Principal Component Analysis Report: Heart Disease and Wine Quality Datasets

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

May 11, 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 **3**
1.1 Heart Disease Dataset 3
1.2 Wine Quality Dataset 3

2 Data Characteristics and Preprocessing **3**
2.1 Feature Classification 3
2.1.1 Heart Disease Dataset 3
2.1.2 Wine Quality Dataset 3
2.2 Data Quality Assessment 4
2.2.1 Heart Disease Dataset 4
2.2.2 Wine Quality Dataset 4
2.3 Target Variable Identification 5
2.3.1 Heart Disease Dataset 5
2.3.2 Wine Quality Dataset 5

3 Exploratory Data Analysis **5**
3.1 Heart Disease Dataset 5
3.1.1 Key Statistical Measures 5
3.2 Wine Quality Dataset 6
3.2.1 Key Statistical Measures 6

4 Correlation Analysis **6**
4.1 Heart Disease Feature Correlations 6
4.1.1 Most Correlated Feature Pair 7
4.2 Wine Quality Feature Correlations 7
4.2.1 Most Correlated Feature Pair 8

5 Principal Component Analysis Results 9
5.1 PCA Methodology 9
5.2 Variance Explanation 9
5.2.1 Heart Disease Data 9
5.2.2 Wine Quality Data 9
5.3 PCA Projections 10
5.3.1 Heart Disease Data 10
5.3.2 Wine Quality Data 10
5.4 Highest vs. Lowest Principal Components 11
5.4.1 Heart Disease Data 11
5.4.2 Wine Quality Data 11
5.5 Reconstruction Error Analysis 12
5.5.1 Heart Disease Data 12
5.5.2 Wine Quality Data 12

6 Dimensionality Reduction Implications 13
6.1 Optimal Component Selection 13
6.1.1 Heart Disease Data 13
6.1.2 Wine Quality Data 14
6.2 Feature Importance Insights 14
6.2.1 Heart Disease 14
6.2.2 Wine Quality 15
6.3 Data Structure Comparison 15

7 Conclusions and Recommendations 16
7.1 Model Development 16
7.1.1 Heart Disease Prediction 16
7.1.2 Wine Quality Assessment 16
7.2 Feature Engineering 16
7.2.1 Heart Disease 16
7.2.2 Wine Quality 17
7.3 Component Selection Trade-offs 17
7.4 Domain-Specific Insights 18
7.4.1 Medical Applications 18
7.4.2 Wine Industry Applications 18

8 Limitations and Future Work 18
8.1 Current Limitations 18
8.2 Future Research Directions 19

9 Summary and Key Takeaways 19
9.1 Heart Disease Dataset 19
9.2 Wine Quality Dataset 19
9.3 Comparative Insights 20

10 References 20

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 Data Characteristics and Preprocessing

2.1 Feature Classification

2.1.1 Heart Disease Dataset

Classification criteria:

- **Discrete vs. Continuous**: Features with more than 10 unique numeric values are considered continuous; otherwise, they are discrete.
- **Quantitative vs. Qualitative**: Numeric features are classified as quantitative; non-numeric features are qualitative.
- **Numerical vs. Categorical**: Features that are numeric with multiple unique values are considered numerical; otherwise, they are categorical.

2.1.2 Wine Quality Dataset

Classification criteria:

- **Discrete vs. Continuous**: Numeric features with less than 20 unique values are considered discrete; otherwise, they are continuous.
- **Quantitative vs. Qualitative**: All numeric features are classified as quantitative.
- **Numerical vs. Categorical**: Numeric features that are not integers with fewer than 10 unique values are classified as numerical; otherwise, they are categorical.

Table 1: Heart Disease Dataset Feature Classification

Feature	**Discrete/Continuous**	**Quantitative/Qualitative**	**Numerical/Categorical**
age	Continuous	Quantitative	Numerical
sex	Discrete	Qualitative	Categorical
cp	Discrete	Qualitative	Categorical
trestbps	Continuous	Quantitative	Numerical
chol	Continuous	Quantitative	Numerical
fbs	Discrete	Qualitative	Categorical
restecg	Discrete	Qualitative	Categorical
thalach	Continuous	Quantitative	Numerical
exang	Discrete	Qualitative	Categorical
oldpeak	Continuous	Quantitative	Numerical
slope	Discrete	Qualitative	Categorical
ca	Discrete	Qualitative	Categorical
thal	Discrete	Qualitative	Categorical
target	Discrete	Qualitative	Categorical

Table 2: Wine Quality Dataset Feature Classification

Feature	**Discrete/Continuous**	**Quantitative/Qualitative**	**Numerical/Categorical**
fixed acidity	Continuous	Quantitative	Numerical
volatile acidity	Continuous	Quantitative	Numerical
citric acid	Continuous	Quantitative	Numerical
residual sugar	Continuous	Quantitative	Numerical
chlorides	Continuous	Quantitative	Numerical
free sulfur dioxide	Continuous	Quantitative	Numerical
total sulfur dioxide	Continuous	Quantitative	Numerical
density	Continuous	Quantitative	Numerical
pH	Continuous	Quantitative	Numerical
sulphates	Continuous	Quantitative	Numerical
alcohol	Continuous	Quantitative	Numerical
quality	Discrete	Quantitative	Categorical

2.2 Data Quality Assessment

2.2.1 Heart Disease Dataset

The heart disease dataset exhibited some missing values:

- 'ca' (number of major vessels): 4 missing values
- 'thal': 2 missing values
- All other features: No missing values

These missing values were appropriately handled during preprocessing to ensure data quality and analytical integrity.

2.2.2 Wine Quality Dataset

The wine quality dataset was found to be complete with no missing values across all 1,599 observations and 12 variables.

2.3 Target Variable Identification

2.3.1 Heart Disease Dataset

The target variable is located at column index 13, indicating the presence of heart disease:

- 0 = no heart disease
- 1, 2, 3, 4 = increasing levels of heart disease presence/severity

For analytical purposes, the target was treated as a binary variable, with all non-zero values consolidated to represent the presence of heart disease.

2.3.2 Wine Quality Dataset

The target variable 'quality' represents wine quality ratings on a scale from 3 to 8, making this an ordinal regression problem rather than a classification task.

3 Exploratory Data Analysis

3.1 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

3.1.1 Key Statistical Measures

Table 3: Mean and Variance for Numerical Features (Heart Disease Data)

Feature	Mean	Variance
age	54.438944	81.697419
trestbps	131.689769	309.751120
chol	246.693069	2680.849190
thalach	149.607261	523.265775
oldpeak	1.039604	1.348095

3.2 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.

3.2.1 Key Statistical Measures

Table 4: Mean and Variance for Features (Wine Quality Data)

Feature	Mean	Variance
fixed acidity	8.319637	3.031416
volatile acidity	0.527821	0.032062
citric acid	0.270976	0.037947
residual sugar	2.538806	1.987897
chlorides	0.087467	0.002215
free sulfur dioxide	15.874922	109.414884
total sulfur dioxide	46.467792	1082.102373
density	0.996747	0.000004
pH	3.311113	0.023835
sulphates	0.658149	0.028733
alcohol	10.422983	1.135647
quality	5.636023	0.652168

4 Correlation Analysis

4.1 Heart Disease Feature Correlations

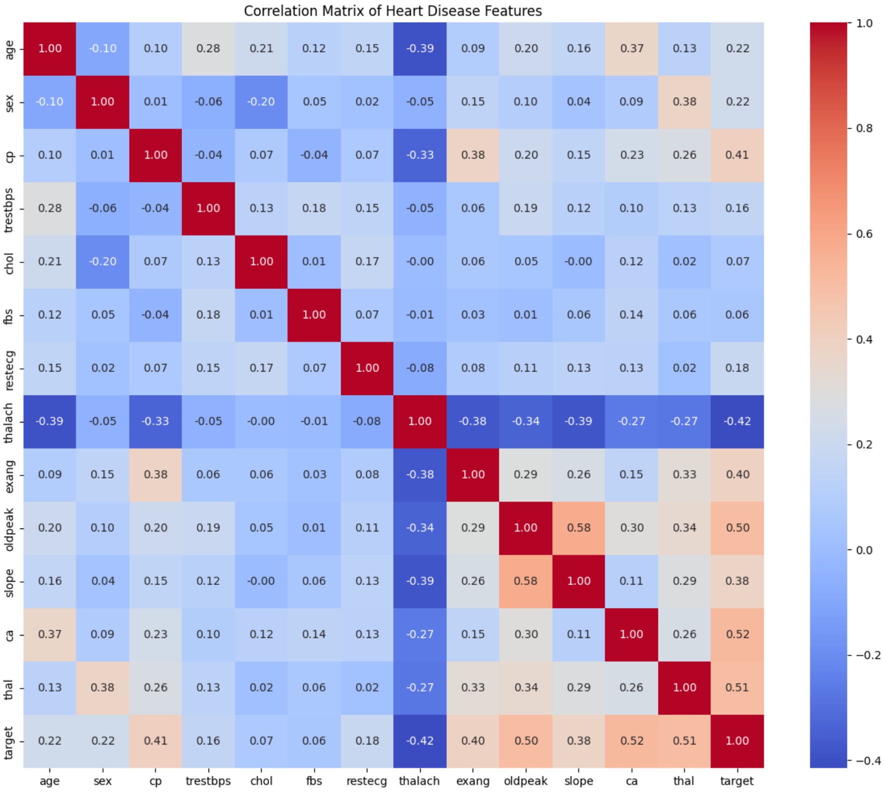
The correlation matrix visualization (Figure 1) reveals several significant relationships:

- **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.

Figure 1: Heart Disease Dataset Correlation Matrix



4.1.1 Most Correlated Feature Pair

The analysis identified that the most strongly correlated pair of features in the heart disease dataset is:

- Feature 1: ’oldpeak’ (ST depression induced by exercise)
- Feature 2: ’slope’ (the slope of the peak exercise ST segment)
- Correlation coefficient: 0.577537

This strong positive correlation aligns with clinical understanding, as both features relate to aspects of the ST segment in electrocardiograms, which are important indicators of heart function during stress testing.

4.2 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.

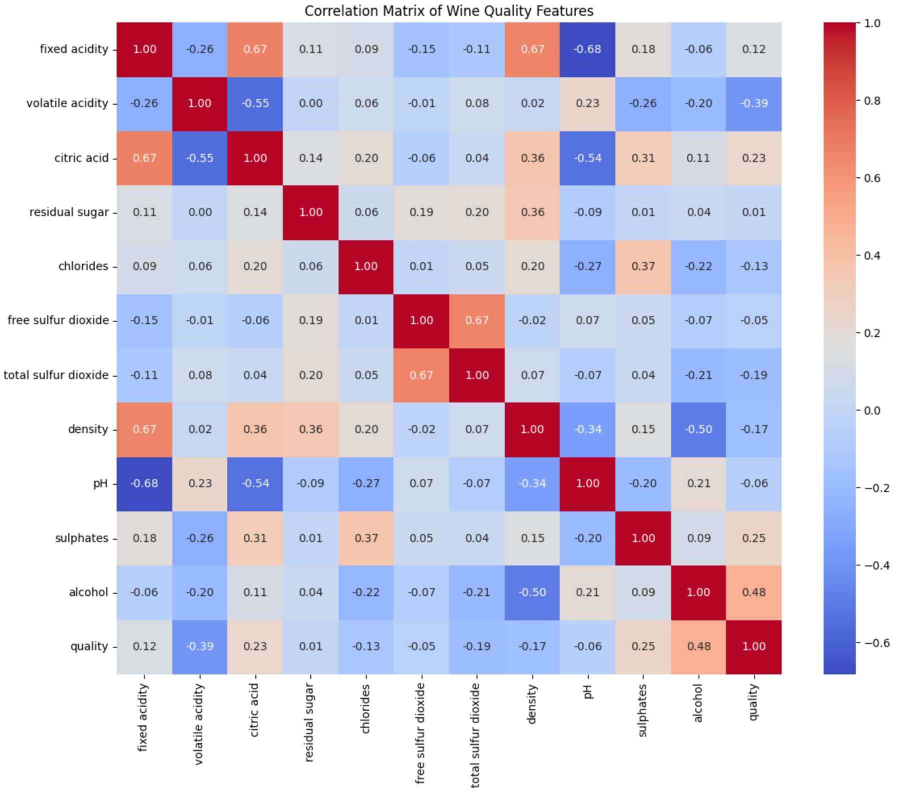
- **Chemical interactions**: Strong correlation (0.67) between fixed acidity and density, and between free and total sulfur dioxide (0.67).

- **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.

Figure 2: Wine Quality Dataset Correlation Matrix



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.05
total sulfur dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	1.00

4.2.1 Most Correlated Feature Pair

The analysis identified that the most strongly correlated pair of features in the wine quality dataset is:

- Feature 1: ’fixed acidity’

- Feature 2: ’pH’

- Correlation coefficient: -0.682978

This strong negative correlation is expected from a chemical perspective, as higher fixed acidity typically results in lower pH values (more acidic solutions). This relationship is fundamental to wine chemistry and affects both taste and stability.

5 Principal Component Analysis Results

5.1 PCA Methodology

Our approach to Principal Component Analysis focused on systematically evaluating component selection across both datasets:

- **Component Selection Methods**: We examined explained variance ratios, analyzed cumulative variance, tested specific component counts (2, 3, 5, 8), and calculated reconstruction errors to determine the optimal dimensionality.
- **Selection Criteria**: We prioritized components with high explained variance, used cumulative variance thresholds for sufficient representation, balanced reconstruction error against the number of components, and conducted visual inspection of 2D projections.
- **Comparative Analysis**: Rather than using fixed thresholds, we tested multiple component counts to understand the trade-off between dimensionality reduction and information preservation.

5.2 Variance Explanation

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

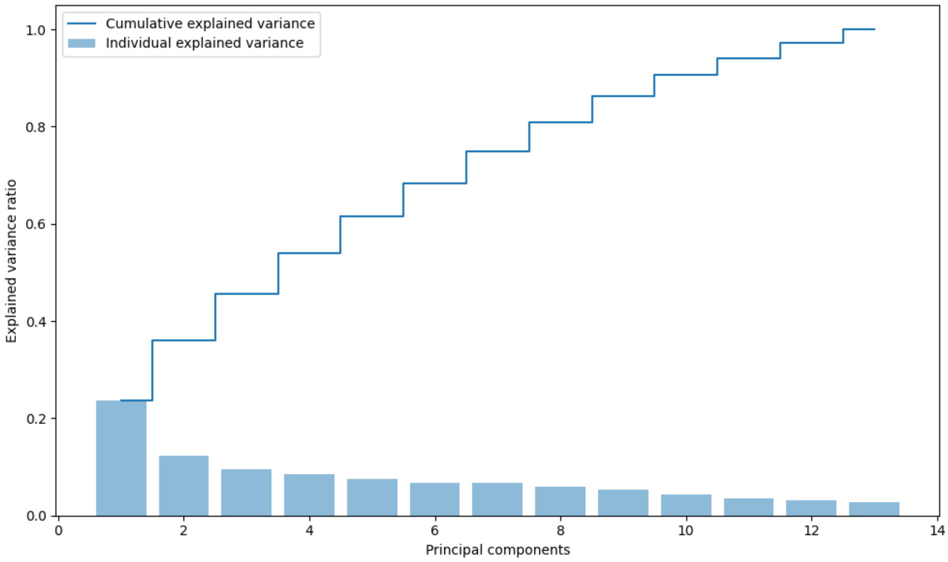
5.2.1 Heart Disease Data

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

5.2.2 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

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



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.

5.3 PCA Projections

5.3.1 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 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.

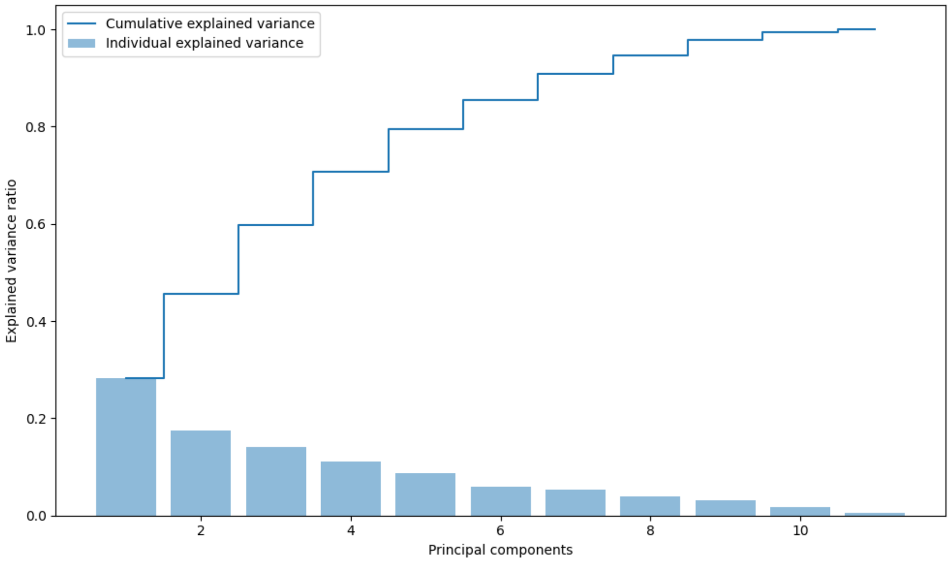
- **Least significant PCs** (Figure 6): PC12 and PC13 (explaining just 3.16% and 2.72% of variance respectively) show minimal separation between classes, with points from both classes thoroughly mixed throughout the projection space. This confirms these components capture primarily noise or very specific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

5.3.2 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.

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



- **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.

5.4 Highest vs. Lowest Principal Components

5.4.1 Heart Disease Data

Our comparative analysis of the highest and lowest principal components revealed several key differences:

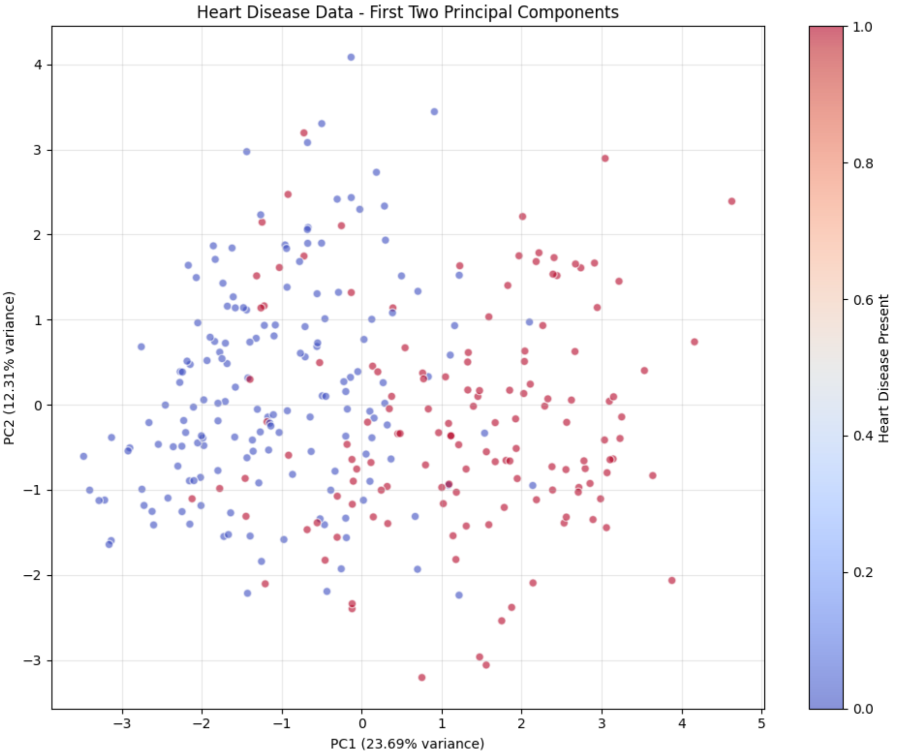
- **Information Content**: The first few principal components captured most of the dataset's variability, while the lowest principal components contained minimal information about overall data patterns.
- **Classification Utility**: When visualizing data with the top components, clearer separation between heart disease cases was evident, while the least significant components showed poor separation.
- **Reconstruction Error**: Higher components were essential for accurate data reconstruction, while lower components contributed minimally.

5.4.2 Wine Quality Data

Similar patterns were observed in the wine quality dataset:

- **Information Capture**: The first principal components (PC1, PC2) captured the largest proportion of variance, while the lowest components captured only a very small portion of the total variance.

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



- **Discrimination Ability**: The scatter plot using the top two principal components showed better ability to differentiate wine quality levels, while the plot using the least significant components demonstrated minimal differentiation.
- **Practical Value**: High-value principal components enabled effective dimensionality reduction while preserving important information, whereas low-value components typically contained noise or unimportant features.

5.5 Reconstruction Error Analysis

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

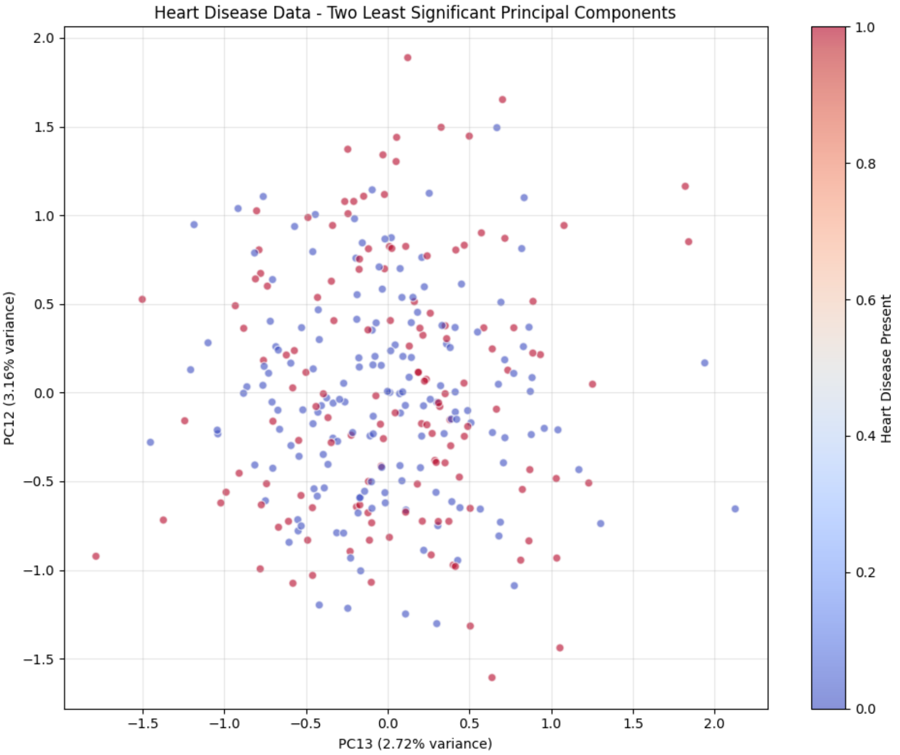
5.5.1 Heart Disease Data

- 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

5.5.2 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)

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



- 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.

6 Dimensionality Reduction Implications

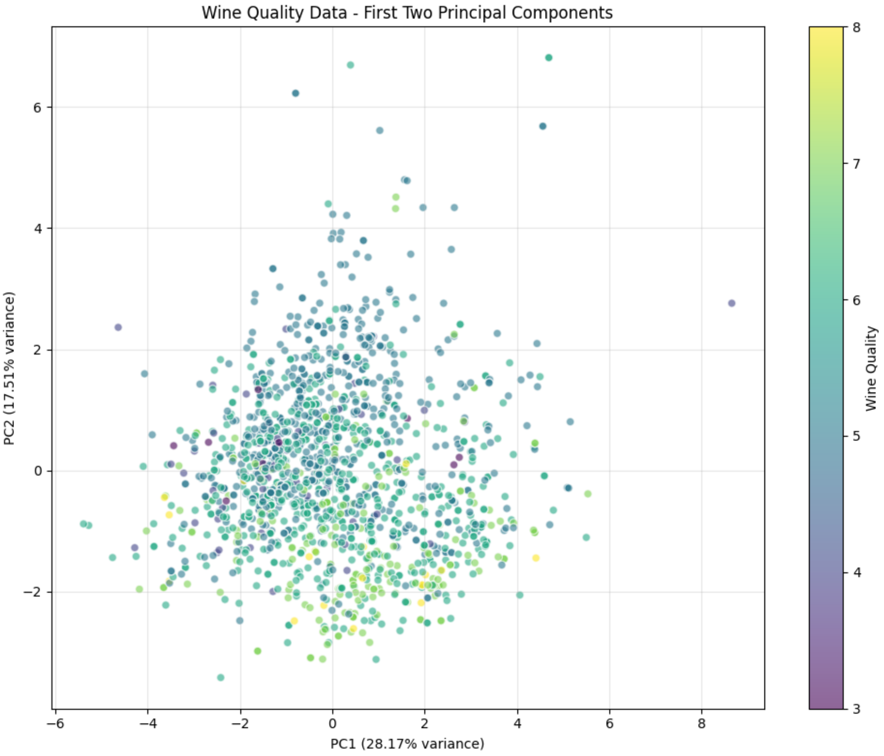
6.1 Optimal Component Selection

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

6.1.1 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

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



6.1.2 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

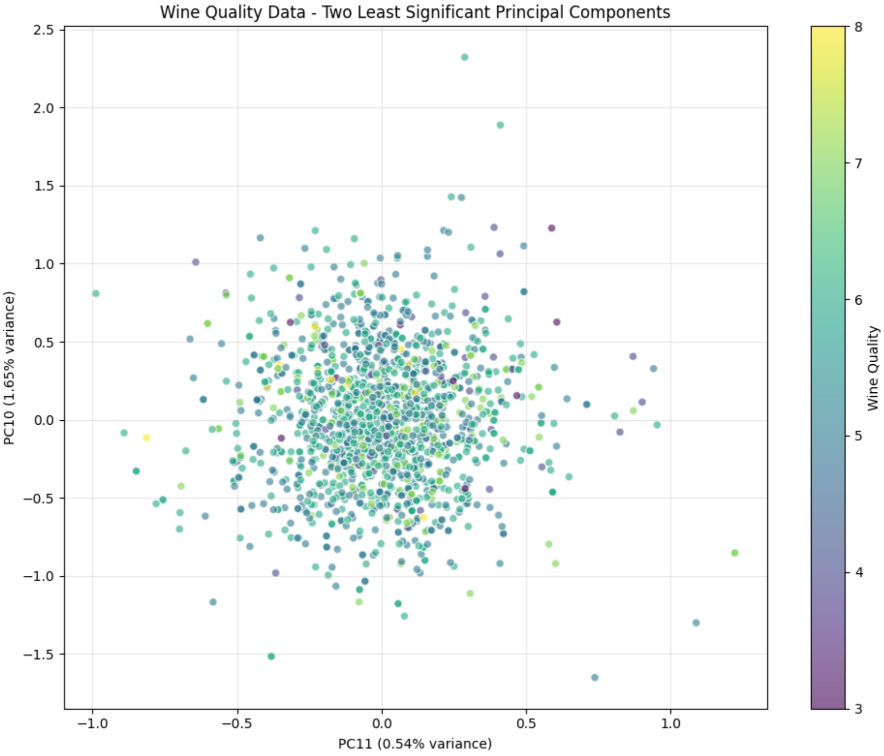
6.2 Feature Importance Insights

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

6.2.1 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

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



6.2.2 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

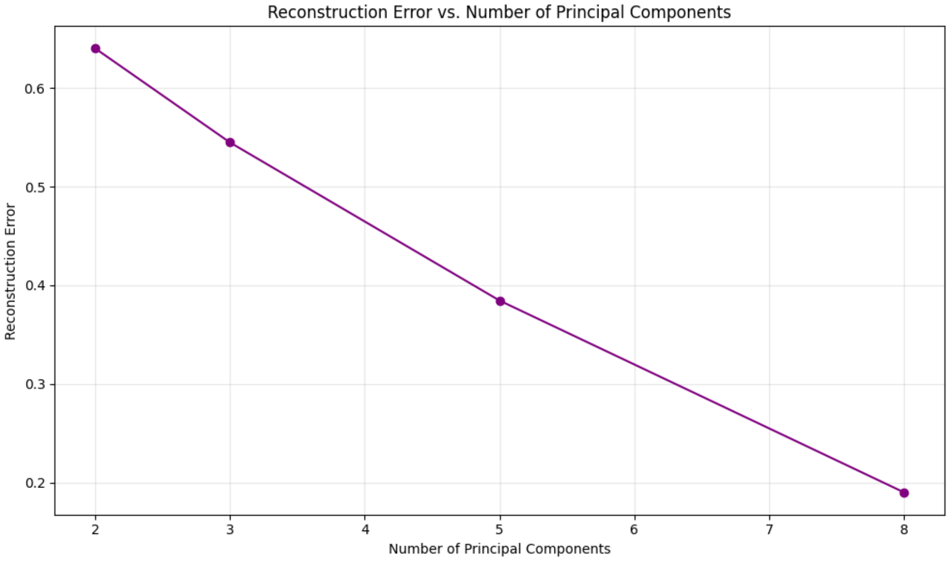
6.3 Data Structure Comparison

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.

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



7 Conclusions and Recommendations

7.1 Model Development

7.1.1 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

- 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

7.1.2 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)

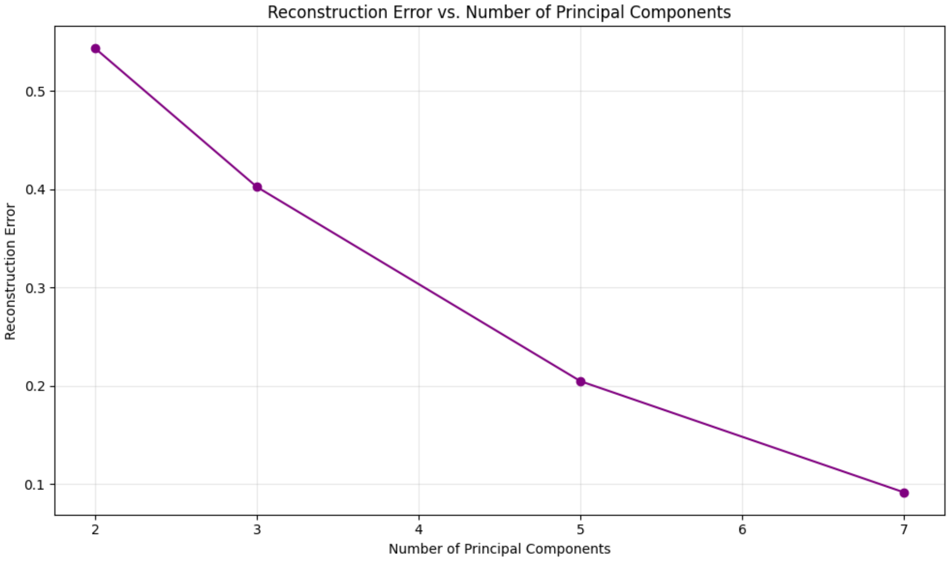
- The gradual blending of quality levels in the PCA projection suggests that precise quality prediction may be challenging

7.2 Feature Engineering

7.2.1 Heart Disease

- Focus on heart rate, major vessels data, and thalassemia indicators as primary inputs

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



- 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
- The significant relationship between 'oldpeak' and 'slope' (correlation of 0.577537) suggests creating a derived feature that captures this interaction

7.2.2 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 (−0.682978) visible in Figure 2
- The strong correlation between free and total sulfur dioxide suggests using their ratio rather than absolute values
- Feature engineering should focus on chemical balance aspects rather than individual components

7.3 Component Selection Trade-offs

Our analysis revealed important trade-offs in dimensionality reduction decisions:**Variance retention vs. dimensionality**: For heart disease data, using 5 components retains approximately 70% of variance with significant dimensionality reduction, while 8 components capture 90% but with less reduction. **Reconstruction error vs. model complexity**: Wine quality data showed steep error reduction up to 5 components (error of 0.20), with diminishing returns for additional components. This suggests 5 components as a reasonable compromise between accuracy and simplicity. **Interpretability considerations**: Using fewer components generally improves interpretability but may sacrifice predictive accuracy. For medical applications like heart disease prediction, higher accuracy (using more components) may be warranted.

Computational efficiency: Lower-dimensional representations reduce computational requirements for subsequent modeling, particularly beneficial for the larger wine quality dataset (1,599 observations) compared to the heart disease dataset (303 observations).

7.4 Domain-Specific Insights

7.4.1 Medical Applications

The heart disease PCA analysis provides several clinically relevant insights:

- **Diagnostic efficiency**: Using 6-8 principal components may streamline clinical assessments by focusing on the most informative measurements.
- **Risk factor identification**: The strong correlations of 'thalach', 'ca', and 'thal' with heart disease status confirm these as significant risk indicators, aligning with medical literature.
- **Age-related patterns**: The correlation between age and target (0.37) reaffirms age as a significant risk factor, but not definitive on its own.
- **Stress test importance**: The strong correlation between 'oldpeak' and 'slope' highlights the value of exercise stress testing in heart disease diagnosis.

7.4.2 Wine Industry Applications

For the wine quality dataset, PCA results offer valuable insights for production and quality control:

- **Quality determinants**: Alcohol content, volatile acidity, and sulphates emerge as the most significant factors affecting perceived wine quality.
- **Chemical balance**: The strong negative correlation between fixed acidity and pH (-0.68) emphasizes the importance of acid balance in wine quality.
- **Production guidance**: Wine producers could focus quality improvement efforts on optimizing alcohol content and minimizing volatile acidity.
- **Quality prediction**: Using 5-7 principal components could enable efficient quality prediction systems for quality control processes.

8 Limitations and Future Work

8.1 Current Limitations

- **Linear assumptions**: PCA assumes linear relationships between variables, which may not fully capture complex nonlinear interactions, particularly in biological systems like heart disease pathophysiology.
- **Categorical variables**: Both datasets contain categorical variables, which PCA does not optimally handle. Alternative methods like multiple correspondence analysis might be more appropriate for categorical data.
- **Sample size constraints**: The heart disease dataset's relatively small sample size (303 observations) may affect the stability of PCA results.

- **Outlier sensitivity**: PCA is sensitive to outliers, which may disproportionately influence component directions. Our analysis did not include robust outlier detection and treatment.
- **Interpretability challenges**: While PCA reduces dimensionality effectively, the resulting principal components lack direct physical interpretability, making it difficult to translate findings directly to practical recommendations.

8.2 Future Research Directions

- **Nonlinear dimensionality reduction**: Exploring nonlinear techniques such as t-SNE or UMAP could potentially reveal more complex patterns, particularly in the heart disease dataset where biological relationships may be nonlinear.
- **Feature-specific PCA**: Conducting separate analyses for continuous and categorical variables might yield more nuanced insights, especially for the heart disease dataset with its mix of variable types.
- **Comparative modeling**: Developing predictive models using both original features and PCA-reduced features to quantify the practical impact of dimensionality reduction on predictive performance.
- **Time series extension**: For wine quality assessment, incorporating temporal data from the fermentation process could enhance understanding of how quality develops over time.
- **Cross-validation**: Implementing cross-validation in PCA component selection would improve the robustness of dimensionality decisions, particularly for the smaller heart disease dataset.
- **Hybrid approaches**: Combining PCA with domain-specific feature engineering might yield more interpretable and powerful predictive features for both medical diagnostics and wine quality assessment.

9 Summary and Key Takeaways

9.1 Heart Disease Dataset

- **Optimal dimensionality**: 6-8 principal components (from original 13) capture 80-90% of variance while maintaining reasonable reconstruction error (0.19 at 8 components).
- **Key predictors**: Maximum heart rate ('thalach'), number of major vessels ('ca'), ST depression ('oldpeak'), and thalassemia indicator ('thal') emerge as the most significant predictors of heart disease.
- **Visualization insights**: PCA projections show partial separation between disease and non-disease cases, suggesting PCA-based dimensionality reduction preserves clinically relevant information.
- **Clinical application**: The PCA results suggest a potential streamlined diagnostic approach focusing on the most informative measurements, potentially reducing the number of tests needed for initial screening.

9.2 Wine Quality Dataset

- **Optimal dimensionality**: 5-7 principal components (from original 11) capture 80-90% of variance with low reconstruction error (0.09 at 7 components).

- **Quality indicators**: Alcohol content, volatile acidity, and sulphates demonstrate the strongest relationships with wine quality ratings.
- **Continuous nature**: PCA projections reveal wine quality as a continuous spectrum rather than discrete categories, emphasizing the subjective and gradualistic nature of quality assessment.
- **Industry application**: The analysis suggests that wine producers could focus quality improvement efforts on specific chemical parameters, particularly optimizing alcohol content and minimizing volatile acidity.

9.3 Comparative Insights

- **Dataset structure differences**: Heart disease data shows more discrete clustering (categorical outcome), while wine quality demonstrates a continuous gradient (ordinal outcome).
- **Variance concentration**: Wine quality data shows more variance concentrated in fewer components (28% in PC1) compared to heart disease data (23% in PC1), suggesting more pronounced correlations among wine chemical properties.
- **Dimensionality reduction potential**: Both datasets demonstrate similar dimensionality reduction potential (approximately 50%), despite their different domains and structures.
- **Methodology applicability**: PCA proves effective for dimensionality reduction across both medical diagnostic and product quality assessment domains, demonstrating its versatility.

10 References

- UCI Machine Learning Repository. (2019). Heart Disease Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease
- UCI Machine Learning Repository. (2009). Wine Quality Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/wine+quality