sommAIlier

AI WINE RECOMMENDER

AI Wine Recommender System

Table of Contents

1. Project Overview

1.1 Company Background

In today's increasingly digital world, artificial intelligence (AI) is making significant inroads across various industries. AI technologies are reshaping decision-making processes, notably in the realm of wine selection. This project aims to build an AI-powered wine recommendation model for the conceptual startup “SommAIlier”. By leveraging machine learning, we aim to enrich and enhance the relationship between wine and consumers, striving to deliver tailored wine suggestions that align seamlessly with individual tastes and preferences (Vientur, 2023).

The advantages of AI-driven wine recommendations include personalization, the discovery of new wines and regions, and enhanced accuracy. The future of AI-powered wine recommendations holds exciting possibilities, with potential integrations of real-time data and augmented reality for an even more immersive wine selection journey. Embracing AI has the potential to elevate the wine experience for enthusiasts and casual drinkers alike. Furthermore, another notable benefit of leveraging AI is the increase in predictive accuracy of recommendations. What traditional recommender systems are not capable of is solving data sparsity and cold start problems. Implementation of AI can facilitate more efficient recommendations and revolutionize how consumers buy and learn about wine (Senevirathne, 2023).

1.2 Business Opportunity

The wine industry, characterized by its dynamic trends and diverse attributes, presents challenges for those new to the world of wine. Its vast variety of options, each with its own unique set of characteristics, creates a formidable barrier to entry. Factors such as acidity, body, ABV, wine type, and food pairings add to the complexity of wine selection. Novice wine enthusiasts often find the task of choosing the right wine daunting, and even experienced individuals may struggle to navigate the wine market, especially when exploring new brands or types.

In response to the intricate nature of the wine market and the challenges it poses, we have identified a promising business opportunity: the development of an AI-driven Wine Recommender System. By leveraging the capabilities of artificial intelligence, our objective is to streamline the wine selection process and elevate the overall wine experience. With the creation of “SommAIlier”, we aim to empower users to utilise this AI recommendation to make well-informed choices that align with their preferences. Simultaneously, we hope this tool will serve as an informative resource, assisting consumers in broadening their understanding of wine varieties.

1.3 Target Audience

“SommAIlier” is designed to cater to the needs of two primary audience segments: restaurants and individuals.

1.3.1 Restaurants

For restaurants, the system opens a valuable opportunity to enhance the dining experience for their customers. It has the potential to assist restaurants in the selection of wines that complement the restaurant’s menu, aligning with the diverse tastes and preferences of their customers. Furthermore, it aids in crafting curated wine lists that harmonize with the culinary offerings and ensure that the wines individuals want are not out-of-stock, maximising customer satisfaction.

1.3.2 Individuals

The system is equally beneficial for individual consumers, whether they are wine enthusiasts or casual wine drinkers. This is particularly beneficial for those who may need to gain more extensive knowledge and expertise in the field of wine. It simplifies the wine selection process, offering recommendations for wines that pair harmoniously with their meals and their preferences in wine.

Additionally, it can provide insights into wines that share similar characteristics, aiding in the exploration of wines.

2. Dataset

2.1 Source

The dataset was obtained from a recently published journal named “X-Wines: A Wine Dataset for Recommender Systems and Machine Learning”. The authors have published their dataset under a free license for wider use. Initially comprising 100,646 wine instances and 17 attributes, the dataset was sufficiently expansive to train our recommendation model. We made adjustments to the original dataset, streamlining it to remove complexities and retain only the columns pertinent to our business concept (Appendix A).

2.2 Descriptive Statistics

2.2.1 Numerical Variables

The average number of types of grapes used in making a wine is about 1.5. While the most complex wine is composed of 16 grape types, the third quartile is 2.0, indicating that at least 75% of wines are made of 2 or fewer grapes (Appendix B Table 2).

The mean and median ABV are 13.3 and 14.0 respectively, which is within the ABV range of a bottle of wine known as 5.5% to 20% (Gold, 2023). The average ABV of our dataset is slightly higher because it contains more red wines, which have relatively higher ABVs than other types of wine in general (Vinifera, 2018).

The number of vintages ranges from 1 to 73. 50% of the wine comes in a variation of 19 or fewer vintage years, and 75% of the wine is available in 27 or fewer vintage years.

2.2.2 Categorical Variables

There are 60,981 unique wine names in the dataset, and the most frequent wine name is ‘Cabernet Sauvignon’. The most common wine types, body, acidity and country are ‘Red’, ‘Full-bodied’, ‘High’ and ‘France’ respectively (Appendix B Table 3).

There are 346 unique combinations of food paired with each wine in the dataset. The combination of [‘Game Meat’, ‘White Meat’, ‘Red Meat’] is most frequently paired with the wines.

3. Exploratory Data Analysis

3.1 ABV by Wine Type

Port wines have the highest average alcohol content, which aligns with their reputation for having ABV levels typically running between 19% and 22% (Kranemann, 2021) (Appendix C Figure 1). The average ABV observed for table wines also corresponds with industry norms, as these wines generally fall within the 5.5% to 20% ABV range. Notably, Red and White table wines tend to have a relatively higher ABV, while Sparkling and Rosé wines typically have a lower ABV (Gold, 2023). However, it is imperative to acknowledge the presence of outliers within the category of table wines, displaying exceptionally low ABV levels, as well as dessert wines with notably elevated ABV. Addressing these outliers adequately is essential in the context of model development.

3.2 Wine Body and Acidity

It is widely understood that wines with higher acidity tend to create a lighter sensation on the palate for the drinker (Hale, 2023). While our dataset doesn't distinctly highlight this trend (Appendix C Figure 2), it is not a cause for concern. This is because there are numerous other factors intertwined with a wine's body and acidity that collectively contribute to the overall tasting experience, such as oak ageing, tannins, and glycerol (Maria, 2022). Upon further examination of the wine count for each body type (Appendix C Figure 3), we deduced that our dataset's inability to reflect the inverse correlation between wine body and acidity is due to an imbalance in data. Specifically, the dataset is skewed, with a predominance of full-bodied and medium-bodied wines.

3.3 Distribution of Wine Type Across Countries

France predominantly produces Red wines, with a count of 12,383, closely followed by Italy (Appendix C Figure 4). While countries known for their rich wine heritage like France and Italy tend to lead across most wine types, a few countries have a pronounced edge in some wine types. For instance, Portugal tops the dataset for Dessert wines, nearly doubling the listings of France and Italy, which aligns with its fame for Port wines (*What Is Port Wine? | Wine Folly*, n.d.). Moreover, if we take the dataset as an accurate reflection of global wine production, it appears that countries with more wineries tend to produce higher quantities of wine regardless of type (Appendix C Figure 5).

4. Model

4.1 Description

Our classification models assume users do not know their preference for the type of wine. They provide inputs such as wine body, acidity, country of production, grape variety, alcohol by volume (ABV), and food pairing. These are the 6 features of our wine recommendation model. “SommAllier” then recommends the best wine type (target) based on these features, ensuring the recommended wine type is suitable for the users' needs. The top 3 wine names of the recommended wine type will be provided, along with other relevant information such as the vintages available, the name of the winery, and its website, among others.

4.2 Cleaning and Preprocessing

4.2.1 Cleaning

Body: For the body of wine, the substring ‘-bodied’ was removed, such that the values now become: Full, Medium, Light, Very Light and Very Full.

Country: In the original dataset, multiple countries appeared very few times (Appendix D Figure 6), so predictions might be inaccurate using those countries as inputs. To combat this, we filtered countries that occurred less than 100 times, ensuring that the country appeared in both the training and test data.

Grapes: With the idea that novice wine drinkers are unfamiliar with the many sorts of grapes used in a wine, we opted to mention the number of grapes used instead. The number of grapes used could indicate the wine's complexity (Wang & Spence, 2018). Hence, for each row, we counted the number of grapes used to make the wine.

Harmonise: As wines can be paired with multiple foods at the same time (Appendix D Figure 7), we decided to categorise the food into the following food types: Red Meat, White Meat, Game Meat, Cheese, Seafood, Italian, Dessert, Vegetarian, Snacks, Appetiser, Cured Meat, Spicy Food, and Others. This was done to avoid an overly extensive training dataset and to prevent overwhelming users with too many choices, which could complicate the model and reduce its user-friendliness.

4.2.2 Preprocessing

Data Splitting: We first split the dataset into catalogue and training data. 80,000 rows were randomly allocated to the catalogue dataset and 19,867 rows to the training dataset. The catalogue data acted as the wine catalogue of our business for later use when making wine recommendations to system users. The training data was again split into train (80%) and test (20%) datasets to build a robust recommendation model.

One-hot Encoding: All categorical variables in the dataset (Body, Acidity, Country, Harmonise) were one-hot encoded before model training. While Body, Acidity, and Country were one-hot encoded together by defining a transformer function, Harmonise was encoded separately from other variables because it contained multiple values in one row. For instance, we aimed to prevent a row with ‘Seafood, Red Meat’ becoming a singular category rather than two separate categories of ‘Seafood’ and ‘Red Meat’.

4.3 Model Building

To address the data imbalance highlighted in Appendix D Figure 8, we employed ensemble techniques, as they have demonstrated higher efficiency compared to sampling methods such as undersampling and oversampling (Feng et al., 2018). We used the following 3 random ensemble classification algorithms: Random Forest, Multi-Layer Perceptron (MLP), and XGBoost.

Random Forest (*RandomForestClassifier*)**:** By choosing ‘entropy’ as the model’s *criterion*, we aimed to build a decision tree that prioritizes splits, providing the greatest information gain about wine preferences. *class_weight* = ‘balanced’ was used to address the class imbalance by inversely adjusting the weights of the classes based on their occurrence frequency in the data.

Multi-Layer Perceptron (*MLPClassifier*)**:** *max_iter* = 25000 was set to prevent potential infinite loops in case the model does not converge (i.e., does not find an optimal solution). *random_state* = 100 was to ensure that any process in the code that requires randomness produced consistent results across runs and the same clusters every time.

XGBoost (*XGBClassifier*)**:** Unlike the previous algorithms, which could handle categorical targets with explicit encoding, XGBoost required both input features and target labels in numerical format. Hence, the target variable (‘type’) was label encoded with values between 0 and n-1 before building the model. *random_state* = 100 was also used for the same reason.

Upon building the models, k-fold cross-validation, with a k value of 10, was utilized to mitigate overfitting and ensure that the test scores were a true representation of a model’s performance (Appendix E).

4.4 Hyperparameter Tuning

Hyperparameter tuning was done using *GridSearch* to find the best possible combination of hyperparameters for each model used. The top 5 hyperparameter combinations of each model by mean test scores were compared in a table (Appendix F) to see in better detail how a change in a single hyperparameter affects the model performance.

Random Forest: The hyperparameters tuned were *n_estimators*, *max_depth*, *min_samples_split* and *min_samples_leaf*.

Multi-Layer Perceptron: The hyperparameters tuned were *hidden_layer_sizes*, *activation function*, *alpha* (strength of L2 regularization), and *learning_rate*.

XGBoost: The hyperparameters tuned were *n_estimators*, *max_depth*, *learning_rate*, *min_child_weight* and *gamma* (minimum loss function).

4.5 Model Performance

We assessed the performance of each model using a classification report and a confusion matrix. Additionally, to gauge the improvement in performance after tuning, we compared the AUC of the initial model with that of the best model and calculated the percentage growth in AUC (Appendix G).

Random Forest: The model achieved a weighted average F1 score or accuracy of 0.93. For specific wine types, the F1 score reached 0.97 for Red wines, given their significant support of 2202, while it was only 0.60 for Rosé wines, reflecting their lower support of 146. After tuning, the model's performance saw an improvement of 3.372%, with the AUC rising from 0.949 to 0.981.

Multi-Layer Perceptron Classifier: The model achieved a weighted average F1 score or accuracy of 0.94. For a specific wine type, the F1 score was as high as 0.98 for Red wines and as low as 0.68 for Rosé wines. After tuning, the model’s performance saw an improvement of 1.363%, with the AUC rising from 0.954 to 0.967

XGBoost Classifier: The model achieved a weighted average F1 score or accuracy of 0.95. For a specific wine type, the highest F1 score was 0.98 for Red wines and lowest for Rosé wines (0.69). After tuning, the model’s performance saw an improvement of 0.826%, with the AUC rising from 0.969 to 0.977.

4.6 Model Comparison

All three models performed very well in predicting wine types as shown in their high F1 scores or accuracy (Appendix G). Some models performed slightly better in predicting certain wine types than others. For instance, the Multi-Layer Perceptron Classifier was better in predicting Red, Rosé, and White wines than the Random Forest Classifier, but the Random Forest Classifier was better in predicting Dessert and Sparkling wines than the Multi-Layer Perceptron Classifier.

Using AUC as the evaluation metric, the Random Forest Classifier initially exhibited the weakest performance in its base model. Yet, after refining its hyperparameters, it emerged as the best model with an AUC score of 0.981 (Appendix H). Thus, we decided to use the Random Forest Classifier model to predict wine types for “SommAIlier” users.

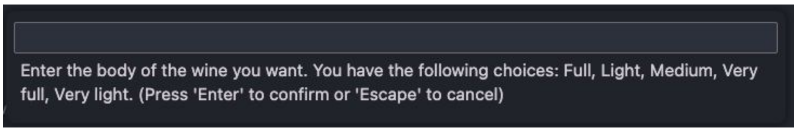
Furthermore, for all three models, whether the wine harmonizes with seafood is very important and would have a huge implication on the type of wine recommended. Random Forest seems to be the only model that places high importance on Alcohol by Volume (ABV) (Appendix I).

4.7 Recommendation Function

The ratings dataset shows the customers’ ratings (that range from 1 to 5) after tasting the wines they had. The dataset contains 21 Million rows with each row representing one customer’s rating for the specific wine that was consumed. It includes information such as the ‘RatingID’, ‘UserID’, ‘WineID’, ‘Vintage’, ‘Rating’ and ‘Date’ (Appendix J Table 4).

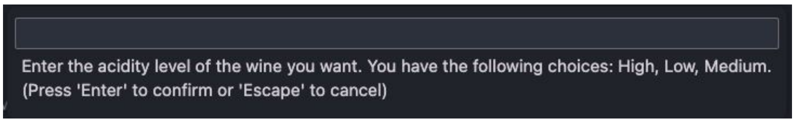
In order to define the recommender function, the team decided to focus on the ‘WineID’ and ‘Ratings’ columns from the ratings dataset. As reflected in the final dataset used for the function, the ‘Ratings’ column from the ratings dataset was used to compute the average ratings of wines, and this computation is now known as ‘Avg_Ratings’ to predict the type of wine (Appendix K).

Our AI-driven wine recommendation system is crafted to suggest wine types tailored to users' preferences, ensuring they receive the best wine suggestions aligned with their tastes. The user can input one’s preference in terms of wine body, acidity, country of production, and number of grapes used in making the wine. The function also takes into account the user’s food pairing choices as well as the Alcohol by Volume (ABV) to suggest wines that complement the food’s flavors and textures. The pictures below are the order of prompts users will see:



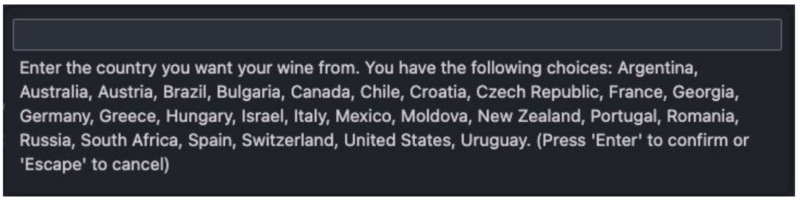
Enter the body of the wine you want. You have the following choices: Full, Light, Medium, Very full, Very light. (Press 'Enter' to confirm or 'Escape' to cancel)

Figure 1: Users Entering the Body of Wine



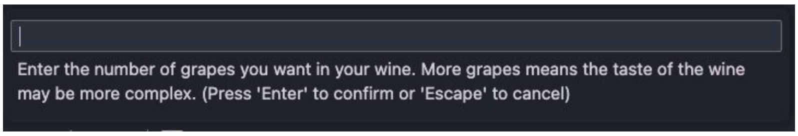
Enter the acidity level of the wine you want. You have the following choices: High, Low, Medium. (Press 'Enter' to confirm or 'Escape' to cancel)

Figure 2: Entering the Acidity Level



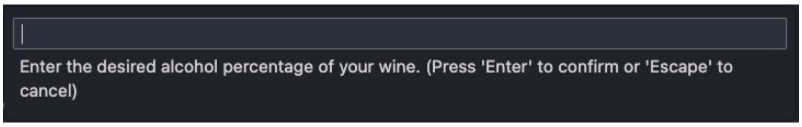
Enter the country you want your wine from. You have the following choices: Argentina, Australia, Austria, Brazil, Bulgaria, Canada, Chile, Croatia, Czech Republic, France, Georgia, Germany, Greece, Hungary, Israel, Italy, Mexico, Moldova, New Zealand, Portugal, Romania, Russia, South Africa, Spain, Switzerland, United States, Uruguay. (Press 'Enter' to confirm or 'Escape' to cancel)

Figure 3: Entering the Country of Wine



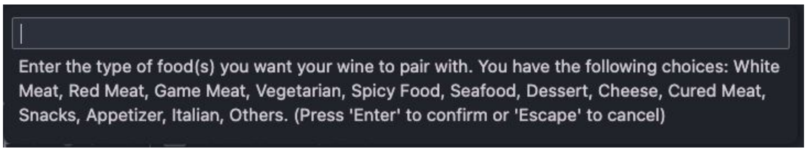
Enter the number of grapes you want in your wine. More grapes means the taste of the wine may be more complex. (Press 'Enter' to confirm or 'Escape' to cancel)

Figure 4: Stating the Number of Grapes (Indicator of Wine Complexity)



Enter the desired alcohol percentage of your wine. (Press 'Enter' to confirm or 'Escape' to cancel)

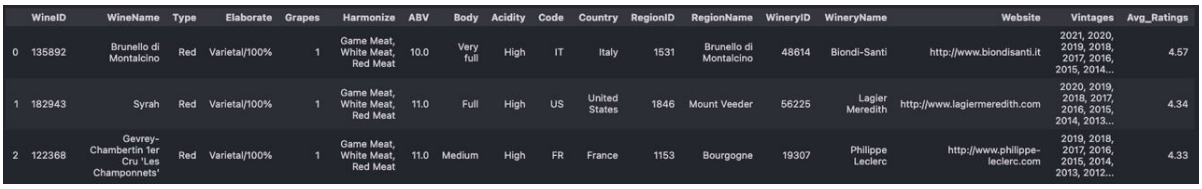
Figure 5: Desired Alcohol Percentage



Enter the type of food(s) you want your wine to pair with. You have the following choices: White Meat, Red Meat, Game Meat, Vegetarian, Spicy Food, Seafood, Dessert, Cheese, Cured Meat, Snacks, Appetizer, Italian, Others. (Press 'Enter' to confirm or 'Escape' to cancel)

Figure 6: Type of Food to Pair With

For example, the inputs were: Full (body), High (acidity), Italy, 1 (grape), 10.9 (alcohol percentage), Red Meat. After inputting these answers, the user will see the top 3 wines recommended by the system and will choose from them:



	WineID	WineName	Type	Elaborate	Grapes	Harmonize	ABV	Body	Acidity	Code	Country	RegionID	RegionName	WineryID	WineryName	Website	Vintages	Avg_Ratings
0	135892	Brunello di Montalcino	Red	Varietal/100%	1	Game Meat, White Meat, Red Meat	10.0	Very full	High	IT	Italy	1531	Brunello di Montalcino	48614	Biondi-Santi	http://www.biondisanti.it	2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014...	4.57
1	182943	Syrah	Red	Varietal/100%	1	Game Meat, White Meat, Red Meat	11.0	Full	High	US	United States	1846	Mount Veeder	56225	Lagier Meredith	http://www.lagiermeredith.com	2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013...	4.34
2	122368	Gevrey-Chambertin 1er Cru 'Les Champonnets'	Red	Varietal/100%	1	Game Meat, White Meat, Red Meat	11.0	Medium	High	FR	France	1153	Bourgogne	19307	Philippe Leclerc	http://www.philippe-leclerc.com	2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012...	4.33

Figure 7: Output Table for the User

For added user personalization, only ABV values that have a 1 standard deviation around the ABV as given by the user will be considered. For example, if a user inputs 11% as their desired ABV, the recommender function will only output the top 3 wines that are around 11% ABV.

5. Potential Recommendations

The first recommendation to improve the wine recommender system is to obtain more data to balance the wine types provided in the dataset. Currently, we have a limitation where the number of wines for each wine type is disproportionate, which can lead to bias in the system.

The second recommendation is to incorporate more wine brands into the dataset by avid drinkers who can suggest drinks that will fit the food choices available. By doing so, the recommender system can provide more accurate and diverse recommendations that cater to individual tastes and preferences.

The third recommendation is to bring in UI/UX designers to design an interface for the recommender system to look more appealing, interactive, and easy to use. Currently, the recommender system relies on the guidance of knowledgeable representatives to input customer preferences and interpret the generated recommendations. Enhancing the user interface allows for a more direct customer interaction with the recommender system. For instance, instead of restaurant staff operating the system post customer preference indication, customers can immediately engage with it through iPads installed at their tables. As for the individual customers, a better user interface will also improve the accessibility and intuitiveness of the system, extending the system's appeal to a wider audience. As a result, this enhancement will not only elevate the user experience but also expand the potential reach of the target market for the recommender.

6. Benefits to Key Stakeholders

6.1 Restaurants

Restaurants that act as our wine retailers would use this system to increase their revenue and offer personalised recommendations to customers. Furthermore, restaurants can improve the efficiency of their inventory management by predicting demand for the different types of wines that they can offer to customers. By identifying common trends in the types of wines that complement their menu, restaurants can use this insight to expand their wine selection. This could involve sourcing a broader range of wines from various countries that pair effectively with their dishes.

6.2 Individuals

The recommender system can help restaurant patrons discover new wines that match their wine preferences and food pairing choices. This would help enhance their drinking experience and dining satisfaction in the restaurants. Moreover, customers who lack wine knowledge and would like to learn how to enjoy and appreciate drinking wine would find this experience fruitful and eye-opening. This also helps restaurants attract more customers.

7. Limitations

One of the limitations is that we did not take into consideration the number of reviews or ratings for each of the wines. When implementing a wine recommendation system, some choose to rely on the wine ratings to make sure the wines are of high-quality and satisfy most the customers' tastes. Despite this, to keep our recommendation's objective and reduce the popularity bias, we focused on providing content-based recommendations based on the consistent attributes of wine input by users (i.e., type, body, acidity, etc.).

Another limitation is that the prices for each wine were not considered as price information was not available in the dataset we used. There might be occasions when customers are more price-conscious and have a certain budget set aside for wine. This would mean that the price could be an important factor that they consider when purchasing the wine. Despite this missing factor, customers who are planning to drink wines would generally have a rough estimation of how much they would cost. Hence, to improve the recommender function further, we will include the prices of wines and ensure that these prices are able to be automatically updated.

With regards to our model, though the ensemble classification model we employed is better at avoiding overfitting compared to single models, there may still be chances of overfitting. This can happen if the model is too complex for the data (data is not diverse enough). Adding to that, the model requires hyperparameter tuning to find the optimal values for model performance. Thus, we tried to mitigate these limitations by engaging in k-fold cross validation techniques and an extensive hyperparameter tuning process to offer accurate wine recommendations to our target audience.

8. Conclusion

As the team had a common interest in wanting to make wine selections more enjoyable, especially for those relatively new to drinking wines, the team decided to use AI to develop a recommender system. With that, the team decided to establish our startup company called "SommAIllier" and build an AI-powered wine recommendation model to help those who are new to the wine industry.

After conducting the various analyses, cleaning, and preprocessing, as well as developing and tuning models, three models were developed to identify the best one out of the three. Random forest was the best model that obtained the highest AUC score of 0.981 which suggests that it can successfully classify 98.1% of true positives as positive. This also indicates that the model can accurately distinguish between the different types of wines, making the recommender a reliable and accurate system.

Furthermore, by addressing the unique needs of restaurants and individual customers as well as highlighting the benefits stakeholders can reap from, "SommAIllier" aspires to make the world of wine more approachable and enjoyable, simplifying the selection process and enriching the overall wine experience.

9. Pitch Write-Up

In today's rapidly evolving digital landscape, artificial intelligence (AI) is revolutionizing various industries, including the world of wine. The team would like to introduce "SommAIllier", a cutting-edge AI-powered wine recommendation model aimed at enhancing and personalizing the wine experience for restaurants and individuals. Our innovative solution leverages machine learning algorithms to provide tailored wine suggestions that align seamlessly with individual tastes and preferences. This pitch will discuss various segments of our findings from the report.

Firstly, the team would give an overview of the current wine industry and the importance of adopting AI in wine industries, specifically for wine recommendation. Secondly, a brief explanation of how our key value propositions align with the two primary target audiences, restaurants and individuals, will be provided. This will answer one big question: "Why would our target audience be interested in our product?". Thirdly, we have performed exploratory data analysis to delve deeper into understanding the dataset and investigate its main characteristics. These characteristics include ABV, body, wine type, and food pairing - essential information to be fed into the model and to provide accurate recommendations to wine consumers. Fourthly, we explained some of the assumptions we undertook for our classification model, methodologies of cleaning and preprocessing, and employment of ensemble techniques. This section serves to guide the readers through our thought process and justify why we have used the three popular ensemble classification algorithms. We have also determined the best model to predict wine types for "SommAIlier" users, which is the Random Forest Classifier model. Fifthly, after developing our wine recommender system, we have come up with some potential recommendations to better improve our product in terms of predictive quality and interactivity. Sixthly, we summarised the overall benefits to key stakeholders to show why there is a need to use AI to solve this problem. Lastly, before we conclude, we give some limitations to our overall model and justify why this is not a major issue for our final outcome.

"SommAIllier" has the potential to revolutionise the world of wine, making it more approachable and enjoyable for everyone. With the increasing demand for wine, there has never been a better time to embrace the power of "SommAIllier". By using our system, you can unlock a world of new flavors and experiences, discovering the perfect wine for any occasion. Don't miss out on this incredible opportunity to elevate your wine game and join the ranks of the world's most discerning wine connoisseurs. Try "SommAIllier" today and experience the future of wine selection for yourself, as wine selections will never be the same again!

10. References

Bossart, C. (2023, June 21). *Why you should be drinking dessert wine with dinner*. Food & Wine. https://www.foodandwine.com/sweet-wine-7550176#:~:text=This%20sweetness%20is%20often%20balanced,or%20enjoyed%20on%20their%20own.

Feng, W., Huang, W., & Ren, J. (2018). Class imbalance ensemble learning based on the margin theory. *Applied Sciences*, *8*(5), 815. https://doi.org/10.3390/app8050815

Gold, B. (2023, September 11). Here's How Much Alcohol Is in Wine, From Lowest to Highest. *Real Simple*. https://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine

Hyken, S. (2015). CX and EX (Customer Experience and Employee Experience). *Shep Hyken | Customer Service Expert*. https://hyken.com/customer-experience/cx-ex-customer-experience-employee-experience/

Mei, F. L. (2018, January 4). Pop-up power: Short-term shops stoke FOMO. *Campaign Asia*. https://www.campaignasia.com/article/pop-up-power-short-term-shops-stoke-fomo/441906

Senevirathne, R. (2023). Unleashing the power of artificial intelligence in wine selection: Smart Wine Recommendations. https://www.linkedin.com/pulse/unleashing-power-artificial-intelligence-wine-smart-ruchira#:~:text=AI%2Dpowered%20wine%20recommendations%20have,wines%2C%20regions%2C%20and%20flavors.

Trivium Packaging. (2022, April 22). New Data Reveals Preference for Sustainable Packaging Remains Strong in a Changing World. *CISION PR Newswire*. https://www.prnewswire.com/news-releases/new-data-reveals-preference-for-sustainable-packaging-remains-strong-in-a-changing-world-301530676.html

Vinetur. (2023, September 19). How Artificial Intelligence (AI) is transforming the wine industry. *Vinetur the Ultimate Wine Magazine*. https://www.vinetur.com/en/2023091975351/how-artificial-intelligence-ai-is-transforming-the-wine-industry.html

Vinifera. (2018, May 11). Does red wine or white wine have more alcohol? | Wine Spectator. *Wine Spectator*. https://www.winespectator.com/articles/which-wine-type-has-more-alcohol-56456#:~:text=There%20are%20exceptions%20but%2C%20in,grapes%20when%20they%20were%20harvested.

Wang, Q. J., & Spence, C. (2018). Wine complexity: An empirical investigation. *Food Quality and Preference*, *68*, 238–244. https://doi.org/10.1016/j.foodqual.2018.03.011 *What is Port Wine? | Wine Folly*. (n.d.). Wine Folly. https://winefolly.com/deep-dive/what-is-port-wine/

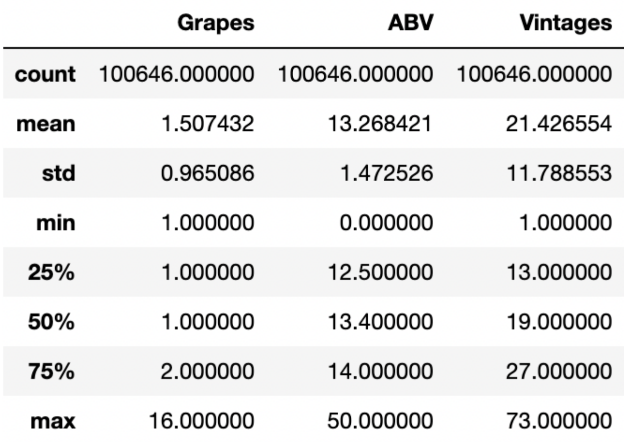
11. Appendices

11.1 Appendix A

WineName	Name of a wine
Type	Type of a wine [‘Red’, ‘White’, ‘Sparkling’, ‘Rosé’, ‘Dessert’, ‘Dessert/Port’]
Grapes	Number of unique grape types used in producing a wine
Harmonize	List of food that best pairs with a wine
ABV	The alcohol content in a wine represented in percentage
Body	Richness and weight of a wine in a drinker’s mouth [‘Very light-bodied’, ‘Light-bodied’, ‘Medium-bodied’, ‘Full-bodied’, ‘Very full-bodied’]
Acidity	Tartness of a wine [‘Low’, ‘Medium’, ‘High’]
Country	The country where a wine has been produced
Vintages	Number of years that grapes were harvested for a wine

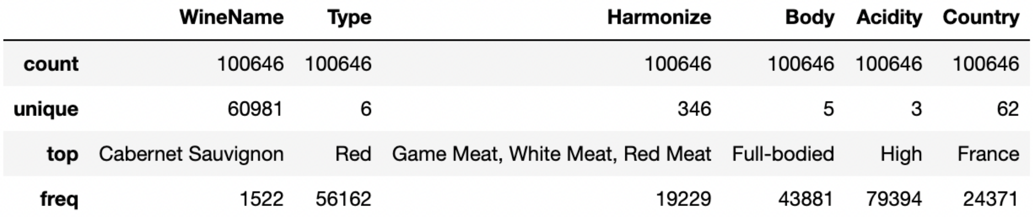
Table 1. Columns in the dataset after cleaning

11.2 Appendix B



	Grapes	ABV	Vintages
count	100646.000000	100646.000000	100646.000000
mean	1.507432	13.268421	21.426554
std	0.965086	1.472526	11.788553
min	1.000000	0.000000	1.000000
25%	1.000000	12.500000	13.000000
50%	1.000000	13.400000	19.000000
75%	2.000000	14.000000	27.000000
max	16.000000	50.000000	73.000000

Table 2. Descriptive statistics for numerical variables



	WineName	Type	Harmonize	Body	Acidity	Country
count	100646	100646	100646	100646	100646	100646
unique	60981	6	346	5	3	62
top	Cabernet Sauvignon	Red	Game Meat, White Meat, Red Meat	Full-bodied	High	France
freq	1522	56162	19229	43881	79394	24371

Table 3. Descriptive statistics for categorical variables

11.3 Appendix C

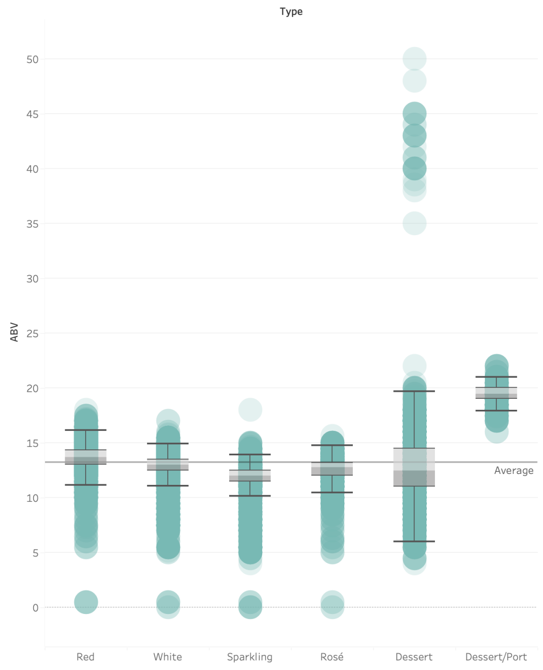


Figure 1. Box Plot of ABV by Wine Type

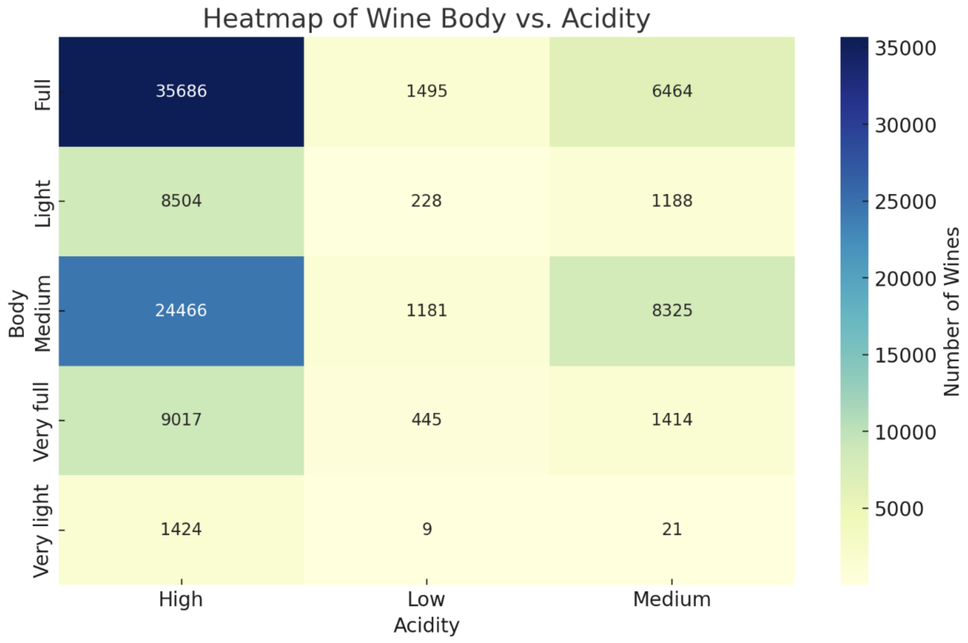


Figure 2. Heatmap of Wine Body and Acidity

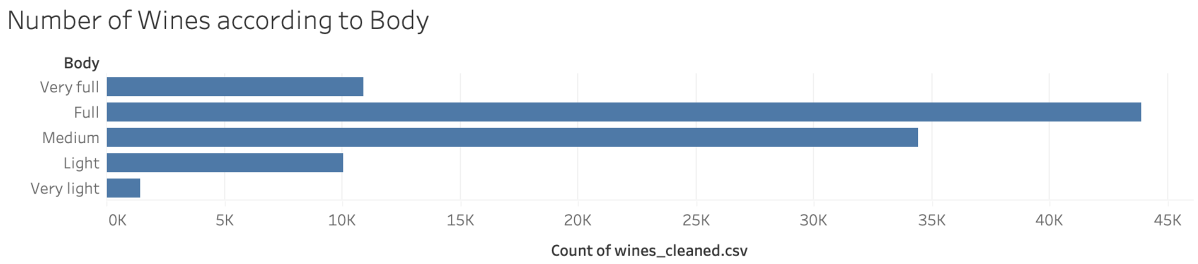
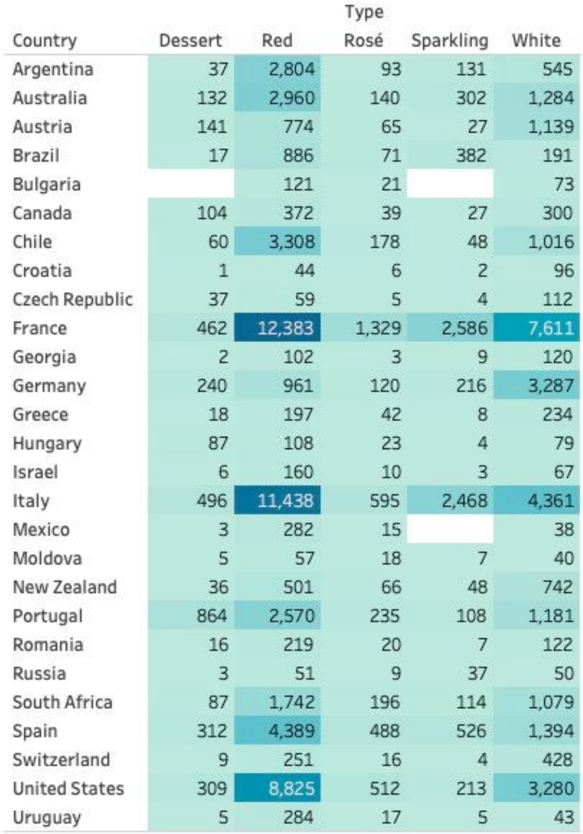


Figure 3. Number of Wines for Different Wine Bodies



	Type				
Country	Dessert	Red	Rosé	Sparkling	White
Argentina	37	2,804	93	131	545
Australia	132	2,960	140	302	1,284
Austria	141	774	65	27	1,139
Brazil	17	886	71	382	191
Bulgaria		121	21		73
Canada	104	372	39	27	300
Chile	60	3,308	178	48	1,016
Croatia	1	44	6	2	96
Czech Republic	37	59	5	4	112
France	462	12,383	1,329	2,586	7,611
Georgia	2	102	3	9	120
Germany	240	961	120	216	3,287
Greece	18	197	42	8	234
Hungary	87	108	23	4	79
Israel	6	160	10	3	67
Italy	496	11,438	595	2,468	4,361
Mexico	3	282	15		38
Moldova	5	57	18	7	40
New Zealand	36	501	66	48	742
Portugal	864	2,570	235	108	1,181
Romania	16	219	20	7	122
Russia	3	51	9	37	50
South Africa	87	1,742	196	114	1,079
Spain	312	4,389	488	526	1,394
Switzerland	9	251	16	4	428
United States	309	8,825	512	213	3,280
Uruguay	5	284	17	5	43

Figure 4. Distribution of Wine Type Across Countries

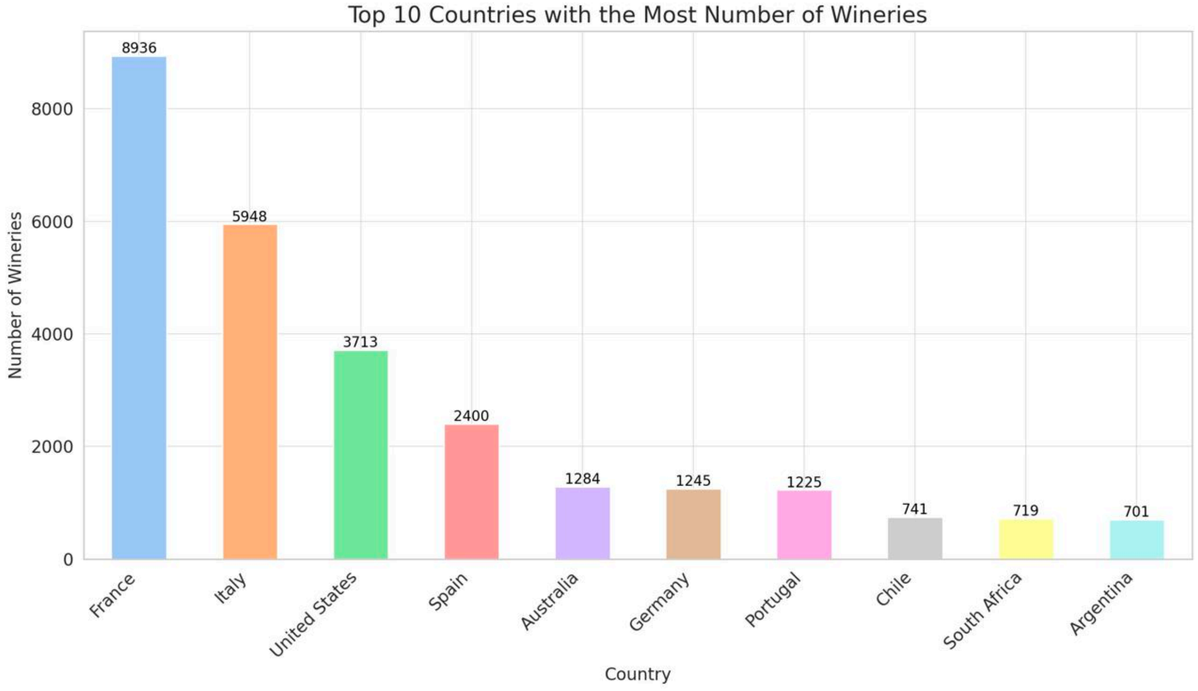


Figure 5. Top 10 Countries with the Most Number of Wineries

11.4 Appendix D

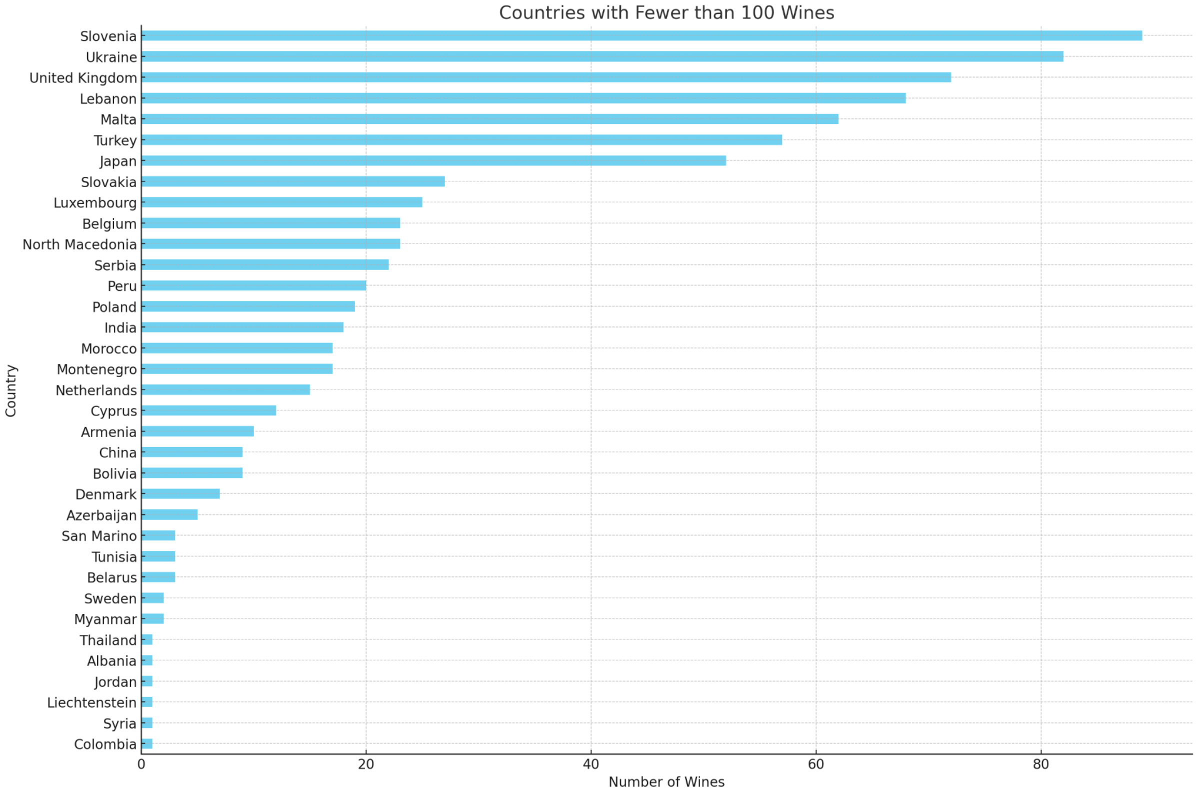
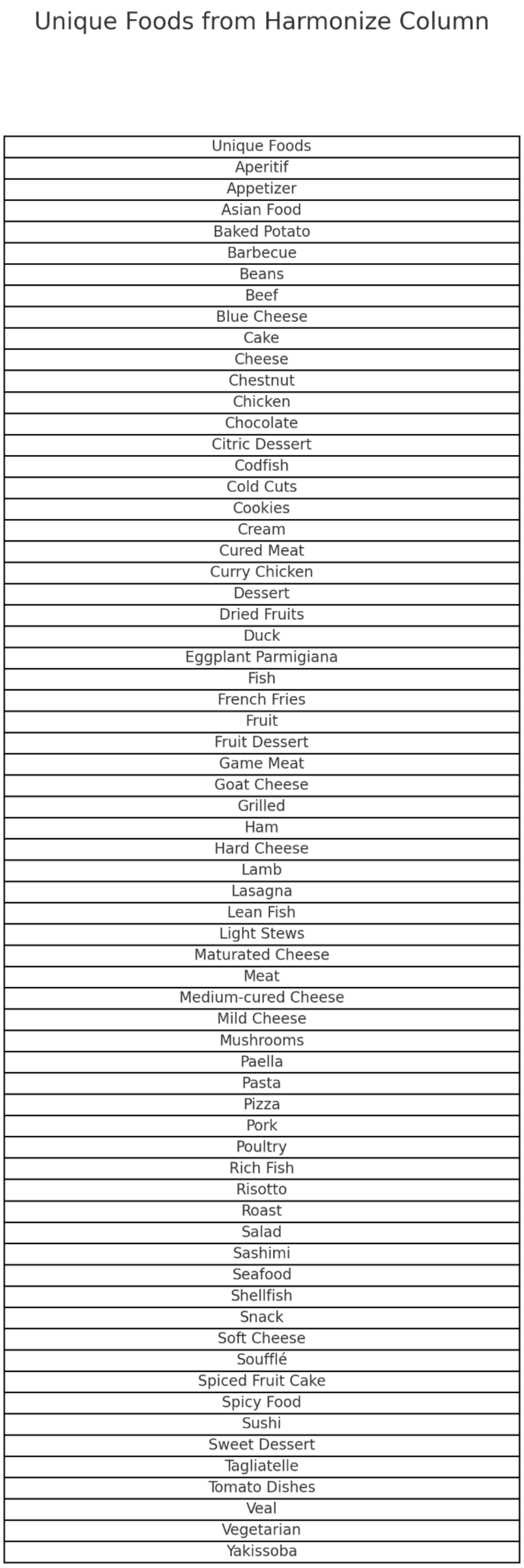


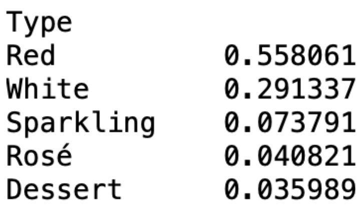
Figure 6. Countries with Fewer than 100 Wines



Unique Foods from Harmonize Column

Unique Foods
Aperitif
Appetizer
Asian Food
Baked Potato
Barbecue
Beans
Beef
Blue Cheese
Cake
Cheese
Chestnut
Chicken
Chocolate
Citric Dessert
Codfish
Cold Cuts
Cookies
Cream
Cured Meat
Curry Chicken
Dessert
Dried Fruits
Duck
Eggplant Parmigiana
Fish
French Fries
Fruit
Fruit Dessert
Game Meat
Goat Cheese
Grilled
Ham
Hard Cheese
Lamb
Lasagna
Lean Fish
Light Stews
Maturated Cheese
Meat
Medium-cured Cheese
Mild Cheese
Mushrooms
Paella
Pasta
Pizza
Pork
Poultry
Rich Fish
Risotto
Roast
Salad
Sashimi
Seafood
Shellfish
Snack
Soft Cheese
Soufflé
Spiced Fruit Cake
Spicy Food
Sushi
Sweet Dessert
Tagliatelle
Tomato Dishes
Veal
Vegetarian
Yakissoba

Figure 7. Unique Foods from ‘Harmonize’ Column



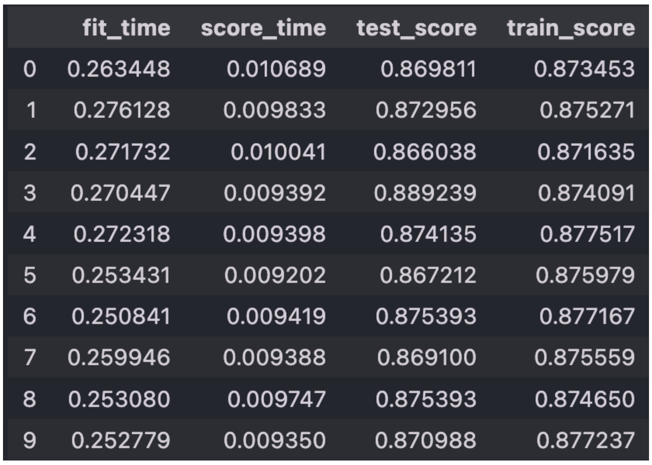
```
Type
Red          0.558061
White        0.291337
Sparkling    0.073791
Rosé         0.040821
Dessert      0.035989
```

Figure 8. Evidence of imbalance data

11.5 Appendix E

Cross-Validation (CV = 10)

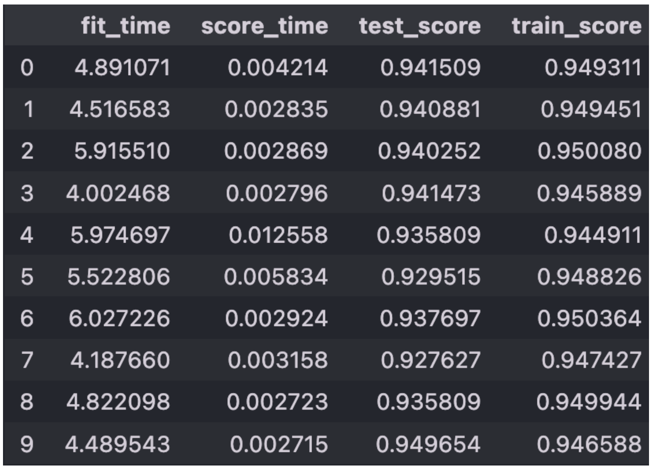
Random Forest Classifier



	fit_time	score_time	test_score	train_score
0	0.263448	0.010689	0.869811	0.873453
1	0.276128	0.009833	0.872956	0.875271
2	0.271732	0.010041	0.866038	0.871635
3	0.270447	0.009392	0.889239	0.874091
4	0.272318	0.009398	0.874135	0.877517
5	0.253431	0.009202	0.867212	0.875979
6	0.250841	0.009419	0.875393	0.877167
7	0.259946	0.009388	0.869100	0.875559
8	0.253080	0.009747	0.875393	0.874650
9	0.252779	0.009350	0.870988	0.877237

Mean test score: 0.873 (3 d.p)

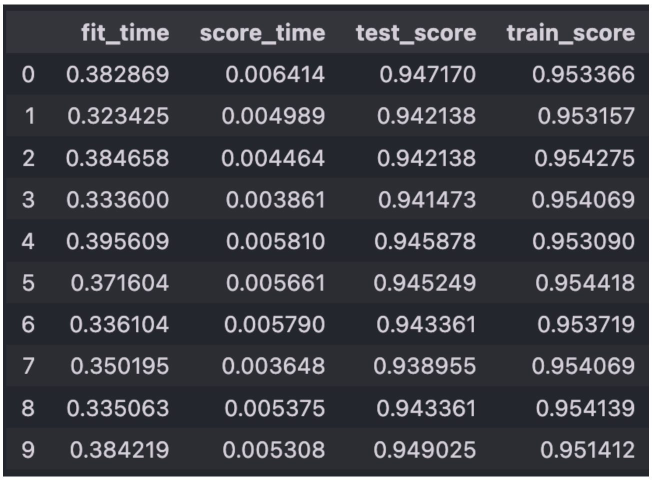
Multi-Layer Perceptron Classifier



	fit_time	score_time	test_score	train_score
0	4.891071	0.004214	0.941509	0.949311
1	4.516583	0.002835	0.940881	0.949451
2	5.915510	0.002869	0.940252	0.950080
3	4.002468	0.002796	0.941473	0.945889
4	5.974697	0.012558	0.935809	0.944911
5	5.522806	0.005834	0.929515	0.948826
6	6.027226	0.002924	0.937697	0.950364
7	4.187660	0.003158	0.927627	0.947427
8	4.822098	0.002723	0.935809	0.949944
9	4.489543	0.002715	0.949654	0.946588

Mean test score: 0.938 (3 d.p)

XGBoost Classifier

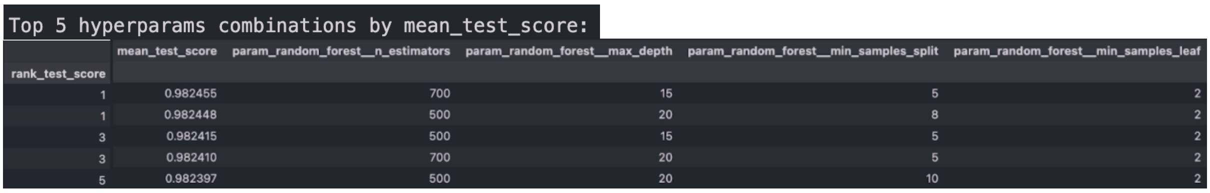


	fit_time	score_time	test_score	train_score
0	0.382869	0.006414	0.947170	0.953366
1	0.323425	0.004989	0.942138	0.953157
2	0.384658	0.004464	0.942138	0.954275
3	0.333600	0.003861	0.941473	0.954069
4	0.395609	0.005810	0.945878	0.953090
5	0.371604	0.005661	0.945249	0.954418
6	0.336104	0.005790	0.943361	0.953719
7	0.350195	0.003648	0.938955	0.954069
8	0.335063	0.005375	0.943361	0.954139
9	0.384219	0.005308	0.949025	0.951412

Mean test score: 0.944 (3 d.p)

11.6 Appendix F

Model Hyperparameter Tuning



Top 5 hyperparams combinations by mean_test_score:

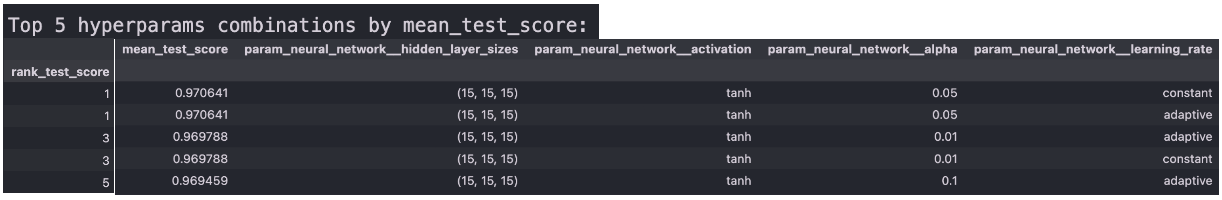
	mean_test_score	param_random_forest__n_estimators	param_random_forest__max_depth	param_random_forest__min_samples_split	param_random_forest__min_samples_leaf
rank_test_score					
1	0.982455	700	15	5	2
1	0.982448	500	20	8	2
3	0.982415	500	15	5	2
3	0.982410	700	20	5	2
5	0.982397	500	20	10	2

Random Forest Classifier

Best hyperparameters:
N_estimators = 700, Max_depth = 15, Min_samples_leaf = 2, Min_samples_split = 5

Best mean test score = 0.982

Multi-Layer Perceptron Classifier



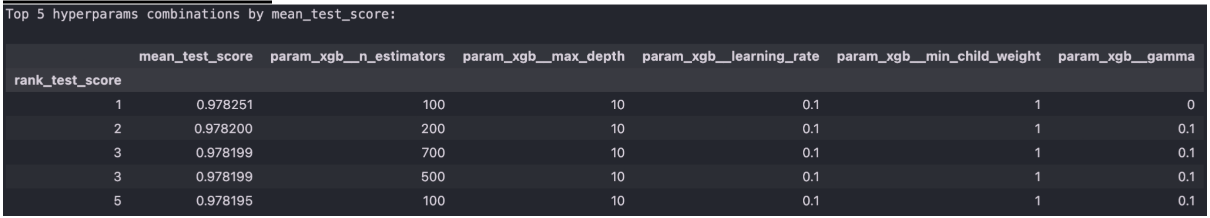
Top 5 hyperparams combinations by mean_test_score:

	mean_test_score	param_neural_network__hidden_layer_sizes	param_neural_network__activation	param_neural_network__alpha	param_neural_network__learning_rate
rank_test_score					
1	0.970641	(15, 15, 15)	tanh	0.05	constant
1	0.970641	(15, 15, 15)	tanh	0.05	adaptive
3	0.969788	(15, 15, 15)	tanh	0.01	adaptive
3	0.969788	(15, 15, 15)	tanh	0.01	constant
5	0.969459	(15, 15, 15)	tanh	0.1	adaptive

Best hyperparameters:
Activation = ‘tanh’, alpha = 0.05, hidden_layer_sizes = (15, 15, 15), learning_rate = ‘constant’

Best mean test score = 0.971

XGBoost Classifier



Top 5 hyperparams combinations by mean_test_score:

	mean_test_score	param_xgb__n_estimators	param_xgb__max_depth	param_xgb__learning_rate	param_xgb__min_child_weight	param_xgb__gamma
rank_test_score						
1	0.978251	100	10	0.1	1	0
2	0.978200	200	10	0.1	1	0.1
3	0.978199	700	10	0.1	1	0.1
3	0.978199	500	10	0.1	1	0.1
5	0.978195	100	10	0.1	1	0.1

Best hyperparameters:
N_estimators = 100, Max_depth = 10, Learning_rate = 0.1, Min_child_weight = 1, gamma = 0

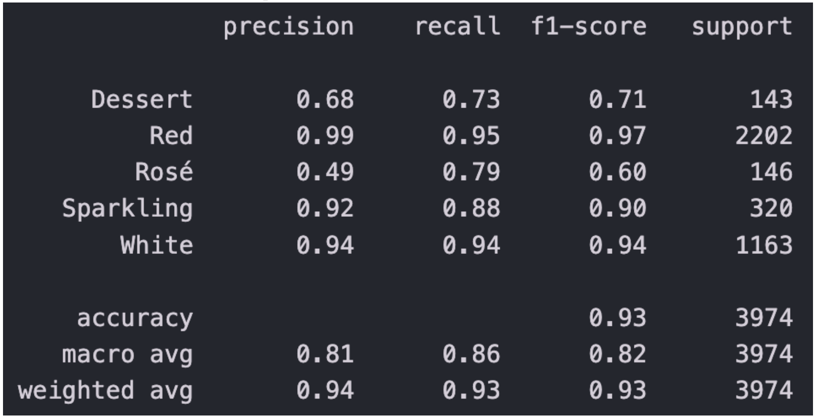
Best mean test score = 0.978

11.7 Appendix G

Model Performance

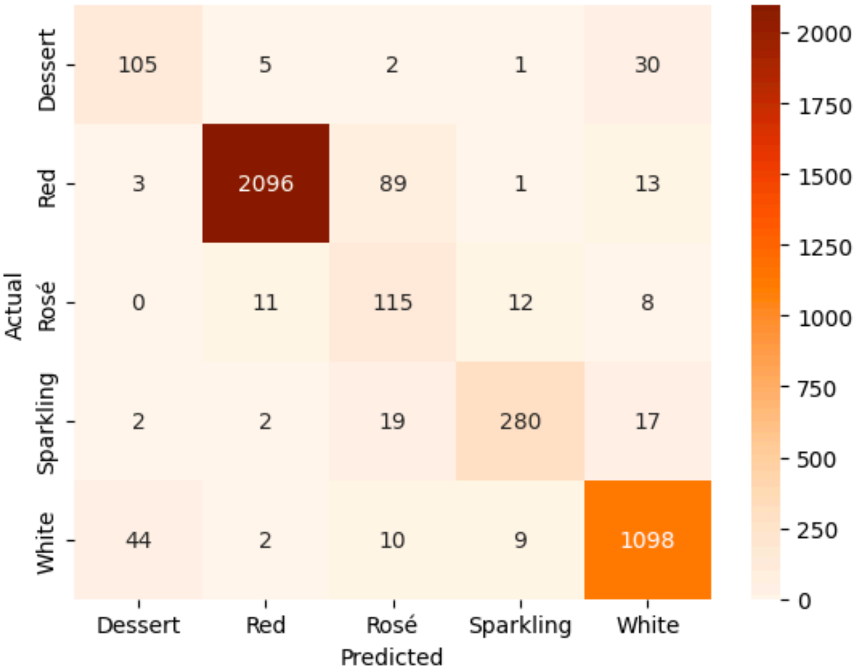
Random Forest Classifier

Classification Report



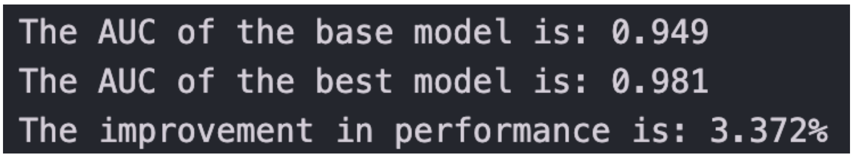
```
              precision    recall  f1-score   support

     Dessert       0.68      0.73      0.71       143
         Red       0.99      0.95      0.97      2202
        Rosé       0.49      0.79      0.60       146
   Sparkling       0.92      0.88      0.90       320
       White       0.94      0.94      0.94      1163

    accuracy                           0.93      3974
   macro avg       0.81      0.86      0.82      3974
weighted avg       0.94      0.93      0.93      3974
```

Confusion Matrix



Actual \ Predicted	Dessert	Red	Rosé	Sparkling	White
Dessert	105	5	2	1	30
Red	3	2096	89	1	13
Rosé	0	11	115	12	8
Sparkling	2	2	19	280	17
White	44	2	10	9	1098

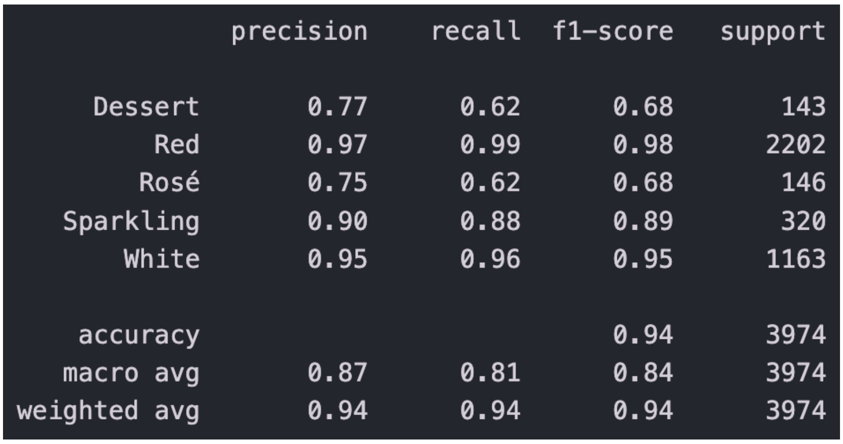
AUC before and after tuning



```
The AUC of the base model is: 0.949
The AUC of the best model is: 0.981
The improvement in performance is: 3.372%
```

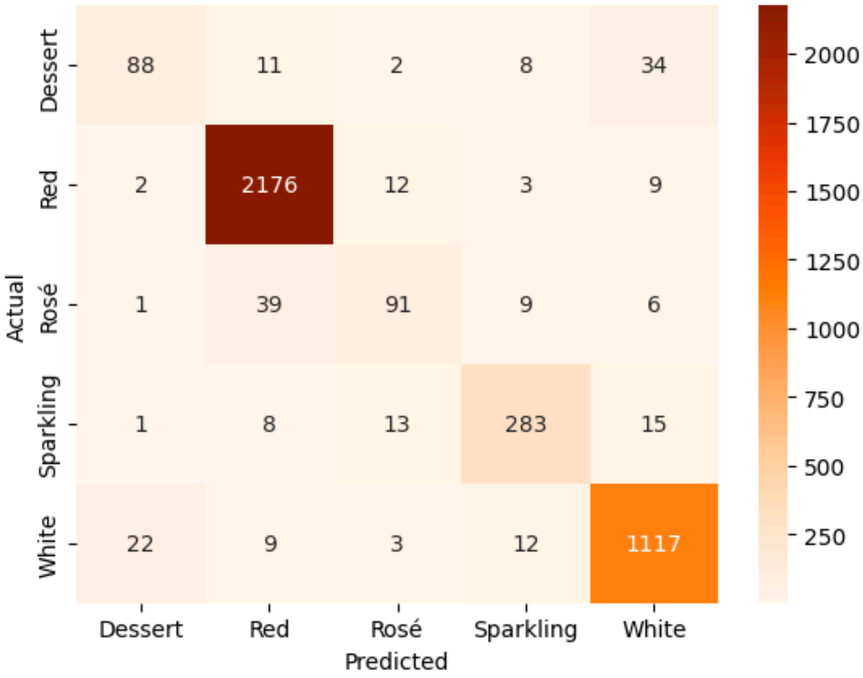
Multi-Layer Perceptron Classifier

Classification Report

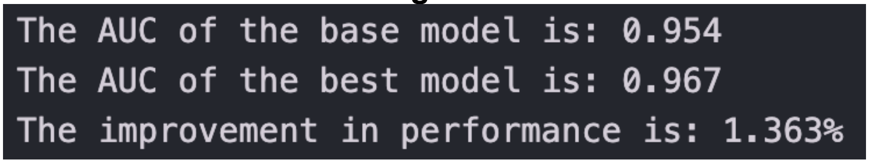


```
              precision    recall  f1-score   support

     Dessert       0.77      0.62      0.68       143
         Red       0.97      0.99      0.98      2202
        Rosé       0.75      0.62      0.68       146
   Sparkling       0.90      0.88      0.89       320
       White       0.95      0.96      0.95      1163

    accuracy                           0.94      3974
   macro avg       0.87      0.81      0.84      3974
weighted avg       0.94      0.94      0.94      3974
```

Confusion Matrix



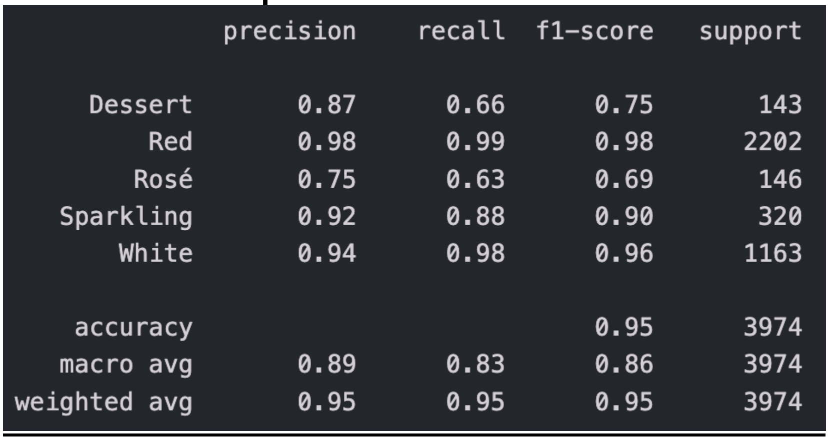
AUC before and after tuning



```
The AUC of the base model is: 0.954
The AUC of the best model is: 0.967
The improvement in performance is: 1.363%
```

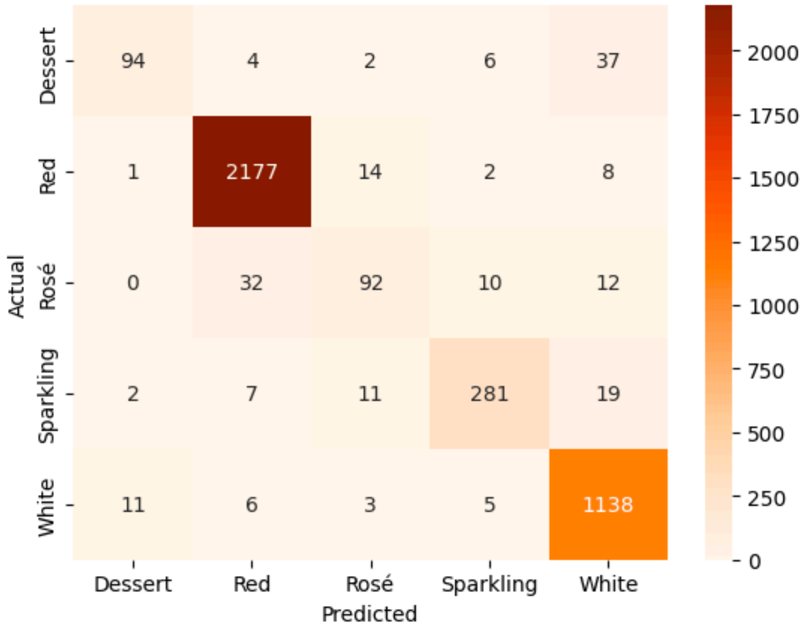
XGBoost Classifier

Classification Report



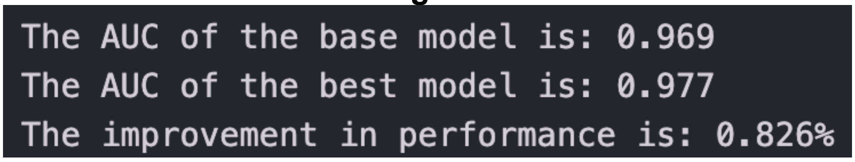
```
              precision    recall  f1-score   support

     Dessert       0.87      0.66      0.75       143
         Red       0.98      0.99      0.98      2202
        Rosé       0.75      0.63      0.69       146
   Sparkling       0.92      0.88      0.90       320
       White       0.94      0.98      0.96      1163

    accuracy                           0.95      3974
   macro avg       0.89      0.83      0.86      3974
weighted avg       0.95      0.95      0.95      3974
```

Confusion Matrix



Actual \ Predicted	Dessert	Red	Rosé	Sparkling	White
Dessert	94	4	2	6	37
Red	1	2177	14	2	8
Rosé	0	32	92	10	12
Sparkling	2	7	11	281	19
White	11	6	3	5	1138

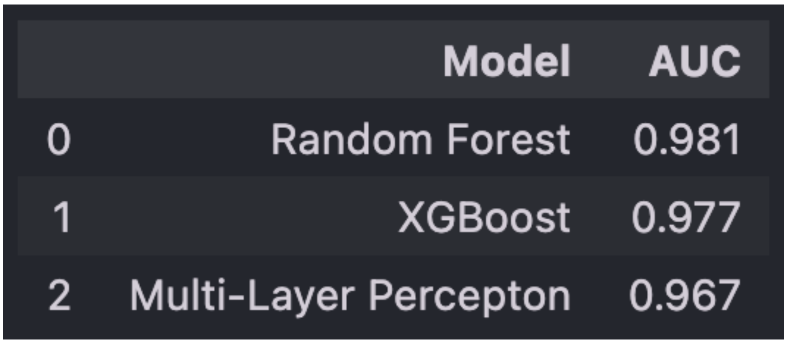
AUC before and after tuning



```
The AUC of the base model is: 0.969
The AUC of the best model is: 0.977
The improvement in performance is: 0.826%
```

11.8 Appendix H

Model Comparison



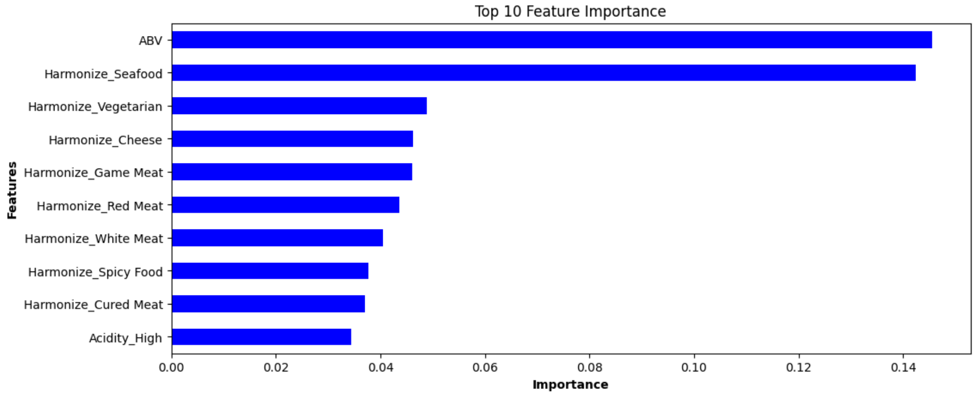
	Model	AUC
0	Random Forest	0.981
1	XGBoost	0.977
2	Multi-Layer Percepton	0.967

Random Forest is the best model out of the 3.

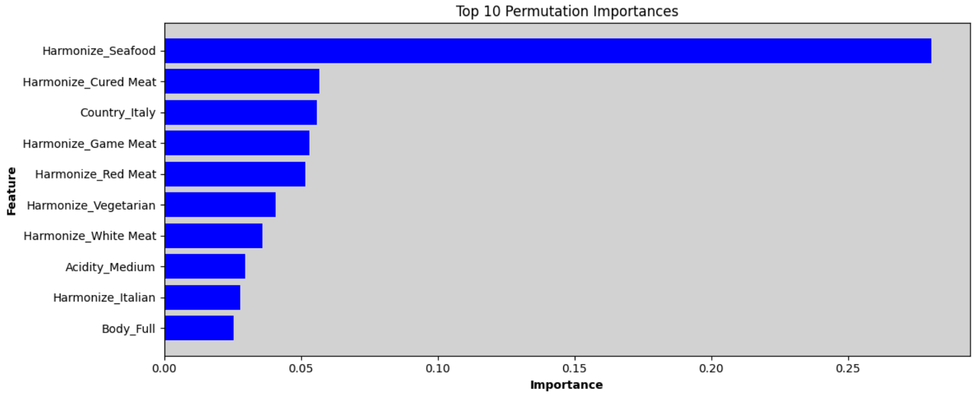
11.9 Appendix I

Model Interpretation

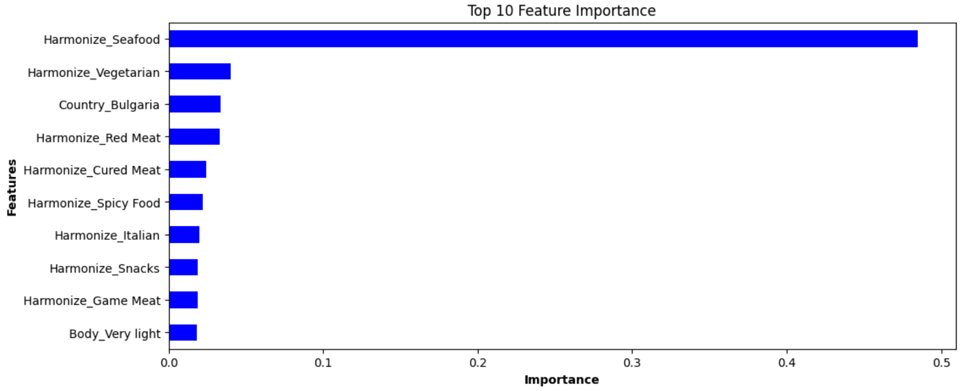
Random Forest



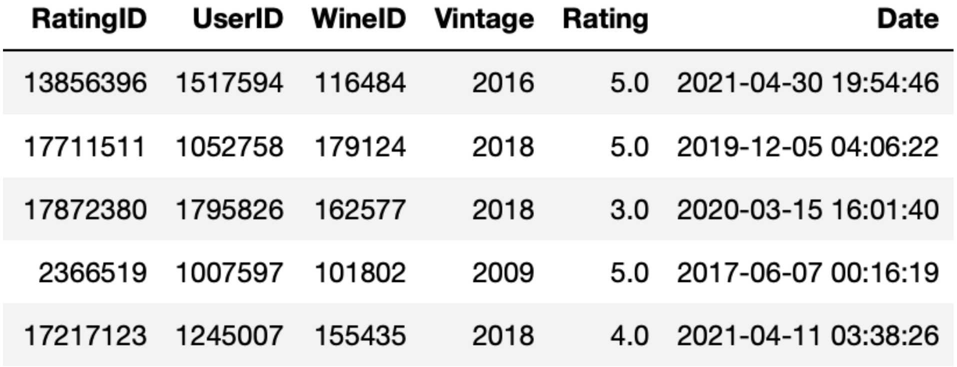
Multi-Layer Perceptron Classifier



XGBoost Classifier



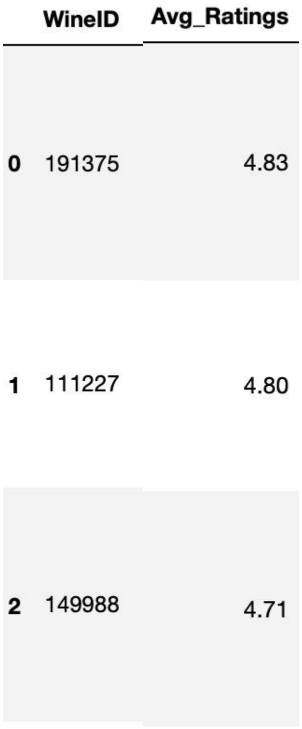
11.10 Appendix J



RatingID	UserID	WineID	Vintage	Rating	Date
13856396	1517594	116484	2016	5.0	2021-04-30 19:54:46
17711511	1052758	179124	2018	5.0	2019-12-05 04:06:22
17872380	1795826	162577	2018	3.0	2020-03-15 16:01:40
2366519	1007597	101802	2009	5.0	2017-06-07 00:16:19
17217123	1245007	155435	2018	4.0	2021-04-11 03:38:26

Table 4. Top 5 rows of ratings dataset

11.10 Appendix K



	WineID	Avg_Ratings
0	191375	4.83
1	111227	4.80
2	149988	4.71