Prediction of Wine Quality

Yinuo Song

***Abstract*—Wine quality refers to the factors that go into producing a wine, as well as the indicators or characteristics that tell you if the wine is of high quality. This report focuses on the prediction of wine quality. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. 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.). Using the Linear Regression model, we analyze the effects of ten features, e.g., alcohol and pH, on the quality scores of wines. We also use the Support Vector Machine and Neural Networks with Tensorflow approaches to replicate the prediction. For all of the methods, we calculate the mean square error and evaluate their fitness accuracy on the dataset.**

I. Introduction

Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wider range of consumers. Wine is an alcoholic drink typically made from fermented grapes. The earliest known traces of wine are from Georgia (c. 6000 BC), Iran (Persia) (c. 5000 BC), and Sicily (c. 4000 BC) although there is evidence of a similar alcoholic drink being consumed earlier in China (c. 7000 BC). Wine reached the Balkans by 4500 BC and was consumed and celebrated in ancient Greece, Thrace and Rome. Throughout history, wine has been consumed for its intoxicating effects.

To support its growth, the wine industry is investing in new technologies for both wine making and selling processes. Wine certification and quality assessment are key elements within this context. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices). When you know what influences and signifies wine quality, you'll be in a better position to make good purchases. You'll also begin to recognize your preferences and how your favorite wines can change with each harvest. Your appreciation for wines will deepen once you're familiar with wine quality levels and how wines vary in taste from region to region.

In this report, we present a case study for modeling wine quality based on two datasets related to red and white variants of the Portuguese "Vinho Verde" wine. The report is organized as follows: Section 2 decribes the task; in Section 3, we discuss the major challenges and solutions; Section 4 presents the data processing, including the major results and analysis, and evaluation metrics of dataset description, missing value, correlation plot, descriptive statistics, outliers; Section 5 presents the machine learning experiments, including the major results and analysis of linear regression model, support-vector machines and Neural Network Tensorflow; our conclusion and future works are drawn in Section 6.

II. Task Description

In this case study, we choose Python as our statistical software.

The first part is preprocessing and exploratory data analysis. We set an appropriate working environment and import two datasets, including the red wine dataset and the white wine dataset. We then treat with the raw data and check if there're missing values. After that, we draw the correlation plots of pairwise variables and reduce the potential multicollinearity. We then detect the outliers and remove them to improve the performance of our models.

The second part is model building and analysis. For each experiment, we split the dataset into training set and testing set at different ratios, e.g., 80-20 and 60-40. We first run our Linear Regression model on the combined dataset of red wine and white wine. We then run our model on segmented dataset. Also, we use the alternative approach, Support Vector Machine, to build another model. For each model, we analyze the mean square error (MSE) to evaluate its performance. Last, we use the Neural Networks with Tensorflow to build another model. And we calculate the fitness accuracy for each model to evaluate its prediction performance.

III. Major Challenges

One of the challenges is the multicollinearity. Severe multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct model.

To reduce the multicollinearity, we focus on the correlation plots and drop some variables that are highly correlated to each other. Also, we run our models on segmented datasets, which can also reduce the multicollinearity.

Another challenge is the outliers. They are data records that differ dramatically from all others, and they distinguish themselves in one or more characteristics. In other words, an outlier is a value that escapes normality and can (and probably will) cause anomalies in the results obtained through algorithms and analytical systems. There, they always need some degrees of attention.

To detect the outliers, we write some codes to classify the

observations into normal ones and abnormal ones. To be specific, we use the Interquartile Range (IQR) method to detect outliers. If a data point is below Q1 – 1.5×IQR or above Q3 + 1.5×IQR, it is viewed as being too far from the central values to be reasonable. We then delete these observations and solve the problem.

After remove the outliers, we further make Box Cox transformation for both red and white wine datasets to transform the non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques. In this way, the model will have higher accuracy than before.

IV. DATA PROCESSING

Oftentimes, data can be quite messy, especially if it hasn't been well-maintained. Data processing is, generally, the collection and manipulation of items of data to produce meaningful information.

A. Dataset Description

We have two datasets, the red wine dataset with 1599 observations and the white wine dataset with 4898 observations, having 6497 observations in total.

In each dataset, there're 12 variables: fixed acidity, the fixed acids involved with wine that do not evaporate readily; volatile acidity, the amount of acetic acids in wine, which at too high of levels can lead to an unpleasant, vinegar taste; citric acid, which is found in small quantities but can add "freshness" and flavor to wines; residual sugar, the amount of sugar remaining after fermentation stops; chlorides, the amount of salt in the wine; free sulfur dioxide, the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion, which prevents microbial growth and the oxidation of wine; total sulfur dioxide, amount of free and bound forms of SO2; 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); sulphates, a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant; alcohol, the percent alcohol content of the wine; and quality, the sensory data, median of at least 3 evaluations graded by wine experts between 0 (very bad) and 10 (very excellent).

B. Missing Value

Missing values are always a problem while analyzing the data and also building the models using that data. Missing values might affect the model and analysis to greater extent by resulting in errors in the code and sometimes wrong insights about the data. Thus, we look for missing values in both the data. The result (Table 1) clearly shows that we have the data intact 100% and there is no missing data in our dataset.

C. Correlation Plot

We draw the correlation plots of two datasets (Figure 1 & Figure 2). We find that fixed acidity and citric acidity, density are highly correlated. Also, free sulfur dioxide and total sulfur dioxide are highly correlated. Hence, we drop the variables citric acidity, density and free sulfur dioxide to reduce the multicollinearity. After that, we also draw the correlation plot of the difference between red wine and white wine (Figure 3). Here, we show the pair plots of two datasets (Figure 4 & Figure 5).

TABLE 1
MISSING VALUE

	Red Wine	White wine
fixed_acidity	0	0
volatile_acidity	0	0
citric_acid	0	0
residual_sugar	0	0
chlorides	0	0
free_sulfur_dioxide	0	0
total_sulfur_dioxide	0	0
density	0	0
pH	0	0
sulphates	0	0
alcohol	0	0
quality	0	0
color	0	0

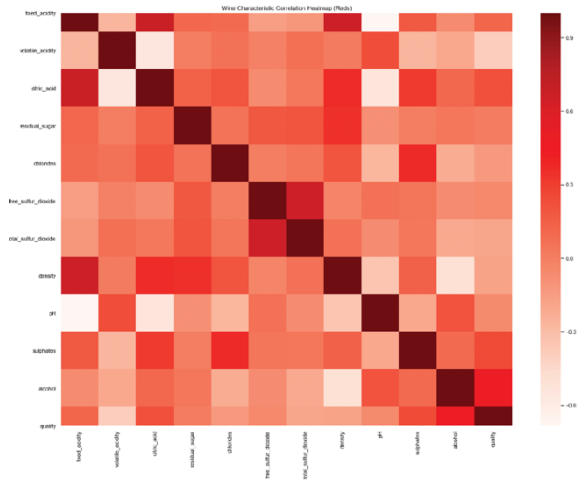

Fig. 1. Correlation plot of red wine.

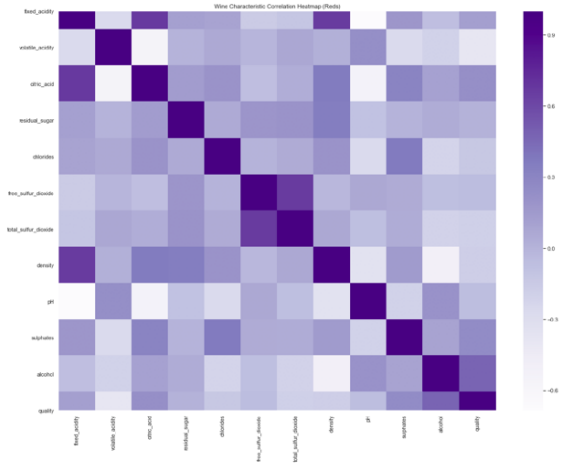

Fig. 2. Correlation plot of red wine.

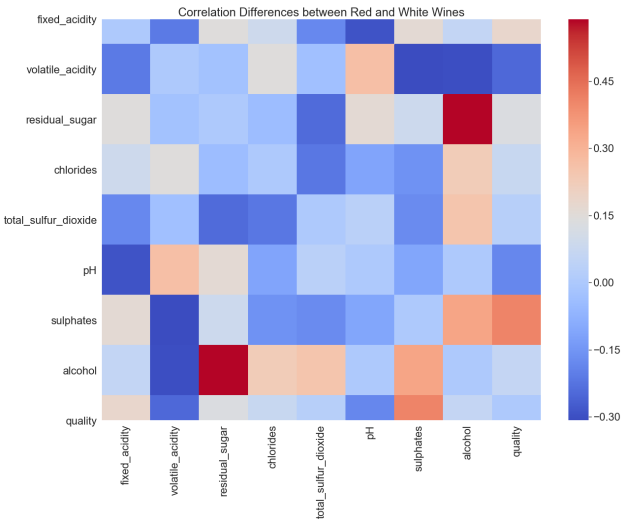

Fig. 3. Correlation plot of difference between red wine and white wine.

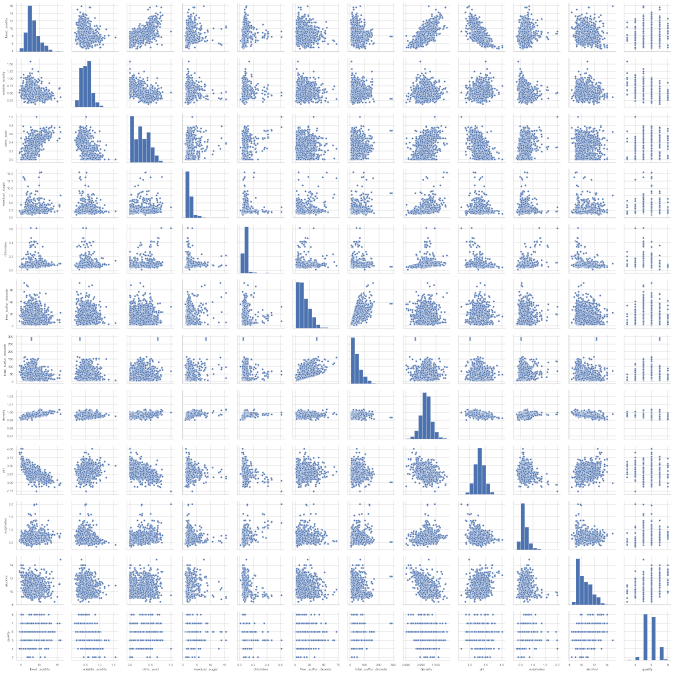

Fig. 4. Pair plot of red wine.

Heatmap of red wine dataset shows the correlation between the variables in its dataset. High concentration of color representation high correlation. Dark red represents positive and light color represents negative correlation. And the correlation plot shows the relationship between the different variables of a dataset.

From the above heatmap and correlation plot on the red wine, we have the following inference: 1) It looks like we have pH and fixed acidity has inverse relationships between them. The same in case of citric acid and volatile acidity too. 2) There is a strong positive relationship between total sulfur dioxide and free sulfur dioxide.

Heatmap of white wine dataset shows the correlation between the variables in its dataset. High concentration of color representation high correlation. Dark purple represents positive and light color represents negative correlation. And the correlation plot shows the relationship between the different variables of a dataset.

From the above heatmap and correlation plot on the white wine, we have the following inference: 1) We see that low density contents are high on alcohol content. 2) Sugar content in the wine might represents the wine density. 3) Just like red wine, here also total and free sulfur dioxide are related together.

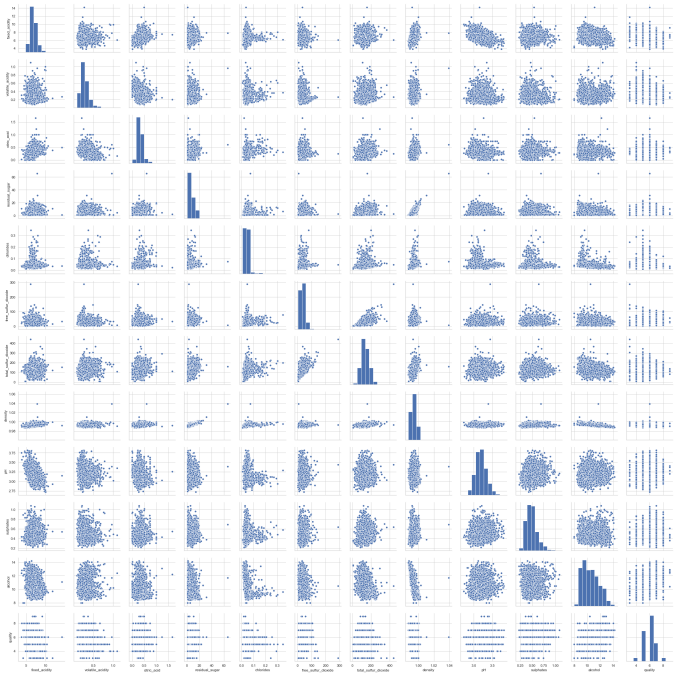

Fig. 5. Pair plot of white wine.

D. Descriptive Statistics

We look at the descriptive statistics of two datasets (Table 2 & Table 3). We find that the mean quality score of red wine is 5.64, and most independent variables have some outliers. We also know that the mean quality score of white wine is 5.88. Then we use the IQR method mentioned above to remove the outliers.

We also find that there're some significant differences between red wine and white wine. For example, the red wine has higher average fixed acidity, volatile acidity, chlorides, pH and sulphates than the white wine. The white wine has higher residual sugar, total SO2, alcohol and quality than the red wine.

TABLE 2
DESCRIPTIVE STATISTICS OF RED WINE

	fixed acidity	volatile acidity	residual sugar	chlorides
mean	8.32	0.53	2.54	0.09
min	4.60	0.12	0.90	0.01
25%	7.10	0.39	1.90	0.07
median	7.90	0.52	2.20	0.08
75%	9.20	0.64	2.60	0.09
max	15.90	1.58	15.50	0.61

TABLE 2 (CONT.)
DESCRIPTIVE STATISTICS OF RED WINE

	Total SO2	pH	sulphates	alcohol	quality
mean	46.47	3.31	0.66	10.42	5.64
min	6.00	2.74	0.33	8.40	3.00
25%	22.00	3.21	0.55	9.50	5.00
median	38.00	3.31	0.62	10.20	6.00
75%	62.00	3.40	0.73	11.10	6.00
max	289.00	4.01	2.00	14.90	8.00

TABLE 3
DESCRIPTIVE STATISTICS OF WHITE WINE

	fixed acidity	volatile acidity	residual sugar	chlorides
mean	6.85	0.28	6.39	0.05
min	3.80	0.08	0.60	0.01
25%	6.30	0.21	1.70	0.04
median	6.80	0.26	5.20	0.04
75%	7.30	0.32	9.90	0.05
max	14.20	1.10	65.80	0.35

TABLE 3 (CONT.)
DESCRIPTIVE STATISTICS OF WHITE WINE

	Total SO2	pH	sulphates	alcohol	quality
mean	138.36	3.19	0.49	10.51	5.88
min	9.00	2.72	0.22	8.00	3.00
25%	108.00	3.09	0.41	9.50	5.00
median	134.00	3.18	0.47	10.40	6.00
75%	167.00	3.28	0.55	11.40	6.00
max	440.00	3.82	1.08	14.20	9.00

E. Outliers

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Therefore, before we build regression models and use machine learning method to make the predictions about the wine quality, we detect whether there are outliers in the datasets and remove them from the datasets (Figure 6 & Figure 7).

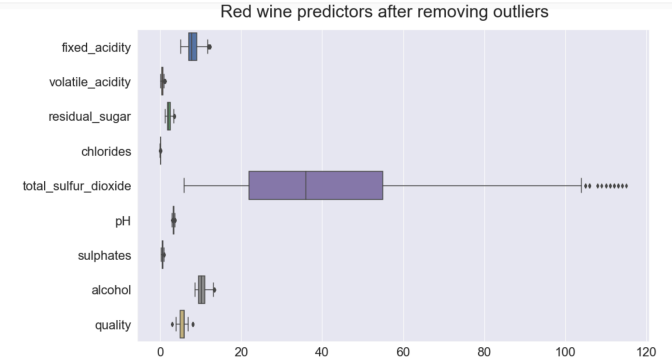

Fig. 6. Red wine predictors after removing outliers.

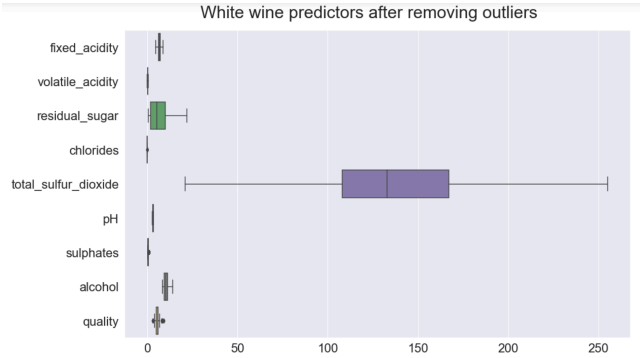

Fig. 7. White wine predictors after removing outliers.

Last, we also use the Box Cox transformation method to transform the non-normal dependent variables into a normal shape (Figure 8 & Figure 9).

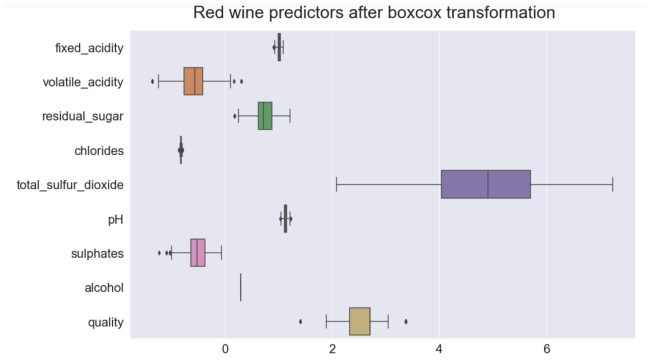

Fig. 8. Red wine predictors after boxcox transformation.

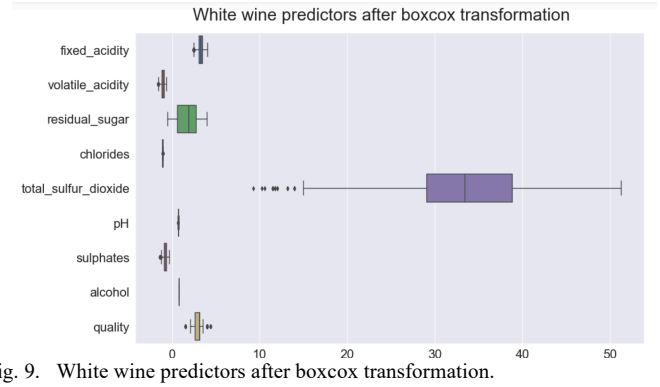

Fig. 9. White wine predictors after boxcox transformation.

V. MACHINE LEARNING METHOD

We will use three approaches for building models that could predict quality of the red and white wine. One is using the traditional linear regression model; one is using the classification Machine Learning algorithm and another by using Neural Networks using TensorFlow. For the traditional model we use Support vector Machines.

A. Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is

called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

We first use the whole dataset regression to find the coefficient estimates of different factors (Table 4). In this model, we generate a dummy variable to label the color of the wine. For the whole dataset: The in-sample Mean squared error: 0.11; Out-of-sample Mean squared error: 0.11 (Table 5).

TABLE 4
LINEAR REGRESSION RESULT OF COMBINED DATASET

	Coefficient	t	P>|t|
intercept	-20.973	-29.264	0.000
total_sulfur_dioxide	0.002	2.142	0.032
residual_sugar	0.064	12.182	0.000
fixed_acidity	-0.046	-2.533	0.011
chlorides	-2.731	-6.663	0.000
Volatile_acidity	-0.443	-18.289	0.000
alcohol	25.529	28.603	0.000
sulphates	0.188	7.825	0.000
pH	0.114	0.529	0.597
color_R	13.465	29.661	0.000

TABLE 5
MEAN SQUARE ERROR OF WHOLE DATASET

	Mean square error
in-sample	0.11
out-of-sample	0.11

We then use the segmented dataset regression to run the regression again. For each model, we calculate the accuracy and evaluate its performance.

For the red wine dataset: In-sample Mean squared error: 0.05; Out-of-sample Mean squared error: 0.05. The linear regression results for red wine dataset are as follows (Table 6 & Table 7).

TABLE 6
LINEAR REGRESSION RESULT OF RED WINE

	Coefficient	t	P>|t|
intercept	-919.828	-13.142	0.000
total_sulfur_dioxide	-0.004	-0.679	0.498
residual_sugar	-0.000	-0.008	0.994
fixed_acidity	-0.520	-1.674	0.094
chlorides	-2.039	-2.608	0.009
Volatile_acidity	-0.219	-7.250	0.000
alcohol	3124.071	13.156	0.000
sulphates	0.407	10.746	0.000
pH	-1.196	-4.407	0.000

TABLE 7
MEAN SQUARE ERROR OF RED WINE

	Mean square error
in-sample	0.05
out-of-sample	0.05

For the white wine dataset: In-sample Mean squared error: 0.12; Out-of-sample Mean squared error: 0.12. The linear regression results for white wine dataset are as follows (Table 8 & Table 9).

TABLE 8
LINEAR REGRESSION RESULT OF WHITE WINE

	Coefficient	t	P>|t|
intercept	-21.417	-27.191	0.000
total_sulfur_dioxide	0.002	2.500	0.012
residual_sugar	0.064	11.370	0.000
fixed_acidity	-0.036	-1.796	0.073
chlorides	-2.359	-4.988	0.000
Volatile_acidity	-0.473	-14.664	0.000
alcohol	25.831	26.733	0.000
sulphates	0.104	3.610	0.000
pH	0.718	1.963	0.050

TABLE 9
MEAN SQUARE ERROR OF WHITE WINE

	Mean square error
in-sample	0.12
out-of-sample	0.12

B. Support Vector Machine

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). 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. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

The accuracy of the SVM model on Red wine dataset with 10-fold cross validation is 59.47 percentage with the best cost value of 100.

The accuracy of the SVM model on White wine dataset with 10-fold cross validation is 59.36 percentage with the best cost value of 100 (Table 10).

TABLE 10
SVM ACCURACY

	Accuracy
red wine	59.47%
white wine	59.36%

Now we have the model for predicting the quality scale of the red and white wine separately. We also build a model that can make prediction on whether it is a red or white wine. For this we merge both the dataset into one single dataset by generating new column that represent '0' for Red and '1' white

wine. Then we follow the same procedure as previously to use the custom SVM model we build before.

The accuracy of the SVM model on whole dataset with 10-fold cross validation is 100.0 percentage with the best cost value of 100.

This result is amazing, we have 100% accurate model that we can implement in the production environment to classify any new data points as whether it is a red or white wine.

C. Neural Networks with TensorFlow

Created by the Google Brain team, TensorFlow is an open source library for numerical computation and large-scale machine learning. TensorFlow bundles together a slew of machine learning and deep learning (aka neural networking) models and algorithms and makes them useful by way of a common metaphor.

By using this Neural Networks with TensorFlow method, we get the model with 99.779% accuracy. The process is as follows (Table 11).

TABLE 11
TensorFlow Result

Number			Value
Epoch:	100	average cost:	0.252
Epoch:	200	average cost:	0.179
Epoch:	300	average cost:	0.145
Epoch:	400	average cost:	0.117
Epoch:	500	average cost:	0.099
Epoch:	600	average cost:	0.085
Epoch:	700	average cost:	0.076
Epoch:	800	average cost:	0.067
Epoch:	900	average cost:	0.061
Epoch:	10000	average cost:	0.056
		Final accuracy:	99.779%

VI. CONCLUSION

In conclusion, all of the three machine learning methods have got the wine quality predictive models with high accuracy.

For linear regression method, the whole dataset regression model generates In-sample Mean squared error: 0.11 and Out-of-sample Mean squared error: 0.11. The red wine dataset regression model generates In-sample Mean squared error: 0.05 and Out-of-sample Mean squared error: 0.05. The white wine dataset regression model generates In-sample Mean squared error: 0.12 and Out-of-sample Mean squared error: 0.12.

For Support Vector Machine method, the accuracy of the SVM model on Red wine dataset with 10-fold cross validation is 59.47 percentage with the best cost value of 100. The accuracy of the SVM model on White wine dataset with 10-fold cross validation is 59.36 percentage with the best cost value of 100. The accuracy of the SVM model that can make prediction on whether it is a red or white wine on whole dataset with 10-fold cross validation is 100.0 percentage with the best cost value of 100.

For Neural Network TensorFlow method, we get the prediction model with 99.779% accuracy.

Therefore, we find that in this experiment, the Support Vector Machine method has the highest accuracy with 100%, followed by Neural Network TensorFlow method with 99.779% and then by linear regression model with MSE 0.11.

We also can understand one more thing from this. That is if you look back our SVM model provided an accuracy of 100% where Neural Networks being praised as superior Machine Learning algorithm nowadays have got less on the SVM results. This shows that in general machine learning world, it is better to try the simpler ones at first before heading your way to hard problem.