**Target attributes analysis of white wines**

**Methods used:**

* Naïve Bayes
* Generalised Linear model
* Decision Tree
* Random Forest
* Gradient Boosted Trees

1. Target variable – level of preservatives

In this particular exercise we are doing analysis to figure out how changing the average of one of measurement affects another.

This attributes has been created for this purposes – Level of preservatives in white wine. It is categorical attribute.

Rule : =IF(AND(J44<=0.45,E44<=0.045),"Low",IF(AND(J44<=0.6,E44<=0.06),"Medium","High"))

There is 3 kind of output Low, Medium and High. Ordinal Polynomial attributes.

The aim of this analysis is to check what will happen with one variable if we will change or delete the other. In that case we have deleted chlorides.

Training set was extracted at 70% of dataset and 30% was a test set.

Training:

A screenshot of a cell phone

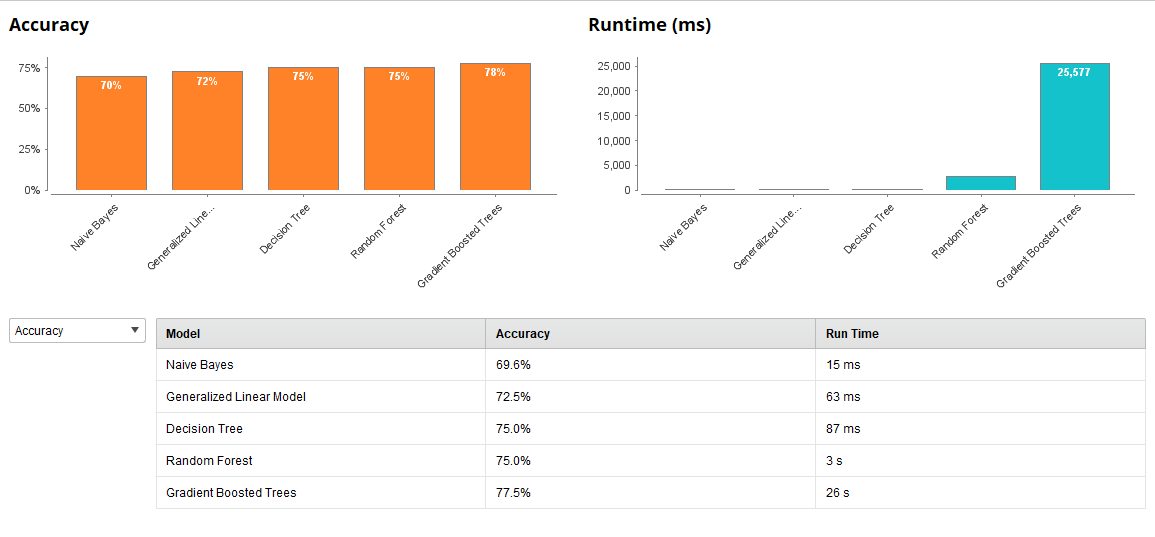
Description generated with very high confidence

Summary of model outputs.

In test set we have delete chloride attribute and then try to predict the outcome of target variables based only on other attributes.

Cross – validation:

Test:

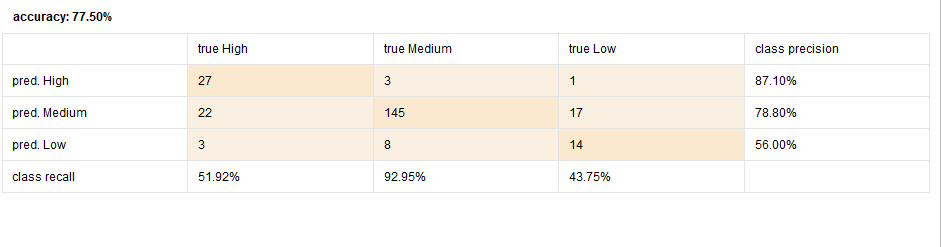


Summary of models outputs.

As we can see the Gradient Boosted Trees method had the highest accuracy at 77.5%. But it was the slow as it takes 26 seconds compared to others.

That’s mean that even without chloride attribute the other attributes of white wines can give us information about level of preservatives at 77.5% chances that we are correct.

Training set had higher score of 89%, but still model is quite good.



Let’s see the outcome of Gradient Boosted Trees method.

We can see that model had no problem with medium level of preservatives but had problem with the High and Low level.

1. Target attribute Sugar/PH ratio.

This ratio was picked not without discussion and further research. It is proven that this ratio should be the lowest to make sure that wine will be tasty.

Are we able to prove it with our sample of wines?

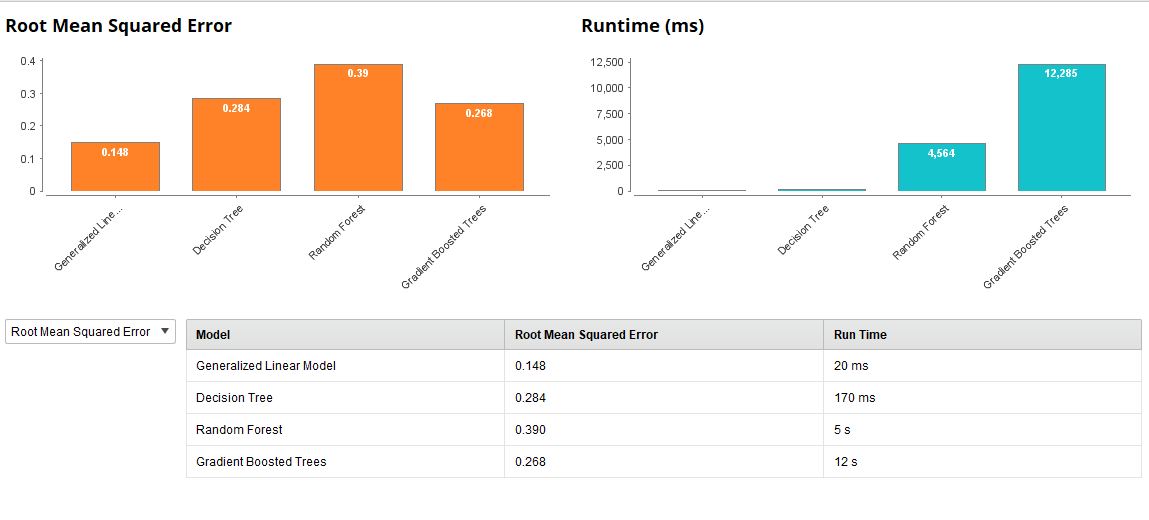
Methodology used as in previous example: training 70%, test 30% and we have deleted PH attribute.

A screenshot of a social media post

Description generated with very high confidenceTraining:

Summary of model outputs with given attribute. – root mean squared error

Cross -validation:

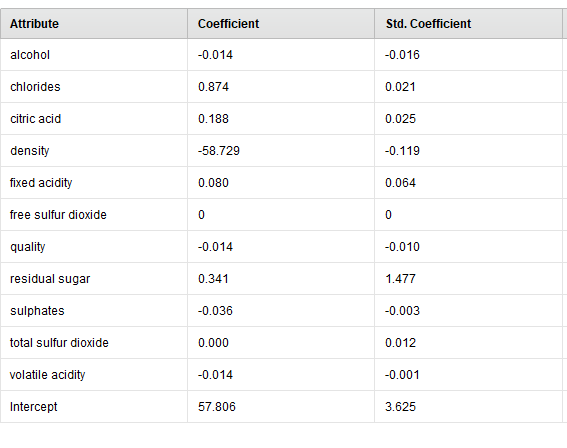
Test:

Summary of model outputs with deleted attribute – Ph.

In this case the best was simple Generalised Linear Model with the smallest error.

Error in both training 0.043 and test 0.268 is quite small. We can assume that model is quite good.

Output below:



That’s is very interesting finding. We can see here that even without given PH of wine we are able to predict sugar/Ph ratio in almost 90% accuracy.

The biggest impact on that ratio will have density -58.72. But few others are also very important like chlorides or residual sugar.

1. Target attributes - level of alcohol.

That is a very interesting one. Are we able to predict level of alcohol without given alcohol content?

Level of alcohol was set up under rule:

=IF(K2<=9,"Low",IF(K2<=11,"Medium","High"))

There is 3 kind of output Low, Medium and High. It is Ordinal Polynomial attributes.

Methodology used as in previous example: training 70%, test 30% and we have deleted alcohol content attribute.

Training:

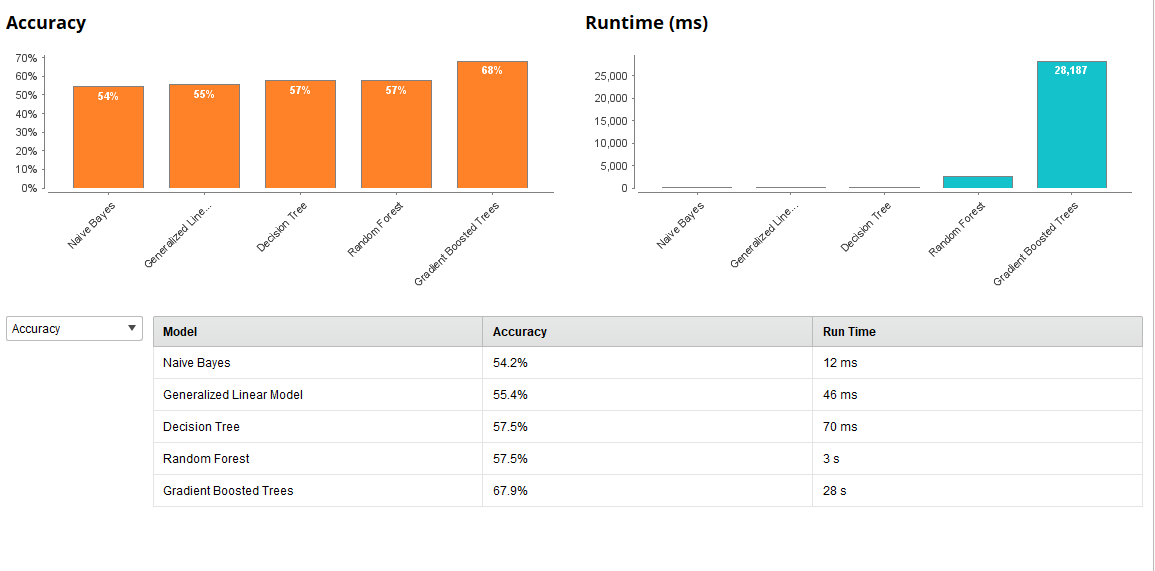
A screenshot of a cell phone

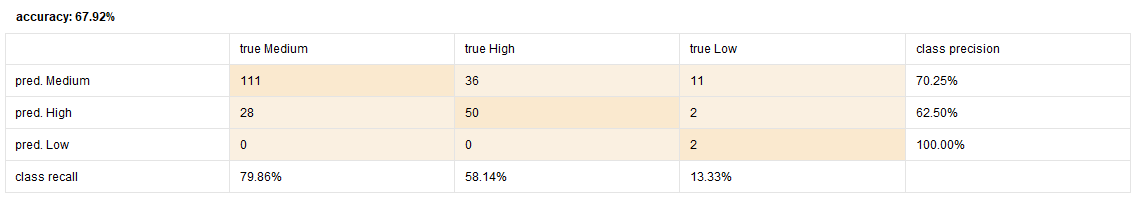
Description generated with very high confidenceSummary of model outputs with given attribute.

Test:

Cross-validation:

Summary of model outputs with deleted attribute – alcohol content.



Only Gradient Boosted Trees could give us 68% of accuracy in both training and test set. The conclusion is that even without knowing the alcohol content we still can predict alcohol level in wines when other attributes are known in 68%.

1. Target variable – Balanced wine

The aim was classify wines under label balanced and unbalanced. So that is binominal attribute.

Classification rule:

If total sulfur dioxide <=200, and sugar/ph ratio<=3, and level of alcohol = medium

The unknown here is free and total sulfur dioxide.

Is machine able to correctly classify wines with given attributes but unknown total sulfur dioxide?

Methodology used as in previous example: training 70%, test 30% and for test we have

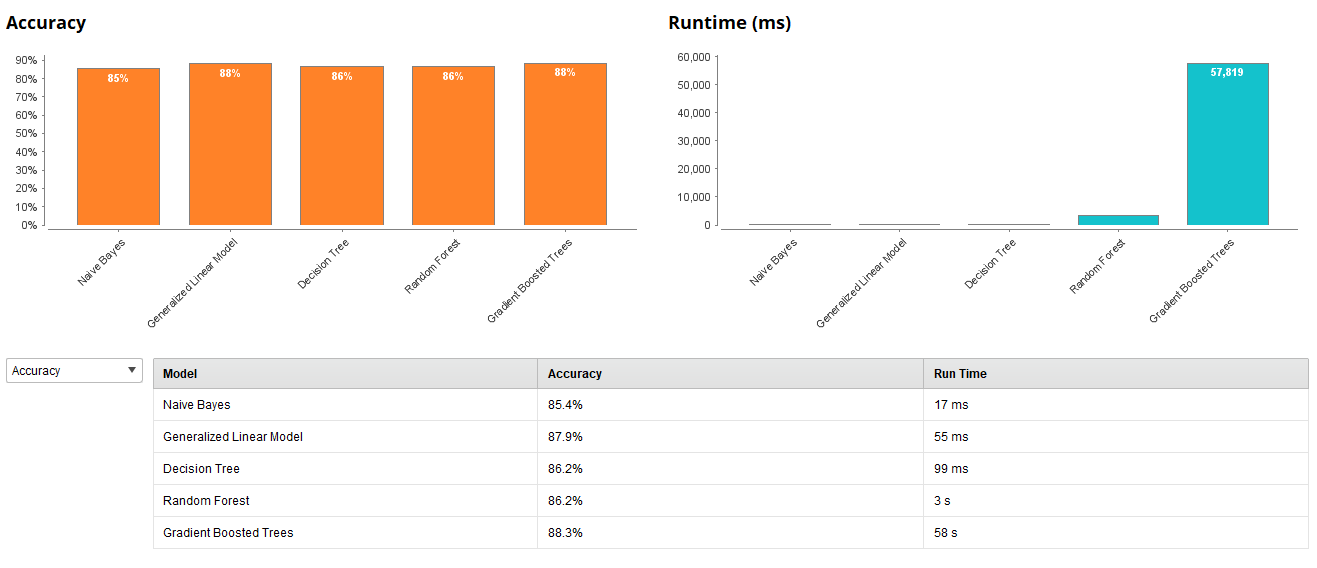
A screenshot of a cell phone

Description generated with very high confidenceTraining:

Rapidminer – accuracy of training model – balanced wine

Test

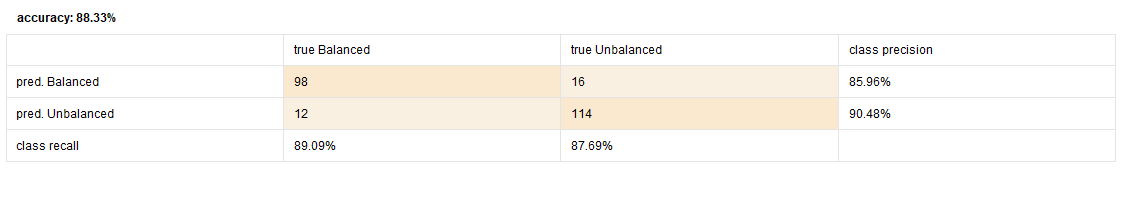
Cross – validation:



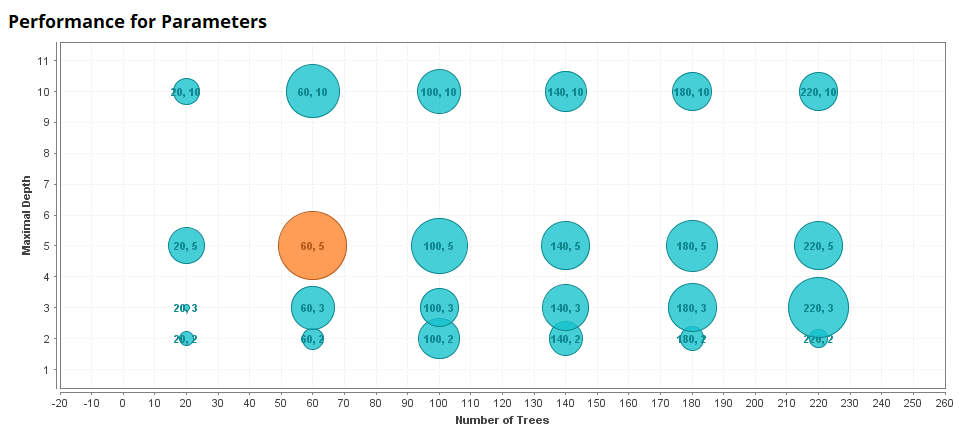
Rapidminer – Accuracy of test – balances wine

The highest accuracy of models we can see in Gradient Boosted Tree and Generalised Linear model in test set– 88% and 98% in training set. It is quite good classification model.

Gradient Boosted tree:

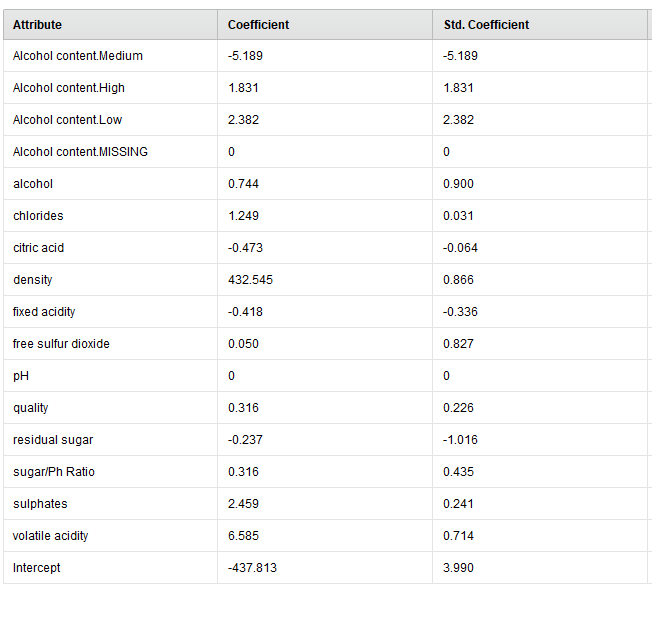


Rapidminer - Gradient Boosted tree accuracy



Model also shows us that the best result was achieved with 60 trees and 5 level of depth.

Generalised Linear model:



Rapidminer - Generalised Linear model outcome – balanced wine

Results of Generalised Linear models are much clearer for us to understand but it’s not mean it is better even if accuracy is high. We can see many drawbacks of this technique.

As it can see data only in linear position, it can classify wines even with dummy variables.

But overall information which we can take from this model that:

Even without known free and total sulfur dioxide we can classify wines with 88% accuracy. Very important attributes which affects balance of wine will be density, volatile acidity and sulphates.