**Contents**

[**Approach** 2](#_Toc40029935)

[**To build the model:** 2](#_Toc40029936)

[**1.** **Descriptive Statistics** 2](#_Toc40029937)

[**2.** **Feature Extraction & Pre-processing** 2](#_Toc40029938)

[b. Tree Based Decision Tree Classifier 2](#_Toc40029940)

[c. Random Forest Classifier 2](#_Toc40029941)

[**4.** **Model Selection** 2](#_Toc40029942)

[**5.** **Optimization of Model** 3](#_Toc40029943)

[a. Changing Attribute selection measure and max\_depth. 3](#_Toc40029944)

[b. Plotting Accuracy at each combination. 3](#_Toc40029945)

[3](#_Toc40029946)

[I built the final model with the entropy criterion and max depth of 27. 3](#_Toc40029947)

[**6.** **Prediction** 3](#_Toc40029948)

[Finally, I Made the predictions on the test set. Which is loaded in the zip file predicted.csv 3](#_Toc40029949)

[**Insights** 4](#_Toc40029950)

[Frequency of Reviews with Different Varieties of Wine 4](#_Toc40029951)

[Average Wine Prices in Each Variety 4](#_Toc40029953)

[Average Wine Prices based on its country" 5](#_Toc40029954)

[Distribution of Prices and score points 5](#_Toc40029955)

[Trend Line between Prices and Score Point 5](#_Toc40029956)

**Conclusion**

Based on The Problem Statement I have built a predictive model using the features provided in the dataset Which has an average Accuracy of 93% and F1\_score of 92%.

**Approach**

# **To build the model:**

## **Descriptive Statistics**

## **Feature Extraction & Pre-processing**

1. I combined features extracted from the categorical and text-based columns (Review Title, Review Text) with the numerical column (Price, points) combined them into a sparse matrix.
2. Then included them into the model with our labels to train the model.
3. I trained the model with 70% of existing training data and tested it with the 30% labels which gave the accuracy 93%.

## **Model Building**

1. Probability based Multinomial Naïve Bayes model

Accuracy: 0.79

F1\_score: 0.79

Training Time: 2 secs

1. Tree Based Decision Tree Classifier

Accuracy: 0.93

F1\_score: 0.93

Training Time: 7.2 sec

1. Random Forest Classifier

Accuracy: 0.9303878632437271

F1\_score: 0.9247663532996585

Training Time: 44.7 sec.

## **Model Selection**

Between Multinomial Naive Bayes, Random Forest, and Decision tree. both Random Forest and Decision tree have equal metrics so we will move forward with our prediction Using Decision Tree because of its lesser training time.

## **Optimization of Model**

## Changing Attribute selection measure and max\_depth.

## Plotting Accuracy at each combination.

## 

From the above Fig, we can see that entropy criterion has better accuracy but both curves intersect when the max\_depth increases so we will choose criterion = entropy because of shorter training time and max depth = 27

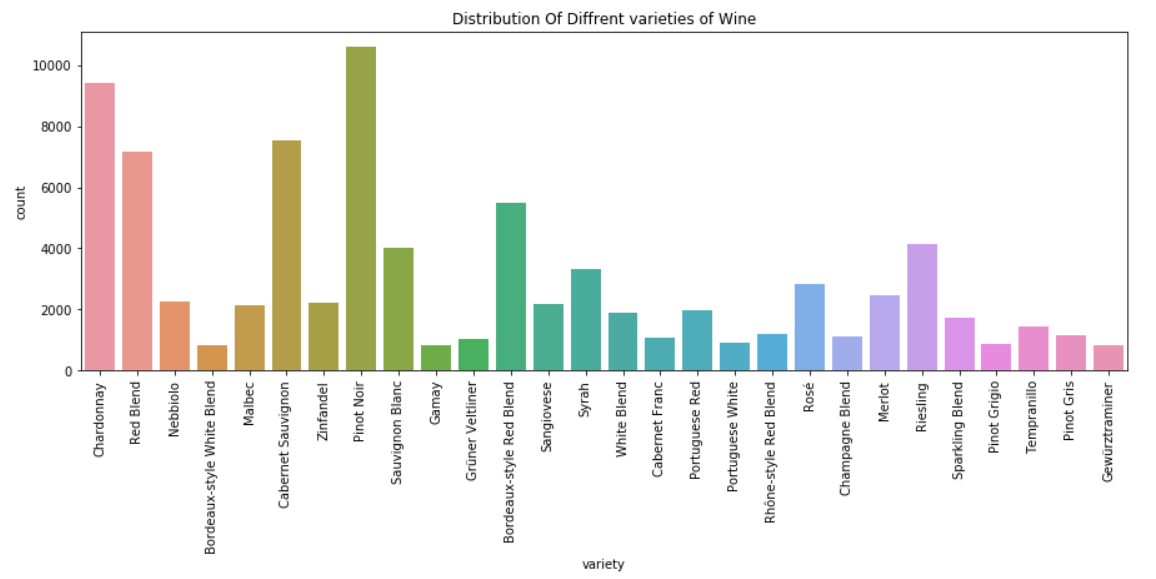
I built the final model with the entropy criterion and max depth of 27.

## **Prediction**

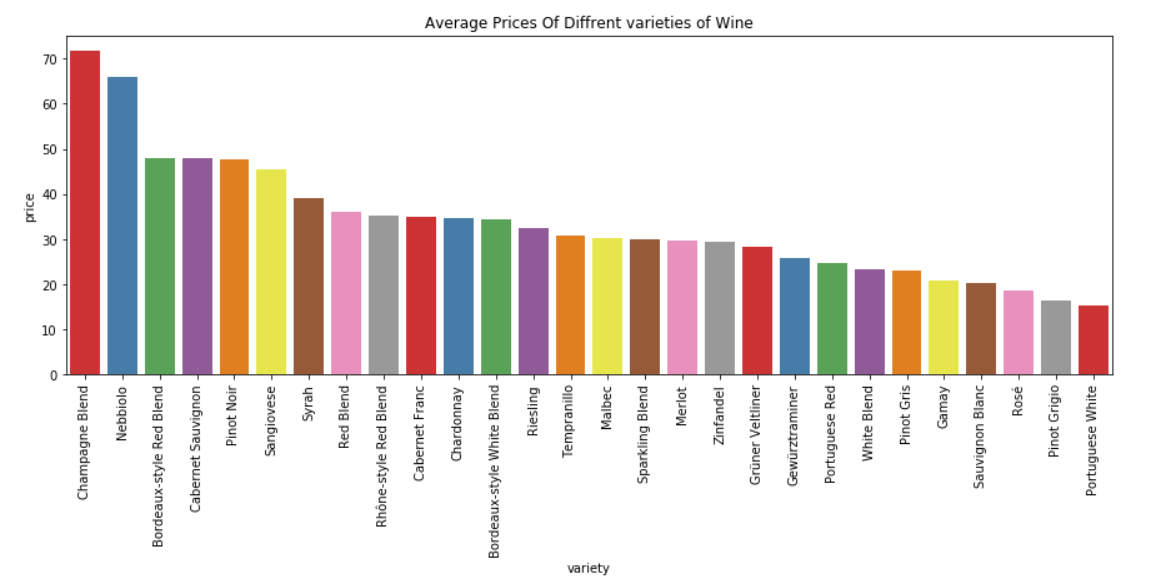
Finally, I Made the predictions on the test set. Which is loaded in the zip file predicted.csv

# **Insights**

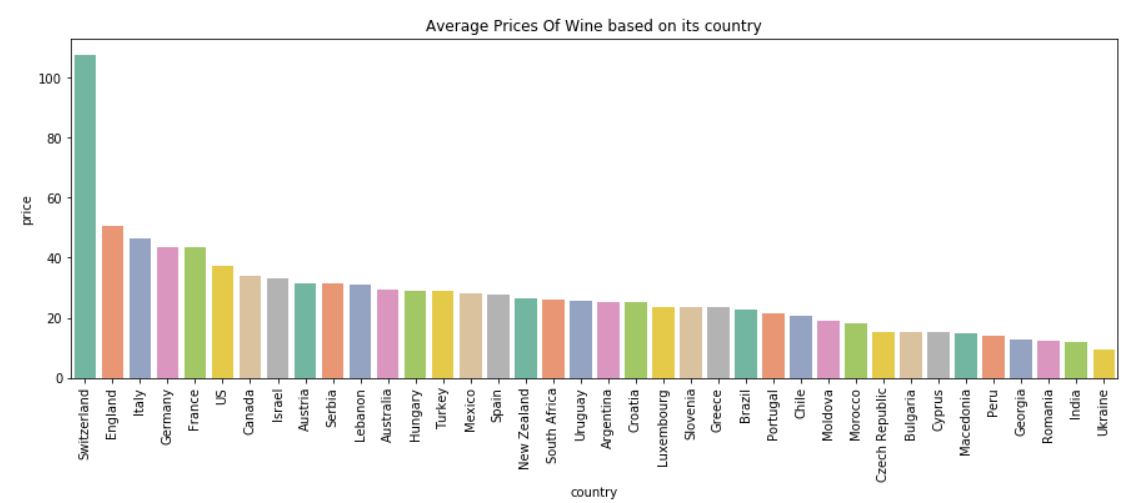
## **Frequency of Reviews with Different Varieties of Wine**



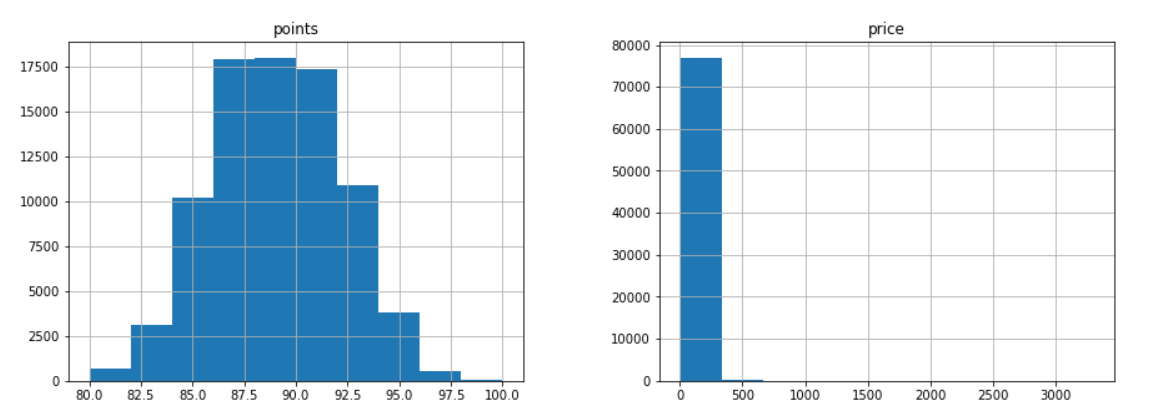
## **Average Wine Prices in Each Variety**



## **Average Wine Prices based on its country"**



## **Distribution of Prices and score points**



## **Trend Line between Prices and Score Point**

