Wine Type Prediction from chemical composition using Deep Learning

Ravi Munde| munde.r@husky.neu.edu

Abstract: Red and white wines are distinct due the process involved in their manufacturing. The chemical composition might not be the factor which decides the color of the wine, however, the composition might be the result of the type of the wine in contention. It is, thus, possible to predict if the wine is either red or white using the chemical composition. We built a deep neural network model that was trained on the chemical attributes of the wine. The quality of the wine provided was skipped as it is having no contribution and is independent of the wine.

Introduction: We have used the Portugal wine dataset. The dataset is separated into two separated sets, white wine and red wine. We studied the correlation between all the attributes and observed that each attribute has a relation with the color of the wine. The dataset used here is unbalanced in a proportion of 75.4% for "White wine" and 24.6% for "Red wine". But our model could learn and predict wine type with a high accuracy (99%).

Data:

UCI Machine learning respository: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal.

Original source: http://www3.dsi.uminho.pt/pcortez/wine/

Data Attributes:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulphur dioxide
7. total sulphur dioxide
8. density
9. pH
10. Sulphates
11. Alcohol
12. quality (score between 0 and 10, based on sensory data)

Data Exploration:

We checked for null values by plotting missing values for each column and observed no missing values. Also, the data types of the all the attributes were as expected and no modification was required. Since we have two datasets, we have combined them together with an additional column *'is_red'* to indicate wine type (0=white, 1=red).

Data Distribution:

Plotting each columns distribution graph

- **fixed acidity** - normally distributed, mean around 7
- **volatile acidity**- normally distributed, mean 0.25, but due to excessive outliers, data is positively skewed
- **citric acid** - normally distributed, few outliers with values greater than mean but the distribution is almost symmetric
- **residual sugar** - almost normal distribution, most values are close to 0 but some outliers has skewed it positively
- **chlorides** - normally distributed, some ouliers on right of the mean, most data close to 0
- **free sulfur dioxide** - normally distributed
- **total sulfur dioxide** - too many outliers on the left of mean making the distribution negatively skewed
- **density** - symmetric normal distribution, few outliers present on right of the mean
- **pH** - symmetric normal distribution
- **sulphates** - normally distributed but some outliers extending to right of the mean
- **alcohol** - exact values were not really well distributed, however, a tan transformation of the data seems

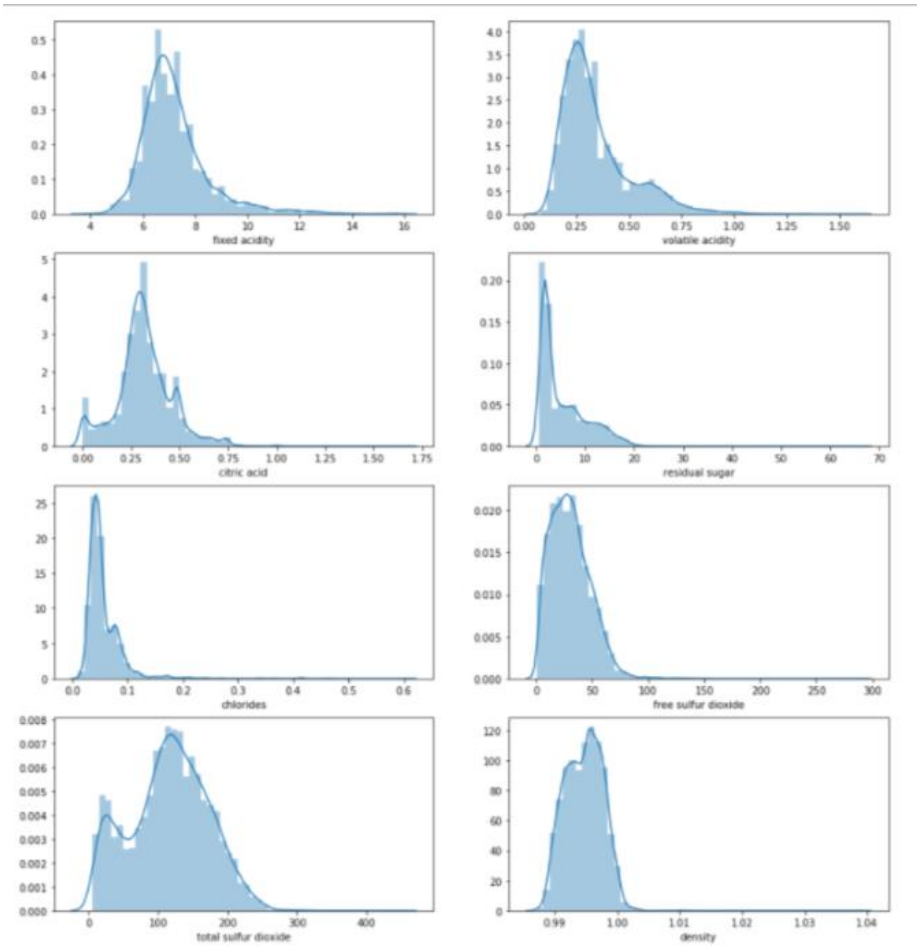

Figure 1 Data Distributions

Outliers: Plotting graphs for outliers, we see that almost all the columns have some outliers. However, they cannot be removed or changed as they are actual business values and need to be considered while training our model

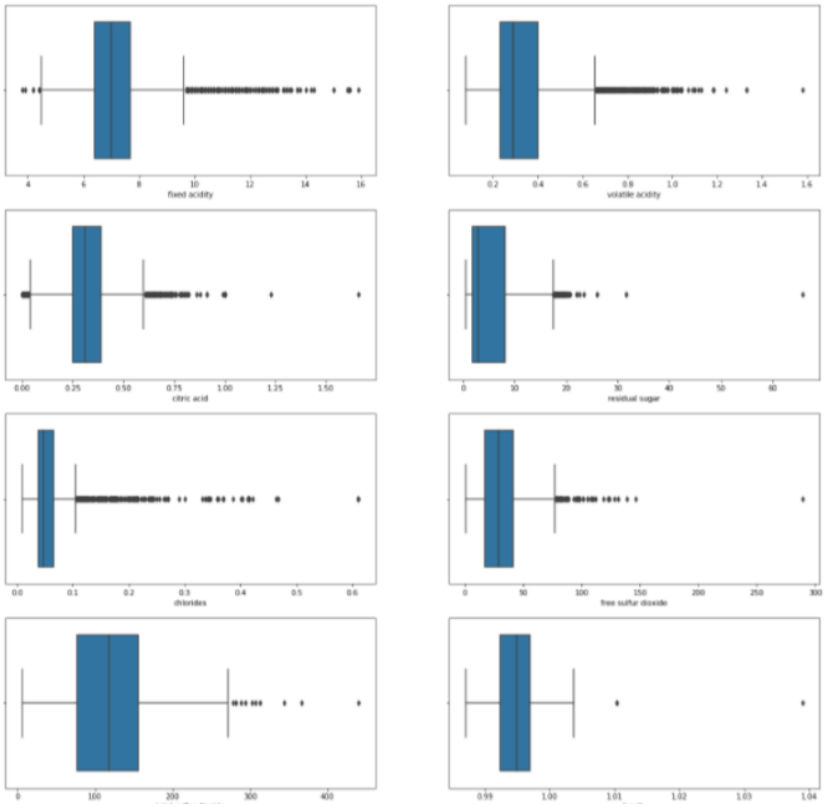

Figure 2 Outliers

Attribute variations: For our network to learn the features, it is required that all attributes have equal weightage while learning. We plot the average of all the columns to see the variation. We consider data normalization to train model accordingly.

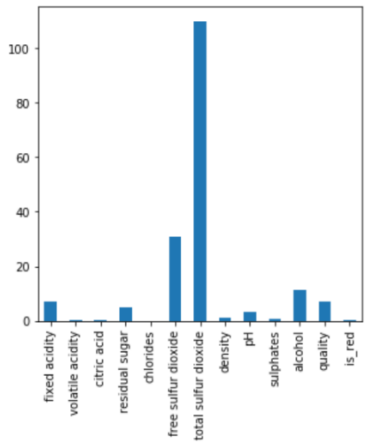

Figure 3 Mean values of each column

Chemical Composition Analysis: Plotting high quality wines(ratings>6) and some of their properties (quality >6)

- **Fixed Acidity -** Red Wine seems to have higher fixed acidity as compared to white wine, especially wines with quality 5,6 & 7
- **Citric Acid -** Red Wines have mostly lower citric acid content when high quality wines are considered, however, some have higher content just like most of the white wines.
- **Residual Sugar -** Mostly white wines have residual sugar content, red wines, if present have comparatively low residual sugar
- **Chlorides -** Both, white and red wine contain moderate amount of chloride content.
- **Free sulphur dioxide -** White wines have higher free sulphur dioxide content and low sulphur dioxide content means higher quality wine as per the graph. Since there are no red wines with quality 9, this interpretation cannot be applied to red wines.

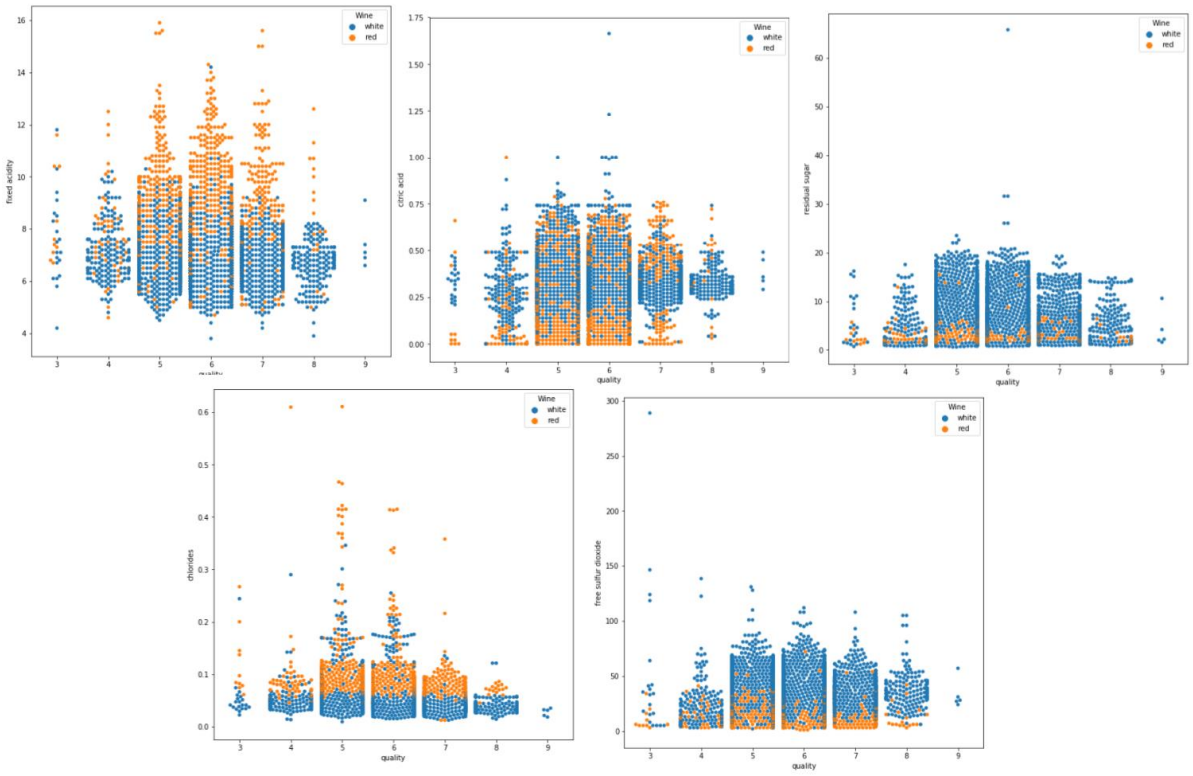

Figure 4 Swarm plot analysis

Correlation: Plotting pair plot for all columns with hue for red and white wine. Density seems to be independent with all the columns.

The plot also shows possible linear relation between different columns:

1. is_red - volatile acidity
2. is_red - total sulphur dioxide
3. density - alcohol

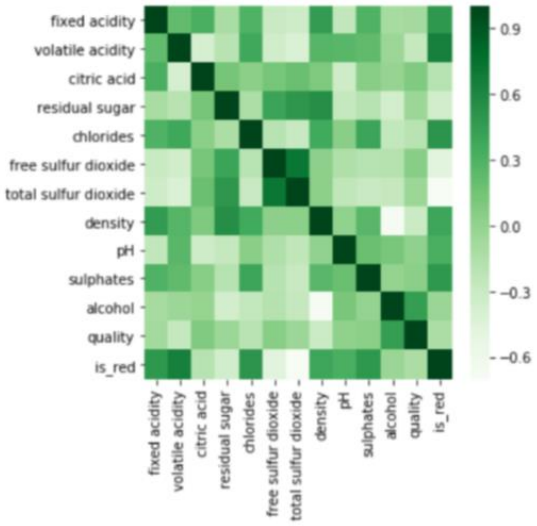

Figure 5 Correlation heat map

Aim: Given some contents of the wine, we will try to predict the type of the wine. However, the quality of the wine in no way is a predictor if the wine is red or white, hence ignoring it in our model is a good idea.

Background Work:

- Predicting Wine Quality, Ilker Karakasoglu : The published article has detailed Data Exploration and predictions generated by Support Vector Machine(Regression). Three different models were tried, one for white wine, another for red and the third model was fitted for the combined dataset of wines. The individual dataset models scored better than the combined (RMS = 0.71). The color of the wine was not a significant factor for deciding factor for the quality of the wine. Alcohol and Sulphates were the most important factor that decided the quality.
- Wine Classification This project develops 5 algorithms of machine learning to classify the wines in white or red according to the 11 variables that characterize the wine subject to classification. The modelling results were similar across all datasets. The same model - Random Forest (RF)- was selected for "df" and "dff" datasets. On the other hand, Support Vector Machine (SVM) model was selected for "dfffull" dataset.
- Analysis of Wine Quality Data : Classification was done on the quality of the wines (Split into high, medium and low quality). Random forest with ntree = 150 gave an accuracy of 67.7% as compared to 50% given by simple Tree-based model. However, the results were inconclusive as the model did not show any significant dependency between wine quality and its chemical composition.

Deep Learning Model:

- We built the deep learning model to predict the color of the wine depending on its chemical composition. Since the datasets were separated into two files, we added an extra parameter on the combined dataset to indicate the color of the wine(is_red). Our model was trained to predict this variable, 0 indicating a white wine while 1 being the red wine. Even though the dataset was imbalanced with smaller number of red wines, the model performed well to predict wine (accuracy around 99%).

- We used Keras to implement the model in python for its simplicity to plug and use hyperparameters. Input was the 11 attributes (quality column was not considered for training). We first tried a two-hidden layer model which seemed to perform well. However, it stayed on the accuracy of 94% and showed no improvement over epochs. Single hidden layer had an accuracy of 99% on training data.
- The output of the model was a single neuron showing a probability of a wine being red or white(1=red). Since the probability can lie between 0-1, anything above 0.5 was considered 1(red wine prediction) and anything less 0(white wine prediction).
- The confusion matrix generated on the test data validated the accuracy of the network (99.53%) meaning the model was not overfitting on the train data.

Conclusion: The color of the wine may not be a result of the chemical composition in discussion but we observed that the composition was unique to each wine type. The pattern in the distribution of these chemicals is enough to predict the type of the wine. However, by analysing the data, it was inconclusive whether the chemicals have the effect on the type of the wine. We tried to implement a neural network to predict the type of wine and got a high accuracy of 99% without overfitting. The distribution of the wine types was unbalanced, the white wines being greater in number but had no effect on the accuracy of the model.

References:

- Code Sample: https://www.udemy.com/deeplearning/learn/v4/overview
- Machine Learning: https://github.com/arqmain/Machine_Learning/blob/master/R_MLearning/MLearning_Classification_Portugal_Wine_TwoClass_RedWhite_R_KFold/README.md?lipi=urn%3Ali%3Apage%3Ad_flagship3_pulse_read%3Br6RIXJl2Tv%2BJgakD0OCNfA%3D%3D
- Predicting Wine Quality, Ilker Karakasoglu: https://web.stanford.edu/~ilker/doc/wine_Stats315A.pdf
- Analysis of Wine Quality Data https://onlinecourses.science.psu.edu/stat857/node/223:
- UCI Machine learning https://archive.ics.uci.edu/ml/datasets/Wine+Quality
- Paulo Cortez, University of Minho, Guimarães, Portugal http://www3.dsi.uminho.pt/pcortez