STR459

Artificial Intelligence & Robotics

Exam

01.03.23 – 19.04.23

NHH

Candidate Numbers:

5, 32, 33, 37, 48

Total Number of Pages: 18

Total Number of Words: 4895

Question: Do you allow your report to be used as an example in the future: YES

Table of Contents

Distribution of Work

In terms of work distribution, two of us focused mainly on the Python script whilst the other three had the responsibility of writing the report. The line between these two areas became blurred as the work went on, so we would say that everyone contributed about equally.

Wine Quality AI Prediction

1.0 Introduction

Quality is a crucial and complex factor for both wine producers and consumers. This project aims to predict wine quality by analyzing chemical properties measured during the wine production cycle, utilizing machine learning (ML) models. Current wine production methods include both qualitative and quantitative tests to ensure that required quality metrics are met for each batch. The project will evaluate the accuracy and reliability of the ML models compared to these traditional wine quality assessment methods. The ultimate goal is to improve wine quality, consumer satisfaction, and potentially reduce the number of qualitative tests performed on each batch of wine. The ML models will be trained on a set of chemical properties and expert reviews of each wine. Considering the wine industry is a crucial industry in several economies around the world, this project has the potential to play an active role in developing wine industries around the world.

2.0 Data Analysis

2.1 Data Gathering

Studying the possibility of creating a ML model in order to solve any type of problem, one must make sure that applied data to train and test the model is appropriate, which is crucial to achieving usable results. As the aim of this project is creating a model which can sort out good batches of wine from bad batches of wine based on the chemical properties of a sample, there are a few key requirements which need to be met. First, the data must be from reliable sources and indicative of the measurements actual wine producers need to test wine batches. In addition to this, the data must contain information about the wines' quality in order to train the ML model.

One aspect of determining the quality of a given vintage or batch is acknowledging the subjectiveness of taste, which is inherent in determining how quality is perceived in a given wine. Conversely, quality can be determined by subjective reviews or objective measurements such as acidity and sweetness. Choosing between these two measurements determines which perspective one must look upon the output of the model in order to get any meaningful information out of it. If the basis of quality is reviewing scores, the output of the model suggests which wine tastes the best to a given number of wine reviewers. An objective measurement of quality is therefore very hard to create. When choosing the dataset to base this project on, there are certain specifications which are important in order to account for the subjectiveness of review scores. The review scores have to be based on the same

type of wine, and the reviewers have to use the same criteria in order to grade the various batches. With these criteria met, it is more likely that (though not guaranteed) the quality of a given batch is graded in a way which makes it possible to assume that they are unbiased.

One dataset which met these criteria is a set of white wine measurements from the Portuguese "Vinho Verde"-wines collected by P. Cortez et al. (2009). The data set consists of 11 columns of various chemical properties from a given batch of wine and its associated quality set by expert wine tasters. The data is loaded from a URL and imported into Python, the tool used to conduct further analysis of the data.

2.2 Cleaning the Data

Both data wrangling and cleaning are fundamental aspects of transforming raw data into a usable format for subsequent analysis. The primary objective of this stage is to ensure that the data's structure can serve as a foundation for machine learning models. The used data set was already presented in a format suitable for machine learning models. One important aspect to consider in this case is the possibility to implement additional datasets.

Missing data is a pervasive issue in research studies, and it is crucial to remove them before conducting statistical analysis. Missing data can distort the output of a model, making it challenging to identify patterns, relationships, and trends in the data accurately. Additionally, missing data can lead to biased estimates, thereby reducing the overall quality of the statistical analysis. As such, the removal of missing data is necessary to obtain an accurate and reliable view of the dataset as a whole. In the case of this data there were no NA values to be removed, and the datatypes were assigned correctly.

2.3 Descriptive Statistics

Prior to deploying a machine learning model to a collected dataset, it is imperative to undertake preliminary evaluations of the data. The purpose of this exercise is to mitigate the risk of any inherent biases within the dataset which is impacting the model's outputs, and which might subsequently alter the perception of the resulting outputs.

Understanding the distribution of the variables in the dataset is important when trying to identify underlying trends and patterns. Disregarding these underlying features of the data might lead to a misinterpreted analysis. Visualization is a powerful tool when determining the distribution of data. By

plotting the observed values against the number of occurrences of the given value, it is possible to get an overview of the distribution of the data. This is done for all variables in the dataset.

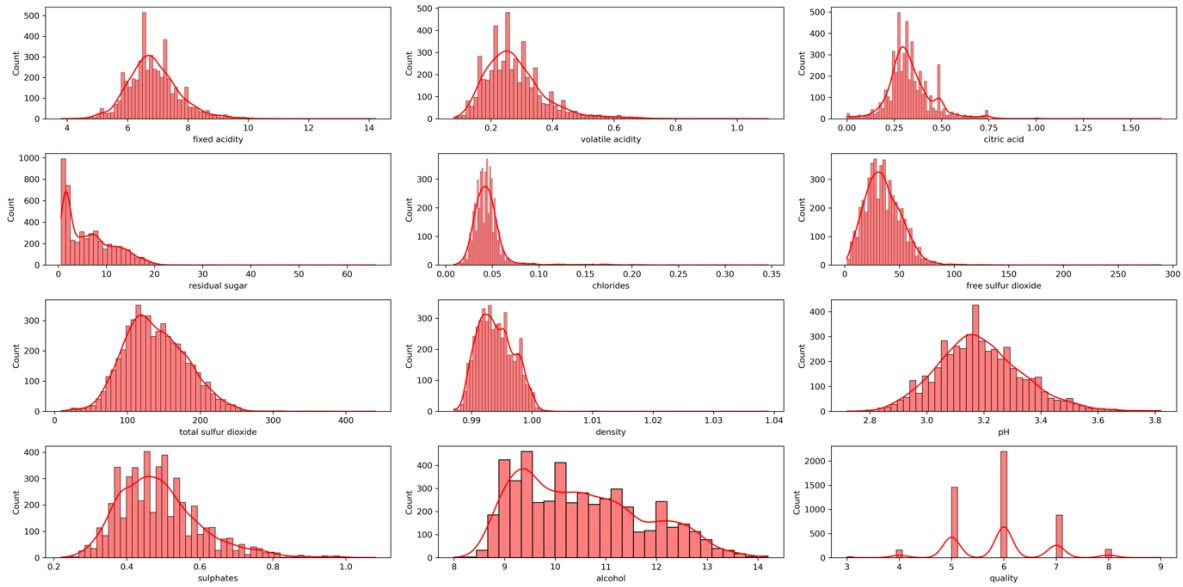

Figure 1 - Variable Distributions

Although most variables in the dataset resemble a normal distribution, many exhibit skewness and kurtosis that differ from it. Therefore, it might be more applicable to use quantiles rather than standard measurements when determining the distribution of the variables in question. It is also important to keep these outliers in mind when examining the output of a ML model trained on this dataset.

With the distribution of the variables established, the measurements for the central tendency of the data can be presented in order to gather more detailed knowledge. These measurements are used to describe the center of a given variable. They are also helpful when identifying anomalies in the data which might not be present in the visualizations.

	count	mean	std	min	25%	50%	75%	max
fixed acidity	4898	6.85	0.84	3.80	6.30	6.80	7.30	14.20
volatile acidity	4898	0.28	0.10	0.08	0.21	0.26	0.32	1.10
citric acid	4898	0.33	0.12	0	0.27	0.32	0.39	1.66
residual sugar	4898	6.39	5.07	0.60	1.70	5.20	9.90	65.80
chlorides	4898	0.05	0.02	0.01	0.04	0.04	0.05	0.35
free sulfur dioxide	4898	35.31	17.01	2	23	34	46	289
total sulfur dioxide	4898	138.36	42.50	9	108	134	167	440
density	4898	0.99	0.00	0.99	0.99	0.99	1.00	1.04
pH	4898	3.19	0.15	2.72	3.09	3.18	3.28	3.82
sulphates	4898	0.49	0.11	0.22	0.41	0.47	0.55	1.08
alcohol	4898	10.51	1.23	8	9.50	10.40	11.40	14.20
quality	4898	5.88	0.89	3	5	6	6	9

Figure 2 - Central Tendencies of Variables and Count

The quality variable which is the output of the eventual model is especially interesting. Even though it represents wine quality on a scale from 1-10, only 3-9 is represented as actual observations in the

data. The total number of observations is important to identify, as the size of the dataset might impact the predictive power of a model, and the ability to test its accuracy. The number of observations is represented by the *count* column, which is 4898.

Correlating variables are an important concept to consider when developing a machine learning model. More specifically we are interested in investigating the linear correlation between every variable. High correlation between predictors can indicate multicollinearity, which can cause instability and unreliable estimates in machine learning models. In contrast, a low or zero correlation between variables may indicate that the variables are not related or may require further exploration to understand their relationships. As such, investigating the correlation between variables is important for identifying relationships, selecting predictors, addressing multicollinearity, and developing accurate machine learning models. The correlation between the wines chemical variables is shown in the correlation matrix in Figure 3.

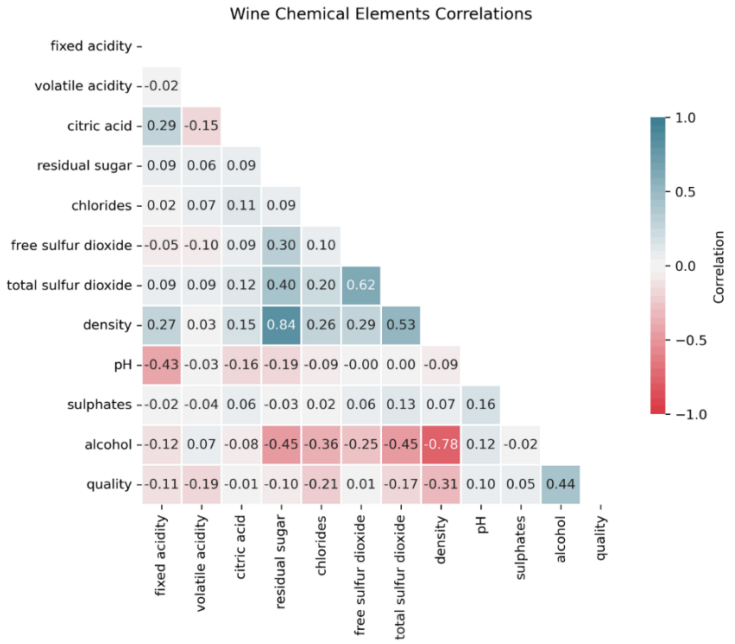

Figure 3 - Correlation Matrix

Following the correlation matrix analysis of the dataset's 12 variables, a significant correlation (>0.7) was identified between *density* and several other variables, indicating a possible overlap in their measurements. This correlation might lead to multicollinearity issues when training a ML model on this dataset, compromising its reliability (Witten, James, Tibshinari, & Hastie, 2017). To avoid multicollinearity and maintain a diverse range of variables in the statistical models, *density* was removed from the dataset. This approach ensures greater accuracy and facilitates meaningful conclusions from the data.

Based on the descriptive statistics presented, one could conclude that the dataset is suitable for applying in a machine learning environment. However, it is essential to consider certain inherent weaknesses that might affect the performance of the eventual model. The visualizations illustrate that many variables have outliers leading to difficulties in the machine learning process. Outliers can skew and mislead the training process of machine learning algorithms resulting in longer training time, less accurate models, and ultimately poorer results. The quality measurement might also cause some problems, as they are not entirely objective which in turn could lead to misclassifications in the model.

3.0 Building AI models

3.1 Classification Models

The goal of this project is to differentiate a bad batch of wine from a good one based on its chemical composition. This can be done by classifying a wine batch as either good or bad, whereas good wines are sent to bottling and the bad batches for further inspection. One way of determining which class should be assigned to each observation is using a machine learning classification model (Witten, James, Tibshinari, & Hastie, 2017). Classification is a supervised learning algorithm that learns from labeled data to predict class labels of new, unseen data. It finds the best decision boundary to separate the data into different classes and make accurate predictions.

Classification models typically use supervised learning techniques to create a robust model. This involves training the model on labeled data that consists of input and the corresponding output. The classification models then learn to map the input features to the output by minimizing the difference between its predicted output and the actual output. In this project four types of classification models including KNN, Logistical Regression, Decision Tree and Random Forest are applied, and their results are compared. The framework presented by Witten et al. (2017) is the basis of the following model dissertations. These models differ in their assumptions about the distribution of the data and the computational resources required for training the models.

3.1.1 KNN

K - Nearest Neighbors (KNN) is a machine learning classification algorithm that creates predictions for new data points by finding the k nearest neighbors in the training set and using their labels or values to estimate the label of the new data point. This classification model is non-parametric, and it does not assume any form of distribution on the data. Instead, it stores the entire training set and computes the distance between data points to determine the k nearest neighbors. One of the

strengths of the KNN algorithm is its simplicity and interpretability. Being computationally intensive, choosing K and the distance metric are weaknesses of the model as it might strongly impact the results and performance.

3.1.2 Logistic Regression

Logistic regression is a statistical machine learning classification algorithm. The logistic regression algorithm works by minimizing a cost function that measures the difference between the predicted probabilities and the actual labels in the training data. The output of logistic regression is always between 0 and 1, which is suitable for a binary classification task. Logistic Regression uses one out of two possible penalties: L1 or L2. The L1 has features where it can reduce the model's complexity, do feature selection and it works well with high-dimensional data. However, the main defect of choosing L1 is its sensitivity to correlated features and its tendency to be unstable. On the other hand, the L2 penalty can reduce the impact of statistical noise and improve the model stability. Nonetheless, L2 lacks efficiency in the feature selection compared to its L1 counterpart. (Scikit learn, 2023).

3.1.3 Decision Tree

Another supervised machine learning algorithm is the decision tree classification model. It is a hierarchical model that represents a sequence of decisions and their potential consequences making it similar to a tree-like structure. Each node of the tree represents a decision based on one of the input features, and each leaf node represents a predicted class or value. The algorithm works by splitting the data based on the features providing the most information. It splits the data and creates new nodes until a stopping criterion is met. One of the advantages of the decision tree model is that it can handle both numerical and categorical datapoints like KNN which is also a non-parametric model. Two disadvantages of the model, however, are its susceptibility to overfitting and sensitivity to small variations in the data set.

3.1.4 Random Forest

The last machine learning algorithm that will be presented in this paper is the random forest classification model. This algorithm builds a large number of decision trees on a random subset of training data, which is called a bootstrap sample, and it combines their predictions to improve accuracy and reduce overfitting. This machine learning model can handle missing data, outliers, and noise. It is also robust to nonlinear relationships between input and output. Although the model is highly accurate and robust, it also has some weaknesses. The algorithm can be computationally intensive for large datasets and may not perform as well on sparse or imbalanced datasets.

Furthermore, the results may also be less interpretable than a single decision tree as the predictions can be hard to trace back.

3.2 Building AI Models

The goal of this AI model is to use machine learning tools to predict which batches of wine is acceptable to serve to customers based on its chemical composition. When it comes to building the models there are certain assumptions that need to be established to secure the reliability of the model. In section 3.1 some of these assumptions were described. The choices to create the models are presented in the following sections.

3.2.1 KNN

Before creating the KNN model there are some assumptions and choices that must be disclosed. Specifically, the value of K and the distance measure. The choice of K is dependent on the input data. Data with more outliers and noise will perform better with a higher number of K as each observation receives less decision power. Generally, it is also recommended to have an odd number to avoid tied classifications. In determining the value of K, some analysis on accuracy, precision and error rate were performed. According to Figure 4, the best value of K is 5. The distance measure chosen for this model was Euclidean distance, which is the default option in the SKlearn library.

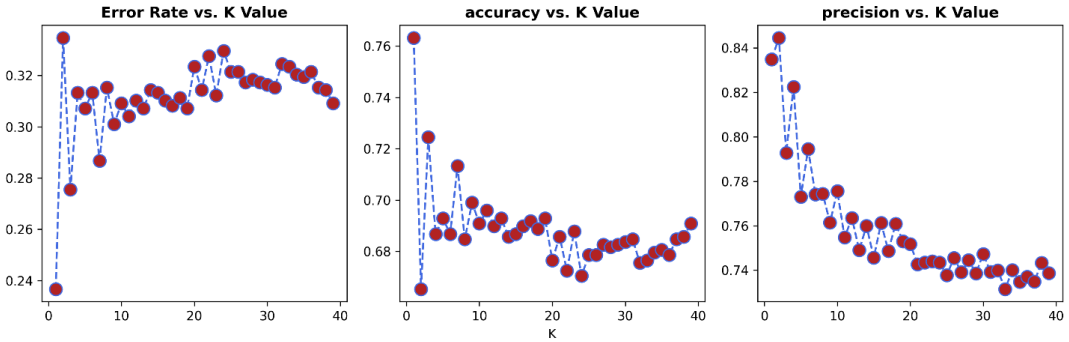

Figure 4 - K-value metrics

3.2.2 Logistic Regression

The *n* number of iterations chosen for the logistic regression was 100. It is important to know the repercussions by choosing too high or too low number of iterations. If the chosen *n* set too high, the regression could become overfitted. If it is set too low, the regression could struggle to find the optimal solution. The solver used in the logistic regression is liblinear, which solves binary classification fast and well for small and medium sized datasets (<10 000) (Scikit learn, 2023). Liblinear can exert two possible penalties on the data: L1 and L2. For this machine learning task, both the L1 and L2 penalty were implemented as their own models.

3.2.3 Decision Tree

As with the other models, assumptions and choices must be disclosed before creating the decision trees (Arora, 2020). The algorithm uses recursive binary splitting to grow a large decision tree using the node splitting function; high quality wine > 5.5. A large tree is grown on the training data using recursive binary splitting that stops when each terminal node has fewer than a minimum number of 2 observations. Then cost complexity pruning is applied to the large tree to obtain the sequence of the best subtrees as a function of alpha. The process of pruning is executed as a measure to avoid overfitting. To decide the variables used in the decision tree, K-fold cross validation is used. The cross validation optimizes the variables max depth to 15, criterion to Gini impurity, max features to square root and minimum sample split to 2. The subtree corresponding to these values will be the optimal decision tree for the given data set and called the tuned decision tree. The default decision tree will also be further investigated in part 4.1.

3.2.4 Random Forest

Many of the choices and assumptions of the random forest are similar to the decision tree due to the nature of the random forest model. However, the amount of randomness added to the model is determined through the parameter $mtry$. This determines the creation process of the decision trees, and specifically it controls how many of the input features a decision tree has available to consider. By default, random forest uses $\sqrt{p}$ variables when building the classification trees. In the random forest models created here $mtry = \sqrt{p}$. When tuning the random forest hyperparameters, the accuracy did not become better. This is common and an argument for using the default hyperparameters. Both the tuned and the default model will be investigated further in section 4.1.

All four models use an algorithm to predict the output variable using the input variables. Further, they also divide the data into a train and a test data set to determine the accuracy of their prediction. In the dataset there is given a quality score to each batch. These scores are set by professionals after tasting a given wine (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009). The quality range is from 1-10 and the definition of an acceptable batch is a score greater than or equal to 5.5. The output variable is defined as a binary quality score based on the acceptance rate. Although the output variables must be categorical, the input variables can be continuous in these classification models.

The input variables are structured into a matrix which the machine learning algorithm will use to predict the binary quality score. By excluding density data as a predictor, the machine learning algorithm is also spared from possible multicollinearity issues due to the high correlation between it

and several other variables. This will also increase the machine learning algorithms' accuracy to predict the binary wine quality score.

4.0 Evaluating AI Model

The models created in section 3.0 will now be compared and evaluated against each other. The comparison and evaluation of machine learning algorithms is based on the predicted values against the actual outcome. The main criteria used when determining the quality of the AI models are how well they can classify observations which were not a part of their training set. The optimal model for this application will be used to develop a business case that demonstrates how it can lower the expenses of quality control in wine production.

4.1 Comparing the Models

4.1.1 Confusion Matrix

The confusion matrix is a measure to evaluate the accuracy of a specific model. It displays the predicted and the observed values in a table and provides a statistical summary of the results. The tables below illustrate the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) to assess the accuracy of each model. The most important factor to consider when evaluating this confusion matrix is the proportion of FP. This is because the negative impact of a FP is greater than the impact of a FN. From the confusion matrices in Figure 5, it can be observed that the default random forest model has the lowest number of FP and FN, indicating that this is the preferred model.

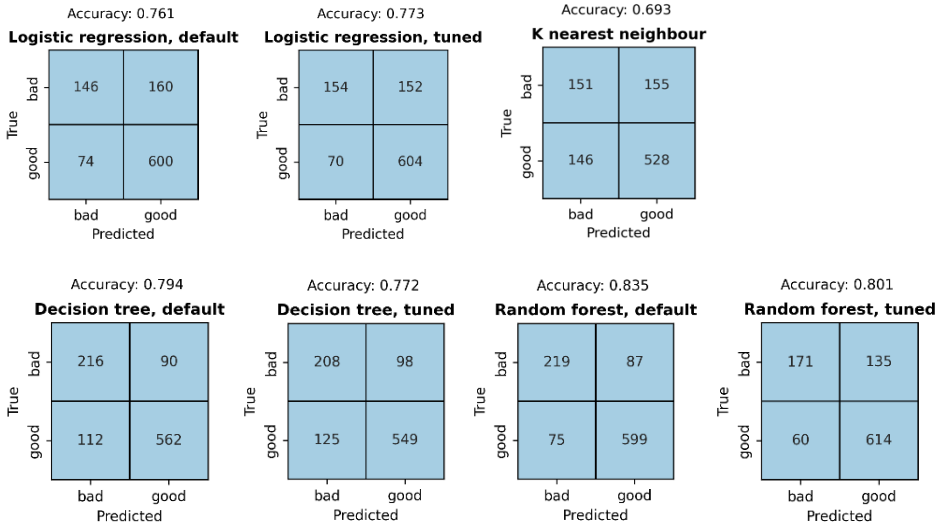

Figure 5 - Confusion Matrices

4.1.2 K-Fold Cross Validation

K-fold cross validation is a good technique when evaluating the performance of the model, as it divides the train and test data sets into random chunks K-Fold times. This provides a higher degree of accuracy and precision of the models' outputs. The data set was divided into 15 validation sets as 15 folds was utilized. The results of the K-fold cross validation are illustrated in Figure 6. These figures compare the models' accuracy and precision. The variance on the accuracy figure is larger than the precision figure. This shows that the models are similar in precision and vary more when it comes to accuracy. The most accurate and precise model is the random forest model.

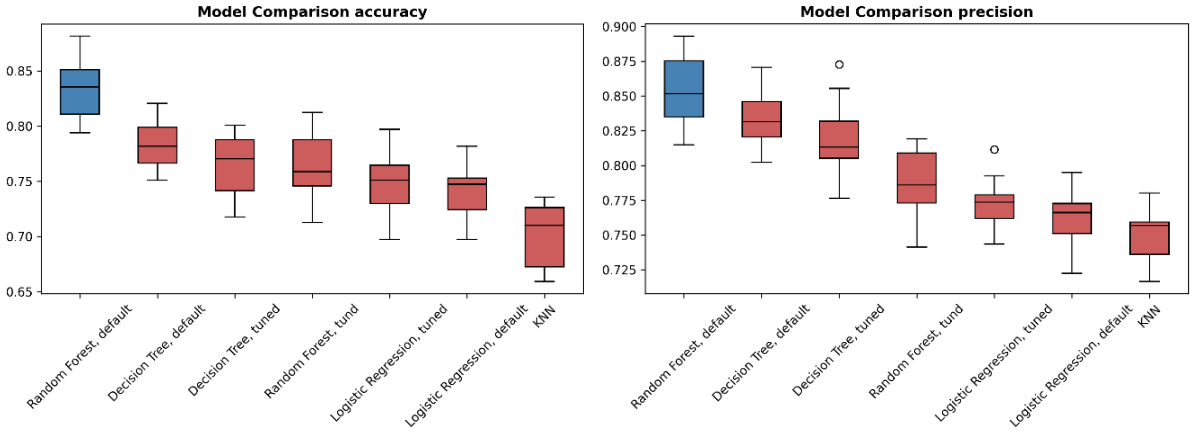

Figure 6 - Accuracy and Precision of the Models

4.1.3 ROC

The Receiver Operating Characteristics Curve (ROC Curve) is a visual way of displaying the two types of errors for all possible thresholds. This makes it a good tool when comparing different models to each other. With an area under the curve (AUC) of 0.91 the default random forest model is the most precise. This is shown in Figure 7. KNN is the worst performing model, and both of the decision tree models are worse than the default random forest model. This means that the default random forest has the highest rate of TP and the lowest rate FN as shown in the confusion matrix in Figure 5. There is a trade-off between sensitivity and specificity when it comes to choosing the optimal model. The ROC curve allows for the visualization of this trade-off. Having higher sensitivity will in this scenario be more important than having lower specificity as a wine brewer would prefer the predictions to be as accurate as possible. However, the default random forest has the best results, and the tradeoff is not necessary.

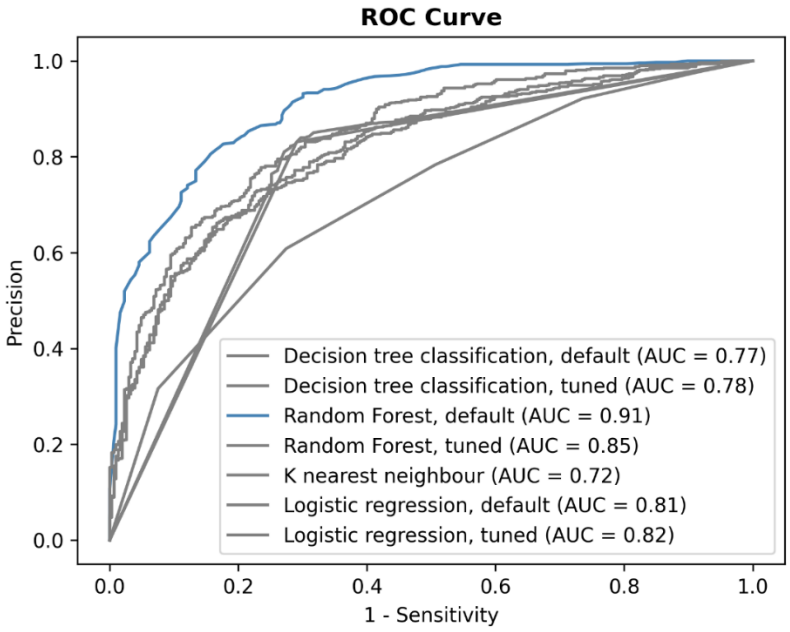

Figure 7 - ROC curve of the Models

After examining the key features that influenced the choice of the best model for this use case, the random forest classification model was recommended for the wine quality control application. This model outperformed the others in predicting a significant proportion of the bad batches with high accuracy, with an accuracy of 83.5%. To summarize, this project has shown that given enough samples, AI models are able to predict to an acceptable accuracy which batches of wine are of high quality. This sets the stage for the last part of this project, where a business application of this model is proposed.

4.2 Business Potential

How can sorting out bad batches of wine with ML Models increase profits for wineries?

Machine learning is a hot topic in several areas of the economies around the world as of late, with wine production being a key area within agriculture where technologic progress is impacting the value chain of wineries. Wine production is a major industry around the world and has traditionally been a cornerstone industry for many economies located in temperate climates around the world. The Mediterranean part of the EU is considered the main production center of the world when it comes to wine, with wine production accounting for around 1% of the EU GDP in 2021. This equates to around 136 billion euros as of 2021 (Statista, 2023).

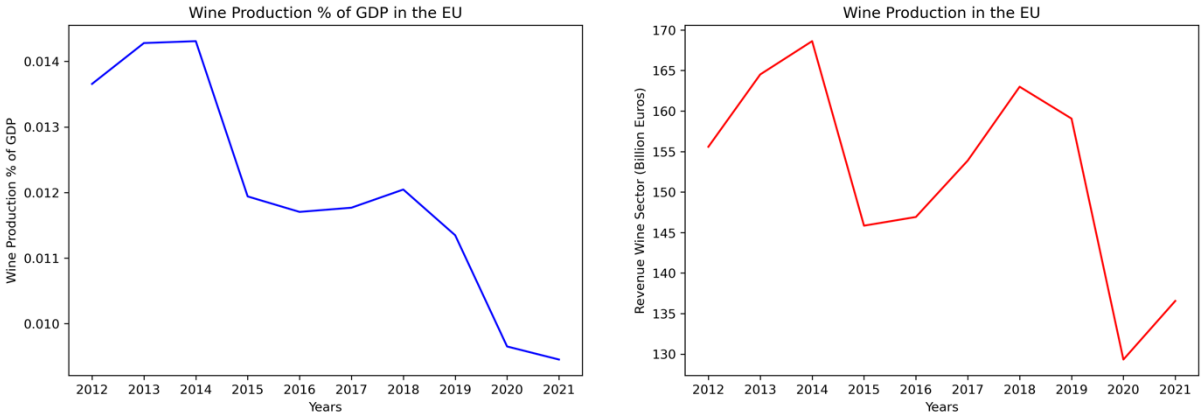

Figure 8 - Wine Production in the EU (Statista, 2023)

The industry contributes significantly to the continent's economy and has a distinctive role in the agricultural field. This wine market exhibits the characteristics of a monopolistic competition, which sets it apart from other agricultural markets. (Rebelo, Gouveia, Lourenço-Gomes, & Marta-Costa, 2017). Various producers are able to differentiate themselves based on production regions, consumer preference, price, and marketing. Thus, different brands of wine are not perfect substitutes for each other, which lends itself to an asymmetric industry where both large and small producers can thrive in their niche.

This separation leads to an air of secrecy around the production process of a given winery. The wine industry is in general very secretive around how they perform quality control, and which parameters are especially important to them. It is therefore hard to get a complete picture of how the current quality control of wineries is practically done. On the other hand, there are multiple job listings from the 10 biggest wine manufacturers (example E. & J. Gallo Winery) for quality control personnel with education in chemistry and engineering (Indeed, 2023). This is indicative of the fact that quality control is done by both chemical measurements and expert testing.

This project proposed a machine learning model that could reduce the reliance on subjective expert testing at wineries and enhance the efficiency of the production process. Wine producers face no shortage of demand for their products, which often leads to producers focusing more on their cost side of their revenue when evaluating ways to improve revenue. A model which is able to make wineries quality control departments run more efficiently would therefore most likely have a high demand in the market, considering the market's relevance, scale, and composition.
This theory is accentuated by the fact that a large portion of actors in the market is producing wine at a smaller scale, which lends itself well to optimizations provided by specialized machine learning models. In addition to this, the marginal cost of running the model on a wine sample is negligible, as

the wineries are already taking such samples of their wine to make sure they are within regulatory standards of the industry, as they produce edible products.

The ML model which has been created in the project is highly adaptable to different wines, given that the amount of data is sufficient. A business proposal for such a model might be a tool used for reducing costs at a winery's quality control department. There are several ways to implement such a model, with an example presented below in Figure 9.

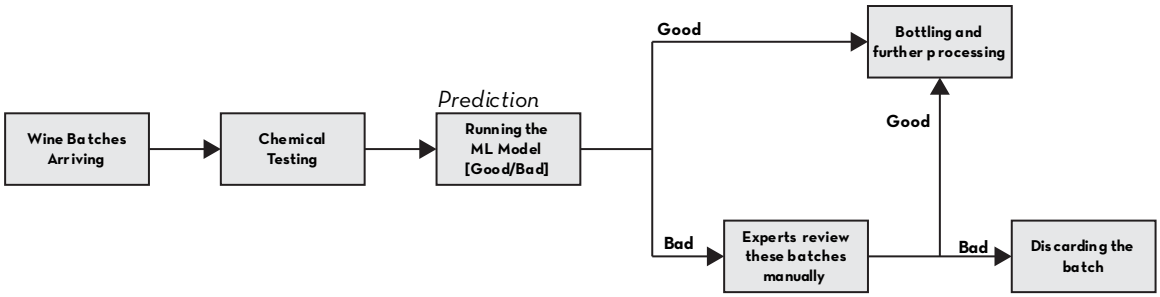

Figure 9 - Visualization of Imagined Production Process with the ML Model

This implementation assumes that the goal of this example winery is to reduce the total costs of running the quality control department whilst keeping the designation accuracy at a high level. Given that the quality control personnel have a 100% detection rate of bad batches, the department has two costs:

C_L: $Cost\ of\ quality\ control\ labor$
C_B: $Cost\ of\ bottling\ a\ bad\ batch\ of\ wine$

The performance of the model is measured in several parameters, but because the predicted bad batches are sent to quality experts, the only cost accrued from a FN is opportunity costs compared to testing all wine batches. On the other hand, bottling and selling a bad batch of wine is caused by mislabeling a bad batch as good, which is a FP. The C_B is defined as the sum of the actual costs of returning wines sold positive and the negative externalities of delivering a bad batch of wine. This leaves the model with these parameters:

N: $Total\ number\ of\ batches\ processed$
FP: $Number\ of\ false\ positives\ (bad\ batches\ sent\ to\ bottling)$
FN: $Number\ of\ false\ negatives\ (good\ batches\ sent\ to\ further\ testing)$
TP: $Number\ of\ true\ positives\ (good\ batches\ sent\ to\ bottling)$
TN: $Number\ of\ true\ negatives\ (bad\ batches\ sent\ to\ further\ testing)$

A_M: $Accuracy\ of\ the\ model$

From here it is possible to estimate the area in which the ML model is advantageous compared to using the quality control personnel on all the batches:

Total cost without ML model is defined as cost of labor times the number of batches controlled:

$$C_L \cdot N \tag{1}$$

Total cost with ML model is given as the total cost of bottling a bad batch times the number of FP added with the cost of quality control labor times the sum of TN and FN:

$$C_B \cdot FP + C_L \cdot (TN + FN) \tag{2}$$

The ML model produces lower costs than the labor driven quality control if the total cost of the department without the model is greater than the total cost with the ML model:

$$C_L \cdot N > C_B \cdot FP + C_L \cdot (TN + FN) \tag{3}$$

The accuracy of the model is measured as the number of TP and TN divided by number of batches:

$$\frac{TP + TN}{N} = A_M \tag{4}$$

Combining the model accuracy measure with the differential equation in (3), it is possible to simplify the equation to:

$$C_L \cdot N - C_L(TN + FN) > C_B \cdot FP \tag{5}$$

$$\frac{C_L(N - TN - FN)}{FP} > C_B \tag{6}$$

$$N - TN - FN = TP + TN \tag{7}$$

$$\frac{C_L \cdot A_M \cdot N}{FP} > C_B \tag{8}$$

It is therefore possible to examine whether or not the model is worth implementing for a given winery based on their total costs of bottling bad batches and the accuracy of the model. If the amount which is saved $(C_L \cdot A_M \cdot N)$ by implementing the model per unit of FP is greater than the C_B cost, then it becomes a worthwhile investment for a winery. Whilst it might not be viable for all wineries to implement this system based on their costs of bottling bad batches, the accuracy of the model will continue to improve based on the assumption that the data set used to train the model is

growing over time. This data collection and model implementation could therefore become an interesting business opportunity within the wine industry. However, it should be mentioned that the assumption that quality control experts have a 100% detection rate is a simplification which could impact the exact implementation of the model in a real setting.

5.0 Summary of findings

The aim for this project was to increase the efficiency of the determining wine quality of a given batch using machine learning models. Several classification models were applied to a data set of real observations of the chemical properties of Vinho Verde white wines and their performance was compared to identify the most suitable model for this problem. The random forest classification model outperformed the other models in accuracy and precision, which were the key metrics for this problem. This provided the foundation for the business application. By discussing the structure of the wine industry in the context of possible entrances of a company selling an implementation of this model, this project managed to find an equation which made it possible to assess whether or not implementing the model were economically feasible. This project demonstrates the feasibility and utility of machine learning models for distinguishing good and bad batches of wine based on their chemical properties. It also proposes a practical business application of the model for improving wine quality and reducing costs. This project highlights the importance of data collection as a key factor for enhancing the performance and scalability of the model in the wine industry.

References

Arora, S. (2020, 10 2). *Analytics Vidhya*. Retrieved from Let's Solve Overfitting! Quick Guide to Cost Complexity Pruning of Decision Trees: https://www.analyticsvidhya.com/blog/2020/10/cost-complexity-pruning-decision-trees/

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *In Decision Support Systems, Elsevier*, 47(4):547-553.

Indeed. (2023, April 16). *Find jobs: Wine Chemistry*. Retrieved from Indeed: https://www.indeed.com/q-Wine-Chemistry-jobs.html?vjk=4fe7bfb625408769

Rebelo, J., Gouveia, S., Lourenço-Gomes, L., & Marta-Costa, A. A. (2017). Wine Firm's Size and Economic Performance: Evidence from Traditional Portuguese Wine Regions. *InTech*, DOI: 10.5772/intechopen.71320.

Scikit learn. (2023, April 15). *sklearn.linear_model.LogisticRegression*. Retrieved from Scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Scikit learn. (2023, April 16). *sklearn.linear_model.Ridge*. Retrieved from Scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Statista. (2023, April 16). *Revenue of the wine market in Europe from 2012 to 2025*. Retrieved from Statista: https://www.statista.com/forecasts/1242555/europe-wine-market-revenue

Witten, D., James, G., Tibshinari, R., & Hastie, T. (2017). *Introduction to Statistical Learning.* Springer-Verlag New York Inc. .