Analyzing Factors Affecting Red Wine Quality

Buse Tolunay
Dağlar Eren Tekşen
Yusuf Aygün

Project Description

In this project, we aim to investigate the relationships between the **chemical properties** of red wines and their **quality ratings**. The dataset we are using contains physicochemical properties of wine such as **alcohol content**, **acidity**, and **sugar levels**, alongside **a quality score** assigned by wine experts. The goal is to build predictive models using data analysis and machine learning techniques to understand how these features impact **wine quality**.

We will apply both **regression models** and **unsupervised clustering techniques** to identify patterns in the data, and use feature importance analysis to determine which chemical properties are most significant in **predicting wine quality**.

The dataset, sourced from Kaggle, consists of 1599 red wine samples with 12 chemical attributes:

These features will allow us to explore both **linear** and **non-linear** relationships between the chemical compositions of wine and their quality ratings. It offers a rich dataset for applying both **regression models** and **clustering algorithms**, providing a good platform for dimensionality reduction and feature importance analysis.

01. Fixed Acidity

02. Volatile Acidity

03. Citric Acid

04. Residual Sugar

05. Chlorides

06. Free Sulfur Dioxide

07. Total Sulfur Dioxide

08. Density

09. pH

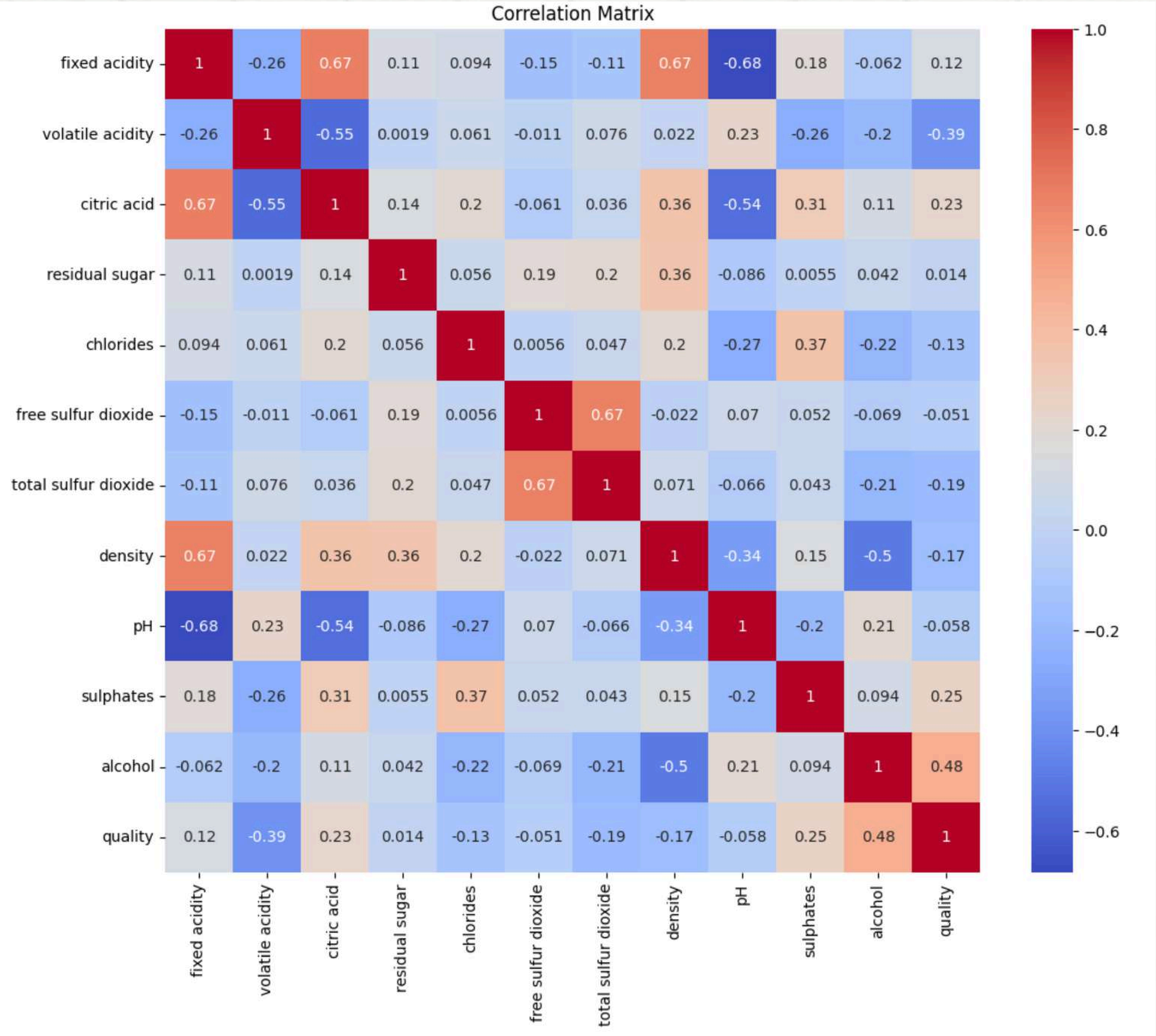
10. Sulphates

11. Alcohol

12. Quality Score

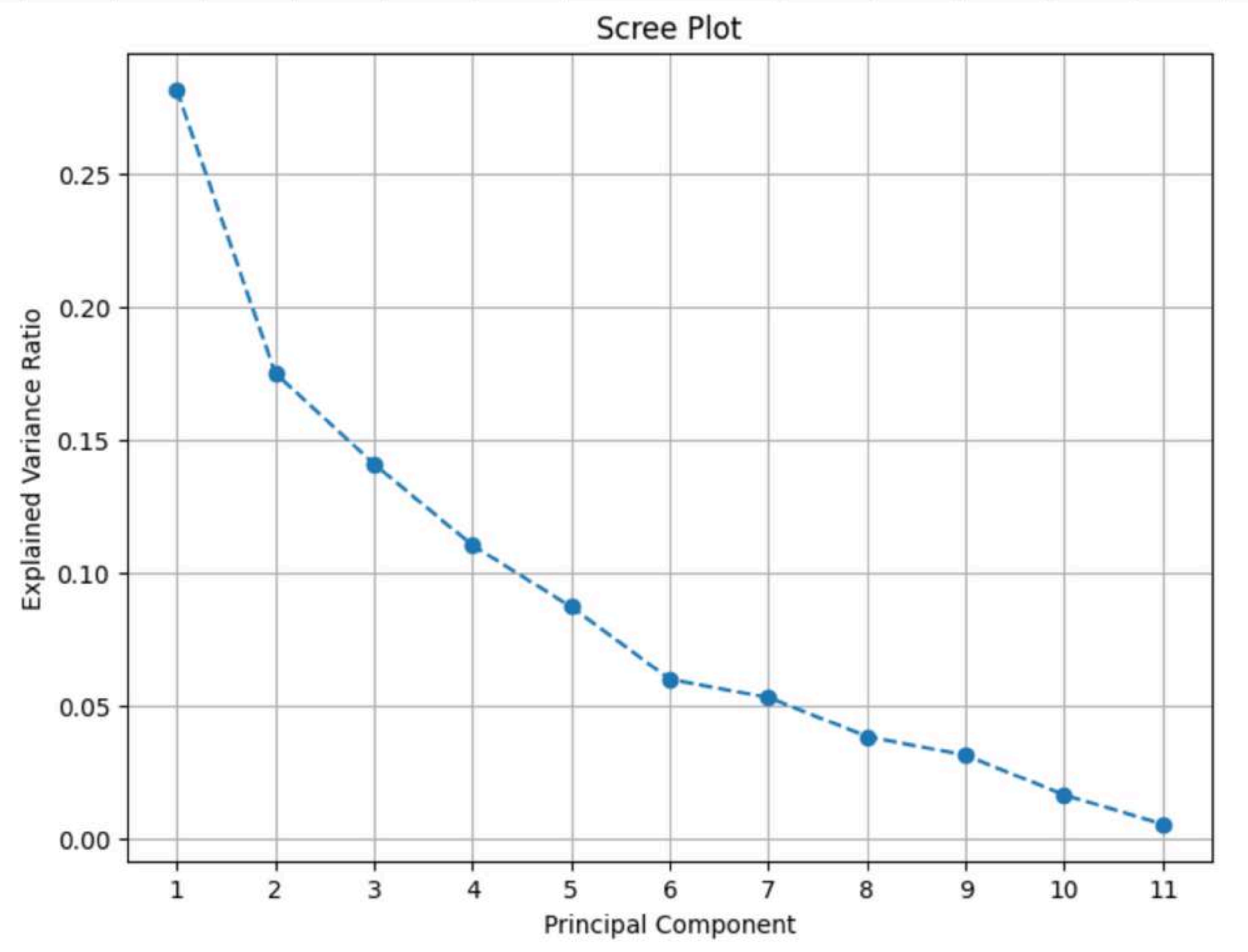
Exploratory Data Analysis (EDA)

- Correlation Matrix:
  - Highlights strong correlations among features.
  - Example: Density correlates positively with fixed acidity (0.67).
  - Quality correlates moderately with alcohol (0.48) and sulphates (0.25).
- Observations:
  - Alcohol and volatile acidity are significant predictors for quality.
  - High multicollinearity observed (e.g., citric acid vs fixed acidity).



Correlation Matrix

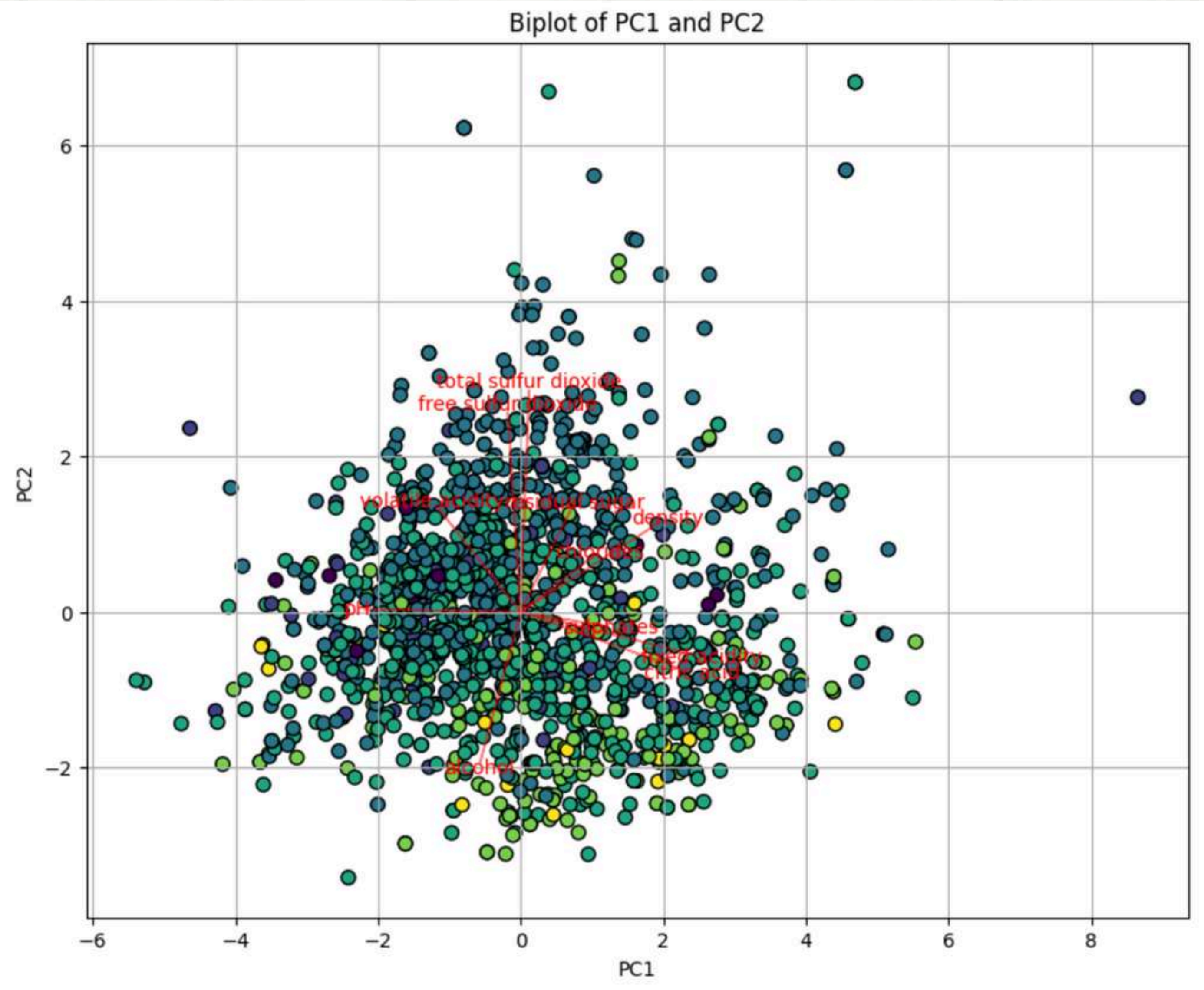
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

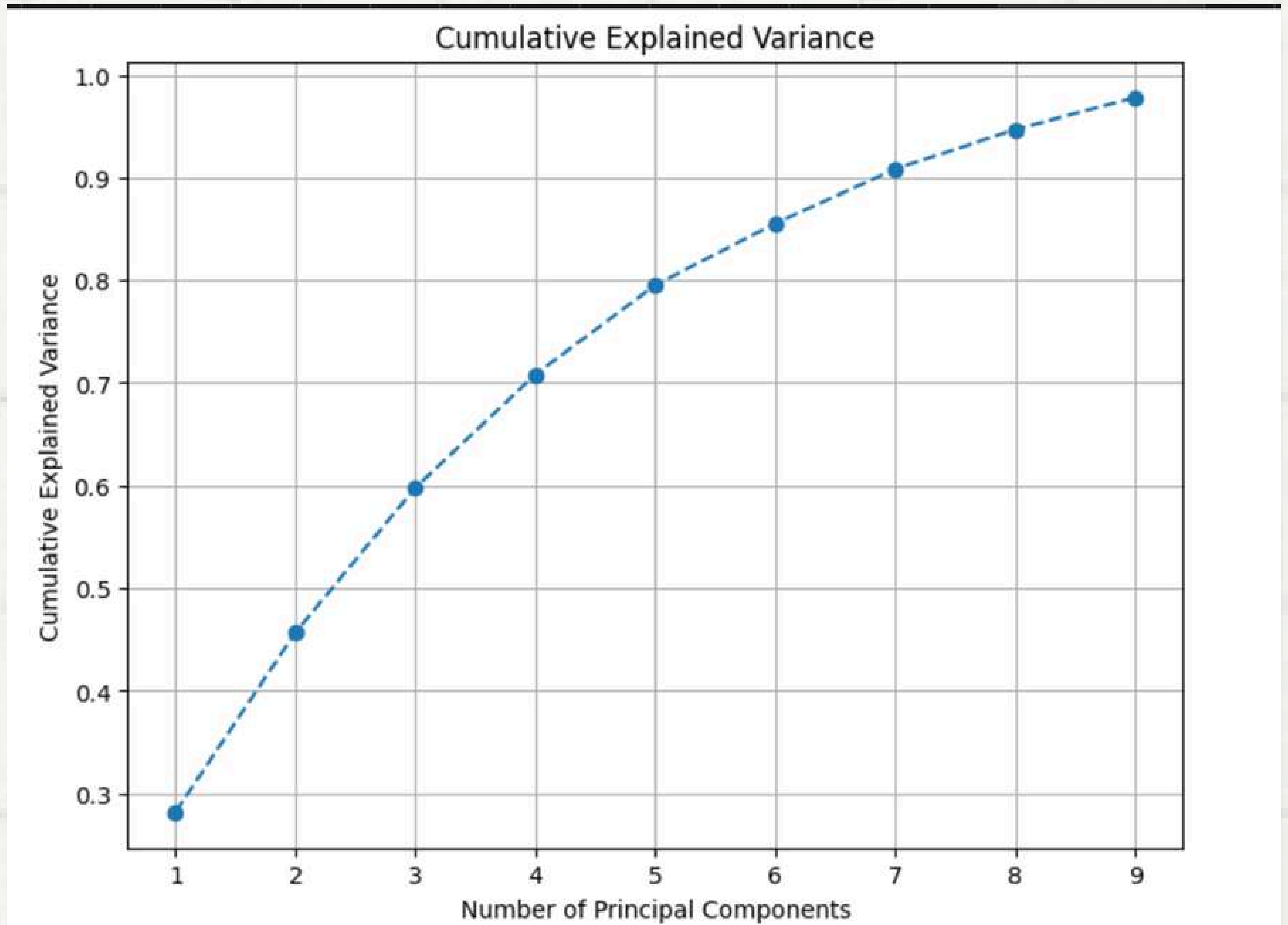


Dimensionality Reduction with PCA

PCA Scree Plot

- The scree plot shows a sharp drop-off after the first three components, capturing 59.78% of the variance.
- Retaining nine components explains 95% of the variance, balancing information retention with reduced dimensionality.





PCA Biplot

- PC1 and PC2 together explain 45.68% of the variance.
- Key insights:
  - Alcohol and density align closely with PC1, suggesting a strong influence.
  - Volatile acidity shows a negative relationship in PC1, consistent with its correlation to wine quality.
  - Citric acid and chlorides cluster in PC2, indicating a distinct influence compared to PC1.

Linear Regression

Model Training and Optimization

- The Linear Regression model was trained using both PCA-transformed data and the original feature set for comparison.
- Training and Test Split:
    - **80%** of the data was used for training, and **20%** for testing.
- Steps:
    a. Data normalization was applied to standardize the features.
    b. For the PCA approach, dimensionality was reduced to retain **95%** of the variance, resulting in 9 principal components.
    c. For the non-PCA approach, all original features were retained.
    d. The Linear Regression model was trained separately on the PCA-transformed and original training data.
- Performance Evaluation:
    - The model was evaluated using three metrics:
        - **Mean Squared Error (MSE):** Measures the average error in predictions.
        - **Mean Absolute Error (MAE):** Reflects the absolute average error in predictions, making it more interpretable.
        - **R-squared (R^2):** Indicates the proportion of variance in wine quality explained by the model.
- No hyperparameter optimization was applied since Linear Regression has no tunable parameters.

Model Performance

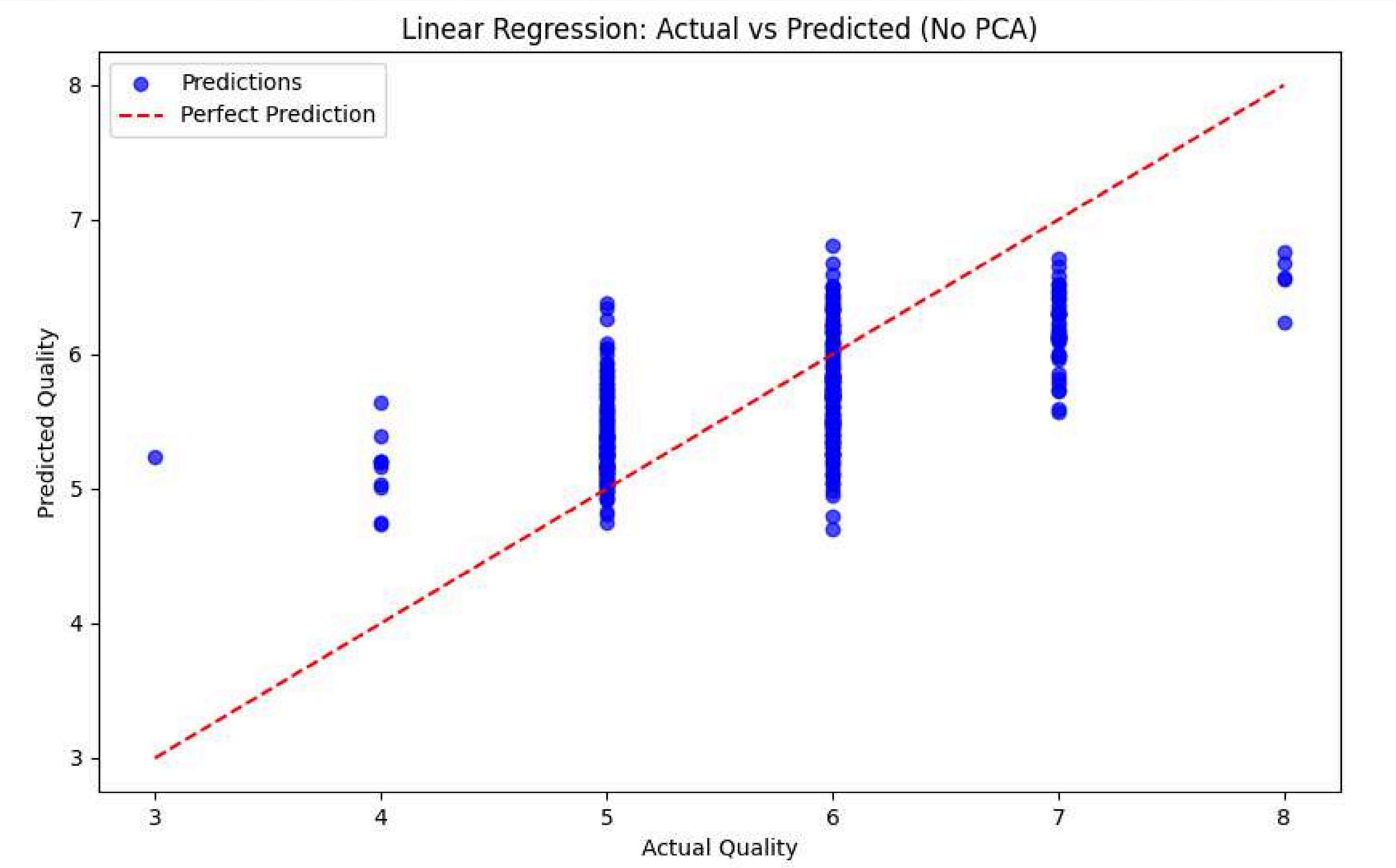
With PCA:

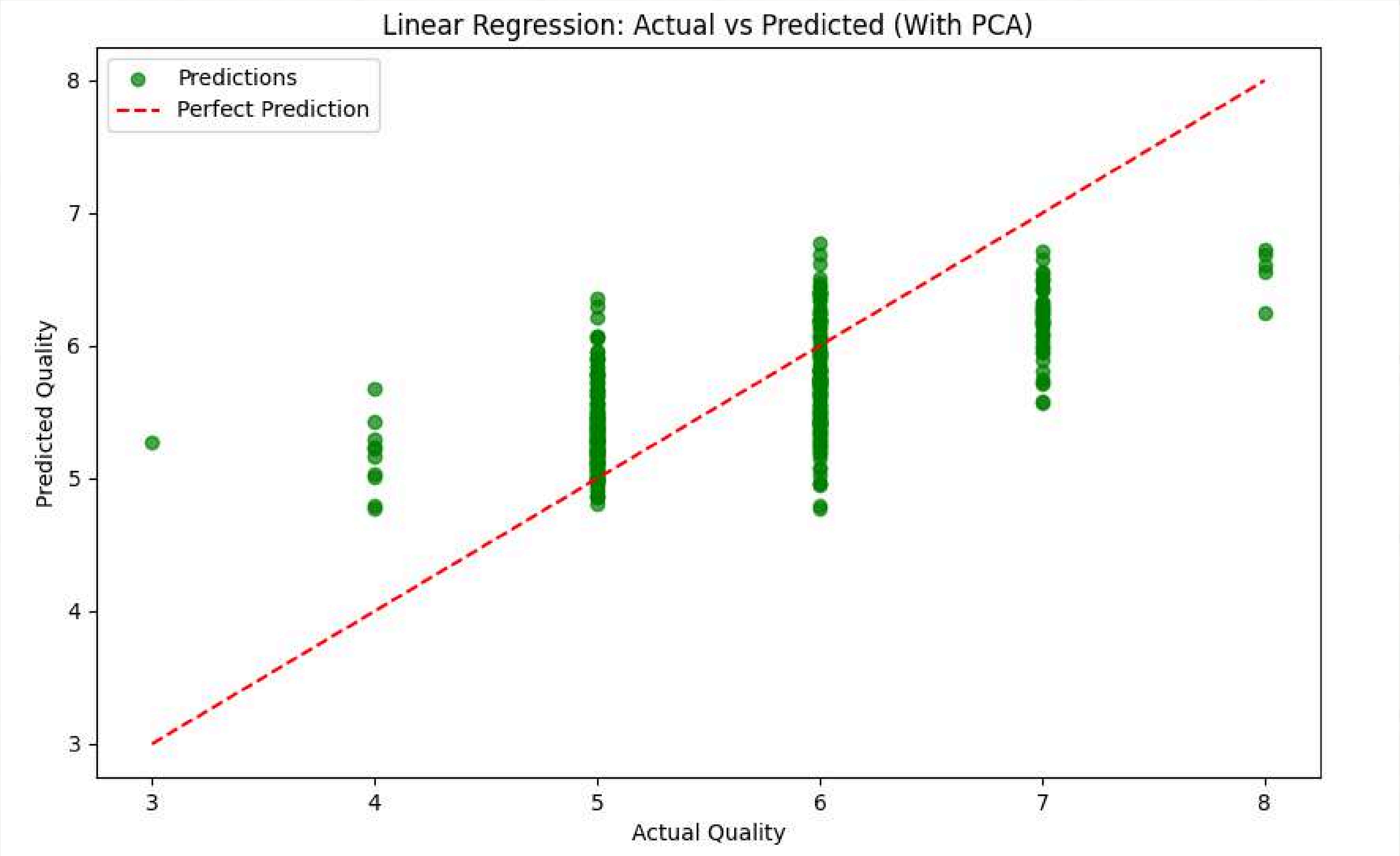
- **Mean Squared Error (MSE): 0.3926**
  - The model achieved moderate accuracy, capturing general trends but struggling with extreme values.
- **Mean Absolute Error (MAE): 0.5046**
  - On average, the predictions deviated by **~0.50** quality points.
- **R-squared (R^2): 0.3992**
  - The model explains approximately **39%** of the variance in wine quality.

Without PCA:

- **Mean Squared Error (MSE): 0.3900**
  - Lower error compared to the PCA version, benefiting from retaining the original feature interactions.
- **Mean Absolute Error (MAE): 0.5035**
  - Predictions deviated by **~0.50** quality points on average, slightly better than the PCA version.
- **R-squared (R^2): 0.4032**
  - The model explains approximately **40%** of the variance, showing improved performance without dimensionality reduction.

Actual vs Predicted Quality

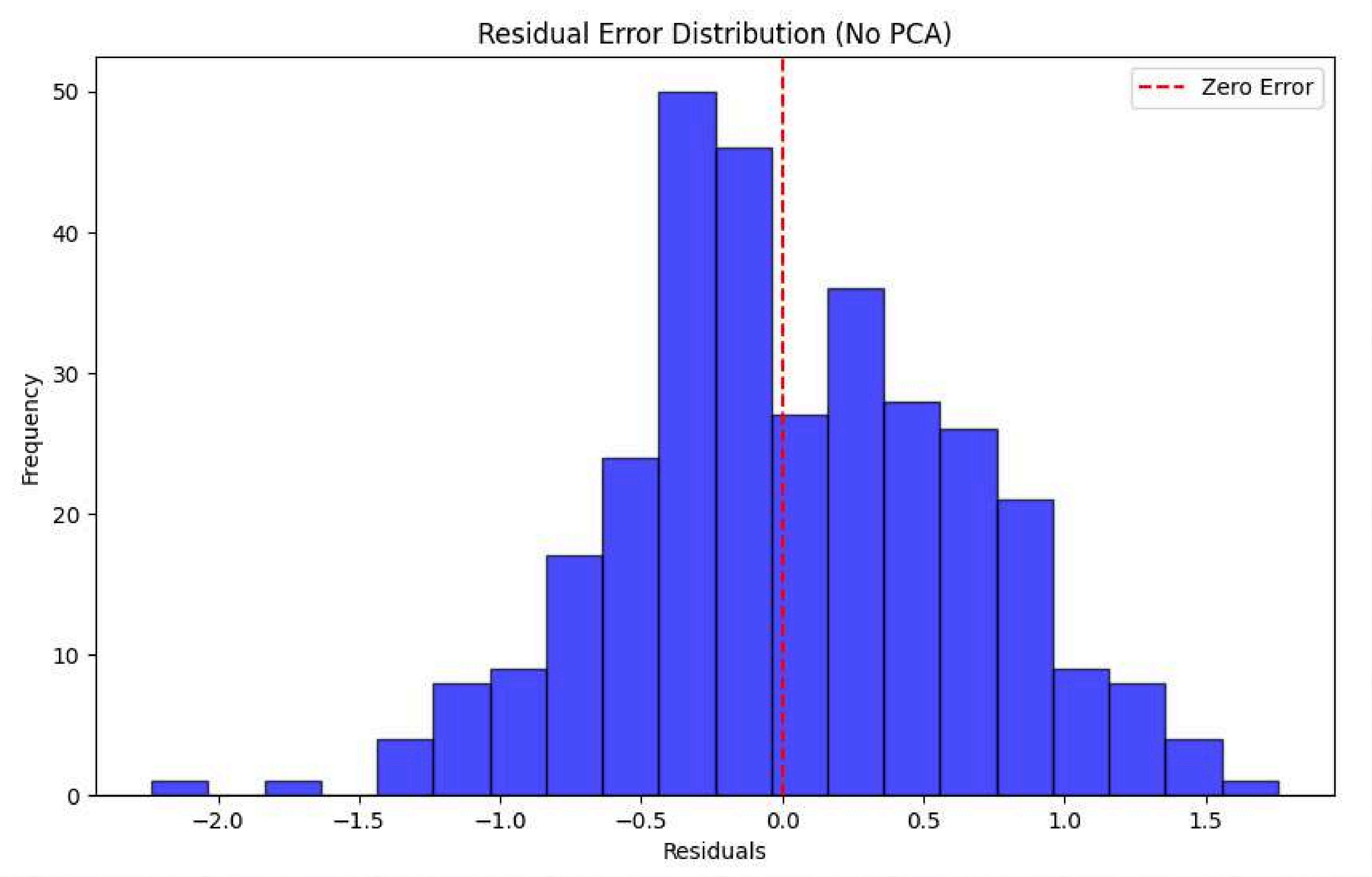


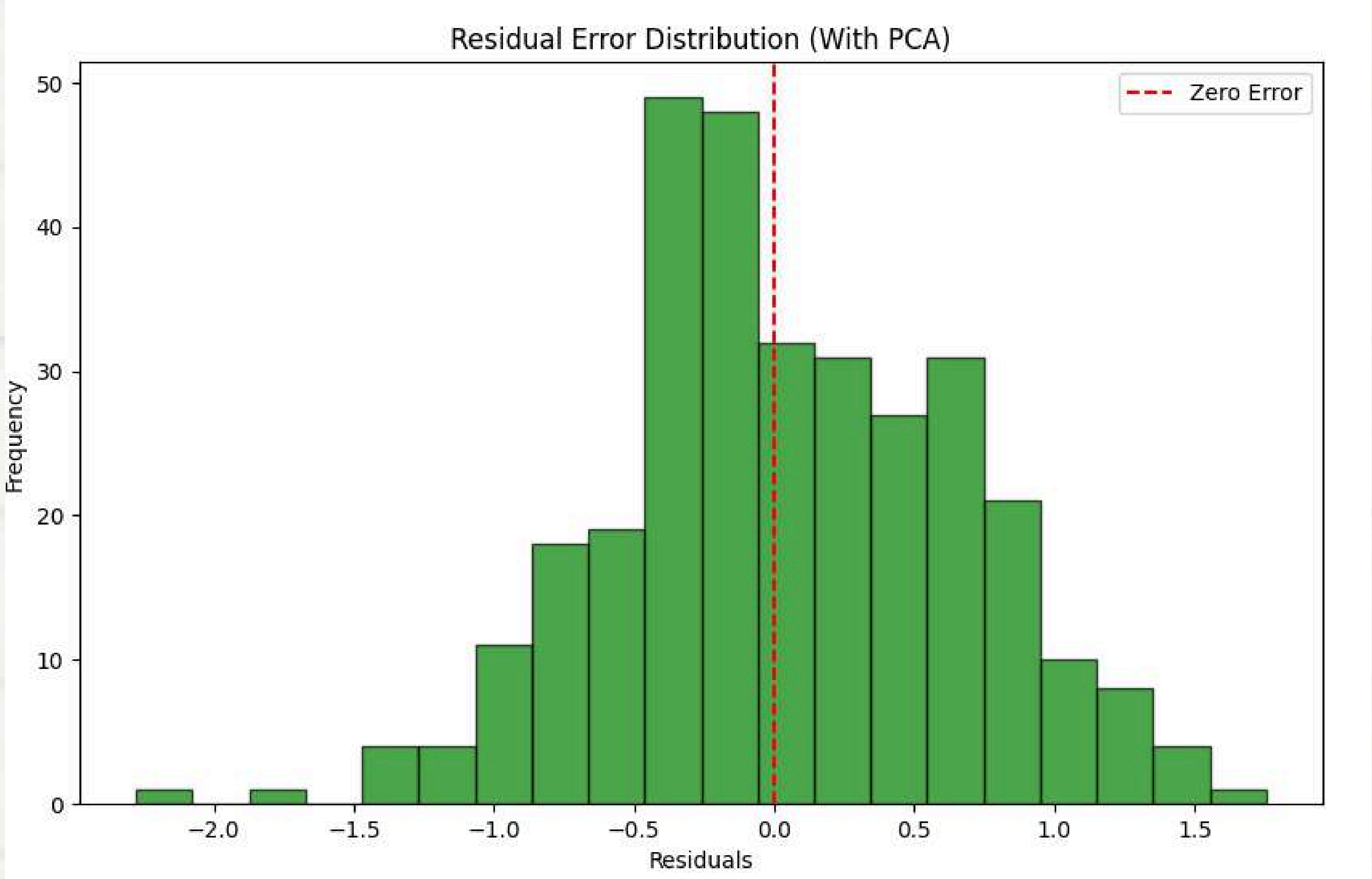


- Predictions align better with actual values compared to the PCA version, particularly for quality scores around the mean.
- Predictions are **closer to the actual values for average quality** wines (5 and 6).
- Higher quality (7–8) and lower quality (3–4) predictions deviate significantly from the actual values.
- Retaining all features without transformation benefits linear regression, given its sensitivity to changes in feature relationships.

- Predictions for rare quality scores (e.g., 3 and 8) still show a larger deviation, highlighting the model's limitation in imbalanced datasets.
- PCA simplifies feature space, but this transformation loses the interaction context, leading to reduced accuracy.

Residual Distribution

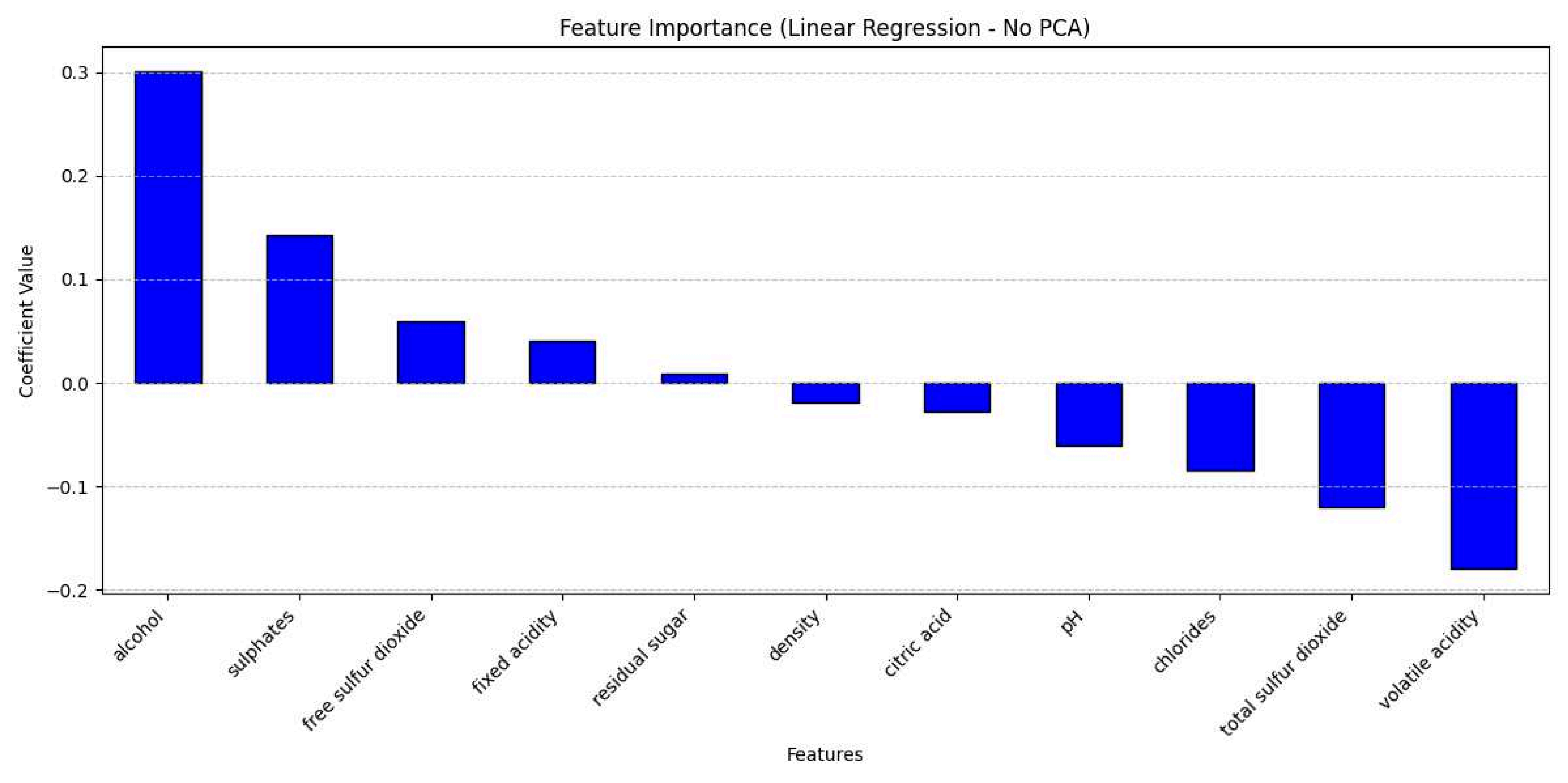




- Residuals are symmetrically distributed around zero, indicating no systematic bias in the model.
- Fewer outliers indicate the model performs better on the original feature set.
- Without PCA, the model has a better fit, particularly for the majority of the dataset.

- Residuals are still centered around zero but with a broader spread compared to the PCA version.
- A broader spread of residuals suggests higher errors for extreme quality scores.
- The PCA transformation introduces some noise, as evident from the wider residual range.

Feature Importance



- **Alcohol and Sulphates contribute the most positively** to wine quality predictions, reflecting their importance in determining sensory scores.

- Free Sulfur Dioxide also shows a positive association, albeit weaker.

- Volatile Acidity and Total Sulfur Dioxide have negative coefficients, aligning with their tendency to degrade wine quality perception.

- Features like pH, Citric Acid, and Residual Sugar contribute minimally, indicating limited impact in this dataset.

Key Insights About Model

- The model **performs best for mid-quality** wines (5-7) but **struggles with extreme cases** (3 and 8). Without PCA, the model retains original feature interactions, improving its performance for mid-range predictions. With PCA, performance slightly decreases due to the removal of feature interactions critical for accurate predictions.
- Key features like alcohol and acidity were critical in both approaches, but PCA limited their interactions.

Strengths:

- **Simplicity:** Linear Regression is easy to implement and interpret.
- **Speed:** The model trains quickly, even with large datasets.
- **Baseline Performance:** Provides a benchmark to compare with more complex models.

Limitations:

- **Limited Accuracy:** The model explains only 40% of the variance ($R^2 = 0.4032$), leaving room for improvement.
- **Linear Assumptions:** Assumes linear relationships, which might not capture the complexity of the data.
- **Edge Case Performance:** Struggles with extreme values (e.g., wine quality = 3 or 8).
- **Outliers Sensitivity:** Residual analysis shows sensitivity to outliers, which can affect predictions.

Linear Regression Conclusion

- Linear Regression achieved moderate accuracy with 40% of variance explained ($R^2 = 0.4032$) in the both non-PCA approach in the PCA-transformed approach ($R^2 = 0.3992$).
- While PCA simplified the data and preserved key features like alcohol and acidity, it limited the model’s ability to capture critical feature interactions, resulting in slightly reduced performance.
- The model's limitations include difficulty handling edge cases and moderate residual errors.
- Future improvements could involve:
  - Trying non-linear models for better edge-case predictions.
  - Exploring additional features, such as sensory data, to enhance the model.

Random Forest Regression

What is Random Forest Regression?

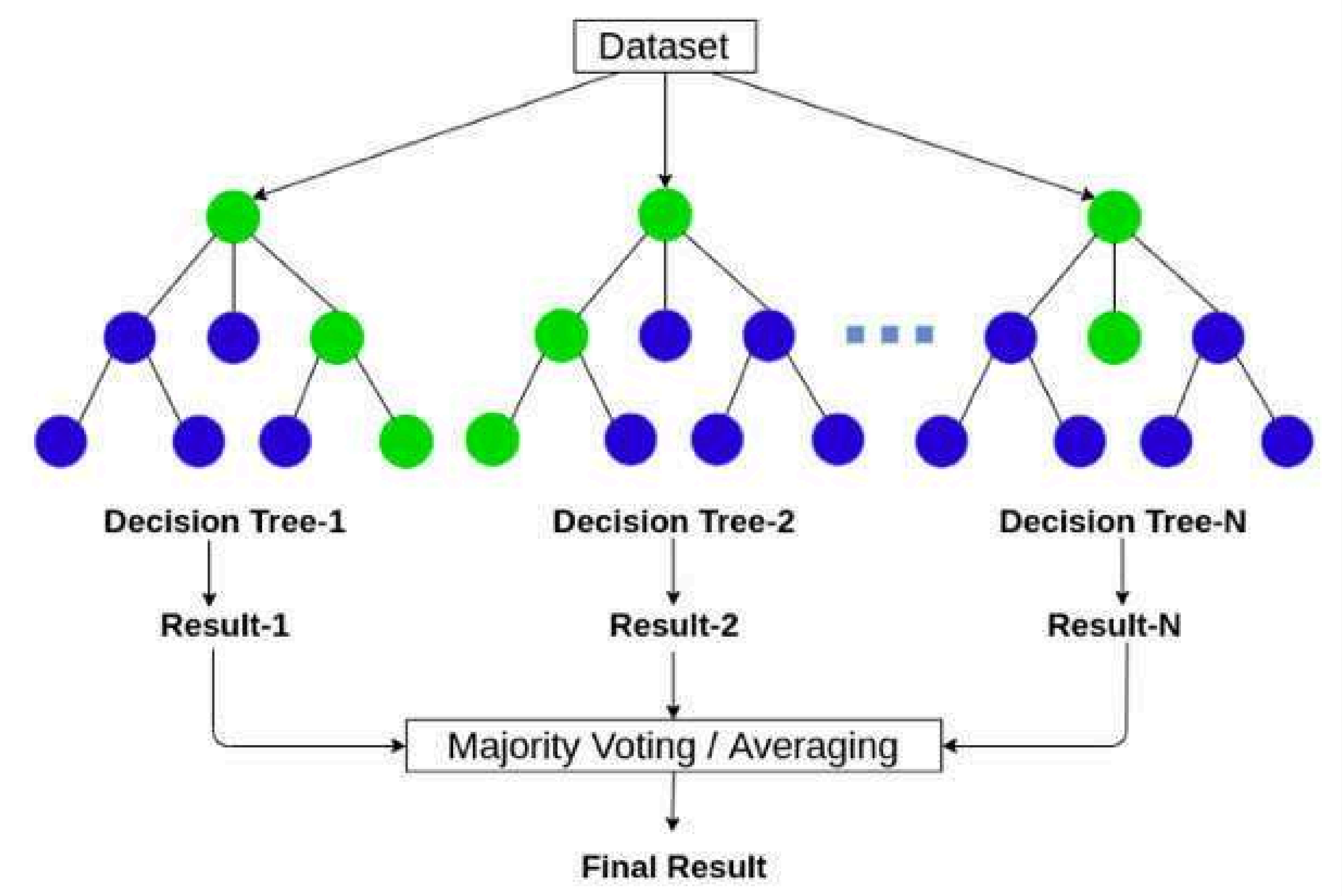
- An ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.
- Builds trees on random subsets of data and features, then aggregates their predictions (average for regression).

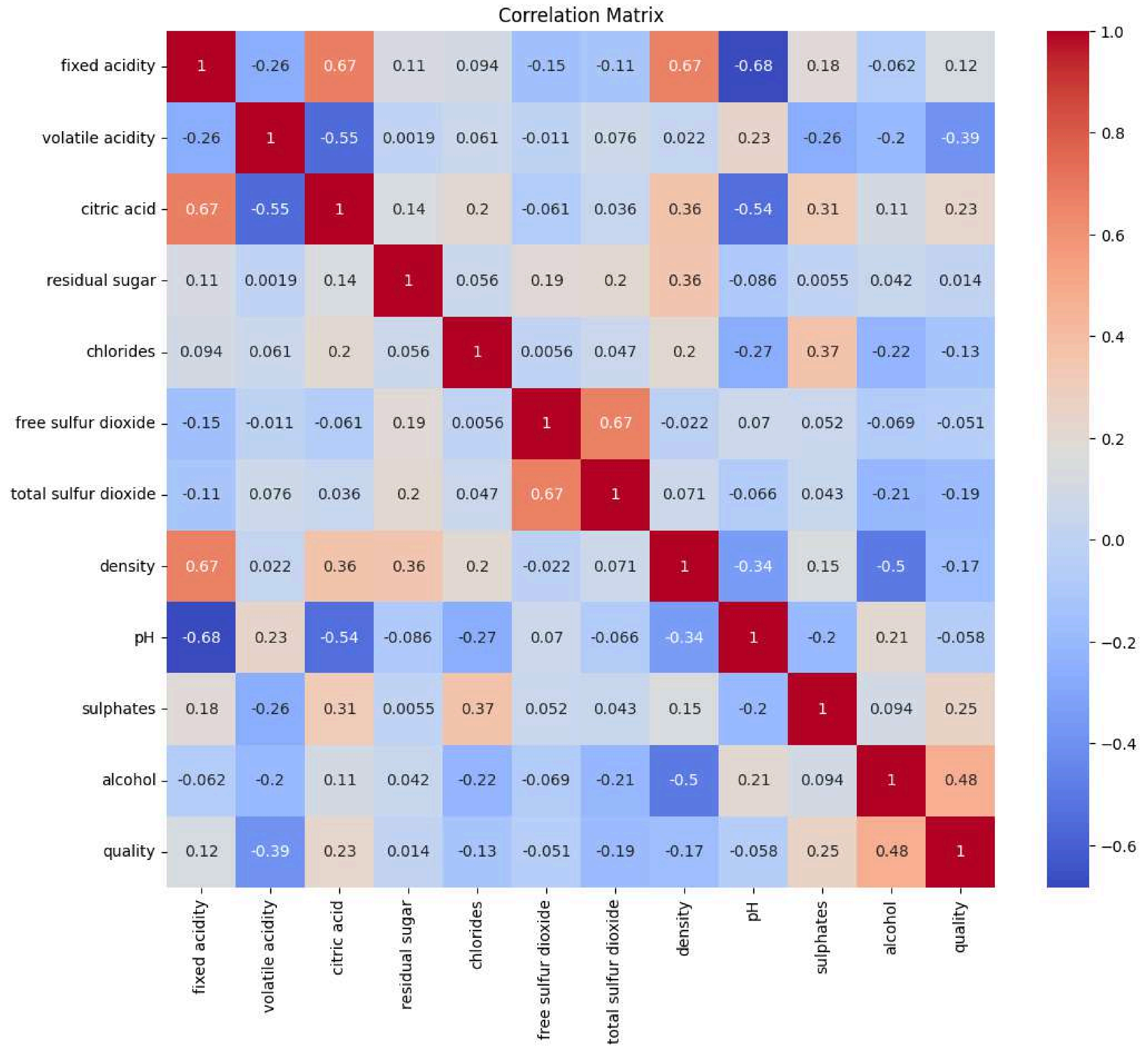
Key Features:

1. Captures Non-Linear Relationships: Handles complex interactions between features and the target variable.
2. Feature Subsampling: Reduces overfitting by training each tree on random subsets of features.
3. Robustness: Manages multicollinearity and missing data effectively.
4. Feature Importance: Identifies the most influential predictors.

Why Use Random Forest for This Dataset?

- Handles Complex Relationships: Models both linear (e.g., alcohol) and non-linear (e.g., volatile acidity) effects on wine quality.
- Manages Multicollinearity: Automatically handles correlated features like density and alcohol.
- Improves Prediction: Handles imbalanced datasets better than linear models like Ridge Regression.





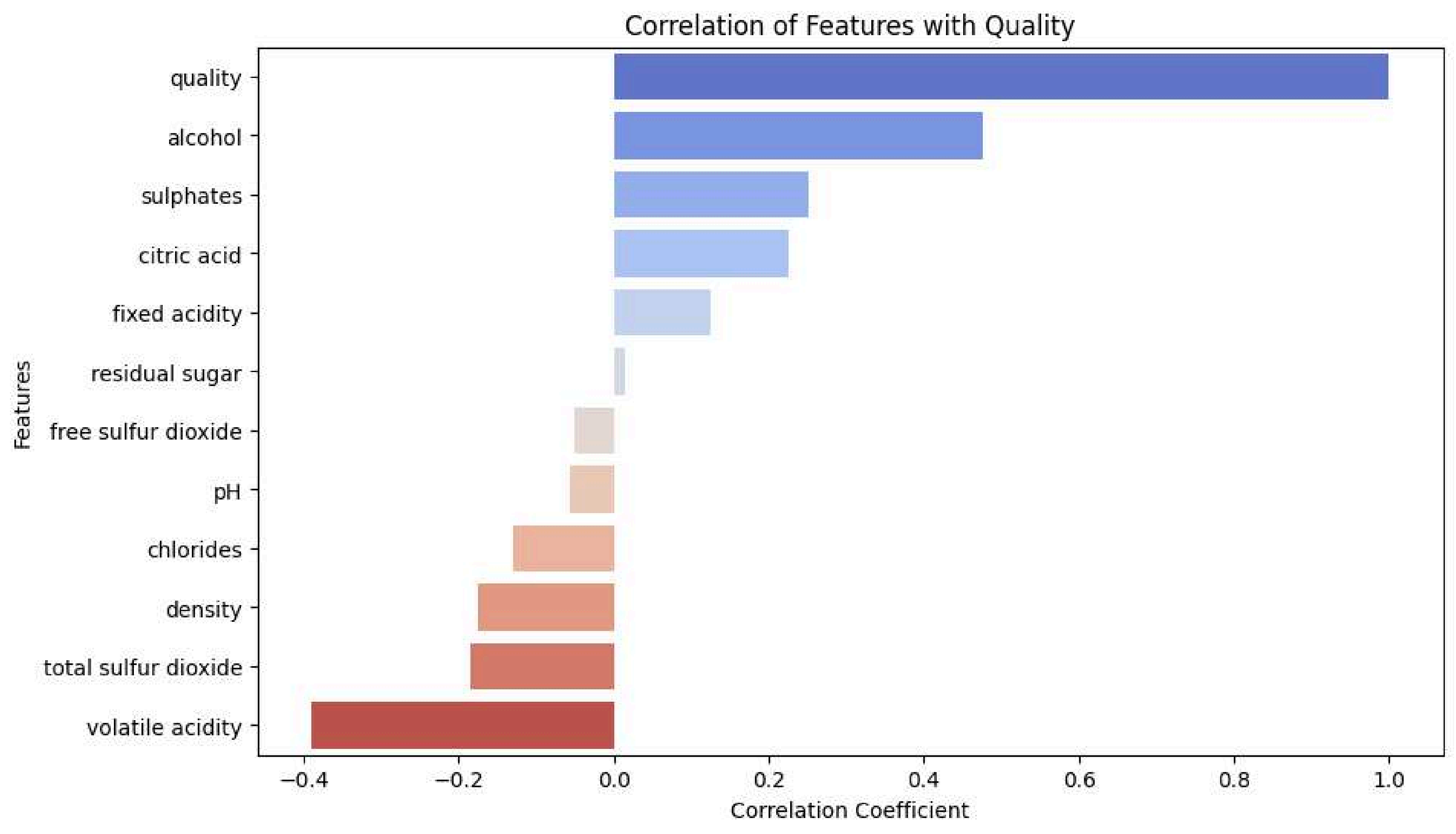
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

Correlation Matrix

- Strongest Correlations with Quality:
  - Alcohol: Highest positive correlation (r=0.48r = 0.48r=0.48).
  - Sulphates and citric acid: Moderate positive correlations.
  - Volatile acidity and density: Negative correlations.

Feature Importance

- Top Principal Components (PCs) contributing to the model:
  - PC2: Highest importance (0.23).
  - PC1 and PC3: Significant contributions (0.18 and 0.16).
  - Later components (e.g., PC6) had minimal impact, confirming the relevance of dimensionality reduction.

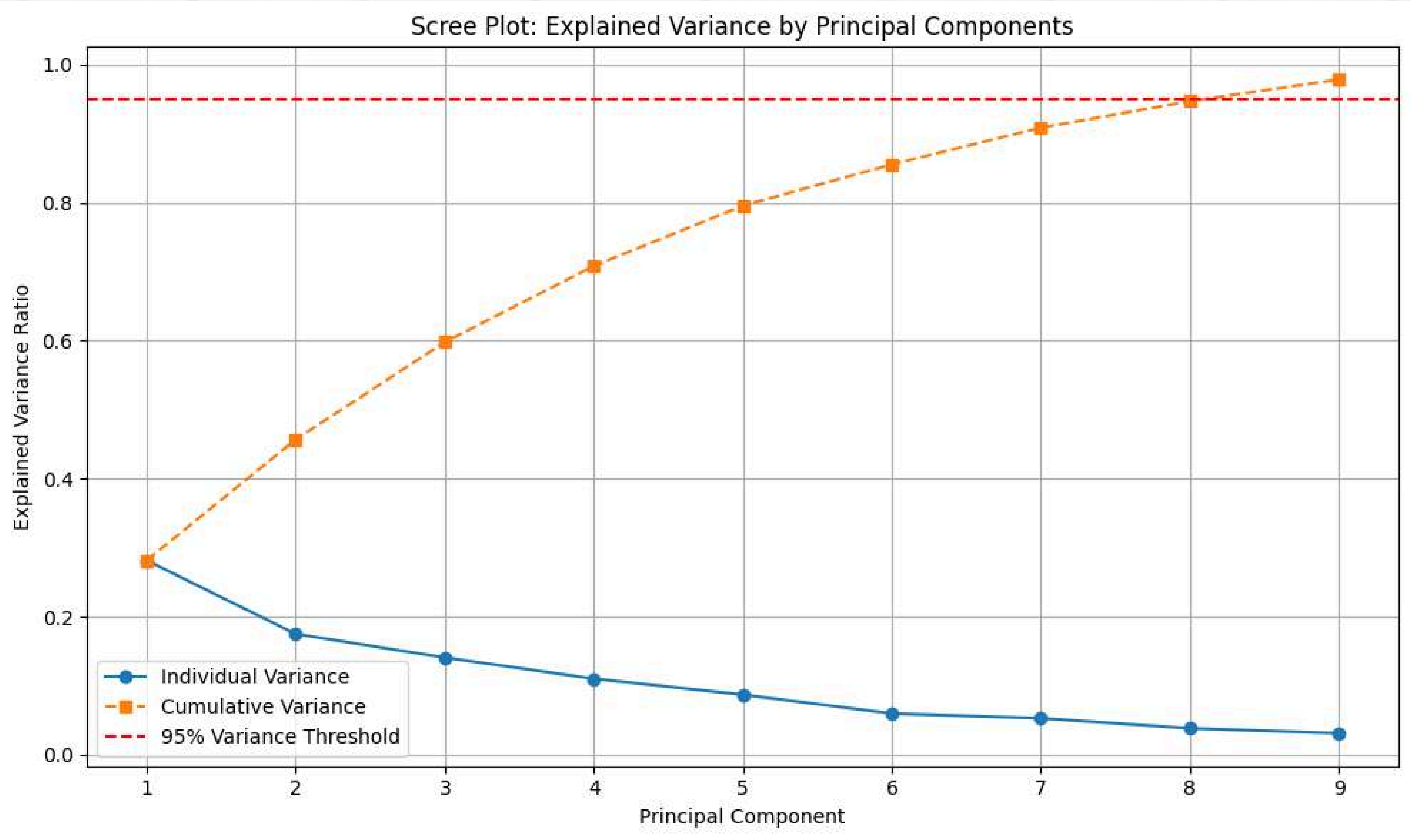


Feature Importance

Random Forest Feature Importance

The bar plot highlights:

- PC2 contributes the most to the model, with an importance score of 0.23.
- PC1 and PC3 are also significant, with scores of 0.18 and 0.16, respectively.
- Later components (e.g., PC6) have minimal impact, confirming the relevance of dimensionality reduction.



PCA’s Role

- PCA reduced the 11 features to 8 principal components, retaining 95% variance.
- Non-PCA models slightly outperformed PCA models due to potential information loss in feature-specific variance.

Hyperparameter Optimization

- Best Parameters (via GridSearchCV):
  - n_estimators: 200, max_depth: None, min_samples_split: 2, min_samples_leaf: 1, max_features: 'sqrt'.

Scenario	n estimators	max depth	min samples split	min samples leaf	max features
With PCA & GridSearch	200	None	2	1	sqrt
Without PCA & GridSearch	200	None	2	1	sqrt

What this table shows:

Both PCA and non-PCA models converged on the same optimal hyperparameters during GridSearchCV. This indicates that PCA does not influence hyperparameter optimization for Random Forest, likely because the dataset's dimensionality (11 features) is manageable.

Performance Metrics

- With PCA & GridSearch:
  - MAE = 0.427, MSE = 0.308, R² = 0.528
- Without PCA & GridSearch:
  - MAE = 0.415, MSE = 0.293, R² = 0.552
- Models without PCA consistently outperformed PCA models, highlighting that Random Forest effectively handles multicollinearity without dimensionality reduction.

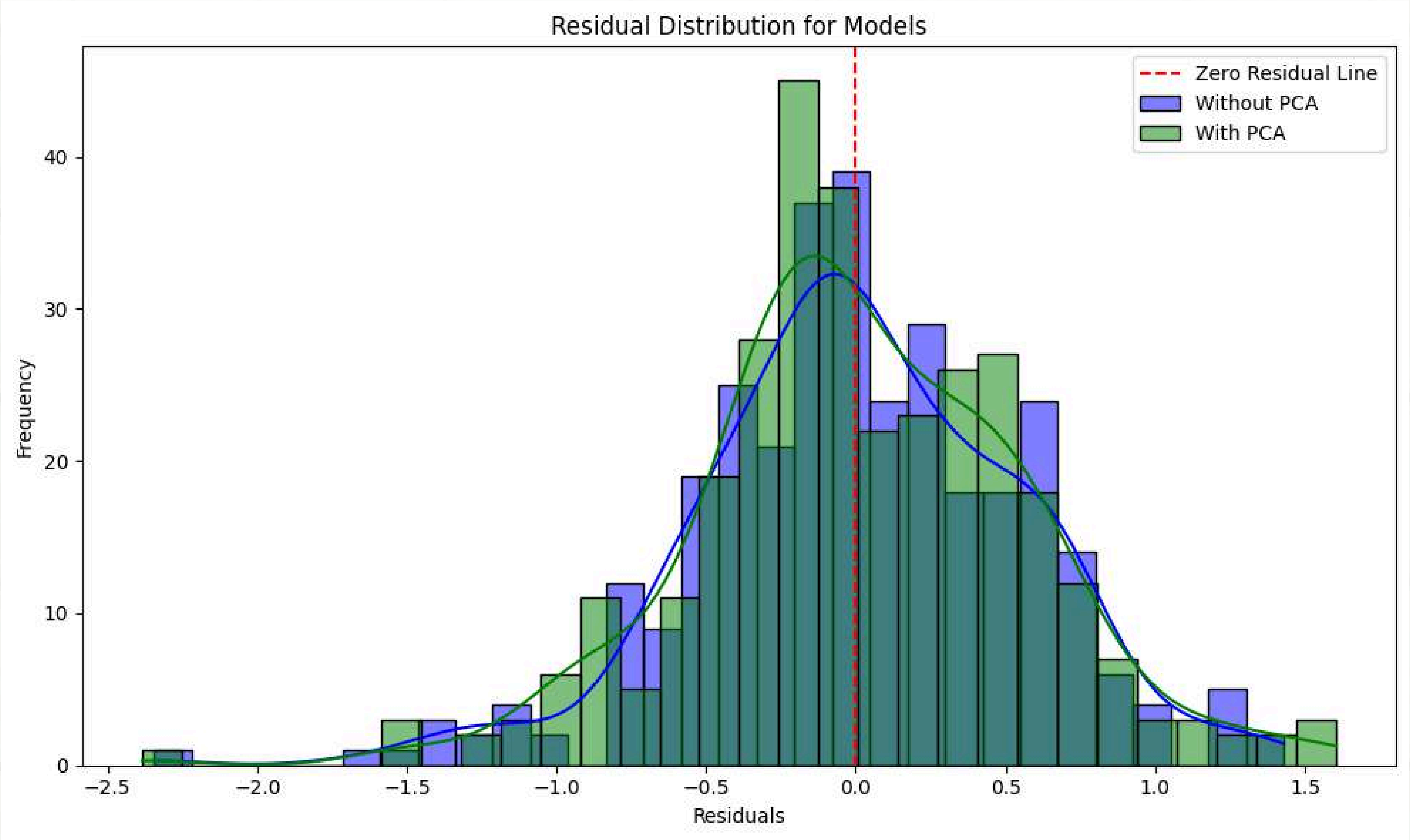
Residual Analysis

- Residuals are approximately normally distributed, centered at 0, indicating a well-fitted model.
- Slight skewness indicates underprediction for extreme quality scores (3 and 8), likely due to dataset imbalance.

Implications of Results

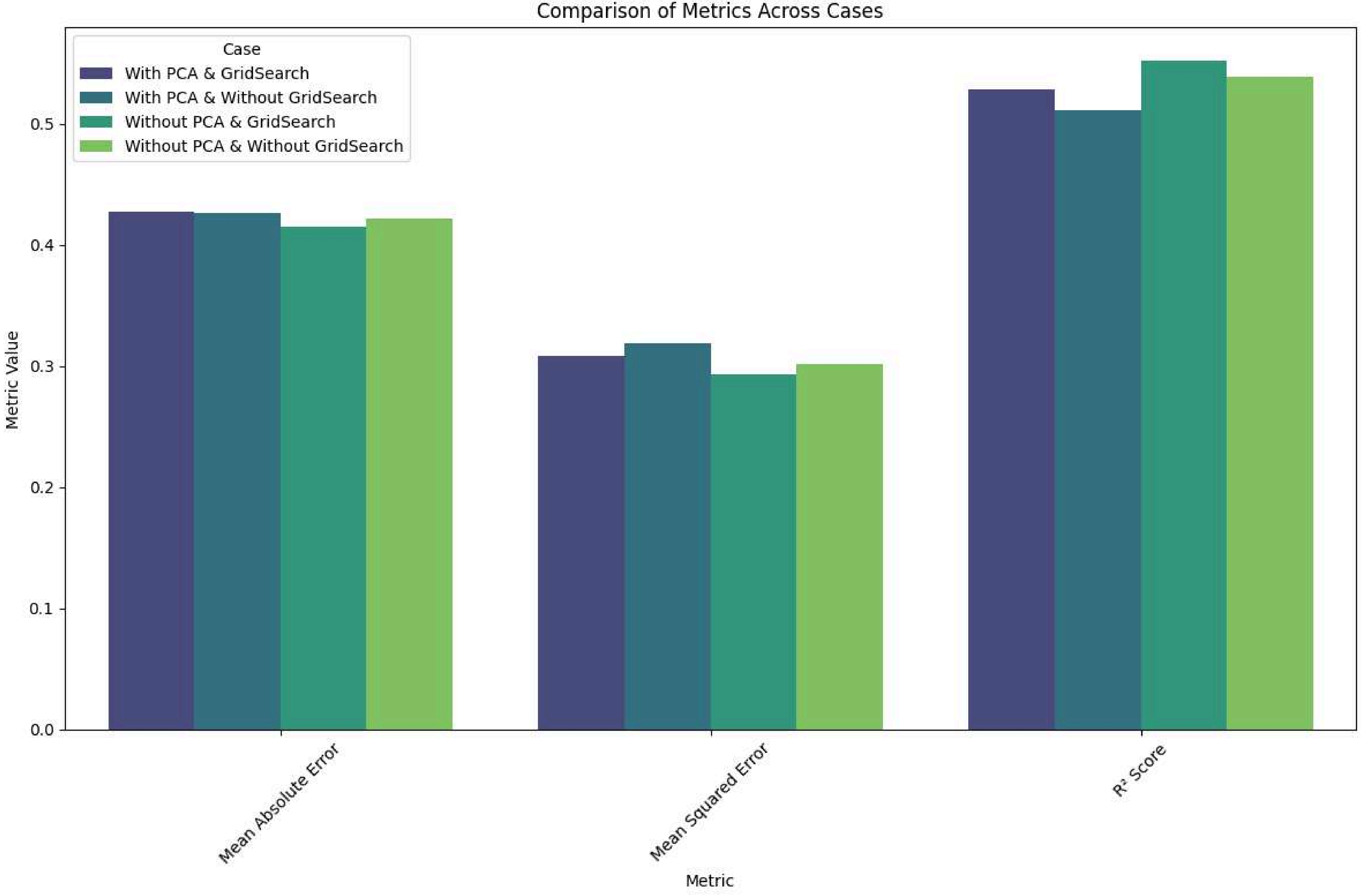
1. Practical Evaluation:
   - Random Forest can accurately predict mid-range wine qualities (scores 5–7), which dominate the dataset.
   - Edge cases (e.g., quality 3 and 8) require further data balancing or model refinement.

Metric	With PCA & GridSearch	With PCA & Without GridSearch	Without PCA & GridSearch	Without PCA & Without GridSearch
Mean Absolute Error	0.43	0.43	0.41	0.42
Mean Squared Error	0.31	0.32	0.30	0.30
R² Score	0.53	0.51	0.56	0.54



Key Insights and Implications

1. PCA's Role:
   - PCA reduced dimensionality but slightly degraded performance due to the loss of feature-specific variance.
   - Implication: PCA is unnecessary for Random Forest when handling low-dimensional datasets like this one (11 features).
2. Optimization's Role:
   - GridSearchCV fine-tuned the model, improving results slightly (e.g., R² = 0.552 without PCA).
   - Implication: Random Forest benefits from hyperparameter tuning for datasets with non-linear relationships.



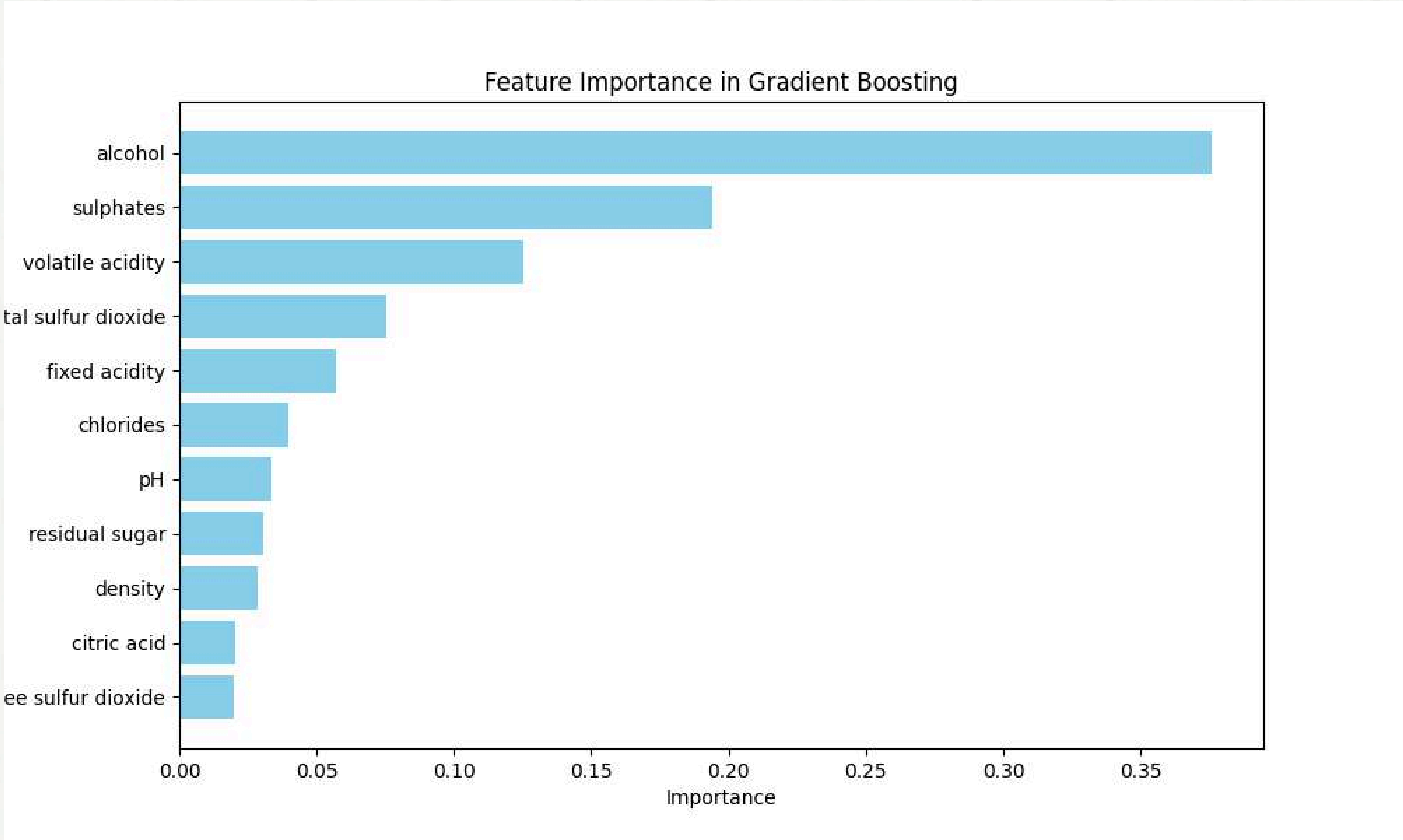
- 3.Random Forest Strengths:
  - Captured non-linear relationships effectively, outperforming Ridge Regression.
  - Implication: Random Forest is well-suited for datasets with complex relationships between features and target variables.
- Limitations:
  - Dataset imbalance affected predictions for extreme cases (quality = 3 and 8).
  - Implication: Future work should address imbalance or add features to better handle these edge cases.

Gradient Boosting Regression

Gradient Boosting Regression is a machine learning technique used to predict a **continuous target variable.** It is part of the **gradient boosting** family, which builds models by sequentially combining weak learners (typically decision trees) to create a stronger model.

Key Characteristics

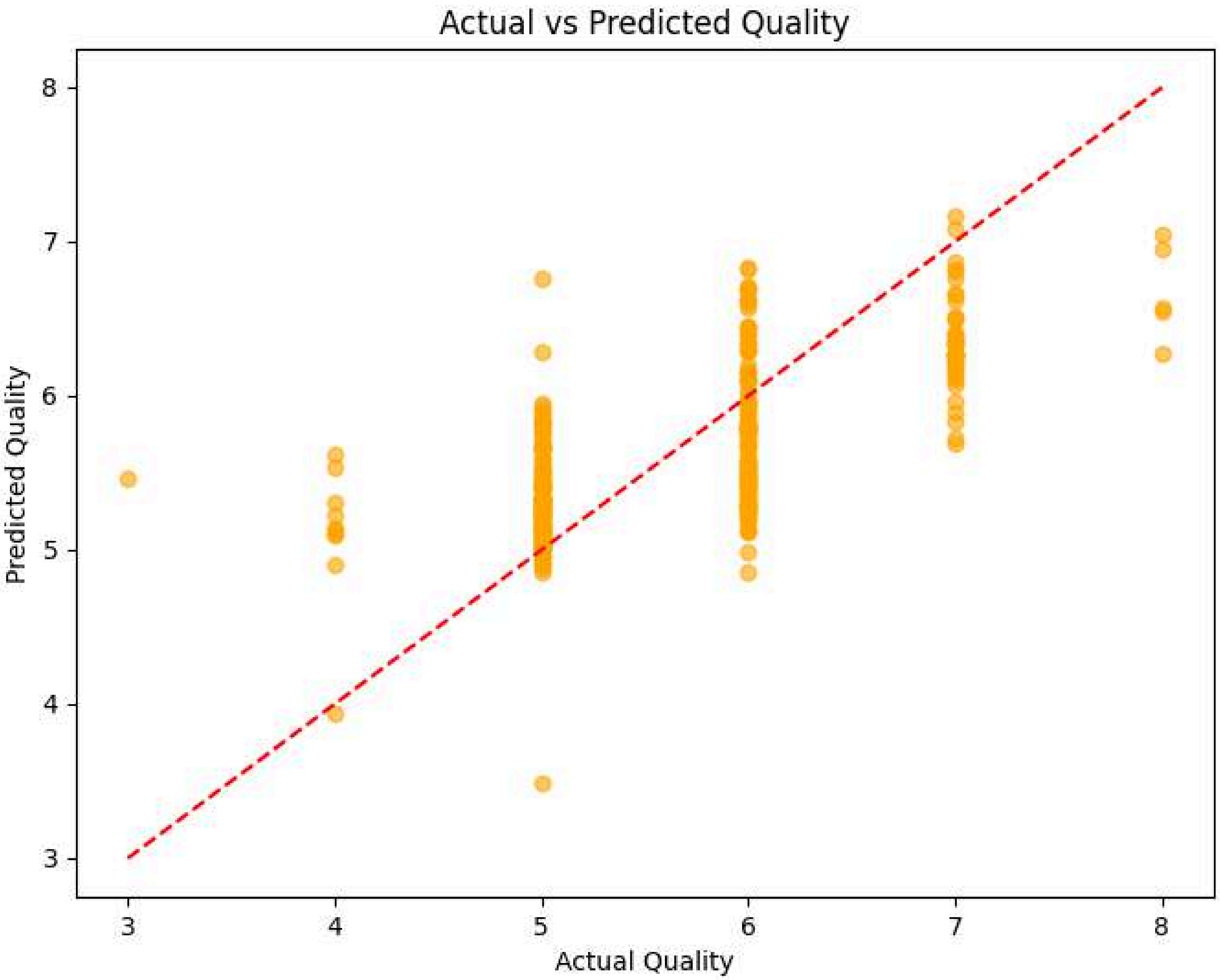
- **Weak Learners:** Typically uses shallow decision trees (small depth, like 3-5).
- **Loss Function:** Optimizes a loss function (e.g., mean squared error) to improve predictions iteratively.
- **Gradient Descent:** Uses gradient descent to minimize the loss function by fitting the new trees to the residuals (the gradient of the loss).



Feature Importance:

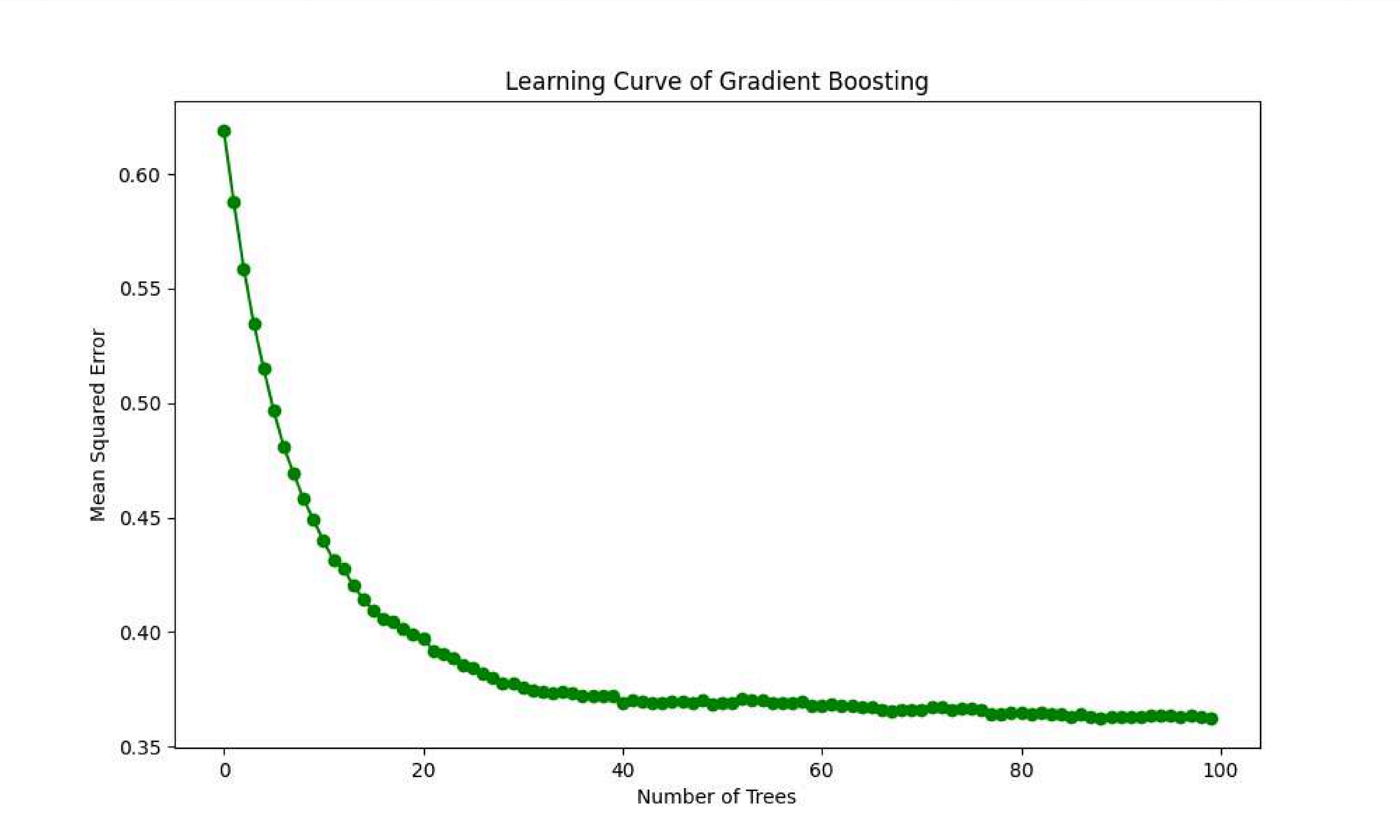
	Feature	**Importance**
10	alcohol	0.375965
9	sulphates	0.194192
1	volatile acidity	0.125359
6	total sulfur dioxide	0.075183
0	fixed acidity	0.056945
4	chlorides	0.039518
8	pH	0.033658
3	residual sugar	0.030401
7	density	0.028772
2	citric acid	0.020179
5	free sulfur dioxide	0.019828

- The alcohol content is by far the most influential factor in predicting wine quality, followed by sulphates and volatile acidity, which are key contributors.
- Citric acid and free sulfur dioxide have the least influence, suggesting they do not strongly determine wine quality in the context of this model.



Actual vs Predicted Plot

- The scatter plot shows a reasonable alignment between predicted and actual wine qualities, particularly for mid-range quality scores (5–7).
- Edge cases (e.g., quality levels 3 and 8) display slight underprediction, possibly due to the dataset's imbalance.
- The majority of predictions are clustered close to the ideal prediction line (y=x), showcasing good accuracy for most samples.



Gradient Boosting Regressor Parameters

The model was trained with the following configuration:

- **n_estimators:** 100 (Number of trees in the model)
- **learning_rate:** 0.1 (Controls the contribution of each tree)
- **max_depth:** 3 (Limits the depth of each decision tree to prevent overfitting)

Without PCA:	**With PCA:**
MAE: 0.48	MAE: 0.52
MSE: 0.36	MSE: 0.44
R^2: 0.45	R^2: 0.33

Ridge Regression

Ridge Regression is a linear regression technique that includes an L2 regularization term to penalize large coefficients and prevent overfitting. For this dataset:

- Problem: Some features are correlated (e.g., density and alcohol, or free sulfur dioxide and total sulfur dioxide).
- Solution: Ridge Regression handles multicollinearity by shrinking feature coefficients, ensuring the model remains stable.

Ridge Regression minimizes the following objective function:

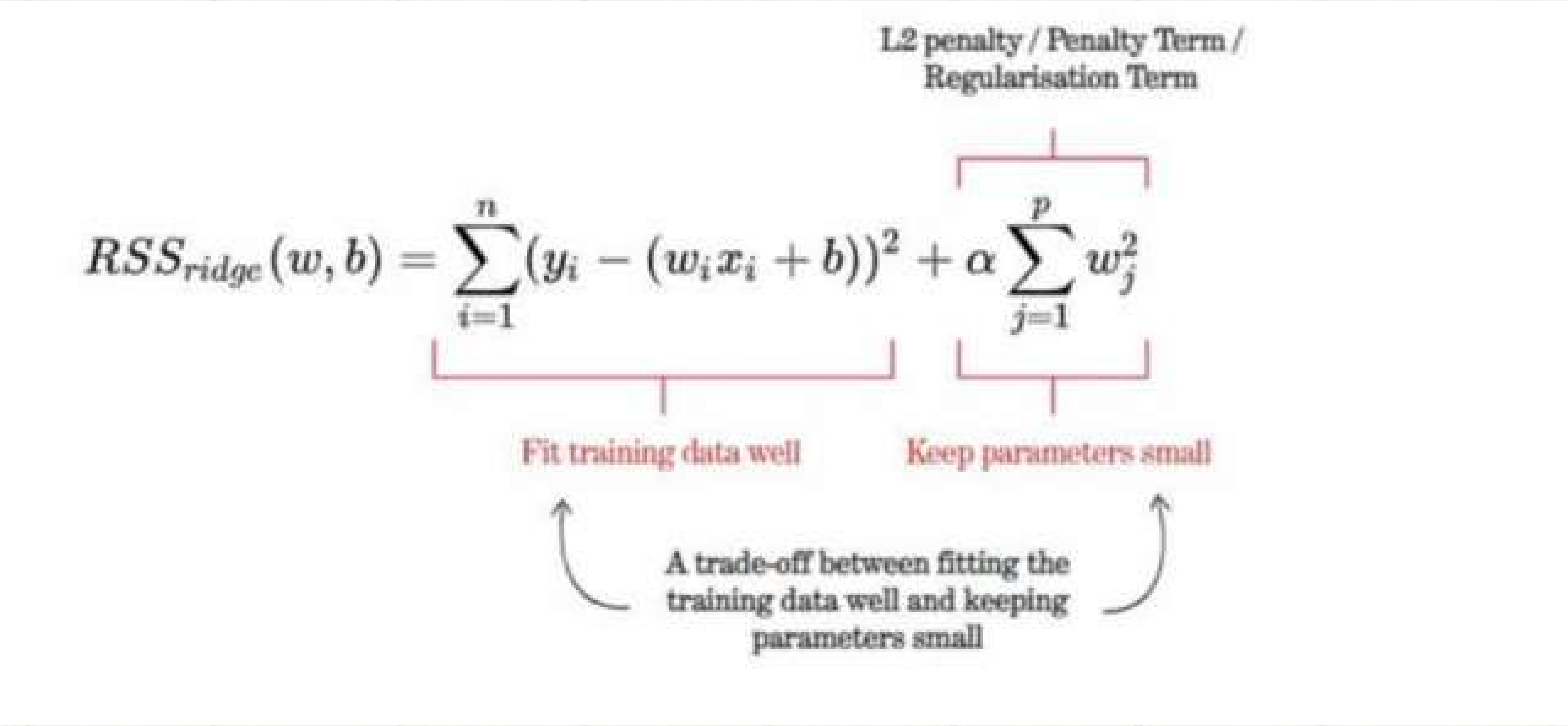
$L(\beta) = RSS + \lambda \Sigma (\beta j^2)$

- RSS: Residual Sum of Squares (the standard error in linear regression).
- λ (lambda): Regularization strength (a hyperparameter).
- βj^2: Squared coefficients of the features.

Purpose of Regularization (λ):

- Penalizes large coefficients, reducing their impact.
- Helps address multicollinearity by limiting the influence of highly correlated features.
- Ridge naturally balances the relationship between correlated features like alcohol and density

- Why Use It for This Dataset?
  - Some features are correlated (e.g., density and alcohol or total sulfur dioxide and free sulfur dioxide), which may destabilize ordinary linear regression.
  - Ridge Regression penalizes large coefficients, ensuring the model remains stable and generalizes well.



L2 penalty / Penalty Term / Regularisation Term

$$RSS_{ridge}(w,b) = \underbrace{\sum_{i=1}^{n}(y_i - (w_i x_i + b))^2}_{\text{Fit training data well}} + \alpha \underbrace{\sum_{j=1}^{p} w_j^2}_{\text{Keep parameters small}}$$

A trade-off between fitting the training data well and keeping parameters small

Why PCA Had No Impact?

- PCA reduced the 11 features to 8 components, retaining 95% variance. However, Ridge Regression results were unchanged with and without PCA.
- Reasons:
- Low Dimensionality:The dataset is low-dimensional (11 features) and manageable without dimensionality reduction.
- Ridge inherently handles multicollinearity, making PCA redundant.

Scenario	Mean Absolute Error	Mean Squared Error	R^2 Score
Without PCA & Before Optimization	**0.50**	**0.39**	**0.40**
With PCA & Before Optimization	**0.50**	**0.39**	**0.40**
Without PCA & After Optimization	**0.51**	**0.39**	**0.40**
With PCA & After Optimization	**0.51**	**0.39**	**0.40**

Why Did Optimization Have No Impact

Optimizing λ:

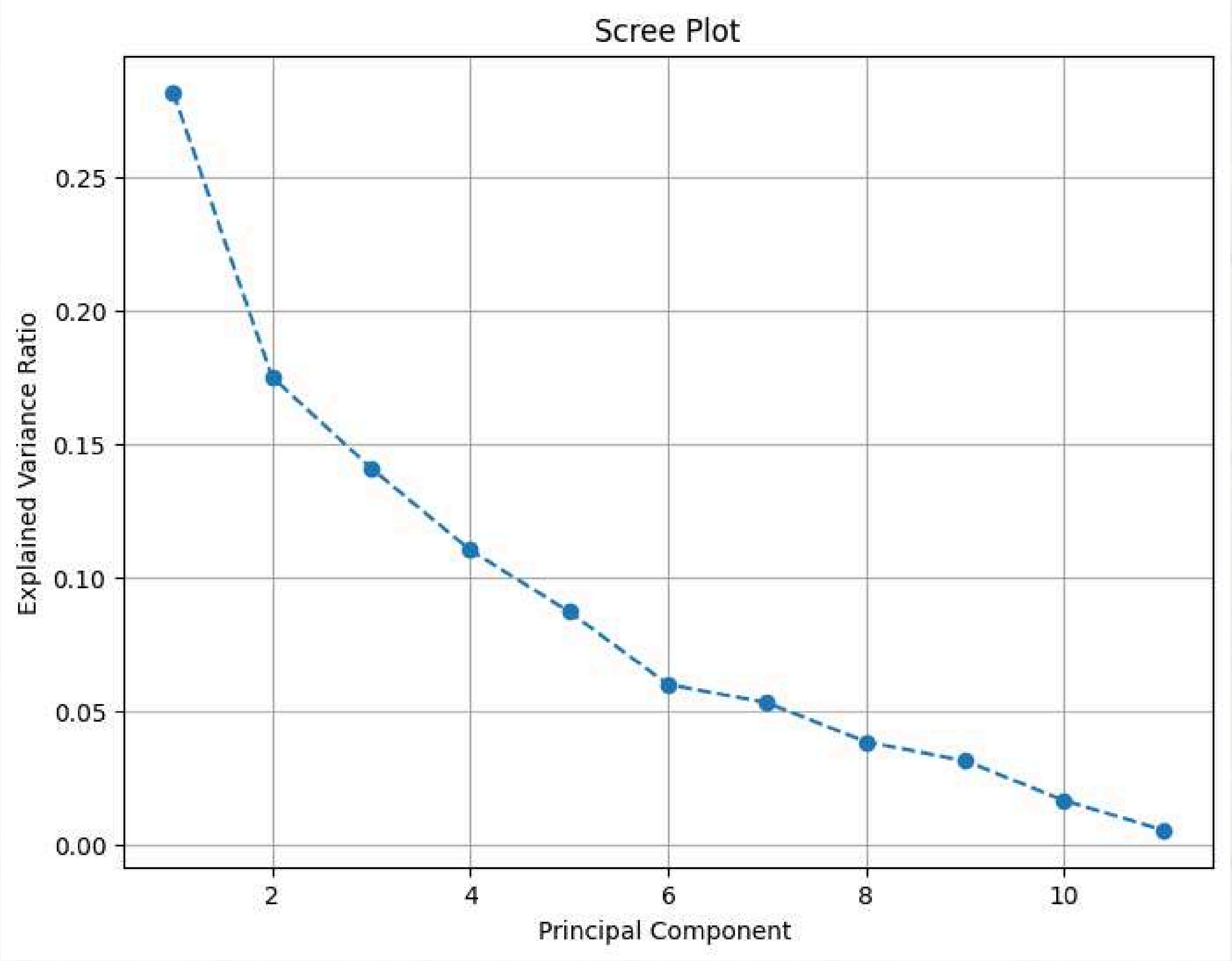
- I used GridSearchCV to find the optimal λ:
  - Without PCA: Best λ=79.06
  - With PCA: Best λ=59.63

However, results before and after optimization were nearly identical:

- Before Optimization: MAE = 0.50, MSE = 0.39, R^2 = 0.40
- After Optimization: MAE = 0.51, MSE = 0.39, R^2 = 0.40

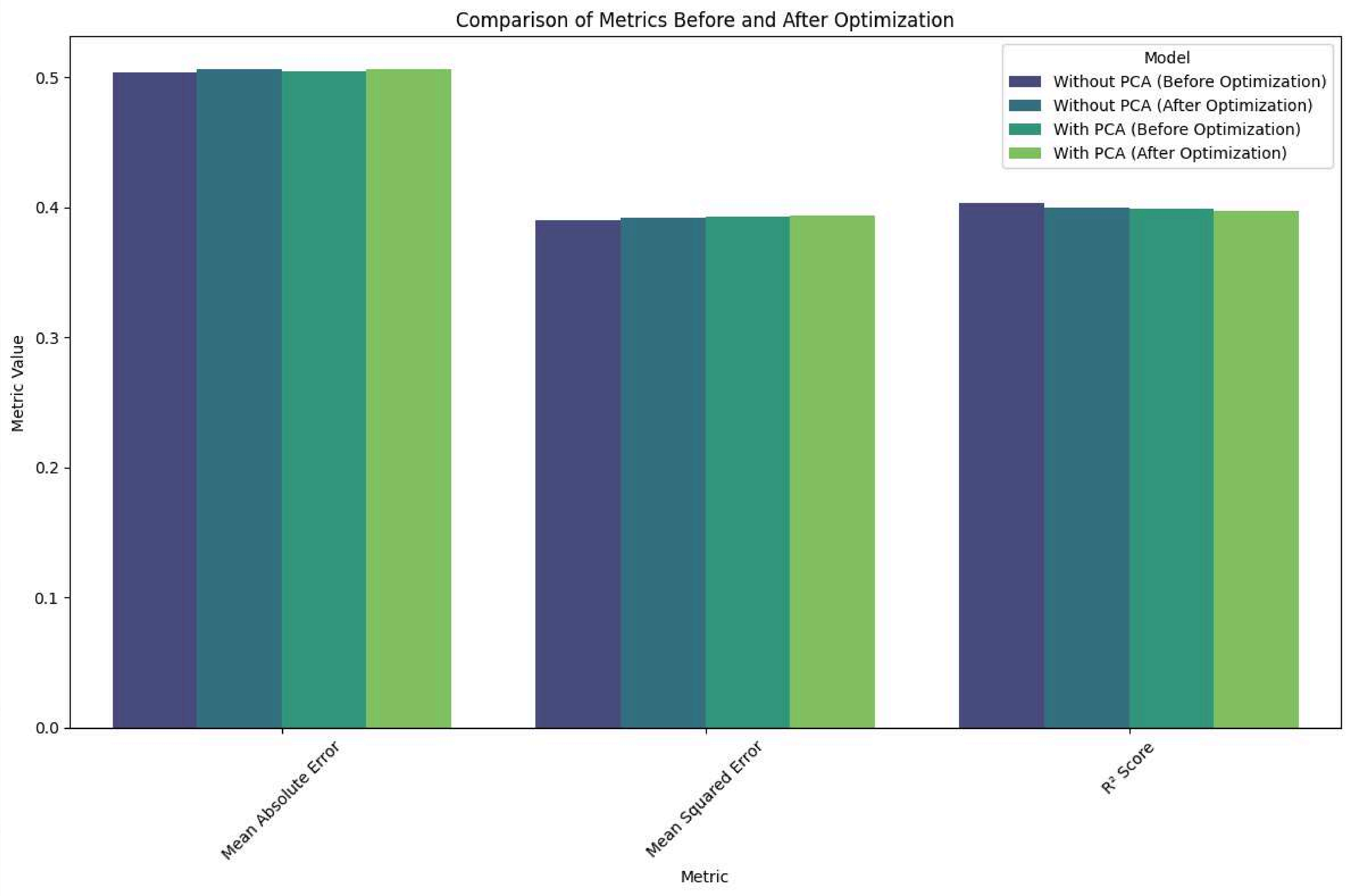
Why Results Didn’t Change:

- Despite tuning, results before and after optimization were nearly identical because the dataset relationships are straightforward and linear.
- Default Hyperparameters Were Already Effective: Ridge's default λ=1.0was already appropriate for my dataset, so tuning didn’t yield significant improvements.



What Did We Learn?

1. Model Stability: Ridge Regression consistently performed the same across all scenarios, highlighting its robustness.
2. PCA's Role:
   - PCA reduced dimensionality but did not improve results.
   - Ridge already handles multicollinearity through regularization, so PCA was unnecessary.
3. Optimization's Role:
   - Hyperparameter tuning provided no measurable improvement because the default λ=1.0 was already effective.
   - Results were capped at $R^2 = 0.40$ due to dataset simplicity.



5..Performance Limitation:

- R^2 = 0.40means only 40% of the variability in wine quality is explained. The remaining 60% is likely due to missing factors like grape type, production processes, or subjective grading criteria.

6. Cross-Validation Results:

- Mean R^2 = 0.29(with and without PCA) confirmed consistent generalization but showed the model struggles to explain all variability.

Ridge Regression Evaluation and Takeaways

Scenario	Mean Absolute Error	Mean Squared Error	R^2 Score
Without PCA & Before Optimization	0.50	0.39	0.40
With PCA & Before Optimization	0.50	0.39	0.40
Without PCA & After Optimization	0.51	0.39	0.40
With PCA & After Optimization	0.51	0.39	0.40

1. Dataset and Model Strengths:
   - The dataset is well-structured and low-dimensional (11 features), with no significant noise or redundancy.
   - Ridge Regression inherently handles multicollinearity effectively, rendering PCA unnecessary.
   - The consistency across scenarios highlights Ridge Regression’s stability and reliability.
2. Why Optimization Didn’t Help:
   - The relationships in the dataset are simple and linear, leaving little room for hyperparameter tuning (λ\lambdaλ) to improve results.
   - The dataset's explanatory power is limited by the absence of critical external variables.
3. Performance Limitation:
   - The R2=0.40R^2 = 0.40R2=0.40 ceiling reflects the dataset’s inability to fully explain wine quality.
   - Ridge Regression’s linear nature restricts its ability to model non-linear relationships, which may exist between features like alcohol, density, and volatile acidity.

CatBoost Regression

CatBoost is a powerful gradient boosting library designed for speed, accuracy, and ease of use. It handles numerical and categorical data natively and is known for preventing overfitting using regularization techniques.

Key Advantages:

- Handles missing data automatically.
- Requires minimal preprocessing (e.g., no need to scale features).
- Highly efficient for small to medium-sized datasets like ours.

Why Use CatBoost?

- The wine quality dataset contains non-linear relationships and interactions among features, which gradient boosting methods excel at capturing.
- CatBoost’s feature importance analysis provides insights into which factors most influence wine quality.

Model Setup and Tuning

Training Process:

- **Dataset split:** 80% training, 20% testing.
- **Hyperparameter Tuning:** Used **Grid Search** for efficient parameter optimization. Explored combinations of:
    - **depth:** Controls tree complexity.
    - **iterations:** Number of boosting rounds.
    - **learning_rate:** Smaller values ensure gradual learning.
    - **l2_leaf_reg:** Regularization parameter.
    - **bagging_temperature:** Adds randomness to reduce overfitting.

- **Best Parameters Achieved:**
    - Learning Rate: **0.05**
    - Depth: **10**
    - Iterations: **375**
    - L2 Regularization: **1.0**
    - Bagging Temperature: **0.5**

Performance Metrics

- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.
- **Mean Absolute Error (MAE):** Reflects the average magnitude of errors.
- **R^2 Score:** Indicates the proportion of variance explained by the model.

Results Summary:

No PCA:

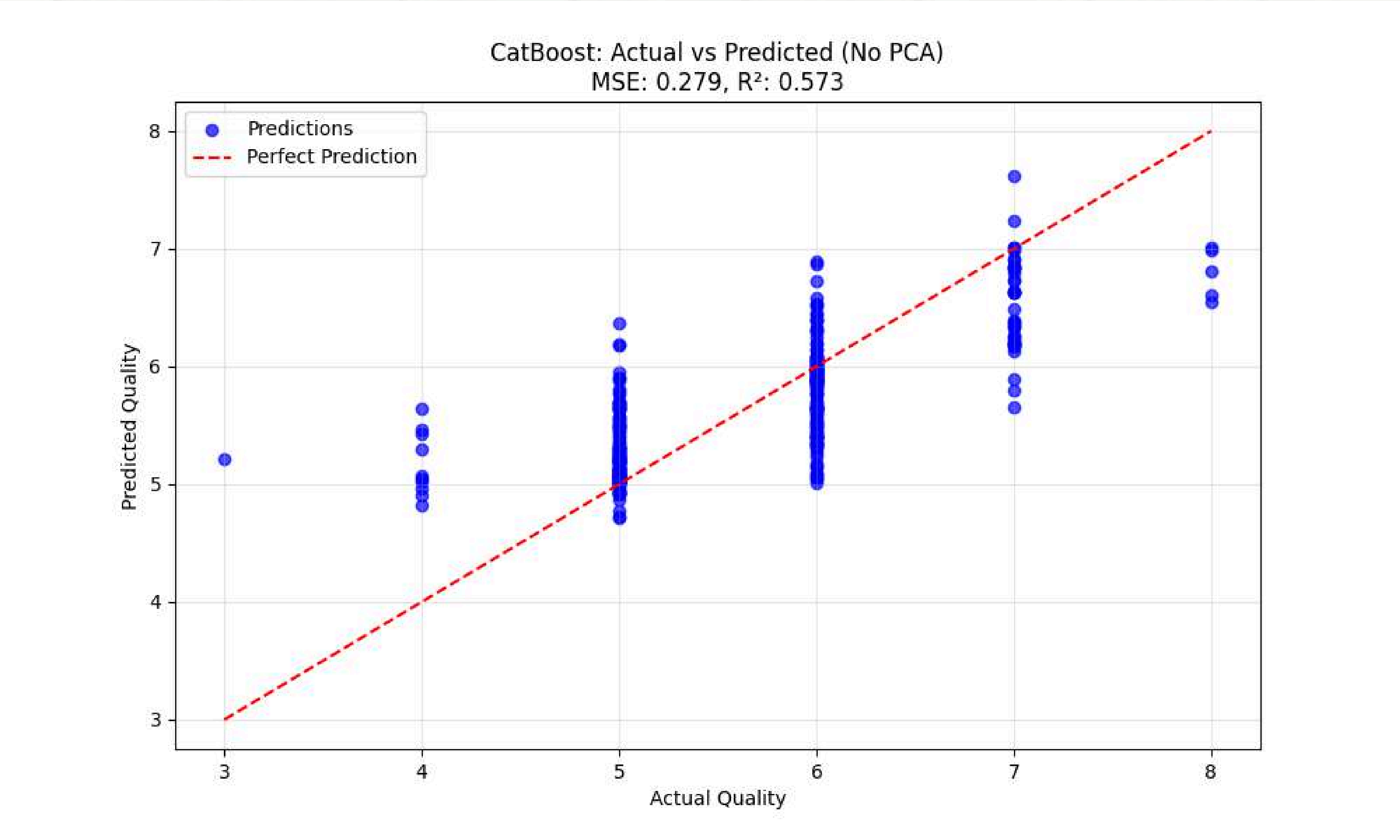
- **MSE: 0.2791**
- **MAE: 0.3889**
- **R^2: 0.5730**

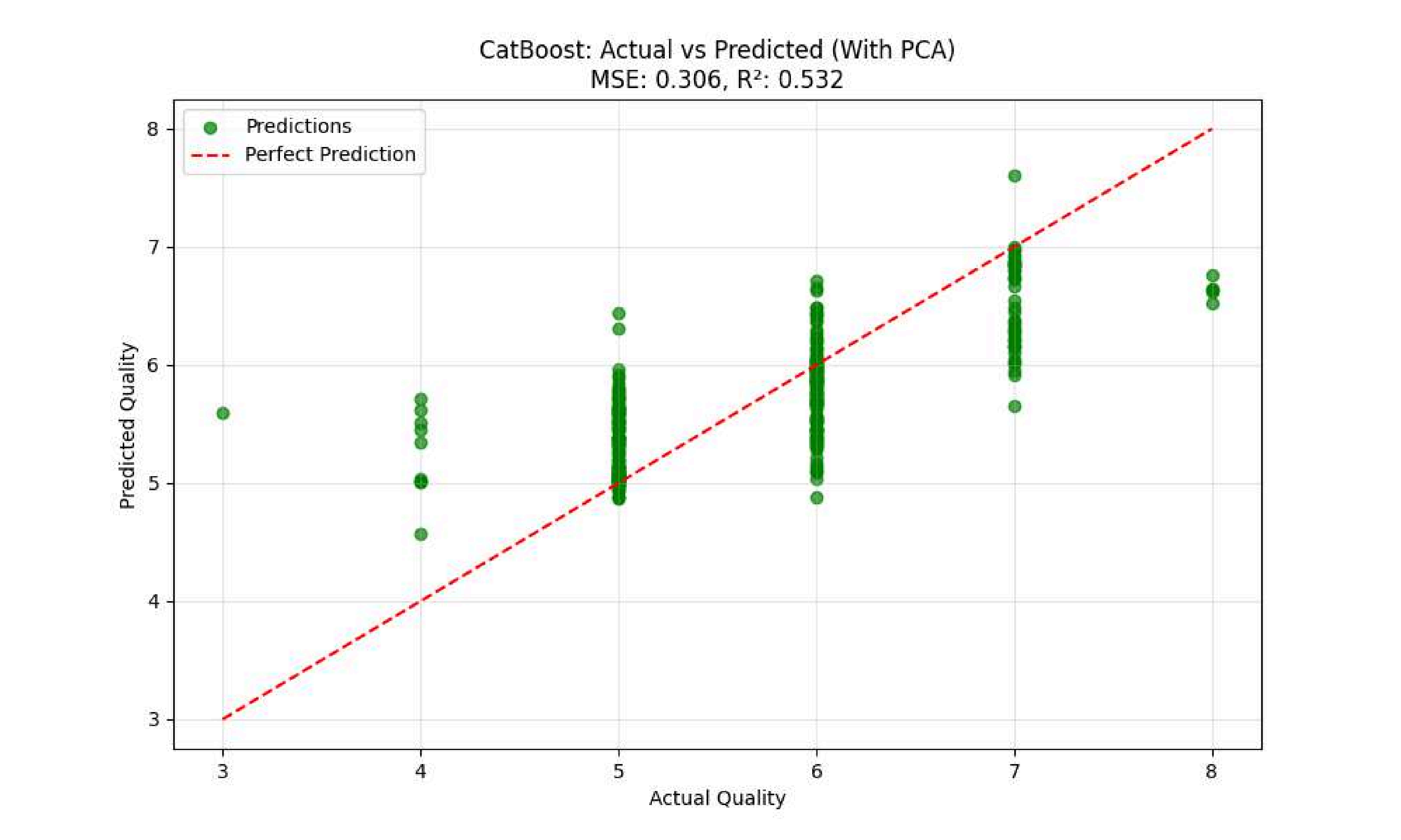
With PCA:

- **MSE: 0.3061**
- **MAE: 0.4074**
- **R^2: 0.5317**

No PCA consistently outperformed PCA across all metrics. MSE and MAE are low, indicating small prediction errors, but R^2 suggests moderate model performance overall.

Actual vs Predicted



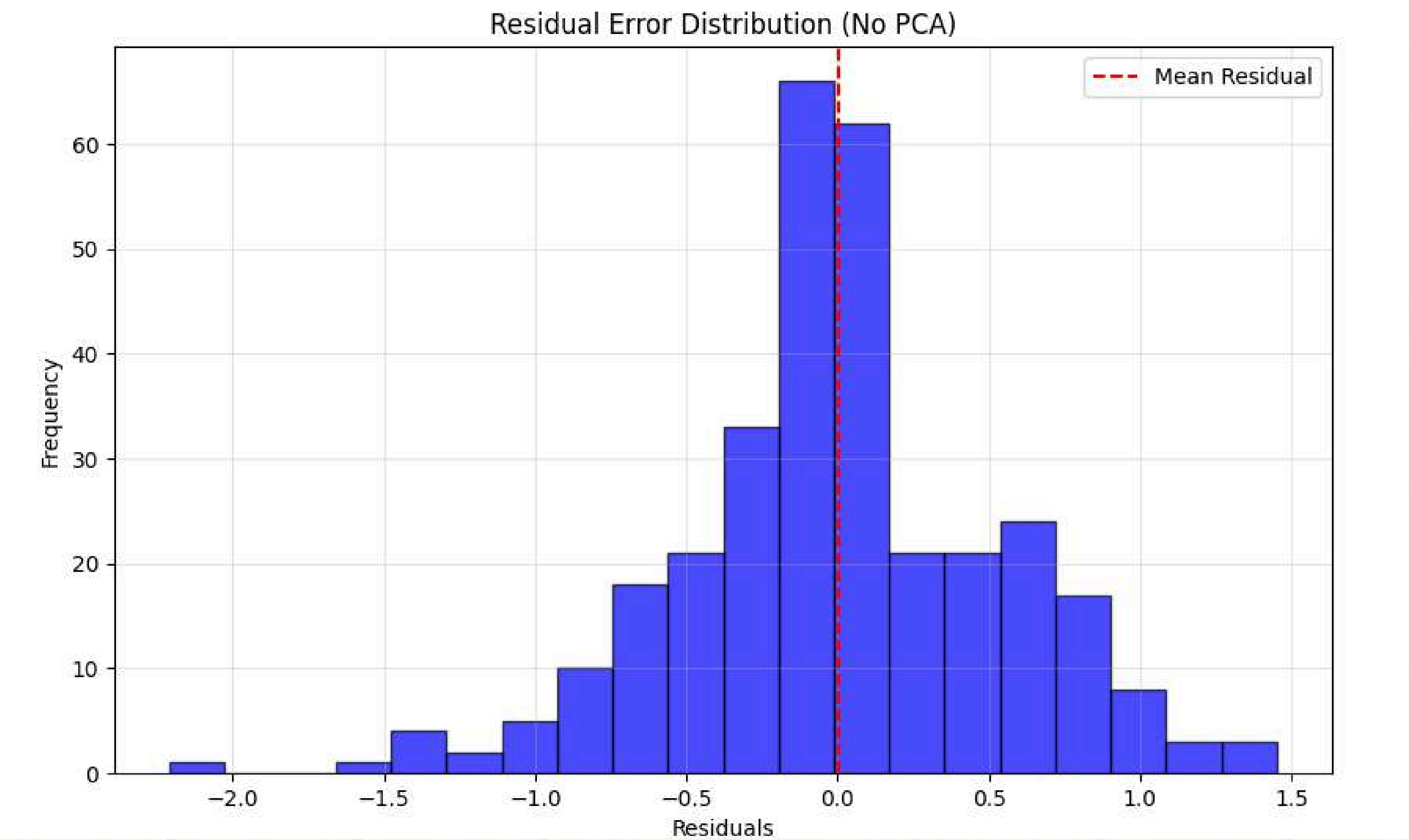


- The predictions align more closely with the actual values compared to PCA.
- This is reflected in the improved R² (0.573) and lower MSE (0.279).
- Highlights CatBoost’s ability to utilize raw feature interactions effectively without dimensionality reduction.

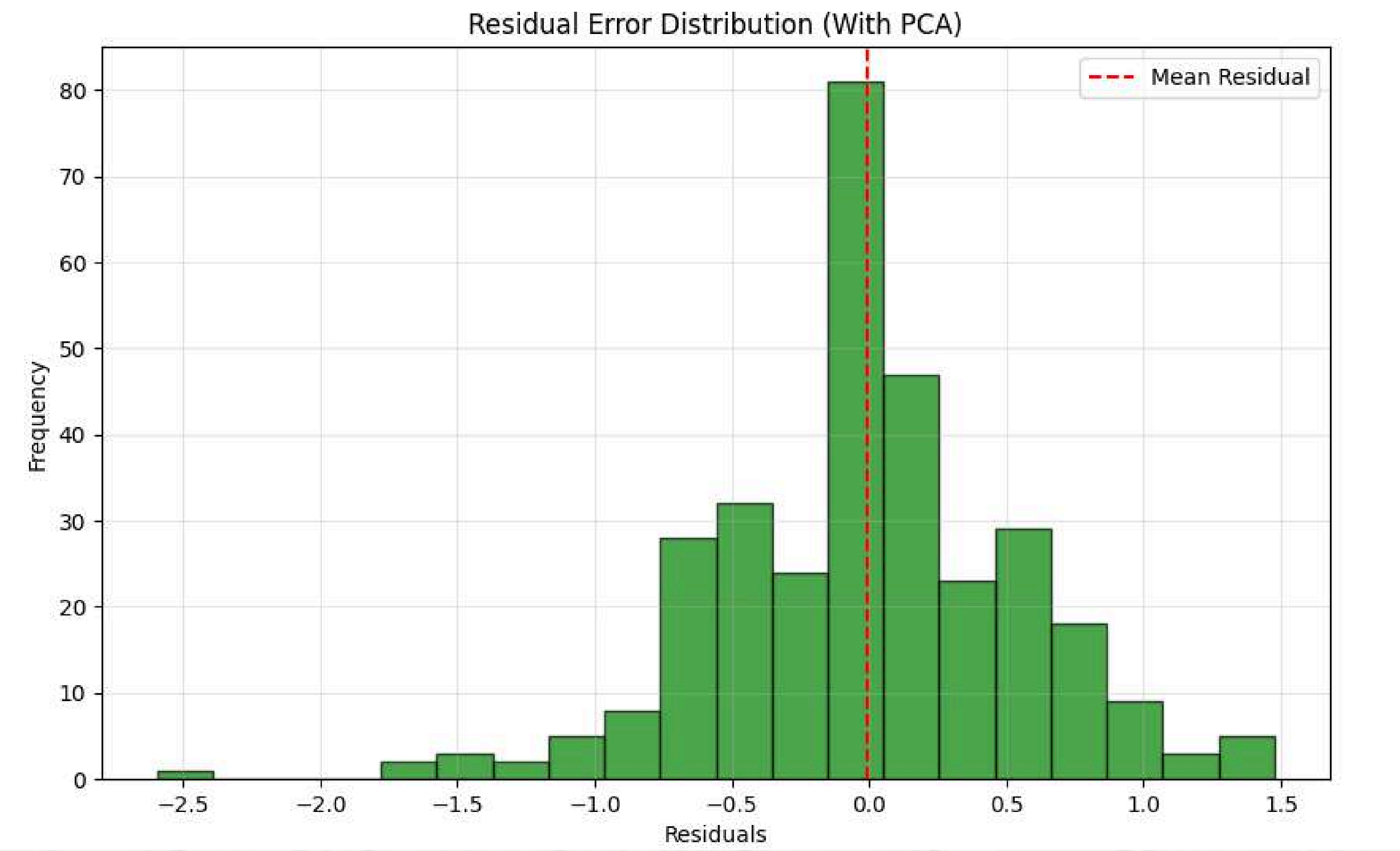
- The scatter plot indicates a noticeable alignment with the perfect prediction line (red dashed line).
- PCA slightly reduces the variance in predictions, as evident in the clustering around the mean values.
- MSE = 0.306, R² = 0.532: Performance slightly lags compared to No PCA due to the loss of feature interaction details.

Residual Distribution Analysis

What Are Residuals?: Residuals represent the difference between actual and predicted values, smaller residuals indicate better model accuracy.

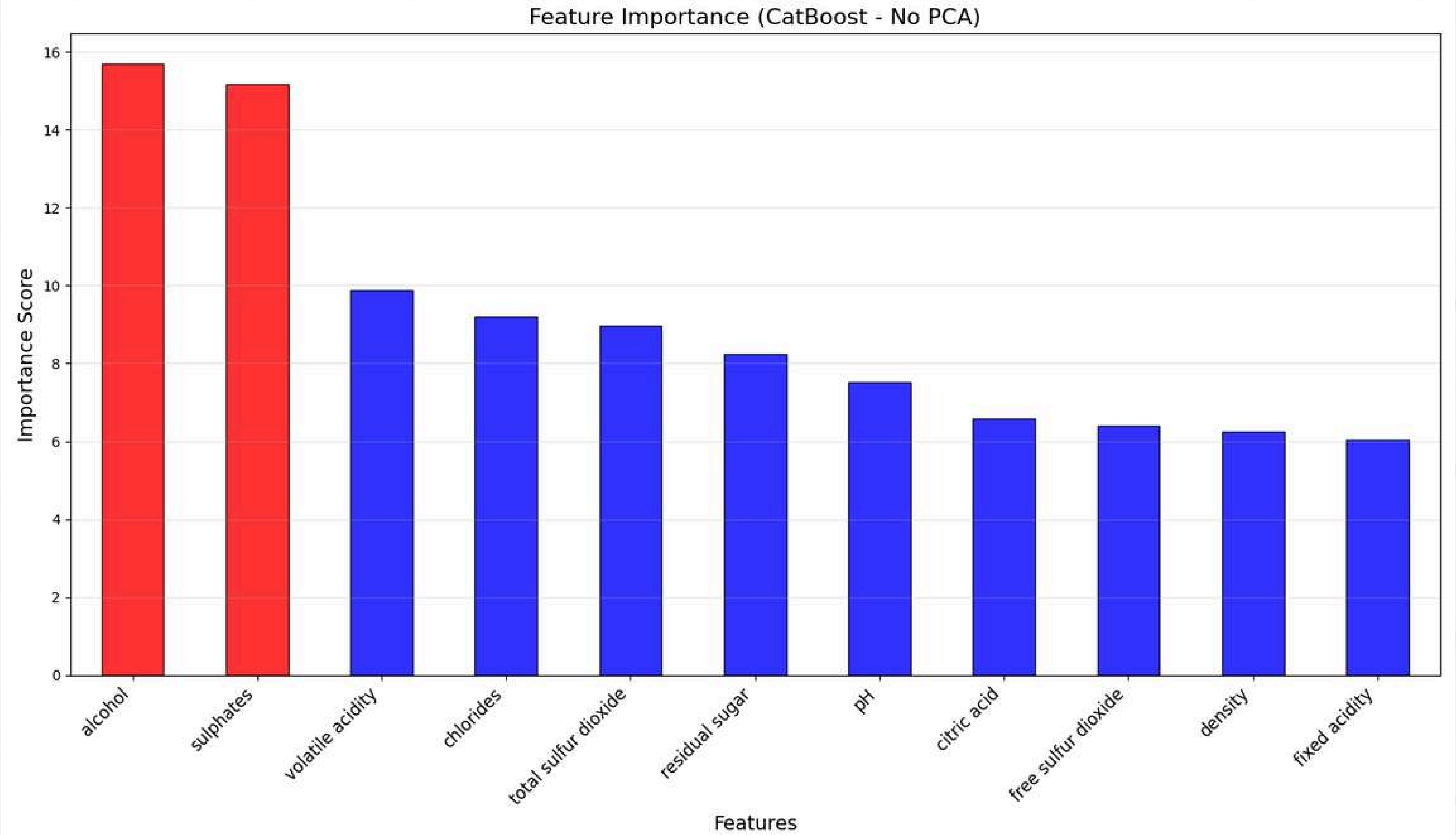


- A tighter and more symmetric residual distribution around 0.
- Indicates better model accuracy and lower prediction error variability.
- Residual range is narrower, highlighting the model's ability to minimize error effectively without PCA.



- The residual histogram shows a near-symmetric distribution around 0, suggesting unbiased predictions.
- Slightly wider distribution compared to No PCA, indicating higher variability in prediction errors.
- Residuals are spread within [-2.5, 1.5], with most values concentrated near 0.

Feature Importance Analysis



- The feature importance highlights the importance of chemical properties in determining wine quality.

- The rankings reaffirm the robustness of CatBoost in capturing complex feature interactions, making it an effective model for this dataset.

Key Observations:

- **Alcohol** and **sulphates** are the top contributors, showcasing their strong influence on wine quality. These features are directly linked to sensory perceptions like flavor and texture.

- Volatile acidity, chlorides, and total sulfur dioxide also rank high, reflecting their impact on taste and preservation quality.

- Features like fixed acidity, density, and pH have lower importance, suggesting lesser direct influence on wine quality in this dataset.

Insights:

- The dominance of alcohol aligns with its well-established role in enhancing wine aroma and body.

- Sulphates are critical for preserving wine freshness and enhancing flavor, justifying their high importance.

- The mid-tier importance of volatile acidity indicates that while it impacts quality, its influence is moderated by other factors like sulphates.

Understanding Model Performance for CatBoost Regresion

1. Model Performance Analysis and Evaluation
2. Why PCA Reduced Performance?
3. How We Could Improve Performance?

Model Performance Analysis and Evaluation

Dataset Challenges Affecting Model Performance:

- **Small Dataset Size:** With only 1,599 samples, the model struggles to generalize predictions. Larger datasets typically allow tree-based models like CatBoost to capture more complex patterns. The limited sample size increases the risk of overfitting or underfitting.
- **Class Imbalance:** Most wines are rated 5 or 6, while very few fall into high-quality (7–8) or low-quality (3–4) categories. The model becomes biased toward predicting the majority class, reducing accuracy for rare outcomes.
- **Missing Contextual Features:** Factors such as grape type, vineyard conditions, and fermentation processes are critical for predicting wine quality but are not included in the dataset. The absence of these variables creates unexplained variance, limiting the model’s ability to fully predict quality.
- **Noise in Output Variable:** Sensory quality (target variable) is inherently subjective, varying due to tasting conditions, human bias, and other unrecorded factors. Noise in the data reduces the precision of the model's predictions.

Dataset Suitability for CatBoost:

- CatBoost is well-suited to handle datasets with non-linear interactions, which aligns with the dataset's physicochemical properties. However, the missing contextual features limit the model’s ability to achieve high R^2 scores.

Performance Metrics:

- **Mean Squared Error (MSE):** Relatively low, showing the model’s ability to minimize large prediction errors. For the best CatBoost model (No PCA), MSE = 0.2791.
- **R^2 Score:** Moderate (~57%), meaning the model explains 57% of the variance in wine quality scores. This leaves 43% of the variance unexplained.
- **Mean Absolute Error (MAE):** Averages at 0.3889, indicating the model's predictions are off by less than 0.4 points on the 10-point scale.

Reasons for Moderate R^2 Score:

- **Complex Feature Interactions:** Features like alcohol, sulphates, and acidity interact in non-linear ways that the model captures, but these relationships are not enough to explain all variance.
- **Feature Limitations:** While physicochemical properties explain part of wine quality, they are not the sole determinants. The lack of contextual variables reduces predictive power.
- **Subjectivity in Quality Ratings:** Human evaluations add variability, introducing noise that even advanced models like CatBoost cannot fully account for.

Overall Assessment:

- **Strengths of the Model:** CatBoost minimizes large errors (low MSE) and provides consistent predictions (low MAE). The model effectively captures non-linear interactions among features.
- **Limitations:** The R^2 score indicates that other unmeasured factors significantly affect wine quality. The dataset’s limited size, imbalance, and missing contextual data restrict the model’s potential.

Why Did PCA Reduce Performance?

Key Characteristics of PCA:

PCA reduces dimensionality by retaining variance and discarding less important information. However, it assumes linear relationships, which can oversimplify the non-linear interactions inherent in many datasets.

Impact on CatBoost Performance:

- **CatBoost's Strengths:** CatBoost naturally captures non-linear relationships and feature interactions without requiring dimensionality reduction. It thrives on high-dimensional data, leveraging its gradient-boosting algorithm to explore complex dependencies between features.
- **Why PCA Falls Short:** PCA transforms features into orthogonal principal components, removing crucial interactions (e.g., between alcohol and sulphates). Tree-based methods like CatBoost rely heavily on feature interactions for accurate predictions, making PCA more of a hindrance than a benefit.

Observation and Conclusion:

PCA is effective for reducing overfitting in linear models, but it offers minimal benefit—and even performance degradation for tree-based models like CatBoost. The dataset's non-linear relationships and feature dependencies (e.g., how acidity and alcohol combine to influence quality) are critical for the model's success. For CatBoost, which excels in capturing these interactions, PCA is unnecessary and counterproductive.

How We Could Improve Performance:

Collecting More Data:

- Increase the sample size, especially for rare quality scores, to improve generalization.

Incorporating Missing Features:

- Add variables like grape variety, fermentation time, and region, which are known to influence wine quality.
- Exploring Classification Tasks:
- Reframe the problem as a binary classification task (e.g., good vs. not good wines). This might yield higher performance, as decision boundaries are easier to learn for simpler classes.

Feature Engineering:

- Derive new features (e.g., acidity-to-sugar ratio, alcohol density ratio) that capture meaningful interactions.

CatBoost Regression Conclusion

Strengths of CatBoost:

- Achieved low MSE and consistent predictions despite dataset limitations.
- Effectively modeled non-linear relationships and feature importance.

Dataset Challenges:

- Small sample size and imbalanced target variable limited generalization.
- Missing contextual features contributed to unexplained variance.

Overall Assessment:

- CatBoost demonstrated strong predictive power for the dataset, but the moderate R^2 score indicates room for improvement.

Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm used for classification and regression tasks. SVM finds a hyperplane that best separates data points into different classes.

Kernel Trick:

SVM uses a kernel function (e.g., Radial Basis Function - RBF) to map data into higher-dimensional space, making it easier to find a hyperplane that can separate complex data patterns.

Model Setup and Tuning

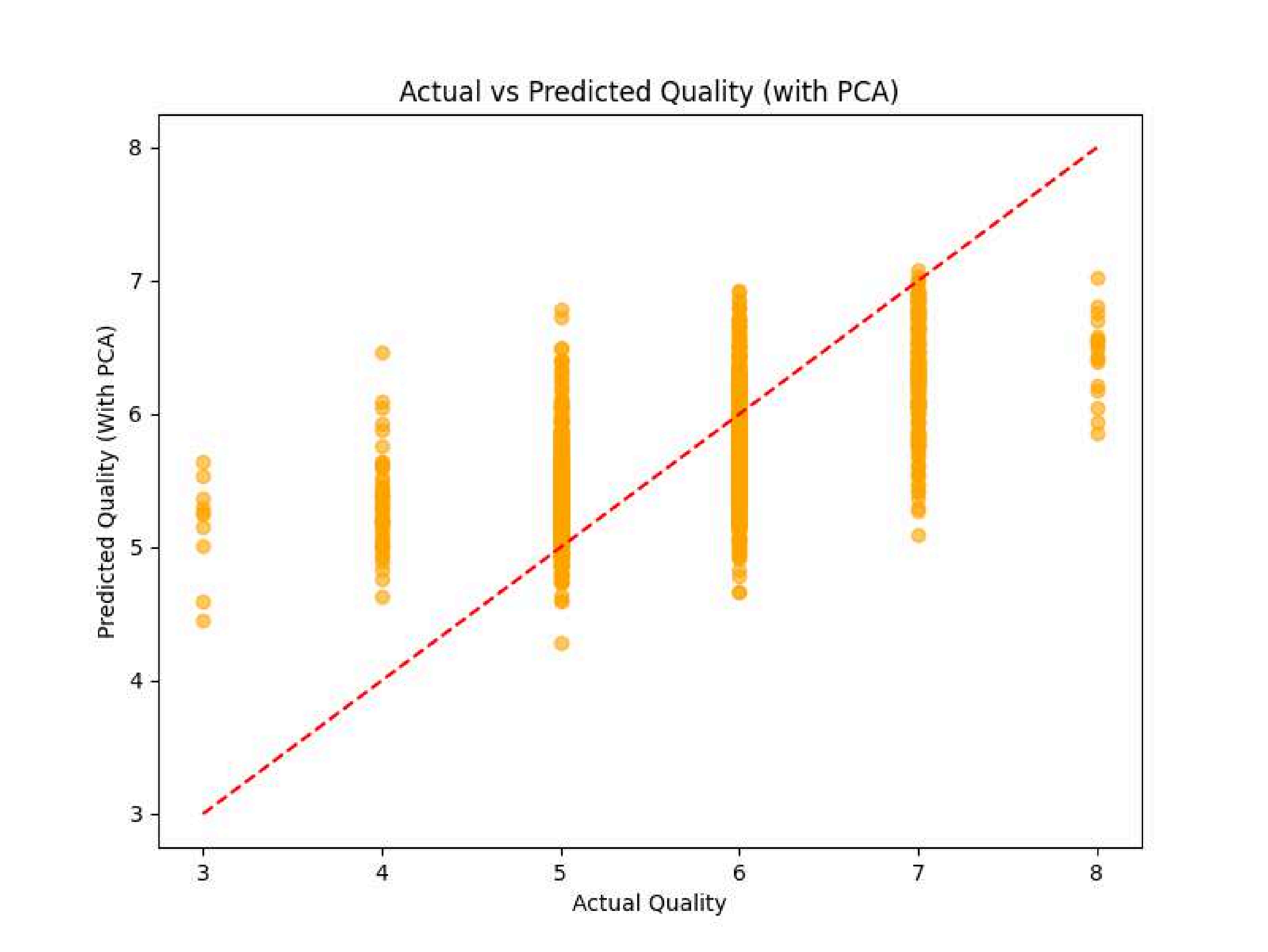
Training Process:

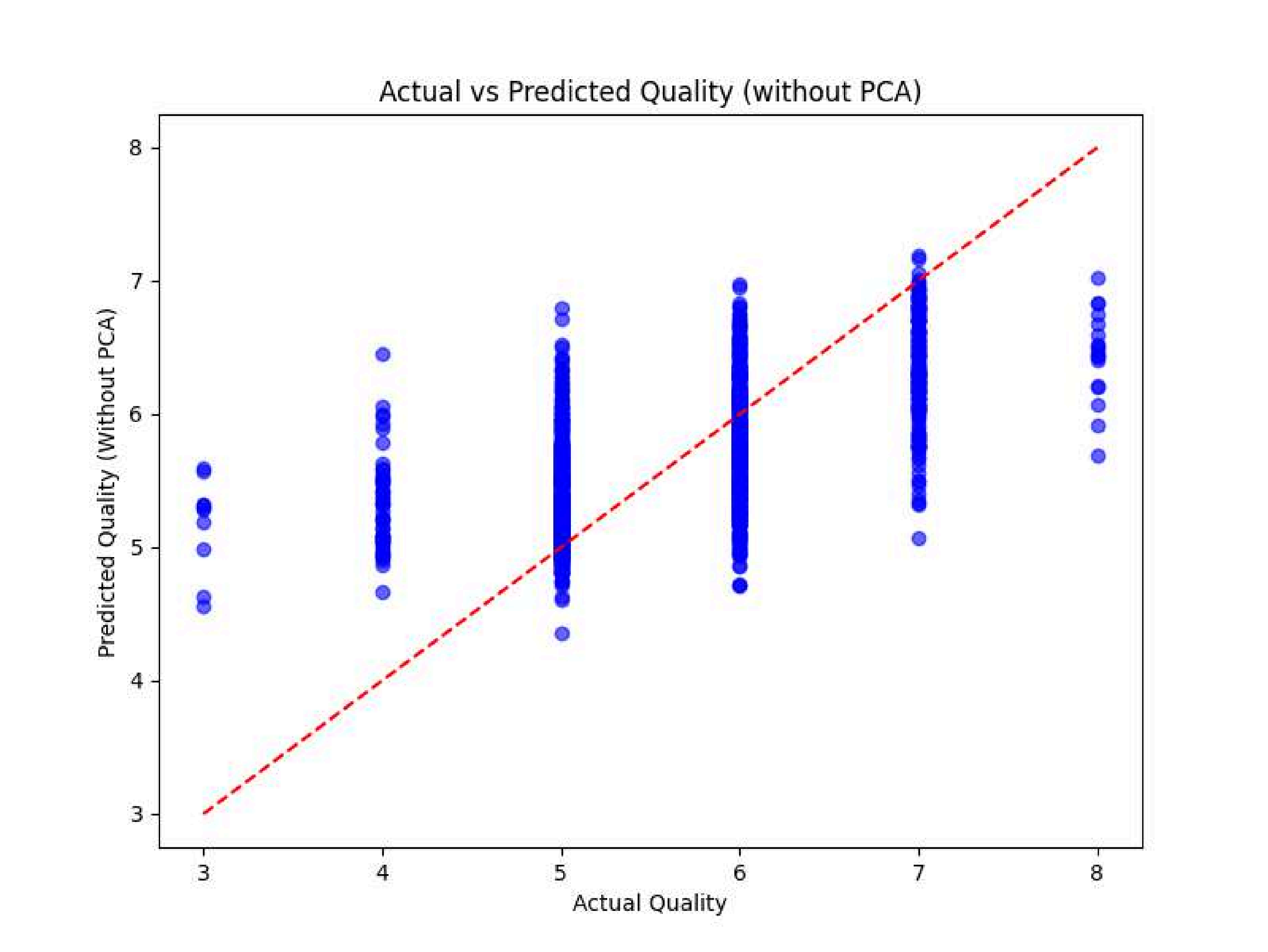
- **Dataset split:** 80% training (validication 5 fold => %16 of data), 20% testing.
- **Hyperparameter Tuning:** Used **Grid Search** for efficient parameter optimization. Explored combinations of:
  - **C (Regularization Parameter):** Controls the trade-off between achieving a low error on the training data and minimizing model complexity (i.e., overfitting).
  - **epsilon:** Specifies the margin of tolerance where no penalty is given for errors.
  - **kernel:** Defines the function used to map the data into a higher-dimensional space to enable linear separation.

Best Parameters Achieved:

- C: 1
- epsilon: 0.3
- kernel: 'rbf'

Actual vs Predicted





Cross-Validation Results (Without PCA):

- MAE: [0.46, 0.48, 0.46, 0.45, 0.48] (Mean: 0.47)
- MSE: [0.40, 0.44, 0.33, 0.35, 0.38] (Mean: 0.38)
- R2: [0.45, 0.41, 0.39, 0.40, 0.35] (Mean: 0.40)

Cross-Validation Results (With PCA):

- MAE: [0.49, 0.51, 0.46, 0.45, 0.48] (Mean: 0.48)
- MSE: [0.41, 0.45, 0.33, 0.35, 0.38] (Mean: 0.39)
- R2: [0.43, 0.40, 0.38, 0.40, 0.35] (Mean: 0.40)

There is not much a difference between with and without PCA.

About PCA

- From these results, we can conclude that PCA (Principal Component Analysis) did not improve the model's performance. In fact, it seems to have slightly worsened the Mean Absolute Error (MAE) and Mean Squared Error (MSE), while the R^2 value remained similar.

- By reducing the dimensionality of the data through PCA, you are essentially simplifying the data. While this can be beneficial in some cases (especially when the data is high-dimensional and noisy), it can also oversimplify the data and remove subtle patterns that might be important for the model’s predictions.

- This underfitting effect can happen if the reduced feature set does not capture the underlying complexity of the data. When the model is trained on this simplified data, it may perform worse than if it were trained on the original, more complex feature set.

About PCA

SVM Sensitivity to Data Transformation:

- Support Vector Machines (SVMs), particularly with the RBF kernel, are sensitive to the representation of the data. PCA might distort the feature space in such a way that the SVM hyperplane becomes less effective at separating or predicting the target variable.
- In some cases, non-linearity or interactions between features might be more easily captured by the original features rather than the principal components. By using PCA, you may have made the decision boundary harder for the SVM to identify effectively.

Summary

While PCA is typically used to reduce dimensionality and remove multicollinearity, it can have adverse effects when the dataset's important features do not align with the directions of maximum variance. For this wine quality dataset, PCA likely removed features that contribute prediction of the target variable, causing a slight decrease in performance in terms of MAE and MSE, while the R^2 value remained similar.

In short, PCA may have oversimplified the data and discarded some information that the SVM model needed to make accurate predictions, leading to worse performance. The key takeaway is that dimensionality reduction methods like PCA are not always beneficial, especially when the data has important features that are less correlated with the target but still relevant for prediction.

About PCA

Feature Engineering and Selection:

- Manual Feature Selection: Instead of relying on PCA, you can manually select important features based on domain knowledge or statistical tests. Features such as citric acid and fixed acidity, which were excluded by PCA, might be valuable for predicting wine quality.
- Correlation Analysis: Analyze the correlation between features and the target variable (wine quality) and keep the most relevant ones. Feature selection techniques like Recursive Feature Elimination (RFE) or L1 regularization can also help identify the most important features.

Conclusion

The performance analysis shows that while PCA was intended to reduce dimensionality and improve generalization, it may have led to a loss of important features that negatively impacted the SVM model's performance. To improve the results, it would be beneficial to focus on better feature engineering, more comprehensive hyperparameter tuning, and possibly trying other dimensionality reduction techniques. Additionally, exploring alternative models or using ensemble techniques could further enhance the predictive accuracy.

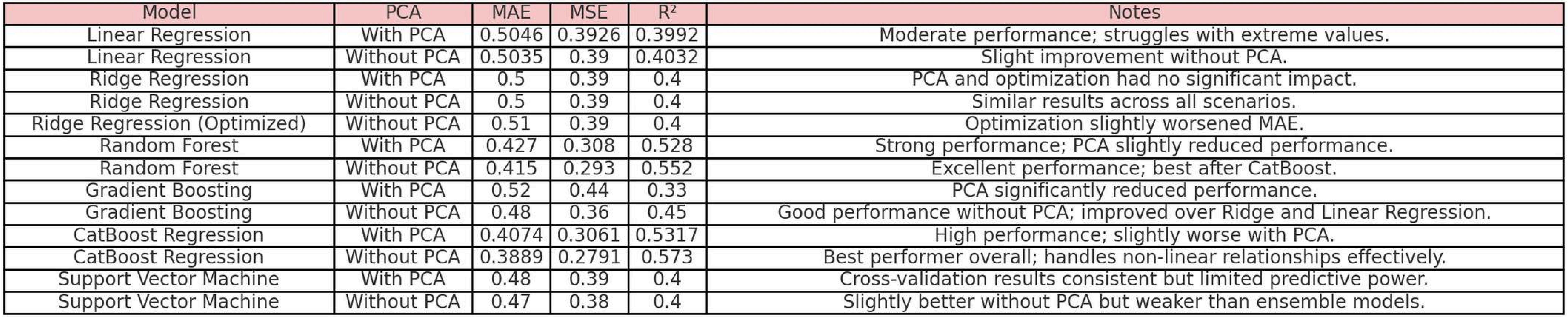
Conclusion

- Best Performer:
  - CatBoost Regression (Without PCA) emerged as the best model with R2=0.5730R^2 = 0.5730R2=0.5730, MSE = 0.2791, and MAE = 0.3889, demonstrating its ability to handle non-linear relationships effectively.
  - Random Forest (Without PCA) followed closely with R2=0.552R^2 = 0.552R2=0.552 and is a strong alternative due to its feature interpretability.
- PCA’s Limited Impact:
  - PCA reduced performance slightly in ensemble models (Random Forest, Gradient Boosting, CatBoost) because it may discard key feature-specific information.
  - For simpler models (Linear and Ridge Regression), PCA had negligible impact, confirming that dimensionality reduction was unnecessary for this dataset.
- Linear Models Underperform:
  - Both Linear and Ridge Regression achieved R2≈0.40R^2 \approx 0.40R2≈0.40, explaining only 40% of the variability in wine quality. These models struggled with non-linear relationships in the data.
- Gradient Boosting:
  - Moderate performance (R2=0.45R^2 = 0.45R2=0.45 without PCA), but significantly degraded with PCA (R2=0.33R^2 = 0.33R2=0.33).
- SVM Consistency:
  - Support Vector Machines were stable but less effective, with R2=0.40R^2 = 0.40R2=0.40. It performed similarly to Ridge and Linear Regression, highlighting its limitations on this dataset.

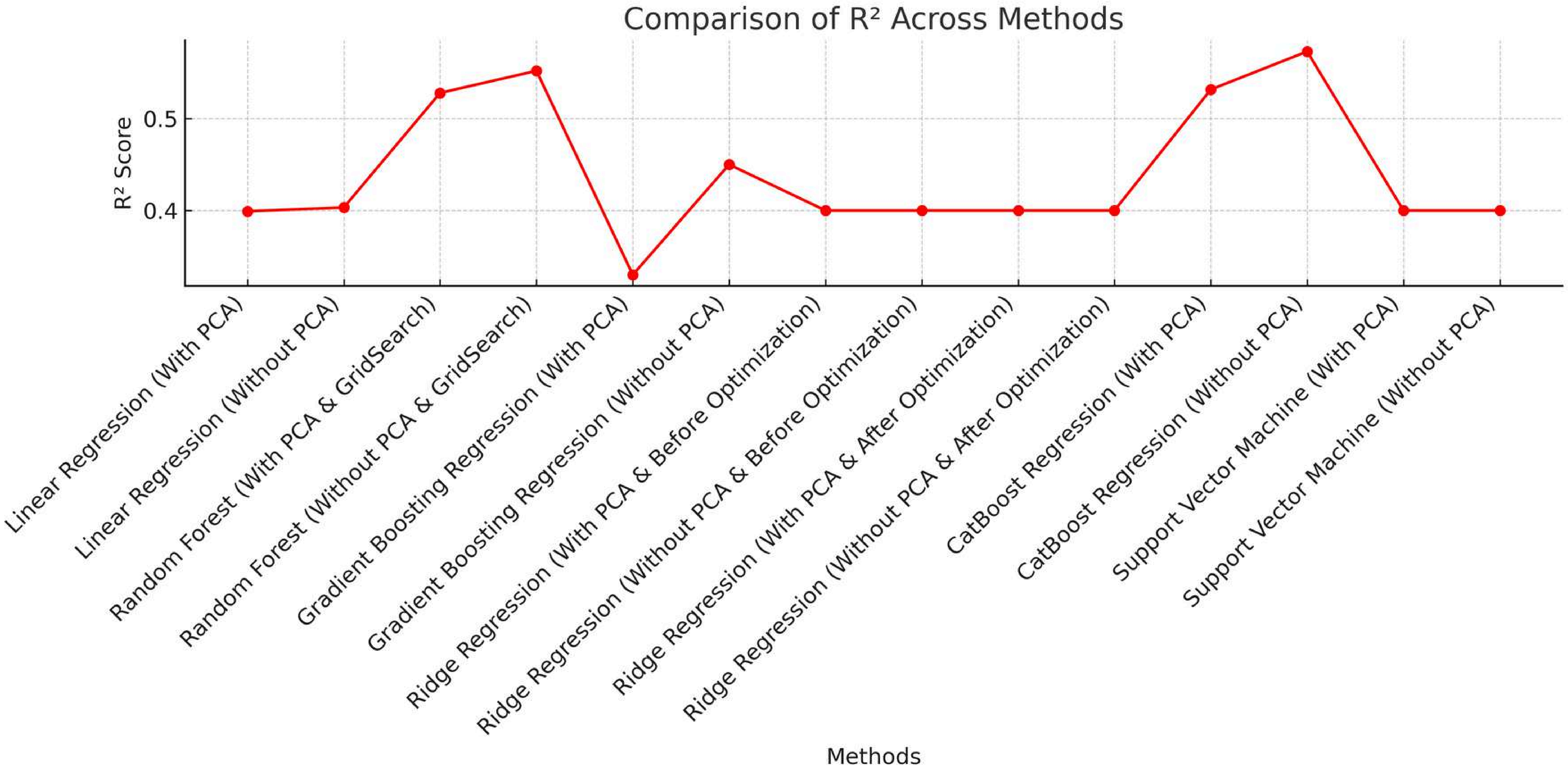
Comments

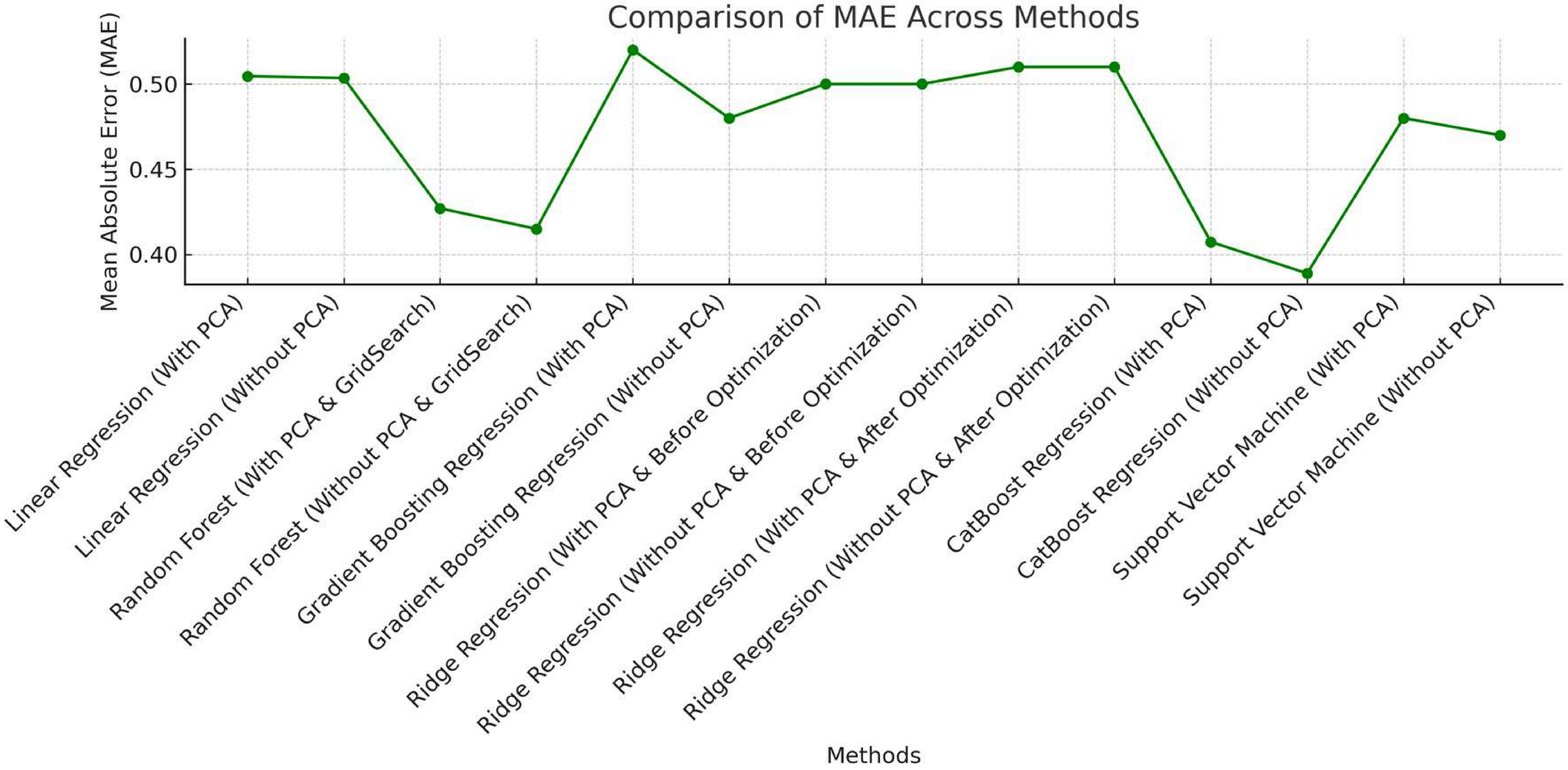
  - Ensemble models (CatBoost, Random Forest) excelled in handling complex, non-linear relationships, making them ideal for datasets with diverse feature interactions.
  - CatBoost's edge comes from its ability to handle categorical variables (if present) and robust built-in regularization.
- Dataset Characteristics:
  - The dataset is low-dimensional (11 features), so dimensionality reduction (PCA) had limited value and sometimes caused information loss.
  - Imbalanced data for extreme quality scores (e.g., 3 and 8) affected model predictions for those cases, evident in residual analyses.
  - Even the best model (CatBoost) explained only 57.3% of the variance in wine quality. This highlights that important factors like grape origin, vineyard conditions, or sensory data (e.g., taste, aroma) are missing.

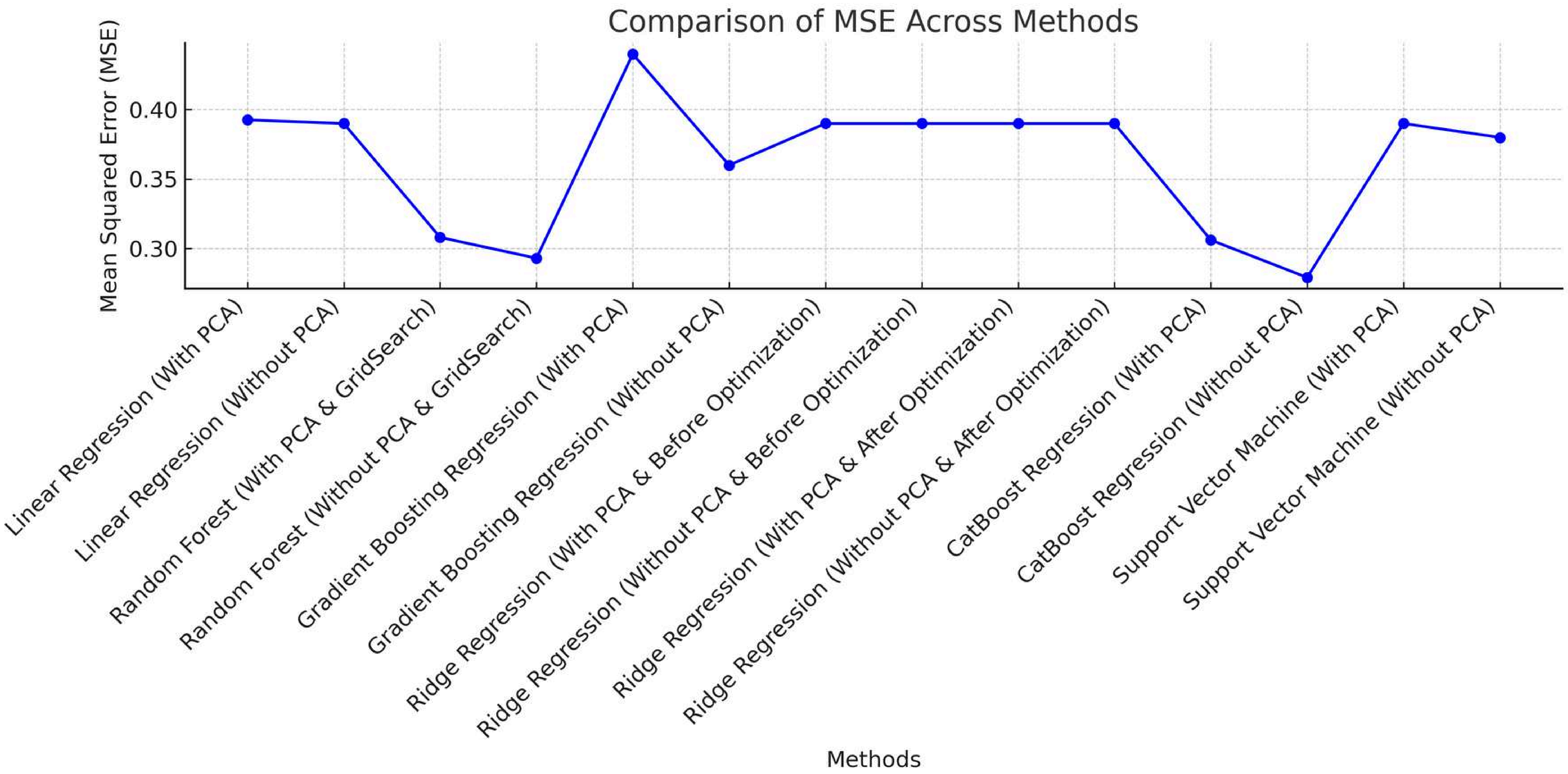
3.



Model	PCA	MAE	MSE	R²	Notes
Linear Regression	With PCA	0.5046	0.3926	0.3992	Moderate performance; struggles with extreme values.
Linear Regression	Without PCA	0.5035	0.39	0.4032	Slight improvement without PCA.
Ridge Regression	With PCA	0.5	0.39	0.4	PCA and optimization had no significant impact.
Ridge Regression	Without PCA	0.5	0.39	0.4	Similar results across all scenarios.
Ridge Regression (Optimized)	Without PCA	0.51	0.39	0.4	Optimization slightly worsened MAE.
Random Forest	With PCA	0.427	0.308	0.528	Strong performance; PCA slightly reduced performance.
Random Forest	Without PCA	0.415	0.293	0.552	Excellent performance; best after CatBoost.
Gradient Boosting	With PCA	0.52	0.44	0.33	PCA significantly reduced performance.
Gradient Boosting	Without PCA	0.48	0.36	0.45	Good performance without PCA; improved over Ridge and Linear Regression.
CatBoost Regression	With PCA	0.4074	0.3061	0.5317	High performance; slightly worse with PCA.
CatBoost Regression	Without PCA	0.3889	0.2791	0.573	Best performer overall; handles non-linear relationships effectively.
Support Vector Machine	With PCA	0.48	0.39	0.4	Cross-validation results consistent but limited predictive power.
Support Vector Machine	Without PCA	0.47	0.38	0.4	Slightly better without PCA but weaker than ensemble models.


Comparison of R² Across Methods
R² Score
0.5
0.4
Methods
Linear Regression (With PCA)
Linear Regression (Without PCA)
Random Forest (With PCA & GridSearch)
Random Forest (Without PCA & GridSearch)
Gradient Boosting Regression (With PCA)
Gradient Boosting Regression (Without PCA)
Ridge Regression (With PCA & Before Optimization)
Ridge Regression (Without PCA & Before Optimization)
Ridge Regression (With PCA & After Optimization)
Ridge Regression (Without PCA & After Optimization)
CatBoost Regression (With PCA)
CatBoost Regression (Without PCA)
Support Vector Machine (With PCA)
Support Vector Machine (Without PCA)


Comparison of MAE Across Methods
Mean Absolute Error (MAE)
0.50
0.45
0.40
Linear Regression (With PCA)
Linear Regression (Without PCA)
Random Forest (With PCA & GridSearch)
Random Forest (Without PCA & GridSearch)
Gradient Boosting Regression (With PCA)
Gradient Boosting Regression (Without PCA)
Ridge Regression (With PCA & Before Optimization)
Ridge Regression (Without PCA & Before Optimization)
Ridge Regression (With PCA & After Optimization)
Ridge Regression (Without PCA & After Optimization)
CatBoost Regression (With PCA)
CatBoost Regression (Without PCA)
Support Vector Machine (With PCA)
Support Vector Machine (Without PCA)
Methods


Comparison of MSE Across Methods
Mean Squared Error (MSE)
0.40
0.35
0.30
Linear Regression (With PCA)
Linear Regression (Without PCA)
Random Forest (With PCA & GridSearch)
Random Forest (Without PCA & GridSearch)
Gradient Boosting Regression (With PCA)
Gradient Boosting Regression (Without PCA)
Ridge Regression (With PCA & Before Optimization)
Ridge Regression (Without PCA & Before Optimization)
Ridge Regression (With PCA & After Optimization)
Ridge Regression (Without PCA & After Optimization)
CatBoost Regression (With PCA)
CatBoost Regression (Without PCA)
Support Vector Machine (With PCA)
Support Vector Machine (Without PCA)
Methods