STEP
fUNCTION

WINE QUALITY PREDICTION MODELLING

-Ankita Sarawgi

Introduction

- Relationship between the chemical properties of wine and its quality
- UCI machine learning repository
- 2 datasets; Red & White variants of Portuguese wine

  Number of instances; Red wine: 1599, White wine: 4898
- 11 Attributes + 1 output
- Output: Sensory data (0=very bad) (10=Excellent)

  Median of evaluations by wine experts

Scope And Limitation

- Usage: Oenologist wine tasting evaluation

  Improve production

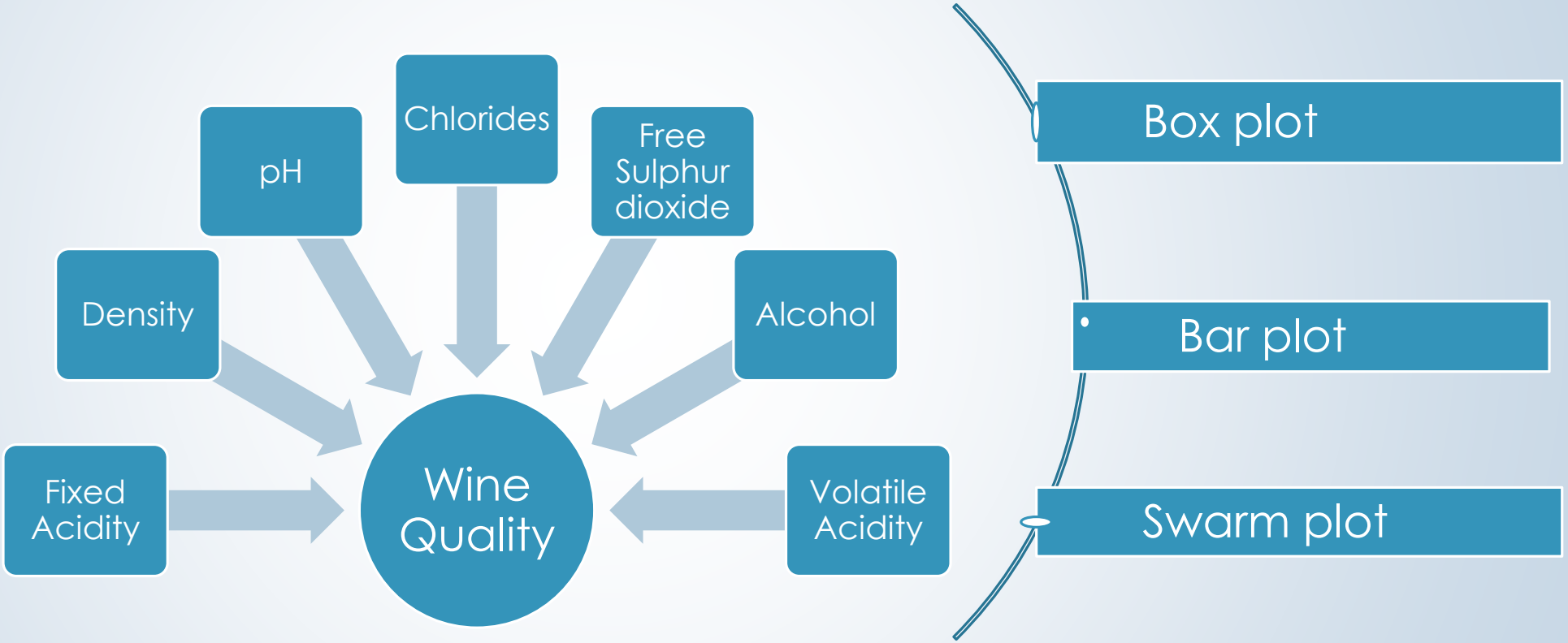
  Controlled pricing
- Opinions of different experts (Not a science)
- Approach: Project divided into two parts

  - Data exploration and preparation

  - Model setup and prediction

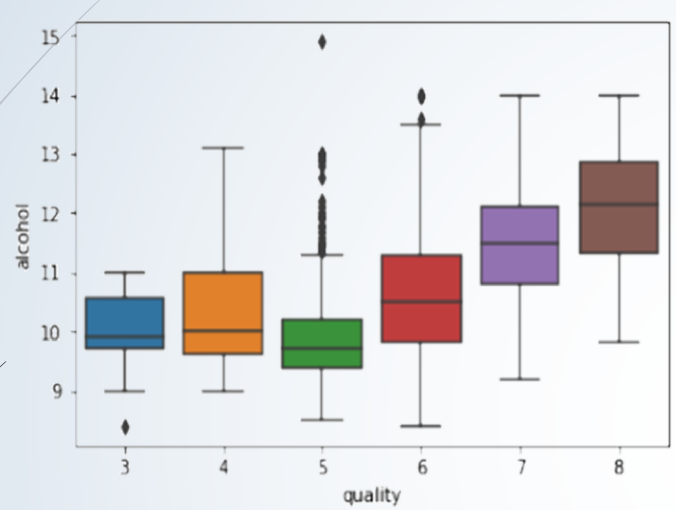
STEP
ƒUNCTION

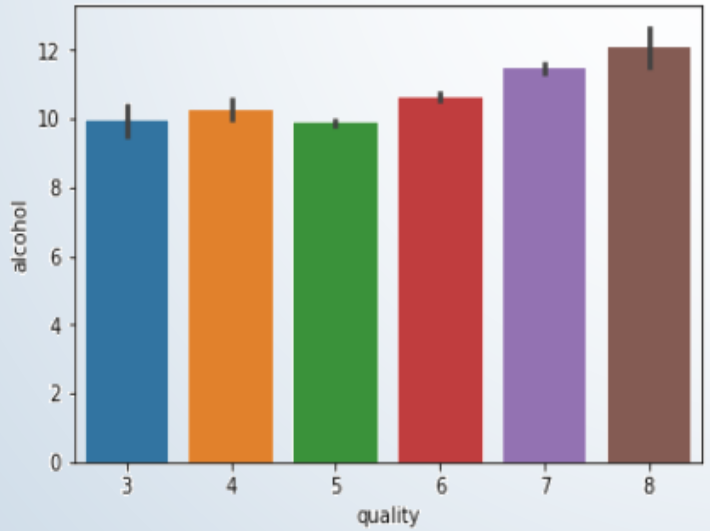
Data Exploration



Why the need for this?

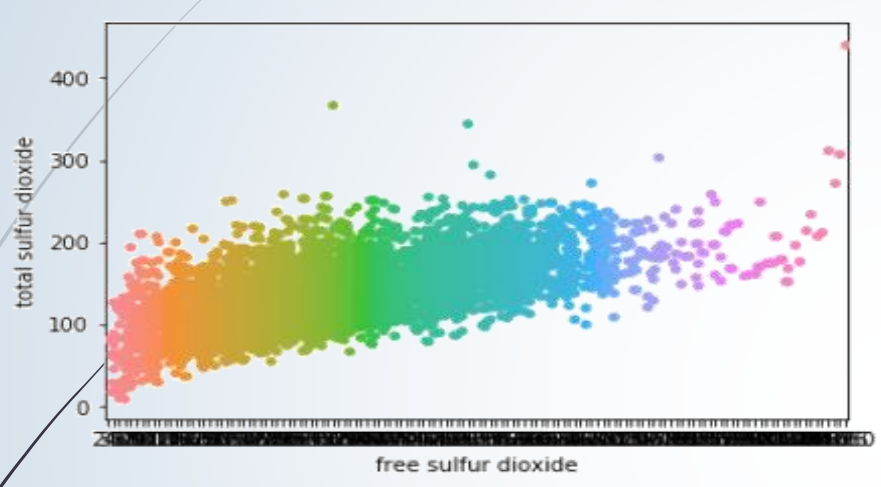
Analyzing relationship

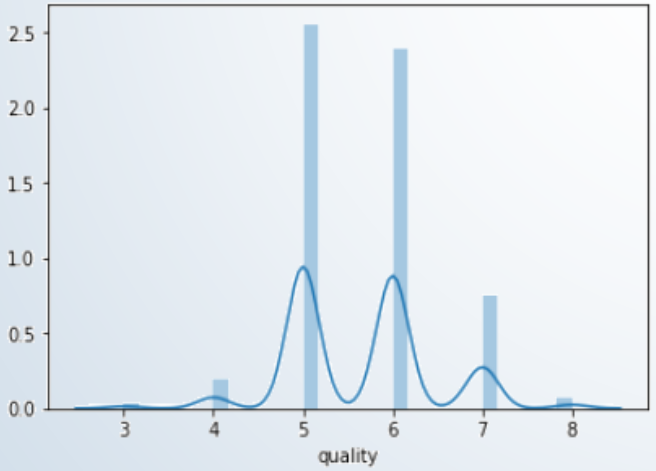




Features	Impact
Fixed Acidity	Medium(+ve)
Volatile Acidity	Medium(-ve)
Citric acid	Not significant
Residual Sugar	Not significant
Chlorides	Medium(-ve)
Free SO_2	Not significant
Total SO_2	Medium(-ve)
Density	Low(-ve)
pH	Low(+ve)
Sulphates	Not Significant
Alcohol	High(+ve)

Correlated Features





Correlation coefficient matrix

[[1 .62]
[.62 1]]

Datasets and Modelling

Training set (70%)

Validation set (15%)

Test Set (15%)

Predicting Values (Numerical)

- Linear Regression
- K Nearest Neighbors
- Decision Tree Regressor ✓
- Support Vector Regressor

Decision Tree Regressor

Sample set

Alcohol	9.8	9.8	9.8	9.8	9.8	9.9	9.9	9.9	9.9	9.9	10	10	10	10	10	10
Quality	5	5	5	5	5	5	6	6	6	6	6	5	7	7	6	7

Case 1

Alcohol	9.8	9.8	9.8	9.8	9.8	9.9	9.9	9.9	9.9	9.9	10	10	10	10	10	10

Case 2

SSE = 10

Predicted	5	5	5	5	5	5	5	5	5	5	7	7	7	7	7	7
Actual	5	5	5	5	5	5	6	6	6	6	6	5	7	7	6	7

SSE = 5

Predicted	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6
Actual	5	5	5	5	5	5	6	6	6	6	6	5	7	7	6	7

Decision Tree Continued....

- For each independent variable, multiple split points are selected
- Lowest yielding SSE split point/node is selected
- Similar process recursively continued
- A number of parameters to be considered : Best/random splitting strategies
  Max depth
  Max features
- Two parameters greatly affect results: **Min sample split** & **Min leaves**
- Considered 36 different scenarios to arrive at the parameter values that gives us the best results

Results

- Different evaluation methods exist for evaluation of Regression problems: MAE, RMSE, R^2, Accuracy percent
- Testing on the remaining data
- RMSE values on the test data

Red_wine:	**.4461**
New_red :	**.4457**
White_wine:	**.6431**
New_white:	**.6528**

Conclusion

- The low RMSE value suggests the models quite accurately predicts the wine quality score based on the chemical properties of wine
- Gives a fairly basic idea to assist in the production and pricing process