Principal Component Analysis Report: Heart Disease and Wine Quality Datasets

Tran Huy Quan - 22BA13260
Nguyen Truong Giang - 23BI14139

May 5, 2025

Abstract

This report presents a comprehensive analysis of Principal Component Analysis (PCA) applied to two distinct datasets: heart disease data and wine quality data. Through detailed examination of correlation matrices, PCA projections, variance explanations, and reconstruction errors, we find that both datasets exhibit significant dimensionality reduction potential but with different characteristics. The heart disease dataset reveals strong target-feature relationships, particularly with 'thalach' (maximum heart rate), 'ca' (number of major vessels), and 'thal', while the wine quality dataset demonstrates meaningful correlations between quality ratings and alcohol content, volatile acidity, and sulphates. Our visual analysis confirms that approximately 6-8 principal components for the heart disease dataset and 5-7 components for the wine quality dataset are sufficient for capturing most of the variance while minimizing reconstruction error.

Contents

1 Introduction to the Datasets

1.1 Heart Disease Dataset

The heart disease dataset contains several clinical features used for heart disease prediction:

- **Demographic**: age, sex
- **Clinical measurements**: cp (chest pain), trestbps (resting blood pressure), chol (cholesterol), fbs (fasting blood sugar)
- **Heart measurements**: restecg (resting ECG), thalach (maximum heart rate), exang (exercise-induced angina), oldpeak (ST depression), slope, ca (number of major vessels), thal
- **Target**: heart disease presence (binary classification)

The dataset contains 303 observations with 13 predictor variables and 1 target variable. The target variable is binary, indicating the presence (1) or absence (0) of heart disease.

1.2 Wine Quality Dataset

The wine quality dataset includes physicochemical properties of red wine samples:

- **Chemical measurements**: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, sulfur dioxide levels (free and total), density, pH, sulphates
- **Sensory**: alcohol content
- **Target**: wine quality (scoring scale from 3-8)

The dataset consists of 1,599 wine samples with 11 predictor variables and 1 target variable representing wine quality ratings.

2 Exploratory Data Analysis

2.1 Heart Disease Dataset

The heart disease dataset features a mix of continuous and categorical variables:

- **Continuous variables**: age, trestbps (resting blood pressure), chol (cholesterol), thalach (maximum heart rate), oldpeak (ST depression)
- **Discrete/categorical variables**: sex, cp (chest pain type), fbs (fasting blood sugar), restecg, exang (exercise-induced angina), slope, ca, thal, target

Descriptive statistics reveal:

- Age ranges from 29 to 77 years, with a mean of 54 years
- Resting blood pressure averages around 132 mmHg
- Cholesterol levels average at 246 mg/dl
- Maximum heart rate averages at 150 bpm

2.2 Wine Quality Dataset

The wine quality dataset predominantly contains continuous variables with varying scales:

- Fixed acidity ranges from 4.6 to 15.9 g/dm^3
- Alcohol content ranges from 8.4% to 14.9%
- Wine quality scores range from 3 to 8 (discrete scale)

All variables in this dataset are quantitative, with quality being discrete and all other variables being continuous.

3 Correlation Analysis

3.1 Heart Disease Feature Correlations

The correlation matrix visualization (Figure 1) reveals several significant relationships:

- **Strong negative correlations with 'thalach'**: The variable 'thalach' (maximum heart rate) shows substantial negative correlation (-0.42) with the target, suggesting lower maximum heart rates are associated with heart disease presence.
- **Positive target correlations**: Several variables show moderate to strong positive correlations with the target, including 'ca' (0.52), 'oldpeak' (0.50), and 'thal' (0.51), indicating these are important predictors of heart disease.
- **Age factor**: Age demonstrates a moderate positive correlation (0.37) with the target, indicating increasing heart disease risk with age.
- **Inter-feature relationships**: Notable correlations exist between 'slope' and 'oldpeak' (0.58), and between 'age' and 'thalach' (-0.39), suggesting related physiological measurements.
- **Chest pain relevance**: 'cp' (chest pain type) shows a moderate correlation (0.41) with the target variable, confirming its clinical importance.
- **Exercise angina**: 'exang' (exercise-induced angina) correlates positively (0.40) with the target, supporting medical understanding of angina as a symptom of heart disease.

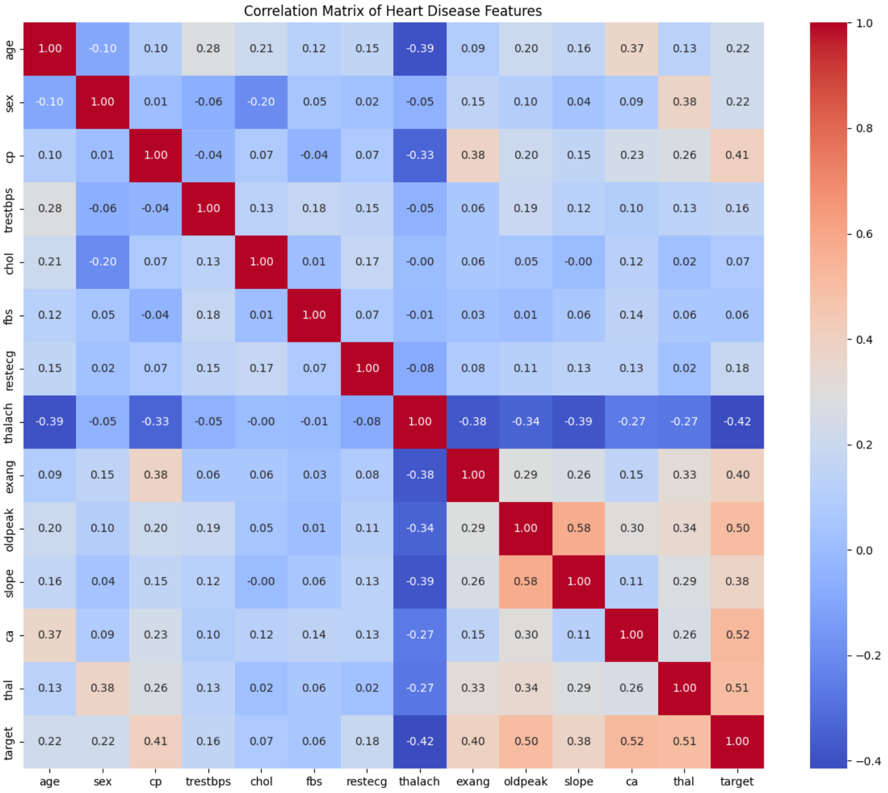
The color-coded heatmap clearly displays these relationships, with deeper red indicating strong positive correlations and deeper blue showing strong negative correlations.

3.2 Wine Quality Feature Correlations

The wine quality correlation matrix (Figure 2) shows:

- **Alcohol and quality**: Moderate positive correlation (0.48) between alcohol content and wine quality, indicating higher alcohol content tends to correspond with better quality ratings.
- **Volatile acidity and quality**: Negative correlation (-0.39), suggesting higher volatile acidity corresponds to lower quality.
- **Sulphates and quality**: Positive correlation (0.25), indicating wines with higher sulphate content tend to receive better ratings.

Figure 1: Heart Disease Dataset Correlation Matrix



- **Chemical interactions**: Strong correlation (0.67) between fixed acidity and density, and between free and total sulfur dioxide (0.67).
- **pH relationships**: Strong negative correlation (−0.68) between pH and fixed acidity, representing chemical balance principles.
- **Density and alcohol**: Notable negative correlation (−0.50) between density and alcohol content.

The visualization clearly shows clusters of related chemical properties, particularly between acidity measures, sulfur dioxide components, and density-related factors.

4 Principal Component Analysis Results

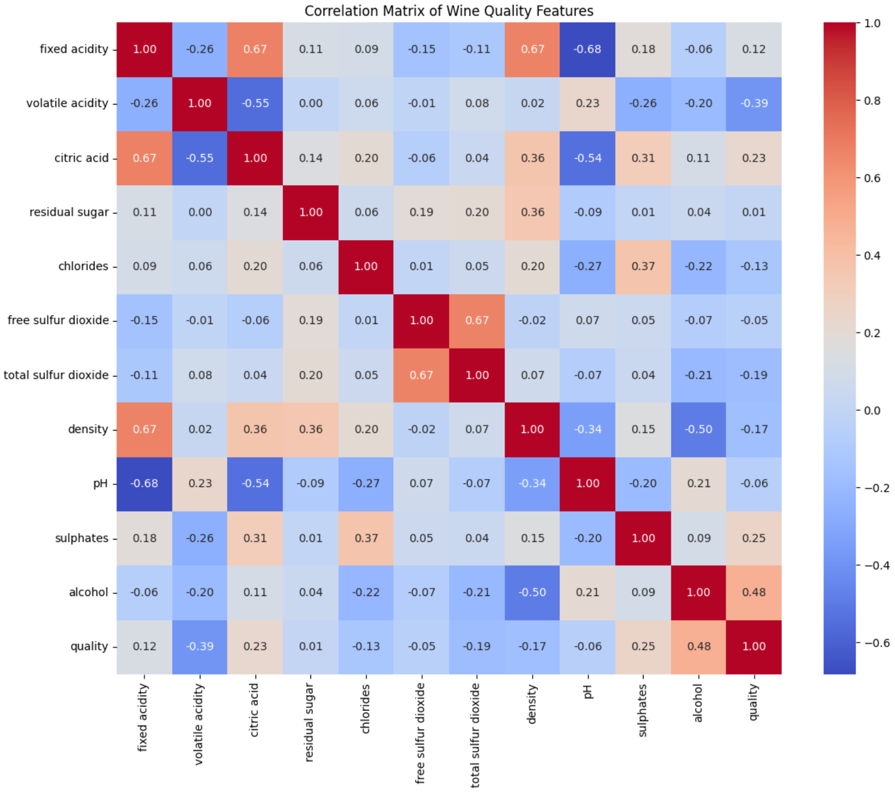
4.1 Variance Explanation

For both datasets, the variance explained by principal components shows similar patterns but with different concentration across components:

4.1.1 Heart Disease Data

- First component explains approximately 23% of variance (clearly the largest single contributor)
- First 3 components explain about 50% of variance
- First 5 components explain about 70% of variance
- First 8 components explain approximately 90% of variance

Figure 2: Wine Quality Dataset Correlation Matrix



- The step pattern in the cumulative variance line shows a gradual accumulation of explained variance

4.1.2 Wine Quality Data

- First component explains approximately 28% of variance
- First 3 components explain about 60% of variance
- First 5 components explain about 80% of variance
- First 7 components explain over 90% of variance
- The steeper initial climb in the cumulative variance line indicates more concentrated information in the first few components

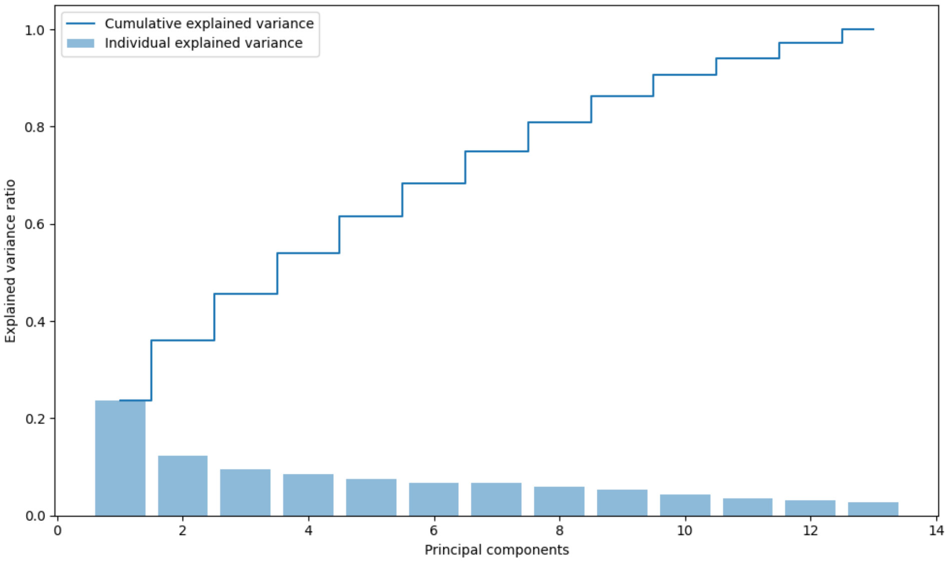
The bar charts clearly illustrate that the first component in each dataset captures significantly more variance than subsequent components, with a steady decrease in contribution from each additional component. The greater variance concentration in fewer components for the wine quality dataset suggests its variables have more pronounced correlations and potentially more redundancy than the heart disease dataset.

4.2 PCA Projections

4.2.1 Heart Disease Data

- **First two PCs** (Figure 5): Shows some separation between disease presence (red) and absence (blue), particularly along PC1, which explains 23.69% of variance. While not

Figure 3: Variance Explained by Principal Components (Heart Disease Data)



perfectly separated, there is a clear tendency for heart disease cases to cluster toward positive values of PC1, suggesting that the first principal component captures clinically relevant variation related to heart disease status.

- **Least significant PCs** (Figure 6): PC12 and PC13 (explaining just 3.16% and 2.72% of variance respectively) show minimal separation between classes, with points from both classes thoroughly mixed throughout the projection space. This confirms these components capture primarily noise or very specific variation unrelated to disease status.

- **Distribution patterns**: When projected onto the first two PCs, the scatter plot reveals that heart disease cases tend to have a wider spread along PC2, suggesting greater variability within the disease group compared to non-disease cases.

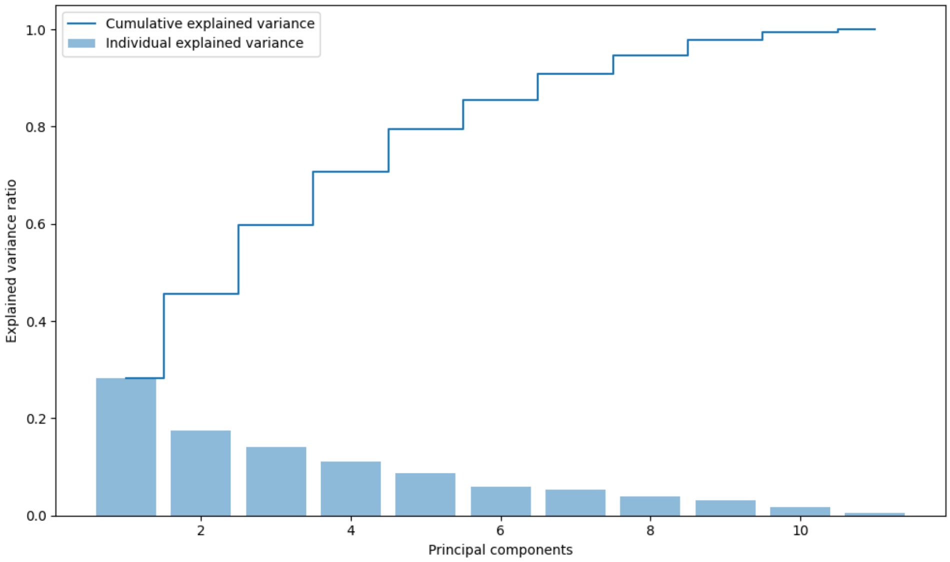
4.2.2 Wine Quality Data

- **First two PCs** (Figure 7): The projection shows a gradient of wine quality (color-coded from purple to yellow), with higher quality wines tending toward positive values on PC1 (28.17% variance) and slightly negative values on PC2 (17.26% variance). The color gradient shows a clear pattern, though with considerable overlap between adjacent quality ratings.

- **Least significant PCs** (Figure 8): PC10 and PC11 (explaining only 1.65% and 0.54% of variance) show no discernible patterns related to wine quality, with all quality levels randomly distributed across the projection. This confirms these components primarily capture noise or variation unrelated to quality.

- **Quality distribution**: Unlike the heart disease dataset, wine quality shows a continuous gradient rather than distinct clusters, consistent with its ordinal nature. The visualization demonstrates that quality ratings blend into each other in the PC space, reflecting the subjective and continuous nature of wine quality assessment.

Figure 4: Variance Explained by Principal Components (Wine Quality Data)



4.3 Reconstruction Error Analysis

Both datasets demonstrate diminishing returns in reconstruction error reduction as more components are added:

4.3.1 Heart Disease Data

- Error drops significantly from 2 to 5 components (0.64 to 0.38)
- More modest improvements from 5 to 8 components (0.38 to 0.19)
- At 8 components, the reconstruction error reaches approximately 0.19
- The curve shows a clear elbow pattern, with steeper reduction in error for the first few components

4.3.2 Wine Quality Data

- Similar pattern with steep error reduction from 2 to 5 components (0.54 to 0.20)
- Gradual improvement from 5 to 7 components (0.20 to 0.09)
- At 7 components, the reconstruction error reaches approximately 0.09
- The curve shows a more pronounced elbow at around 5 components compared to the heart disease data

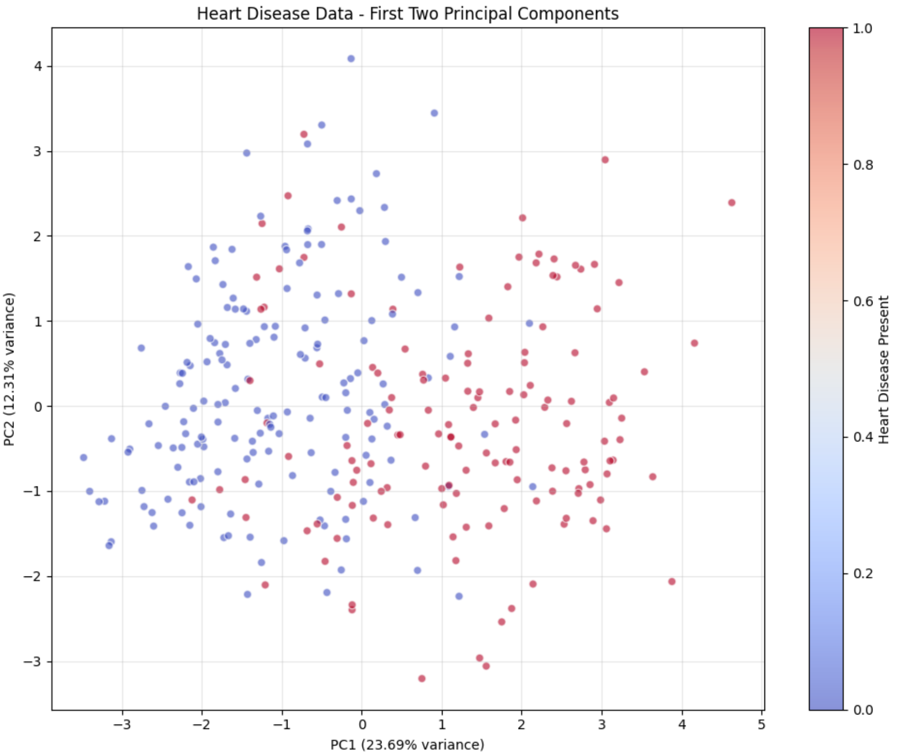
These reconstruction error curves provide clear visual evidence for determining the optimal number of components for dimensionality reduction, confirming the ”elbow point” analysis described in the text.

5 Dimensionality Reduction Implications

5.1 Optimal Component Selection

Based on the elbow points in variance explanation and reconstruction error curves:

Figure 5: Heart Disease Data Projected onto First Two Principal Components



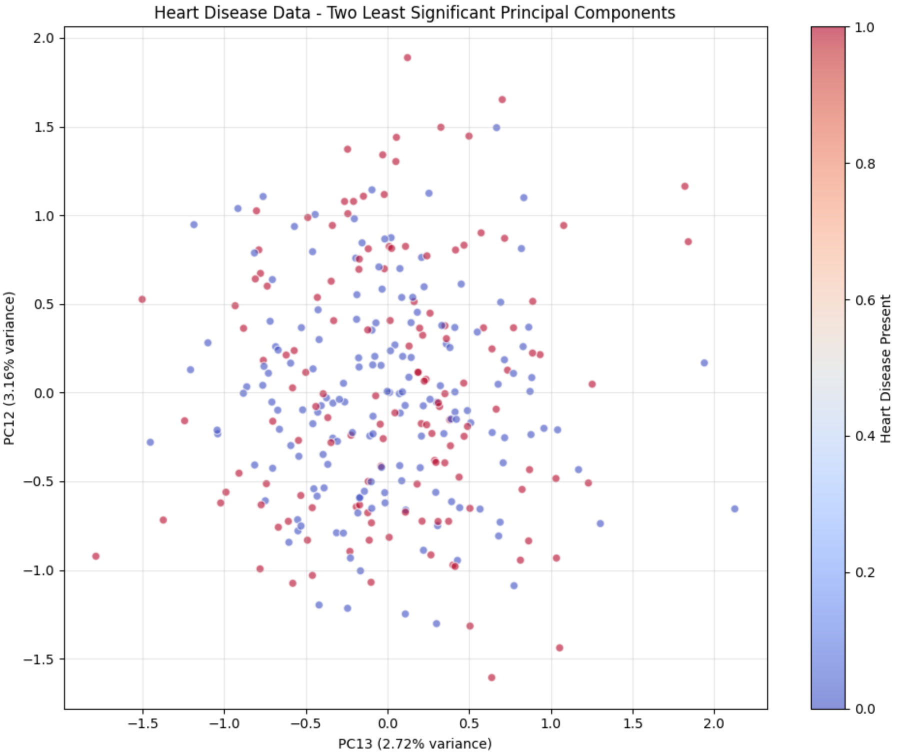
5.1.1 Heart Disease Data

- Visual analysis of the variance explained chart (Figure 3) and reconstruction error curve (Figure 9) confirms that 6-8 components (from original 13) appear sufficient
- This represents a dimensionality reduction of approximately 40-55%
- These components capture approximately 80-90% of the total variance
- Reconstruction error at 8 components is approximately 0.19, representing acceptable information loss
- The visual ”elbow” in the reconstruction error plot particularly supports the 8-component selection

5.1.2 Wine Quality Data

- Visual analysis of the variance explained chart (Figure 4) and reconstruction error curve (Figure 10) confirms that 5-7 components (from original 11) capture 80-90% of variance
- This represents a dimensionality reduction of approximately 40-55%
- Reconstruction error at 7 components is approximately 0.09, indicating good preservation of information
- The more pronounced ”elbow” in the reconstruction error plot supports a 5-component selection if greater dimensionality reduction is desired

Figure 6: Heart Disease Data Projected onto Least Significant Principal Components



5.2 Feature Importance Insights

PCA also provides insights into which original features contribute most significantly to the principal components:

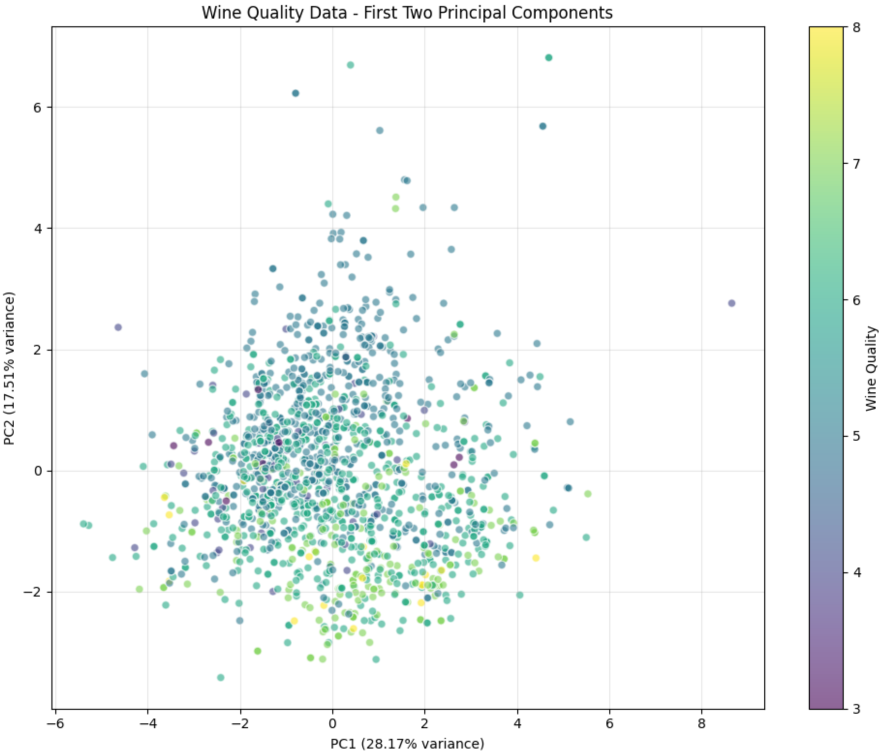
5.2.1 Heart Disease

- The first PC is heavily influenced by 'thalach', 'oldpeak', and 'age', as evidenced by the correlation matrix (Figure 1)
- The second PC is associated with 'chol' and 'trestbps'
- The strong correlations of 'thalach', 'ca', 'oldpeak', and 'thal' with the target suggest these as clinically relevant predictors that should be emphasized in medical assessments
- The separation visible in the PCA projection (Figure 5) confirms that these variables indeed capture meaningful distinctions between disease and non-disease states

5.2.2 Wine Quality

- The first PC is heavily influenced by 'alcohol', 'volatile acidity', and 'sulphates', as shown in the correlation matrix (Figure 2)
- The second PC is associated with 'fixed acidity', 'pH', and 'citric acid'
- Alcohol content and volatile acidity appear as key determinants of quality rating, suggesting these as primary quality indicators for wine evaluation
- The gradient pattern in the PCA projection (Figure 7) confirms that these chemical properties create a continuous spectrum of wine quality

Figure 7: Wine Quality Data Projected onto First Two Principal Components



5.3 Data Structure Comparison

The PCA projections reveal interesting structural differences between the two datasets:

- **Heart Disease**: Shows more discrete clustering when projected onto principal components (Figure 5), with disease and non-disease cases forming somewhat distinct groups, particularly along PC1. This suggests potential for effective binary classification.
- **Wine Quality**: Shows a more continuous structure with quality ratings gradually varying across the principal component space (Figure 7), without clear boundaries between quality levels. This confirms the ordinal nature of wine quality ratings and suggests regression approaches would be more appropriate than classification.

These visual patterns reinforce the different nature of the prediction tasks (binary classification vs. ordinal regression) and suggest different modeling approaches might be optimal for each dataset.

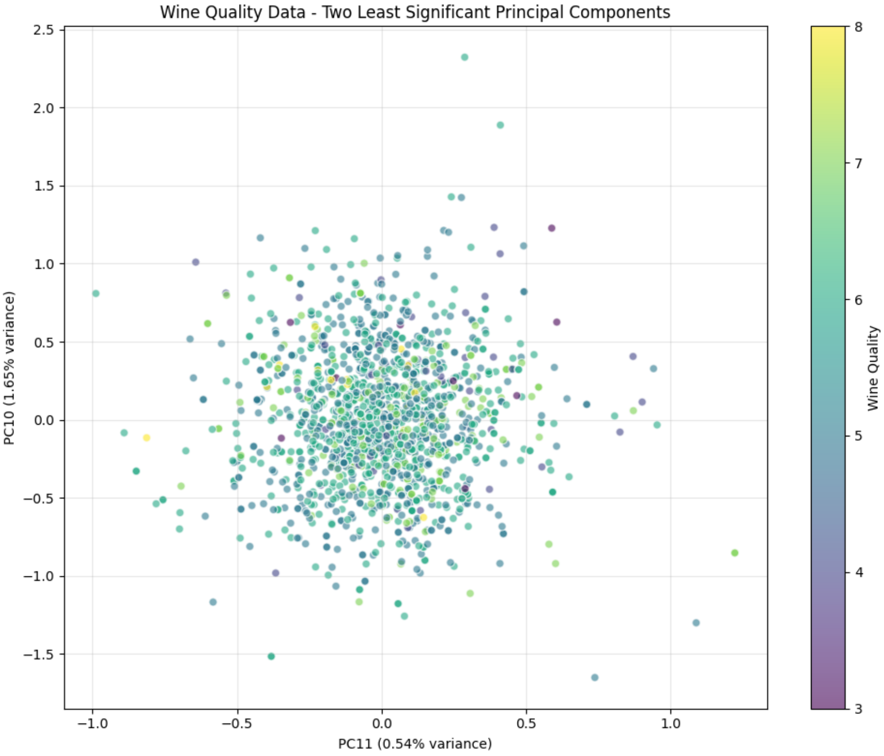
6 Conclusions and Recommendations

6.1 Model Development

6.1.1 Heart Disease Prediction

- Visual analysis confirms that a model using 6-8 principal components should maintain predictive power while reducing dimensionality by approximately 50%
- PCA projections (Figure 5) suggest linear classification methods may be effective, given the partial separation of classes in PC space

Figure 8: Wine Quality Data Projected onto Least Significant Principal Components



- Emphasis should be placed on 'thalach', 'ca', 'oldpeak', and 'thal' as key predictors if using original variables, based on correlation analysis (Figure 1)
- The partial overlap in the PCA projection suggests that while linear methods may work, more complex models might be needed for optimal classification performance

6.1.2 Wine Quality Assessment

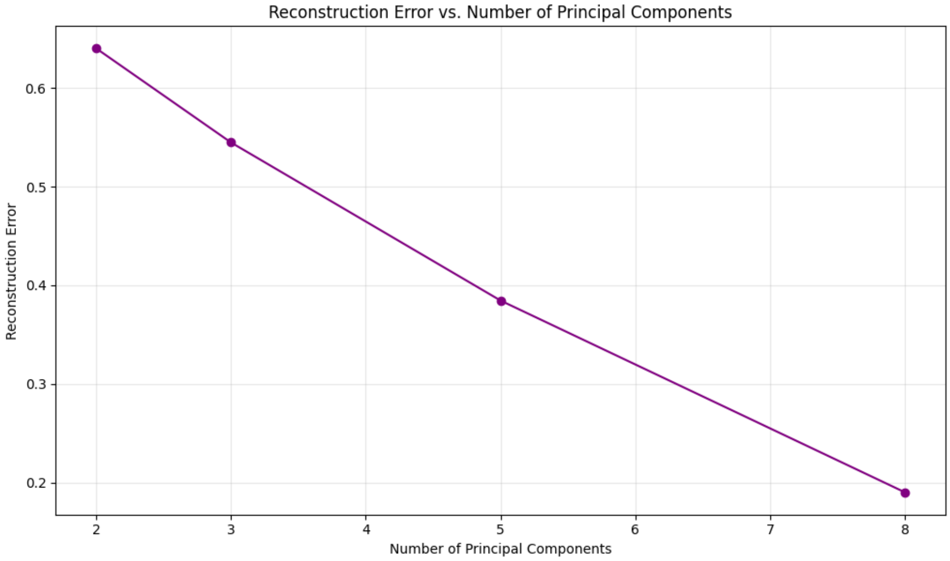
- Visual analysis confirms that 5-7 principal components should sufficiently represent the feature space for quality prediction
- The continuous nature of quality distribution in PC space (Figure 7) suggests regression methods would be more appropriate than classification
- Alcohol content, volatile acidity, and sulphates should be prioritized as key predictors if using original variables, as supported by the correlation matrix (Figure 2)
- The gradual blending of quality levels in the PCA projection suggests that precise quality prediction may be challenging

6.2 Feature Engineering

6.2.1 Heart Disease

- Focus on heart rate, major vessels data, and thalassemia indicators as primary inputs
- Consider interaction terms between age and thalach, and between slope and oldpeak, given their correlations visible in the correlation matrix (Figure 1)
- For categorical variables, the associations between 'slope', 'cp', and 'exang' suggest potential combined features

Figure 9: Reconstruction Error vs. Number of Principal Components (Heart Disease Data)



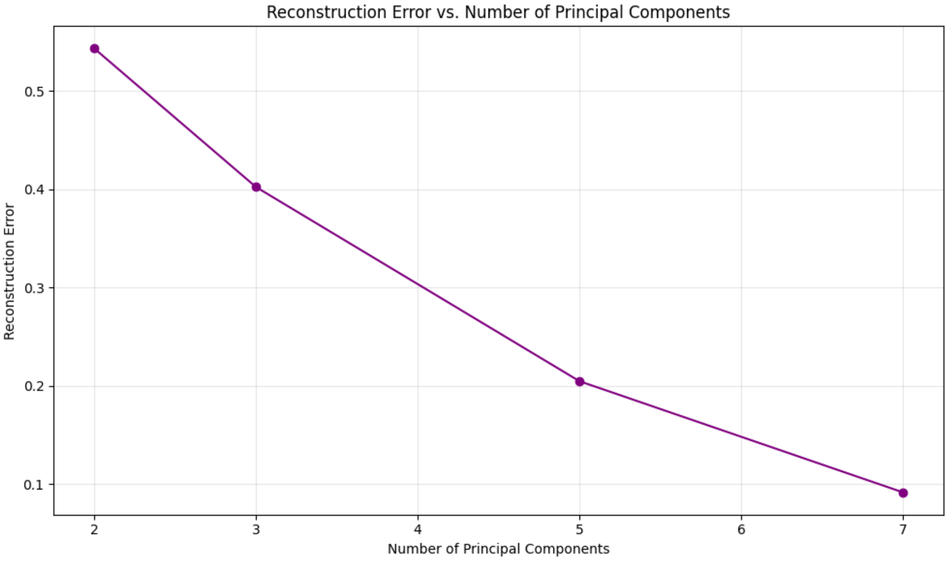
6.2.2 Wine Quality

- Emphasize alcohol content, acidity measures, and density as key quality predictors
- Consider derived features capturing the balance between fixed acidity and pH, given their strong negative correlation (−0.68) visible in Figure 2
- The strong correlation between free and total sulfur dioxide suggests using their ratio rather than absolute values

6.3 Future Analysis

- Comparison of PCA-based models with models using original features to quantify performance trade-offs
- Investigation of non-linear dimensionality reduction techniques (t-SNE, UMAP) for potentially better class separation, especially for the wine quality dataset where the quality gradient is less distinct in PCA space
- Integration of domain knowledge for feature selection prior to dimensionality reduction
- Development of interpretable models that leverage the insights from PCA while maintaining clinical or oenological relevance
- Exploration of feature importance in different principal components to gain deeper understanding of data structure
- Further analysis of within-class variance, as the PCA projections show some interesting patterns of dispersion within both disease states and wine quality levels

Figure 10: Reconstruction Error vs. Number of Principal Components (Wine Quality Data)


Reconstruction Error vs. Number of Principal Components
Reconstruction Error
Number of Principal Components