Introduction / Problem Statement

Our data science team has been contracted by a white wine company to help them understand which variables impact the quality of their white wine. The company has collected data on various characteristics of their white wines, such as alcohol content, acidity, and sugar levels, as well as quality ratings for each wine. However, they are unsure which variables are most important in determining the quality of their white wine and want our team to provide insights.

More specifically, the goals of the project are to:

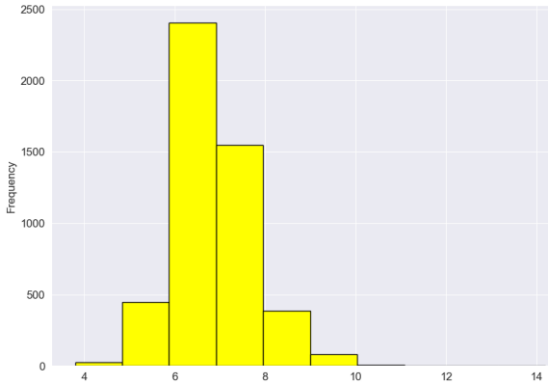
- Identify the key variables that significantly impact the quality of the white wine.
- Determine how well these variables can predict the quality of the wine.
- Provide actionable recommendations to the company on how to optimize their white wine production process to improve quality.

Here are some of the features in the dataset:

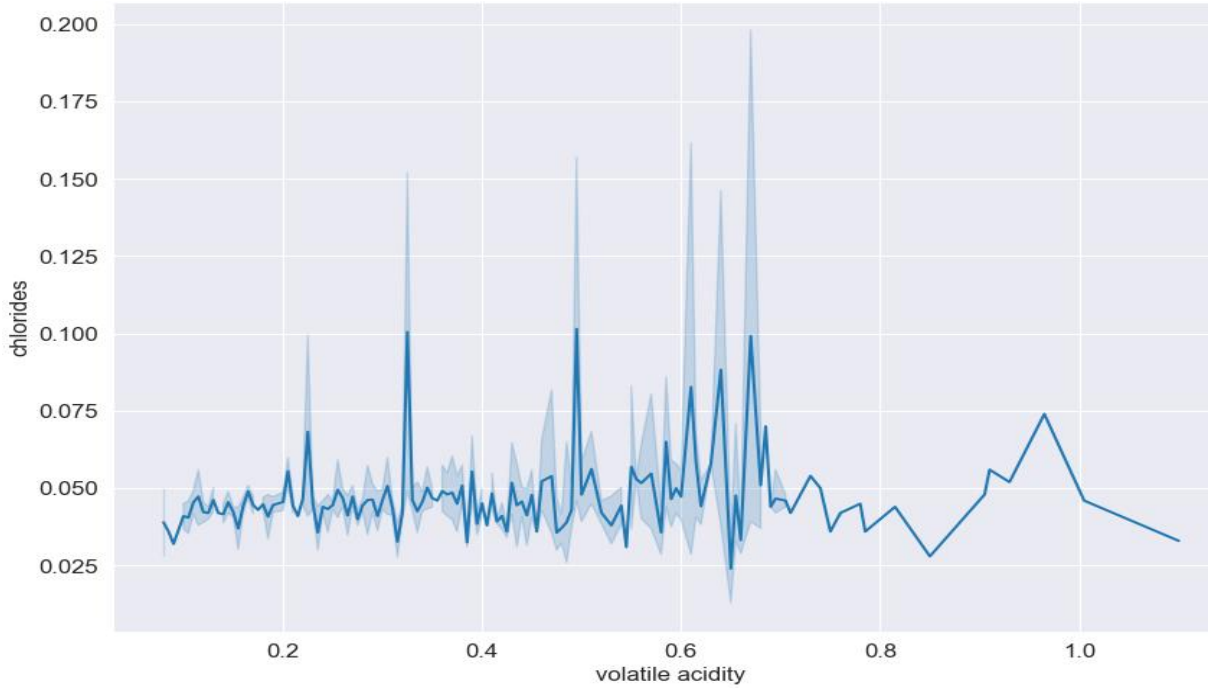
Fixed acidity: the amount of acids that do not easily evaporate in a wine, affecting its taste and acidity.
Volatile acidity: the amount of volatile acids in a wine, which can cause an unpleasant, vinegar-like taste and aroma.
Citric acid: the amount of citric acid in a wine, which can affect its acidity and fruitiness.
Residual sugar: the amount of sugar left in a wine after fermentation, which affects its sweetness.
Chlorides: the amount of salt in a wine, which can affect its taste and preservation.
Free sulfur dioxide: the amount of SO2 gas present in a wine, which acts as a preservative and prevents oxidation.
Total sulfur dioxide: the total amount of SO2 (free + bound) present in a wine.
Density: the mass of a wine per unit volume, which can indicate its alcohol and sugar content.
pH: a measure of the acidity or basicity of a wine.
Sulphates: the amount of sulfates in a wine, which can affect its taste and preservation.
Alcohol (%): the percentage of alcohol in a wine, which affects its body and flavor.
Quality: a score assigned by wine experts based on the overall quality of a wine, taking into account its various characteristics.

Methods / Implementation

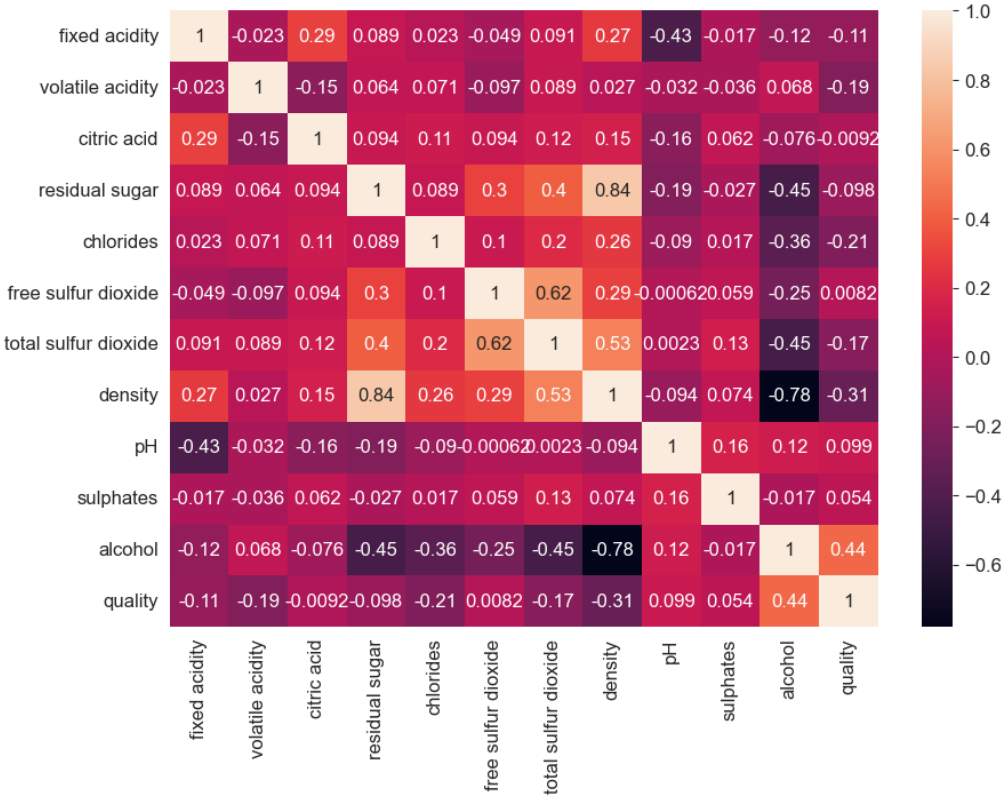
We began by analyzing the relationships between certain features to see if they were correlated with each other.



The distribution of fixed acidity appears relatively normal and most of the fixed acidity values are between 6 and 7. We can also see that the minimum fixed acidity value was around 4 and the maximum was around 14.



Volatile acidity and chloride content appear to have no correlation with each other. However, certain volatile acidity values are associated with a spike in chloride content. As we can see, volatile acidity values of 0.32, 0.5, and 0.67 correlate to about 0.1 chlorides.



Although most of the variables appear to have a weak negative or no correlation with the quality of white wine, alcohol has a moderately strong positive correlation with the quality of white wine. This means that alcohol content is more significant in predicting the quality of white wine since it has the strongest positive correlation to quality than the other variables. The heat map also reveals that there is a relatively weak positive correlation between white wine density and fixed acidity. Additionally, there is a relatively strong positive correlation between white wine density and total sulfur dioxide. However, there is a strong negative correlation between white wine alcohol content and density.

Results

After binning the features by quantile and fitting each model, the random forest was the most accurate model, followed by the decision tree and followed by naïve bayes. The accuracy scores were 0.66, 0.55, and 0.48 respectively. To achieve better performance, we could use different binning techniques or different preprocessing methods for converting continuous data to categorical data.

Discussion / Explanation

The random forest ended up being the only model that can moderately predict whether a patient with heart failure will survive or not based on certain features. This can be useful for the white wine company in making decisions about their wine formula. When selecting a model, it is important to consider the nature of the problem, the size of the dataset, the type of data, and the desired outcome. In this case, we have a classification problem with a relatively large dataset of 4898 observations and 12 features. We aim to predict the quality of white wine based on certain features. Based on these considerations, we chose to use random forests, decision trees, and Naive Bayes. Decision trees and Naive Bayes are models that can handle both categorical and continuous data, which is important when dealing with complex datasets, such as the white wine data given. Decision trees are flexible in terms of model complexity and can capture both linear and nonlinear relationships between variables. They create a clear decision path with a series of if-else statements, making them easy to interpret. However, decision trees tend to overfit the training data, resulting in poor performance on unseen data. Naive Bayes, on the other hand, is a simple yet effective probabilistic model that is based on the Bayes theorem. It assumes that the features are conditionally independent given the target variable, which can be a limitation when dealing with correlated features. Random forests, which are an ensemble of decision trees, combine multiple decision trees to reduce overfitting and improve performance on unseen data. They can handle both categorical and continuous data and are suitable for datasets of this size. Random forests can capture nonlinear relationships between variables and are less prone to overfitting than individual decision trees. They can also provide feature importance scores, which can be used to identify the most important features in predicting wine quality. However, random forests are less interpretable than decision trees and Naive Bayes, as they involve multiple decision trees working together.

Here are the classification reports for each model:

	precision	recall	f1-score	support
3	0.000000	0.000000	0.000000	7.000000
4	0.666667	0.100000	0.173913	40.000000
5	0.679814	0.687793	0.683781	426.000000
6	0.637592	0.776946	0.700405	668.000000
7	0.698492	0.496429	0.580376	280.000000
8	0.750000	0.306122	0.434783	49.000000
accuracy	0.659864	0.659864	0.659864	0.659864
macro avg	0.572094	0.394548	0.428876	1470.000000
weighted avg	0.662930	0.659864	0.646209	1470.000000

Random forest

Accuracy could definitely be improved. We binned by quantiles, meaning that each bin receives 10% of the data, but adding more bins and using a different binning technique could improve the accuracy. Still, there doesn't appear to be overfitting to the training data.

	precision	recall	f1-score	support
3	0.000000	0.000000	0.000000	7.00000
4	0.148936	0.175000	0.160920	40.00000
5	0.565217	0.579812	0.572422	426.00000
6	0.608764	0.582335	0.595256	668.00000
7	0.519298	0.528571	0.523894	280.00000
8	0.333333	0.346939	0.340000	49.00000
9	0.000000	0.000000	0.000000	0.00000
accuracy	0.549660	0.549660	0.549660	0.54966
macro avg	0.310793	0.316094	0.313213	1470.00000
weighted avg	0.554511	0.549660	0.551884	1470.00000

Decision tree

Random forest did 10% better. Highest precision and f1-score were associated with wine quality 6 Looking at the confusion matrix, it appears that the decision tree had a good portion of incorrect predictions around the labels of 2 and 3 representing wine qualities of 5 and 6 on the wine quality scale from 3 - 9.

	precision	recall	f1-score	support
3	0.000000	0.000000	0.000000	7.000000
4	0.125000	0.025000	0.041667	40.000000
5	0.496479	0.661972	0.567404	426.000000
6	0.558577	0.399701	0.465969	668.000000
7	0.372881	0.550000	0.444444	280.000000
8	0.000000	0.000000	0.000000	49.000000
9	0.000000	0.000000	0.000000	0.000000
accuracy	0.478912	0.478912	0.478912	0.478912
macro avg	0.221848	0.233810	0.217069	1470.000000
weighted avg	0.472134	0.478912	0.461968	1470.000000

Naïve bayes

Naive bayes was outperformed by both random forest and decision tree. Random forest and decision tree classifiers can outperform naive Bayes classifiers in situations where the relationship between features and target variables is complex or nonlinear. This is because decision trees and random forests can capture complex interactions and non-linearities between features and target variables, while naive Bayes classifiers assume that features are independent of each other, and that the relationship between features and target variables is linear.