Model Used

I used Random Forest Classifier because
1. Very large data
2. High dimensionality
3. Many Outliers
4. Multi-Class Target

I used 200 n_estimators since we had a lot of features with less correlation with the target variable i.e. variety and used 'sqrt' feature with bootstrapping turned effective

Random forest improves on bagging because it decorrelates the trees with the introduction of splitting on a **random** subset of features. This means that at each split of the tree, the model considers only a small subset of features rather than all of the features of the model. This feature helped in solving the problem of lot of categorical variables with outliers.

Features Extracted

Apart from the given features:

I used feature engineering for extracting some important features like
1. points*winery (for seeing whether points given for some wines is high or low)
2. user*points (the problem when some users assign more/less points for exact same details)
3. Price*winery (some winery produces costly wines)
4. province* price*points (for dependency on price and province)

Model Accuracy in train – 66.08%
Can be further improved using ensemble methods like bagging and boosting and developing more features and removing less correlated features.

All important visualizations which are used by me are available in python notebook.

Actionable insights:

1. Bordeaux Style Red Blend is high prices (Always an outlier)
2. France produces costliest wines.
3. Most data (~20%) is of Pinot-Noir which is relatively cheap (18 units)
4. 19.85% of the price of wines that were present in data lies in a range of (19-30 units)
5. USA and France produce 51% of wines given in data.