Predicting Wine: Color and Quality

Brad Schaeffer & Connor Beaton

Related Works and Other Discoveries

Although no research directly mimics the dataset / research we found, there has been similar research conducted on the chemical structure of red / white wine.

- Similarity of Red / White Wine and Polyphenols:
  - When making white wine, if you let grape skin have contact with alcohol, then polyphenols grant the white wine red wine like antioxidant properties
- Comparisons of Antioxidant Properties of Red / White Wine
  - In comparison of red / white wines, red have significantly more phenols than white wines
- Antiatherogenic Effects of Red / White Wines
  - Studies have been done to compare the level of platelet activating factors in red / white wines, which have been shown to have anti-inflammatory properties. In this study, red wines had more platelet activating factors.

Our Question

Within the scope of this project, our goal is:

- Find the chemical components that have large effects on type of wine(color) and quality
- Use these factors to predict the type of wine or quality using machine learning algorithms including, KNN, Decision Trees,Random Forests, Linear and Logistic regression

Wine Data

	type	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6
1	white	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6
2	white	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6
3	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
4	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6492	red	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	NaN	11.2	6
6494	red	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

6497 rows × 13 columns

Cleaning Up

```
nan_value = float("NaN")
newdata = wine_data
newdata.replace("", nan_value, inplace=True)
newdata.dropna(axis = 0 ,inplace = True)
```

Wine Stats

Total Number of wines = 6457

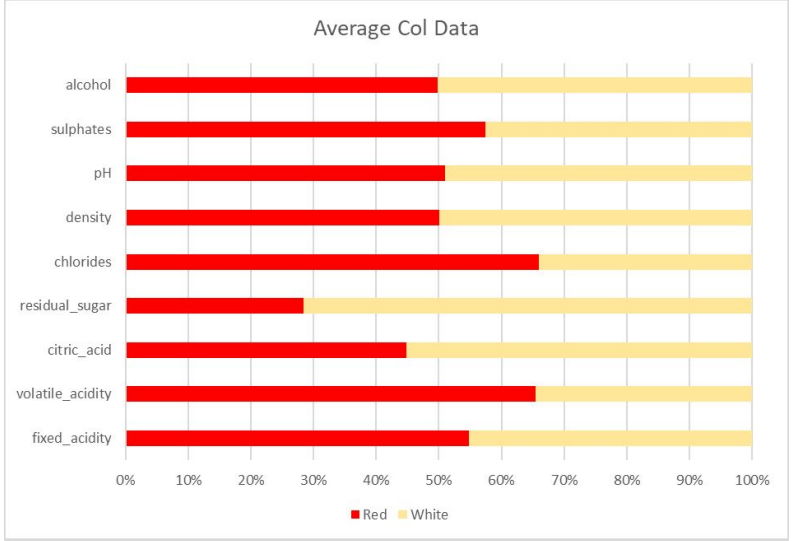
13 Columns of data

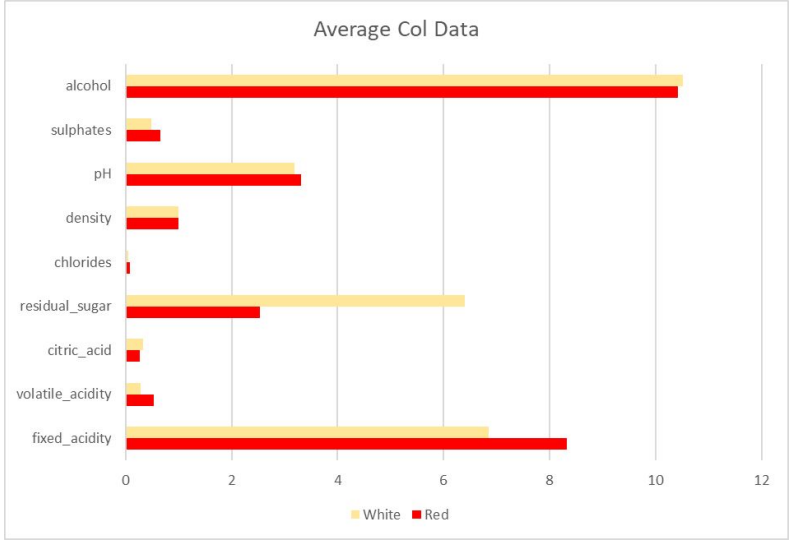
Number of white wines = 4898 (75.86%)

Number of red wines = 1559 (24.14%)

Comparing with Averages

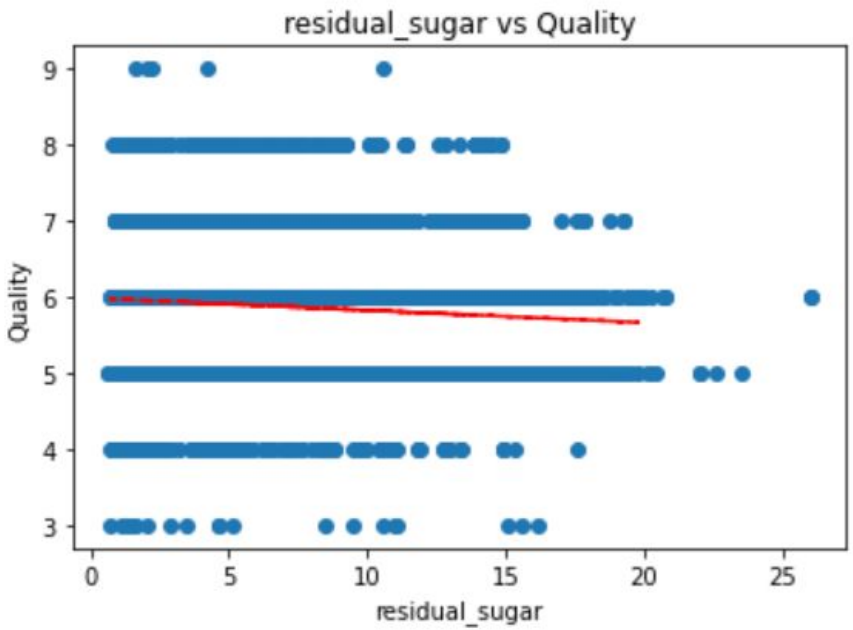
- Fixed Acidity
  - White: 6.854
  - Red: 8.322
- Volatile Acidity
  - White: 0.278
  - Red: 0.527
- Residual Sugars
  - White: 6.393
  - Red: 2.538
- Average Chlorides
  - White: 0.045
  - Red: 0.087
- Sulfur Dioxide
  - White: 138.360
  - Red: 46.467



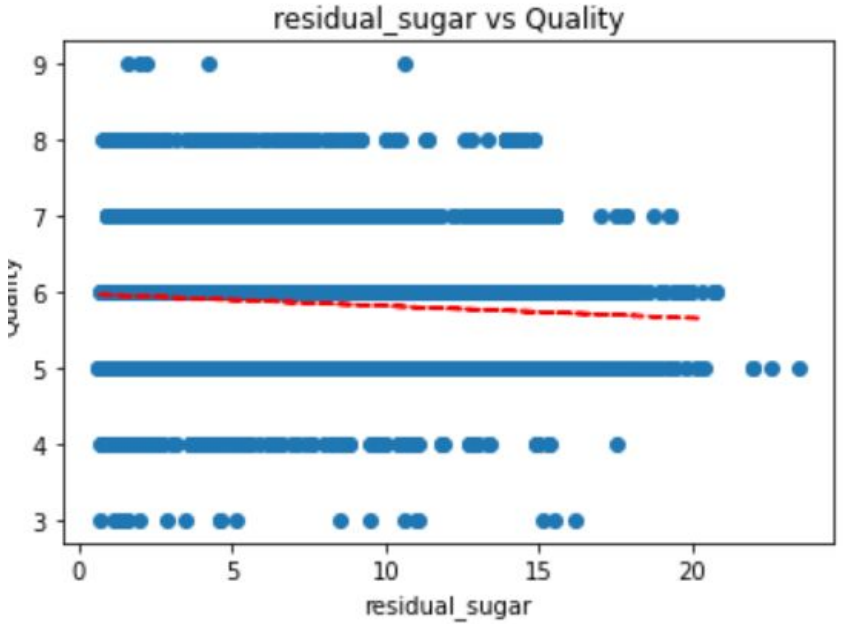


Linear Regression -- Residual Sugar vs. Quality

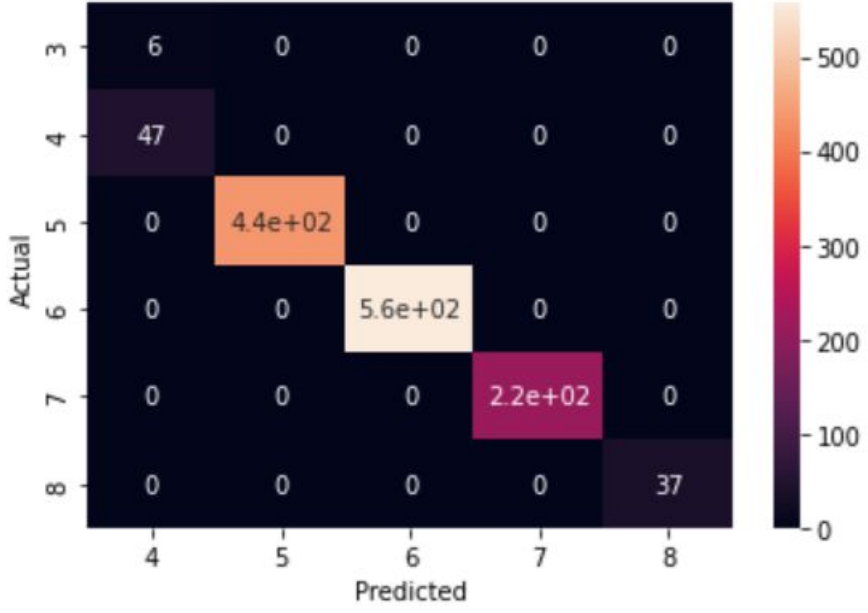
Red



White



Logistic Regression -- Type vs Quality



Type Column Changed

- White = 0
- Red = 1

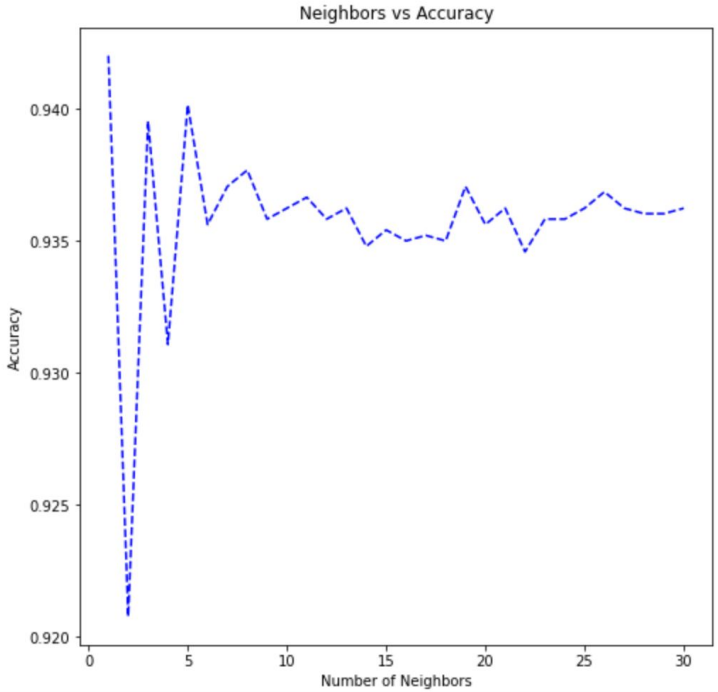
```
wine_data = wine_data[["type", "quality"]]
wine_data = wine_data.dropna()
wine_data["type"] = wine_data["type"].apply(lambda x: 0 if x == "white" else 1)
```

KNN Algorithm

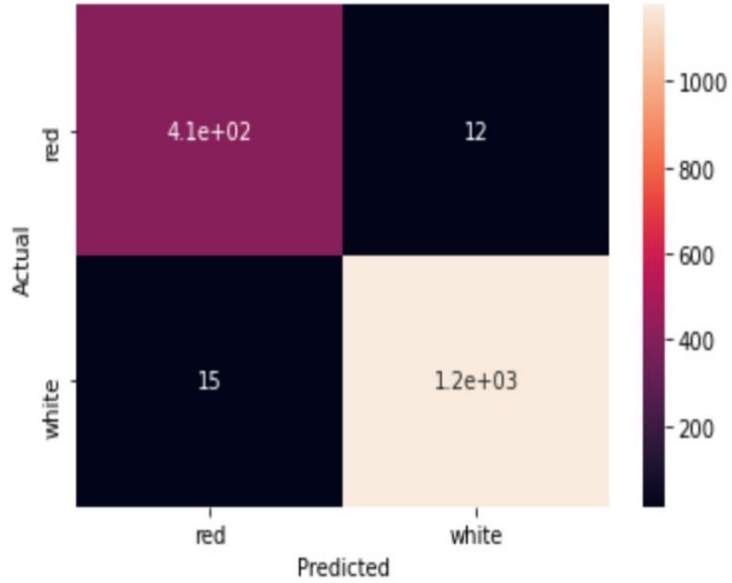
```
drop_type = newdata.drop("type" , axis = 1)
X_train, X_test, y_train, y_test = train_test_split(drop_type,newdata["type"],test_size = .25)
acc = []

#5 Fold cross validation
for i in range(1,31):
    knn = KNeighborsClassifier(n_neighbors=i)
    cross_val_acc = sum(cross_val_score(knn,X_train,y_train,cv=5))/5
    acc.append(cross_val_acc)

plt.figure(figsize=(8,8))
plt.xlabel("Number of Neighbors")
plt.ylabel("Accuracy")
plt.title("Neighbors vs Accuracy")
plt.plot(range(1,31),acc,color="blue",linestyle="dashed",markerfacecolor="red",markersize=10)

k_star = np.argmax(acc)+1
print("Highest performing k value occcurs at:",k_star)

knn = KNeighborsClassifier(n_neighbors=k_star)
knn.fit(X_train,y_train)
pred = knn.predict(X_test)

## compare pred vs y_test
print("The accuracy is: " + str(sum([1 for i in zip(pred,y_test) if i[0]==i[1]])/len(pred)))
```

KNN Algorithm Results

```
Highest performing k value occcurs at: 1
The accuracy is: 0.9467821782178217
```

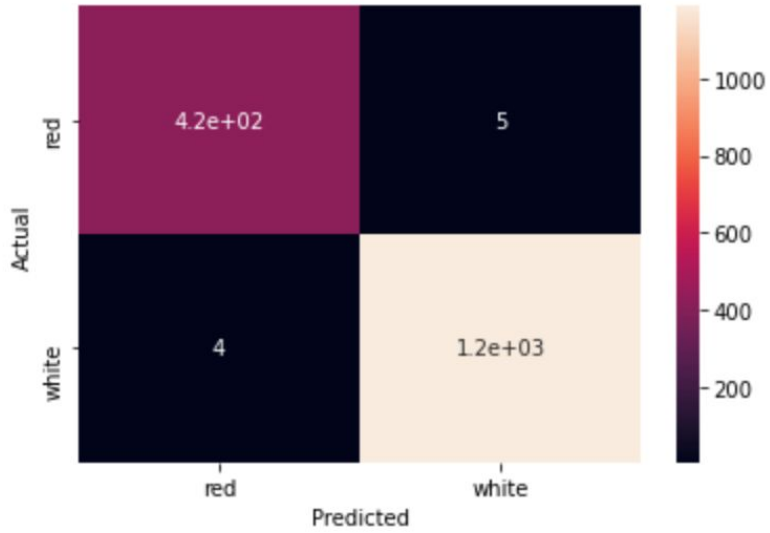


Decision Tree




```
Tree took 0.004595041275024414 amount of time.
0.9832920792079208
```

Random Forest



Forest took 0.04619884490966797 amount of time.
0.994430693069307

Conclusions

- The Wine data set was interesting to work with, and had a lot of columns , which varied the data. There were some NaN values which needed cleaning, but that was the only special attention required.
- Using Linear Regression, we found that although the White wine has slightly more sugar, the quality for both averaged around 6
- Logistic Regression, when separated by type, did not tell us much about the quality of the wine.

Conclusions

- Using KNN, we found that accuracy was not poor but did vary wildly depending on # of neighbors, and would vary a fair amount per run.
- When comparing Decision Trees vs a Random Forest, we found that there was a slight accuracy increase for a slight time increase, although the scale of the time increase was not large for our data set.
- Lastly, the acidity and the chloride content were the largest determining factors to predict red / white wine.

Thank You!

Sources:

- Similarity of Red / White Wine Over Polyphenols
  - https://pubs.acs.org/doi/abs/10.1021/jf001378j?casa_token=N1jVoYfXRFgAAAAA%3AKTIGhVK7NVz9MlSgpSAmSkdxytkQhjetyeo-YGzTaOLONeFYRZF5QdWK0znywfSmr5u6kDLguiHK5Q&
- Antioxidant Effectiveness Between Red / White Wine
  - https://pubs.acs.org/doi/pdf/10.1021/jf00050a027?casa_token=_cZMT8_0KaYAAAAA:VyUT5Kezp21xUTuBZAt5mJkZw_nHo86N0c2inR4vQ01gX-KzzgWjARc_v9fFDYGIvESdbavY8CX8yg
- Antiatherogenic Effects of Red / White Wine
  - https://www.sciencedirect.com/science/article/abs/pii/S1388198103000660