AttrIBUtes Matter:
Classifying Homebrewed Beers

ISYE-7406 – Data Mining & Statistical Learning

4/17/2022

Mihaiela Vieru – x340, mvieru3@gatech.edu
Samantha Virgil – x249, svirgil3@gatech.edu
Lauren Wilson – x012, lwilson85@gatech.edu
Joel Zapata – x169, jzapata9@gatech.edu
Patrick Zdunek – x426, zdunek3@gatech.edu

Project Group 33

AttrIBUtes Matter:
Classifying Homebrewed Beers

Group 33 - Mihaiela Vieru, Samantha Virgil, Lauren Wilson, Joel Zapata, and Patrick Zdunek 4/17/22

Abstract

This analysis focuses on the use of seven different types of classification models to accurately predict the style of a homemade beer based on its recipe. The methods used are Decision Tree, Random Forest, K-Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Naïve Bayes, and Support Vector Machine. Each model's parameters were tuned (when applicable) using the training dataset and evaluated for model performance using the testing dataset. Overall, all the models were within a testing accuracy range of 10% of each other, except for the Naïve Bayes model, which far outperformed the other models. After investigating the models and dataset, this seems to be due to the nature of how the Naïve Bayes model assigns conditional probabilities to the observations differently than the other models explored. Suggestions for future improvement include potential data transformation to improve the representation of relationships between variables, utilizing more complex feature selection methods, and possibly increasing the number of response groups to reduce overlap in features of similar but different beer families.

Introduction

Beer is one of the oldest and most widely consumed alcoholic drinks in the world, and the third most consumed drink overall after water and tea [1]. It is produced by brewing and fermentation of starches, such as barley, wheat, maize, rice, and oats. Modern recipes also include hops and other flavoring agents (herbs or fruits). Over the last four decades the number of breweries has exploded, as has the population of people making their own brews. There are many elements that go into brewing, from the ingredients to the order of operations to the temperatures and methods used, all of which impact the taste and category of the resulting beer. Based on a collection of thousands of these unique recipes, there is an opportunity to look for patterns amongst these recipes to see if any combination of factors can be used to predict the type of beer that will be produced from the next new recipe.

The purpose of this analysis is to test various multi-class models using a dataset of homemade beer recipes. We will explore key components of these recipes and build a variety of models to see which, if any, can be used to accurately predict the style classification of the beer recipe. The methods used for comparison are Decision Tree, Random Forest, K-Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Naïve Bayes, and Support Vector Machine. The styles of beer will be categorized into twelve groups to increase the sample sizes of each response classification, and the model performance will be measured by comparing the testing accuracy of each classification model. This report will explore the dataset and any unique aspects of it and discover any usable patterns for prediction using the methods listed above. Finally, the suspected causes of the quantitative results, lessons learned, and potential future improvements will be discussed in the conclusion.

Data Source

The data was sourced from *Kaggle*, a web-based hub for publicly shared datasets [2]. This dataset has nearly 74,000 observations, each representing a different homemade beer recipe with 21 attributes and 2 columns representing the beer style (*Style* and *StyleID*) as the response. After examining the dataset, some columns were determined not to be useful in the prediction of the style of beer based on features of the recipe, such as the *BeerID*, *URL*, *UserID*, etc. Thus, to simplify the dataset for the purpose of this analysis, the attributes will be reduced only to the potentially indicative variables with consistent values, as listed in Table 1 below.

Attribute	Description	Data Type
ABV	Alcohol by volume	Decimal
BoilSize	Amount of fluid in liters at beginning of boil	Decimal
BoilTime	Time wort is boiled (in minutes)	Integer
BrewMethod	Various techniques for brewing	Categorical
Color	Standard Reference Method - light to dark ex. 40 = black	Decimal
FG	Specific gravity of wort after fermentation	Decimal
IBU	International Bittering Units	Decimal
OG	Specific gravity of wort before fermentation	Decimal
Size.L	Amount in liters brewed for recipe	Decimal
Style	Type of beer (Ex. American IPA, Cream Ale, etc.)	Categorical
Efficiency	Beer mash extraction efficiency - extracting sugars from the grain during mash	Integer

Table 1: Dataset Attributes, Descriptions, and Data Types of *beerRecipe* dataset [2]

With the dataset as is, there are over 176 different classifications of *Style*. To simplify the models and potentially increase their accuracy, these styles are categorized into twelve larger families of beers commonly known in the industry based on their characteristics [3]. Any recipe that did not fit into one of the twelve groups was put into "Other" and removed from this analysis. The twelve groups, in a derived column called *General.Group*, will be used as the response variable in the analysis in place of *Style* and the values are listed below in Table 2. The full breakdown of the styles of beer included in each group can be found in Appendix A.

Group Name
Belgian-Style Ale
Brown Ale
Cider
Dark Lager
German Bock
IPA
Pale Ale
Pilsner
Porter
Stout
Wheat
Wild & Sour Ale

Table 2: Groups used to categorize styles and used as response in analysis

Exploratory Data Analysis

Variable Correlation

To compare the relationships between the predicting variables, a correlation matrix will be used, as shown in Figure 1. In the correlation matrix, the large, dark blue dots with values close to 1 represent a high positive correlation, while the dark yellow dots with values less than 0 represent a negative correlation.

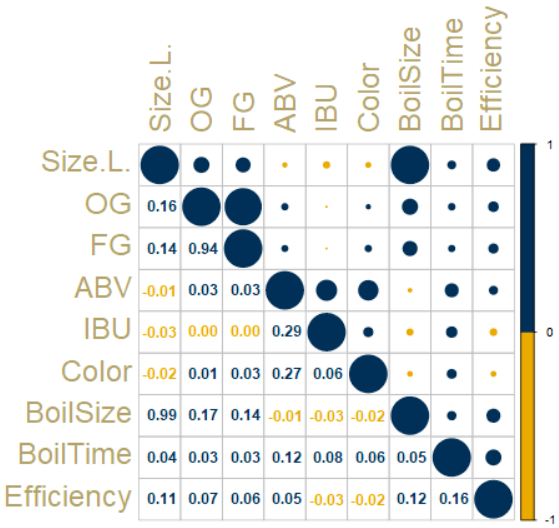

Figure 1: Correlation matrix of quantitative variables in *beerRecipe* dataset

For this dataset, the most notable positive correlations are between *BoilSize* and *Size.L*, and *OG* and *FG*. It is logical that *BoilSize* and *Size.L* are correlated because they both relate to the amount of fluid being produced by the recipe. It also makes sense that *OG* and *FG* are correlated since they represent the "before and after" of the specific gravity of the wort [4] in relation to fermentation. To combat the effects of these correlations, two new variables are created to capture the difference without being redundant. The formulas for the new variables are shown in Equation 1 and 2 below.

$$Size.Change = BoilSize - Size.L$$

Equation 1: Calculation of *Size.Change* variable

$$G.Change = OG - FG$$

Equation 2: Calculation of *G.Change* variable

The new correlation matrix with the two new variables, *Size.Change* and *G.Change*, added and the redundant variables, *Size.L*, *OG*, and *FG*, removed are shown in Figure 2 below.

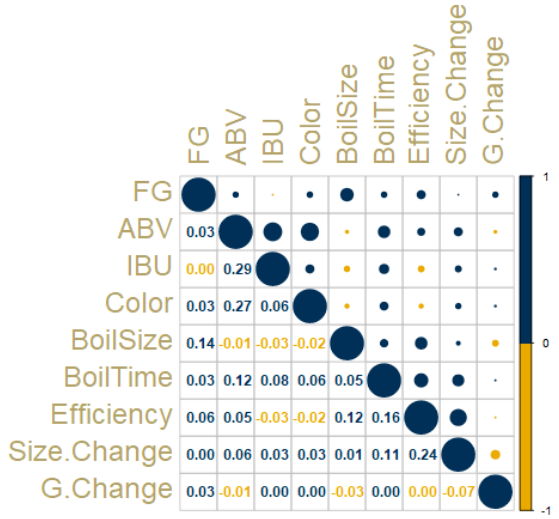

Figure 2: Correlation matrix of updated quantitative variables in *beerRecipe* dataset

The newly calculated variables have removed the strong correlations from the dataset. Based on Figure 2, there are no other outstanding positive correlations, and no strong negative correlations between the continuous variables.

Variable Distributions

The boxplots shown in Figure 3 below visually represent the distributions of the variables *Gravity Change*, *ABV*, *Color*, *IBU*, *Size Change*, and *Boil Time*. The inner quartile range is marked by the blue bar on each boxplot with the independent blue points showing the observations that fell out of the range of most of the values for that variable. All the variables have long tails of points, indicating there are a lot of points that fall far from the mean for that variable that will require further exploration to see if any are in fact outliers.

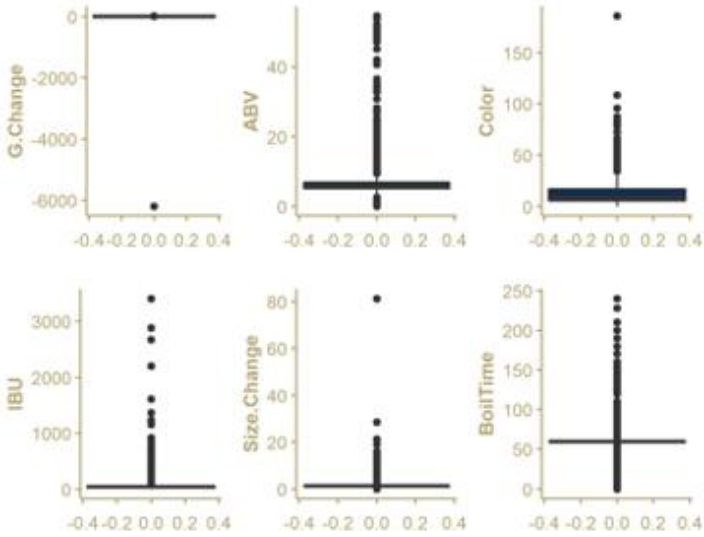

Figure 3: Initial distribution of predicting variables

After further investigation, there turned out to be many outliers in the dataset. By removing a few of the most egregious outliers, such as the large negative value in *G.Change*, and the large value in *Size.Change*, the distribution becomes clearer. More outliers could have been removed but were kept because they seemed to be legitimate and contributing to the full picture of the data.

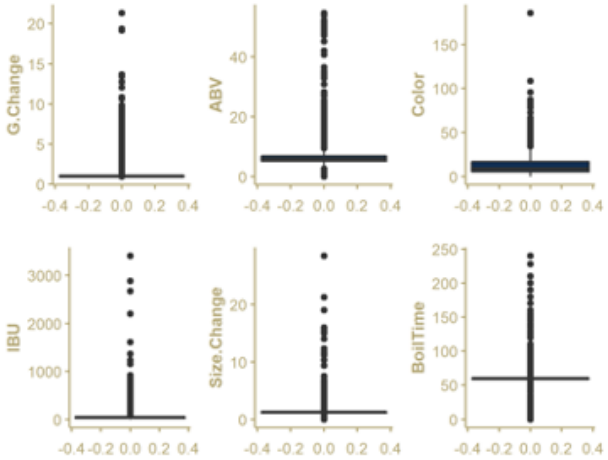

Figure 4: Improved distribution of predicting variables

To better understand the data, empirical analysis of the variables should be conducted. The summary statistics of each of the variables in the dataset are shown in Figure 5 below.

```
    Size.L.              OG                FG                ABV               IBU              Color            BoilSize          BoilTime
 Min.   :   1.00   Min.   : 1.000   Min.   : 0.4254   Min.   : 0.000   Min.   :   0.00   Min.   :  0.00   Min.   :   1.00   Min.   :  0.00
 1st Qu.:  18.93   1st Qu.: 1.051   1st Qu.: 1.0110   1st Qu.: 5.080   1st Qu.:  23.91   1st Qu.:  5.15   1st Qu.:  21.00   1st Qu.: 60.00
 Median :  20.82   Median : 1.058   Median : 1.0130   Median : 5.780   Median :  36.36   Median :  8.29   Median :  27.63   Median : 60.00
 Mean   :  44.48   Mean   : 1.407   Mean   : 1.0756   Mean   : 6.125   Mean   :  45.01   Mean   : 13.31   Mean   :  50.31   Mean   : 65.19
 3rd Qu.:  24.00   3rd Qu.: 1.068   3rd Qu.: 1.0170   3rd Qu.: 6.820   3rd Qu.:  57.29   3rd Qu.: 16.46   3rd Qu.:  30.00   3rd Qu.: 60.00
 Max.   :9200.00   Max.   :34.035   Max.   :10.3414   Max.   :54.720   Max.   :3409.30   Max.   :186.00   Max.   :9700.00   Max.   :240.00
   Efficiency          BrewMethod     Size.Change          G.Change
 Min.   :  0.0   All Grain   :46825   Min.   : 0.002857   Min.   : 1.000
 1st Qu.: 65.0   BIAB        :11381   1st Qu.: 1.125165   1st Qu.: 1.038
 Median : 70.0   extract     : 8007   Median : 1.241744   Median : 1.044
 Mean   : 66.4   Partial Mash: 3197   Mean   : 1.206773   Mean   : 1.136
 3rd Qu.: 75.0                        3rd Qu.: 1.363593   3rd Qu.: 1.052
 Max.   :100.0                        Max.   :28.500000   Max.   :21.358
```

Figure 5: Summary statistics of the *beerRecipe* dataset

ABV, *IBU*, *Color*, and *Efficiency* all have minimum values of zero, indicating they are likely missing data for one of more observations. Additionally, most of the maximum values of the variables seem to be significantly larger than the respective mean and third quartile values, again providing evidence of outliers that will need to be monitored during the analysis.

Figure 6 below shows the distribution of the response variable *General.Group*. The plot indicates that the data is skewed, which may be another factor that could affect the performance of the models.

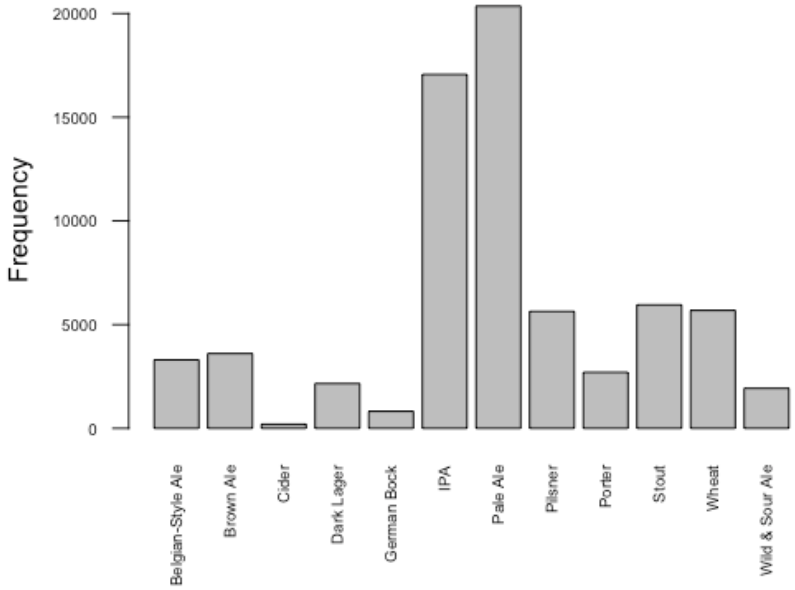

Figure 6: Distribution of *General.Group*

Table 3 below summarizes the changes made to the dataset through the exploratory data analysis process. This will be the final set of attributes used in the classification models in the next section. For further investigation into the dataset, see Appendix B.

Attribute	Description	Data Type
ABV	Alcohol by volume	Decimal
BoilSize	Amount of fluid in liters at beginning of boil	Decimal
BoilTime	Time wort is boiled (in minutes)	Integer
BrewMethod	Various techniques for brewing	Categorical
Color	Standard Reference Method - light to dark ex. 40 = black	Decimal
G.Change	Difference in specific gravity of wort before and after fermentation	Decimal
IBU	International Bittering Units	Decimal
Size.Change	Amount of fluid at beginning of boil vs. final amount produced	Decimal
General.Group	Response - type of beer (12 families of beer styles)	Categorical

Table 3: Final Dataset Attributes, Descriptions, and Data Types of *beerRecipe* dataset [2]

Proposed Methodology

Splitting the Data

To build the classification models, the data must first be divided into training and test sets. This was done by randomly sampling 75% of the dataset to be in the training set and the remaining 25% to be the test set. This distribution is common as it is beneficial to have a larger training set than test set to reduce overfitting.

Classification Models

Overall, this scenario is a multi-class classification task, as there are more than two types of responses. Luckily, many algorithms that were designed for binary classification can also be used for multi-class situations as well. The models selected for this analysis include Decision Tree, Random Forest, K-Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Naïve Bayes, and Support Vector Machine.

a. *Decision Tree* – a class discriminator that recursively partitions the training dataset until each partition consists entirely or dominantly of instances of one class. Each non-leaf node of the tree contains a split point that is a test on one or more attributes and determines how the data is partitioned. Decisions work best when the response variable is categorical, which is true for the *beerRecipe* dataset.

b. *Random Forest* – a method built upon decision trees; random forest models are a collection of decision trees that are combined to classify responses. Each individual tree's classification of the observations is considered for the final classification, and the most voted category is selected. Random Forests are known to be more accurate than decision trees due to the amount of variability represented in the model. However, unlike decision trees, they are difficult to explain as they do not produce a single visible output.

c. *K-Nearest Neighbors* – an algorithm that classifies unknown data points by finding the most common class among the *k* closest data points; The most common category among the points then becomes the classifying category of the specified point. To find the most effective value of *k*, several values between 1 and 50 will be used and evaluated using the resulting training accuracies.

d. *Linear Discriminant Analysis (LDA)* – LDA is a classification method used to find linear combinations of features that characterize the response classes. The resulting linear combination can then be used as a linear classifier.

e. *Logistic Regression* – multinomial logistic regression is a generalized form of logistic regression allowing for more than two responses, as necessary in this scenario. In both binary and multinomial forms, logistic regression models predict the probabilities of the responses based on the predicting variables. However, instead of outputting a single probability value as in the binary form, the model will output one probability for each class type and will assign a particular observation to the class with the

highest probability. This method assumes the multicollinearity to be low, as was confirmed during the exploration and preparation of this dataset.

f. *Naïve Bayes* – a nonlinear classification algorithm based on Bayes Theorem. This method assumes that all the predictors are independent (earning the prefix "Naïve") and assigns data to the classes by finding the highest posterior probabilities for each class using the Naïve Bayes equation [5]. In terms of multi-class classification, this is one of the most ideal methods to use due to the process in which the model assigns conditional probabilities to the observations.

g. *Support Vector Machine (SVM)* - SVM performs classification tasks by constructing hyperplanes in a multidimensional space that separates different class labels. The objective is to find a hyperplane that maximizes the separation of the data points. The data points with the minimum distance to the hyperplane are called Support Vectors. The separation can be computed using different kernels (I.e., Linear, Polynomial, Gaussian, Radial and Sigmoid).

Analysis and Results

Table 4 below shows the training and testing accuracies for each of the models. Though all the models were run with various combinations of parameters to maximize the accuracy, the Random Forest, KNN, and LDA models each had a few models that performed similarly which is why they are listed in the table more than once. Overall, all the testing accuracies except one were within 8.9% percent of each other, which are relatively close. The model with the highest testing accuracy by far was the Naïve Bayes model, with a testing accuracy of 94.80%. This is very high, especially in comparison to the next best model, the Random Forest model with 7 predictors, which had a testing accuracy of 61.02%. The Random Forest models were not able to get meaningful training accuracy measurements due to the model implementation and code. An interesting observation between comparing the training and testing accuracies is that the testing accuracy was greater than the training accuracy for a few of the models, which is unusual and could be due to the particular split used between the training and test datasets. While the LDA models were able to run to completion, the results were found to be unusable due to the conditions required for a valid LDA model. More details of this decision can be found in Appendix C.

Model	Training Accuracy	Testing Accuracy
Decision Tree	54.31%	55.20%
Random Forest with 7 predictors	99.99% *	61.02%
Random Forest with 3 predictors	99.99%*	59.23%
K-Nearest Neighbors (k=5)	NA	53.00%
K-Nearest Neighbors (k=10)	NA	55.35%
K-Nearest Neighbors (k=15)	NA	55.95%
LDA with 7 predictors	51.32%	52.17%
LDA with 3 predictors	51.11%	52.12%
Logistic Regression	52.36%	53.46%
Naïve Bayes	94.40%	94.80%
SVM	56.88%	58.49%

Table 4: Results of Classification Models with Testing Accuracy

Decision Tree

The Decision Tree model was trained using all seven predictors, and the resulting tree is shown in Figure 7 below. Since this model has a multi-class response, each node shows the predicted class (Pale Ale, IPA, etc.), the predicted probability of each class, and the percentage of observations in the node. Based on the final leaves of the tree (the bottom row), the Pale Ale and IPA nodes ended with the largest percentage of

the observations, with 41% and 26% respectively. This aligns with the expected results based on the class distributions shown in Figure 6. Based on the splits, the *IBU* and *Color* predictors proved to be the most divisive in separating the classes, which is likely due to the greedy approach of this algorithm. One major drawback of this model is that it only separated 6 of the 12 response classes, which is likely what led to the testing accuracy of only 55%.

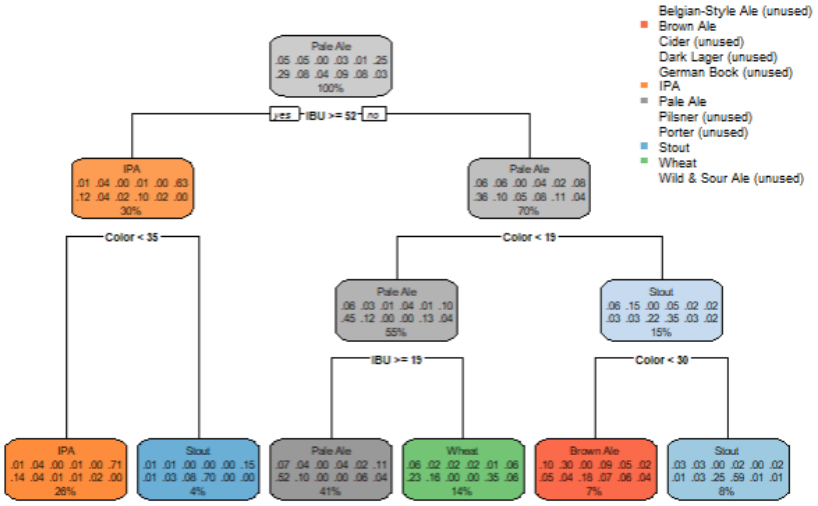

Figure 7: Decision Tree

Random Forest

Building on the process of the decision tree, the major advantage of the Random Forest method is that as an ensemble method, it considers the results of several variations of decision trees, lowering the risk of overfitting and increasing the accuracy of the classification. It is also robust to outliers and runs efficiently on large datasets, which are both important features for this dataset. These advantages are reflected in the testing accuracy of the model, which was found to be 61.02%, the second highest of all the models run. One drawback of Random Forests is that they are hard to interpret as they do not produce one single output model. However, the model does give a glimpse into the importance of the variables in determining the response class, as shown in Figure 8 below. The plot shows that *Color* and *IBU* again proved to be the most crucial deciding factors, which is consistent with the results from the Decision Tree model.

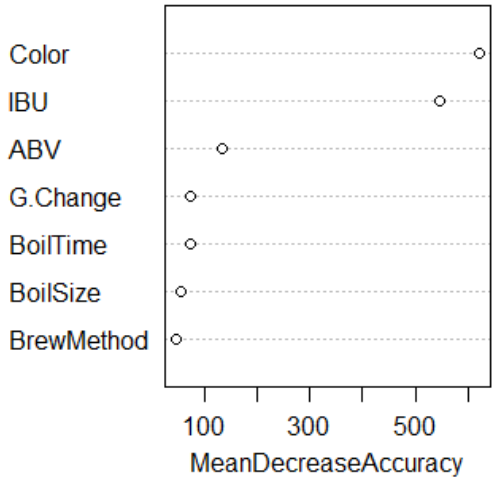

Figure 8: Predicting Variable Importance Based on Random Forest Model

K-Nearest Neighbors (KNN)

There were three KNN models that performed very similarly and had made up the middle of performance range compared to the other models. The accuracy of the model began to plateau around *k*=13 neighboring points and remained in the 55% to 57% range. A plot of the accuracy vs. change in the value of *k* is shown in Figure 9 below:

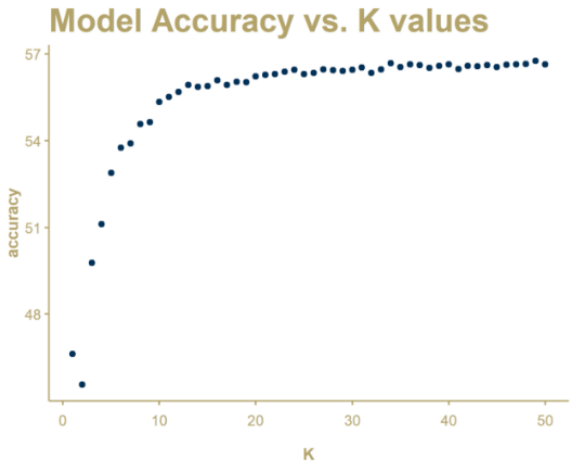

Figure 9: *k* vs. Accuracy

The K-Nearest Neighbors algorithm becomes computationally more complex in high-dimensional space as it is more difficult to calculate the distances between points, which is the assumed cause for the 55%-57% accuracy range.

Linear Discriminant Analysis (LDA)

The LDA model performed slightly below average at about 52% accuracy on the test data. The best models found were those using 3 predictors and 7 predictors to classify the model's response; however, the additional 4 predictors only increased the accuracy by about 0.05%. The left plot in Figure 10 below displays the true classification of the test data and the figure on the right displays the predicted values from the LDA model on the test data. Despite reaching termination, the LDA model was determined to be illogical to use for this scenario and removed from the final comparison. The criteria of this decision are discussed more in Appendix C.

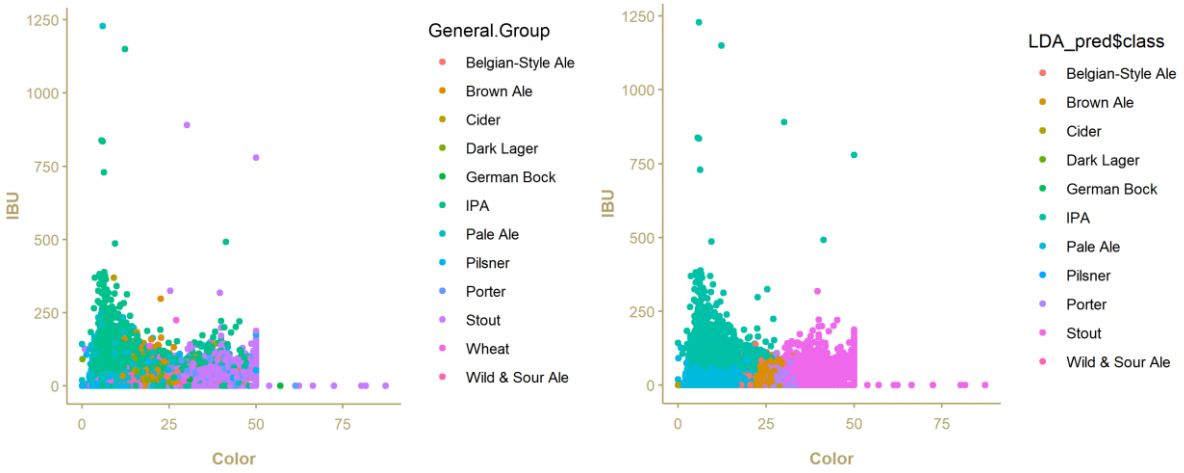

Figure 10: True Responses (left) vs. Predicted Responses (right) from LDA Model

Logistic Regression

Of the models used, the logistic regression model had one of the lowest test accuracies at 53.46%. After further investigation of the model, it seemed that this result was likely due to the major drawbacks of logistic regression – the assumption of linearity between the independent and dependent variables and the linear decision surface of the model. Up to this point, the models that have been performing the best have been those that can handle non-linearity in some capacity, namely the Random Forest and KNN models. However, like LDA, logistic regression is not meant for these scenarios, which may describe why these two models have the lowest accuracies.

Naïve Bayes

The Naïve Bayes model's testing accuracy of 94.80% was 33.78% better than the next highest performer. This substantial difference could be due to the model's capability to handle multinomial classification without being sensitive to irrelevant attributes. This model can classify data by having a range of acceptable values per predicting variable for each response category. To better visualize how the model is making the class decisions, probability density plots such as Figure 11 below can be used. Figure 11 shows the accepted probability density for each type of beer for the *Color* variable. These probability densities are used in the overall model to determine the probability of an observation being a particular class based on just that variable. The culmination of these probabilities from all the predictors is what ultimate influences the model to pick the class for the observation.

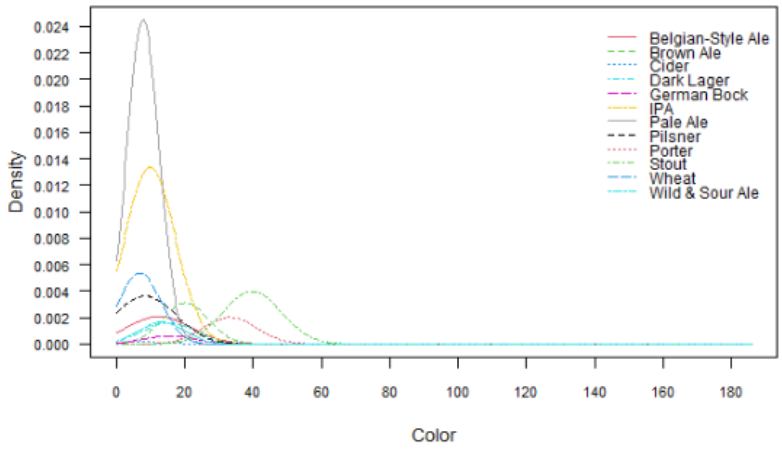

Figure 11: Probability Density Plot for each Family of Beer in Relation to *Color*

SVM

The SVM's model accuracy was 58.49%, which is topped by Naïve Bayes and Random Forest. Since there are twelve classes, the multi-class SVM tries to separate the problem into multiple binary problems: a class versus all other classes. This means the separation considers all data points and it will try to maximize the hyperplane between one class and the rest and repeat this for each class. This can of course lead to inferior performance, especially for this data set with overlapping class attributes. While SVMs can work well on high-dimensional data and are suitable for both linear and non-linear separable data points, for the *beerRecipes* dataset and with the current classes in place, it does not seem to be a great model to use. Figure 12 shows sample graphs of SVM classification boundaries between two attributes, where "x" represents a support vector and "o" represents a data point. As seen from the graphs a lot of the data points are also support vectors, which also suggests that it's hard to linearly separate this dataset due to the similarity of all these classes.

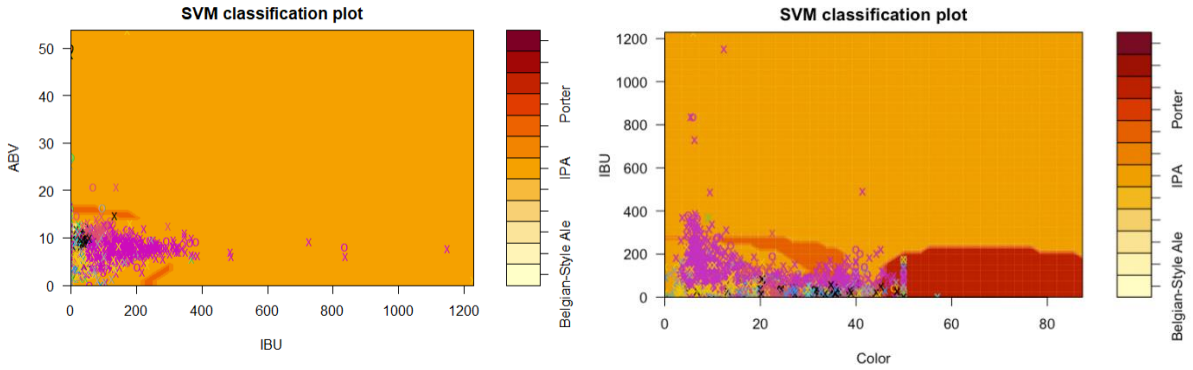

Figure 12: SVM Classification Plots: *IBU* vs. *ABV* (left); *Color* vs. *IBU* (right)

Conclusion

The Naïve Bayes model performed the best with an accuracy of 94.80%, followed by the Random Forest model at 61.02%. The results of the Random Forest model were expected based on the robustness of the model and ability for highly correlated variables to play nearly equivalent roles in the classification. However, the success of the Naïve Bayes model was a little more unexpected due to the fundamental assumption that all the predictors are independent, which is unlikely in this scenario. Thus, a suggestion for future improvement would be to explore the details of this model further to ensure the results and model assumptions agree. Another suggestion would be to explore data transformation and feature engineering methods to potentially improve the representation of the relationships between the variables.

The similar accuracy across such different models also suggests that the models may not be at fault for the low performance rate but rather the data itself. The styles of beer were originally categorized into the 12 families in hopes of improving model accuracy. However, because there are many recipes that combine characteristics of multiple families, there is likely to be a lot of overlap of similar predicting variable values that lead to different family definitions. For example, while Stouts tend to be easier to classify due to their very dark brown color, other families such as IPAs, Pilsners, and Lagers share a lot of similarities in color and other factors. Thus, future analysis could involve separating the families into smaller groups of styles, which could provide more useful and accurate models. This could also help reduce the imbalance in response data, which is likely to impact the results of most classification models. Other future improvements to consider that would address this issue are oversampling, undersampling, or weighting.

Lessons Learned

- Tuning a model's parameters is very important to finding the best model as the default settings of a model may, and often are not optimal.
- Cross validation is essential to make sure your model perform well in multiple circumstances.
- The importance of choosing appropriate models to test depending on the problem type and data available.
- Sometimes you can be surprised by a model's performance, which in turn leads to new questions/investigation (what is cause for this unexpected result? Do we need to transform data? What other methodologies might work better?)
- Model accuracy is often not telling the whole story of the model; thus, it is important to understand the assumptions and limitations of a model to ensure its validity in the situation.

References

[1] Wikipedia contributors. (2022, April 3). Beer. Wikipedia. https://en.wikipedia.org/wiki/Beer
[2] Brewer's Friend Beer Recipes. (2018, April 12). Kaggle. https://www.kaggle.com/datasets/jtrofe/beer-recipes
[3] W. (2021b, September 23). Different Types & Styles of Beer: The Ultimate Guide. WebstaurantStore. https://www.webstaurantstore.com/article/27/different-types-of-beers.html
[4] C. (2021a, April 23). Beer Glossary. CraftBeer.Com. https://www.craftbeer.com/beer/beer-glossary#:%7E:text=Wort%20The%20bittersweet%20sugar%20solution,which%20becomes%20beer%20through%20fermentation
[5] naiveBayes function - RDocumentation. (n.d.). RDocumentation. https://www.rdocumentation.org/packages/e1071/versions/1.7-9/topics/naiveBayes

Appendix

Appendix A: Breakdown of Styles within each Beer Group

Belgian-Style Ale	Brown Ale	Cider
• Belgian Dark Strong Ale • Belgian Dubbel • Belgian Golden Strong Ale • Belgian Pale Ale • Belgian Specialty Ale • Belgian Tripel • Fruit Lambic • Gueuze • Lambic • Straight (Unblended) Lambic	• Altbier • American Barleywine • American Brown Ale • British Brown Ale • Dusseldorf Altbier • English Barleywine • Flanders Brown Ale/Oud Bruin • Kentucky Common • London Brown Ale • Mild • North German Altbier • Northern English Brown • Oud Bruin • Southern English Brown • Strong Scotch Ale • Wee Heavy	• Apple Wine • Common Cider • English Cider • French Cider • Fruit Cider • New England Cider • Other Specialty Cider or Perry • Traditional Perry
Dark Lager	**German Bock**	**IPA**
• Czech Amber Lager • Czech Dark Lager • Dark American Lager • Dark Mild • International Amber Lager • International Dark Lager • Kellerbier: Amber Kellerbier • Märzen • Oktoberfest/Märzen • Old Ale • Rauchbier • Schwarzbier	• Doppelbock • Dunkles Bock • Eisbock • Helles Bock • Maibock/Helles Bock • Traditional Bock • Weizenbock	• American IPA • Double IPA • English IPA • Imperial IPA • Specialty IPA: Belgian IPA • Specialty IPA: Black IPA • Specialty IPA: Brown IPA • Specialty IPA: Red IPA • Specialty IPA: Rye IPA • Specialty IPA: White IPA

• Vienna Lager		
Pale Ale	**Pilsner**	**Porter**
• American Amber Ale • American Pale Ale • Australian Sparkling Ale • Belgian Blond Ale • Best Bitter • Bière de Garde • Blonde Ale • British Golden Ale • British Strong Ale • Cream Ale • Czech Pale Lager • Czech Premium Pale Lager • Dortmunder Export • Extra Special/Strong Bitter (ESB) • International Pale Lager • Kölsch • Kellerbier: Pale Kellerbier • Ordinary Bitter • Pre-Prohibition Lager • Saison • Scottish Light • Scottish Light 60/- • Special/Best/Premium Bitter • Standard/Ordinary Bitter • Strong Bitter • Trappist Single	• American Lager • American Light Lager • Bohemian Pilsener • California Common • California Common Beer • Classic American Pilsner • Classic Rauchbier • Festbier • German Helles Exportbier • German Leichtbier • German Pils • German Pilsner (Pils) • Light American Lager • Munich Helles • Premium American Lager • Standard American Lager	• American Porter • Baltic Porter • Brown Porter • English Porter • Pre-Prohibition Porter • Robust Porter
Stout	**Wheat**	**Wild & Sour Ale**
• American Stout • Dry Stout • Foreign Extra Stout • Imperial Stout • Irish Extra Stout • Irish Stout • Oatmeal Stout • Russian Imperial Stout • Sweet Stout • Tropical Stout	• American Strong Ale • American Wheat Beer • American Wheat or Rye Beer • Berliner Weisse • Dunkelweizen • Dunkles Weissbier • Gose • Munich Dunkel • Piwo Grodziskie • Weissbier • Weizen/Weissbier • Wheatwine • Witbier	• Braggot • Brett Beer • Cyser (Apple Melomel) • Dry Mead • Flanders Red Ale • Irish Red Ale • Lichtenhainer • Metheglin • Mixed-Fermentation Sour Beer • Open Category Mead • Pyment (Grape Melomel) • Semi-Sweet Mead • Sweet Mead • Wild Specialty Beer
Other		
• Alternative Grain Beer • Alternative Sugar Beer	• Mixed-Style Beer • N/A	• Scottish Heavy • Scottish Heavy 70/-

• Autumn Seasonal Beer • Classic Style Smoked Beer • Clone Beer • Experimental Beer • Fruit and Spice Beer • Fruit Beer • Holiday/Winter Special Spiced Beer	• Other Fruit Melomel • Other Smoked Beer • Roggenbier • Roggenbier (German Rye Beer) • Sahti • Scottish Export • Scottish Export 80/-	• Specialty Beer • Specialty Fruit Beer • Specialty Smoked Beer • Specialty Wood-Aged Beer • Spice Herb or Vegetable Beer • Winter Seasonal Beer • Wood-Aged Beer

Appendix B – Further Dataset Exploration

K-Means Clustering – Unsupervised Learning

For the sake of exploration, a K-means clustering model was constructed, which is a form of unsupervised learning that partitions the data into clusters based on similarities between the observations. By using K-means clustering with the variables *FG*, *ABV*, *IBU*, *Color*, *BoilTime*, *Efficiency*, *Size.Change*, and *G.Change*, two clusters were able to be identified. The first two principal components explain 22% and 19.3% of the variability for a total of 41.3%. No clear definition can be drawn from this yet, but it may be a useful when evaluating the various classification methods. A plot of these clusters is shown in Figure B1.

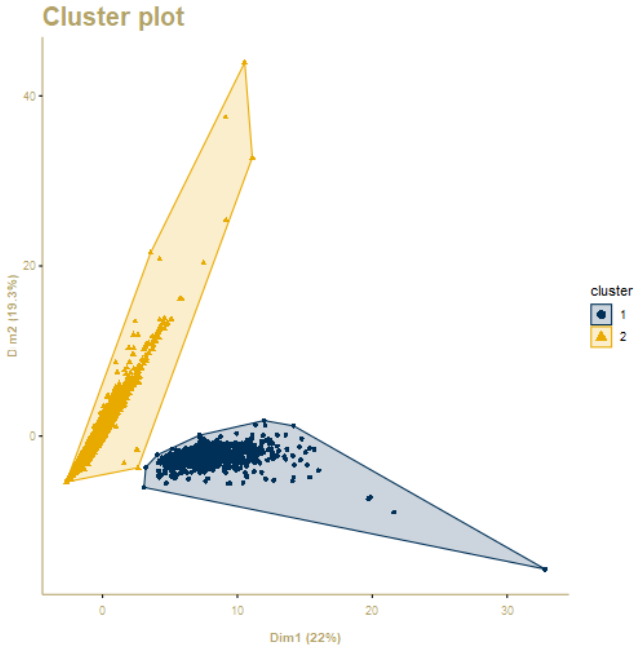

Figure B1: K-means Clustering Plot of *beerRecipe* dataset

Appendix C: Elimination of LDA

LDA was tested in this project to better understand the model. LDA was not an appropriate fit for our problem as LDA models are not effective for multinomial classification. LDA models also require that the data be on a normal fit, and the following Q-Q plot shows how the predictor ABV (this is common for other predictors as well in our dataset) does not fit a normal distribution.

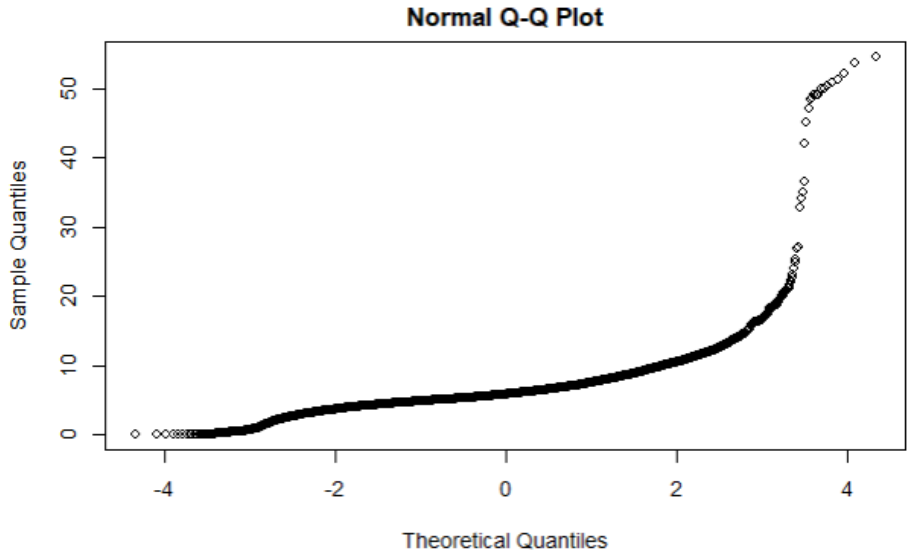

Figure C1: Normal Q-Q Plot of Predicting Variable *ABV*