Wine classification

Jovana Jovanovic; sw26/2015

1. Motivation

Motivation for this project comes from the idea of making software that will be used in markets and restaurants with goals of helping people to select high-quality wines.

2. Research questions

In this project I want to compare some algorithms for classification and choose the best for the given problem. Dataset[https://archive.ics.uci.edu/ml/datasets/Wine+Quality] is found on the internet. It contains 12 characteristics of wine, which are:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
12 - quality (score between 0 and 10)

3. Related work

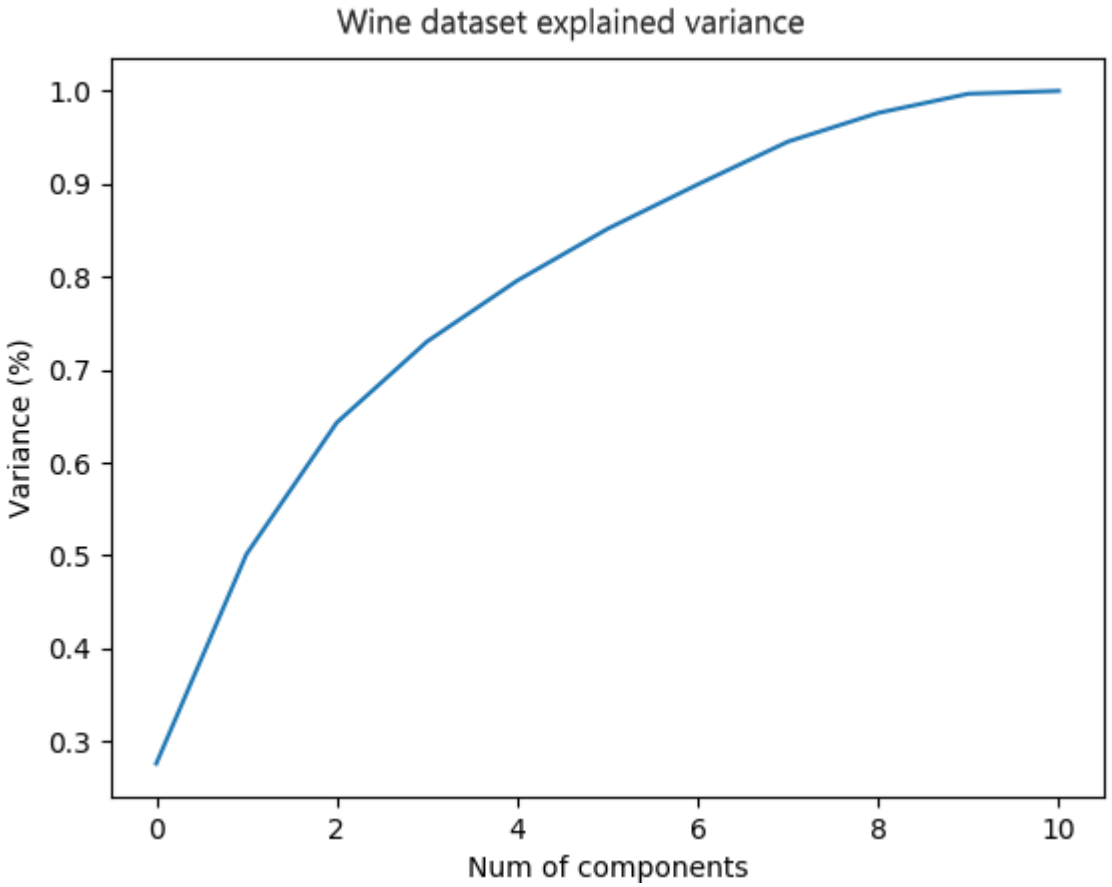
I found some projects on this theme, but they classified only red or only white wines. The y used Support vector machine, Decision Tree, Random Forest, KNeighbors, GaussianNB and XG Boost and they calculated accurancy measure. And they got the best result for Random Forest al gorithm. [https://www.kaggle.com/mathvv/prediction-of-red-wine-quality-93-215]

4. Methodology

In this project for data processing I used Standar Scaler for collapse data and PCA algorithm to reduce dimensionality. For classification I used Support vector machine (svm)[https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python], AdaBoost[https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c], Random Forest[https://towardsdatascience.com/understanding-random-forest-58381e0602d2], Bagging algorithm with Decision tree and Extra tree classifier, Extra tree classifier, KNN[https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn]and Voting. Also GridSearch and cross validation[https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/] are used for determination of hyperparameters.

5. Discussion

Data set contains red and white wines. I split the set to training and test set in proportions 80:20. Training set is used for determination of hyperparameters because I had used cross validation[https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f] and had split training set on n sets where one is used for test, n-1 for training. Number of components for PCA algorithm[https://towardsdatascience.com/an-approach-to-choosing-the-number-of-components-in-a-principal-component-analysis-pca-3b9f3d6e73fe] is determined on the basis of data variance.



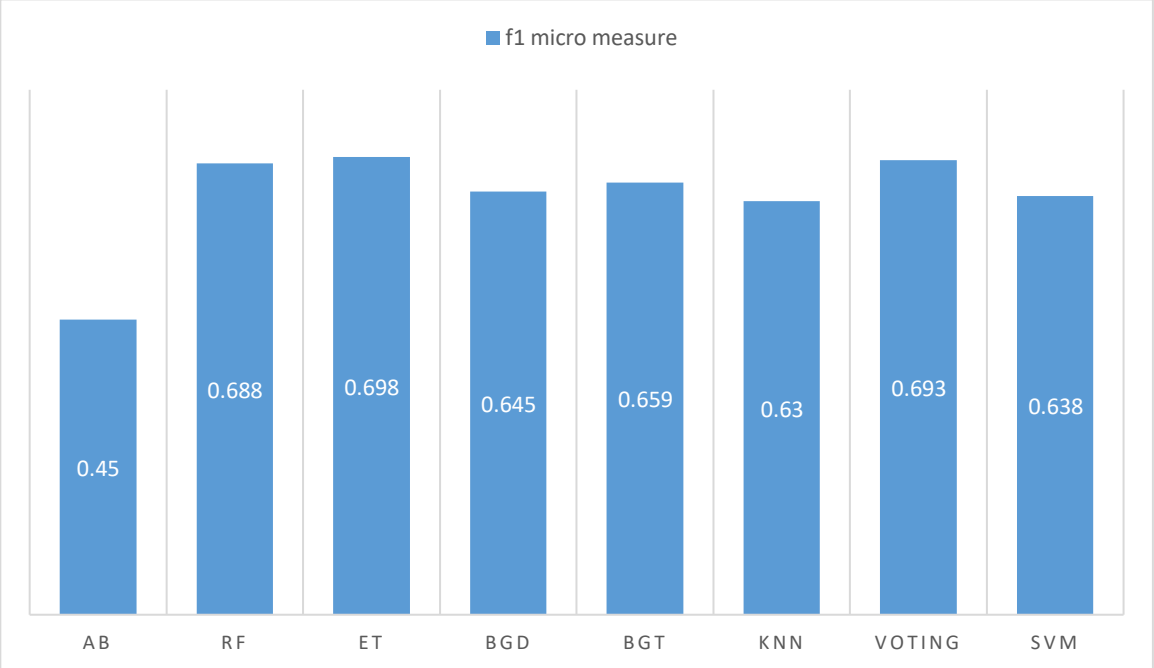
From the graphic above we can conclude that dimensionality can be reduced to 9.
Hyperparameters for alghoritms:

1. AdaBoost [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html]
    a. learning_rate=0.01
    b. n_estimators=500
2. Random Forest [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html]
    a. bootstrap=True
    b. criterion='gini'
    c. n_estimators=500
3. Bagging [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.htmlhttps://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html]
    a. With Decision tree
    b. With Extra tree classifier

4. Extra tree classifier[https://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html]
    a. bootstrap=True
    b. criterion= 'entropy'
    c. n_estimators=500

5. KNN[https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html]
    a. n_neighbors=1
6. Voting[https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html]
    a. All of the above mentioned algorithms are used in voting
7. SVM[https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html]
    a. gama =1
    b. C =1
    c. Kernel = 'rbf'

For each algorithm I calculated f1 micro measure and compared the result obtained.



Finally, Random Forest, Extra tree classifier and Voting showed the best results.

6. References

1. https://archive.ics.uci.edu/ml/datasets/Wine+Quality
2. https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f
3. https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c
4. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
5. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python
6. https://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html
7. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

8. https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/
9. https://towardsdatascience.com/an-approach-to-choosing-the-number-of-components-in-a-principal-component-analysis-pca-3b9f3d6e73fe
10. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
11. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
12. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
13. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
14. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html
15. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
16. https://www.kaggle.com/mathvv/prediction-of-red-wine-quality-93-215