Predicting Wine Quality: Using Machine Learning

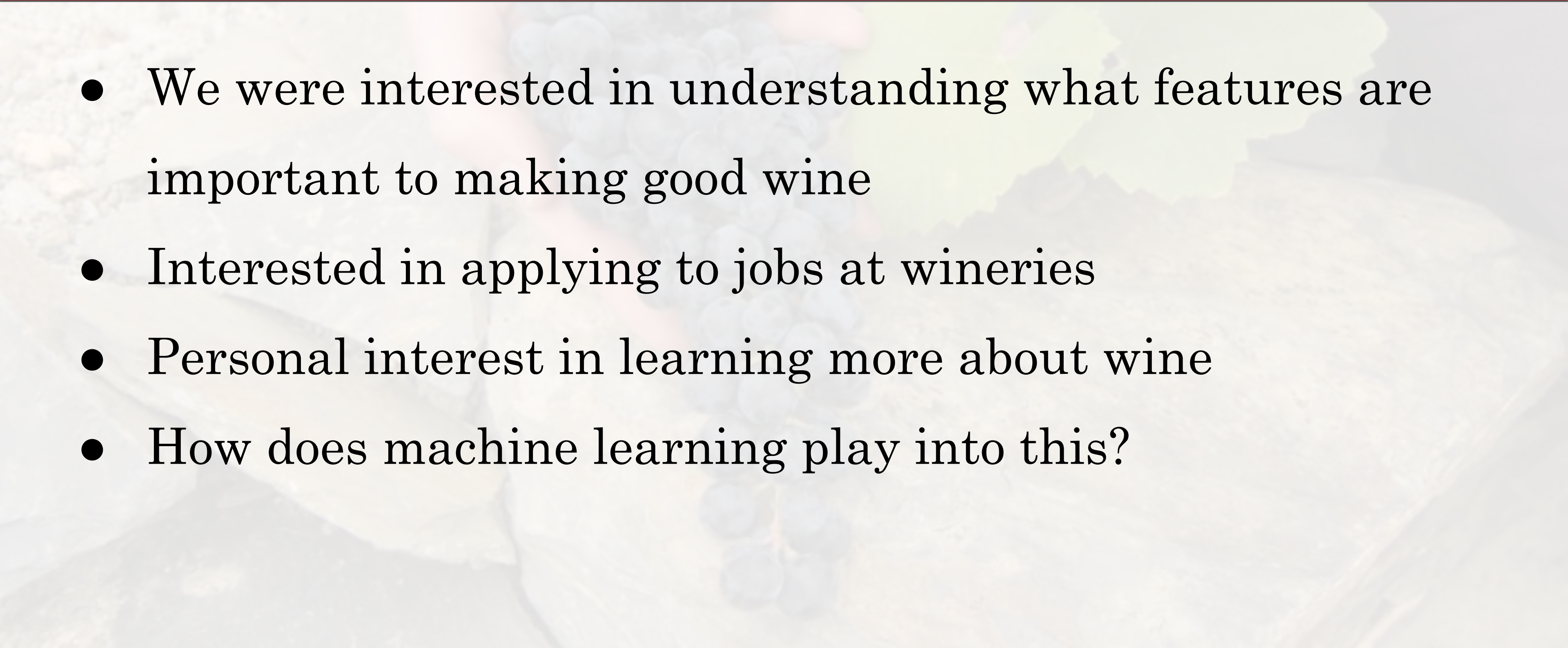
Satvik Ajmera and Kenneth Tanaka

Berkeley Data Analytics Boot Camp

UC Berkeley Extension

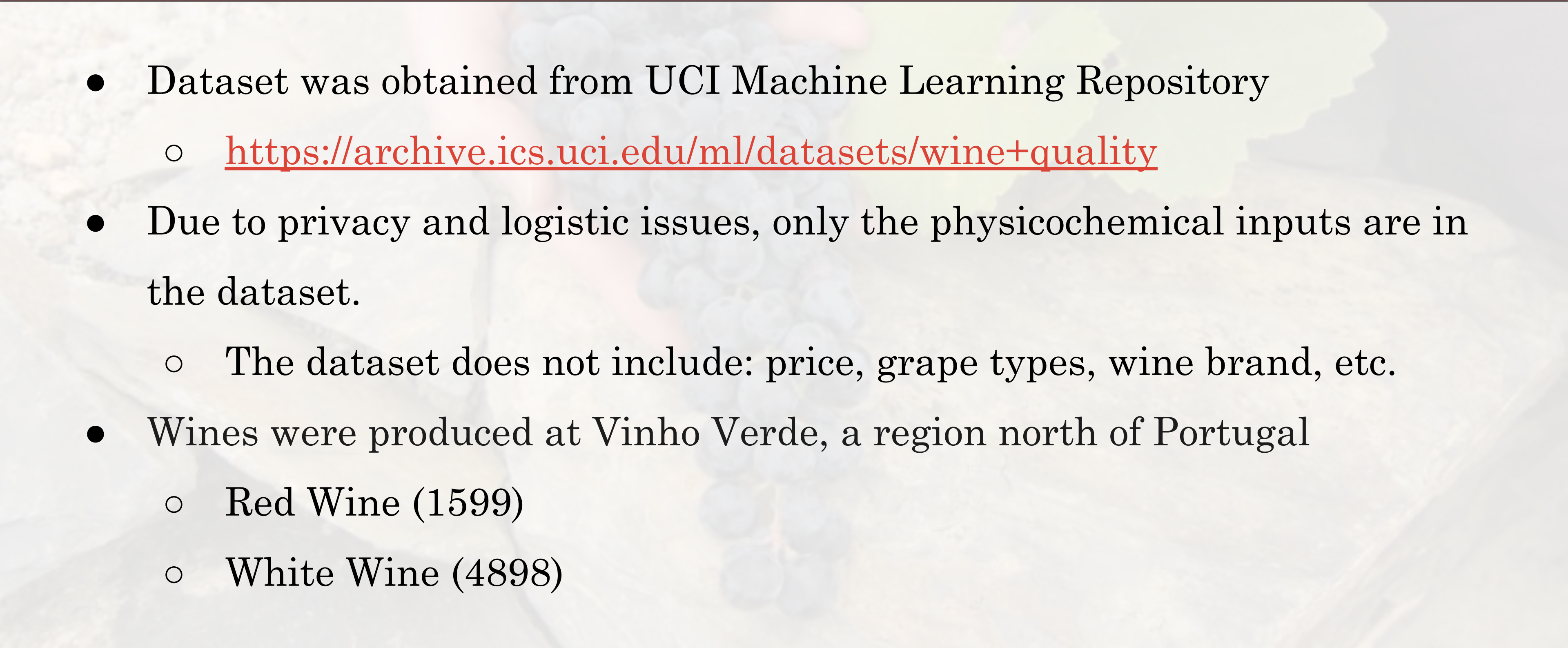
May 9, 2020

Motivation



- We were interested in understanding what features are important to making good wine
- Interested in applying to jobs at wineries
- Personal interest in learning more about wine
- How does machine learning play into this?

About the Data



- Dataset was obtained from UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/datasets/wine+quality
- Due to privacy and logistic issues, only the physicochemical inputs are in the dataset.
  - The dataset does not include: price, grape types, wine brand, etc.
- Wines were produced at Vinho Verde, a region north of Portugal
  - Red Wine (1599)
  - White Wine (4898)

Features of the Wine Dataset

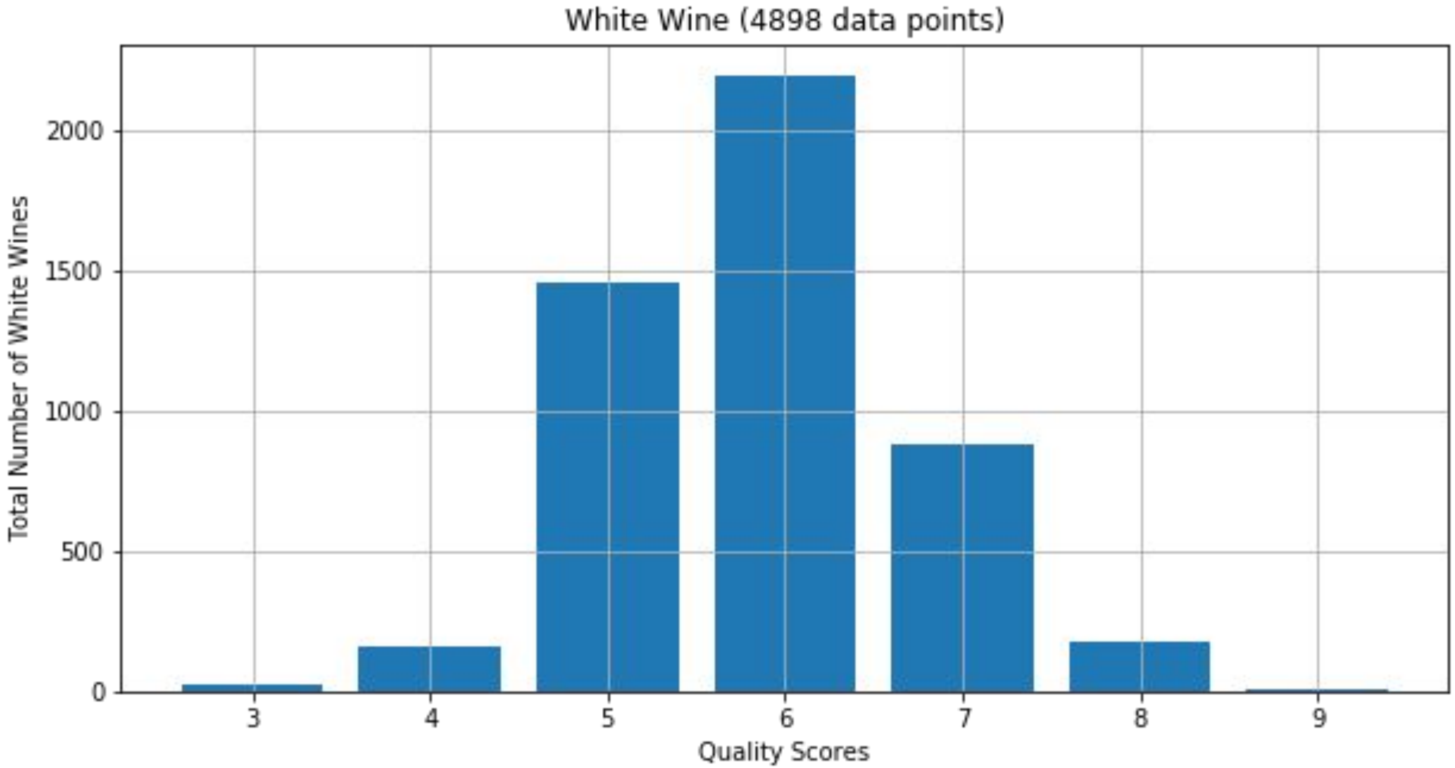
Input variables (based on physicochemical tests):

1. **Fixed Acidity** - a measurement of the total concentration of titratable acids and free hydrogen ions
2. **Volatile Acidity** - caused by bacteria in the wine creating acetic acid
3. **Citric Acid** - acts as a preservative and is added to wines to increase acidity, complement a flavor
4. **Residual Sugar** - natural grape sugars that are left over after fermentation ceases
5. **Chlorides** - indicator of “saltiness”
6. **Free Sulfur Dioxide** - preservative used to protect wine from negative effects of exposure to air
7. **Total Sulfur Dioxide -** is the portion of SO2 that is free in the wine plus the portion that is bound to other chemicals in the wine such as aldehydes, pigments, or sugars
8. **Density** - can be used to measure the alcohol concentration in wines
9. **pH** - values less than 7 are considered acidic, values above 7 are considered alkaline or basic
10. **Sulphates** - salts of sulfuric acid
11. **Alcohol -** the percentage of alcohol content in the wine

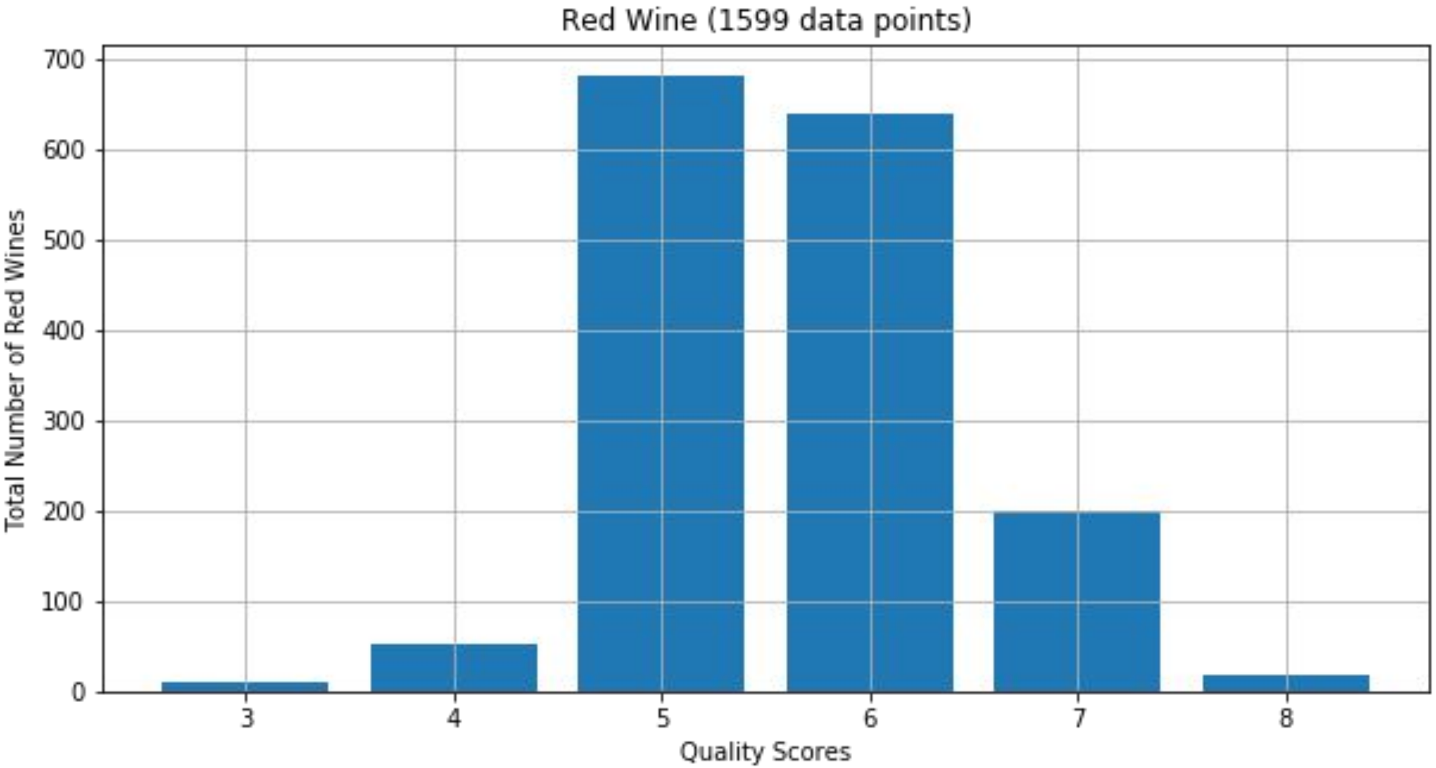
Output variable (based on sensory data):

1. **Quality** (score between 0 - very bad and 10 - very excellent)

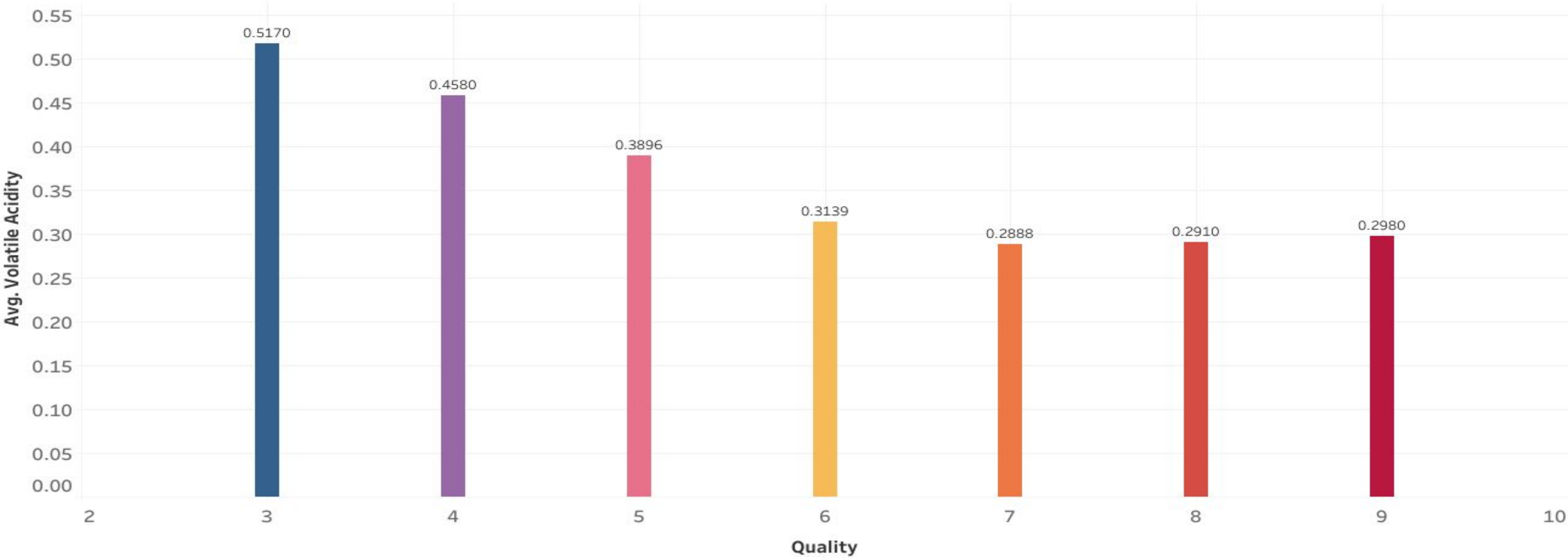
Exploratory Data Analysis (EDA):
White Wine Histogram



Red Wine Histogram

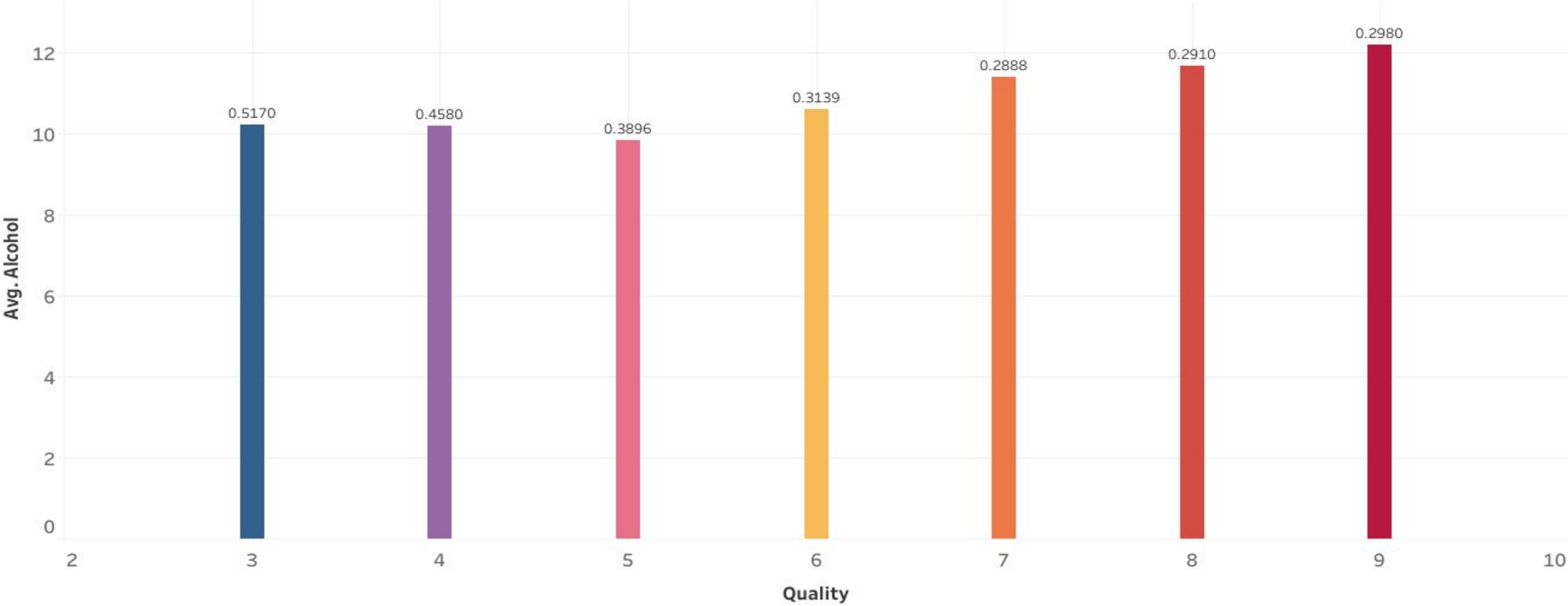


Quality and Volatile Acidity



Higher quality is associated with low volatile acidity levels

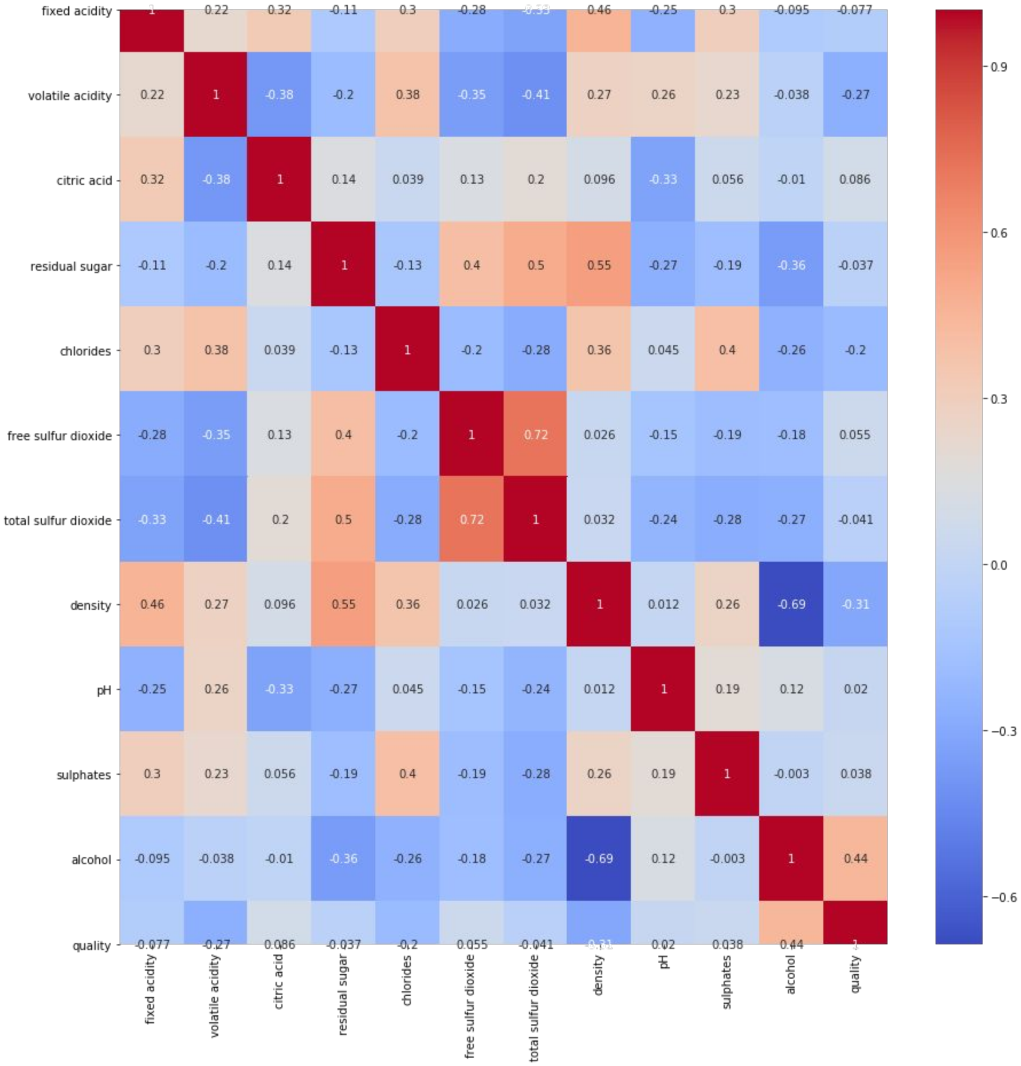
Quality and Alcohol



Higher quality is associated with higher alcohol levels

Red and White Wine: Correlation Heatmap Plot

- Red Wine: 1599 wines
- White Wine: 4898 wines
- Total wines: 6,497 wines



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	0.22	0.32	-0.11	0.3	-0.28	-0.33	0.46	-0.25	0.3	-0.095	-0.077
volatile acidity	0.22	1	-0.38	-0.2	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.038	-0.27
citric acid	0.32	-0.38	1	0.14	0.039	0.13	0.2	0.096	-0.33	0.056	-0.01	0.086
residual sugar	-0.11	-0.2	0.14	1	-0.13	0.4	0.5	0.55	-0.27	-0.19	-0.36	-0.037
chlorides	0.3	0.38	0.039	-0.13	1	-0.2	-0.28	0.36	0.045	0.4	-0.26	-0.2
free sulfur dioxide	-0.28	-0.35	0.13	0.4	-0.2	1	0.72	0.026	-0.15	-0.19	-0.18	0.055
total sulfur dioxide	-0.33	-0.41	0.2	0.5	-0.28	0.72	1	0.032	-0.24	-0.28	-0.27	-0.041
density	0.46	0.27	0.096	0.55	0.36	0.026	0.032	1	0.012	0.26	-0.69	-0.31
pH	-0.25	0.26	-0.33	-0.27	0.045	-0.15	-0.24	0.012	1	0.19	0.12	0.02
sulphates	0.3	0.23	0.056	-0.19	0.4	-0.19	-0.28	0.26	0.19	1	-0.003	0.038
alcohol	-0.095	-0.038	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.003	1	0.44
quality	-0.077	-0.27	0.086	-0.037	-0.2	0.055	-0.041	-0.31	0.02	0.038	0.44	1

What Did We Find in Our Initial Analysis?

- Would adding wine color as a feature into our model improve our model?

- We had unbalanced data (more white wines than red wines). How would this affect our model? How do balance this aspect of our data?

- The red wine quality scores actually ranged from 3 to 8, while the white wine quality scores ranged from 3 to 9.

- Because there are such an uneven number of quality scores, how do we feature engineer to improve accuracy?

Linear Regression : Red and White dataset

Two Datasets: White Wine: 4898 ; Red Wine: 1599
Train/Test: 80/20

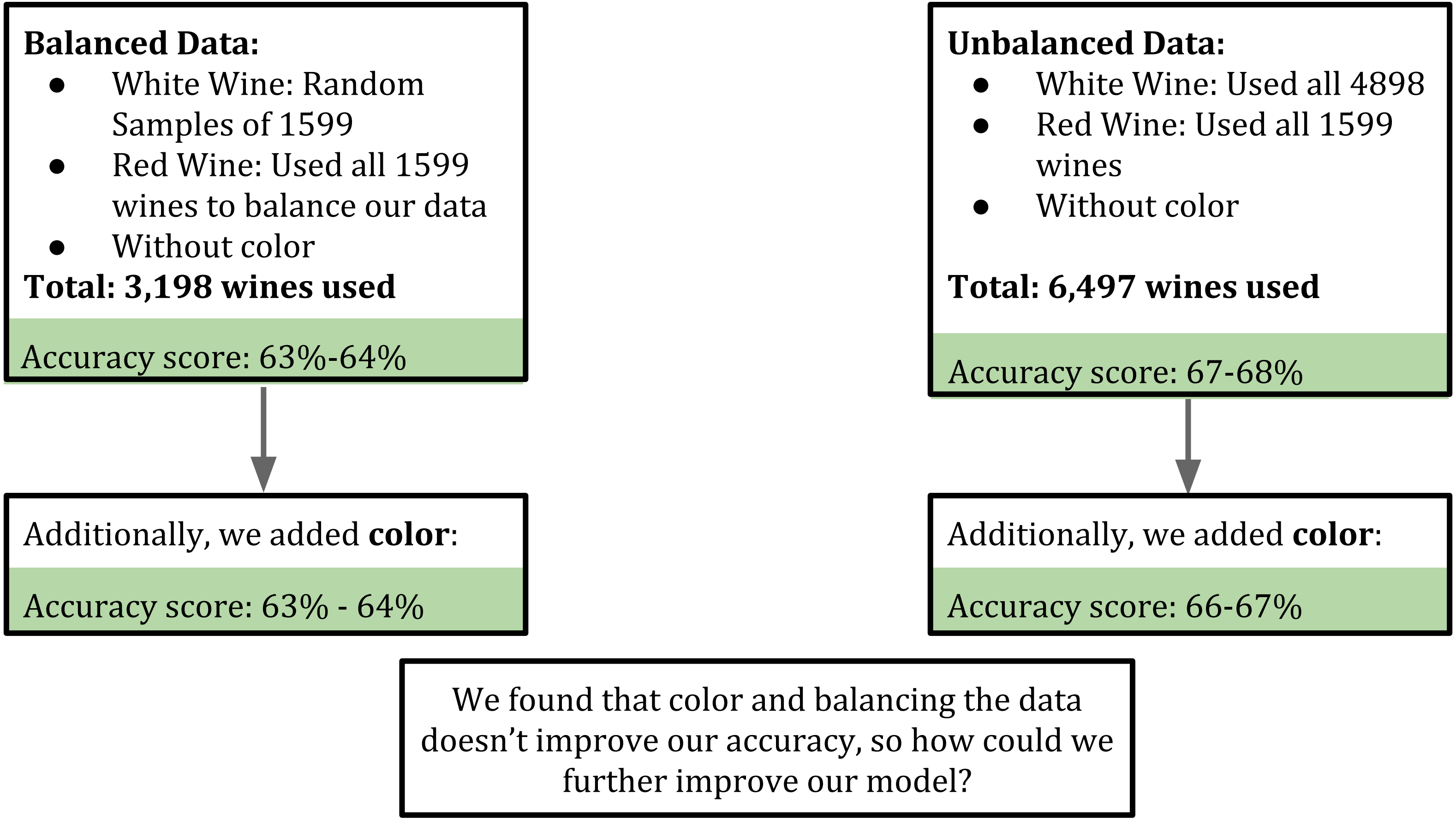
	RED			WHITE	
	MSE	**R2**		**MSE**	**R2**
LINEAR	**0.5861**	**0.3722**		**0.7178**	**0.2727**
LASSO	**0.5906**	**0.3673**		**0.7228**	**0.2676**
RIDGE	**0.5861**	**0.3722**		**0.7178**	**0.2727**
ELASTIC NET	**0.5891**	**0.3691**		**0.7209**	**0.2696**

Selecting a Model

Combined Unbalanced Data: White Wine: 4898 ; Red Wine: 1599
Train/Test: 80/20

	ACCURACY
LOGISTIC	0.47
DEEP LEARNING	0.56
RANDOM FOREST	0.68

Data Manipulation: Random Forest Classifier



Balanced Data:

- White Wine: Random Samples of 1599
- Red Wine: Used all 1599 wines to balance our data
- Without color

Total: 3,198 wines used

Accuracy score: 63%-64%

Additionally, we added **color**:

Accuracy score: 63% - 64%

Unbalanced Data:

- White Wine: Used all 4898
- Red Wine: Used all 1599 wines
- Without color

Total: 6,497 wines used

Accuracy score: 67-68%

Additionally, we added **color**:

Accuracy score: 66-67%

We found that color and balancing the data doesn't improve our accuracy, so how could we further improve our model?

Feature Engineering: Three Bins

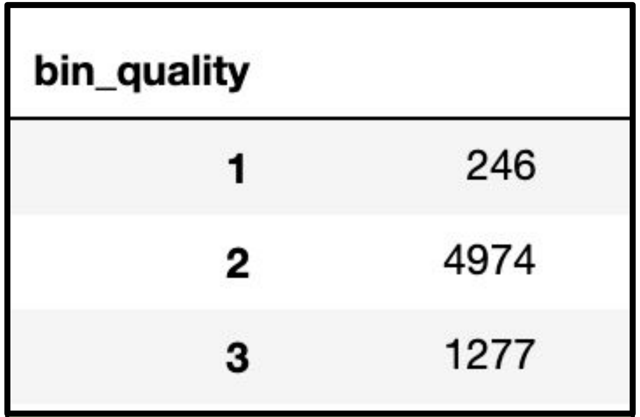
Three Bins:

- Terrible (1), Mediocre (2), and Great (3)
- The quality scores range from 3 to 4, 5 to 6, and 7 to 9.
- White Wine: 4898 wines
- Red Wine: 1599 wines

Total: 6,497 wines used.

This is the highest accuracy score we tested for, because most of the quality scores are 5 and 6. The model is better at predicting mediocre wines.

Number of Wines Per Bin



bin_quality	
1	246
2	4974
3	1277

Accuracy score: 85%

Feature Engineering: Four Bins

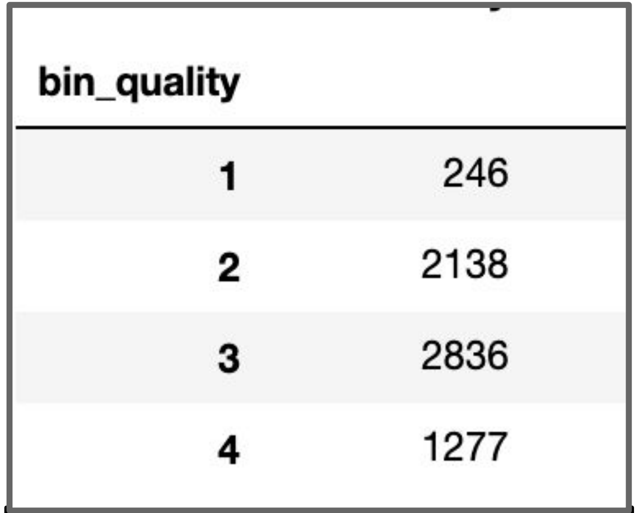
Four Bins (added one additional bin):

- Terrible (1), Mediocre (2), Good (3) and Great(4)
- The quality scores range from 3 to 4, 5, 6, and 7 to 9.
- White Wine: 4898 wines
- Red Wine: 1599 wines

Total: 6,497 wines used.

This experiment had a lower accuracy than before. We binned the quality scores as they are.
Our model may be getting a more realistic accuracy.

Number of Wines Per Bin



bin_quality	
1	246
2	2138
3	2836
4	1277

Accuracy score: 68%

Feature Engineering: Balanced Bins

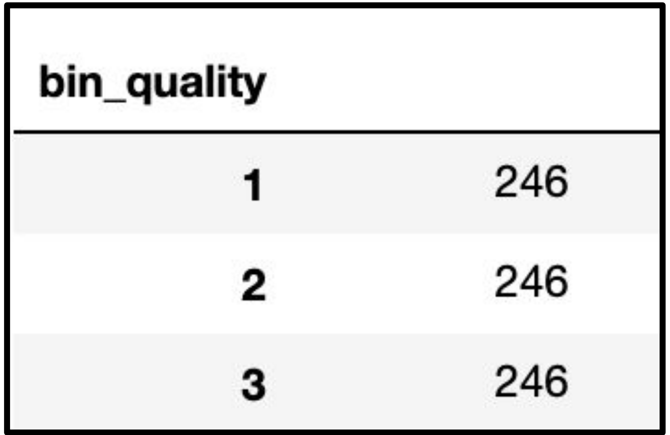
Balanced Bins (three):

- Terrible (1), Mediocre (2), and Great (3)
- The quality scores range from 3 to 4, 5 to 6, and 7 to 9.
- Same number of wines used for each bin
- Indifferent to wine color.

Total: 738 wines used.

When compared to unbalance bins, accuracy was lower. When compared to four bins, accuracy was higher. (Accuracy could be easier to achieve when there is a bin with higher quantity of wines and or when there are fewer bins to choose from.)

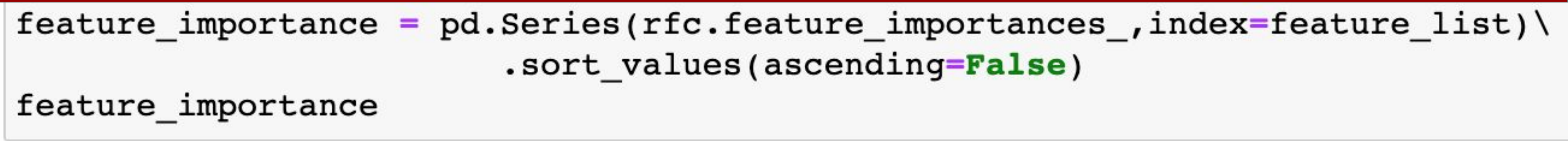
Number of Wines Per Bin



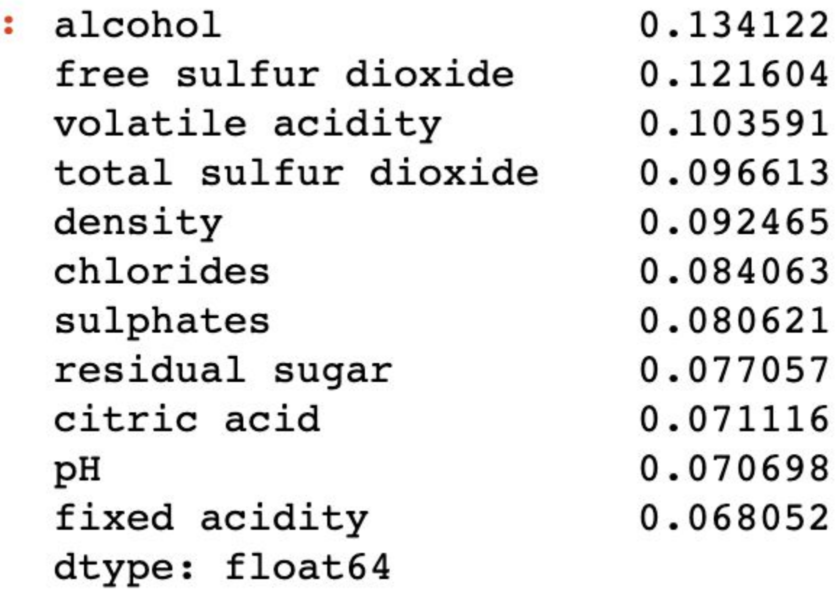
bin_quality	
1	246
2	246
3	246

Accuracy score: 67%

Feature Importance: Feature Engineering Challenges



```
feature_importance = pd.Series(rfc.feature_importances_,index=feature_list)\
                        .sort_values(ascending=False)
feature_importance
```



```
alcohol                 0.134122
free sulfur dioxide     0.121604
volatile acidity        0.103591
total sulfur dioxide    0.096613
density                 0.092465
chlorides               0.084063
sulphates               0.080621
residual sugar          0.077057
citric acid             0.071116
pH                      0.070698
fixed acidity           0.068052
dtype: float64
```

- Alcohol percentage is determining factor of wine quality relative to the other 10 features
- For the most part, for our experiments, this is how the feature importances were ranked.
- Overall, even distribution of feature importance

Conclusion

- Random Forest model resulted in the highest accuracy
- Normalize dataset showed no accuracy improvement
- Adding color as an input showed no accuracy improvement
- Feature engineering did impacted accuracy:
  - Number of bins - fewer bins increased accuracy
  - Unbalanced bins - increased accuracy; bias to “good”
  - Balanced bins - lower accuracy
- Quality is subjective and requires a human expert

Further Exploration

- Improve Dataset
  - More “bad” and “great” quality
  - Other Countries
  - Type of wine
  - Name of winery
  - Price
- Only consider select features in prediction instead of using all
- Challenging to find similar datasets to further test model, because trade secrets?
- Create website for user input
- Explore other machine learning techniques

Thank you!

Citations

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. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016
[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf
[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib