DATA ANALYTICS REPORTRED WINE QUALITY TESTING

Akash Rao and Shreyas C Reddy

Abstract—**This dataset is related to red variants of the Portuguese "Vinho Verde" wine. Certification and quality assessment are crucial issues within the wine industry. At the present time, wine quality is majorly derived by the physicochemical and sensory tests. In this paper, we propose various data mining approaches to predict wine preferences that are based on popular analytical tests at the wine certification step. A large dataset is considered with red "Vinho Verde" samples from the Minho region of Portugal. Explanatory information is provided in terms of a sensitivity analysis, which measures the response changes when an input variable is changed through its domain. Multiple regression techniques were applied, under systematic and efficient procedure that performs simultaneous variable and model selection. The support vector machine achieved promising results, along with decision tree and random forest models. Such models are useful for explaining the effects of physicochemical tests on the sensory preferences. Moreover, it will support wine expert evaluations and gradually improve the production.**

I. INTRODUCTION

A wide range of the population are currently consuming wine. Portugal is one of the top promising country for the exports of its Vinho Verde wine produced in the northern region of the country with the percentage of export having increased by 40% swell percentage in a span of a decade. The industry is looking to invest in newer and more modern technology to produce and export the wine to keep up with the increasing demand. This requires that the certification and quality assessment processes by ever more scrutinized as the certification process ensures the genuinity of the product to make it safe for consumption and ensure the quality of the business. Quality evaluation is usually a part of the certification process and is used to upgrade the quality of the product.

Wine certification is assessed by the physicochemical and sensory tests. The tests are run often to differentiate the various wine products based on the density, alcohol percentage, pH values etc. whereas the sensory tests are still relied upon humans senses which makes the classification of wine a tedious and tiring process. The entwinement between the two methods of classification also make it a complex task to perform.

With the introduction of advanced technologies, it is easier to collect, store and retrieve the enormous datasets which hold valuable information of patterns and trends which help in easing the process of decision making and sky rocketing the chances of success.

Support Vector Machines, decision trees, random forest and KNN models have been gaining attention due to their learning capabilities and their high predictive and accuracy rate only help in establishing them as the forerunners in classification algorithms under Data Mining techniques.

One of the most important notes to follow while implementing these algorithms is that one should be vary that the performance of the said algorithms will depend on the variable chosen and the model selection as simple models may crumble while trying to map the data while complex ones may tend to overfit the provided data.

In the current paper, we're presenting a real world scenario where the testing preferences are modelled by various algorithms which run on the analytical data available only during the certification step. The current dataset is large with over 16000 samples and the quality of wine is modelled in such a way that the order of the grades is maintained. Model selection and variable selection are performed in parallel. The output of the analysis to be undergone will provide the impact of various models used for the classification of wine.

II. LITERATURE REVIEW

One of the simplest methods to provide a good classification of the data is SUPPORT VECTOR MACHINES. With the introduction of newer classification models like the RANDOM FOREST and XGBoost and BAYES Classification there are evermore options available to perform the tests.

The DECISION TREE, RANDOM FOREST models were the ones which provided the most accuracy followed by SVM, KNN and BAYES classifiers.

The parameters SVM was improved by providing sufficient options to the model which would eventually select the best permutation and combination of values and hence would have tremendously improved accuracy. For the random forest and decision tree algorithms, the number of trees and the branching, skewness and selection of nodes played a major part, while for the KNN classifier it was all about efficiently choosing the right value of 'k' with the BAYES classifier being the base model for comparison.

III. DATASET AND INITIAL INFORMATION

The undergoing study will consider vinho verde, a product from the northern province of Portugal. Medium in alcohol, it is well sought after due to the freshness it has in the warmer part of the calendar year. This particular wine accounts for almost $1/6^{th}$ of the total production in Portugal and over 70% of it is eventually exported to various consumers

Akash Rao is with Computer Science, PES University Bangalore, India SRN: PES1UG19CS041

Shreyas C Reddy is with PES University, Bangalore, India

around the globe. the data which will undergo the analysis was collected from may/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the main obsession of improving the quality and marketing of the wine. The data was recorded by a computerized system which automatically manages the process of wine sample testing from producer requests sent to the laboratory. Each entry denotes a given test and the final database was exported into a single sheet.

The dataset consists of 12 attributes of importance to the testing of the wine. These include Fixed Acidity (most acids involved with wine or fixed or non-volatile), volatile acidity (the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste), citric acid (found in small quantities, citric acid can add 'freshness' and flavour to wines), residual sugar (the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter), chlorides (the amount of salt in the wine), free sulphur dioxide (the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulphite ion), total sulphur dioxide (amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine), density (the density of water is close to that of water depending on the percent alcohol and sugar content), 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), sulphates (a wine additive which can contribute to sulphur dioxide gas (S02) levels, which acts as an antimicrobial).

The data provided in the dataset is highly imbalanced.

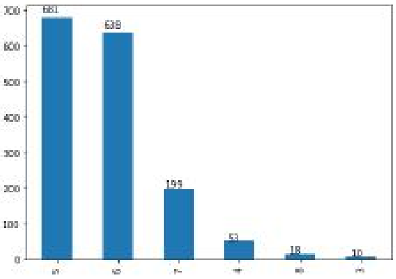


Fig. 1.

IV. Proposed Solution

The goal of the analysis is to be able to classify the quality of the wine based on the given physicochemical values in the dataset. The correlation of the various values of the dataset with respect to the quality of the wine can be seen in the chart displayed below

For achieving the said goal, we train various models with a part of the dataset (around 72.5%) and test it against the remaining test data (27.5%) of the dataset to obtain models with high accuracy, precision and recall. The various models that were trained have performed extremely well and have each been trained with over 80% accuracy.

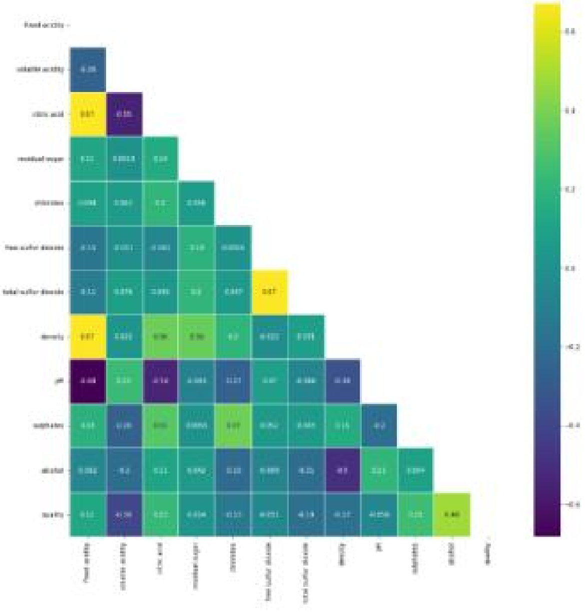


Fig. 2.

a. Pre Processing

The dataset was highly imbalanced as only around 5% of the entire data was involved in providing quality of certain values. Hence there would be a chance of having performed oversampling or under-sampling which would result in having redundant rows or duplicate rows causing issues by being skewed or removing important values. The box plots provide an excellent view of the outliers present in the dataset.

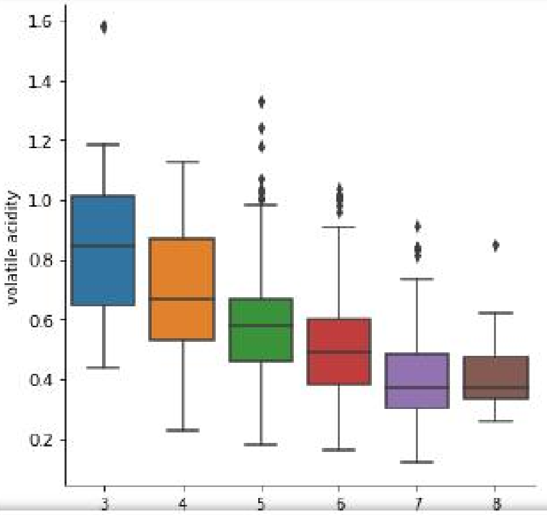


Fig. 3.

Volatile acids vs quality

Citric acid vs quality

The number of outliers present in the dataset can be seen below

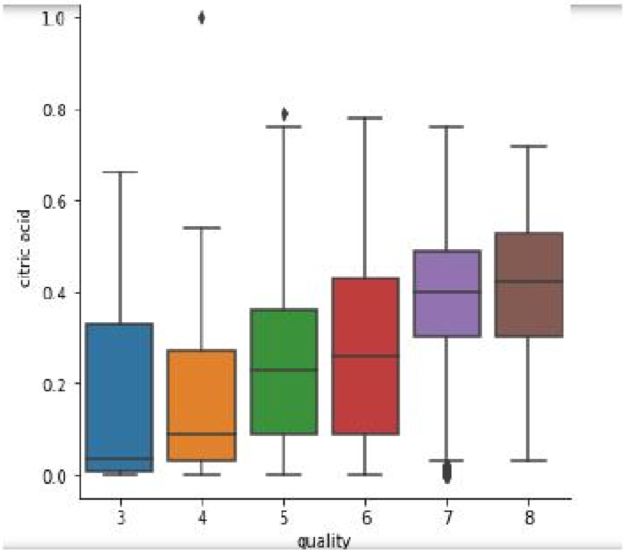
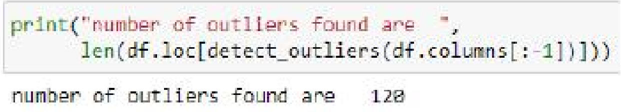


Fig. 4.

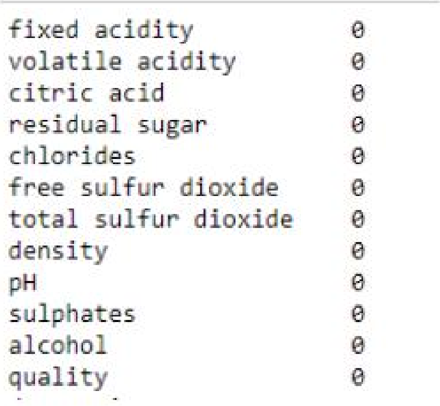


```
print("number of outliers found are  ",
      len(df.loc[detect_outliers(df.columns[:-1])]))
```

```
number of outliers found are   120
```

Fig. 5.

The dataset was also scanned through for identification of any missing values and none were found



```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

Fig. 6.

V. Building a Model

As new data input regarding the wine quality keeps getting added into the dataset, there is a need to constantly update the existing entries to the model and make sure that is improving as a learner or at least remains as efficient as it initially was. To facilitate this, we build multiple models for learning the dataset and select the model which provides us with the most accurate results.

1. Decision Trees:

Decision trees are supervised models which will classify the oncoming data into a class based on the learning of its previous experiences. It models data as a tree of hierarchical branches and each node divides the dataset till it reaches a leaf node which is prominently giving us the classification result of the data. The splitting of data happens based on the best attribute and they are chosen based on the gini index value or entropy values of the dataset, average information and information gain of the attribute in question. The best attribute is one which has maximum information gain or minimum entropy (chaos in the dataset). As these models are known to handle non linearity they are expected to perform better than the logistic regressors.

The model will choose the best possible parameter options (criteria = {gini, entropy}, features = {auto, sqrt, log2}) based on least error

The AUC of the decision tree model is shown below:

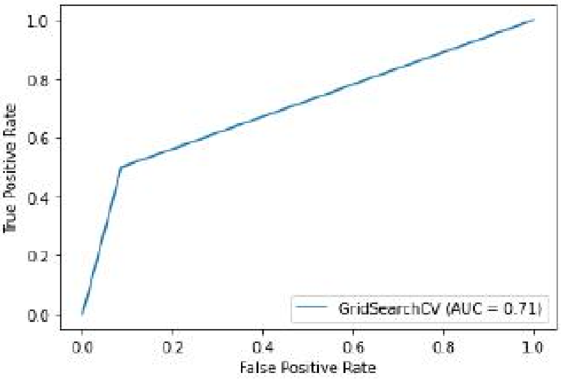


Fig. 7.

2. KNN:

This is another of the supervised algorithm which is also a lazy learner (the algorithm does not train when the dataset is provided but rather does so when it is to test a given value). KNN algorithm works on the principle of nearest neighbours (datapoints) to the point in question, the distance is calculated by minkowski / Euclidean distances and points closer to the given point are assigned the same class while points farther away are given a different class. The method of selection of 'k' is done by elbow method where the value of chosen 'k' provides for the least error and the chosen value is not to high that it might cause overfitting nor too low to be swayed by the noise present in the dataset.

Another alternate version of the algorithm is called the weighted KNN where the points are assigned classes with respect to the distance with the given point and also the weight assigned to the point. This process is much more efficient as compared to vanilla KNN.

In the model designed, the model with the best parameter constants is chosen after a thorough comparison with the different values given parameter list and the model will be ready to take values from the dataset to start classifying.

The AUC of the model is shown below:

3. Random forest:

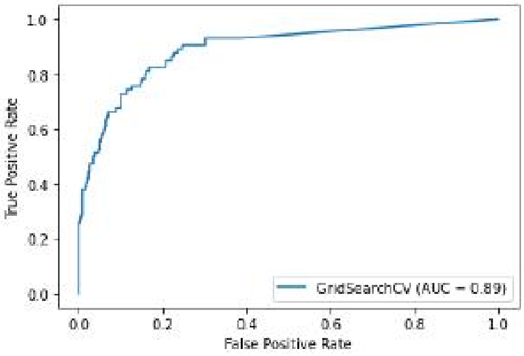


Fig. 8.

Also known as random decision trees are an ensemble learning method for classification, regression and other tasks that operate by construction of multitude of decision trees at training time. For classification tasks, the output of random forest is the class selected by most trees and for regression tasks, the mean prediction of the individual trees is considered. These models often have a tendency to overfit the data, they do have an advantage of being able to outperform their counterparts (decision trees).

The given model will select the best parameters possible to enable it to provide us with the most accuracy and precision possible. The parameters being (criteria = {gini, entropy}, features = {auto, sqrt, log2}, number of estimators = {defaulted to 100})

The AUC graph for the model is shown below:

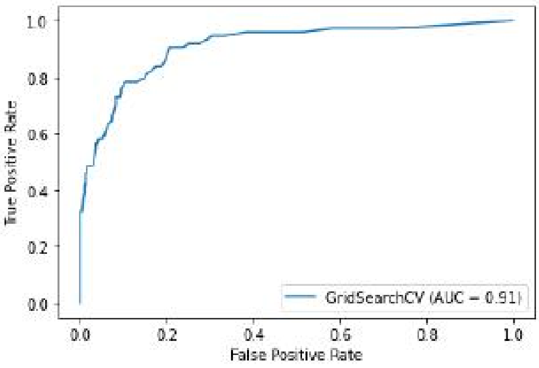


Fig. 9.

4. Support Vector Machines:

These models are supervised learning models that analyse data for classification and regression. Considered to be one of the most advanced and well off models, it is a non-probabilistic binary linear classifier. The model is also known to be able to classify non-linear data with the help of kernel (mathematical functions that help transform the non-linear data to various formats to help ease the classification process). A support vector machine is known to construct a hyperplane or set of hyperplanes which are then used for classification or outlier detection. A good separation is one which has maximum distance to the nearest training data point of any class (margin) larger the margin, lower the error.

The AUC for the model is shown below:

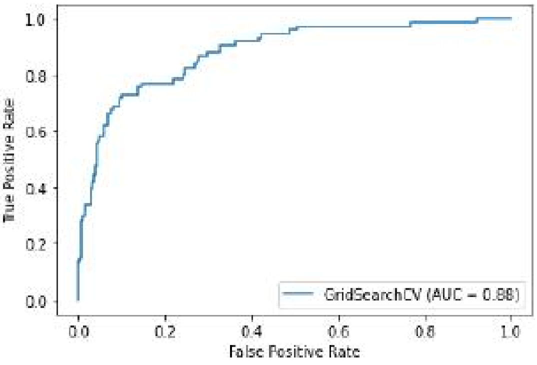


Fig. 10.

5. Naïve Bayes:

This is a part of the probabilistic classifier family, when bayes' theorem is applied with strong independence assumptions between features. Bayes classifier is a highly scalable while requiring a number of parameters with regards to number of variables, the maximum likelihood training is done by evaluation of a closed form expression taking linear time. It is a simple technique for construction of classifiers, models that assign class labels to problem instances.

Posterior = (prior x likelihood)/ evidence

The AUC of the model is shown below:

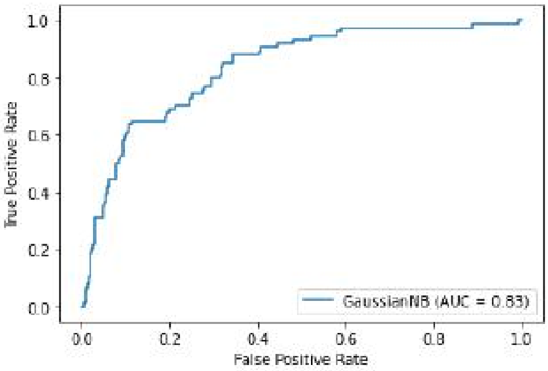


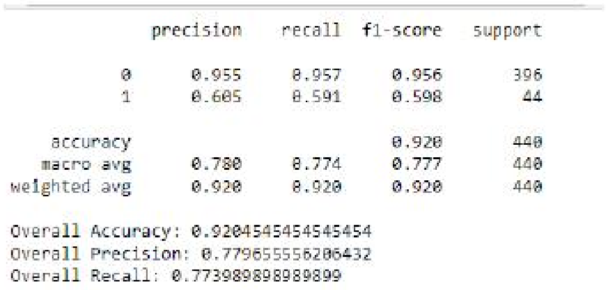
Fig. 11.

c. EXPERIMENTAL RESULTS

The dataset was savaged off of Kaggle.

Model evaluation was done using AUC graphs, accuracy, precision, recall and f1 scores. The higher the accuracy and precision, the better the model performs in distinguishing the quality of the wine for the provided attributes.

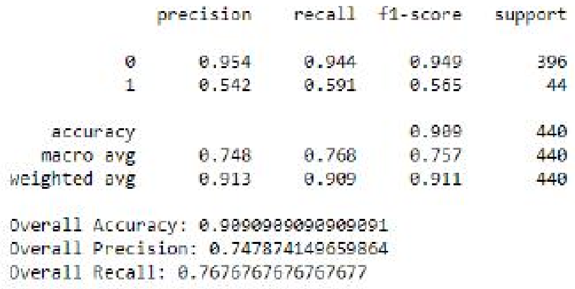
The dataset is first run with KNN model and the corresponding scores of accuracy, precision are noted down, and then the model is run with the remaining models of SVM, Bayes classifier, Decision Tree, XGBoost, Random Forest and their corresponding values are stored for comparison with each other.



```
              precision    recall  f1-score   support

           0      0.955     0.957     0.956       396
           1      0.605     0.591     0.598        44

    accuracy                          0.920       440
   macro avg      0.780     0.774     0.777       440
weighted avg      0.920     0.920     0.920       440

Overall Accuracy: 0.9204545454545454
Overall Precision: 0.779655556206432
Overall Recall: 0.773989898989899
```

Fig. 12.

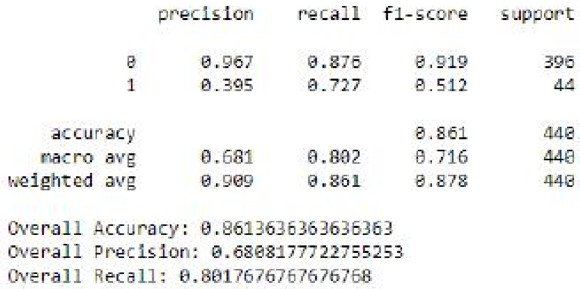
KNN



```
              precision    recall  f1-score   support

           0      0.954     0.944     0.949       396
           1      0.542     0.591     0.565        44

    accuracy                          0.909       440
   macro avg      0.748     0.768     0.757       440
weighted avg      0.913     0.909     0.911       440

Overall Accuracy: 0.9090909090909091
Overall Precision: 0.747874149659864
Overall Recall: 0.7676767676767677
```

Fig. 13.

Support Vector Machines



```
              precision    recall  f1-score   support

           0      0.967     0.876     0.919       396
           1      0.395     0.727     0.512        44

    accuracy                          0.861       440
   macro avg      0.681     0.802     0.716       440
weighted avg      0.909     0.861     0.878       440

Overall Accuracy: 0.8613636363636363
Overall Precision: 0.6808177722755253
Overall Recall: 0.8017676767676768
```

Fig. 14.

Bayes Classifier

Decision Tree

Random Forest

XGBoost

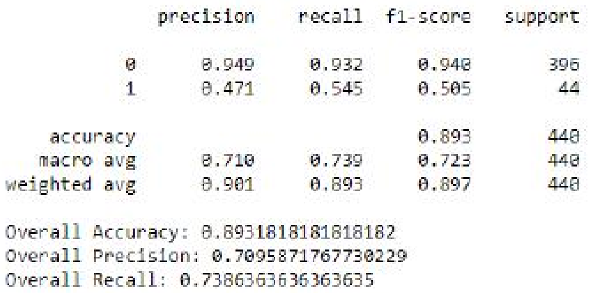
Accuracy of different classifiers

Precision of the different classifiers

Recall of the various models

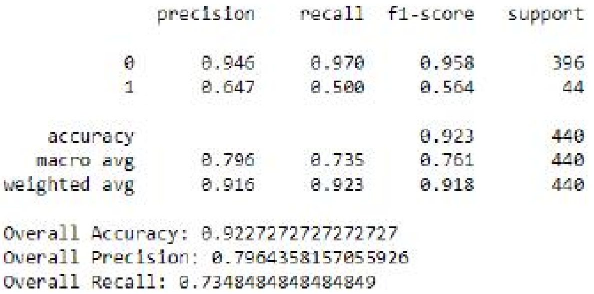
VI. CONCLUSION:

The final Goal of the analysis is to reduce the time consumed in classifying the given data on wine and also to accurately find out the quality of the product at the same time.



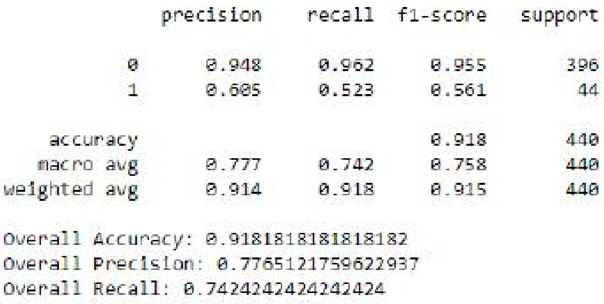
```
              precision    recall  f1-score   support

           0      0.949     0.932     0.940       396
           1      0.471     0.545     0.505        44

    accuracy                          0.893       440
   macro avg      0.710     0.739     0.723       440
weighted avg      0.901     0.893     0.897       440

Overall Accuracy: 0.8931818181818182
Overall Precision: 0.7095871767730229
Overall Recall: 0.7386363636363635
```

Fig. 15.



```
              precision    recall  f1-score   support

           0      0.946     0.970     0.958       396
           1      0.647     0.500     0.564        44

    accuracy                          0.923       440
   macro avg      0.796     0.735     0.761       440
weighted avg      0.916     0.923     0.918       440

Overall Accuracy: 0.9227272727272727
Overall Precision: 0.7964358157055926
Overall Recall: 0.7348484848484849
```

Fig. 16.



```
              precision    recall  f1-score   support

           0      0.948     0.962     0.955       396
           1      0.605     0.523     0.561        44

    accuracy                          0.918       440
   macro avg      0.777     0.742     0.758       440
weighted avg      0.914     0.918     0.915       440

Overall Accuracy: 0.9181818181818182
Overall Precision: 0.7765121759622937
Overall Recall: 0.7424242424242424
```

Fig. 17.

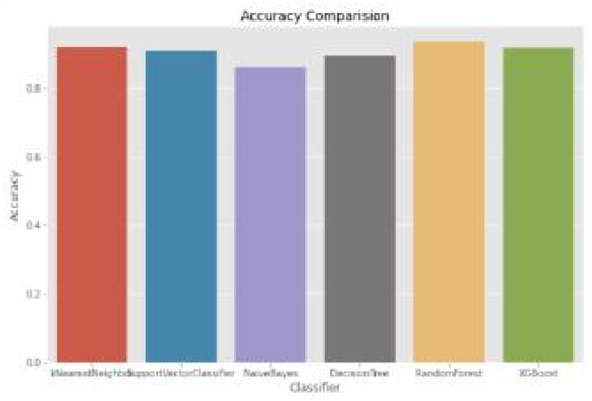


Fig. 18.

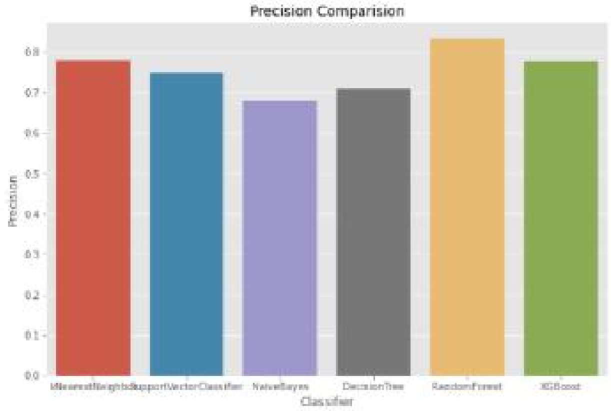


Fig. 19.

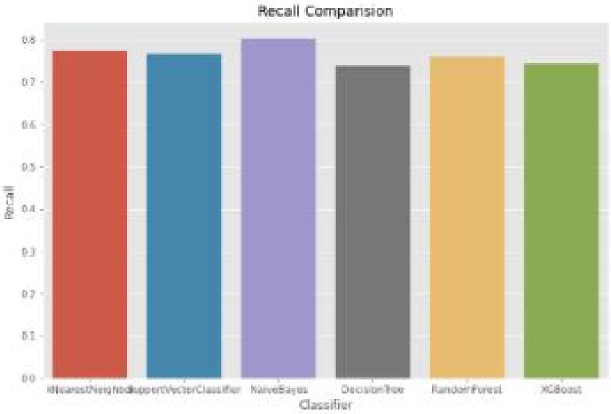


Fig. 20.

Due to the increase in interest in wine, companies are investing in modern technologies to improve their production and selling processes. Quality certification is a critical step for both the processes and is currently heavily depended on wine tasting by human experts. This analysis aims at having least human intervention to determine the preference of wine at the certification step. The approach preserves the order of classes while allowing the evaluation of various distinct accuracies, with an acceptable tolerance on error.

With the improvement in the technology of data mining, it is easier to retrieve information from the raw data that is available.

Encouraging results were achieved with the various models all having an accuracy of over 80% while it be noted that the dataset contains 6 to 7 various classes and that these accuracies are much better than the ones expected by a random classifier.

The result of this testing is relevant to the wine science domain with helping in the understanding of how physicochemical characterization affects the end quality of the product. The proposed data driven approach is based on objective tests and thus can be integrated into a decision support system, aiding in speed, quality and repeatability of the process.

A. Contribution:

The analysis started off with the prior knowledge of data cleaning and pre-processing along with model building, but that alone would not provide us with the required accuracy and precision to conclude and set up for a model. Hence the two of us researched about the different possibilities of combinations of models with the parameters and eventually agreed to settle with the models chosen in the analysis made.

B. References:

1) A. Smola and B. Scholkopf. A tutorial on support vector regression. Statistics and Computing, 14:199–222, 2004.
2) C. Blake and C. Merz. UCI Repository of Machine Learning Databases, 1998.
3) V. Cherkassy and Y. Ma. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. Neural Networks, 17(1):113–126, 2004.
4) D. Smith and R. Margolskee. Making sense of taste. Scientific American, 284:26– 33, 2001.
5) S. Ebeler. Flavor Chemistry - Thirty Years of Progress, chapter Linking flavour chemistry to sensory analysis of wine, pages 409–422. Kluwer Academic Publishers, 1999.
6) R. Kewley, M. Embrechts, and C. Breneman. Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks. IEEE Transactions on Neural Networks, 11(3):668–679, May 2000.