Predicting Heart Disease Using Logistic Regression and Neural Networks

Adam Lopaška

Project Repository: GitHub Repository

Abstract

Cardiovascular diseases are a leading cause of mortality worldwide, making early prediction essential. This project aimed to develop machine learning models specifically Logistic Regression and Neural Networks to predict heart disease based on clinical indicators. Principal Component Analysis (PCA) was also applied to explore the impact of dimensionality reduction in terms of model performance. The results demonstrate unchanged performance for Logistic Regression with and without PCA, while the Neural Network achieved better results without PCA.

1 Introduction

Heart disease prediction is vital in medical diagnostics as early detection can save lives and reduce healthcare costs. This project aimed to predict the presence of heart disease using machine learning models based on clinical measurements such as blood pressure, cholesterol levels, and ECG results.

The dataset, containing 303 patient records, includes five features: age, resting blood pressure, cholesterol, maximum heart rate and ST depression. The target variable indicates whether a patient has heart disease (1) or not (0). The dataset is relatively balanced, with 54.46% positive and 45.54% negative samples. An 80%-20% train-test split was used for model evaluation.

This study investigates the performance of Logistic Regression and Neural Networks on the original dataset and its PCA-transformed version to compare their effectiveness in predicting heart disease.

2 Context

Machine learning models, including Logistic Regression and Neural Networks, have been widely used for heart disease prediction, with promising results.

In first study [5], researchers compared Logistic Regression, SVM, and Neural Networks on the Cleveland Heart Disease Dataset. Their findings highlighted the effectiveness of simpler models like Logistic Regression and SVM in achieving high accuracy, while Neural Networks provided flexibility for capturing complex patterns.

In second study [2], deep learning models were shown to outperform traditional algorithms for heart disease prediction. Neural Networks, in particular, demonstrated superior accuracy, showcasing the potential of deep learning in medical diagnostics.

These studies emphasize the importance of model selection and the potential benefits of dimensionality reduction, such as PCA, in enhancing prediction accuracy and reducing computational complexity.

3 Methods

This section outlines the methods used to preprocess the data, machine learning models and metrics to evaluate their performance.

3.1 Data Preprocessing

The dataset was preprocessed to ensure optimal performance of the machine learning models:

- **Normalization:** Feature values were normalized to a range between 0 and 1. This was necessary because the features, such as age, cholesterol levels, and blood pressure, are measured on different scales. Normalization ensures that all features contribute equally to the model's learning process [6].
- **Train-Test Split:** The dataset was split into a training set (80%) and a testing set (20%) to evaluate model performance on unseen data.

3.2 Methodology

Principal Component Analysis (PCA) To explore the effect of dimensionality reduction on model performance, Principal Component Analysis

(PCA) was applied. PCA is a technique used to reduce the number of features in a dataset while retaining as much information as possible. It works by identifying new axes, called principal components, that capture the maximum variability in the data. These components are linear combinations of the original features and are ranked based on the amount of variance they explain.

In this project, PCA was used to transform the dataset into a smaller set of uncorrelated features while retaining 91% of the original variance. This process helps eliminate redundant information and can potentially improve the performance of machine learning models by reducing noise and computational complexity. The use of PCA allowed us to test whether a simpler representation of the data could enhance model performance without sacrificing accuracy [4].

Two machine learning models were used to predict heart disease:

Logistic Regression: was chosen because it is a widely used and well understood machine learning model, especially in medical fields such as heart disease detection. Its simplicity and interpretability make it a popular choice for binary classification problems, where the goal is to predict one of two possible outcomes based on input features. Logistic regression models the relationship between input features and the probability of the target outcome using a sigmoid function, which outputs a value between 0 and 1. This probability is then thresholded to classify the instance into one of the two classes. Regularization techniques, such as $l1$ and $l2$, were also employed to prevent overfitting. These methods penalize large coefficients, ensuring the model generalizes better to unseen data [3].

Neural Network (Multilayer Perceptron): was chosen to capture complex nonlinear relationships in the data that logistic regression might miss. MLP is a type of feedforward neural network that consists of multiple layers of interconnected nodes, or neurons. Each neuron applies a non-linear activation function to the weighted sum of its inputs, enabling the network to model intricate patterns in the data. The flexibility of MLP and its ability to handle complex patterns make it a good option to deal with problems with a variety of clinical indicators [7].

3.3 Hyperparameter Optimization

GridSearchCV was employed to optimize hyperparameters for both models. This technique systematically evaluates all possible combinations of spec-

ified hyperparameter values using cross-validation. For logistic regression, parameters such as the type of regularization ($l1$ or $l2$) and the regularization strength were optimized. For neural networks, parameters including the number of hidden layers, neurons per layer, activation functions, and learning rates were tuned. Grid search ensures that the chosen models achieve the best possible performance on the training data [1].

3.4 Evaluation Metrics

To assess model performance, various metrics were used:

- **Accuracy:** Proportion of correctly classified samples.
- **Precision and Recall:** Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced evaluation metric.
- **ROC-AUC:** Evaluates the model's ability to distinguish between classes by measuring the area under the receiver operating characteristic (ROC) curve.

4 Results

4.1 Analysis

To better understand the dataset and ensure the reliability of our models, we performed an initial analysis focusing on class balance and feature correlations.

4.1.1 Class Balance

We first checked the balance of the target variable, which represents the presence or absence of heart disease. Class balance is crucial in machine learning because imbalanced datasets can bias the model toward the majority class, leading to poor performance on the minority class. Fortunately, the dataset is relatively balanced, with 54.46% of the samples labeled as 1 (heart disease) and 45.54% labeled as 0 (no heart disease). This balance ensures that the model is trained on well balanced examples of both classes, improving its ability to generalize.

4.1.2 Feature Correlation Analysis

We visualized feature correlations using a heatmap (Figure 1) to assess relationships between features and identify multicollinearity, which can affect model performance, especially in Logistic Regression.

Key observations include:

- **Age** has a moderate negative correlation with **thalach (maximum heart rate achieved)** (-0.40), reflecting the expected decline in cardiovascular performance with age.
- **Oldpeak (ST depression)** shows a moderate negative correlation with **thalach** (-0.34), suggesting that patients with higher heart rates during exercise tend to have lower ST depression levels.
- Weak correlations, such as between **chol (cholesterol)** and other features, indicate minimal relationships, suggesting largely independent contributions from most features.

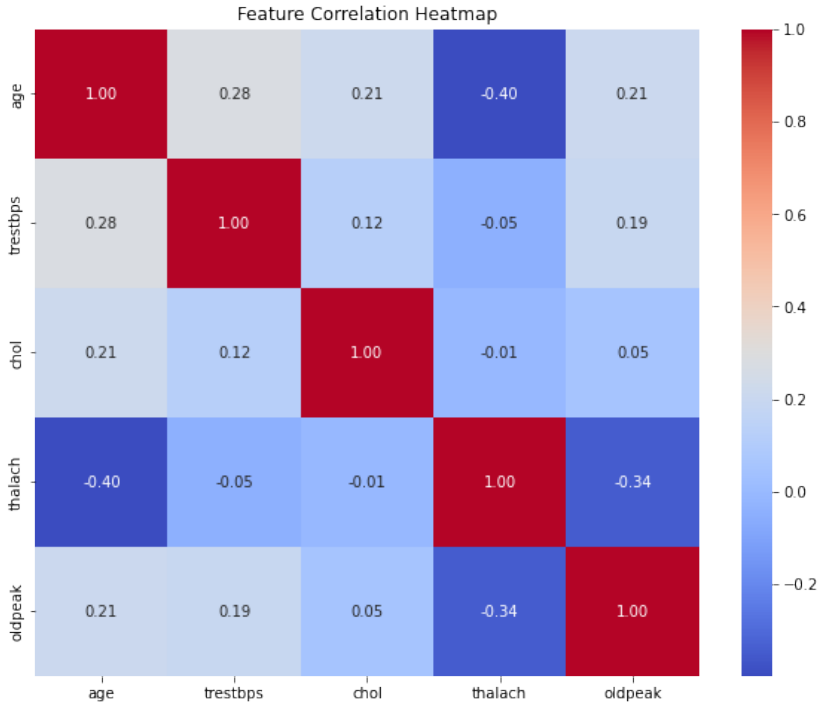

Figure 1: Correlation Heatmap of Features

These insights provide confidence in the dataset's quality and inform the

selection of machine learning models. The low multicollinearity indicates that no immediate feature elimination is required at this stage.

4.1.3 PCA Results

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the dataset and assess whether fewer features could capture most of the variance in the data. Originally, the dataset consisted of 5 features. After applying PCA, the number of features was reduced to 4 while retaining 91.15% of the total variance in the data.

The explained variance ratio for each principal component is as follows:

- **Component 1:** 42.47% of the variance (cumulative: 42.47%).
- **Component 2:** 20.05% of the variance (cumulative: 62.51%).
- **Component 3:** 18.03% of the variance (cumulative: 80.54%).
- **Component 4:** 10.61% of the variance (cumulative: 91.15%).

Figure 2 shows the cumulative explained variance as a function of the number of components. The first three components capture over 80% of the variance, demonstrating that a significant amount of information can be retained with fewer dimensions. This reduction simplifies the dataset while preserving most of its structure, which is beneficial for reducing noise and computational complexity during model training.

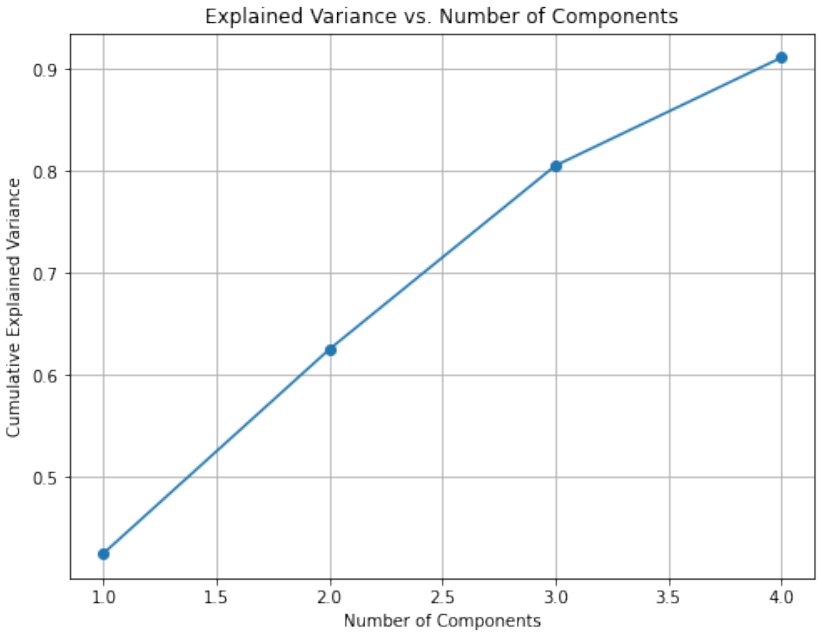

Figure 2: Cumulative Explained Variance vs. Number of Components

4.1.4 Logistic Regression: Comparative Results (With and Without PCA)

To evaluate the impact of dimensionality reduction, logistic regression was applied to the dataset both with and without PCA. The best hyperparameters for each scenario were chosen using GridSearchCV with 5-fold cross-validation.

Hyperparameters and Their Meaning

- **Without PCA:** The best hyperparameters identified were:
  - **C:** 1.62, which controls the strength of regularization. Lower values of C imply stronger regularization to prevent overfitting, while higher values relax this constraint.
  - **Penalty:** $l1$, which performs feature selection by driving some coefficients to zero. This helps simplify the model and potentially improve generalization.
- **With PCA:** The best hyperparameters identified were:

- **C:** 4.28, which indicates weaker regularization compared to the model without PCA, likely because PCA reduces feature redundancy.
- **Penalty:** $l1$, which again helps in feature selection by shrinking less relevant coefficients to zero.
- **Solver:** `liblinear`, an efficient solver for small datasets, particularly suited for $l1$ regularization.

Metric	**Without PCA**	**With PCA**
Testing Accuracy (%)	72.13	72.13
ROC-AUC	0.7736	0.7769
Precision (%)	73.68	73.68
Recall (%)	80.00	80.00
F1-Score (%)	76.71	76.71

Table 1: Comparison of Logistic Regression Performance With and Without PCA

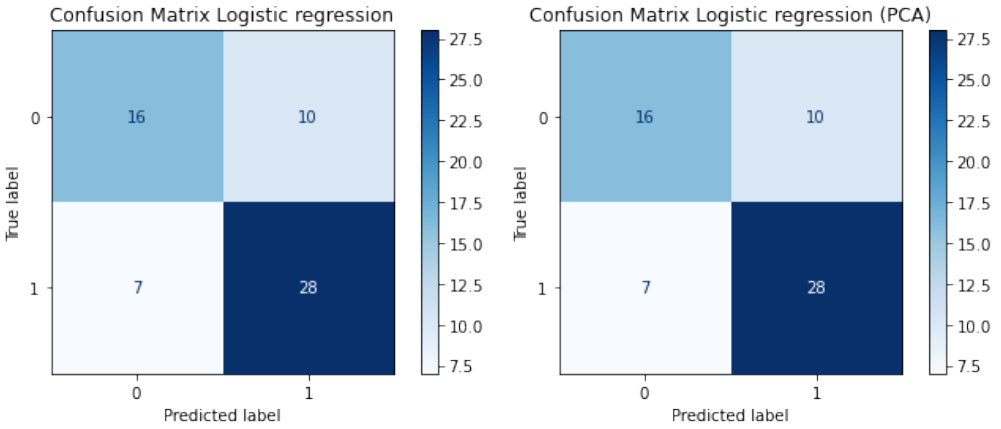

Figure 3: Confusion Matrices for Logistic Regression (Left: Without PCA, Right: With PCA)

- The darker blue cells along the diagonal represent the correctly classified cases:
  - **True Negatives (16 cases):** These are patients correctly identified as not having heart disease. In a medical context, this means these individuals would not undergo unnecessary treatments or tests.

  - **True Positives (28 cases):** These are patients correctly identified as having heart disease. This is critical, as these individuals can receive timely medical intervention or further testing.

- The lighter blue cells off the diagonal represent the misclassified cases:
  - **False Positives (10 cases):** These are patients incorrectly classified as having heart disease when they do not. In a real scenario, this could lead to unnecessary stress for the patients and additional, potentially costly, diagnostic tests or treatments.
  - **False Negatives (7 cases):** These are patients incorrectly classified as not having heart disease when they actually do. This is particularly concerning in medical diagnostics, as these individuals may not receive the necessary care, potentially leading to worsened health outcomes or missed opportunities for prevention.

Overall, the confusion matrix shows that the model performs reasonably well, but reducing false negatives should be a priority. Since the heatmap remained the same we may conclude that PCA had almost zero effect on logistic regression in this application.

Interpretation of Results The results indicate that applying PCA did not significantly alter the performance of logistic regression. The testing accuracy, precision, recall, F1-score, and ROC-AUC values remain almost identical with and without PCA. This suggests that the original dataset did not have significant redundancy or noise that PCA could eliminate, and the reduced dimensionality did not meaningfully enhance the model's predictive ability. The relatively higher recall (80%) indicates that the model performs well in identifying patients with heart disease, which is crucial in medical settings to minimize missed diagnoses. However, the false positives could lead to some unnecessary follow-up actions.

4.1.5 Neural Network: Comparative Results (With and Without PCA)

To evaluate the performance of a Neural Network (Multilayer Perceptron, MLP) in predicting heart disease, we applied the model to both the original dataset and the PCA-transformed dataset. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation.

Hyperparameters and Their Meaning

- **Without PCA:** The best hyperparameters identified were:
  - **Activation:** ReLU (Rectified Linear Unit), a widely used activation function that introduces non-linearity into the model while being computationally efficient.
  - **Alpha:** 0.0001, the L2 regularization term to prevent overfitting by penalizing large weights.
  - **Hidden Layer Sizes:** (128, 64), indicating two hidden layers with 128 and 64 neurons, respectively. This structure allows the network to capture complex patterns in the data.
  - **Learning Rate:** 0.01, controlling the step size for weight updates during optimization.
  - **Solver:** SGD (Stochastic Gradient Descent), an optimization algorithm that iteratively updates weights to minimize the loss function.
- **With PCA:** The best hyperparameters identified were:
  - **Activation:** Tanh, which maps input values to the range [-1, 1], providing a smoother gradient than ReLU.
  - **Alpha:** 0.0001, as with the model without PCA.
  - **Hidden Layer Sizes:** (128, 64), the same as the non-PCA model.
  - **Learning Rate:** 0.001, a smaller step size compared to the non-PCA model, likely reflecting the reduced feature dimensionality.
  - **Solver:** Adam, a more advanced optimization algorithm that adapts learning rates for faster convergence.

Metric	**Without PCA**	**With PCA**
Testing Accuracy (%)	75.41	67.21
ROC-AUC	0.7670	0.7725
Precision (%)	75.00	71.43
Recall (%)	85.71	71.43
F1-Score (%)	80.00	71.43

Table 2: Comparison of Neural Network Performance With and Without PCA

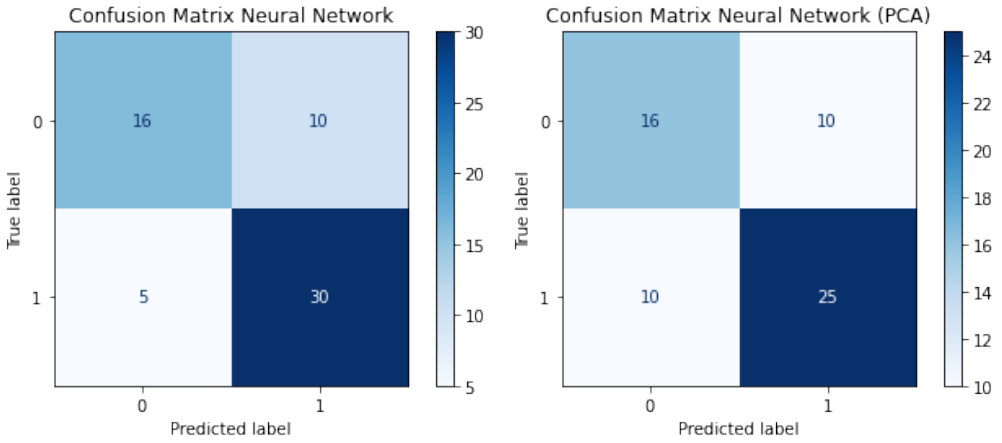

Figure 4: Confusion Matrices for Neural Network (Left: Without PCA, Right: With PCA)

Interpretation of the Confusion Matrices The confusion matrices highlight notable differences in the Neural Network's performance with and without PCA. Without PCA, the model achieves higher recall, correctly identifying more patients with heart disease (30 true positives compared to 25 with PCA). This is crucial in a medical context, as missing true cases (false negatives) can lead to delayed treatment and worsened health outcomes.

However, with PCA, the model shows an increased number of false negatives (10 compared to 5 without PCA), indicating a reduced ability to detect heart disease. This suggests that PCA may have removed important information critical for the neural network's classification. The number of false positives and true negatives remains unchanged, demonstrating that PCA primarily impacts the model's ability to correctly identify positive cases.

Overall, the confusion matrices suggest that while PCA simplifies the dataset, it negatively affects the neural network's ability to capture patterns necessary for accurate heart disease prediction.

The performance of the MLP model without PCA outperforms the PCA-transformed model in most metrics. Notably, the non-PCA model achieves higher testing accuracy (75.41% vs. 67.21%) and F1-score (80.00% vs. 71.43%).

4.2 Final Results: Comparison of Model Accuracies

The bar chart in Figure 5 provides a comparative summary of the test accuracies achieved by both models (Logistic Regression and Neural Network) with and without PCA.

Observations:

- **Logistic Regression:** The test accuracy remains unchanged with or without PCA, achieving an accuracy of 72.13%. This suggests that dimensionality reduction via PCA did not significantly impact the performance of logistic regression, likely because the original dataset's dimensionality was already optimal for this model.
- **Neural Network:** The model without PCA achieved a higher accuracy of 75.41% compared to 67.21% with PCA. This indicates that the PCA-transformed dataset lost critical information necessary for the Neural Network to perform well, highlighting the sensitivity of neural networks to feature representation.

The results indicate that while PCA can simplify datasets, it does not always lead to improved performance. In this study, PCA had no impact on logistic regression but degraded the performance of the Neural Network.

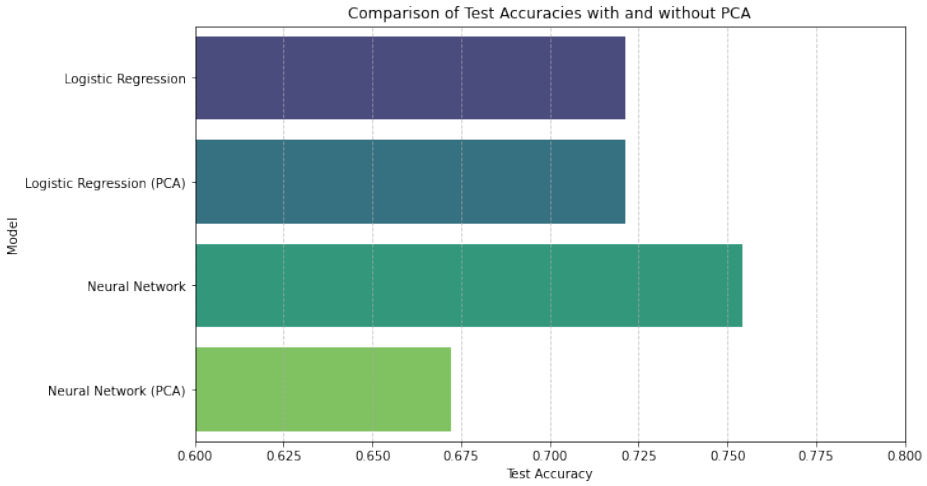

Figure 5: Comparison of Test Accuracies for Logistic Regression and Neural Network Models With and Without PCA

5 Conclusion

This project successfully demonstrated the use of Logistic Regression and Neural Networks for predicting heart disease based on clinical indicators. Both models achieved reasonable performance, with Logistic Regression achieving 72.13% accuracy consistently with and without PCA, while the Neural Network performed better without PCA, achieving 75.41% accuracy.

5.1 Future Directions

To further improve results:

- Incorporate additional clinical features for better predictive accuracy.
- Explore advanced neural network architectures.

[1]

References

[1] James Bergstra and Yoshua Bengio. A comprehensive introduction to grid search for model optimization. *Journal of Machine Learning Research*, 2012.

[2] Mohammadreza Hajiarbabi. Heart disease detection using machine learning methods: a comprehensive narrative review. *Journal of Medical Artificial Intelligence*, 7(0), 2024.

[3] David W. Hosmer and Stanley Lemeshow. Logistic regression: Theory and application. 2000.

[4] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[5] Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande. Comparative study of classification techniques (svm, logistic regression and neural networks) to predict the prevalence of heart disease. *International Journal of Machine Learning and Computing*, 5(5):414, 2015.

[6] Fabian Pedregosa et al. Normalization in machine learning. *Journal of Machine Learning Research*, 2011.

[7] David E. Rumelhart and James L. McClelland. Multilayer perceptrons and their applications to pattern recognition. *Nature*, 1986.

[1] AI tools were used for code troubleshooting, hyperparameter tuning and for finding a relevant resources.