Project Big Data - Group 20
Wine Reviews Analysis Report

A. Ilbay (2798233)
B. Yalçın (2797623)
Ç. Pınarbaşı (2658187)
D. Gülal (2797094)

July 2, 2023

Contents

1 Introduction

This paper focuses on a collection of wine reviews in the Wine Reviews Dataset provided by Kaggle (Wine Reviews , 2017). This research will aim to answer interesting questions ranging from determining the years that produced most high quality wines, quality price ratio of the wines and more. This research will also explore the correlation between the features and will attempt to find the most important features that affect the ratings of the wines. Methods such as classification and regression models will be used for the analyses.

1.1 Research Questions

This paper will be focusing on the following research questions:

- What was a good wine year?
- What feature(s) seems to be most import for the quality of a wine?
- Which countries have irregularities in wine quality?
- Which regions have irregularities in wine quality?
- Which countries on average produce the best wines?
- Which provinces on average produce the best wines?
- Which countries have the best quality/price rate?
- Which wine varieties have the best quality/price rate?
- What are the keywords that describe the wine characteristics?
- Which classification models fit best for predicting best wines?

2 Background

The wine reviews dataset is available publicly on Kaggle. The data was collected from the WineEnthusiast magazine. The magazine consists of wine reviews. The dataset contains reviews of wine tasters from different regions of the world who are experts in their fields. The dataset is used for explore and analyze different types of wine from different regions, wineries, wine varieties and prices. The dataset contains 130k wine reviews. The dataset contains subjective opinions held by the reviewers therefore the consistency of the ratings are likely to differ. It is important highlight that differences in reviews and the abundance of categorical data will introduce challenges for this research.

3 Dataset

The datasets provided by Kaggle consists of 280k reviews in total. The two wine review datasets are named "winemag-data_first150k.csv" and "winemag-data-130k-v2.csv" and contain the relevant information about the reviewed wines, their reviewers and the reviews in text form. The size of the datasets are approximately 49 MBs and 52MBs respectively. The CSV formatted datasets include comma separated columns for the items in the dataset. The dataset contains the following features:

- "country" depicts the country of origin for the wine.
- "description" column contains the review made by the wine taster.
- "designation" column indicates the vineyard that the grapes were grown.
- "points" column depicts the ratings given by the reviewer for the the wine and it will be interpreted as an integer type for this research.
- "price" column depicts the prices of the wines and it will be interpreted as a float type for this research.

- "province" specifies the province that the wine was produced in.
- "region_1" and "region_2" provide the information about the region the wine originated from.
- "taster_name" shows the reviewers full name.
- "taster_twitter_handle" is the Twitter handle of the wine reviewer.
- "title" depicts the name of the reviewed wine.
- "variety" column indicates the various types of grapes that were used for the production of the wine.
- "winery" contains the name of the winery that produced the wines.

4 Analysis

4.1 Data Cleaning

Two datasets "winemag-data_first150k.csv" and "winemag-data-130k-v2.csv" were merged however some of the features were removed from the merged dataset as they weren't present in the other dataset. The removed features are "taster_name" and "taster_twitter_handle" as these columns only existed in the "winemag-data-130k-v2.csv" dataset. During the analysis stage t the entries that contained empty data had to be removed from the remaining database to ensure meaningful data analysis.

4.2 What was a good wine year?

As shown in Figure 1 the average points per year for wines was explored to determine the year that had the most amount of positive reviews for wines. As the data didn't contain a column for the years a new column had to be created for years by extracting the years from the title of the wines. The years had to be limited from 1950 to 2023 as extracting 4 digit numbers from the title of the wines did not always correspond to a valid year. For example, if a wine had 7000 in its title the extracted year would be the year 7000 and because the wine year couldn't be greater than current year the year limit was put in place. The results show that the year of 1969 was the best year for wines while there was a great dip in quality in the following years. Although it improves around 1980s again the average points for wines never reaches the heights late 1960s and the quality keeps degrading as years approach 2020.

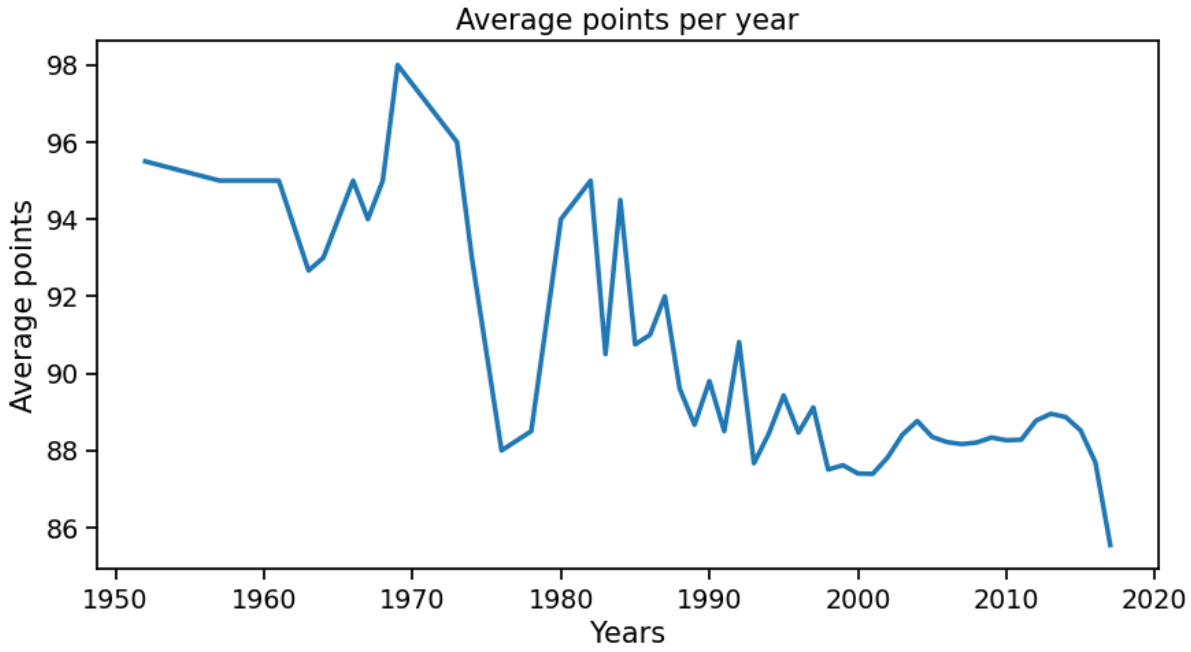

Figure 1: Graph of average points per year

4.3 What feature(s) seems to be most import for the quality of a wine?

In order to determine the features that affect the data most a correlation matrix was constructed. From the matrix it was discovered that price and description were the most important features to affect the rating of the wines. As shown in Figure 2 price was the most important with 0.19 followed by description at 0.03. Other than region_2 and and designation the features seem to be less important for determining quality of the wine. Although one might assume that because price has the greatest correlation to the points it would be the determining factor for predicting the points of wines the later sections of the research will outline how this notion didn't hold true when using classification and regression models. Correlation of description and region_2 would prove to be more useful for gaining an insight about the dataset

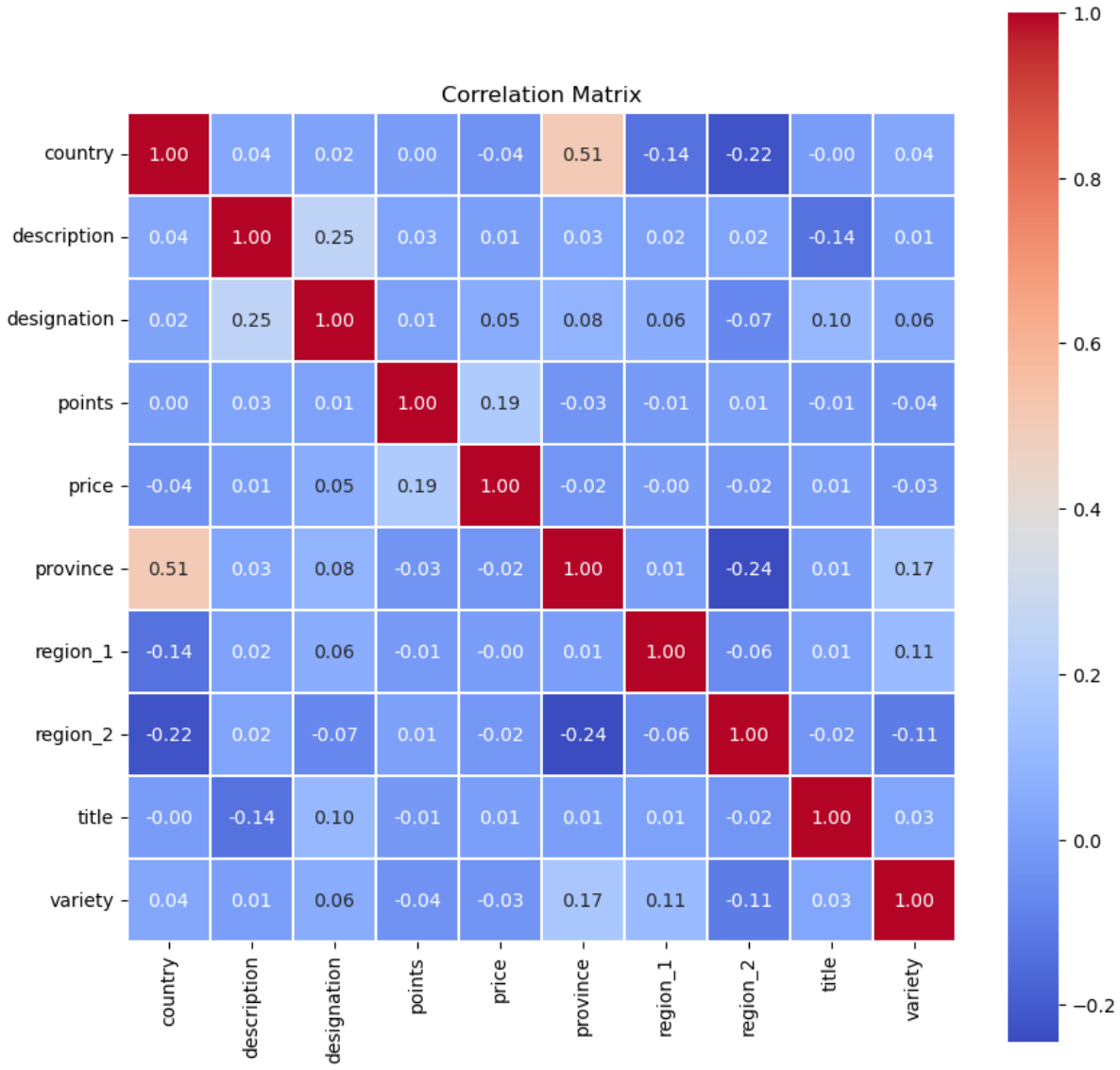

Figure 2: Heatmap of correlations between features

4.4 Which countries have irregularities in the wine quality?

To measure the irregularities in wine qualities for every country the standard deviation of wine points was calculated. In Figure 3 top 25 countries with the most standard deviation of wine quality are displayed. In countries such as China and United States the quality of the wines differed drastically. The results were intriguing as it brought up a new question to be answered for the varying wine qualities. Were the irregularities in the wine quality related to the geographical size of the countries?

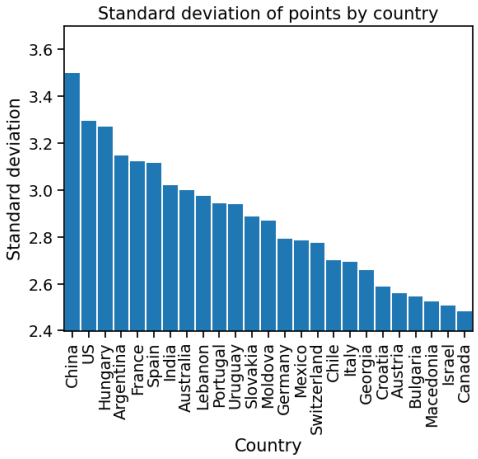

Figure 3: Wine points standard deviations

4.5 Which regions have irregularities?

As mentioned before, after investigating the irregularities in wine qualities of countries next step was questioning whether the deviations in quality for China and US were due to great land mass the two countries covered. Similar to how it was done for the countries the standard deviation of wine scores were calculated. As shown in Figure 4 most irregularities observed belonged to regions from the United States. However, regions of China weren't in the lead for the list of regions with most irregularities. Although not conclusive this shows that for United States the varying quality of wines in the regions seemed to have an affect on the overall quality consistency for the country while for China regions weren't the main cause of the irregularities. This also relates back to the correlation matrix in Figure 2 as region_2 had an impact on the quality of the wine.

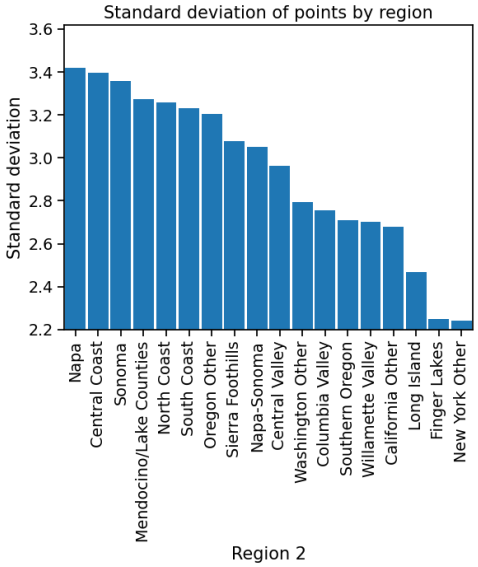

Figure 4: Irregularities in wine quality of regions

4.6 Which country/province on average produces the best wines?

Another aspect of the data that was analyzed was countries and provinces that on average produced the best wines. Average score of wines for each country and province was calculated from the mean of the wine points and later the top 25 countries were plotted as shown in Figure 5. The country produced the best wine over the years was England which was followed by Austria and Germany. However, as shown in Figure 6 when the average points per provinces are considered England as the region was in the 5^{th} place while Südenburgland region of Austria took the lead.

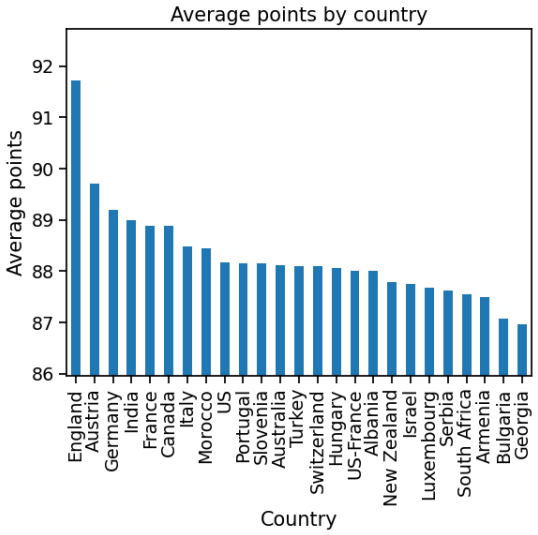

Figure 5: Countries best

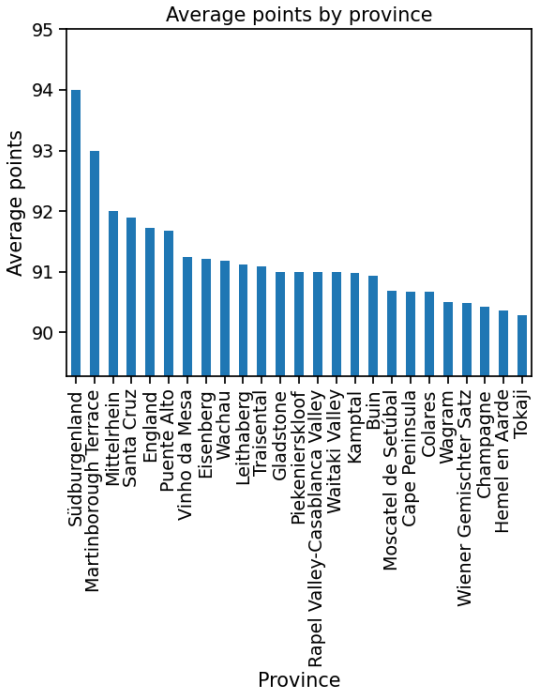

Figure 6: Provinces best

4.7 Which countries have the best quality/price rate?

For each entry the point/price rate was calculated and then the boxplot was generated according to the countries of those entries. The boxplot shown in Figure 7 illustrates the top 25 countries that have the best quality/price rate of the wines they produced. Overall the country that had the best price/points ratio was Switzerland which was calculated by comparing the average point/price rates of all countries. However, the boxplot provides a better perspective as it shows that some of the countries produce wines that provide better taste for their price. It is important to note that a wine that has bad quality but also has the cheapest price would be considered the same as a wine that is high in quality but also equally high in price. For this dataset it is not a concern as previously seen the dataset contains wines that are rated between 80 to 95 therefore none of the wines are considered bad by the reviewers. This means that for customers it is more likely than not that the wine of choice with highest points/price rates would be satisfactory.

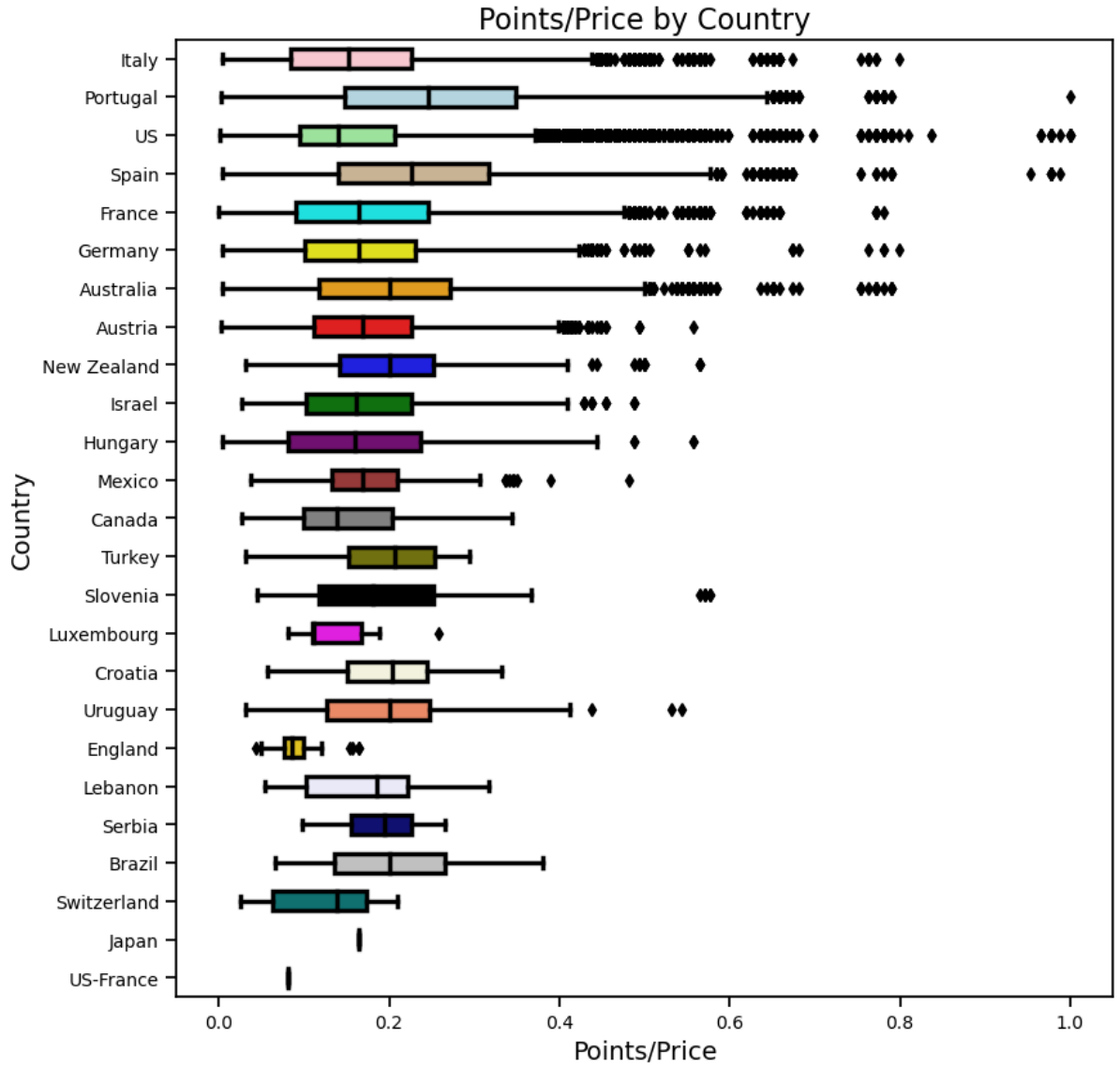

Figure 7: Boxplot of price/performance of countries the wines were produced in.

4.8 Wine Variety Analysis

The reviews were grouped by regions then the number of reviews were counted for each region. As there are on average 1.1 reviews per wine title, and only 12 titles with more than 5 occurrence, it is safe to assume that each review corresponds to a different wine. Then these counts are shown on Figure 8, indicating the regions with the highest diversity, which provides some insight on the distribution of varieties. As Figure 9 shows, most regions produce few varieties, while other regions produce more than 60 different varieties. Its relation to price and points can be seen on Figure 10. As seen from Figure 8, 9 out of 10 regions were in the US. Overall, the data contains 1,332 regions, and 756 varieties. These regions on average contain 6 different varieties of wine with standard deviation of 10. The maximum of these is Paso Robles with 86 different varieties. Each variety occurs on average 342 times with standard deviation of 1,933. This shows there is high variation between varieties.

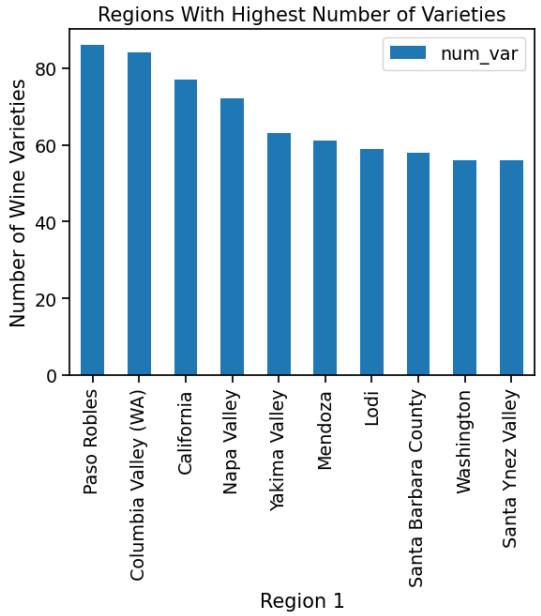

Figure 8: Regions with highest number of unique wine varieties

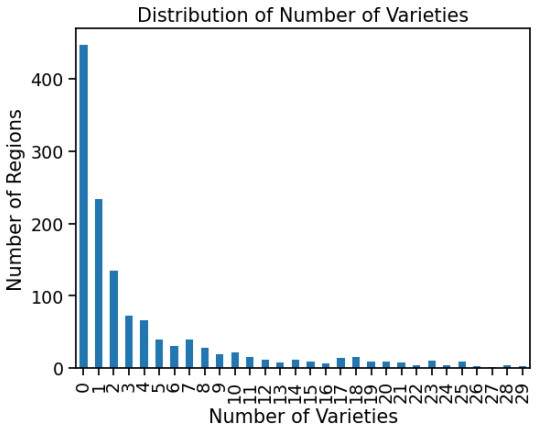

Figure 9: Distribution of Varieties

To see if number of varieties a region produces is correlated with price or points, variety counts of regions were grouped by certain intervals and the price and points were averaged as shown in Figure 10. Furthermore, Figure 10 indicates a correlation between the number of varieties and prices, with higher prices observed when a region focuses on a smaller number of different varieties. Although points seem unrelated, it seems to be higher when a region produces less number of varieties. For the quality/price rates of wine varieties the points for the wines were divided by the prices of the wines similar to quality/price rates of countries. The results were fit into a 0 to 1 scale and later the top 25 varieties were plotted as shown in Figure 11 in decreasing order. From the results it is gathered that Ramisco variety, a Portuguese brand, had the best quality/price rate.

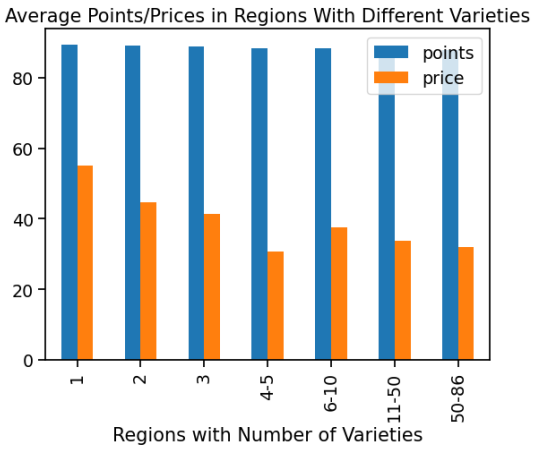

Figure 10: Point/Price Rates For Regions Grouped by Number of Wine Varieties

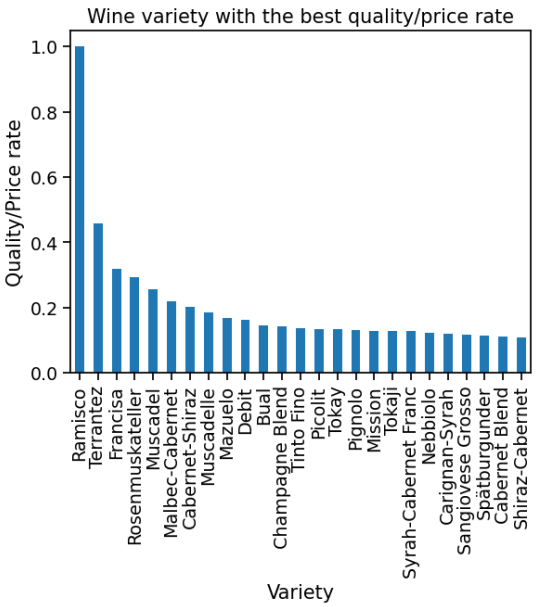

Figure 11: Variety Quality/Price rates

To gain some insight about the popular varieties themselves, reviews were grouped by varieties then occurence of varieties were counted. The most common variety Pinot Noir occurs 26,415 times. This can be seen on figure Figure 12. The top five regions with the most occurence of Pinot Noir can be seen on figure Figure 13. On average, each region has 115 occurence of Pinot Noir, with standard deviation of 418. The region with the maximum number of Pinot Noir is Napa Valley with 5400 occurences.

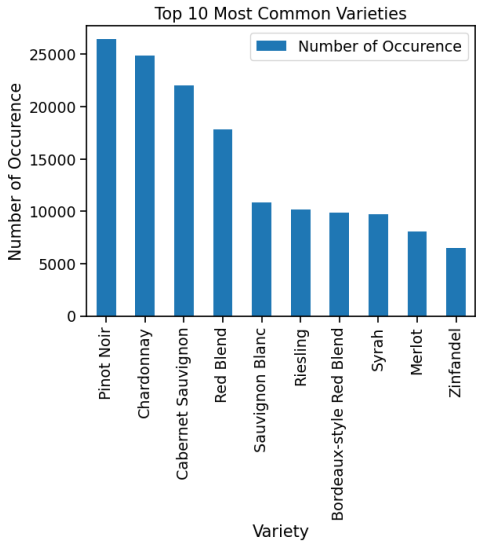

Figure 12: Common Wine Varieties

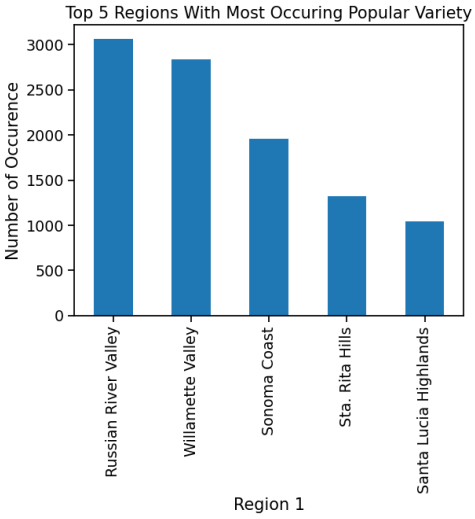

Figure 13: Regions with most occurrence of the most popular wine Pinot Noir

4.9 What are the keywords that describe the wine characteristics?

To answer this, description of the reviews were taken and put through pre-processing steps, first removing punctuation, then stemming the words. Then a CountVectorizer has been used to calculate word occurences, and then later to train LDA and Clustering models. The word frequency analysis entailed summing up token counts for each word, allowing for the identification of the most common words and their corresponding occurrences. The results of this analysis provided valuable insights into prevalent terms and a deeper understanding of the dataset. Unsurprisingly most common word that was used in the reviews was "wine" as shown in Figure 14. Reviewers seemingly focused on the flavor, fruitiness and the finish of the wines. According to reviews, flavor approximately doubles the finish in word count. The color of the wine was also mentioned less in the reviews. Texture and the smell were less important compared to acidity and aroma of the wines.

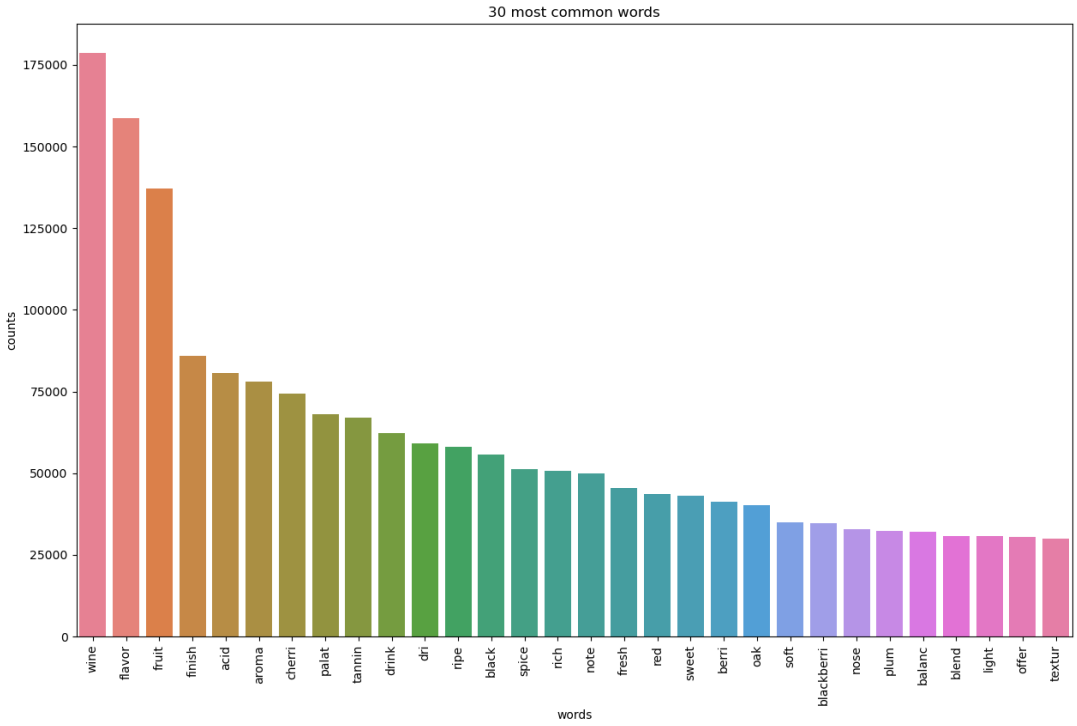

Figure 14: Most common words used to describe wine characteristics

In order to explore the connections among words, two approaches, namely Latent Dirichlet Allocation (LDA) and K-Means clustering were employed. Both models performed poorly. While LDA did identify certain correlations, such as grouping "cherry" and "apple" with "fruit", "tannins" with "age", clear distinctions between the groups were lacking. Moreover, the groups shared common words such as "wine" and "flavor", which made it hard to differentiate between groups. This can be confirmed by Figure 15 where the words are ordered to indicate significance to determine the group. K-Means performed similarly.

```
Topic #0:
wine fruit tannin black drink ripe age acid structur firm

Topic #1:
cherri black wine flavor fruit blackberri cabernet spice blend chocol

Topic #2:
wine fresh fruit palat white appl acid flavor aroma finish

Topic #3:
flavor wine acid drink rich dri fruit sweet ripe cherri

Topic #4:
flavor finish aroma palat fruit note berri nose feel plum
```

Figure 15: Groups found by the model.

5 Method

In order to investigate whether the dataset could be fit into a model various regression and classification methods were used. This section will provide the theoretical background for the methods.

5.1 Regression models

5.1.1 Logistic Regression

Logistic Regression is a statistical model used for binary classification tasks. The algorithm calculates the probability of the binary outcome using a logistic or sigmoid function. By fitting the model to the training data, logistic regression estimates the coefficients associated with the independent variables (Hosmer et al., 2000). During prediction, logistic regression uses the learned coefficients to compute the probability of the binary outcome for new data. Logistic regression is widely used due to its simplicity and effectiveness in dealing with linearly separable data.

5.1.2 Linear Regression

Linear regression is a statistical modeling technique that establishes a linear relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting line that minimizes the difference between predicted and actual values (Khuri , 2013).

5.2 Classification

5.2.1 Gaussian Naive Bayes

A Bayes classifier is probabilistic model that applies Bayes' theorem to find for a set of feature values the target class it is most likely to belong to. A Naive Bayes classifier simplifies finding the conditional probability of the label by assuming all features are independent conditional on the class (Rish , 2001).

5.2.2 Gradient Boosting

Gradient boosting is a machine learning technique that combines multiple weak predictive models, typically decision trees, to create a strong predictive model. It works by sequentially adding models that correct the errors made by previous models. Each subsequent model is trained to minimize the residual errors of the previous models, using a gradient descent optimization algorithm. By iteratively improving the model's predictions, gradient boosting achieves high predictive accuracy (Bentéjac et al. , 2020).

5.2.3 Multinominal Naive Bayes

Multinomial Naive Bayes is a probabilistic classification algorithm that is commonly used for text classification tasks. The algorithm calculates the likelihood of observing each feature given the class and combines it with prior probabilities of the classes using Bayes' theorem. It assigns the class label with the highest posterior probability as the predicted class (McCallum et al. , 1998).

5.2.4 Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm used for both classification and regression tasks. It aims to find an optimal hyperplane in a high-dimensional space that separates data points of different classes with the maximum margin. The algorithm works by transforming the input data into a higher-dimensional feature space using a kernel function. In this transformed space, SVM identifies the hyperplane that best separates the data points of different classes while maximizing the distance, or margin, between the closest data points from each class (Steinwart et al. , 2008).

5.2.5 Random Forest

Random Forest is an ensemble method used for classification. It combines multiple decision trees trained on random subsets of data and features. Each tree independently predicts the outcome, and the final prediction is made by aggregating the tree predictions through majority voting. Random Forest mitigates overfitting, improves accuracy, and excels in handling high-dimensional datasets with complex feature relationships (Liaw et al. , 2002).

5.2.6 Decision Tree

A decision tree is a machine learning algorithm that uses a tree-like structure to make predictions. The decision tree algorithm works by recursively splitting the data based on the selected features, aiming to maximize the separation between different classes or minimize the variance within each class. The splitting process continues until a stopping criterion is met, such as reaching a maximum depth, a minimum number of samples at a node, or when no further improvement can be achieved (Hastie et al. , 2009).

5.2.7 Extra Tree

Extra Trees is an ensemble learning algorithm similar to Random Forest, but with additional randomness. Each tree in the Extra Trees ensemble is built using a random subset of features and splitting thresholds. Unlike Random Forest, Extra Trees introduces randomness not only in selecting features but also in choosing splitting thresholds. This increased randomness reduces bias and increases variability compared to Random Forest (Geurts et al. , 2006).

5.2.8 K Nearest

KNN is an algorithm used for classification and regression tasks. It calculates the distance between a new data point and all training set points, selecting the K nearest neighbors. In classification, the majority class among these neighbors is assigned as the predicted class. The parameter K determines the number of neighbors considered, with smaller K values increasing sensitivity to local variations and larger K values providing robustness to noise but potentially overlooking local patterns (Mucherino et al. , 2009).

5.3 Text Data Representation

5.3.1 Label Encoding

Label encoding is the process of assigning an integer value to every possible value of a categorical variable. For a dataset that has categorical variable with values drawn from the set “fruity”, “acidic”, “clear” then label encoding might assign the mapped values from the set 0,1,2 respectively. (Hancock et al. , 2020)

5.3.2 TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) is a word encoding technique extensively employed in natural language processing (NLP) tasks. It combines the TF component, which quantifies the frequency of words within a document, with the IDF component, which measures the rarity of words across a collection of documents. By considering the relative frequency of a word in a document and its uniqueness across the entire document set, TF-IDF captures the importance and significance of each word. (Qaiser et al. , 2018) For instance, in the context of wine descriptions, TF-IDF can be utilized to extract relevant keywords such as “fruit,” “tannin,” and “cherry,” enabling the inference of points and prices associated with the wines.

5.4 Clustering and LDA

5.4.1 K-Means

K-means, is widely recognized as one of the most commonly used clustering techniques. Initially, an imperfect clustering is established. Each data point is then reassigned to its nearest center, and the clustering centers are updated by computing the mean of the points assigned to each center. This reassignment and updating process is repeated until certain convergence criteria are met, such as a predefined number of iterations or a specified difference in the distortion function value. (Jin et al. , 2011)

5.4.2 LDA

Latent Dirichlet Allocation (LDA) is a hierarchical Bayesian network that represents the generative model of a corpus of documents. It is used for automatically generating topics from text corpora and assigning words in the corpus to these topics (AlSumait et al. , 2009).

6 Results

For regression models and classification models the previously mentioned methods were used. The wide range of classification models and the two regression models were used in order to find the best model for predicting the scores of the wines. The data was split into three; train set, test set and validation set. The test and validation sets were 10% of the data and this was done to ensure predictions to be unbiased. The data was split into the 80/10/10 shape for training, test, and validation sets. For models that used label encoding all of the features were used in order to get the best results. Although the correlation matrix implied that price played a bigger role compared to other features fitting the data without other features resulted in diminishing returns. For the label encoding the best results were achieved by encoding all of the features. As an alternative to label encoding TF-IDF was applied to the descriptions of the wine reviews as that was the second most important feature according to the correlation matrix.

6.1 Regression models

From the two regression models that were used logistic regression that utilized Label Encoding performed worse compared to linear regression that utilized TF-IDF. As illustrated in Figure 16 predictions made by logistic regression on most cases failed to follow the line. The inaccuracy of the regression model can be caused by the label encoding method's failure to recognize a pattern in the encoded values. On the other hand, linear regression with TF-IDF depicted in Figure 17 shows a better fitting regression model where the predictions made by the model follow the line more closely compared to logistic regression. This is significant since the data is mostly comprised of categorical data generated by the subjective beliefs of the reviewers.

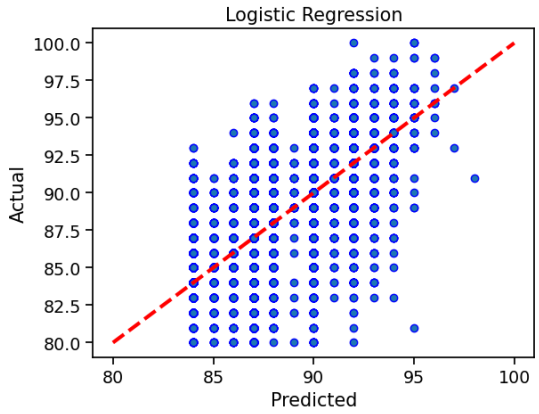

Figure 16: Logistic Regression with Label Encoding

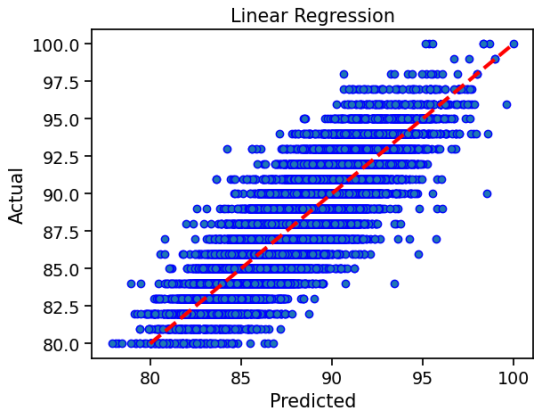

Figure 17: Linear Regression with TF-IDF

Linear regression with TF-IDF method yielded the best accuracy of 0.72 compared to all other models. This was in accordance to the assumptions made from the correlation matrix and it showed that TF-IDF was a better method than label encoding as it was able to recognize the patterns in the descriptions from the words that were used by the reviewers to describe the wine quality.

Models	**Accuracy**
Linear Regression TF-IDF	0.72
Logistic Regression	0.17

Table 1: Regression Accuracy

6.2 Classifiers

As Table 2 shows most of the models struggled to achieve an accuracy that was above 50%. This was likely due to the fact that majority of the data was categorical and Label Encoding the features wasn't enough for most of the classifiers to make accurate predictions. While random forest and decision tree had comparable accuracy, random forest on multiple runs yielded the consistent accuracy results while decision tree's accuracy dropped or increased between different executions. Random forest with TF-IDF performed significantly better than all classification methods.

Models	**F1 Scores**
Random Forest TF-IDF	0.68
Random Forest	0.55
Decision Tree	0.54
Extra Tree	0.51
Gradient Boosting	0.20
K Nearest	0.19
Gaussian Naive Bayes	0.17
Support Vector Machine	0.16
Gradient Boosting TF-IDF	0.14
Multinomal Naive Bayes	0.04

Table 2: F1 scores

From the Label Encoded classifications the resulting ROC curve depicted in Figure 18 showed that random forest classifier converged in a favorable rate while others were far from ideal. Although the F1 score for Gradient Boosting with and without TF-IDF fell far below the top performing models ROC curve implies that gradient boosting is the a better fit for the dataset than decision tree and extra tree classifiers.

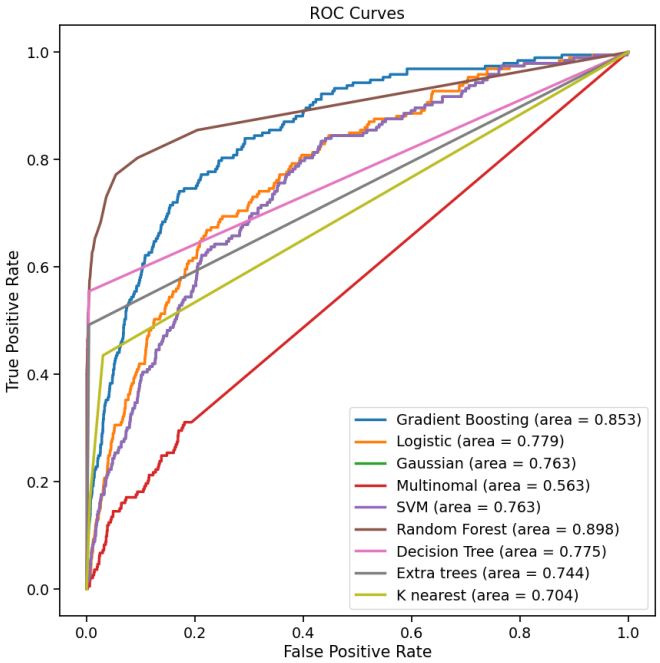

Figure 18: ROC Curves for Label Encoded classifications

The confusion matrix given in Figure 19 shows the drawbacks of the Label Encoding method for random forest classifier as this made it more difficult for random forest to come up with correct predictions. In Figure 19 it shows that the true values of the labels and predicted values different considerably.

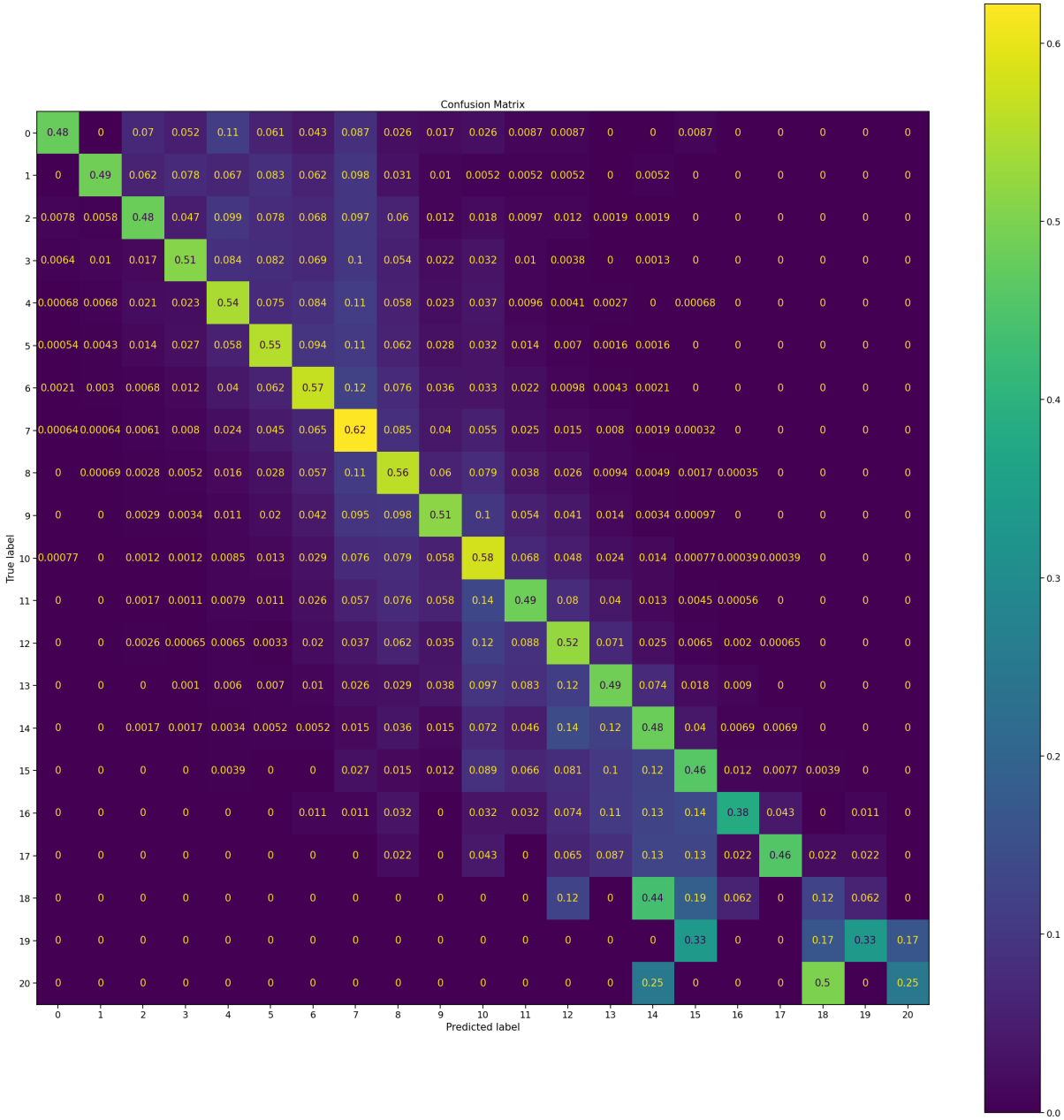

Figure 19: Confusion matrix for Label Encoded Random Forest

On the other hand, from Figure 20 it's clear that random forest classification that used the TF-IDF methodology has the most favorable convergence. This meant that the random forest classifier with the appropriate encoding method is able to correctly predict the quality of the wines from the descriptions of the reviews alone. Although the accuracy for the linear regression was still better than the random forest with TF-IDF for classification models the results showed that random forest was ideal.

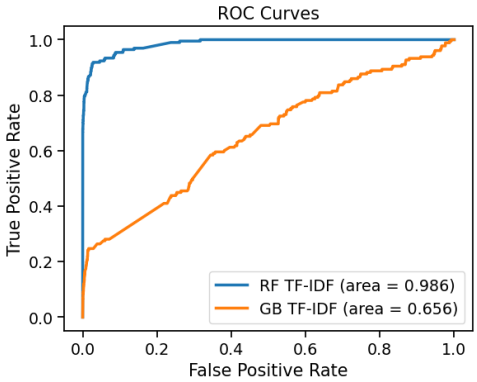

Figure 20: ROC Curves for TF-IDF classifications

The confusion matrix given in Figure 21 shows that compared to the confusion matrix of label encoding given in Figure 19 TF-IDF performance was well above label encoding for predicting the scores of the wines.

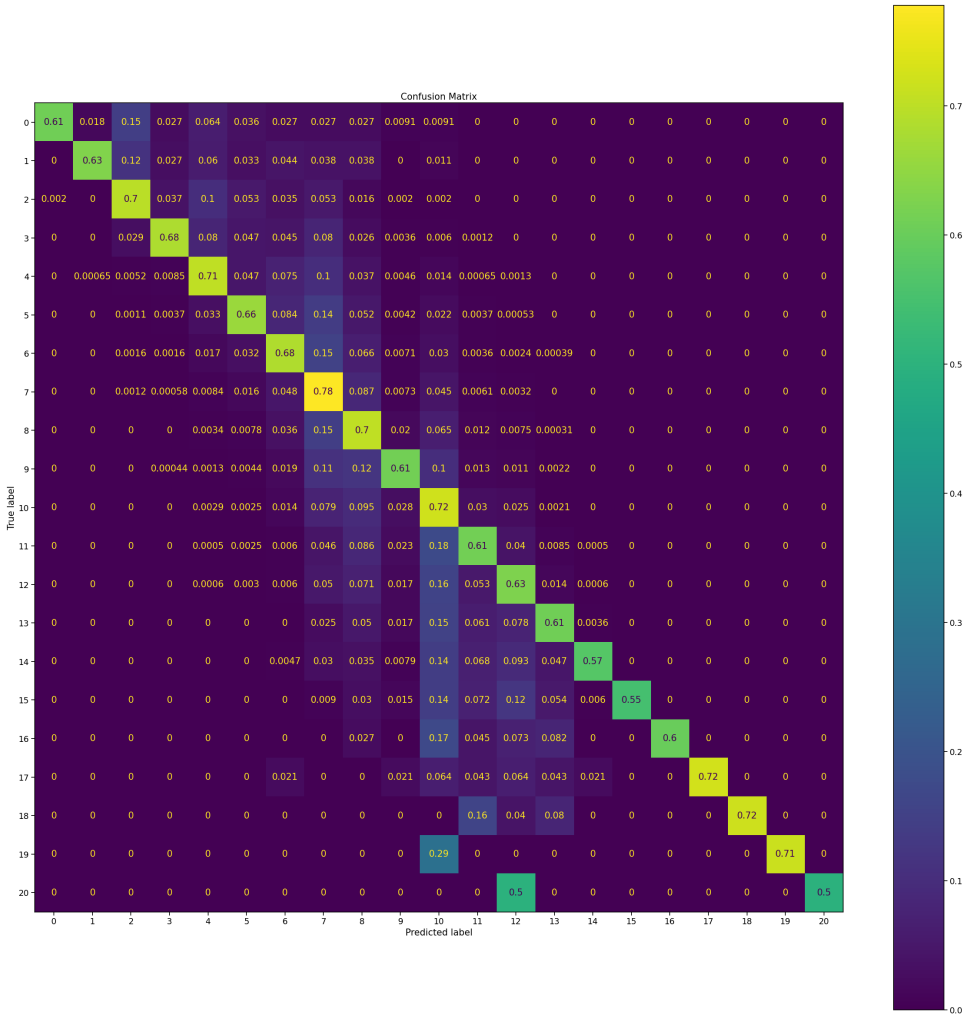

Figure 21: Confusion matrix for TF-IDF Random Forest

7 Discussion and Limitations

The aim of this research was to explore and analyze the Wine Reviews Dataset to answer various research questions related to wine characteristics, quality, and factors affecting ratings. The findings from the analysis provide valuable insights into the dataset and shed light on the relationships between different features.

One of the research questions focused on determining the years that produced the most high-quality wines. The analysis showed that the year 1969 had the highest average points for wines, indicating it as a good wine year. However, the quality declined in the following years and never reached the heights of the late 1960s. This suggests that there might be a specific period in the past that produced exceptional wines, and the quality has been decreasing over time. The analysis also explored the irregularities in wine quality across different countries. The standard deviation of wine points was used as a measure of irregularities. The findings revealed that countries like China and the United States had significant variations in wine quality, indicating irregularities within their wine production. This suggests that these countries might have diverse wine-making practices or variations in grape quality, leading to inconsistent wine ratings. Furthermore, the analysis investigated the countries and provinces that, on average, produced the best wines. England was found to be the country that produced the best wine over the years, followed by Austria and Germany. However, when considering the average points per province, the Südenburgland region of Austria emerged as the leader. This indicates that specific regions within a country can have a significant impact on the overall wine quality. The research also examined the quality/price ratio of wines and identified Switzerland as the country with the best ratio. The boxplot analysis displayed the top 25 countries with the best quality/price rates, providing insights into the affordability and quality of wines across different regions. To determine the features that have the most influence on wine quality, a correlation matrix was constructed. The results indicated that price and description were the most important features affecting wine ratings. This suggests that the cost of the wine and the description provided by the reviewers play a crucial role in determining the perceived quality of the wine. The analysis also delved into the characteristics of different wine varieties. It was observed that the most common variety was Pinot Noir, occurring 26,415 times in the dataset. Regions such as Napa Valley had the highest occurrence of Pinot Noir wines. The analysis also revealed variations in the number of wine varieties and their quality/price rates across different regions. Finally, the research explored the keywords that describe wine characteristics based on the reviews. The most common words used by reviewers included “wine,” “flavor,” “fruitiness,” and “finish.” This provides insights into the aspects of wines that reviewers focused on and considered important.

While this research provides valuable insights into the Wine Reviews Dataset and addresses several research questions, certain limitations were identified during the model fitting process that need to be addressed in future research:

- **Subjectivity of reviews**

  The dataset consists of subjective opinions and reviews from different wine tasters. The ratings and descriptions provided by the reviewers might vary based on individual preferences, experiences, and biases. This subjectivity could introduce inconsistencies in the analysis and interpretations.

- **Missing data**

  The dataset might contain missing values or incomplete entries, which could affect the accuracy and reliability of the analysis. For this project, extracting years from the title is not a reliable option. Entries with missing data had to be removed from the dataset, potentially leading to a loss of valuable information.

- **Categorical data challenges**

  The dataset contains categorical features such as country, region, variety, and winery. Analyzing and interpreting categorical data can be challenging, as it requires appropriate techniques such as encoding and feature engineering. One alternative to the encoding methods would be Target Encoding which as of writing this report is only available in the unstable development version of the scikit learn package. Additionally, for the subsequent studies hyperparameter tuning could be used to improve the accuracy of the models. Simply removing features with negative or zero correlation to wine quality wasn’t enough to improve the accuracy of the models.

8 References

Wine Reviews. (2017, November 27). Kaggle. https://www.kaggle.com/datasets/zynicide/wine-reviews

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. In Journal of Applied Science and Technology Trends (Vol. 1, Issue 4, pp. 140–147). Interdisciplinary Publishing Academia. https://doi.org/10.38094/jastt1457

Hosmer, D. W., & Lemeshow, S. (2000). Introduction to the Logistic Regression Model. In Applied Logistic Regression (pp. 1–30). John Wiley & Sons, Inc. doi: https://doi.org/10.1002/0471722146.ch1.

Khuri, A. I. (2013). Introduction to Linear Regression Analysis, Fifth Edition by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. In International Statistical Review (Vol. 81, Issue 2, pp. 318–319). Wiley. https://doi.org/10.1111/insr.12020_10

Rish, I. (2001) An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

McCallum, A., & Nigam, K. (1998) A comparison of event models for naive Bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

Steinwart, I., & Christmann, A. (2008) Support vector machines. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-77242-4

Liaw, A., & Wiener, M. (2002) Classification and regression by randomForest. R news, 2(3), 18-22.

Hastie, T., Tibshirani, R., & Friedman, J. (2009) The elements of statistical learning (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-84858-7

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. In Machine Learning (Vol. 63, Issue 1, pp. 3–42). Springer Science and Business Media LLC. https://doi.org/10.1007/s10994-006-6226-1

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J., & Zhang, X. (2011). K-Means Clustering. In Encyclopedia of Machine Learning (pp. 563–564). Springer US. https://doi.org/10.1007/978-0-387-30164-8_425

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models. In Machine Learning and Knowledge Discovery in Databases (pp. 67–82). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04180-8_22

Qaiser, S., Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. In International Journal of Computer Applications (Vol. 181, Issue 1, pp. 25–29). Foundation of Computer Science. https://doi.org/10.5120/ijca2018917395.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). k-Nearest Neighbor Classification. In Data Mining in Agriculture (pp. 83–106). Springer New York. https://doi.org/10.1007/978-0-387-88615-2_4

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. In Journal of Big Data (Vol. 7, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-020-00305-w.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. In Artificial Intelligence Review (Vol. 54, Issue 3, pp. 1937–1967). Springer Science and Business Media LLC. https://doi.org/10.1007/s10462-020-09896-5