Binary Classification

Tomas Olarte Hernandez
School of Engineering
EAFIT University
Medellin, Colombia
tolarteh@eafit.edu.co

Juan Gonzalo Quiroz Cadavid
School of Engineering
EAFIT University
Medellin, Colombia
jquiro12@eafit.edu.co

Ronald Cardona Martinez
School of Engineering
EAFIT University
Medellin, Colombia
rcardo11@eafit.edu.co

I. Introduction

Through the application of three classification algorithms, it is intended to analyze and determine when a given wine is likely to be good or not. For these we will use binary classification by means of Logistic Regression and Decision Tree classifiers. Additionally with the information obtained from those classifications we made a Linear Regression Model that given some wine data predicts its quality.

II. Exploring the Data

A. Checking the Integrity of the Data

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009].

These data set contain the following information about wines. Input variables (based on physicochemical tests):

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

Output variable (based on sensory data):

- Quality (score between 0 and 10)

Preliminary approaches to the data show that the data set is small (only 4200 entries) but there are no missing values or bad data. We also found that the *quality* column is a number between 1(very bad) and 10(excellent) [1], meaning the wine's quality. That score in the quality column needed to be transformed into a 1 (good) or a 0 (bad) wine to enable us to perform binary classification techniques.

Listing 1. Transforming scores into 0's and 1's

```
y = y.applymap(lambda x: 1 if x > 5 else 0)
```

We apply a mapping function over each value of the solution set *y*.

B. A Priory Probabilities

We found that 2815 out of 4200 wine instances were good (67.02%) and the remaining 1385 were bad wines (32.97%).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant.[1]

It could be interesting to test feature selection methods but that is out of the topic of these study. We'll select our features manually by analyzing the data.

III. Analyzing the Data and Variable Selection

Before applying any classification algorithm, we need to understand the data, and more important, when applying these methods, variable selection is a critical issue. So we'll get to know better our data before through some empirical and visual techniques. These analysis will be oriented to measure the importance of the inputs and also discard some irrelevant ones, helping the models we will use as classifiers to obtain a better performance.

A. Correlation Matrix

A correlation matrix [2] is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used as a way to summarize data, as an input into a more advanced analysis and as a diagnostic for advanced analyses. In this work, we will use the correlation matrix as a diagnostic for advanced analyses.

The line of 1.00s going from the top left to the bottom right is the main diagonal and were dropped out, it shows that each variable always perfectly correlates with itself.

Correlation analysis is a vital tool for feature selection [3]. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable [4]. Based on the correlation matrix shown in Fig. 1. the variables are not highly correlated to each other. One can say that correlation between Residual Sugar and Density is high (0.84). Or Alcohol and Density (-0.78). So maybe we can drop Alcohol and Residual Sugar because these variables are already represented in Density?. The answer we give is

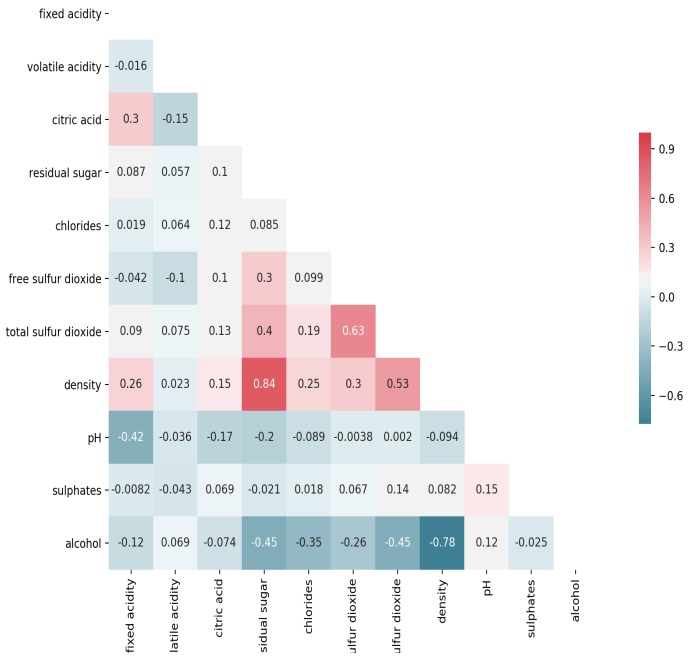

Fig. 1. Correlation matrix for white wine data set

no. Because Alcohol and Residual Sugar do not keep strong relationships with the rest of the variables.

B. Visualizing the distribution of the data set

In figure 2 we can see each variable's distribution in the data set. Here we used *distplot()* to fit a parametric distribution to the data set and visually evaluate how closely it corresponds to the observed data. Based on this distribution plots we can see that the less strongly correlated variables (See Fig 1.) correspond to the closest variables in the respective distribution plot (See Fig 2.). Take Alcohol as an example, its correlation indexes are almost strong and its distribution is more disperse. This can be seen in a two dimensional graph between those variables.

The variables that enable us to divide more the data distributions are the best to use in a model. So if we are planning to remove some variable from the input set, we might remove those less disperse. For example *free sulfur dioxide*.

But maybe only one variable cannot help classifying the data, but that variable combined it with another variable, maybe they both can. And that's the case of *fixed acidity*, where its distribution in Fig 2 is very close but we will see below that it is one of the most important variables appearing in almost every top 5 models.

C. All Possible Feature Combinations

The brute force and less visual approach we've taken to analyze the data was to compute all possible variable

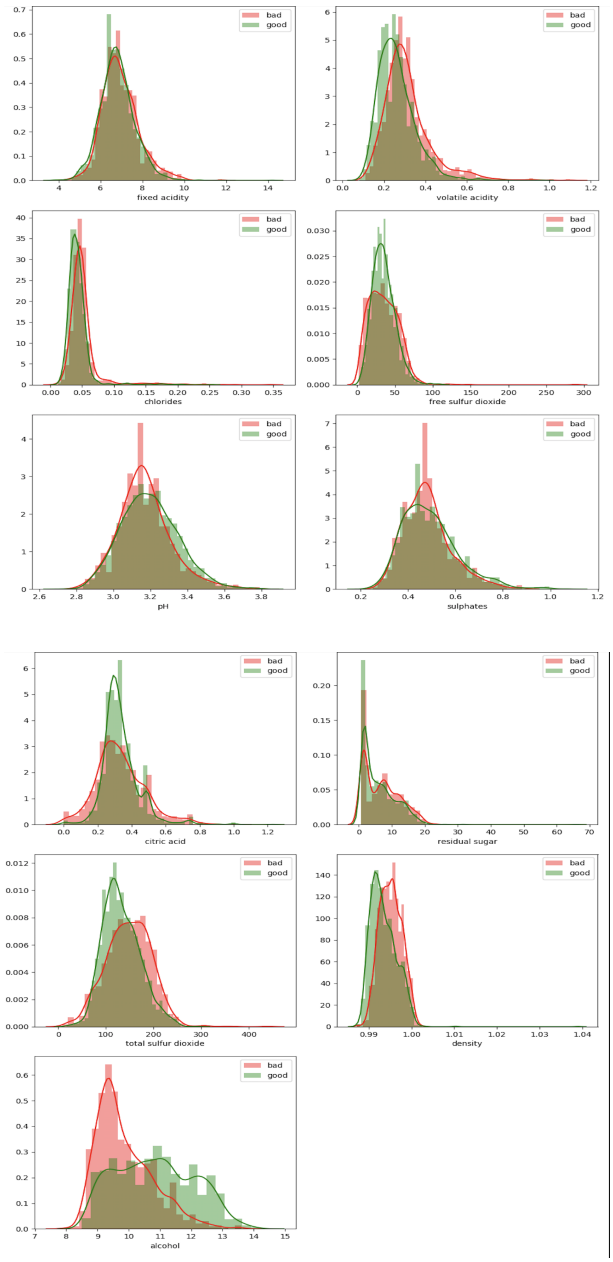

Fig. 2. Distribution Plots

combinations without repetition. To obtain all the variable combinations using at least 6 variables we used:

Listing 2. Computing all combinations

```
combinations = []
i = 12
for j in range(6, i):
  combinations.append(itertools.combinations(df, j))
```

In Listing 1 *df* contains all variable names and *i* contains the number of variables to use, in our case 12 because upper index is exclusive.

$$C_k(n) = (n|k)\frac{n!}{k!(n-k)!} \tag{1}$$

$$C_6(11) = (11|6)\frac{11!}{6!(11-6)!} \quad (2)$$

$$C_6(11) = 462 \quad (3)$$

That leaves us with 462 different models and each one needs to be evaluated in order to obtain its score and select the top models. To select the top models we are going to follow the workflow described in Fig 3.

TABLE I
TOP 5 FEATURE COMBINATIONS USING 5 OR LESS VARIABLES

Logistic Regression	Decision Tree	Variables Used
0.744761905	0.743809524	['volatile acidity', 'sulphates', 'alcohol']
0.751428571	0.741904762	['volatile acidity', 'pH', 'alcohol']
0.737142857	0.754285714	['volatile acidity', 'density', 'pH', 'alcohol']
0.765714286	0.745714286	['volatile acidity', 'residual sugar', 'density', 'alcohol']
0.732380952	0.76952381	['fixed acidity', 'volatile acidity', 'chlorides', 'sulphates', 'alcohol']
0.73047619	0.76952381	['fixed acidity', 'volatile acidity', 'pH', 'sulphates', 'alcohol']

From Table I we can obtain many insights about the best features to fit our models.

- The lower the correlation degree between the variables implemented in the model, the greater its performance. This can be noted in Table I, where the correlation between variables is close to 0, as it is the case of *alcohol* and *volatile acidity* where its correlation is 0.069 , which gives way to conclude that **the more correlated are two variables, the lower their contribution to the model**. Two highly correlated variables reduce the amount of information the model receives.
- There are variables that have more importance on the model, you can note this in Table I, all the top 5 models contain *alcohol and volatile acidity* as well as 3 out of 5 models contain *sulphates*
- The results above show that the lower the correlation grade amount one variable (See Fig 1), the accuracy returned by the model is higher, that means that the variable is more significant to the model. For example all the correlations in Fig 1. for *volatile acidity* are low, and this variable appears in all top 5 models. The same happens with *alcohol*.
- Through obtaining combinations using only 9 nine or more variables, we can see in Table II that for Decision Tree classifiers, the accuracy grows as the number of variables do. This behavior is not the same for Logistic Regression classifiers, where the obtained accuracies are similar to those shown in Table I.
- The variables that enable us to divide more the data distributions are the best to use in a model.

TABLE II
TOP 5 FEATURE COMBINATIONS USING NINE OR MORE VARIABLES

Logistic	Decision Tree	Variables Used
0.792380952	0.743809524	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'density', 'pH', 'sulphates', 'alcohol']
0.756190476	0.786666667	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'free sulfur dioxide', 'sulphates', 'alcohol']
0.743809524	0.786666667	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'pH', 'sulphates', 'alcohol']
0.747619048	0.762857143	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'pH', 'sulphates', 'alcohol']
0.755238095	0.773333333	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'free sulfur dioxide', 'density', 'sulphates', 'alcohol']

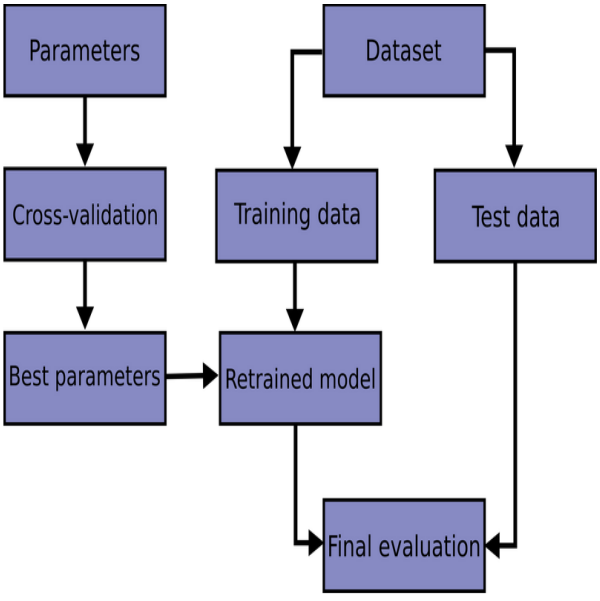

Fig. 3. Grid Workflow. See [5]

IV. MODEL EVALUATION

The next step in Grid Workflow (Fig. 3.) is use the selected parameters listed in tables I and II to retrain the models using

the 70% of the data and keep 30% to obtain the final evaluation and conclude the experiment.

A. Confusion Matrix

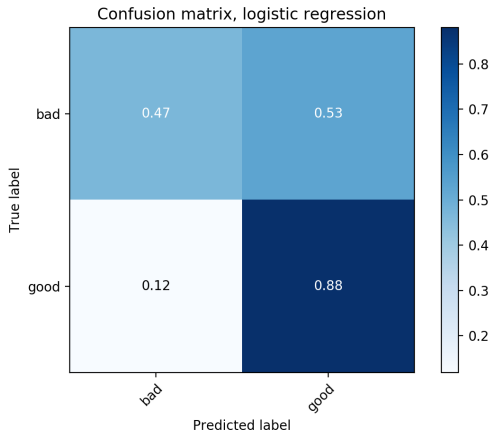

Fig. 4. Variables used: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, density, pH, sulphates, alcohol

In Fig 4 we can see the confusion matrix for a model using 9 parameters that obtained *79.2380952 %*. Given the *a priori* probabilities shown above, the model yields a good performance to predict true positives but false positives are high. Given the nature of our problem it is very important not to fail when classifying a bad wine.

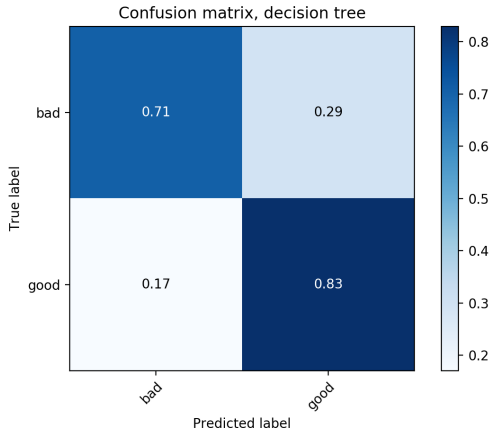

Fig. 5. Variables used: fixed acidity, volatile acidity, citric acid, residual sugar, pH, sulphates, alcohol

In Fig 5. True positives are reduced. This model is more selective because stops classifying almost everything as a good wine and reduces the false positives obtaining an accuracy of *78.6666667 %*

V. Final model

Comparing both tables (Table I and Table II), using the information given about the correlation matrix (Fig. 1) and the Distribution of the variables (Fig. 2) We agree that the best parameters combinations for the model are:

- **Logistic Regression Model:** [fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, density, pH, sulphates, alcohol]. This combination yields *79.2380952 %* of accuracy
- **Decision Tree Model:** [fixed acidity, volatile acidity, citric acid, residual sugar, pH, sulphates, alcohol]. This combination yields *78.6666667 %* of accuracy

We end with two models, now the question is which model would we use?. In order to answer this question we might give a look at the Confusion Matrix, represented in Fig. 4 (Representing the Logistic regression Model) and Fig. 5 (Representing the Decision Tree Model). We infer that our goal as **wine consumer** is to minimize the false positives margin, thus the proper model would be the **Decision Tree Model** specified above. Visually the very first questions in the decision tree are about *alcohol* and *volatile acidity* which confirms our selected variables. The tree can be consulted by following the link in [5].

VI. Linear Model

The same process used to search for the best parameter combinations is used to find those for a linear regression model (Grid Work flow. Fig 3). Starting from a random combination of parameters, validating those models an the picking up the top 5 models. The process resulted in Table III.

Squared Error	Mean Squared Error	Variables used
0.29827098	0.527489545	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'sulphates', 'alcohol']
0.283072253	0.527877314	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'free sulfur dioxide', 'sulphates', 'alcohol']
0.271389213	0.583510509	'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'pH', 'sulphates', 'alcohol']
0.280448803	0.581185907	['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'sulphates', 'alcohol']

A. Evaluation metric

*1) **Squared error (SE):*** The Squared error gives us information about how many points (in terms of percentage) came across the linear function generated.

*2) **Mean Squared error (MSE):*** The Mean Squared error gives us information about the average distance between the predicted points and the actual set of points used to test the function.

*3) **Squared error Vs Mean Squared error**:* This set of metrics help us when finding the best combinations is about. In our case, we chose **Mean Squared error** metrics due to the insights it leads to. From our perspective it's more valuable the averaged distances between the predicted points and the actual points rather than the averaged points that are actually over the function. This metric (Squared Error) can help us but it's not enough information about how well the Linear function is working.

B. Variable selection

The reasoning implemented in Section V apply to the analyze of the parameters used in Table III , having the **Mean Squared Error** as the metric used to compare them we agree that the proper model would be one whose parameters return a small MSE (Small distance between the points predicted and the actual ones). Based on our analysis, we agree the best parameters combination is *['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'pH', 'sulphates', 'alcohol']* with a MSE of *0.5540* $\pm 1 * 10^{-4}$

REFERENCES

[1] *P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.*

[2] *Bock, Tim. What Is a Correlation Matrix? Displayr, 29 Oct. 2018, www.displayr.com/what-is-a-correlation-matrix/.*

[3] *Srishti Saha. Let Us Understand the Correlation Matrix and Covariance Matrix. Towards Data Science, Towards Data Science, 5 Oct. 2018, towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22.*

[4] *Vishal R. Feature selection "Correlation and P-value" Towards Data Science, Towards Data Science, 11 Sep. 2018, https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf.*

[5] *https://scikit-learn.org/stable/modules/crossvalidation.html*

[6] *https://github.com/JuanGQCadavid/knowledge-engineering/blob/master/DecisionTree.ipynb*