

Kennesaw State University
IS 8935 Business Intelligence: Traditional & Big Data Analytics
Dr. Reza Vaezi
Assignment 8
March 12, 2019

Decision Tree Analysis for Wine Quality

By Karis Kim

Executive Summary

The goal of this report is to provide insight for an upscale restaurant manager to predict wine quality based on a number of attributes affecting wine quality. Based on a decision tree analysis, the best attributes are total sulfur dioxide, fixed acidity, chlorides, and sulphates. The model predicted 3 good to 9 excellent wines. However, the training data set does not contain enough examples in the good and excellent category to yield a sound prediction, so the report recommends acquiring more relevant training data for better prediction.

CONTENTS

Contents	2
Business Understanding	3
Data Understanding.....	3
Attribute Information	3
Data Preparation	5
Modeling.....	5
Steps	5
Results	5
Evaluation OF Findings	10
Busines Recommendations	11
References	11

BUSINESS UNDERSTANDING

An upscale restaurant is interested in identifying wines of good and excellent quality based on a number of attributes. The goal of this report is to provide a list of wines that are predicted to be of good and excellent quality.

DATA UNDERSTANDING

The wine quality data set is divided into WineQuality_Training data set and WineQuality_Scoring data set for decision tree prediction analysis. The WineQuality_Training data set contains 540 examples and 12 attributes, while the WineQuality_Scoring data set contains 1,059 examples and all attributes except the *quality* attribute, which will be the label or predictor attribute.

ATTRIBUTE INFORMATION

	Attribute	Description [1]
1	quality (training set)	wine quality (mehh, medium, good, excellent)
	wineID (scoring set)	ID number for each wine in scoring data
2	fixed acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
3	volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
4	citric acid	found in small quantities, citric acid can add 'freshness' and flavor to wines
5	residual sugar	the amount of sugar remaining after fermentation stops; greater than 45 grams/liter are considered sweet
6	chlorides	the amount of salt in the wine
7	free sulfur dioxide	the free form of SO ₂ exists in equilibrium between molecular SO ₂ (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
8	total sulfur dioxide	amount of free and bound forms of SO ₂ ; in low concentrations, mostly undetectable in wine, but at 50 ppm, SO ₂ becomes evident in the nose and taste of wine
9	density	the density of water is close to that of water depending on the percent alcohol and sugar content
10	pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
11	sulphates	a wine additive which can contribute to sulfur dioxide gas (SO ₂) levels, which acts as an antimicrobial and antioxidant
12	alcohol	the percent alcohol content of the wine

Figure 1. Descriptive statistics of Training data set

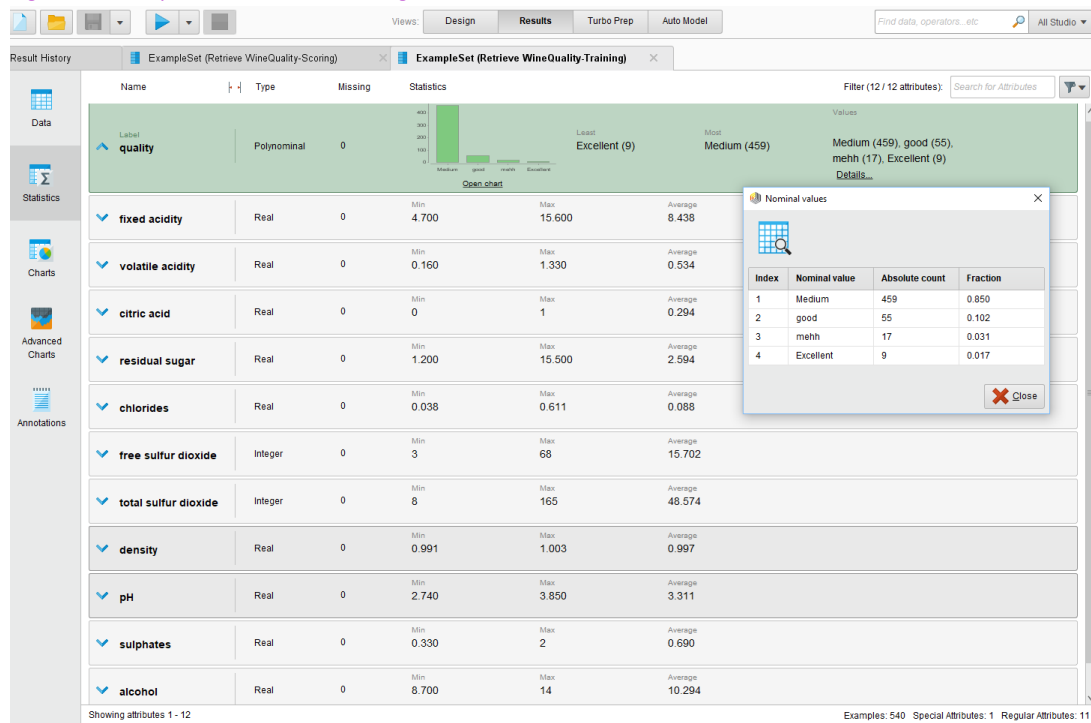


Figure 2. Descriptive statistics of Scoring data set

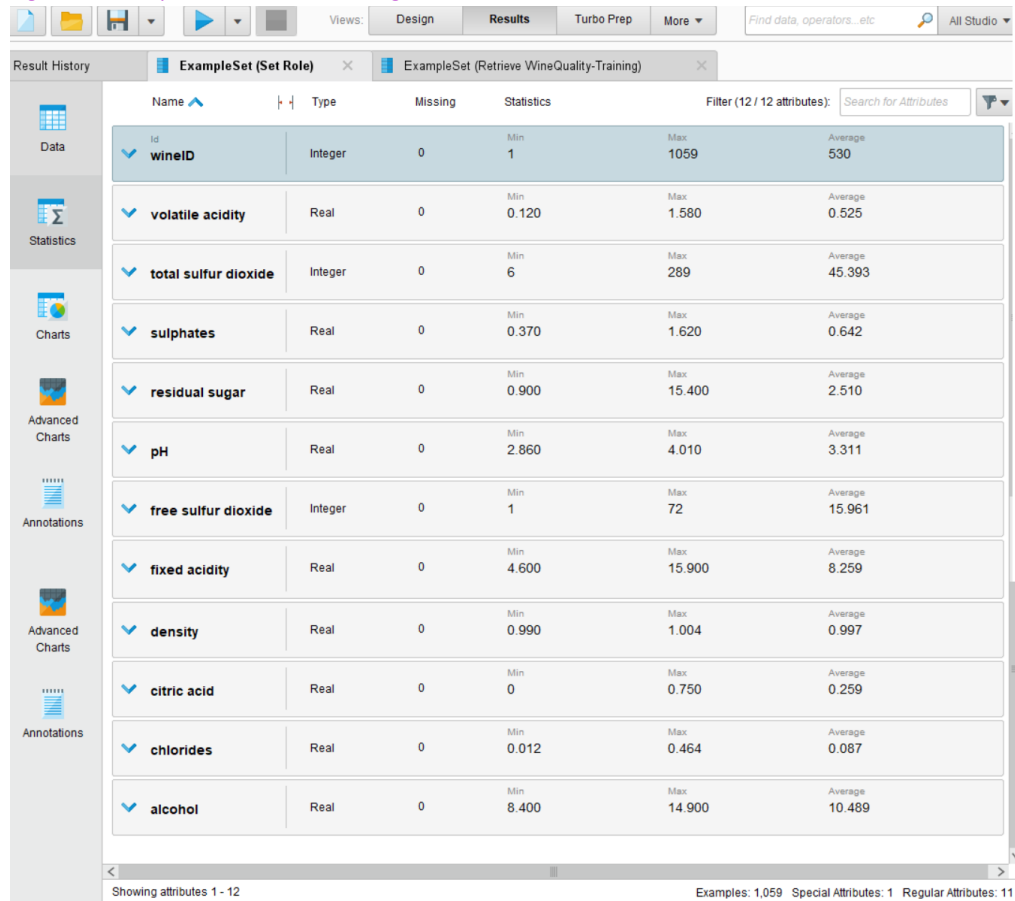


Figure 3. Power BI view of training data

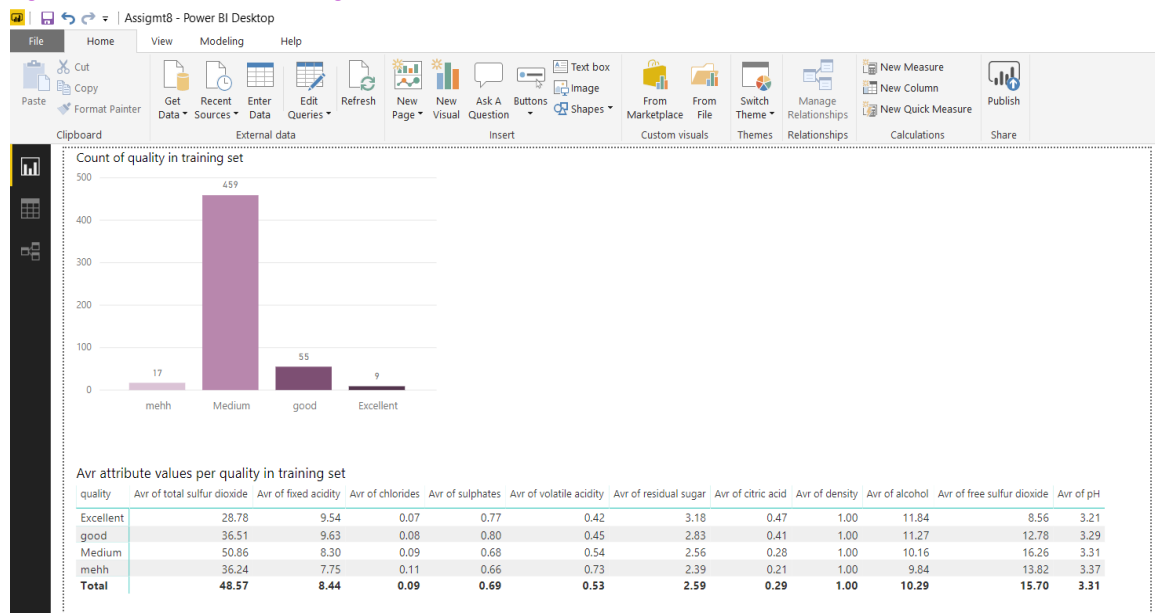


Figure 3 visualization of the training data shows that 85% of the examples (459 examples) in training are of medium quality and only 1.67% (9 examples) are of excellent quality and only 10.19% (55 examples) are of good quality. The matrix table in Figure 3 shows the average value per attribute for each of the 4 qualities in addition to the overall/total average value per attribute.

DATA PREPARATION

Data Type Transformation: At data import of training set, the *quality* attribute was designated as the label (see figure 1). No transformation to data types was needed. At data import of the scoring set, the *wineID* attribute was designated as ID using the Set Role operator (see figure 4), and no transformation of data types was needed.

Data Preparation of Missing Values: No missing values were found in the data set, and decision trees are not sensitive to missing values.

Decision tree involves minimal data preparation because it is not sensitive to missing values or outliers.

MODELING

The decision tree modeling process will involve running the decision tree operator on the training set, then running the Apply Model operator by connecting the model to the scoring set and iterate adjusting the parameters until an optimal decision tree results.

STEPS

First, the Decision Tree operator is added to the training set with parameters at default setting.

Second, the Apply Model operator is added after the Decision Tree operator and unlabeled examples from the scoring set are connected to apply the model to the scoring set.

Next, the process can be run repeatedly at different parameters to generate the decision tree (see figure 4, 7, 10).

RESULTS

The following figures show three iterations of the decision tree model runs with parameters adjusted.

Figure 4. RapidMiner decision tree process with parameter for first iteration

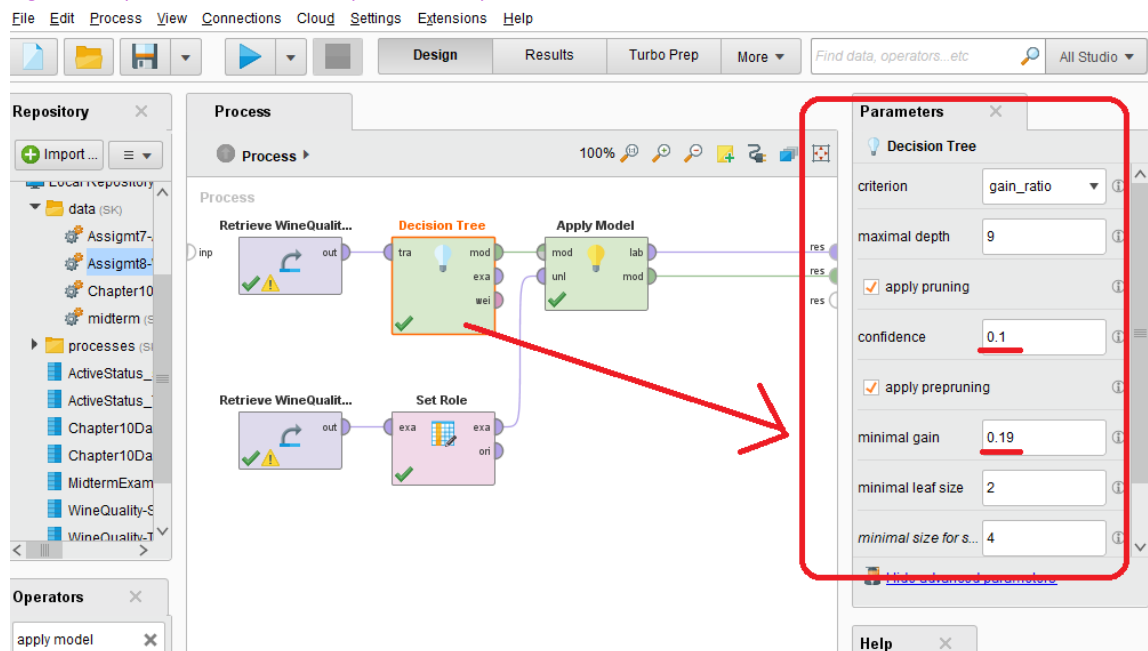


Figure 5. Decision Tree showing the best predictor at top for first iteration

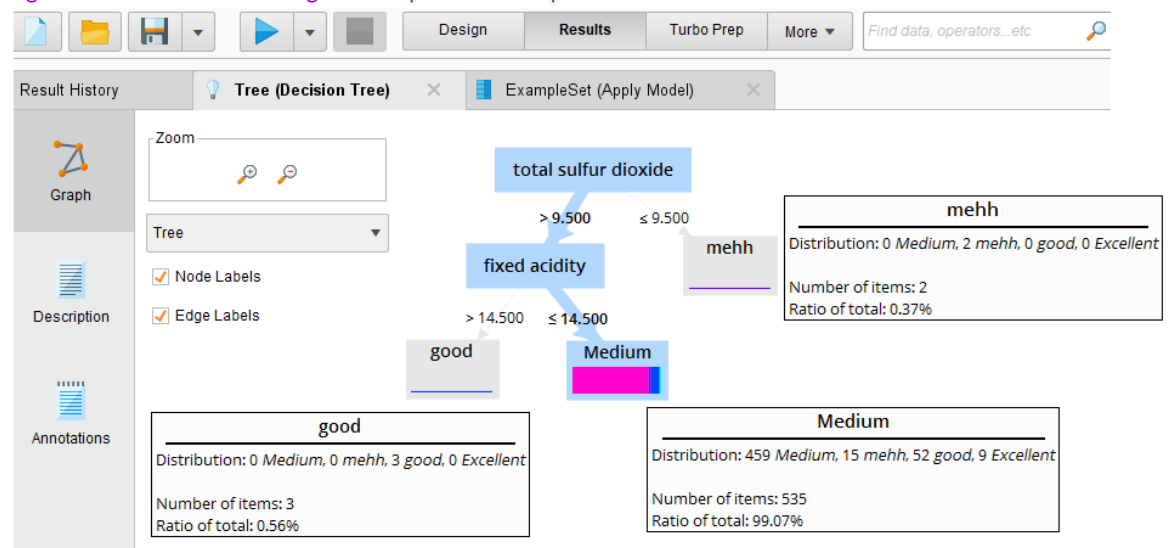


Figure 6. Results showing prediction and confidence for first iteration

Row...	wineID	prediction(quality)	confidence(Medium)	confidence(mehh)	confidence(good)	confidence(Excellent)	fixed acidity	volatile acidity	cit
1	1	Medium	0.858	0.028	0.097	0.017	7.400	0.700	0
2	2	Medium	0.858	0.028	0.097	0.017	7.800	0.880	0
3	3	Medium	0.858	0.028	0.097	0.017	7.800	0.760	0.0
4	4	Medium	0.858	0.028	0.097	0.017	11.200	0.280	0.5
5	5	Medium	0.858	0.028	0.097	0.017	7.400	0.700	0
6	6	Medium	0.858	0.028	0.097	0.017	7.400	0.660	0
7	7	Medium	0.858	0.028	0.097	0.017	7.900	0.600	0.0
8	8	Medium	0.858	0.028	0.097	0.017	7.300	0.650	0
9	9	Medium	0.858	0.028	0.097	0.017	7.800	0.580	0.0
10	10	Medium	0.858	0.028	0.097	0.017	7.500	0.500	0.3
11	11	Medium	0.858	0.028	0.097	0.017	6.700	0.580	0.0
12	12	Medium	0.858	0.028	0.097	0.017	7.500	0.500	0.3
13	13	Medium	0.858	0.028	0.097	0.017	5.600	0.615	0
14	14	Medium	0.858	0.028	0.097	0.017	7.800	0.610	0.2
15	15	Medium	0.858	0.028	0.097	0.017	8.900	0.620	0.1
16	16	Medium	0.858	0.028	0.097	0.017	8.900	0.620	0.1
17	17	Medium	0.858	0.028	0.097	0.017	8.500	0.280	0.5
18	18	Medium	0.858	0.028	0.097	0.017	8.100	0.560	0.2
19	19	Medium	0.858	0.028	0.097	0.017	7.400	0.590	0.0
20	20	Medium	0.858	0.028	0.097	0.017	7.900	0.320	0.5
21	21	Medium	0.858	0.028	0.097	0.017	8.900	0.220	0.4
22	22	Medium	0.858	0.028	0.097	0.017	7.600	0.390	0.3
23	23	Medium	0.858	0.028	0.097	0.017	7.900	0.430	0.2

ExampleSet (1,059 examples, 6 special attributes, 11 regular attributes)

Figure 7. RapidMiner decision tree process with new parameter for second iteration

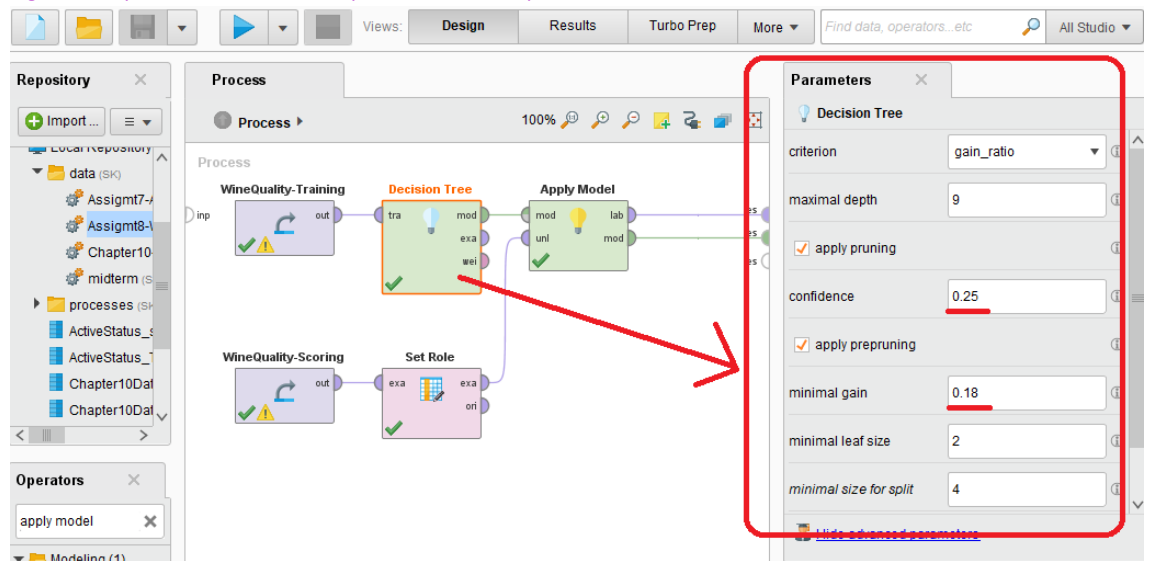


Figure 8. Decision Tree showing the best predictor at top for second iteration

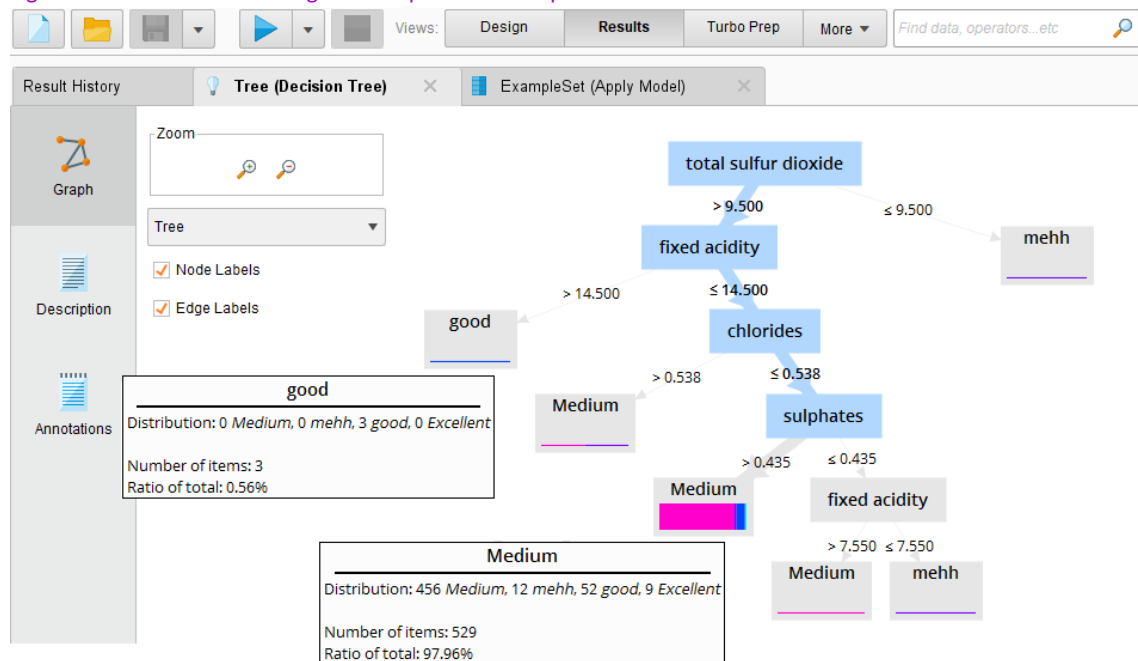


Figure 9. Results showing prediction and confidence for second iteration

Row ...	wineID	prediction(quality)	confidence(Medium)	confidence(mehh)	confidence(good)	confidence(Excellent)	fixed acidity	volatile acidity	ci
9	9	Medium	0.862	0.023	0.098	0.017	7.800	0.580	0.0
10	10	Medium	0.862	0.023	0.098	0.017	7.500	0.500	0.0
11	11	Medium	0.862	0.023	0.098	0.017	6.700	0.580	0.0
12	12	Medium	0.862	0.023	0.098	0.017	7.500	0.500	0.0
13	13	Medium	0.862	0.023	0.098	0.017	5.600	0.615	0.0
14	14	Medium	0.862	0.023	0.098	0.017	7.800	0.610	0.0
15	15	Medium	0.862	0.023	0.098	0.017	8.900	0.620	0.0
16	16	Medium	0.862	0.023	0.098	0.017	8.900	0.620	0.0
17	17	Medium	0.862	0.023	0.098	0.017	8.500	0.280	0.0
18	18	Medium	0.862	0.023	0.098	0.017	8.100	0.560	0.0
19	19	Medium	0.862	0.023	0.098	0.017	7.400	0.590	0.0
20	20	Medium	0.862	0.023	0.098	0.017	7.900	0.320	0.0
21	21	Medium	0.862	0.023	0.098	0.017	8.900	0.220	0.0
22	22	Medium	0.862	0.023	0.098	0.017	7.600	0.390	0.0
23	23	Medium	0.862	0.023	0.098	0.017	7.900	0.430	0.0
24	24	Medium	0.862	0.023	0.098	0.017	8.500	0.490	0.0
25	25	Medium	0.862	0.023	0.098	0.017	6.900	0.400	0.0
26	26	Medium	0.862	0.023	0.098	0.017	6.300	0.390	0.0
27	27	Medium	0.862	0.023	0.098	0.017	7.600	0.410	0.0
28	28	Medium	0.862	0.023	0.098	0.017	7.900	0.430	0.0
29	29	Medium	0.862	0.023	0.098	0.017	7.100	0.710	0.0
30	30	Medium	0.862	0.023	0.098	0.017	7.800	0.645	0.0
31	31	Medium	0.862	0.023	0.098	0.017	6.700	0.675	0.0

ExampleSet (1,059 examples, 6 special attributes, 11 regular attributes)

Figure 10. RapidMiner decision tree process with new parameter for third iteration

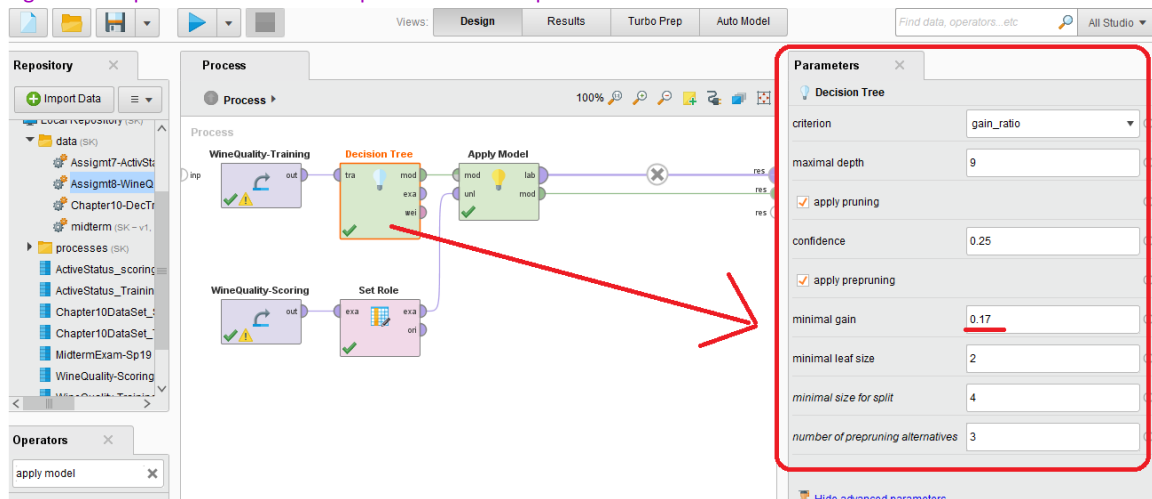


Figure 11. Decision Tree showing the best predictor at top for third iteration

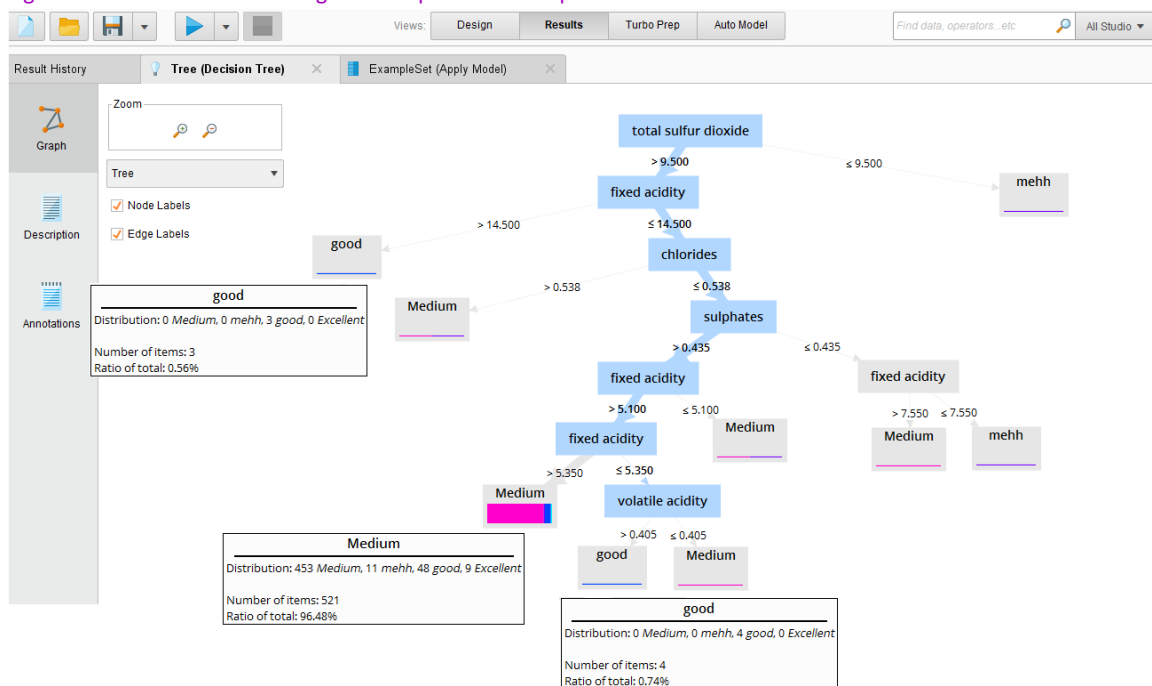
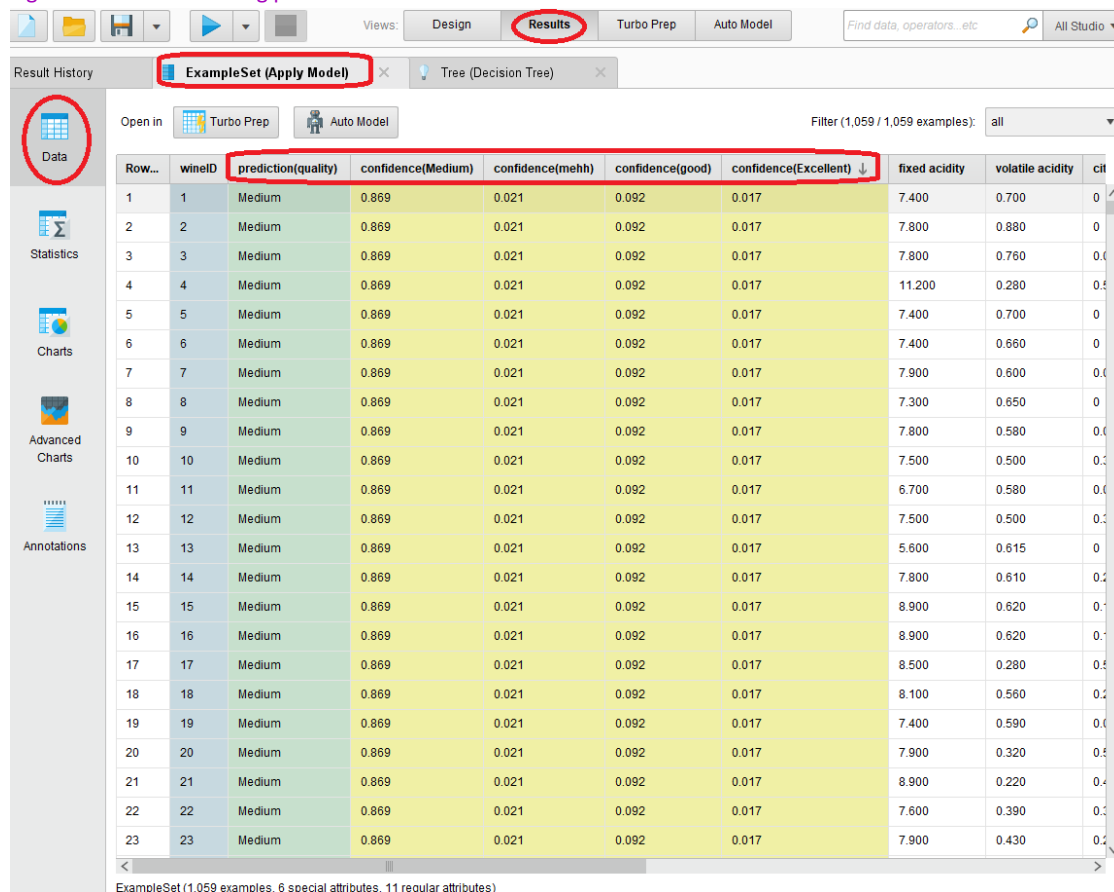


Figure 12. Results showing prediction and confidence for third iteration



Row...	wineID	prediction(quality)	confidence(Medium)	confidence(mehh)	confidence(good)	confidence(Excellent)	fixed acidity	volatile acidity	cit
1	1	Medium	0.869	0.021	0.092	0.017	7.400	0.700	0.400
2	2	Medium	0.869	0.021	0.092	0.017	7.800	0.880	0.300
3	3	Medium	0.869	0.021	0.092	0.017	7.800	0.760	0.400
4	4	Medium	0.869	0.021	0.092	0.017	11.200	0.280	0.500
5	5	Medium	0.869	0.021	0.092	0.017	7.400	0.700	0.400
6	6	Medium	0.869	0.021	0.092	0.017	7.400	0.660	0.400
7	7	Medium	0.869	0.021	0.092	0.017	7.900	0.600	0.400
8	8	Medium	0.869	0.021	0.092	0.017	7.300	0.650	0.400
9	9	Medium	0.869	0.021	0.092	0.017	7.800	0.580	0.400
10	10	Medium	0.869	0.021	0.092	0.017	7.500	0.500	0.300
11	11	Medium	0.869	0.021	0.092	0.017	6.700	0.580	0.400
12	12	Medium	0.869	0.021	0.092	0.017	7.500	0.500	0.300
13	13	Medium	0.869	0.021	0.092	0.017	5.600	0.615	0.400
14	14	Medium	0.869	0.021	0.092	0.017	7.800	0.610	0.200
15	15	Medium	0.869	0.021	0.092	0.017	8.900	0.620	0.100
16	16	Medium	0.869	0.021	0.092	0.017	8.900	0.620	0.100
17	17	Medium	0.869	0.021	0.092	0.017	8.500	0.280	0.500
18	18	Medium	0.869	0.021	0.092	0.017	8.100	0.560	0.200
19	19	Medium	0.869	0.021	0.092	0.017	7.400	0.590	0.400
20	20	Medium	0.869	0.021	0.092	0.017	7.900	0.320	0.500
21	21	Medium	0.869	0.021	0.092	0.017	8.900	0.220	0.400
22	22	Medium	0.869	0.021	0.092	0.017	7.600	0.390	0.300
23	23	Medium	0.869	0.021	0.092	0.017	7.900	0.430	0.200

EVALUATION OF FINDINGS

In the first iteration (figure 4, 5, 6), confidence was set to 0.1 and minimal gain at 0.19. This yielded a simple decision tree with the best predictor being total sulfur dioxide, followed by fixed acidity. When total sulfur dioxide was greater than 9.5, and fixed acidity was greater than 14.5, there were 3 examples predicted to be of good quality, and when fixed acidity was less than or equal to 14.5, there were 52 good and 9 excellent predicted.

In the second iteration (figure 7, 8, 9), confidence was raised to 0.25 and minimal gain to 0.18. So, this yielded a decision tree with more nodes than the first iteration, with total sulfur dioxide still being the top/best predictor, followed by fixed acidity again, then chlorides and sulfates. When total sulfur dioxide is greater than 9.5 and fixed acidity is less than or equal to 14.5 and chlorides is less than or equal to 0.538 and sulfates is greater than 0.435, the model predicted 52 good and 9 excellent, in addition to the 3 good predicted from the first two attributes. In sum, the first iteration and second iteration predicted the same number of good and excellent wines, though there were more nodes in the second iteration.

In the third iteration (figure 10, 11, 12), confidence remained at 0.25 and minimal gain was set to 0.17. This yielded a decision tree with more nodes; total sulfur dioxide was the top/best predictor, followed by fixed acidity, then chlorides and sulfates, and then volatile acidity at the end. While the third iteration brought into account volatile acidity, which may be an important factor affecting wine quality, since levels too high can lead to an unpleasant vinegar taste, the tree includes fixed acidity in 3 levels and can lead to confusion. However, this

iteration predicted a total of 55 good wines in addition to the 9 excellent, which is 3 more than the other two iterations.

Figure 12. Power BI list of wine predictions

wineID	First prediction(quality)	confidence(Excellent)	confidence(good)	confidence(Medium)
142	good	0.00	1.00	0.00
143	good	0.00	1.00	0.00
145	good	0.00	1.00	0.00
240	good	0.00	1.00	0.00
1	Medium	0.02	0.10	0.86
2	Medium	0.02	0.10	0.86
3	Medium	0.02	0.10	0.86
4	Medium	0.02	0.10	0.86
5	Medium	0.02	0.10	0.86
6	Medium	0.02	0.10	0.86
7	Medium	0.02	0.10	0.86
8	Medium	0.02	0.10	0.86
9	Medium	0.02	0.10	0.86
10	Medium	0.02	0.10	0.86
11	Medium	0.02	0.10	0.86
12	Medium	0.02	0.10	0.86
13	Medium	0.02	0.10	0.86
14	Medium	0.02	0.10	0.86
15	Medium	0.02	0.10	0.86
16	Medium	0.02	0.10	0.86
17	Medium	0.02	0.10	0.86
18	Medium	0.02	0.10	0.86
19	Medium	0.02	0.10	0.86
20	Medium	0.02	0.10	0.86
21	Medium	0.02	0.10	0.86
22	Medium	0.02	0.10	0.86
23	Medium	0.02	0.10	0.86
24	Medium	0.02	0.10	0.86
25	Medium	0.02	0.10	0.86
26	Medium	0.02	0.10	0.86
27	Medium	0.02	0.10	0.86
28	Medium	0.02	0.10	0.86
29	Medium	0.02	0.10	0.86
30	Medium	0.02	0.10	0.86
31	Medium	0.02	0.10	0.86
32	Medium	0.02	0.10	0.86
33	Medium	0.02	0.10	0.86
34	Medium	0.02	0.10	0.86
35	Medium	0.02	0.10	0.86
36	Medium	0.02	0.10	0.86
37	Medium	0.02	0.10	0.86
38	Medium	0.02	0.10	0.86
39	Medium	0.02	0.10	0.86

BUSINES RECOMMENDATIONS

Evaluation of findings suggest that the most optimal decision tree model has total sulfur dioxide → fixed acidity → chlorides → sulphates as the best predictor attributes to consider. A list of possible options has been provided in the preceding section. However, the training data contains not enough examples of good and excellent wines and an overwhelming majority of examples in the medium quality, so this is a shortcoming in accurately predicting the wine quality in the scoring data set. This report recommends that another training data set with more examples in the good and excellent quality and/or a more balanced training data set be obtained to rerun the model for prediction.

REFERENCES

1. UCI. (2017, November 27). Red Wine Quality. Retrieved from <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/version/2#winequality-red.csv>