김종윤
2023-28318
협동과정 인공지능 전공

소셜컴퓨팅

과제 7 - 11, 12주차

1. [Week11] Supervised Learning (Red Wine Quality Analysis)

1.1. Task Definition

The various features can be extracted while making red wine. Moreover, a number of reviewers taste the wine and evaluate it in numeric scale. Given the features extracted from the wine, predicting the wine quality is the target task. Although the original evaluation numerics are given in decimal, as this week's lecture was about classification, the numeric values are converted to binary with some reasonable threshold.

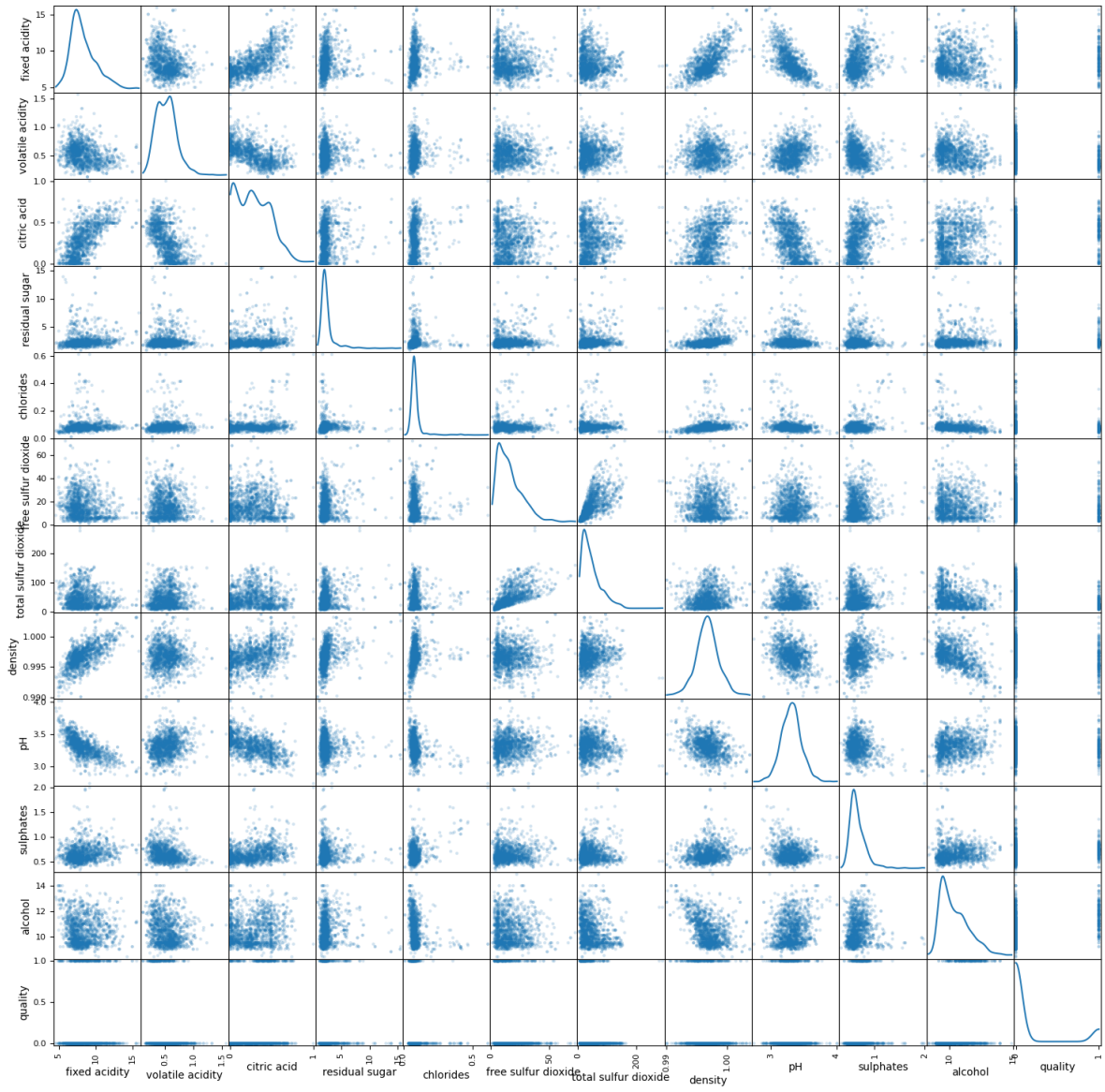
1.2. Dataset

The dataset consists of 12 columns and 1599 records. The 11 features are "fixed acidity", "volatile acidity", "citric acid", "residual sugar", "chlorides", "free sulfur dioxide", "total sulfur dioxide", "density", "pH", "sulphates", "alcohol" and the target prediction is given with column name "quality".

The "quality" is originally given in scale of 2~8 but as the target task is classification, the values are split into two by 6.5. Below 6.5 is considered as bad quality and over 6.5 is good quality.

The distribution of each feature is given below. From the attribute scatter plot matrix, each feature has reasonable distribution which can be described with mixture of Gaussian distribution.

The 25% of the total dataset is set as test dataset and 1/5 of rest of dataset is set as validation dataset and the remaining is used for training. To improve the performance 5-fold Cross-Validation is used.



[Figure1: attribute scatter plot matrix]

1.3. Models

All the models are directly used from scikit-learn framework implementation. As scikit-learn provides Cross-Validation with grid searching of model parameters as `GridSearchCV`, the hyper parameters are explored with it. The list of hyper parameters searched are provided below.

1.3.1. GaussianNB

1.3.1.1. Hyper Parameters

None of hyper parameters are explored for Gaussian model.

1.3.2. Logistic Regression

1.3.2.1.Hyper Parameters

None of hyper parameters are explored for Logistic Regression model.

1.3.3. Support Vector Machine (Classifier)

1.3.3.1.Hyper Parameters

The three hyper parameters are explored for SVC: 'C', 'kernel' and 'gamma'. The values explored for 'C' is[0.1, 0.8, 0.9, 1.1, 1.2, 1.3, 1.4], 'gamma' also has equal values as 'C' and 'kernel' are ['linear', 'rbf'].

1.3.4. Decision Tree

1.3.4.1.Hyper Parameters

The hyper parameter explored is 'max_depth' and two values are explored: 3, None.

1.3.5. Random Forest

1.3.5.1.Hyper Parameters

One hyper parameter is searched, n_estimators, and the values explored are 100, 200.

1.3.6. k - Nearest Neighbors

1.3.6.1.Hyper Parameters

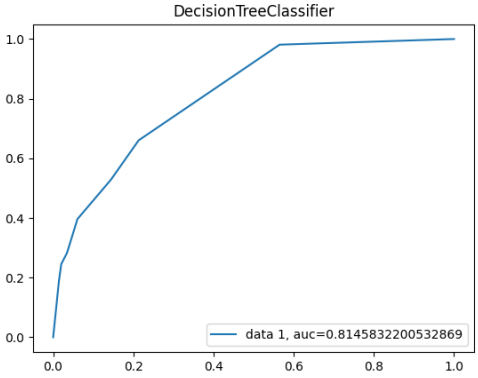
Single hyper parameter is explored, n_neighbors, and 1, 3, 5 are explored.

1.4. Results

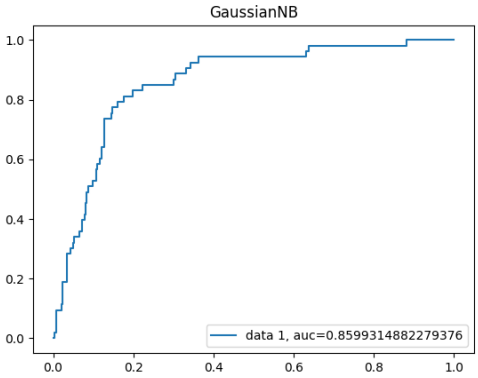
The best model with criteria accuracy is 'SVC' and with precision is also 'SVC'.

The AUC curve is provided on next page.

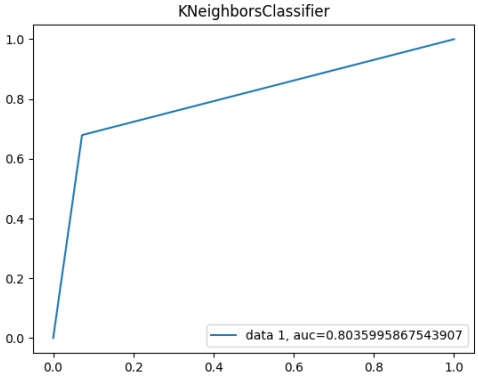
Model	Best Hyper parameter	confusion matrix	accuracy	precision	recall
Gaussian	N/A	[[305 42] [19 34]]	84.75	44.74	64.15
Logistic Regression	N/A	[[338 9] [37 16]]	88.5	64	30.19
SVC	{'C': 1.2, 'gamma': 0.9, 'kernel': 'rbf'}	[[343 4] [32 21]]	91	84	39.62
Decision Tree	{'max_depth': 3}	[[335 12] [38 15]]	87.5	55.55	28.30
Random Forest	{'n_estimators': 200}	[[332 15] [34 19]]	87.75	55.88	35.85
k-NN	{'n_neighbors': 1}	[[322 25] [17 36]]	89.5	59.02	67.92



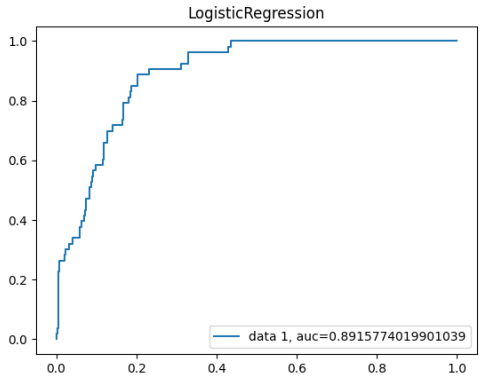
[Figure 2: Decision Tree AUC Curve]



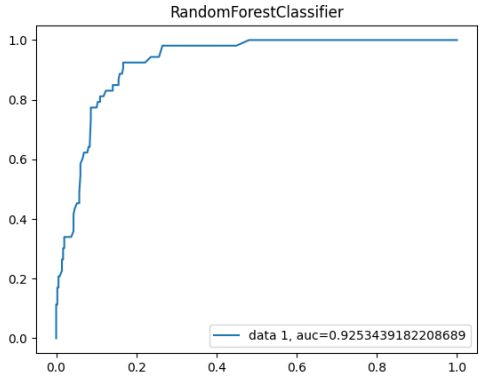
[Figure 3: GaussianNB AUC Curve]



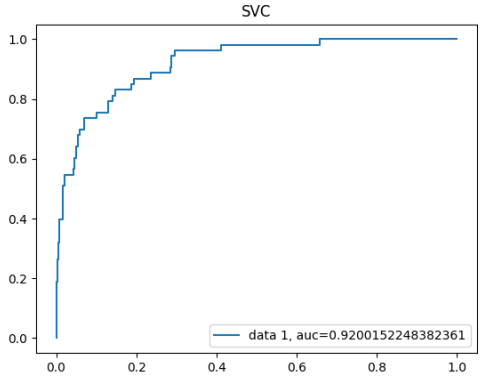
[Figure 4: k-NN AUC Curve]



[Figure 5: Logistic Regression AUC Curve]



[Figure 6: Random Forest AUC Curve]



[Figure 7: SVC AUC Curve]

2. [Week12] Unsupervised Learning - k-Means Clustering (Red Wine Quality, White Wine Quality)

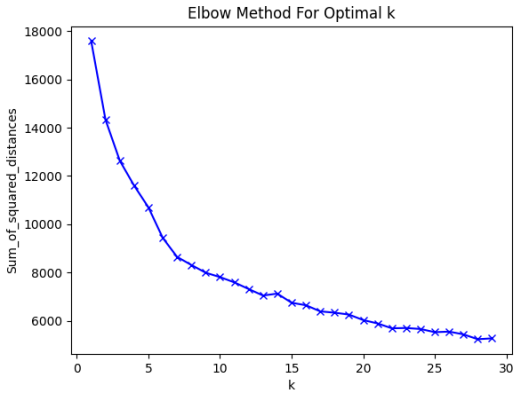
The k-Means clustering is performed on both red wine and white wine dataset. The range of k is from 1 to 30 for 'Red wine dataset' and 'White wine dataset' (not for the combination).

To visually prove which k is optimal to cluster the wine, 'PCA' and 't-SNE' are used.

2.1. Red Wine dataset

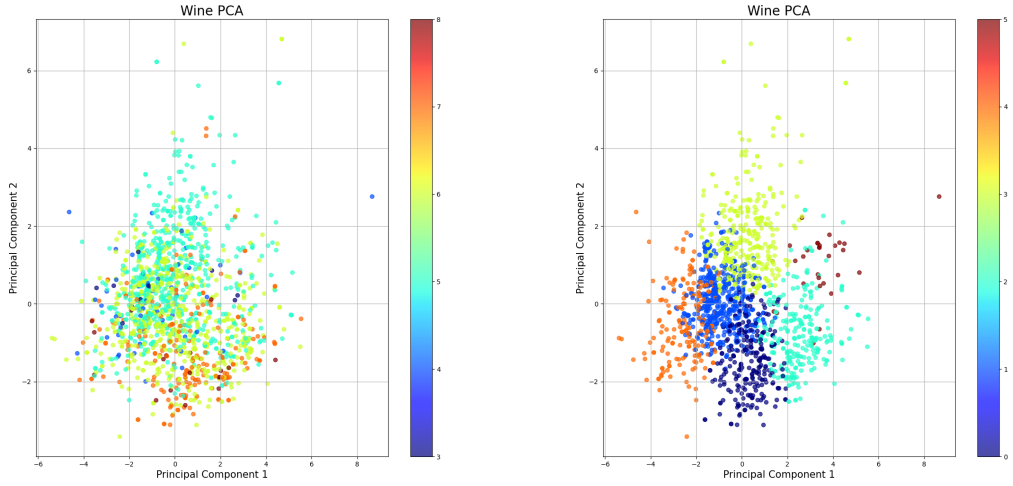
Primarily, the dataset's features' explainability on PCA is about 45.6%.

With 'Elbow Method', the optimal k for the red wine dataset is 6 as the L2 distance does not dramatically drop after k=6.

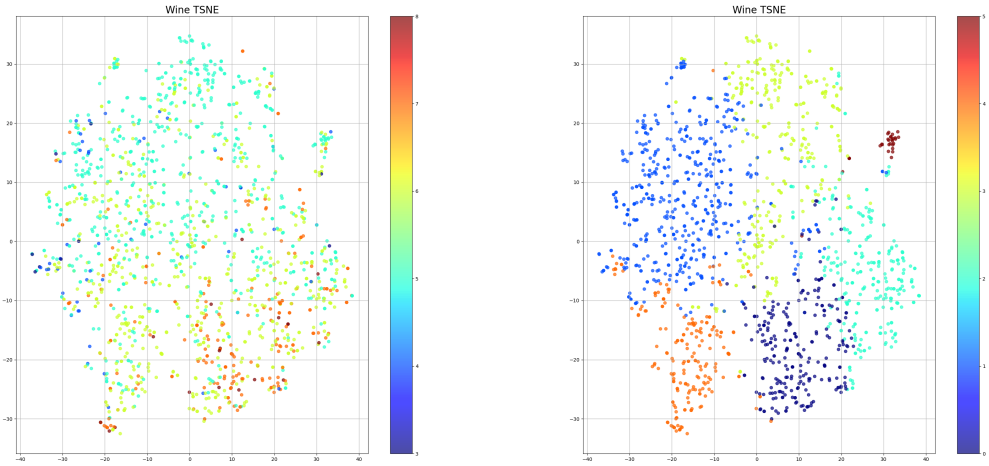


[Figure: Elbow method for finding optimal k in red wine dataset]

The PCA visualization and t-SNE visualization show that even the optimal k is selected, the data points are not well grouped.



[Figure: PCA visualization for red wine dataset (Left: true label, Right: k-Means clustering)]



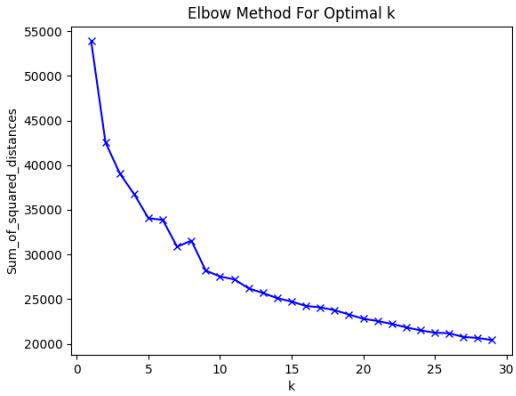
[Figure: t-SNE visualization for red wine dataset (Left: true label, Right: k-Means clustering)]

Therefore, we can say that clustering is not appropriate for this dataset to predict the quality of the wine.

2.2. White Wine dataset

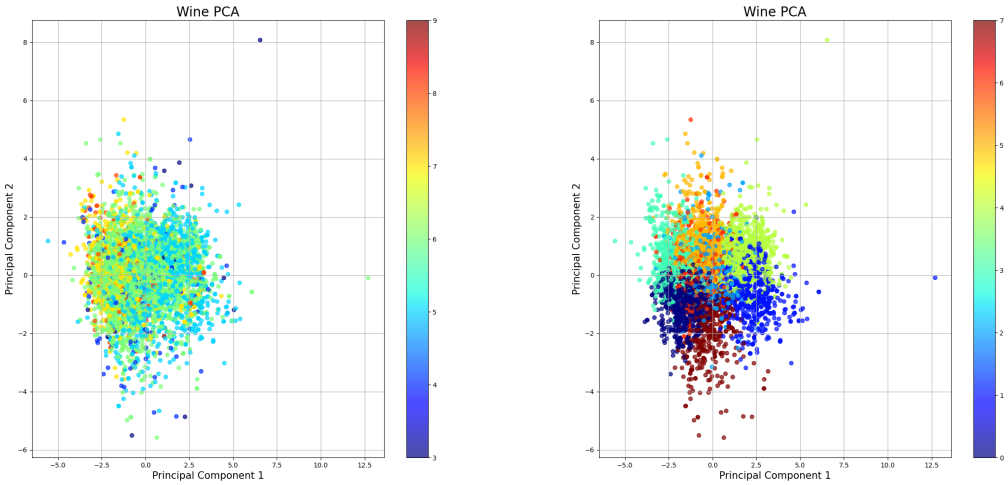
Primarily, the dataset's features' explainability on PCA is about 43.6%.

With 'Elbow Method', the optimal k for the red wine dataset is 8 because until 8 the L2 distance between each datapoint drops drammatically.

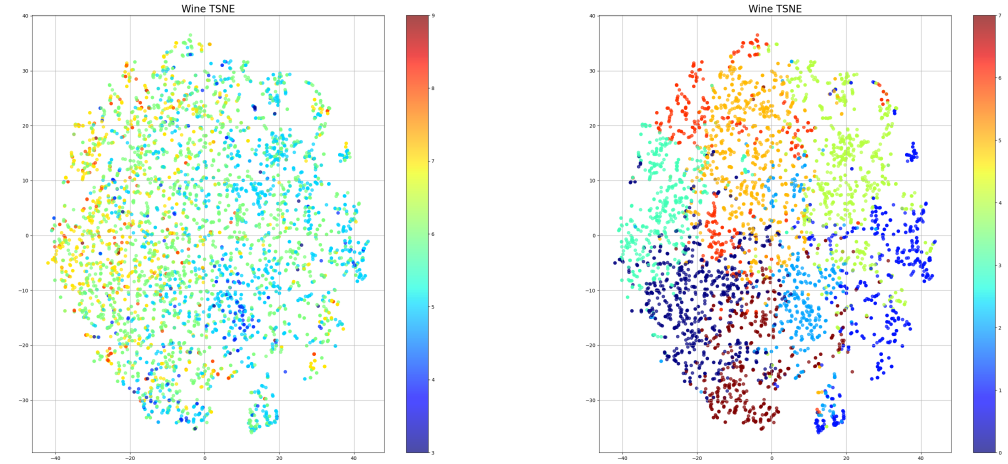


[Figure: Elbow method for finding optimal k in white wine dataset]

The PCA visualization and t-SNE visualization show that even the optimal k is selected, the data points are not clustered as desired.



[Figure: PCA visualization for white wine dataset (Left: true label, Right: k-Means clustering)]



[Figure: t-SNE visualization for white wine dataset (Left: true label, Right: k-Means clustering)]

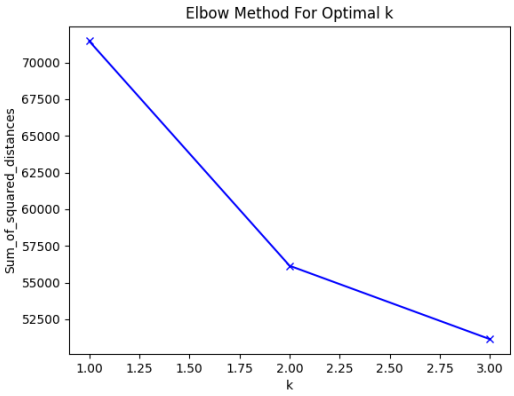
Therefore, we can say that clustering is not appropriate for this dataset to predict the quality of the wine.

2.3. Combination of Red & White wine dataset

Primarily, the dataset's features' explainability on PCA is about 50.2%.

With 'Elbow method', the optimal k is search for 2 and 3, as the dataset is combination of red and white wine dataset.

The optimal k for this dataset combination is obviously 2.

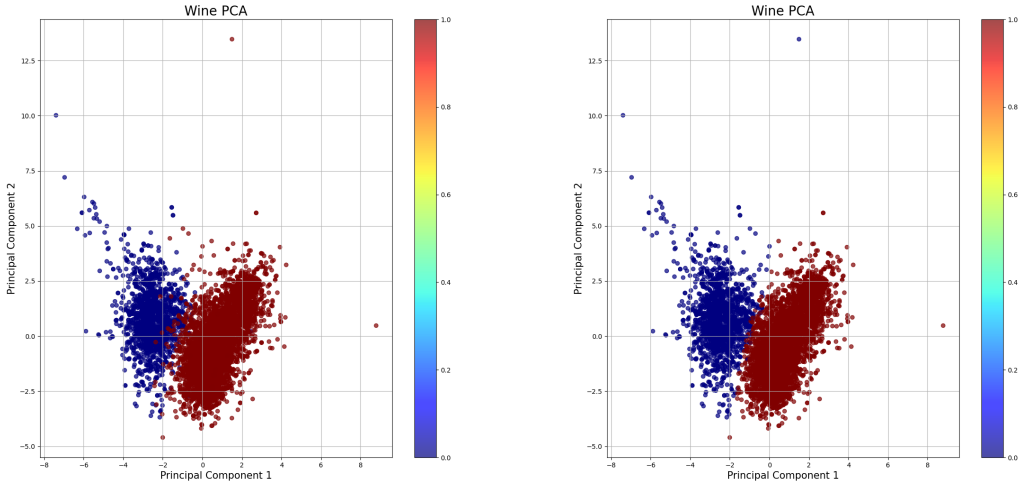


[Figure: Elbow method for red and white wine dataset to find the optimal k]

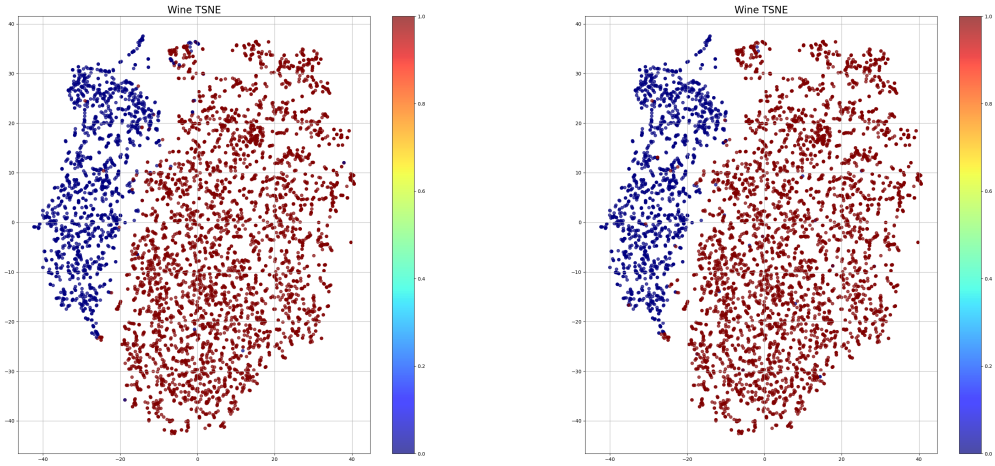
The PCA visualization and t-SNE visualization are drawn below.

The true label (0 for red, 1 for white) visualization for PCA and t-SNE are also provided along with the clustering.

As it can be seen clearly on visualization, the clustering is processed as desired. Some outliers can be found on true label. These data points can be considered as the red wine that has chemical features like white wine and vice versa.



[Figure: PCA visualization for red and white wine dataset (Left: true label, Right: k-Means clustering)]



[Figure: t-SNE visualization for red and white wine dataset
(Left: true label, Right: k-Means clustering)]

3. Conclusion

Therefore, here we can say that the distinguishing the type of wine whether it is red or white can be done with k-Means clustering while the classification or regression can be done with various other supervised classification or regression models such as 'Support Vector Machine' or 'RandomForest'. The clustering result clearly shows that k-Means can distinguish two type of wine with k=2. The SVC (Support Vector Machine for Classification) achieved about 91% in both accuracy and recall for classification of the red wine quality.