Capstone Project Report: Machine Learning for Quality of White Wine

Angela P. Molder

Northwest Missouri State University, Maryville MO 64468, USA
S543719@nwmissouri.edu and apmolder@yahoo.com

Abstract. This project was for the Capstone course in the MS Data Analytics program. The goal was to use Machine Learning to determine what characteristics of white wine will determine quality. Data on almost 5,000 white wines was found with variables including acidity, sugar, chlorides, sulfur dioxide, density, pH, sulphates, alcohol, and wine quality. Exploratory data analysis was used to review the data and determine which variables should be used. The independent variables were narrowed down to volatile acidity, residual sugar, chlorides, pH, and alcohol. The dependent variable, quality, was on a scale of 1 for fair and 10 for great. Machine learning models chosen were Linear Regression, Lasso Regression, Random Forest, and GridSearchCV. After using training and test data, it was determined that none of the models were a good fit with the variables used. It will be necessary to conduct further analysis.

Keywords: wine quality · fixed acidity · volatile acidity · citric acid · residual sugar · chlorides · free sulfur dioxide · density · pH · sulphates · alcohol · exploratory data analysis · machine learning

1 Introduction

The domain for this Capstone Project Report was the quality of white wines. I chose this topic as I enjoy wine, but do not understand the science behind quality wines. Kaggle was searched for data sets on wine quality. [9] The data problem to be analyzed was determining the best features to use, such as acidity, sugar, density, and alcohol content. These features can help determine what type of wine to buy.

The steps for this project started with finding the best data and researching what the features mean for the production of wine. Once the data to be used was confirmed, it may be necessary to clean the data. If there are many features in the data, feature engineering will be used as needed. Machine learning models will be the basis for the data analysis.

2 Dataset

The data source used was Kaggle and the wine quality dataset was downloaded in csv format. [9] It was not necessary to do any data scraping techniques as

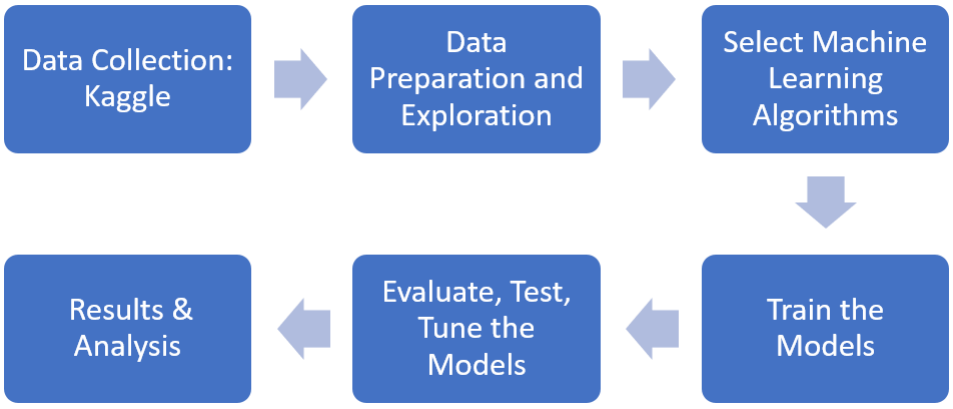

Fig. 1. Project Steps. A. Molder.

several relevant features were already included with a rank for quality. For the white wine data set there are almost 5,000 records and several attributes. These attributes consist of 3 for acidity and 2 for sulfur dioxides as well as one each for residual sugar, chlorides, density, pH, sulphates, alcohol content and the quality ranking. It was not necessary to do further data extraction, but the attributes were researched to find out why they are significant to wine quality.

The initial research was done online and several articles and websites were found to describe what the attributes mean. After the data was prepared, exploratory data analysis and feature engineering were used. This allowed some of the attributes to be filtered out. The Machine Learning (ML) algorithms chosen were Linear Regression, Lasso Regression, Random Forest, and GridSearch CV. After splitting the data into train set and test set, ML model training commenced with the training set. From the test set, the models were analyzed and validated before choosing the final ML model.

2.1 Data Cleaning

Data cleaning is a necessary step in data science as inaccurate data can lead to bad decisions. Some steps include determining what attributes are relevant and if there is any missing data. A decision needs to be made regarding what to do with missing values. Outliers or duplicate rows can also impact the resulting analysis. Excel will be used to determine any missing values or to remove any irrelevant columns. Python will be used for reviewing outliers and potential duplicate values. Some common methods on handling missing data are to drop the row or fill in with a mean, median, or zero value.

There were no rows with missing values. The white wine data set had 4,898 rows to begin, but had 943 duplicates leaving 3,955 rows. Using histograms and box plots, the attributes with the most outliers were determined. Once the outliers were dropped, there were 3,914 rows. As sulfur dioxide limits are set by

law for wine, these 2 attributes were removed leaving a total of 10 attributes. The dependent variable is quality while the independent variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, density, pH, sulphates, and alcohol.

2.2 Attribute Descriptions

The 10 different attributes each contribute varying factors to the quality of the wine. The first 3 that will be described are types of acidity. The sourness or tartness comes from the fixed acidity in wine. Acids from the grapes carry over into the wine and affect the pH, which affects the color, stability to oxidation, or how long the wine will last. Without acid, the wine becomes flat. [7] Volatile acid is measured by looking at mostly acetic acid levels. It has a vinegar-like aroma that indicates spoilage. There are legal limits on levels of volatile acid. [10] Like other acids, citric acid adds to the level of freshness in the wine, but if levels are too high then unwanted microbes can grow. [2]

Natural grape sugars get left over in wine after the fermentation process has ended; this is known as residual sugar. Levels of residual sugar at 10 g/L are considered dry wines compared to 35 g/L or more for sweet wines. [8] Saltiness of wine is determined by the level of sodium chlorides. This can either add to or detract from the flavor profile and quality. There are also legal levels to consider. [6] The level of chloride can be influenced by the distance of the vineyard to the coast. Sulfur dioxide is determined by type of wine, age, and health status. Limits are defined by law as certain levels are not fit for consumption. [4] Density is determined by dissolved solids including alcohol, sugar, and glycerol. More pH levels can push wine to become softer, which is becoming more popular. [3] If a pH is high, a wine is more susceptible to bacterial growth while low pH can cause the tart crisp taste in wines. [11] Sulphates are used as a wine preservative to maintain freshness, but high sulfates can cause some people to have headaches, hives, swelling or stomach pain. [5] Fermentation determines the level of alcohol in the wine. This can affect the taste and bouquet. [1]

3 Exploratory Data Analysis

Exploratory data analysis (EDA) is finding the relationship of the variables in the data. It is during this process that the data is cleaned as the decisions made from the data are only as good as the quality of the data. Without EDA, poor decisions will be made if the data relationships are not understood.

There are many techniques to use during EDA. First, determine what is provided in the data. Digging deeper involves the cleaning of the data, such as missing values or outliers. Once the data has been prepped, the analysis can begin using univariate and bivariate methods. Some of the techniques that were used to clean the data ruled out missing values and removed outliers. Statistical review of the data occurred as well as looking at the correlation of the variables. Scatterplots of each variable were made. Some other items considered were unique values and value counts.

Table 1. Wine Quality Attributes

Attribute	Description
Fixed Acidity	Changes sourness or tartness
Volatile Acidity	Indicates spoilage
Citric Acid	Adds freshness
Residual Sugar	Higher sugar is sweeter
Chlorides	Contributes to saltiness
Sulphur Dioxide	Limits set by law
Density	Concentration of dissolved solids
pH	Lower pH is tart and crisp
Sulphates	Freshness
Alcohol	Alcohol percentage
Quality	Rank from 1 (poor) to 10 (great)

3.1 Feature Engineering

It was not necessary to do any action on missing values as each row and column had a value. For outliers, box plots and histograms were used. Chlorides were heavily skewed due to the number of outliers. From the minimum and maximum sides, 5 percent was removed from each side. Some of the next things reviewed were statistical results, unique values, correlation matrix and heat map, and a pair plot.

Statistical analysis determined that the quality ranges from 3 to 9 making it 7 unique values. Density has the most unique values at 856. The next highest amount was 300 for residual sugar. The average quality is 5.85. Alcohol level ranges from 8 to 14 with an average of 10.58. Residual sugar has the largest range of 0 to 22 with an average of 5.85.

In review of the correlation matrix, citric acid and sulphates have almost zero correlation with quality, the dependent variable. The highest correlation occurred between alcohol and density as well as pH and fixed acidity. Due to these results, citric acid, sulphates, density, and fixed acidity were removed from the list of independent variables to be considered.

The pair plot showed no linear relationship between any pair of variables. Volatile acidity, residual sugar, and chlorides are skewed to the right for their distribution compared to pH, alcohol, and quality that have normal distributions.

	volatile acidity	residual sugar	chlorides	pH	alcohol	quality
count	3914.000000	3914.000000	3914.000000	3914.000000	3914.000000	3914.000000
mean	0.279914	5.845069	0.045925	3.195915	10.583941	5.852325
std	0.102647	4.672021	0.023183	0.151651	1.198902	0.890740
min	0.080000	0.600000	0.012000	2.720000	8.000000	3.000000
25%	0.210000	1.600000	0.035250	3.100000	9.500000	5.000000
50%	0.260000	4.700000	0.042000	3.180000	10.400000	6.000000
75%	0.320000	8.800000	0.050000	3.290000	11.400000	6.000000
max	1.100000	22.600000	0.346000	3.820000	14.050000	9.000000

Fig. 2. Statistics of remaining variables.

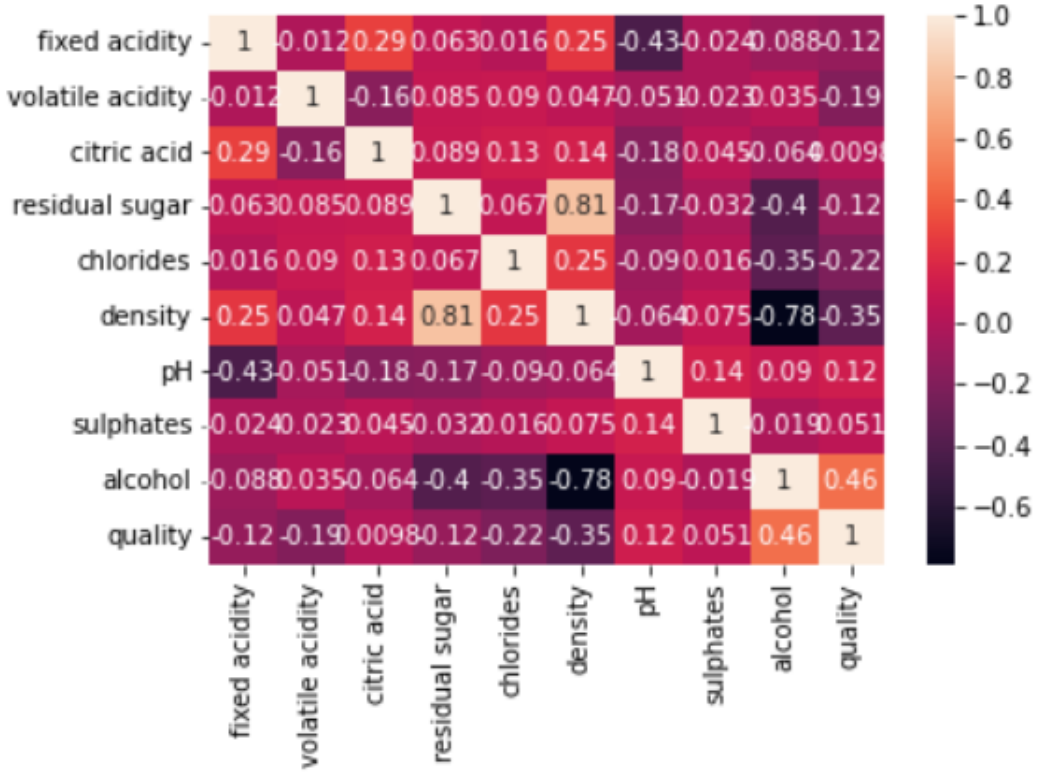

Fig. 3. Correlation Matrix.

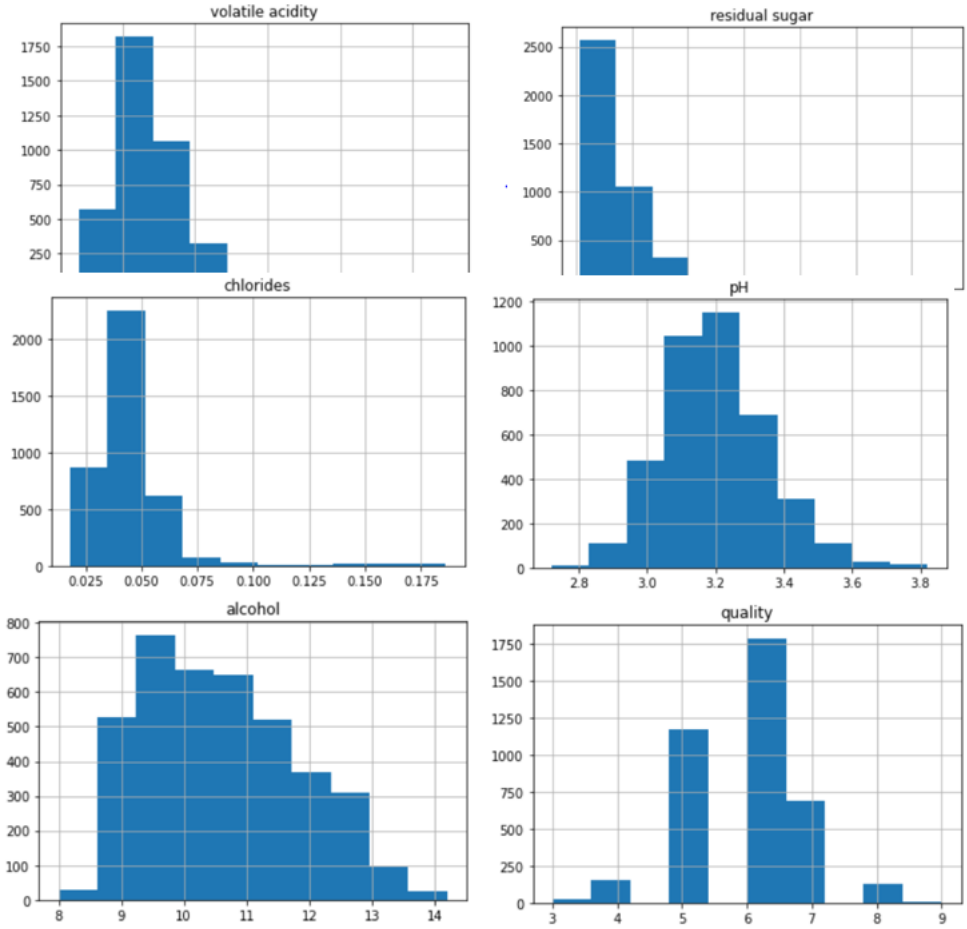

Fig. 4. Histograms of remaining variables.

3.2 Python Coding for Charts

The clickable link for the Python code is: https://github.com/apmolder/Capstone-Final-Project

Python was used for the graphics with the Pandas, Numpy, Scipy, and Seaborn libraries. To drop the duplicates, the code used was drop duplicates. Histograms and box plots were then created using white.hist and white.boxplot. For the outliers in chlorides, quantile .995 and .005 calculations were used. Describe was used for the statistical chart. Two lines of code were used to get the heat map on the correlation matrix - outliers.corr and sns.heatmap. The number of unique values found used nunique.

```
white.drop_duplicates(inplace=True)
white_outliers.describe()

white_outliers = white[(white.chlorides < white.chlorides.quantile(.995)) & (white.chlorides > white.chlorides.quantile(.005))]
white_outliers.boxplot('chlorides')
white_outliers.hist('chlorides')

corelation = white_outliers.corr()
sns.heatmap(corelation, xticklabels=corelation.columns, yticklabels=corelation.columns, annot=True)

data = white_outliers.drop(['citric acid', 'sulphates', 'density', 'fixed acidity'], axis = 1)
sns.distplot(data['volatile acidity'])
```

Fig. 5. Python Code Samples for EDA.

4 Machine Learning Algorithms

The pipeline used for the predictive applications was machine learning algorithms. Most of the steps of this process have been discussed. To summarize, data was collected from Kaggle. It was mostly clean data, but there were some duplicates and outliers that were removed. Using exploratory data analysis, the independent wine variables were reduced to volatile acidity, residual sugar, pH, chlorides, alcohol, and quality. This prepared the data for training and testing of the models before the final analysis.

The data was split into 80 percent training data and 20 percent test data. To initially analyze the models, the training set was used. The test set was used once the models were analyzed. Training set had 3,131 rows compared to 783 rows in the test set.The machine learning algorithms used to analyze the data were Linear Regression, Lasso Regression, Random Forest, and GridSearchCV.

```
#train test splits
X = data.drop('quality', axis=1)
y = data.quality.values
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Fig. 6. Python Code Sample for Training and Test Split.

4.1 Initial Analysis

For the Linear Regression, statsmodels api was imported to get a model summary. It was determined the the R-square value was 0.275 meaning 28 percent of the model was explained. This indicated that it was a poor model fit. For P values less than 0.05, the coefficients are considered significant. With this model, all coefficients are significant except for chlorides. The values of the coefficients indicated that as volatile acidity and chlorides go up, quality goes down. The opposite is true for residual sugar, pH, and alcohol.

```
#linear regression model
import statsmodels.api as sm
X_sm = X = sm.add_constant(X)
model = sm.OLS(y,X_sm)
```

Fig. 7. Python Code Samples for Training Linear Regression.

Table 2. Wine Quality Variables and Coefficients

Variables	Coefficients
Volatile Acidity	-1.8646
Residual Sugar	0.0224
Chlorides	-0.7289
pH	0.5026
alcohol	0.3730

For the next algorithms, the mean absolute error (MAE) was used to analyze which models improved. For Linear Regression, this value was 0.59 compared to Lasso Regression which was 0.68 indicating the Lasso Regression was not an improvement. The next algorithm, Random Forest, was able to be calculated by importing sklearn ensemble Random Forest Regressor. This MAE was 0.59 which means it was about the same as the Linear Regression.

To tune the models, the last algorithm was used. From sklearn model selection, GridSearchCV was imported. Parameters chosen were n estimators, criterion of mse and mae as wells as max features of auto, sqrt, and log2. The best estimators were max features of log2 and n estimators equal to 200. The GridSearchCV MAE was calculated as 0.58. With the smallest MAE, the GridSearchCV provided the best model of all the algorithms for the training data.

```
• #test ensembles
• tpred_lm = lm.predict(X_test)
• tpred_lml = lm_l.predict(X_test)
• tpred_rf = gs.best_estimator_.predict(X_test)
```

Fig. 8. Python Code Samples for Testing Models.

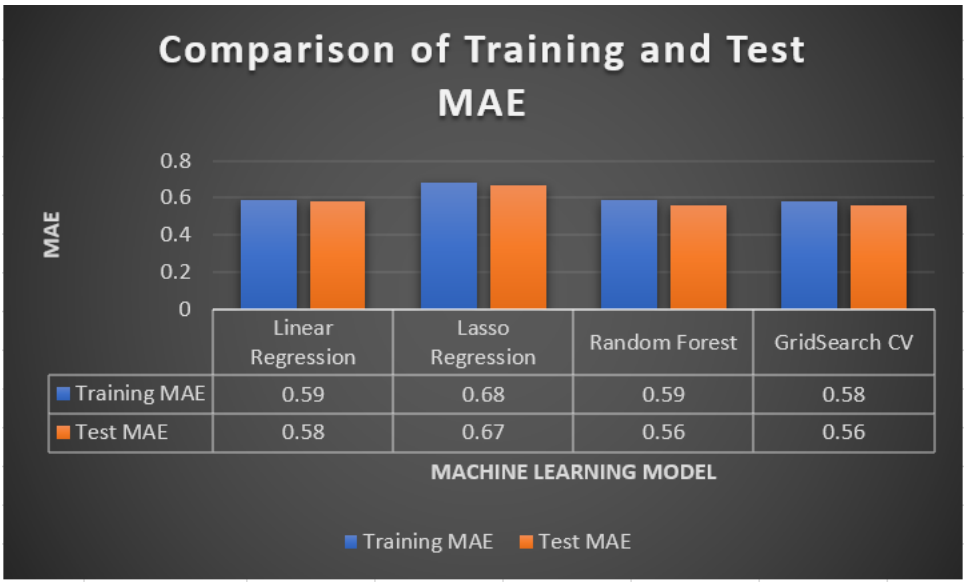

Fig. 9. Comparison of MAE Results with Training and Test Data. (Smaller is better.)

The ensembles were then analyzed with the test data using mean absolute error values. For the Linear Regression model, MAE was similar at 0.58. The worst model with test data was the Lasso Regression at an MAE of 0.67. The final model analyzed calculated an MAE of 0.56 making Random Forest the best model using the test data.

4.2 Secondary Analysis

After the original analysis using MAE as a deciding factor, further analysis was performed looking at Root Mean Square Error (RMSE), Mean Square Error (MSE), and R-squared. It was necessary to import Mean Squared Error and r2 score from sklearn metrics in Python. In the first look at the Linear Regression, R-squared was low at 0.27. Similarly, Lasso Regression was much worse at 0.0. The models with the best R-square values were Random Forest and GridSearchCV. Both had good values at 0.90 which means that 90 percent of the variance was explained by the models. This shows a good fit for the training data.

```
#lasso regression
lm_l = Lasso()
lm_l.fit(X_train,y_train)
print('Results for lasso regression on training data')
print('MAE is ', mean_absolute_error(y_train, y_pred))
print('RMSE is ', np.sqrt(mean_squared_error(y_train, y_pred)))
print('MSE is ', mean_squared_error(y_train, y_pred))
print('R^2 ', r2_score(y_train, y_pred))
```

Fig. 10. Python Code Samples for Training Lasso Regression.

For RMSE, a value of 0.2 to 0.5 is considered a good fit. With this in mind, both the Linear Regression and Lasso Regression were a poor fit as the RMSE values were over 0.75. As with the R-square value, Random Forest and GridSearchCV had 0.28 for the training values. MSE is perfect if the value is zero meaning the lower the value the better. Looking at the training data MSE, Linear Regression had the best fit at 0.57, compared to the other 3 that were all similar around 0.8.

After evaluating the training data, a review of the test data occurred using RMSE, MSE and R-squared. The RMSE for Linear Regression, Random Forest, and GridSearchCV were all close in value to 0.76 versus Lasso Regression which was at 0.89. This indicated that, for the test data, RMSE was not good for any of the models since they were all above 0.5. MSE for those same 3 models was around 0.58 compared to 0.79 for Lasso Regression. That means the test data had better MSE for Linear Regression, Random Forest, and GridSearchCV.

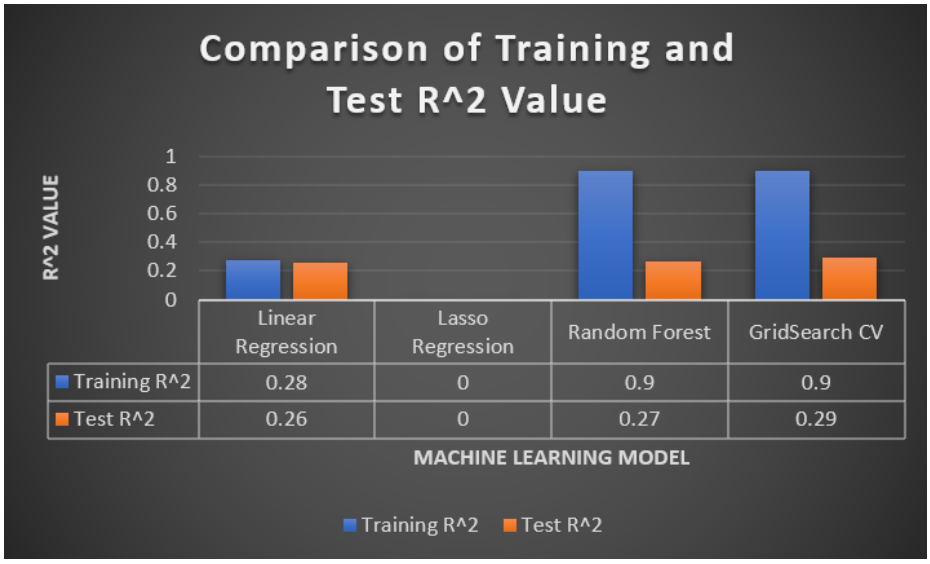

Fig. 11. Comparison of R-square using Training and Test Data. (Larger is better.)

Lastly, R-squared was assessed for the 4 models with the test data. Lasso Regression was the worst with an R-square of zero. The other 3 models were not much better as they all explained less than 30 percent of the variance in the model. Linear Regression was the lowest of the 3 at 0.26, then Random Forest was at 0.27. The best model using the test data ended up being GridSearchCV with R-squared of 0.29.

```
#random forest
rf = RandomForestRegressor()
```

```
#tune models GridSearchCV
from sklearn.model_selection import GridSearchCV
parameters = {'n_estimators':range(10,300,10), 'criterion':('mse','mae'), 'max_features':('auto','sqrt','log2')}
gs = GridSearchCV(rf,parameters,scoring ='neg_mean_absolute_error',cv=3)
gs.fit(X_train,y_train)
```

Fig. 12. Python Code Samples for Training Random Forest and GridSearchCV.

5 Conclusion

The initial project goal was to determine what characteristics of white wine will make a quality wine. Kaggle had a data set to download in csv format. [9] Upon initial review, the data did not have any missing pieces of information. Using Exploratory Data Analysis, duplicates and outliers were removed. It was determined that some features were were not correlated with quality and unnecessary variables were removed. Additionally, others were removed as they were highly correlated with each other. This left volatile acidity, residual sugar, chlorides, pH, and alcohol as the remaining variables to be used in the machine learning algorithms.

The remaining traits determining wine quality make sense based on what each one does for the wine. High volatile acidity indicates spoilage. Sugar determines sweetness. Chlorides can cause saltiness. A tart and crisp taste can be caused by pH. Alcohol level can be a personal preference, but it also impacts the taste greatly. Variable coefficients were calculated with the Linear Regression model further confirming why these variables are important. As volatile acidity and chlorides go down, the quality of the wine goes up. Similarly, residual sugar, pH, and alcohol increasing leads to a higher quality score.

During the training phase, Linear Regression, Lasso Regression, and Random Forest were used. Tuning was done with GridSearchCV on Random Forest. This was because it had the best R-squared value at 0.90. This means that 90 percent of the variance of the model was explained. The other results also indicated Random Forest was the best result.

Table 3. Training Data Results

Training Model	MAE	RMSE	MSE	R^2
Linear Regression	0.59	0.76	0.57	0.28
Lasso Regression	0.68	0.89	0.79	0.0
Random Forest	0.59	0.28	0.08	0.90
GridSearchCV	0.58	0.28	0.08	0.90

Unfortunately, the test data performance was poor on all 4 models. The best was still the GridSearchCV and the worst was Lasso Regression, but the R-squared dropped dramatically to 0.29. RMSE increased from 0.28 to 0.75 on that same model. Similarly, MSE results increased for the training to the test data. In conclusion, GridSearchCV produced the best model, but it was not a good predictor of wine quality with these variables.

Table 4. Test Data Results

Test Model	MAE	RMSE	MSE	R^2
Linear Regression	0.58	0.77	0.59	0.26
Lasso Regression	0.67	0.89	.79	0.0
Random Forest	0.56	0.76	0.58	0.27
GridSearchCV	0.56	0.75	0.56	0.29

6 Limitations

Some initial limitations with the data is that wine brand names were not included in the data. It was purely numerical values of what can change the taste. The project was initially started to learn what makes quality wine, but it would not be possible to buy the quality wine or predict which brands may be the good brands or bad brands. It would be necessary to get a new data set.

For future work, it is possible that the results may improve by doing more Exploratory Data Analysis. For example, the process may show that more variables should be used. The ones that were left out from the original data set included fixed acidity, citric acid, sulphur dioxide, density, and sulphates. Alternatively, additional Machine Learning algorithms could produce better results. It would also be interesting to use other similar data sets. One for red wine has already been found within the same Kaggle download. [9] By using this alternative data set, better Machine Learning results may be achieved.

References

1. Ayza: How alcohol content affects wine, https://ayzanyc.com/blog/how-alcohol-content-affects-wine. Last updated 8/25/2014. (accessed : 10.26.2022)
2. Calwineries: The role of citric acid in wine and winemaking, https://www.calwineries.com/learn/wine-chemistry/wine-acids/citric-acid. (accessed : 10.26.2022)
3. ETS: Density, https://www.etslabs.com/analyses/DEN: :text. Last updated in 2022. (accessed : 10.26.2022)
4. Joze: Analysis of free sulfur dioxide in wine, https://ourvineyardcottage.com/analysis-of-free-sulfur-dioxide-in-wine: :text=Results. Last updated 4/11/2021. (accessed : 10.26.2022)
5. Link, R.: Sulfites, https://www.healthline.com/nutrition/sulfites-in-wine. Last updated 9/9/2019. (accessed : 10.26.2022)
6. Mantech: Chloride in wine by titration, https://mantech-inc.com/wp-content/uploads/2014/07/105-Chloride-in-Wine-by-Titration.pdf. Last updated in 2017. (accessed : 10.26.2022)
7. Nierman, D.: Fixed acidity, https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity. Last updated in 2004. (accessed : 10.26.2022)

8. Puckette, M.: What is residual sugar in wine and where does it come from?, https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/. (accessed : 10.26.2022)
9. Sarahg: Wine datasets, https://www.kaggle.com/datasets/sgus1318/winedata?resource=download. Last updated in 2017. (accessed : 10.18.2022)
10. Vinlab: Volatile acidity - what are the effects on my wine?, https://vinlab.com/blog/2019/08/22/volatile-acidity-what-are-the-effects-on-my-wine/. (accessed : 10.26.2022)
11. Vinny, A.D.: What do ph and ta numbers mean to a wine?, https://www.winespectator.com/articles/what-do-ph-and-ta-numbers-mean-to-a-wine-5035. Last updated 4/15/2009. (accessed : 10.26.2022)