CODE Analytics

BlueBerry Winery Quality Analytics

Georgios Kosmaras
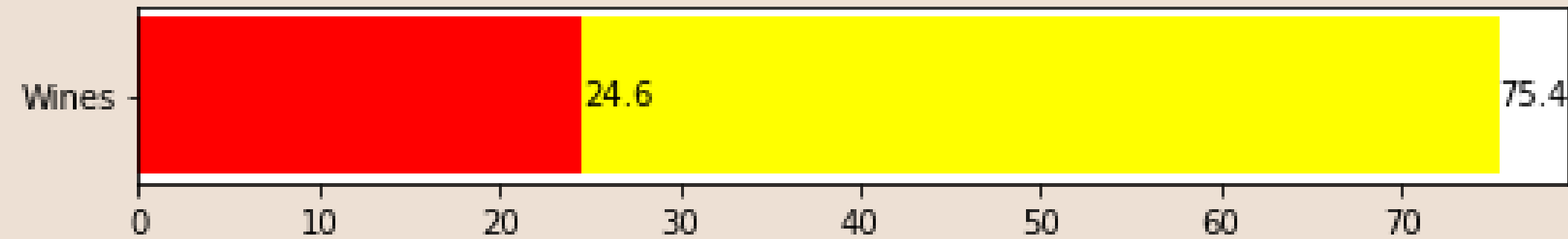
Table of Contents

CODE Analytics

- **We were given a dataset containing 6497 wines, of which 1599 were red**


- **Each wine has been measured for aciditic, sulphric and alcoholic content, as well as other chemical markers**

- **A quality mark determined by the average of three seperate tastings was provided as well**

- **There were no obviously incorrect nor missing values in the dataset**

- **Find whether any relationship exists between chemical composition and wine quality, and identify detrimental indicators**

- **Popular perception of Vinho Verde wines and ways to improve it**

- **Historical analysis of past trends, for price calculation and bussiness insights**

- **From the numerous features and markers of each wine, we determine which are more likely to impact the quality**

- **Crisp refers to the fresh taste, a result of acidity and sugar**


Feature importance visualization
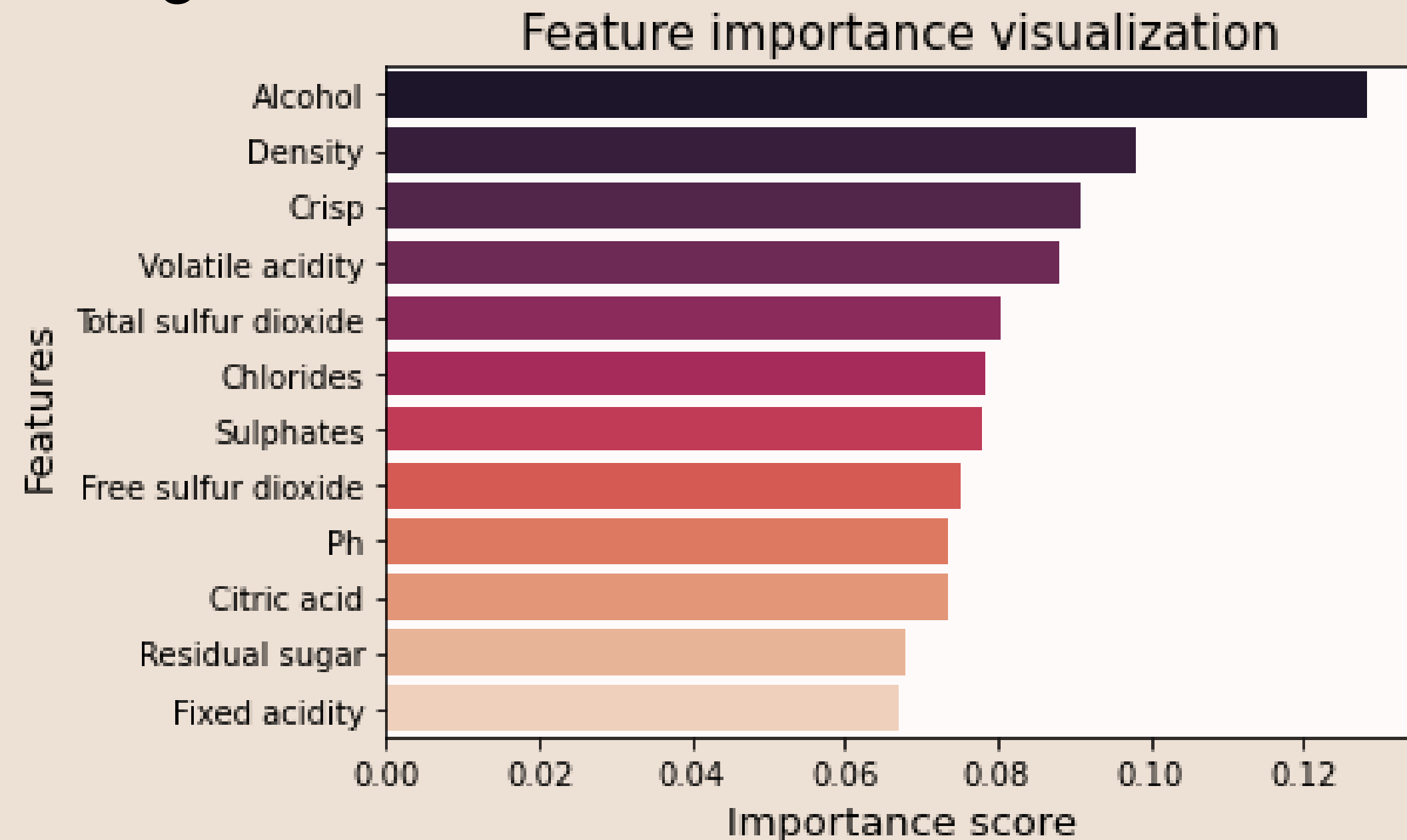
Main indicators for a wine's quality seem to be the following:

1. Alcohol
2. Density
3. Crisp
4. Volatile acidity
5. Sulfuric content

The selection of the indicators was made with both wine types in mind

- **We plan to create a predictive model using Machine Learning to help us determine wine's quality**

- **As we are not interested in individual quality scores, we will divide them in categories:**

1. Low with score less or equal to 5
2. Medium with score equal to 6
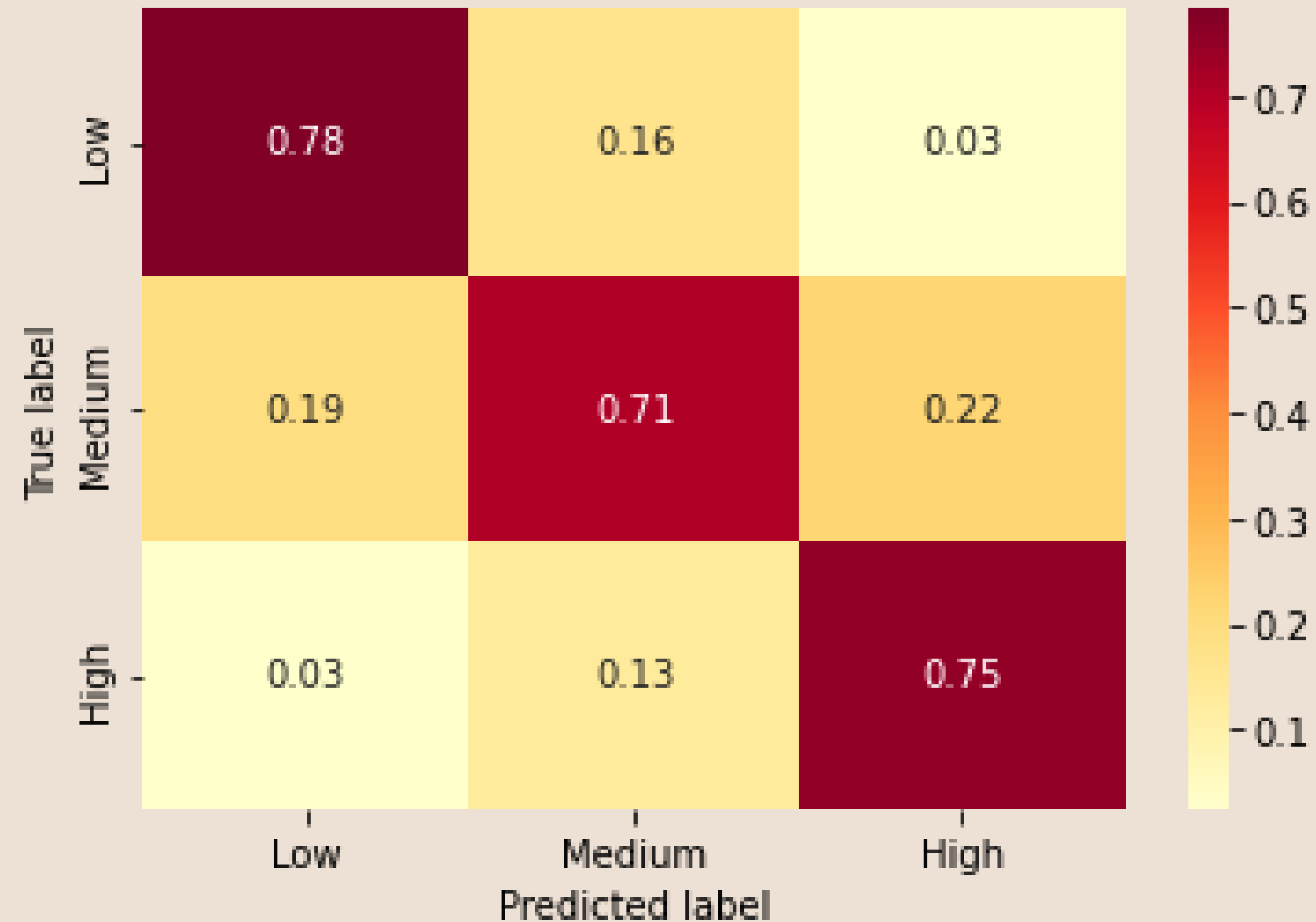3. High quality with score greater of equal to 7

Distribution according to quality label
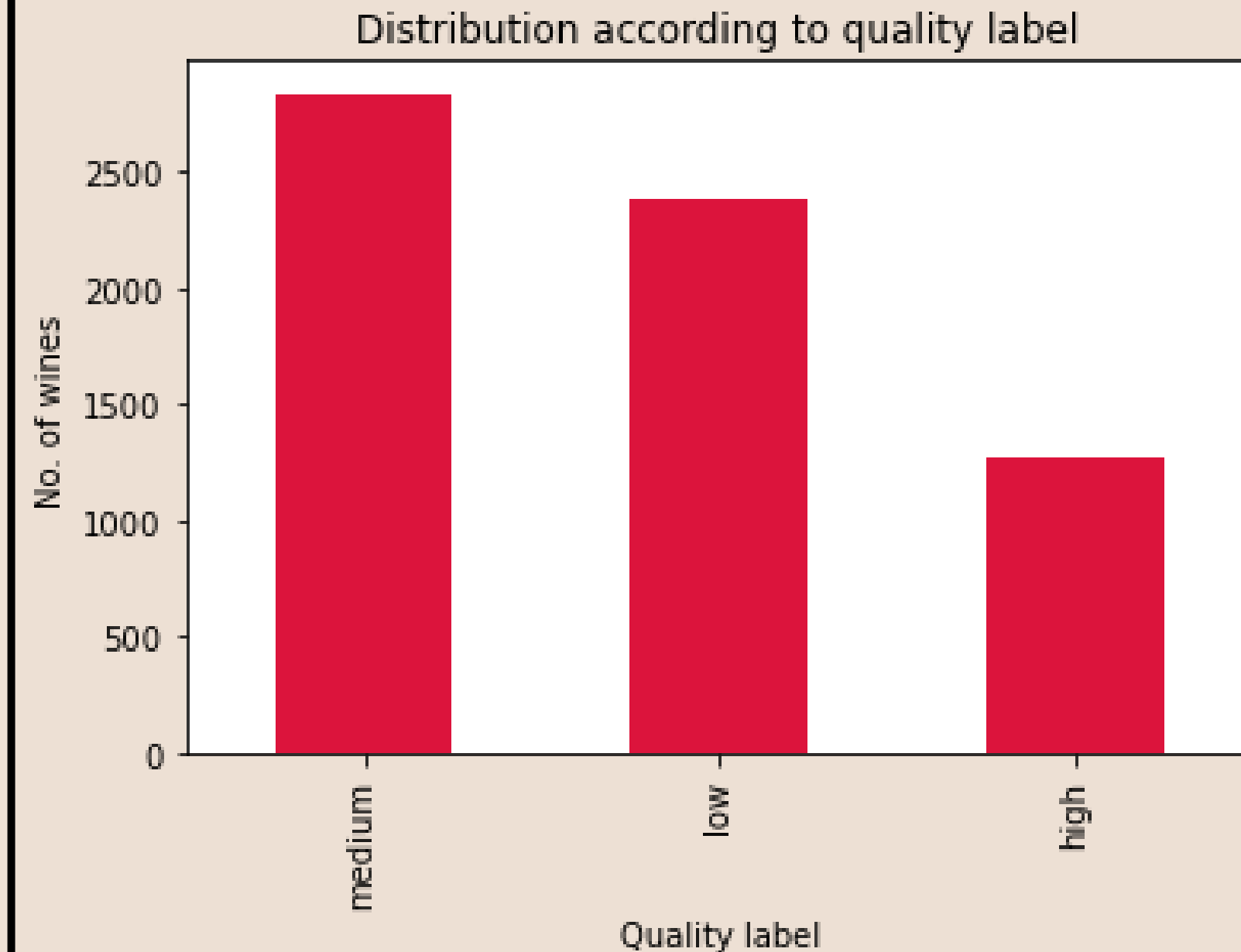
- **For the predictive model we had to choose between the various ML ensembles available**

- **We went with Random Forest, as it offered the best accuracy**

- **Determined the best hyperparameters for optimal results and tested it**

Comperative accuracy of different models
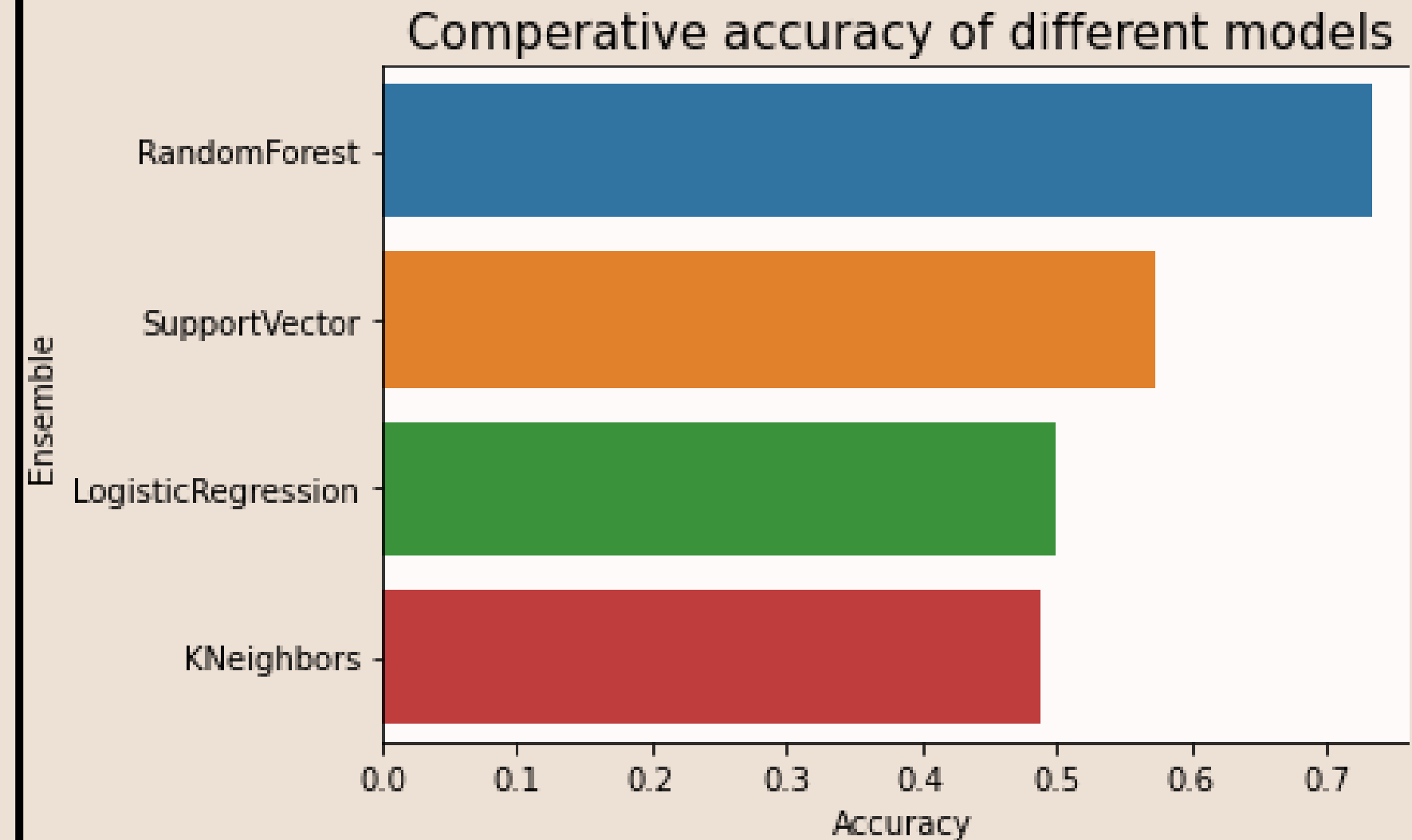
- **Our Random Forest predictive model achieved an average of 74% accuracy rate when determining quality label**

- **15%-20% in each category were misidentified in a neighboring category**
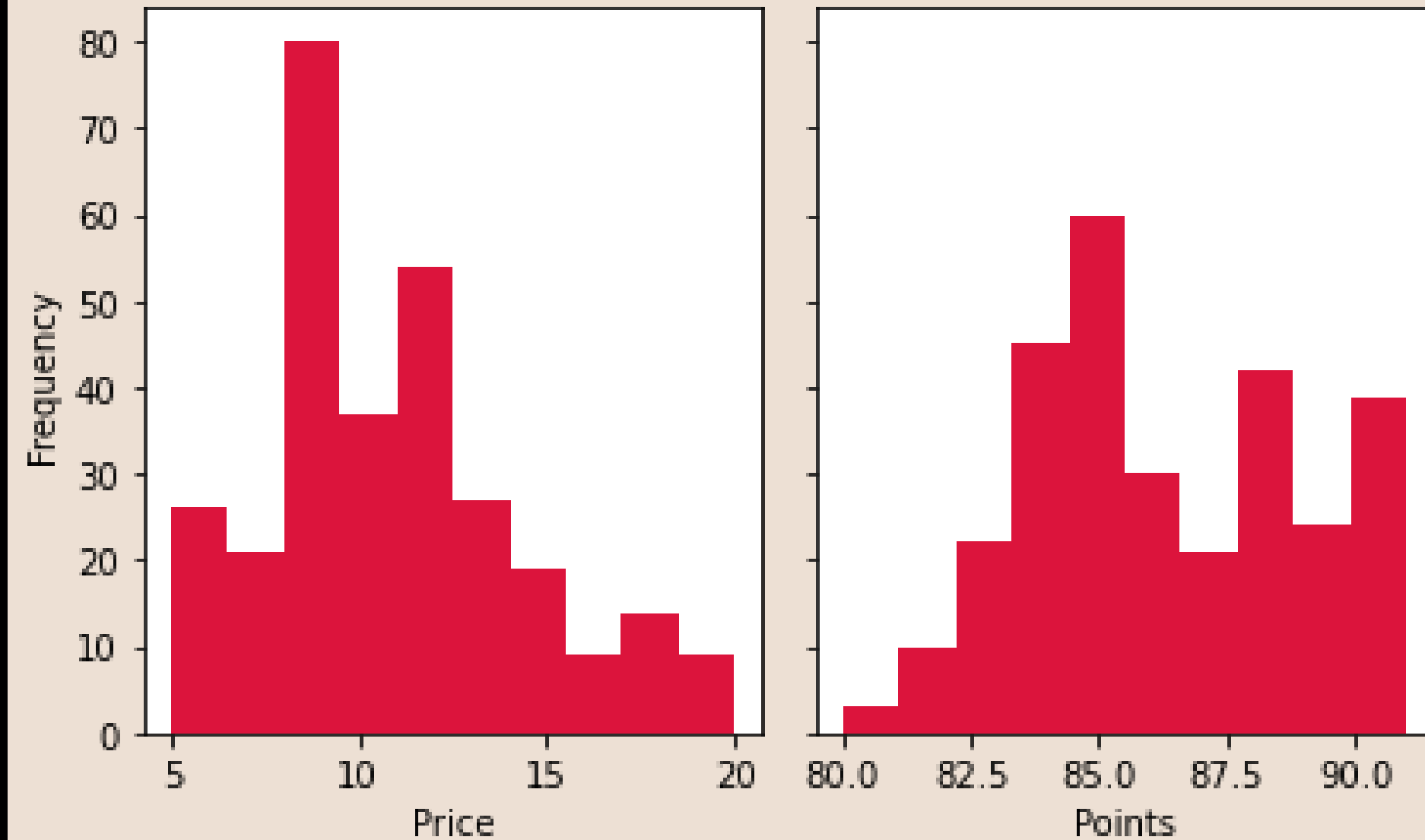
- **Gross misidentification was less than 3%**

- **For determing the best price for the wines we had another dataset available, with the price and a quality ranking (of a different scale)**

- **As before, most of the wines were in the middle range, with a few exceptions of above or below average**

- **We also researched the price history of Vinho Verde wines in various online suppliers**

Frequency analysis of quality points and price for Vinho Verde wines

- **Despite the diffent quality scales used in the datasets, we can still assign quality labels**

- **We managed to achieve a similar distribution to the original dataset**

- **This can help in determining the best price for a wine based on its quality**

Quality label distribution
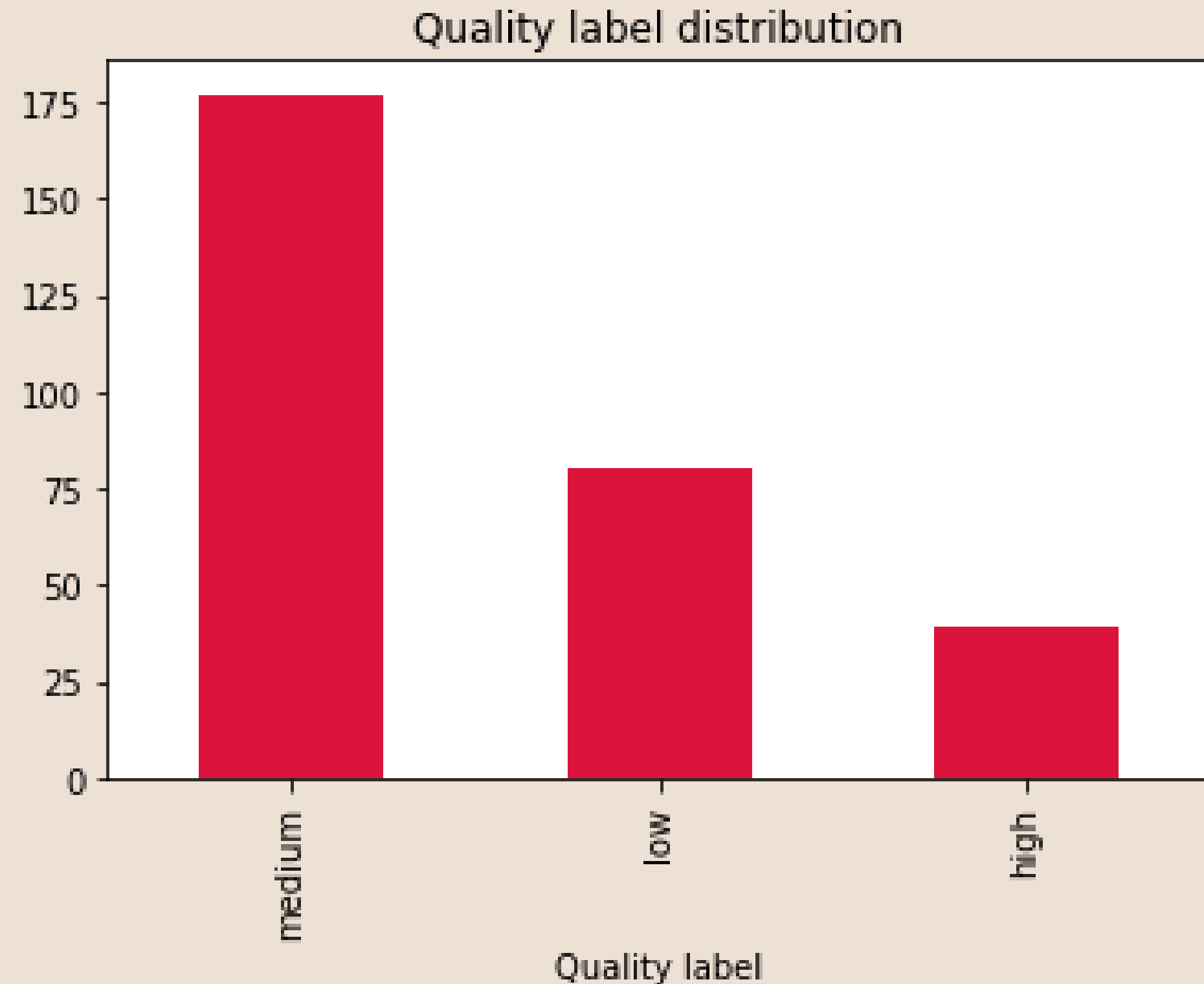
- **Yearly production growth shows a decade average of 2.7%, driven by white wines which increase by 3.5%. Red wines shrink production at 5.8% rate**

- **Recommended for quick consuption, not ideal for aging and storing**

- **Positive consumer attitude towards Vinho Verde, assosiating them with refreshing, fuzzy wines**
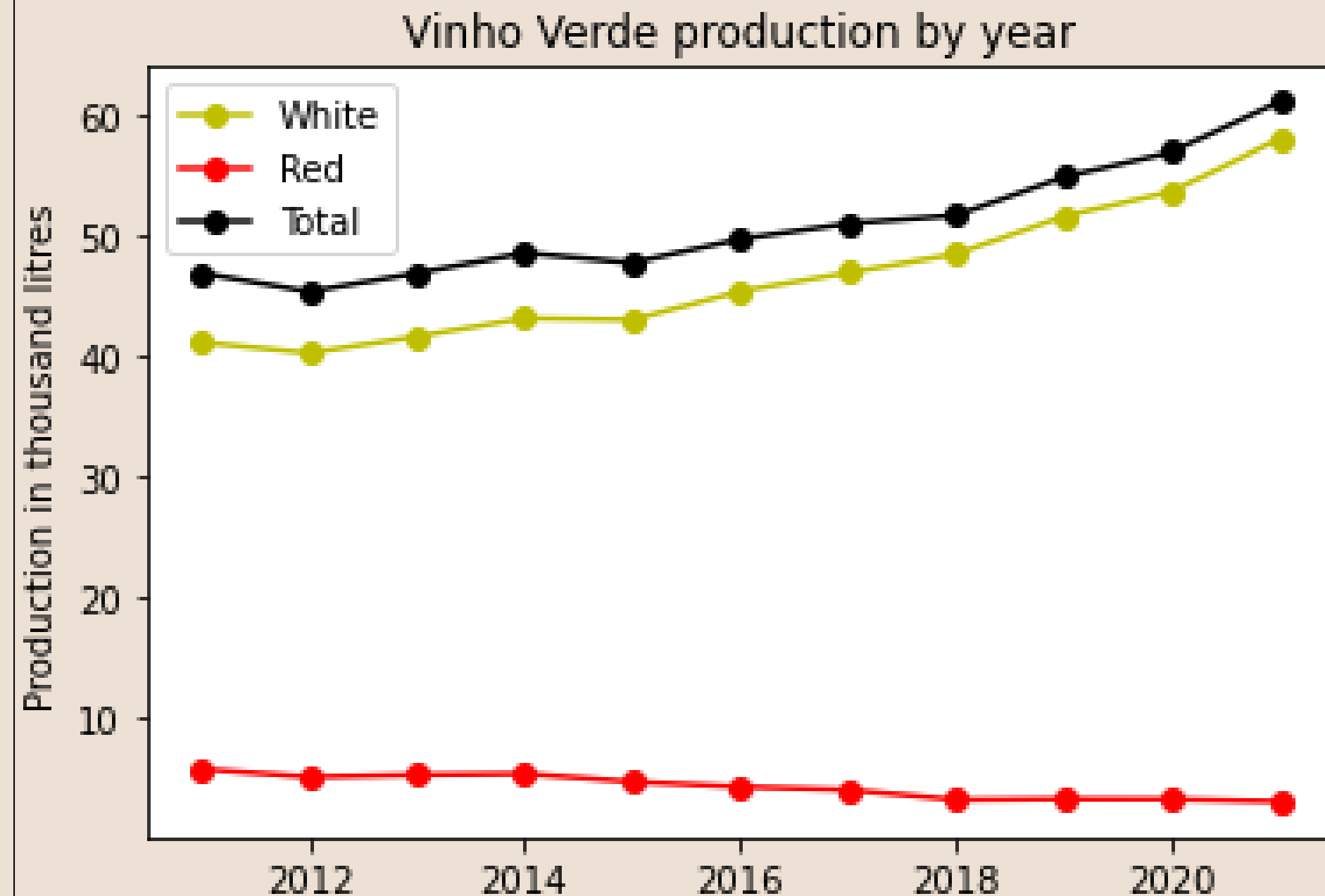
Vinho Verde production by year

- **Exports have been growing at an annual rate of 4.1%; mainly driven by American and European costumers**

- **Preference on those markets for more medium quality, affordable wines**

- **Emerging markets for premium wines in Brazil, Angola and Japan**

Vinho Verde exports by year

- **Based on the aforementioned findings we can recommended a few courses of action for BlueBerry Winery:**

1. According to past trends, Vinho Verde wines can be sold at **6.5-9€**, **9-12.5€** and **13-18€** for low, medium and high quality respectively

2. Main quality indicator that can be affected is the alcoholic content

3. High end wines have higher demand in countries like Japan, Brazil and Angola

4. Medium and low quality ones fare better as exports in Europe and America.

5. The quality label can be deduced with reasonable certainty using our predictive model, making classification and designation for different point of sale much easier