**WINE QUALITY TESTING**

**INTRODUCTION**

Wine is an alcoholic beverage made from fermented grapes. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide, and heat. It is a pleasant tasting alcoholic beverage, loved celebrated. It will definitely be interesting to analyze the physicochemical attributes of wine and understand their relationships and significance with wine quality and types classifications.

The dataset is related to red and white variants of the "Vinho Verde" wine. Vinho verde is a unique product from the Minho (northwest) region of Portugal. Medium in alcohol, is it particularly appreciated due to its freshness (especially in the summer). This dataset is publicly available for research purposes. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

**ATTRIBUTE INFORMATION**

Input variables (based on physicochemical tests):

1.fixed acidity: most acids involved with wine are fixed or nonvolatile (do not evaporate readily).

2.volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.

3.citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines.

4.residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.

5.chlorides: the amount of salt in the wine.

6.free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

7.total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.

8.density: the density of water is close to that of water depending on the percent alcohol and sugar content.

9.pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic). Most wines are between 3-4 on the pH scale.

10.sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.

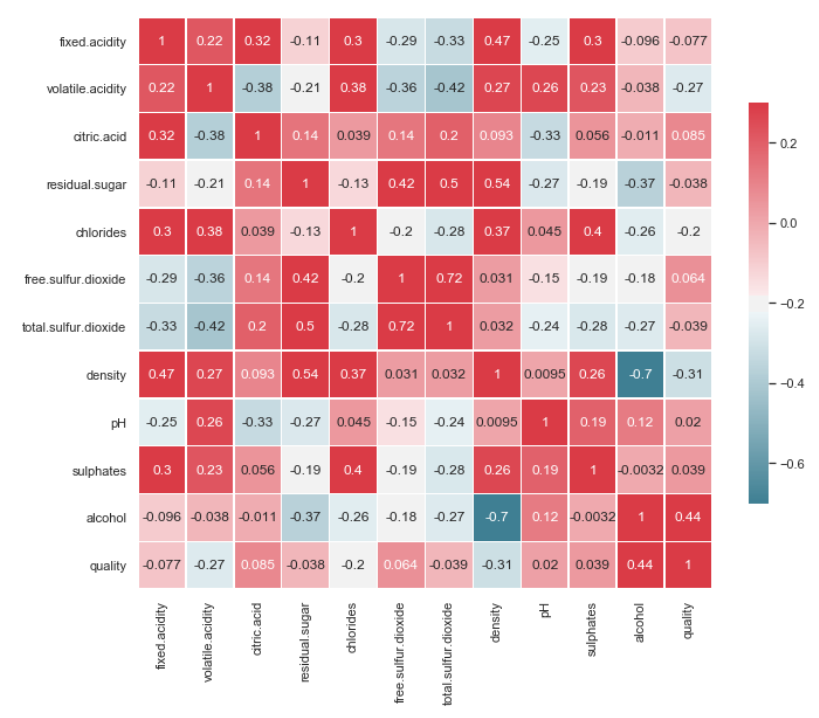
11.alcohol: the percent alcohol content of the wine.

12.quality: output variable (based on sensory data, score between 0 and 10)

**DATASET DESCRIPTION**

* The classes are ordered and not balanced.
* Outlier can be identified by using boxplots and scatterplot.
* Finding out what are the most important input variables for the quality is necessary. So we used random forest is used to identify the importance of the input variables.

**CORRELATION MATRIX**



While most of the correlations are weak, we can see that:

* Total and free sulfur dioxide has the highest correlation. In fact, the second is a parcel of the first and it represents collinearity that can be a problem for linear classifiers, especially when the target predicts the color, in that case we need drop the free sulfur dioxide but in our case it was not a problem.
* The residual sugar has a correlation of 0.5 with total sulfur dioxide. This indicates that wine with more residual sugar and more sulfur dioxide have better quality.
* Density has a relatively positive correlation to residual sugar and relatively negative correlation to alcohol.

**COMPARING THE DEPENDENT VARIABLE WITH THE INDEPENDENT VARIABLES**

A picture containing writing implement, stationary, indoor, pencil

Description automatically generated

**METHODOLOGIES**

**NORMALIZATION AND DATA PARTITIONING**

* The data is first split into X and y data frames where X consists of all the independent variables and Y contains the dependent variable which is quality.
* The data then has been normalized by using the standard scaler and Min-Max scaler.
* The shape of the standardized X data frame is of shape 6495 rows and 12 columns. The shape of the standardized Y is 6495 rows.
* This standardized data is then split into train and the test data where the model is built on the train and test data is used to predict the quality.
* The train data consists of 4546 observations and test data consists of 1949 observations.

**APPLICATION OF THE MACHINE LEARNING MODELS**

**Support vector classification (SVC):**

In machine learning support - vector machines are supervised learning models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. The support-vector clustering[[2]](https://en.wikipedia.org/wiki/Support-vector_machine#cite_note-HavaSiegelmann-2) algorithm applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications.

Advantages:

* It works relatively well when there is clear margin of separation between classes.
* It is more effective in high dimensional spaces.
* It is effective in cases where number of dimensions is greater than the number of samples.
* It is relatively memory efficient.

# Disadvantages:

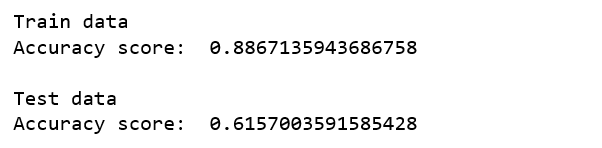
# SVM algorithm is not suitable for large data sets.

# SVM does not perform very well, when the data set has more noise i.e. target classes are overlapping.

# In cases where number of features for each data point exceeds the number of training data sample , the SVM will underperform.

# As the support vector classifier works by putting data points, above and below the classifying hyper plane there is no probabilistic explanation for the classification.

We used grid search in order to determine the optimal values for a given model. In this model we see that the best kernel is RBF and cost of misclassification is 1 and gamma is 1.



After performing the algorithm with the best parameters for the test dataset the accuracy scores for train data and test data are 0.88 and 0.61 respectively.

**Decision Tree Classifier:**

#### Decision Tree are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

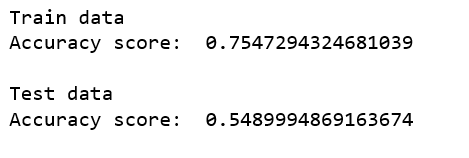
Advantages:

* Simple to understand and to interpret. Trees can be visualized.
* Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
* The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
* Able to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable. See algorithms for more information.
* Able to handle multi-output problems.
* Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
* Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
* Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

Disadvantages:

* Decision-tree learners can create over-complex trees that do not generalise the data well. This is called overfitting. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
* Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
* The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.
* There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.
* Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

We used grid search in order to determine the optimal values for the given model. We applied grid search to obtain the best parameters. The best parameters obtained is max\_depth =10



After performing the algorithm with the best parameters for the test dataset the accuracy scores for train data and test data are 0.755 and 0.549 respectively.

**RANDOM FOREST CLASSIFIER**

A Random Forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement if bootstrap=True (default).

In the random forest model, each base learner is a decision tree model trained on a bootstrap sample of the training data. Besides this, when we want to split a decision node in the tree, the split is chosen from a random subset of all the features instead of taking the best split from all the features. Due to the introduction of this randomness, bias increases and when we average the result from all the trees in the forest, the overall variance decreases, giving us a robust ensemble model, which generalizes well. We will be using the Random Forest Classifier from scikit-learn, which averages the probabilistic prediction from all the trees in the forest for the final prediction instead of taking the actual prediction votes and then averaging it.

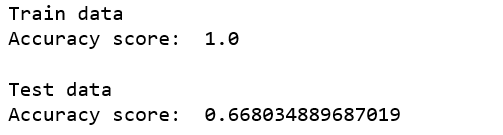
Advantages:

* Random forest can solve both type of problems that is classification and regression and does a decent estimation at both fronts.
* One of benefits of Random Forest which exists me most is, the power of handle large data sets with higher dimensionality.
* It can handle thousands of input variables and identity most significant variables, so it is considered as one of the dimensionality reduction method. Further, the model outputs importance of variable, which can be a very handy feature.
* It has an effective method for estimating missing data and maintains accuracy when large proportion of the data are missing.
* It has methods for balancing errors in data sets where classes are imbalanced. The capability of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
* Random forest involves sampling of the input data with replacement called as bootstrap sampling. Here one third of data is not used for training and can be used to testing.
* These are called the OUT OF BAG samples. Error estimated on these output bag samples is known as out of bag error. Study of error estimates by out of bag, gives evidence to show that the out of bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out of bag error estimate removes the need for a set aside test set.

Disadvantages :

* It surely does a good job at classification but not as for regression problem as it does not give precise continuous nature prediction. In case of regression, it doesn't predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy.
* Random forest can feel like a black box approach for a statistical modeler we have very little control on what the model does. You can at best try different parameters and random seeds.

We used grid search in order to determine the optimal values for the given model. We applied grid search to obtain the best parameters. The best parameters obtained is n\_estimators =50 and max\_features = 12.



After performing the algorithm with the best parameters for the test dataset the accuracy scores for train data and test data are 1.0 and 0.668 respectively.

**FEATURE IMPORTANCE**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

Insights:

1. The alcohol value ranges between 9 to 12 where the wine with quality scores of 5 and 6 have less alcohol content compared to higher quality wines.
2. Bad quality wines have bigger volatile acidity distribution.