Wine Quality Prediction

Supervisor : Dr. Milton King

By:
Kranthika Gurram - 202005785
Sai Teja Reddy - 202106549
Hemanth Chekka - 202006238
Siva Sai Kaliki - 202004650

Introduction

- ➔ Wine is made after fermenting grape juice and some other fruit juices with less alcohol content.
- ➔ Wine is generally tasted and judged if it tastes good or not. This process can be costly and not efficient.
- ➔ Our methods can predict the wine quality with an efficient process reducing both the cost and human interaction.

Dataset

- The dataset was downloaded from the UCI Machine Learning Repository.
- These datasets are related to red and white variants of the Portuguese "Vinho Verde" wine.
- Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

https://www.kaggle.com/datasets/rajyellow46/wine-quality?resource=download
For more information, read [Cortez et al., 2009].

Dataset

- Dataset was created using red and white wine samples.
- The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts).
- Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model
- Number of Attributes: 11 + output attribute
- Number of Instances: red wine - 1599; white wine - 4898.

https://www.kaggle.com/datasets/rajyellow46/wine-quality?resource=download
For more information, read [Cortez et al., 2009].

Dataset

	Type	fixed acidity	volatile acidi	citric acid	residual sug	chlorides	free sulfur di	total sulfur d	density	pH	sulphates	alcohol	quality
2	white	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
3	white	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
4	white	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
5	white	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
6	white	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
7	white	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
8	white	6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6
9	white	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
10	white	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
11	white	8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6
12	white	8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5
13	white	8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5
14	white	7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5
15	white	6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7
16	white	8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5
17	white	6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7
18	white	6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	6
19	white		0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
20	white	7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	6

https://www.kaggle.com/datasets/rajyellow46/wine-quality?resource=download

For more information, read [Cortez et al., 2009].

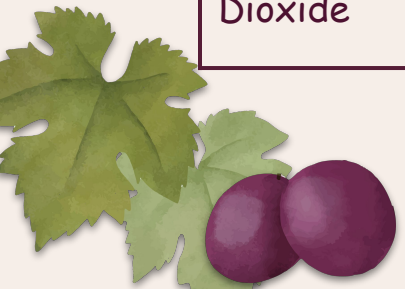
Interesting Findings

- ➔ These datasets can be viewed as classification or regression tasks.
- ➔ The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).
- ➔ Density of wine has strong negative correlation with Alcohol.
- ➔ Quality has considerable positive correlation with Alcohol.

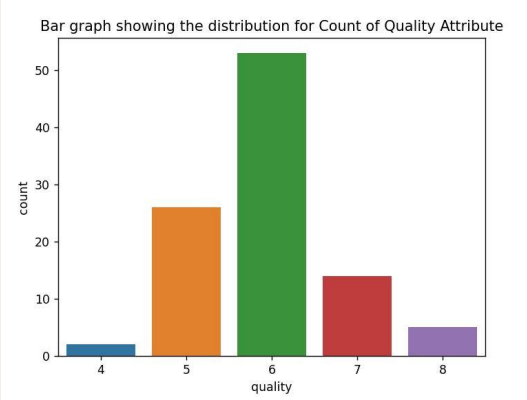
Interesting findings

On calculating the mean, median and mode values for the attributes in the dataset, following findings are discovered:

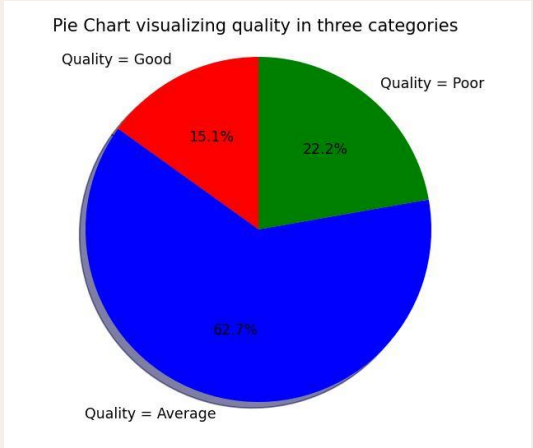
Attributes	Findings
Fixed Acidity, Volatile Acidity, Citric Acid, Chlorides, Density, pH, Sulphates, Alcohol, Quality	• It shows nearness between mean, median and mode values, indicating that these attributes have values that are almost uniformly distributed.
Residual Sugar, Free Sulphur Dioxide, Total Sulphur Dioxide	• It shows mean, median and mode values are distant from each other, indicating that these attributes have values that are asymmetrically distributed.



Interesting findings



➔ This bar graph provides information on count values distribution for the quality attribute. From this we can understand how many wines are associated for the specific quality value.



➔ This pie chart visualizes the quality attribute of dataset into three categories such as Poor, Average and Good.

Task

a) Description:

- ➔ Target Attribute: Quality
- ➔ Output: Categorical Category
- ➔ Task Type: Predictive and Classification

Task

b) Preprocessing:

- ➔ After deciding task, preprocessing step was done.
- ➔ It involves cleaning the dataset by removing inconsistencies.
- ➔ We have handled blank and null values in each column.
- ➔ For every column, missing values are replaced with respective mean value.
- ➔ After Preprocessing the dataset, model training is done.

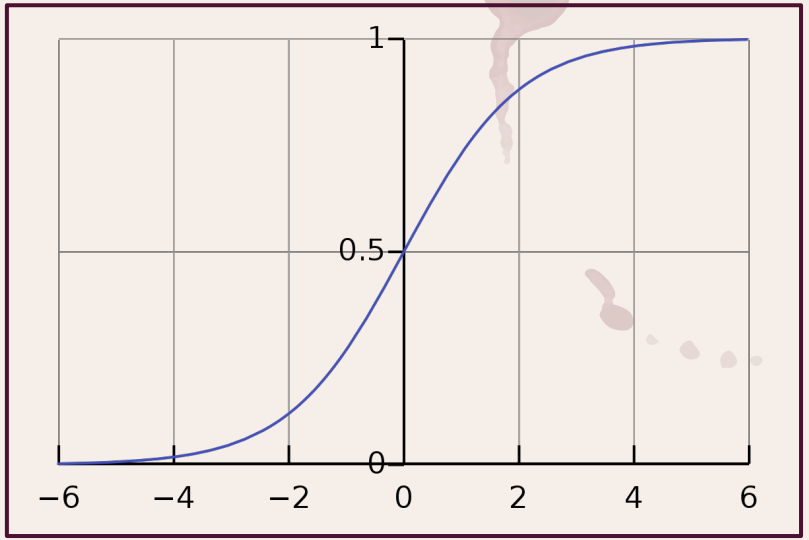
Task

c) Modelling:

- ➔ In this stage, we divided the dataset into two parts
- ➔ One part goes for Training and other for Testing.
- ➔ Around 80 to 85% of data is for training and 15 to 20% consists of testing data.
- ➔ Classifier Models are used for this classification task.
- ➔ After training, we tested the model for accuracy.
- ➔ In this task, we used Logistic Regression and Random Forest Classifiers.

Logistic Regression

- Logistic regression is made up on the logistic model in statistics.
- It models the probability of an event taking place.



Random Forest Classifier

- A random forest fits number of decision tree classifiers on various sub samples of the dataset.
- It averages the outputs and improves the predictive accuracy.
- It also controls overfitting.

Models vs Accuracy

Model	Accuracy Using (accuracy_score())
Random Forest Classifier	0.883846
Logistic Regression	0.823846

Applications

- These results can be used by consumers of wine to look at the bottle and make a selection.
- Results can be used by manufacturers to improve the quality of wine.
- Results can be used by dedicated authorities to control the quality of wine.

Thank You