Predicting Quality of Wine Samples

ISM6136.901F22.90231 Data Mining

Final Paper

Anirudh Voruganti | Bharathi Pemmasani | Sri Guru Achuth Gadicherla | Sravan Kumar Talishetti

1. Introduction

Winemaking is a centuries-old craft that has accumulated a wealth of information about the circumstances that cause grape juice to turn into wine via careful experimentation with yeast-driven fermentation. Wineries worldwide have developed the wide range of wine types we enjoy today by building on this expertise. Vintners produce distinctive blends of flavor, fragrance, color, and clarity by blending grape types, yeast strains, and additives in clearly defined, meticulously watched manufacturing methods. [1]

Through this project, we want to explore a data mining approach to predict the quality of the wine based on analyzing the components present in a wine, such as sugar, alcohol, acidity, and pH levels. We envisioned a scenario where the results of this approach would help a new-age entrepreneur succeed in the winemaking industry with the help of analytics.

The key business questions we intended to address are as follows.

- How can an entrepreneur assess whether their firm is producing quality wine?
- Can predictive modeling help in standardizing quality control measures at wineries?

Who Would Care?

- **Consumers**- Since they are the product's end users, the wine produced by the winery must be of the highest quality possible.
- **Regularity Authorities**- As wine is one of the most consumed beverages, the regularity authorities need to ensure that the wine produced is safe for consumption.
- **Vintners -** In the wine industry, success means retaining customer loyalty by offering the same great wines in the crowded wine landscape. Despite the complex interactions that influence winemaking outcomes, there is a need to achieve that level of consistency and reproducibility in production.

A large dataset was considered for our project, with white and red wine samples. Multiclass Classification models such as Multiclass Neural networks, Multiclass Logistic Regression, and Multi Class Decision Forest algorithms were applied to assess the best fit for our dataset. From our analysis, we observed that multiclass decision forest, an ensemble model provided the highest accuracy for both data sets.

2. Data

Two datasets were investigated, connected to red and white varieties of the Portuguese "Vinho Verde" wine. Due to privacy and logistical concerns, only physicochemical (input) and sensory (output) variables are provided (e.g., there is no information about grape varieties, wine brands, wine selling price, etc.). [2]

Each wine in this dataset is assigned a "quality" score ranging from 0 to 10. We modified the result to a categorical output for this project and created a new column "Quality", where each wine has three distinct outcomes, Low, Medium, and High, but each instance is assigned only one outcome.

The outcomes are segregated as follows.

Low: A "quality" score of less than or equal to four.
Medium: A "quality" score of less than or equal to seven and greater than four.
High: A "quality" score of greater than seven.

The below eleven input factors influence wine quality (based on physicochemical tests).

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulfates
11. Alcohol

Source of Data: [UCI Machine Learning Repository]

Drawbacks:
The main drawback observed in our dataset was that classes in the quality variable are ordered but not balanced. Below is the distribution of Quality variables in both the datasets.

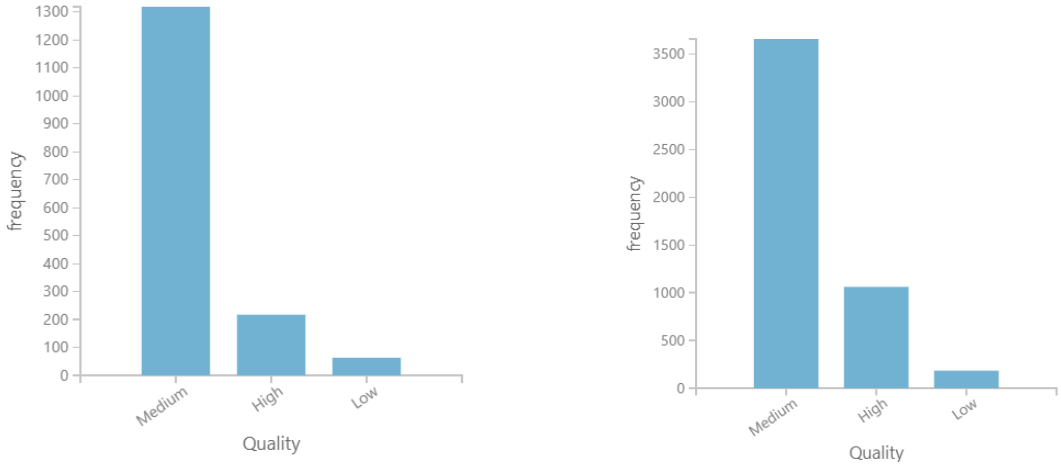


Figure 1: Distribution classes in red wine(left) and white wine (right) Datasets

3. Methodology

3a. High Level Overview

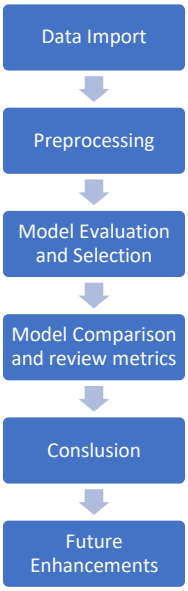


Figure 2 Flow chart of Methodology used

We started by importing both datasets to the Azure ML Studio, then created two diagrams for red and white datasets. Next, we performed a statistical analysis of the input variables.

Summary of Statistics for White Wine raw dataset

rows	4898
columns	13

	Mean	Median	Min	Max	Standard Deviation	Unique Values	Missing Values	Feature Type
fixed acidity	6.8548	6.8	3.8	14.2	0.8439	68	0	Numeric Feature
volatile acidity	0.2782	0.26	0.08	1.1	0.1008	125	0	Numeric Feature
citric acid	0.3342	0.32	0	1.66	0.121	87	0	Numeric Feature
residual sugar	6.3914	5.2	0.6	65.8	5.0721	310	0	Numeric Feature
chlorides	0.0458	0.043	0.009	0.346	0.0218	160	0	Numeric Feature
free sulfur dioxide	35.3081	34	2	289	17.0071	132	0	Numeric Feature
total sulfur dioxide	138.3607	134	9	440	42.4981	251	0	Numeric Feature
density	0.994	0.9937	0.9871	1.039	0.003	890	0	Numeric Feature
pH	3.1883	3.18	2.72	3.82	0.151	103	0	Numeric Feature
sulphates	0.4898	0.47	0.22	1.08	0.1141	79	0	Numeric Feature
alcohol	10.5143	10.4	8	14.2	1.2306	103	0	Numeric Feature

quality	5.8779	6	3	9	0.8856	7	0	Numeric Feature

Summary of Statistics for Red Wine raw dataset

rows	1599
columns	13

	Mean	Median	Min	Max	Standard Deviation	Unique Values	Missing Values	Feature Type
fixed acidity	8.3196	7.9	4.6	15.9	1.7411	96	0	Numeric Feature
volatile acidity	0.5278	0.52	0.12	1.58	0.1791	143	0	Numeric Feature
citric acid	0.271	0.26	0	1	0.1948	80	0	Numeric Feature
residual sugar	2.5388	2.2	0.9	15.5	1.4099	91	0	Numeric Feature
chlorides	0.0875	0.079	0.012	0.611	0.0471	153	0	Numeric Feature
free sulfur dioxide	15.8749	14	1	72	10.4602	60	0	Numeric Feature
total sulfur dioxide	46.4678	38	6	289	32.8953	144	0	Numeric Feature
density	0.9967	0.9968	0.9901	1.0037	0.0019	436	0	Numeric Feature
pH	3.3111	3.31	2.74	4.01	0.1544	89	0	Numeric Feature
sulphates	0.6581	0.62	0.33	2	0.1695	96	0	Numeric Feature
alcohol	10.423	10.2	8.4	14.9	1.0657	65	0	Numeric Feature
quality	5.636	6	3	8	0.8076	6	0	Numeric Feature

Observations based on the above tables and histograms & boxplots from Azure ML Studio

- There were no missing values in our datasets, and given these variables, there is not much room for feature engineering.
- There are outliers in citric acid, volatile acidity, and fixed acidity. The distribution of the variables may be assumed to be symmetrical if those outliers are removed.
- The distribution of residual sugar is positively skewed; even after removing the outliers, the distribution will still be skewed.
- Some of the variables have a few outliers, such as free sulfur dioxide and density, but they are considerably distinct from the rest.
- The majority of outliers are enormous.
- Alcohol's distribution is asymmetrical, although there are no apparent outliers.

3b. Data pre-processing

We have used the "edit metadata" option under data transformation-Manipulation to change our target variable "Quality" column to a categorical feature from a string feature. Since all

models mostly use only numbers, we have converted this variable to a categorical variable to obtain a meaningful analysis. [3]

Also, we have used the "Select Columns in Dataset" option under data transformation-Manipulation to drop the "quality column," which was included in the raw dataset, to use only the "Quality" column, which was created manually.

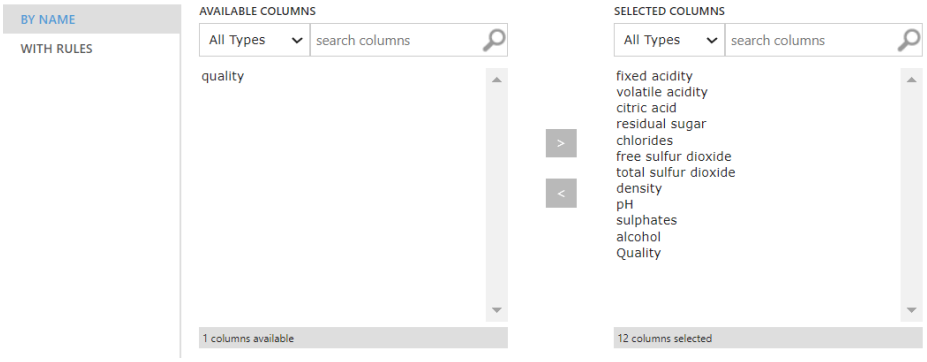


Figure 3 Snapshot of column selector

3c. Methods used

In our use case, classification algorithms were employed to forecast, based on prior observations, the categorical class labels of future cases. Our case is a multi-class classification issue since our target variable, "Quality," has three distinct outcomes: Low, Medium, and High, yet each occurrence is only given one.

We have used the "Filter Based Feature Selection" option with the "Pearson Correlation" Feature scoring method to identify the features in our dataset with the most significant predictive power.

Pearson Correlation: In statistical models, the r-value is another name for Pearson's correlation statistic, often known as Pearson's correlation coefficient. It returns a result representing the degree of correlation for any two variables. The covariance of two variables is divided by the product of their standard deviations to calculate Pearson's correlation coefficient. Scale shifts in the two variables have no impact on the coefficient. [4]

Below is the summary of the results obtained

Quality	alcohol	density	volatile acidity	chlorides	total sulfur dioxide	residual sugar	free sulfur dioxide	fixed acidity	pH	citric acid	sulphates
1	0.385414	0.284002	0.196927	0.184727	0.1712	0.138254	0.106087	0.105657	0.09357	0.058824	0.051141

Figure 4 Snapshot of feature selection for White Wine dataset

Quality	alcohol	volatile acidity	citric acid	sulphates	total sulfur dioxide	density	fixed acidity	pH	free sulfur dioxide	chlorides	residual sugar
1	0.40737	0.33724	0.230655	0.209193	0.163641	0.151575	0.127413	0.107958	0.107336	0.101215	0.053882

Figure 5 Snapshot of feature selection for Red Wine dataset

Looking at the Filter Based Feature Selection output, results, we see that the alcohol variable has the highest predictive power on our target variable in both the datasets.

Since the variables "citric acid and sulphates" have the lowest predictive power on the target variable, we decided to drop those from columns from the white wine model and the "residual sugar" variable from the red wine model. The decision to drop these columns was made to improve classification accuracy and efficiency.

Data split: We divided the data into training and testing sets using the Split Data module. The data was split into two portions using the "Split Rows" option. We specified the 70-30 percent of data to put in each split. We also randomized the selection of rows in each group.

Model Selection

We implemented three multi-class classification algorithms in our model: multi-class neural networks, multi-class logistic regression, and multi-class decision forests to solve our business questions and assess the best fit for our use case.

Multi-class neural networks: Using a standard neural network to simulate a multi-class classification issue is naturally appropriate. It makes probabilistic predictions for several classes of a target variable. The neural network in multi-class classification has the same number of output nodes as classes. Each output node produces a score for the class to which it belongs.

Multi-class logistic regression: The most common method for forecasting binary classes is a binary logistic regression. It uses the logit function to forecast the likelihood that a binary event will occur. It is not intended to mimic a target with several classes, however. The conversion of the multi-class target into one-hot variables and fitting a typical logistic regression model on each one are logical extensions for modeling a multi-class classification issue.

Multi-class decision forests: The decision-tree model has its challenges. It could overfit a specific dataset. For that reason, ensemble models come into play. The random forest approach creates several decision trees by drawing many random samples instead of merely using one decision tree. These sampling and modeling operations are carried out concurrently and separately. The result is the average of what all of the models anticipated.

A tree-based approach naturally models a multi-class classification issue. The random forest may simulate a multi-class classification issue since it inherits the tree-based technique.

4. Related Work

This part comprises journal paper referred for better understanding and designing of the project.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis,Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems [2]

This paper has been the base reference for our project; the researchers of this publication have compiled the dataset.

The researchers applied three regression approaches in a computationally efficient process that simultaneously selects variables and models. They observed that the results from the support vector machine were more positive than those from multiple regression and neural network techniques.

The researchers provided a case study for predicting taste preferences based on analytical data readily accessible during the wine certification process in this publication. Building such a model is advantageous for customers, wine producers, and even certifying organizations. It may be utilized to back up the oenologist's assessments of the wines, thereby enhancing the accuracy and timeliness of their choices. Additionally, assessing how the physicochemical tests affect the finished wine's quality might help refine the manufacturing procedure. Additionally, it can aid in target marketing by modeling customer preferences for profitable and specialized markets using comparable methodologies.

The primary contributions of this study include the following:

- The presentation of a unique method that simultaneously selects variables and models for SVM and NN. Sensitivity analysis, a computationally effective technique that assesses input relevance and directs the variable selection process, is the foundation for the variable selection process.
- The prediction of Vinho Verde wine taste preferences (from the Minho area of Portugal) was tested in a real-world application, demonstrating the approach's effectiveness in this field. In contrast to earlier research, a sizable dataset is considered, including 4898 white and 1599 red samples. They demonstrated how the specification of the tolerance concept helps access various performance levels by modeling wine preferences under a regression technique that maintains the order of the grades.

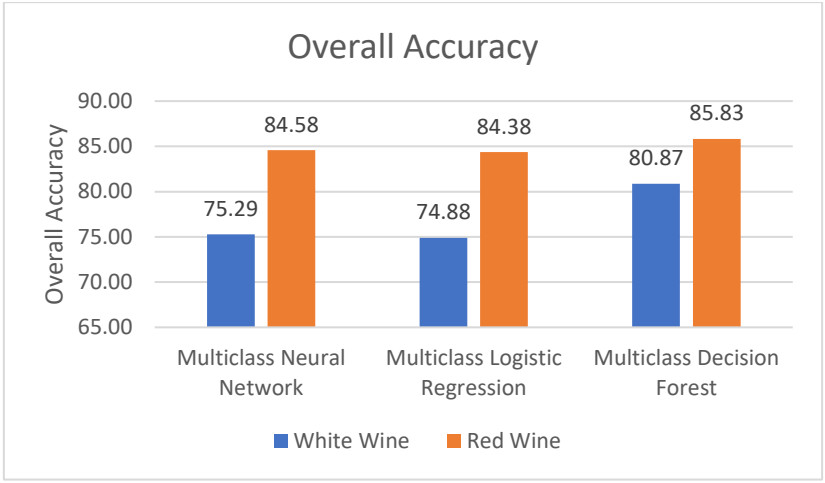
How does our methodology and work compare to what has been done before.

As mentioned earlier, this research has been the basis for our project and the dataset; the researchers have implemented regression models for building the model which predicts human wine taste preferences which is quite the opposite of our methodology and use case; In our project we have implemented classification algorithms to predict the quality of the wine which would help in automated quality control measures. Our model is an extended application of this publication.

5. Results

Below is the summary of results obtained by running our classification algorithms

1) ***Accuracy:*** Simply dividing the number of forecasts by the number of correct predictions gives us a model's overall accuracy. An accuracy score will range from 0 to 1, with 1 being the ideal model.
   When data is skewed, and one class is significantly bigger than another, this measure should only be used with others since the accuracy might be deceptive.



2) Confusion Matrix: A confusion matrix is a beneficial tool for identifying the areas where the model is incorrect (or accurate). It is a matrix that contrasts the proportion of accurate and inaccurate predictions for each class.

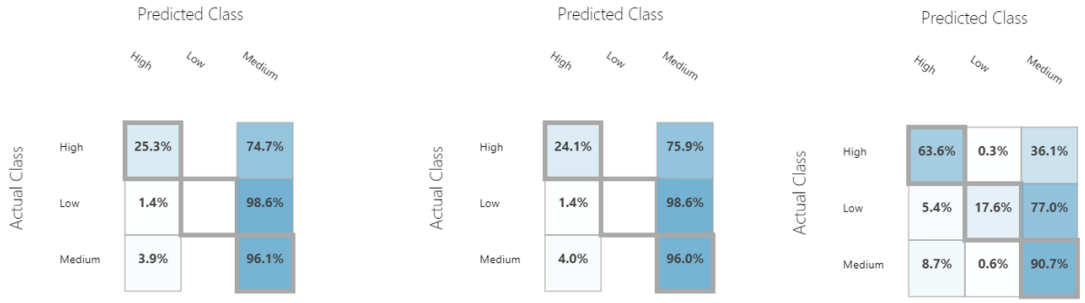


Figure 6 Confusion Matrix for White Wine, Neural Network (L) Logistic Regression (Cen) Decision Forest (R)

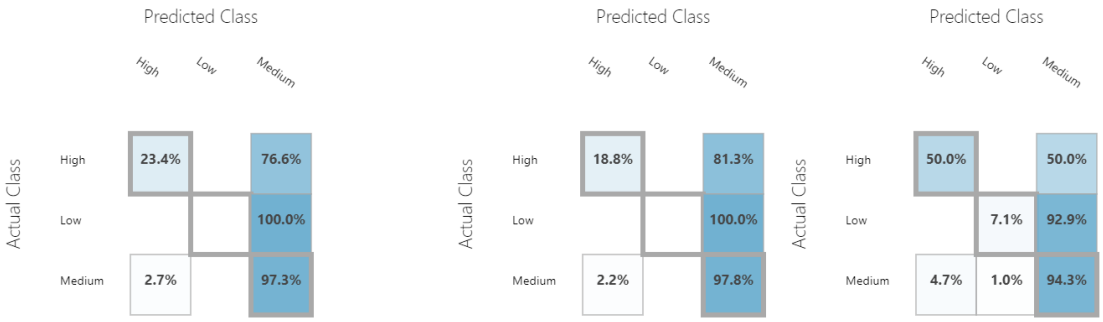


Figure 7 Confusion Matrix for Red Wine, Neural Network (L) Logistic Regression (Cen) Decision Forest (R)

6. Conclusion

Takeaways for Management

From our analysis, we observed that multiclass decision forest, an ensemble model provided the highest accuracy for both data sets.

For our business case, we recommend implementing a predictive modeling system using multiclass decision forests to standardize quality control measures at wineries.

Future Scope

We would like to collect more data samples or perhaps include live and proprietary data to train the model to get better results.

Since we have a class imbalance in the quality variable, we planned to implement oversampling methods such as SMOTE to solve the imbalance problem.

References

[1] R. Hartwell, "Analytical Testing in Wine Making: Transform Quality Control Into Quality Design," 2021.

[2] A. C. F. A. T. M. J. R. Paulo Cortez, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547-553, 2009.

[3] S. I. Inc, Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach, Fourth Edition, Cary, NC, USA: SAS Institute Inc, 2018.

[4] Microsoft, "Filter Based Feature Selection," Microsoft , [Online]. Available: https://learn.microsoft.com/en-us/previous-versions/azure/machine-learning/studio-module-reference/filter-based-feature-selection?redirectedfrom=MSDN.

Appendix

Azure ML Studio Workflow Diagrams

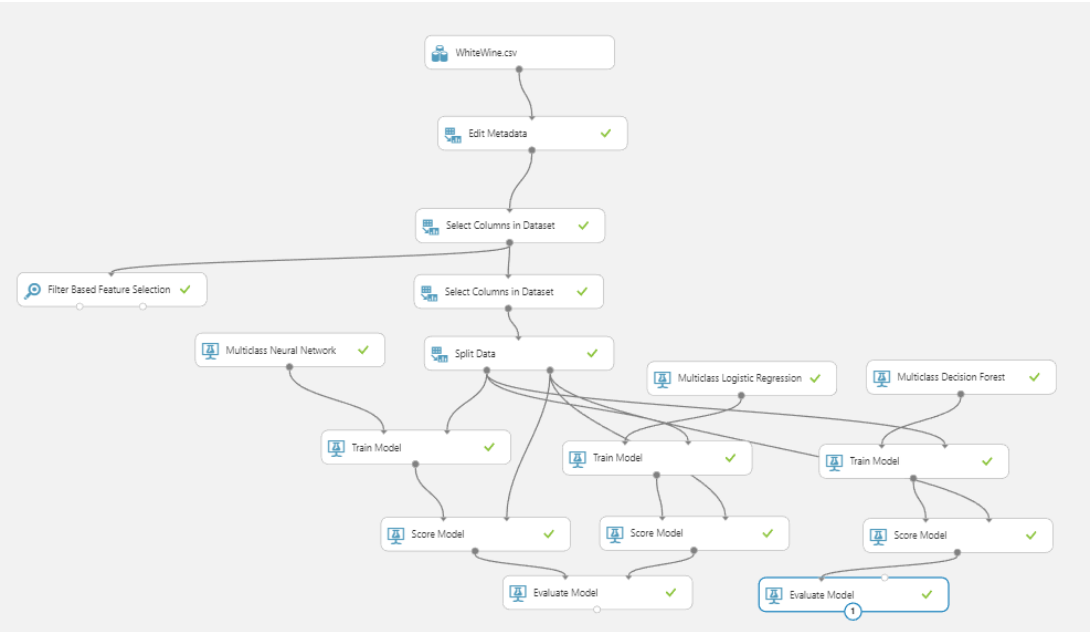


Figure 8 Azure ML Studio Workflow Diagram for White Wine

Properties used for building the Model

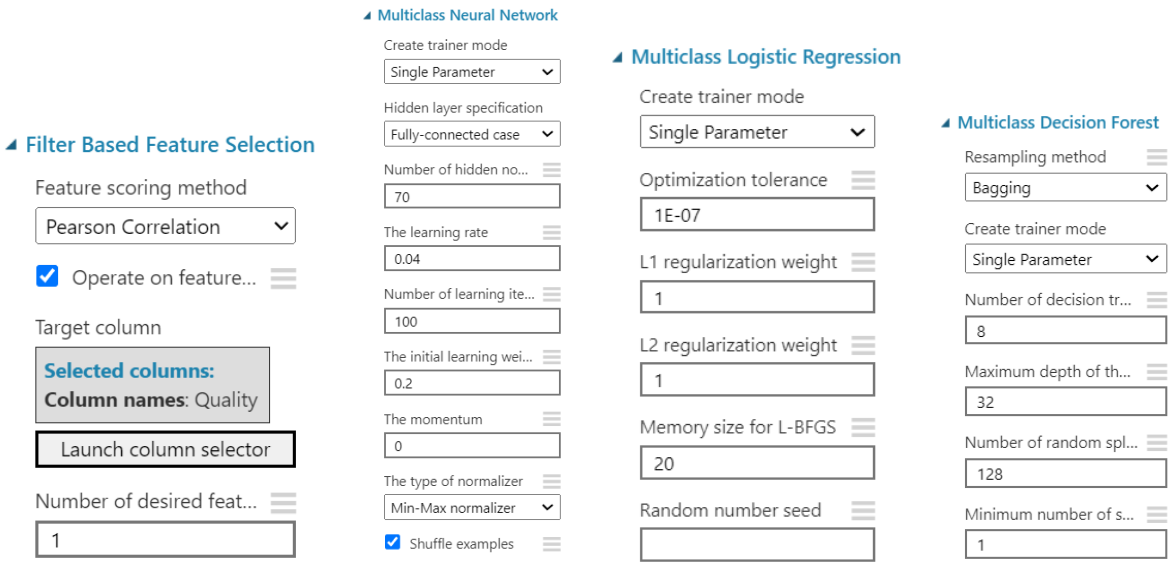


Figure 9 Parameters used for building the Model

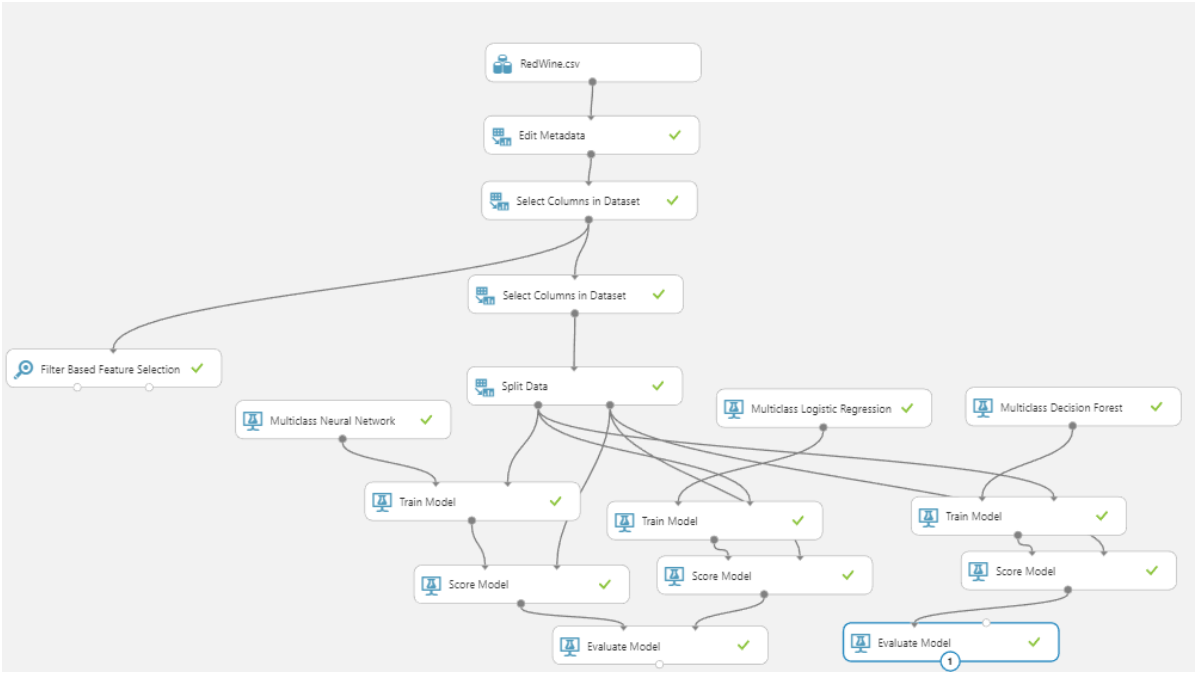


Figure 10 Azure ML Studio Workflow Diagram for Red Wine

Complete Summary of Results

White Wine Dataset with 4898 samples			
Metrics	**Multiclass Neural Network**	**Multiclass Logistic Regression**	**Multiclass Decision Forest**
Overall accuracy	0.752893	0.748809	0.808713
Average accuracy	0.835262	0.832539	0.872476
Micro-averaged precision	0.752893	0.748809	0.808713
Macro-averaged precision	NaN	NaN	0.726646
Micro-averaged recall	0.752893	0.748809	0.808713
Macro-averaged recall	0.404814	0.400171	0.572695

Red Wine Dataset with 1599 samples			
Metrics	**Multiclass Neural Network**	**Multiclass Logistic Regression**	**Multiclass Decision Forest**
Overall accuracy	0.845833	0.84375	0.858333
Average accuracy	0.897222	0.895833	0.905556
Micro-averaged precision	0.845833	0.84375	0.858333
Macro-averaged precision	NaN	NaN	0.573773
Micro-averaged recall	0.845833	0.84375	0.858333
Macro-averaged recall	0.402337	0.388371	0.504738