**Multivariate Analysis of Physicochemical Properties for Wine quality dataset**

* Ravuri Lakshmi

**Abstract:**

This report provides a thorough analysis of addressing multicollinearity in a wine dataset that includes both red and white wines. The report describes an overview of the initial dataset, data visualization using mosaic and line charts, outlier detection using univariate and multivariate boxplots, identification of multicollinearity, and the application of Principal Component Analysis (PCA) to mitigate multicollinearity effects. Regression and classification models were run with and without PCA, and their performance was assessed. The results demonstrate the superiority of PCA in regression analysis and the efficacy of random forest models in both regression and classification. Nonetheless, a slight detrimental effect of PCA on classification was observed. Overall, PCA had a positive effect on regression analysis, whereas random forest models showed promise for addressing multicollinearity.

1. **Dataset overview:**

The dataset analyzed is the Vinho Verde Portuguese Wine dataset, which consists of two separate files: red wine and white wine.

The red wine data set includes 1,599 observations and 12 variables. Each observation represents a unique red wine sample, and the variables represent the wine's chemical properties and quality ratings. The following variables comprise the red wine dataset:

1. fixed.acidity: is a numeric variable that represents the wine's fixed acidity level.

2. volatile.acidity: A numeric variable representing the wine's volatile acidity.

3. citric.acid: A numeric variable representing the wine's citric acid concentration.

4. residual.sugar: A numeric variable representing the wine's residual sugar concentration.

5. chlorides: Numeric variable representing the wine's chloride content.

6. free.sulfur.dioxide: A numeric variable indicating the wine's free sulfur dioxide concentration.

7. total.sulfur.dioxide: A numeric variable that represents the total amount of sulfur dioxide in the wine.

8. density: A numeric variable representing the wine's density.

9. pH: Numeric variable representing the wine's pH level.

10. sulphates: A numeric variable representing the wine's sulphate concentration.

11. alcohol: Numeric variable representing the wine's alcohol content.

12. quality: An integer variable indicating the wine's quality rating (range: 3-9).

The white wine data set includes 4,898 observations and 12 variables. Similar to the red wine dataset, each white wine observation represents a unique sample, and the variables capture the wine's chemical properties and quality ratings. The white wine dataset contains the same variables as the red wine dataset.

There were no missing values in either dataset, and all variables are continuous except for quality, which is discrete.

It is possible to gain valuable insights into the characteristics and qualities of Vinho Verde Portuguese wine by exploring and analyzing these datasets. These insights can be utilized for further wine production, quality control, or consumer preference analysis, modeling, or decision-making processes.

1. **Data visualization:**

**Mosaic plot: Line plot:**

**A picture containing text, screenshot, diagram, parallel

Description automatically generatedA picture containing plot, diagram, text

Description automatically generated**

This plot is between the type of wine and the quality of the wine. This indicates that there are no quality records for red wine (9). Moreover, this indicates that the overall proportion of white wine is greater. Red wine appears to be of lower quality than high quality.

This line graph represents the relationship between various physiochemical characteristics of red and white wine (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol). Except for (total.sulphur dioxide and free sulphur dioxide), the red and white wines exhibited nearly identical characteristics.

1. **Univariate and Multivariate normality:**

A picture containing diagram, line, screenshot, plot

Description automatically generated

Red wine dataset White wine dataset

A picture containing text, diagram, screenshot, plot

Description automatically generated A picture containing text, screenshot, line, plot

Description automatically generated

For the red wine dataset:

• Multivariate Normality Test (Royston's Test): 1230.672, 2.848709e-258. We reject multivariate normality because the p-value is less than 0.05. The red wine dataset is not multivariate normal.

• Univariate Normality Test (Anderson-Darling Test): The Anderson-Darling test statistics and p-values show that all red wine dataset variables deviate from normality. Some variables have non-normal skewness and kurtosis.

For the white wine dataset:

• Multivariate Normality Test (Mardia's Test): 235569.17 skewness and 1461.7 kurtosis for the white wine dataset. Both test statistics have p-values of 0, indicating multivariate normality is violated. The white wine dataset is not multivariate normal.

• Univariate Normality Test (Anderson-Darling Test): Like the red wine dataset, the white wine dataset has significant departures from normality. Some variables exhibit non-normal skewness and kurtosis.

These results indicate that neither the red nor the white wine datasets follow a multivariate normal distribution. Both datasets' individual variables deviate from a normal distribution, according to univariate normality tests.

1. **Outlier detection:**

Royston found 2 outliers in the red wine dataset, which were removed. Royston failed for the white wine dataset with over 2000 observations, so the Mardia test found an outlier that was removed.

**Outliers in red and white wine dataset:**

|  |  |
| --- | --- |
| **Dataset** | **outlier** |
| red | 152 |
| red | 259 |
| white | 2782 |

1. **Multicollinearity:**

Analyzed dataset correlations and found high multicollinearity. We investigate linear relationships between features in the dataset. Correlations between variables reveal their interdependence and help us understand patterns and associations.

Correlation Analysis: The predictor variables' correlation matrix showed both positive and negative correlations. Some variables had strong linear relationships. Some variables were non-linear.

High Linear Relationships:

1. Density and Alcohol: A high negative correlation suggests an inverse relationship. Density decreases alcohol content.

2. Density and Residual Sugar: These variables correlated strongly, indicating a direct relationship. Density increases residual sugar.

Non-Linear Relationships: The following variables had no linear relationships:

1. Alcohol/Volatile Acidity

Density/Volatile Acidity

3. pH/Volatile Acidity

4. Alcohol /Sulphates

Multicollinearity: High correlations between predictor variables may indicate multicollinearity, or interdependence among independent variables. Regression models with high multicollinearity have inflated standard errors and unstable coefficient estimates.

Created a correlation heatmap. Blue indicates negative correlations, red positive, and white no correlation.

High variable correlations suggest multicollinearity. To ensure accurate modeling, further exploration and dimensionality reduction can be used.

**A picture containing text, screenshot, diagram, colorfulness

Description automatically generated A picture containing text, diagram, screenshot, line

Description automatically generated**

1. **Principal component analysis:**

PCA avoids multicollinearity. Two PCA variations with 11 and 12 variables were used on a wine dataset.

**1. Regression PCA for 11 variables:** Since the dataset has only 11 variables (excluding quality), only the first 11 components were taken.

The loadings shows principal component-original variable correlation coefficients. Positive and negative signs indicate correlation direction.

• "Fixed.acidity" has a high loading (0.544) on Comp.1 and moderate loadings on Comp.3 (0.460) and Comp.9 (0.602).

• Comp.2 strongly loads "volatile.acidity" (0.477).

• Comp.3 (-0.706) and Comp.9 (-0.311) load "citric.acid" significantly.

• "Residual.sugar" affects Comp.1 (0.390), Comp.3 (0.354), and Comp.9 (-0.371).

According to the summary table, Comp.1 to Comp.6 explain 77.77% of the dataset's variance. These 6 components be the most important.

**2. Classification PCA for 11 variables:** 12 components.

Loading findings:

• "Fixed.acidity" has a significant loading (0.536) on Comp.1 and moderate loadings on Comp.3 (0.477) and Comp.9 (0.574).

• "Volatile.acidity" has a significant loading on Comp.2 (-0.305) and Comp.3 (0.408).

"Citric.acid" is strongly loaded on Comp.7 (-0.720).

• "Residual.sugar" affects Comp.1, Comp.3, and Comp.9 (-0.337).

• Comp.5 and Comp.6 have high loadings for "chlorides" (-0.596 and -0.650).

• Comp.2 has a significant loading (-0.478) for "free.sulfur.dioxide" and Comp.9 has a moderate loading (0.447).

• Comp.1 and Comp.3 load "total.sulfur.dioxide" at 0.373 and -0.306, respectively.

• "Density" loads Comp.6 (-0.401).

• "pH" loads moderately on Comp.4 (-0.344) and Comp.9 (0.447).

• Comp.4 and Comp.5 load "sulphates" moderately (-0.346 and -0.597).

• Comp.6's "alcohol" variable loads 0.326.

According to the summary table, Comp.1 to Comp.6 explain 77.40% of the dataset's variance. These 6 elements were considered.

The 6 PCA components have no multicollinearity.

**Using 11 components Using 12 components**

A picture containing text, diagram, line, plot

Description automatically generated A picture containing text, diagram, line, plot

Description automatically generated

1. **Regression with PCA using 6 components and using 11 variables:**

Dataset with 11 physicochemical properties as input variables and the quality rating of the wine as the output variable. The dataset has been split into a training set and a testing set, with an 80% training set and a 20% testing set. Evaluation results for the models with 11 variables and with PCA (6 components):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **With 11 variables** | | **With 6 Principal components** | |
| **Model** | **Evaluation Metric** | **Training Set** | **Testing Set** | **Training Set** | **Testing Set** |
| Multivariate Linear Regression | MSE | 0.3528 | 0.4158 | 0.3893 | 0.4082 |
|  | RMSE | 0.8117 | 0.8117 | 0.8205 | 0.8205 |
|  | MAE | 0.4469 | 0.4706 | 0.4647 | 0.4724 |
|  | R-squared | 0.6450 | 0.6061 | 0.6220 | 0.6164 |
| Decision Tree | MSE | 0.0418 | 0.5072 | 0.0410 | 0.5209 |
|  | RMSE | 0.8117 | 0.8117 | 0.8205 | 0.8205 |
|  | MAE | 0.1558 | 0.5586 | 0.1530 | 0.5659 |
|  | R-squared | 0.9666 | 0.5052 | 0.9674 | 0.4903 |
| Random Forest | MSE | 0.0187 | 0.2383 | 0.0185 | 0.2356 |
|  | RMSE | 0.8117 | 0.8117 | 0.8205 | 0.8205 |
|  | MAE | 0.1024 | 0.3423 | 0.1018 | 0.3433 |
|  | R-squared | 0.9898 | 0.7987 | 0.9899 | 0.8004 |
| Support Vector Regression (SVR) | MSE | 0.1704 | 0.2575 | 0.2052 | 0.2781 |
|  | RMSE | 0.8117 | 0.8117 | 0.8205 | 0.8205 |
|  | MAE | 0.2888 | 0.3817 | 0.3397 | 0.3967 |
|  | R-squared | 0.8252 | 0.7428 | 0.7903 | 0.7184 |

The table provides information on the performance of each model on both the training and testing sets. Random Forest appears to perform best in terms of most evaluation metrics for models with 11 variables and PCA (6 components). In addition, PCA-based models had higher R-squared and lower other metrics, indicating their superiority.

**Without PCA:**

A picture containing text, screenshot, diagram, font

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

**With PCA:**

A picture containing text, screenshot, diagram, font

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

1. **Linear regression:**

Linear regression between all the variables separately and the quality variable is performed. Also, with PCA, linear regression between each of the 6 components independently and the quality variable.

**Linear regression without PCA:A picture containing text, screenshot, font, number

Description automatically generated**

• The evaluation metrics for the training set and test set are calculated for each variable independently.

•The performance of the linear regression model varies across different variables.

•None of the variables show a strong relationship with the quality rating of the wine, as indicated by low R-squared values.

•MSE, RMSE, and MAE, are generally high, indicating a poor fit of the model to the data.

A picture containing text, screenshot, line, font

Description automatically generatedA picture containing text, screenshot, font, line

Description automatically generatedA picture containing text, screenshot, line, font

Description automatically generatedA picture containing text, screenshot, font, diagram

Description automatically generated **Linear regression with PCA:**

A picture containing text, screenshot, font, number

Description automatically generated

* The linear regression model is evaluated using principal components
* PC R-squared values show how much variance in quality rating each PC explains.
* The training and test set evaluation metrics are lower than the model without PCA, indicating a better model fit. Even with PCA, the linear regression model may not capture the complexity of the relationship between input variables and wine quality. A picture containing text, screenshot, font, diagram

  Description automatically generated A picture containing text, screenshot, font, number

  Description automatically generated

A picture containing text, screenshot, font, number

Description automatically generated A picture containing text, screenshot, diagram, font

Description automatically generated With or without PCA, both models have low R-squared values and high error metrics. Using these physicochemical properties, linear regression may not be the most accurate way to predict wine quality.

1. **Classification with PCA using 6 components and using 12 variables:**

The dataset has 12 input variables, including acidity, alcohol content, and others. Predicting red or white wine is the objective. It's important to note that red and white wine classes are imbalanced. Red wine is more common than white wine. A classifier may be biased toward the majority class if the classes are imbalanced. To address class imbalance, weights are assigned to each class. In this case misclassifying a red wine instance is three times worse than a white wine instance. White wine gets 1 weight, whereas red gets 3. An 80-20 split divides the dataset into a training and testing set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Random Forest | With PCA | 0.8729792 | 0.7324415 | 0.72039474 | 0.7263682 |
| Random Forest | Without PCA | 0.9946154 | 0.9905956 | 0.98750 | 0.9890454 |
| Gradient Boost | With PCA | 0.6720554 | 0.3780000 | 0.62171053 | 0.4701493 |
| Gradient Boost | Without PCA | 0.9300000 | 0.8021108 | 0.95000 | 0.8698140 |
| SVM | With PCA | 0.8290993 | 0.7662338 | 0.38815789 | 0.5152838 |
| SVM | Without PCA | 0.9246154 | 0.9370079 | 0.74375 | 0.8292683 |
| Naive Bayes | With PCA | 0.7628945 | 0.4666667 | 0.09210526 | 0.1538462 |
| Naive Bayes | Without PCA | 0.7530769 | 0.0000000 | 0.00000 | NaN |
| KNN | With PCA | 0.8306390 | 0.6627907 | 0.56250000 | 0.6085409 |
| KNN | Without PCA | 0.8969231 | 0.8000000 | 0.77500 | 0.7873016 |
| Decision Tree | With PCA | 0.6743649 | 0.3946903 | 0.73355263 | 0.5132336 |
| Decision Tree | Without PCA | 0.9761538 | 0.9313433 | 0.97500 | 0.9526718 |

The models without PCA (using all 12 variables) perform better in terms of accuracy, precision, recall, and F1-score compared to those with PCA, as shown in the table. This indicates that retaining all original variables provides more information and results in improved classification performance. Consequently, in this scenario, models without PCA are regarded as superior. In addition, the Random Forest without PCA is the best model for this classification task among the models compared.

**Without PCA:**

A picture containing text, screenshot, rectangle, diagram

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, rectangle, diagram

Description automatically generated

**With PCA:**

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generated

1. **Conclusion:**

It is observed that PCA showed better performance in regression compared to using all variables, indicating that PCA effectively captured the important information in the dataset and reduced the impact of multicollinearity.

This suggests that PCA can be an effective technique to address multicollinearity in regression analysis. On the other hand, in classification, the Random Forest model performed well in both cases, with and without PCA. However, it is worth noting that PCA had a slightly negative effect on classification metrics such as precision, recall, and F1-score. This suggests that using all variables without PCA might be more suitable for classification tasks in this dataset.