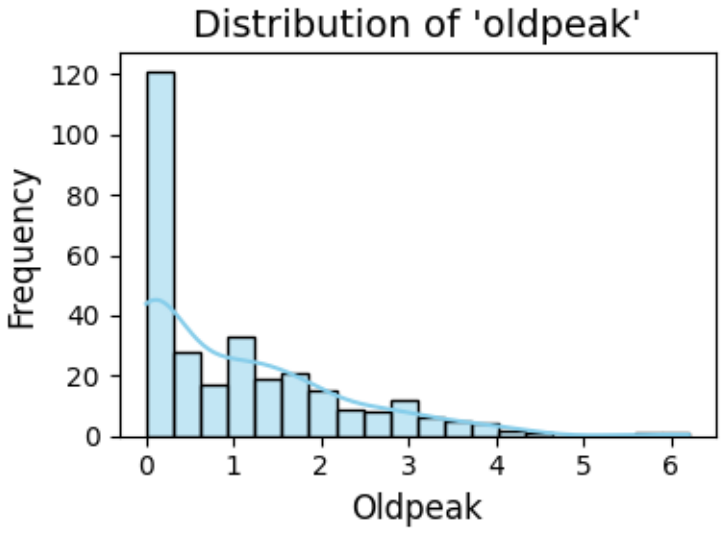

Figure 2: Age Distribution

4 Data Preprocessing

- Replaced '?' with NaN and filled missing values in 'ca' and 'thal' using mode, followed by type conversion to integers.
- One-hot encoded categorical variables: 'thal', 'ca', 'restecg', and 'slope'.
- Standardized numerical features: 'age', 'chol', 'thalach', and 'trestbps' using StandardScaler.
- Outlier detection using Z-score showed minimal impact, so no removal was performed.

5 Modeling

80:20 train-test split was used. Several models were evaluated:

5.1 Modeling Overview and Feature Engineering Decisions

The primary goal was to predict the presence of heart disease using various machine learning models and derive clinical insights. To improve model performance and interpretability, several preprocessing and modeling strategies were explored:

- **Log Transformation:** The `oldpeak` feature (ST depression) exhibited strong right skew.Log transformation was applied to normalize it. While this improved Logistic Regression accuracy slightly, it reduced the performance of Random Forest and XGBoost. Therefore, retained the original values with standard scaling.
- **Feature Removal - Age:** The `age` feature ranked lowest in the feature importance chart from Random Forest. Removing it, however, led to reduced accuracy across all models and was thus retained.
- **Feature Removal - Cholesterol (chol):** Interestingly, removing `chol` improved performance in all models — especially Random Forest, which achieved the highest accuracy post-removal. This suggests that `chol` may have introduced noise or redundant information. Supporting this, outlier analysis revealed that `chol` had the highest number of outliers among all features. These extreme values could have negatively impacted models sensitive to data distribution, such as Logistic Regression and XGBoost.

5.2 Logistic Regression

- Achieved a test accuracy of 88.52%, Precision = 84%, Recall = 93%, F1 Score = 88%.
- Important predictors: **ST depression (oldpeak)**, **chest pain type (cp)**, and **maximum heart rate (thalach)**.

- Patients with typical angina (low cp) and high exercise capacity (thalach) were generally classified as low risk, while higher ST depression (oldpeak) increased the predicted risk.
- Coefficients from the model guided interpretation, confirming medical literature.

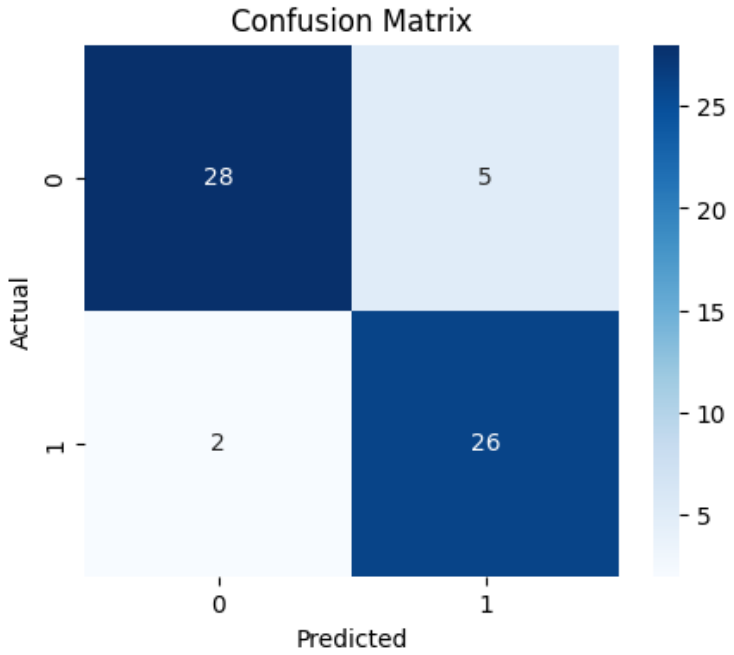

Figure 3: Confusion Matrix: Logistic Regression

5.3 Random Forest

- **Best-performing model**: Test accuracy of **92%**, Precision of **85%**, Recall of **100%**, F1 Score of **92%**.
- Tree-based feature importance confirmed the significance of `cp`, `thalach`, and `oldpeak`.
- Post-processing without the `chol` feature gave the highest gains.

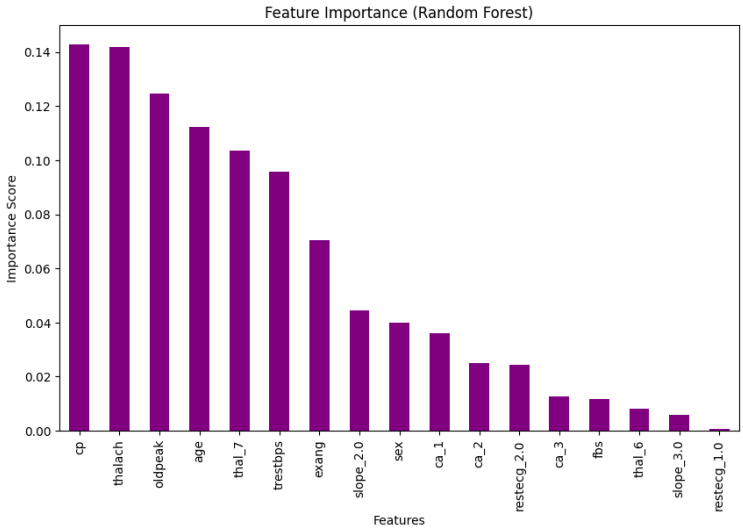

Figure 4: Random Forest Feature Importance

5.4 XGBoost and LightGBM

- **XGBoost:** Accuracy = 84.0%, Precision = 78%, Recall = 89%, F1 Score = 83%.
- **LightGBM:** Accuracy = 82.0%, Precision = 74%, Recall = 93%, F1 Score = 83%.
- Performance was acceptable but consistently below that of Random Forest.
- LightGBM showed slightly higher recall, but at the cost of precision.

5.5 Support Vector Machines (SVM)

- **RBF Kernel:** Accuracy = **90.16%** (on par with Logistic Regression), outperforming the Linear kernel (85.25%).
- SHAP analysis revealed key features: `oldpeak`, `thalach`, and `cp`.
- SVM was included as a baseline model. Despite a small dataset (303 rows, 14 features), SVM with the RBF kernel effectively captured non-linear patterns, proving valuable in evaluation and comparison.

6 Hyperparameter Tuning

To ensure optimal performance, hyperparameter tuning was conducted using GridSearchCV or randomized search depending on model complexity.

6.1 Logistic Regression

- **Best Parameters:** C=10, penalty='l1'
- Test Accuracy after tuning:86.89%

6.2 Random Forest

- **Best Parameters:** {max_depth=10, min_samples_leaf=2, min_samples_split=10, n_estimators=100}
- Test Accuracy after tuning: 90.16%

6.3 Model Comparison Across Preprocessing Variants

Multiple models were evaluated — Logistic Regression, Random Forest, and XGBoost — under three main settings:

- **Log-transformed 'oldpeak'** (rest features scaled normally)
- **Standard scaling without log-transform**, with all features
- **Feature removal experiments** (removing 'age' or 'chol')

Log-transforming only 'oldpeak' resulted in decreased accuracy for scale-sensitive models like Logistic Regression and XGBoost. However, tree-based Random Forest was largely unaffected, showcasing its robustness to feature transformations.

Table 1: Performance Summary Across Different Preprocessing Pipelines

Model Variant	Accuracy	Precision	Recall	F1 Score
RF (log oldpeak)	0.93	0.88	**1.00**	0.93
RF (all features)	0.92	0.87	0.96	0.92
RF (no age)	0.87	0.83	0.89	0.86
RF (no chol)	0.92	0.85	**1.00**	0.92

The best-performing model was the **Random Forest** trained on the dataset with standard scaling and the 'cholesterol' feature removed. This model achieved an accuracy of 0.92 and a perfect recall, making it suitable for high-stakes medical screening tasks where missing a diagnosis could be critical.

Clinical Note: Although a Recall of 1.00 may lead to some false positives, it is often acceptable in clinical settings to ensure all at-risk patients are flagged for follow-up testing. With a high precision of 0.85, this model balances caution with reliability.

7 Final Evaluation

The final evaluation was conducted on the version of the dataset with standard scaling applied and the `chol` feature removed. Results are summarized below:

- **Random Forest (Best Model)**: Accuracy = **92%**, Precision = 85%, Recall = 100%, F1 Score = 92%
- Logistic Regression: Accuracy = 88.52%, Precision = 84%, Recall = 93%, F1 Score = 88%
- XGBoost: Accuracy = 84.0%, F1 Score = 83%
- SVM (RBF): Accuracy = 90.16

8 Conclusion

Multiple models were benchmarked, with Random Forest emerging as the most accurate and balanced model, especially after the removal of the `chol` feature. It achieved a perfect recall score, which is critical in medical diagnosis scenarios.

While Logistic Regression and SVM offered strong interpretability and matched Random Forest in accuracy, the ensemble-based Random Forest model provided more robust generalization.

SHAP analysis further improved transparency by aligning key model insights with established clinical indicators like ST depression and heart rate response.

Future work may explore more complex approaches or training the model on larger clinical datasets.