Classic Machiene LEarning Analyis on Wine Data set using Clustering and Classification

Contents

[Abbreviations Used in the Document: 2](#_Toc150184161)

[Agglomerative Clustering 2](#_Toc150184162)

[DBSCAN 2](#_Toc150184163)

[Gaussian Mixture Model 2](#_Toc150184164)

[Affinity Propagation 2](#_Toc150184165)

[SVM 2](#_Toc150184166)

[Decision Trees 2](#_Toc150184167)

[Gradient Boost 2](#_Toc150184168)

[Random Forest 2](#_Toc150184169)

[Adaboost 2](#_Toc150184170)

[Clustering: 3](#_Toc150184171)

[Classification: 4](#_Toc150184172)

# Abbreviations Used in the Document:

|  |  |
| --- | --- |
| KM | k-Means |
| AC | Agglomerative Clustering |
| DB | DBSCAN |
| GMM | Gaussian Mixture Model |
| AP | Affinity Propagation |
| SVM | Support Vector Machines |
| DT | Decision Trees |
| GB | Gradient Boost |
| RF | Random Forest |
| AB | Adaboost |

# Clustering:

We have performed 6 clustering methods overall. (KM,AC,DB,GMM,AP and BIRCH)

Description:

In, K-means,  K = 3 has good Silhouette Score and the clusters are also pretty even from the below image. Hence, K=3 is the best value for k-means for Wine dataset.

Chart, bar chart

Description automatically generated

PCA has been done for dimensionality reduction to the data and k-means pretty much better clusters data when PCA is done from the below image.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Silhouette score (-1 to 1, ideally should be 1) | Calinski-Harabasz score (ideally should be higher) | Davies-Bouldin score (ideally should be 0) | V measure score (0 to 1, ideally should be 1) | Adjusted rand index score (0 to 1, ideally should be 1) |
| K Means | 0.28 | 70.94 | 1.39 | 0.88 | 0.9 |
| K Means after PCA | 0.31 | 82.75 | 1.27 | 0.88 | 0.9 |
| AC | 0.3 | 79.85 | 1.26 | 0.75 | 0.74 |
| DB Scan | 0.23 | 35.94 | 3.34 | 0.54 | 0.42 |
| GMM- Tied Model | 0.31 | 81.47 | 1.28 | 0.89 | 0.91 |
| GMM- Full Model | 0.31 | 81.14 | 1.28 | 0.86 | 0.88 |
| AF | 0.31 | 80.32 | 1.27 | 0.81 | 0.85 |
| BIRCH | 0.31 | 79.09 | 1.3 | 0.85 | 0.88 |

Chart, scatter chart

Description automatically generated

Looking at the values K-means better performed after PCA as there is an increase in score in all internal metrics though external remained same. Now after doing PCA, the Silhouette score for all metrics is pretty much at 0.31, not really ideal as the ideal score is 1. All algorithms for the Calinski-Harabasz score metric are between 79-81.47, in this K Means after PCA is doing the best with highest score. Ideal value for Davies-Bouldin score should be 0 and the least is AC and the highest is DB accounting it to not performing well. V measure metric should ideally be one almost every algorithm is in range of 0.8+ apart from AC and DB scan. Adjusted rand index score should ideally be 1 closest are K-means and GMM tied model and least performing being DB at a high difference of almost 0.58 from the ideal value.

Conclusion:

By comparing all the ideal values, best performing algorithm will be k-means after PCA.

However, V-measure and adjust rand index score had better values with GMM-tied model along with other metrics being at good performance. Hence, GMM- tied model can be derived as the second best clustering model for this data.

DB scan is the worst performing of all in all metrics and will not be suggested as an suitable algorithm for this data.

Similar analysis can be made by looking at below visualisation, the KNN has the best clustered data and GMM is closer to it. DB scans data looks very scattered and mixed suggesting it to be very unclear and ill clustered.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

# Classification:

All classifications models have been applied to data that is split by train test algorithm and k-fold. We have performed 6 classification methods overall. (SVM,DT,KNN,GB,RF and AB)

Description:

The training set is trained and fit to the SVM classifier. Prediction is applied to the X\_test(testing data).

For the best type of SVM and best parameters grid search is applied. The output clearly states the best Kernelized SVM model is RBF, C value is 100 and gamma is 0.001. The score for these parameters is at 97% which can be considered as a good result.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Balanced Accuracy** | **Precision score** | **Recall score** | **F1 score** | **ROC AUC score** |
|  |  | **Split Protocol** |  |  |  |
| **Normal Parameters** | 0.98-SVM  0.97- DT  0.97- KNN  0.94-GB  0.97-RF  0.89-AB | 0.98- SVM  0.97-DT  0.97-KNN  0.95-GB  0.97-RF  0.9- AB | 0.98-SVM  0.97-DT  0.97-KNN  0.95-GB  0.97-RF  0.9- AB | 0.98-SVM  0.97- DT  0.97-KNN  0.95-GB  0.97-RF  0.9- AB | 1-SVM  0.98-DT  0.98-KNN  1-GB  1-RF  0.94- AB |
| **Best Fit Parameters after grid search** | 0.97-SVM  0.97- DT  0.99-KNN  0.94-GB  0.97-RF  0.96-AB | 0.97-SVM  0.97- DT  0.98-KNN  0.95-GB  0.97-RF  0.97-AB | 0.97-SVM  0.97- DT  0.98-KNN  0.95-GB  0.97-RF  0.97-AB | 0.97-SVM  0.97- DT  0.98- KNN  0.95-GB  0.97-RF  0.97-AB | 1-SVM  0.98- DT  1-KNN  0.99-GB  1-RF  0.99-AB |
|  |  | **K Fold** |  |  |  |
| **Fold 1** | 1-SVM  0.96-DT  1-KNN  0.96-GB  0.96-RF  0.96-AB | 1-SVM  0.95-DT  1-KNN  0.95-GB  0.95-RF  0.95-AB | 1-SVM  0.95-DT  1-KNN  0.95-GB  0.95-RF  0.95-AB | 1-SVM  0.95-DT  1-KNN  0.95-GB  0.95-RF  0.95-AB | 1-SVM  0.97-DT  1-KNN  0.97-GB  1-RF  0.99-AB |
| **Fold 2** | 0.96-SVM  0.96-DT  0.95-KNN  0.92-GB  0.96-RF  0.96-AB | 0.95-SVM  0.95-DT  0.95-KNN  0.91-GB  0.95-RF  0.95-AB | 0.95-SVM  0.95-DT  0.95-KNN  0.91-GB  0.95-RF  0.95-AB | 0.95-SVM  0.95-DT  0.95-KNN  0.91-GB  0.95-RF  0.95-AB | 0.99-SVM  0.96-DT  1-KNN  0.96-GB  0.97-RF  0.99-AB |
| **Fold 3** | 0.95-SVM  0.91-DT  0.95-KNN  0.86-GB  0.95-RF  0.90-AB | 0.95-SVM  0.90-DT  0.95-KNN  0.86-GB  0.95-RF  0.90-AB | 0.95-SVM  0.90-DT  0.95-KNN  0.86-GB  0.95-RF  0.90-AB | 0.95-SVM  0.90-DT  0.95-KNN  0.86-GB  0.95-RF  0.90-AB | 1-SVM  0.95-DT  0.95-KNN  0.98-GB  0.99-RF  0.98-AB |
| **Fold 4** | 0.95-SVM  0.95-DT  0.90-KNN  0.95-GB  0.89-RF  0.95-AB | 0.95-SVM  0.95-DT  0.90-KNN  0.95-GB  0.90-RF  0.95-AB | 0.95-SVM  0.95-DT  0.90-KNN  0.95-GB  0.90-RF  0.95-AB | 0.95-SVM  0.95-DT  0.90-KNN  0.95-GB  0.90-RF  0.95-AB | 1-SVM  0.96-DT  0.99-KNN  0.98-GB  0.99-RF  0.98-AB |
| **Fold 5** | 1-SVM  1-DT  1-KNN  0.94-GB  1-RF  0.94-AB | 1-SVM  1-DT  1-KNN  0.95-GB  1-RF  0.95-AB | 1-SVM  1-DT  1-KNN  0.95-GB  1-RF  0.95-AB | 1-SVM  1-DT  1-KNN  0.95-GB  1-RF  0.95-AB | 1-SVM  1-DT  1-KNN  1-GB  1-RF  1-AB |
| **Fold 6** | 0.96-SVM  0.96-DT  0.96-KNN  0.96-GB  0.96-RF  0.96-AB | 0.95-SVM  0.95-DT  0.95-KNN  0.95-GB  0.95-RF  0.95-AB | 0.95-SVM  0.95-DT  0.95-KNN  0.95-GB  0.95-RF  0.95-AB | 0.95-SVM  0.95-DT  0.95-KNN  0.95-GB  0.95-RF  0.95-AB | 1-SVM  0.97-DT  0.97-KNN  0.96GB  0.99-RF  0.96-AB |
| **Fold 7** | 1-SVM  0.96-DT  0.96-KNN  1-GB  1-RF  1-AB | 1-SVM  0.95-DT  0.95-KNN  1-GB  1-RF  1-AB | 1-SVM  0.95-DT  0.95-KNN  1-GB  1-RF  1-AB | 1-SVM  0.95-DT  0.95-KNN  1-GB  1-RF  1-AB | 1-SVM  0.97-DT  0.99-KNN  1-GB  1-RF  1-AB |
| **Fold 8** | 1-SVM  0.95-DT  1-KNN  1-GB  1-RF  1-AB | 1-SVM  0.95-DT  1-KNN  1-GB  1-RF  1-AB | 1-SVM  0.95-DT  1-KNN  1-GB  1-RF  1-AB | 1-SVM  0.95-DT  1-KNN  1-GB  1-RF  1-AB | 1-SVM  0.96-DT  1-KNN  1-GB  1-RF  1-AB |

Metrics used to calculate the performance of the model are listed above on the top of the table. Ideal value for all these values should be 1. For, SVM(support vector machines) the parameter values after grid search performed less as the value of metrics is less. DT,GB and RF had similar values in both fields. In contrast KNN and AB performed well with best fit parameters.

Coming to K-fold, fold 5 has the best performance amongst all for SVM,DT, KNN and RF. Also, all the algorithms worked well in 8th fold apart from DT.

Correlation Matrix:

The correlation matrix for SVM is shown below. The highlighted boxes denote true positives and higher the number higher the performance of the algorithm. The SVM had lesser performance as the values before grid search are higher than the values after grid search. For DT and RF the values are pretty much same before and after. In KNN and AB the values are better after grid search, which means this performed well after best parameters got fit. For K-fold method, SVM fold did the best. In DT fold 1 and 2 performed the best. Similarly other values can be compared.

Chart

Description automatically generated

ROC Curve for One class Vs rest for split data:

Chart, line chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

The position for ROC is at top left corner and best AOC value is between 0.9-1. Almost all our algorithms are performing similarly. 1 Value denotes that the classifier predicts 100% right positives and negatives. Class 0 data is best predicted by SVM, KNN and RF as their values are 1. Class 2 is best predicted by all and class 1 is best precited by SVM, KNN and RF.

When graphs are plotted similarly to K fold data, class 0 was best predicted by SVM, KNN and RF. Class 2 by SVM and class 3 by SVM, KNN, RF and AB.

Conclusion:

The best scores of accuracy when compared between all models is given by SVM before grid search. After grid search by KNN.

Precession score, recall and F1 were also best given by SVM before grid search and KNN after grid search.

AUC was always 1 for SVM for all classes including grid and k-split data. KNN has almost all 1’s apart from class 1 in k-fold data.

Hence it can be concluded that for the given dataset SVM is the best algorithm to work with and KNN being the second best.