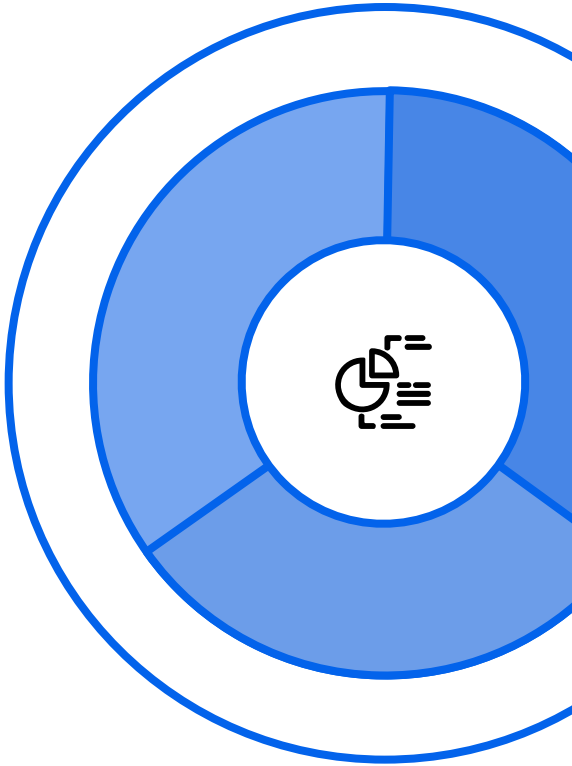
Wine Quality Classification Using Weka

Darshan Pathak

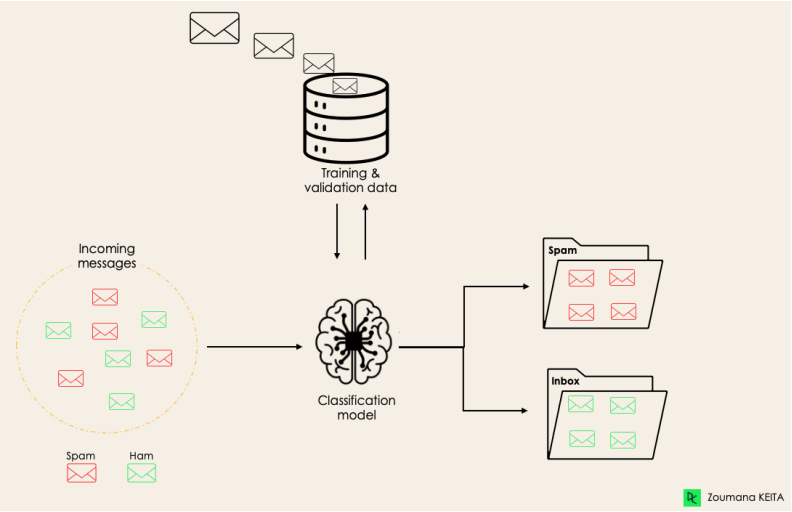


01

Classification

What is Classification

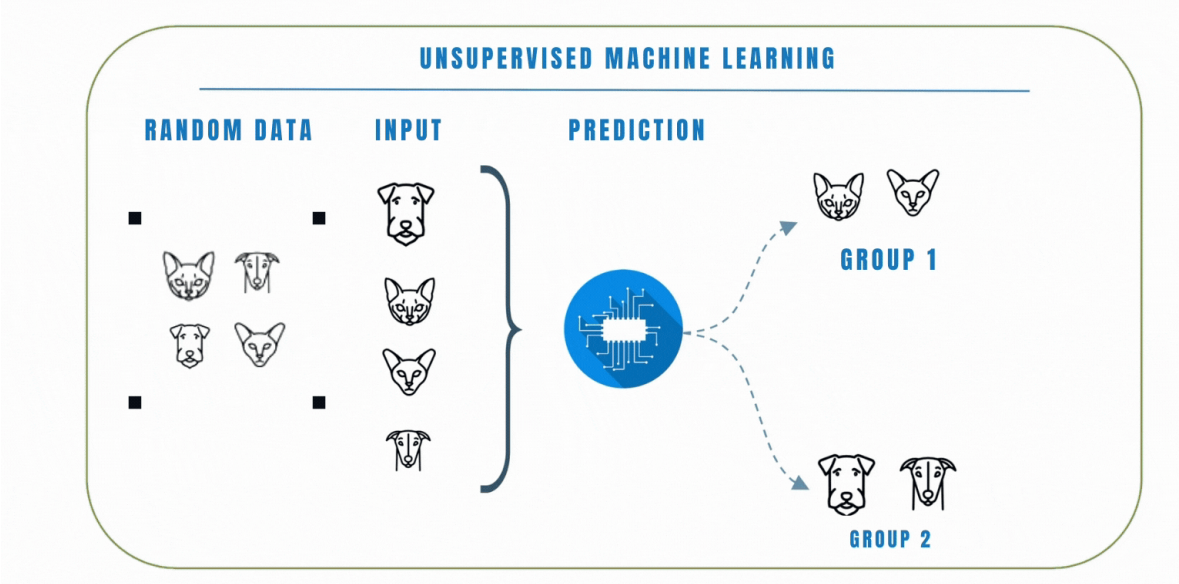
Classification in Machine Learning aims to determine which category an observation by understanding the relationship between the dependent and independent variables.



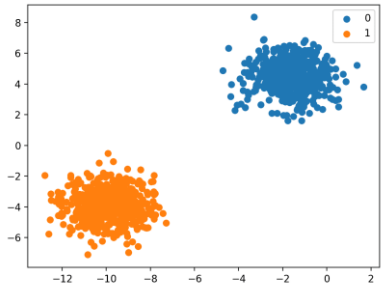
Classification algorithm can learn to predict whether a given email is spam or not span

Working of Classification Algorithms

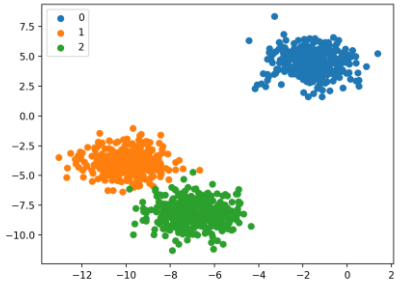
Classification algorithms sort data into predefined categories based on patterns they learn from labeled examples. They use features (data attributes) to create a model during training, and then apply this model to predict the classes of new, unseen data.



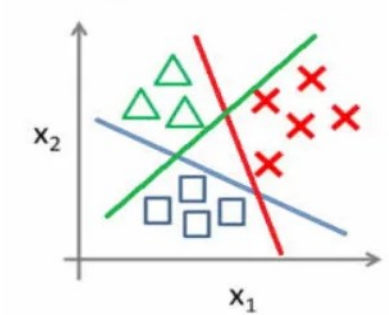
Types of Classification



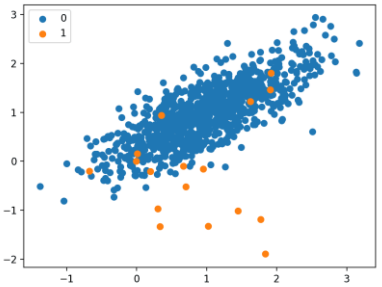
Binary Classification



Multi-Class Classification



Multi-Label Classification



Imbalanced Classification

02

Case Study

Let’s learn with help of

Wine Quality Case Study

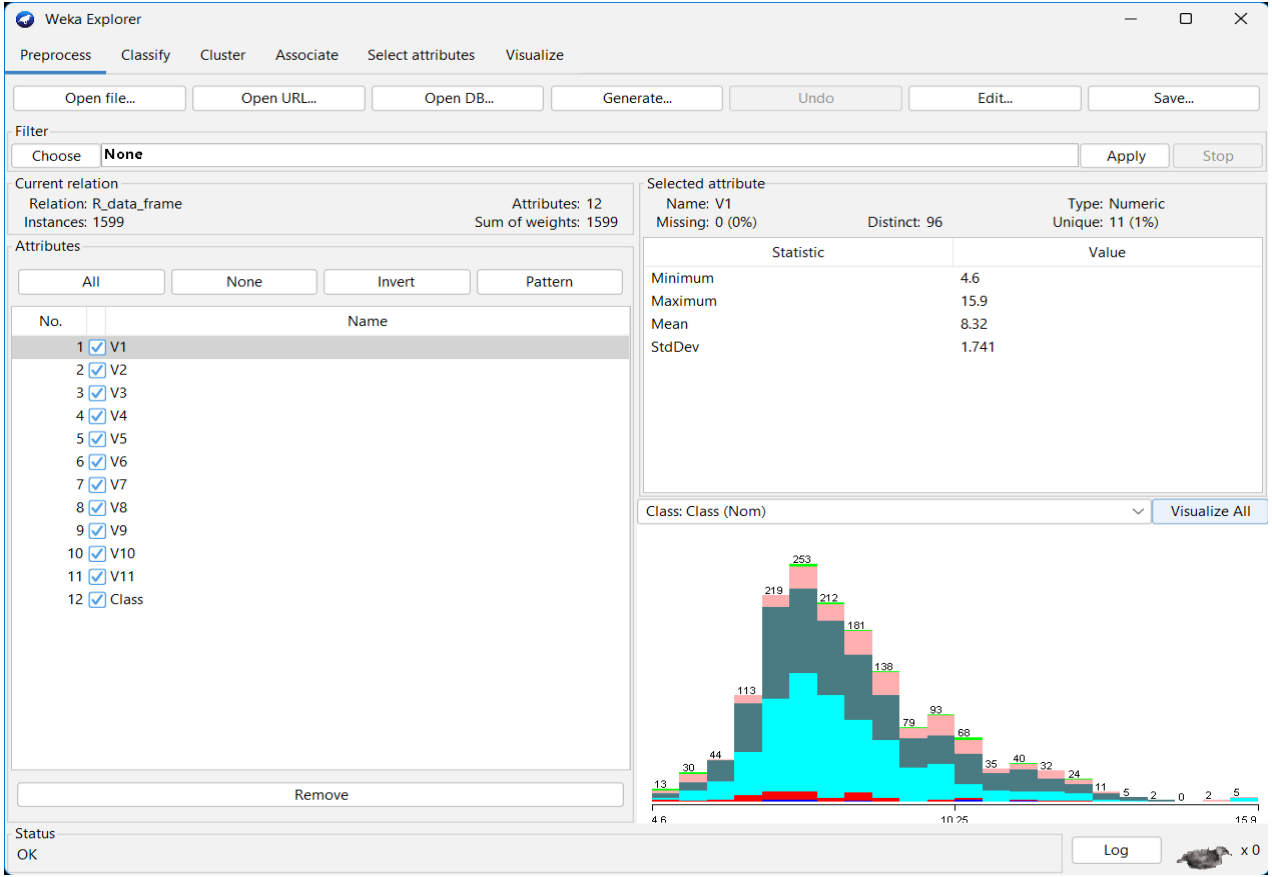
Using Weka Software

What is WEKA Software

- **W**aikato **E**nvironment for **K**nowledge **A**nalysis
- Collection of machine learning algorithms and data processing tools implemented in Java
- Used for the process of experimental data mining
  - Preparation of input data
  - Statistical evaluation of learning schemes
  - Visualization of input data and the result

1. Install Weka

2. Load Data

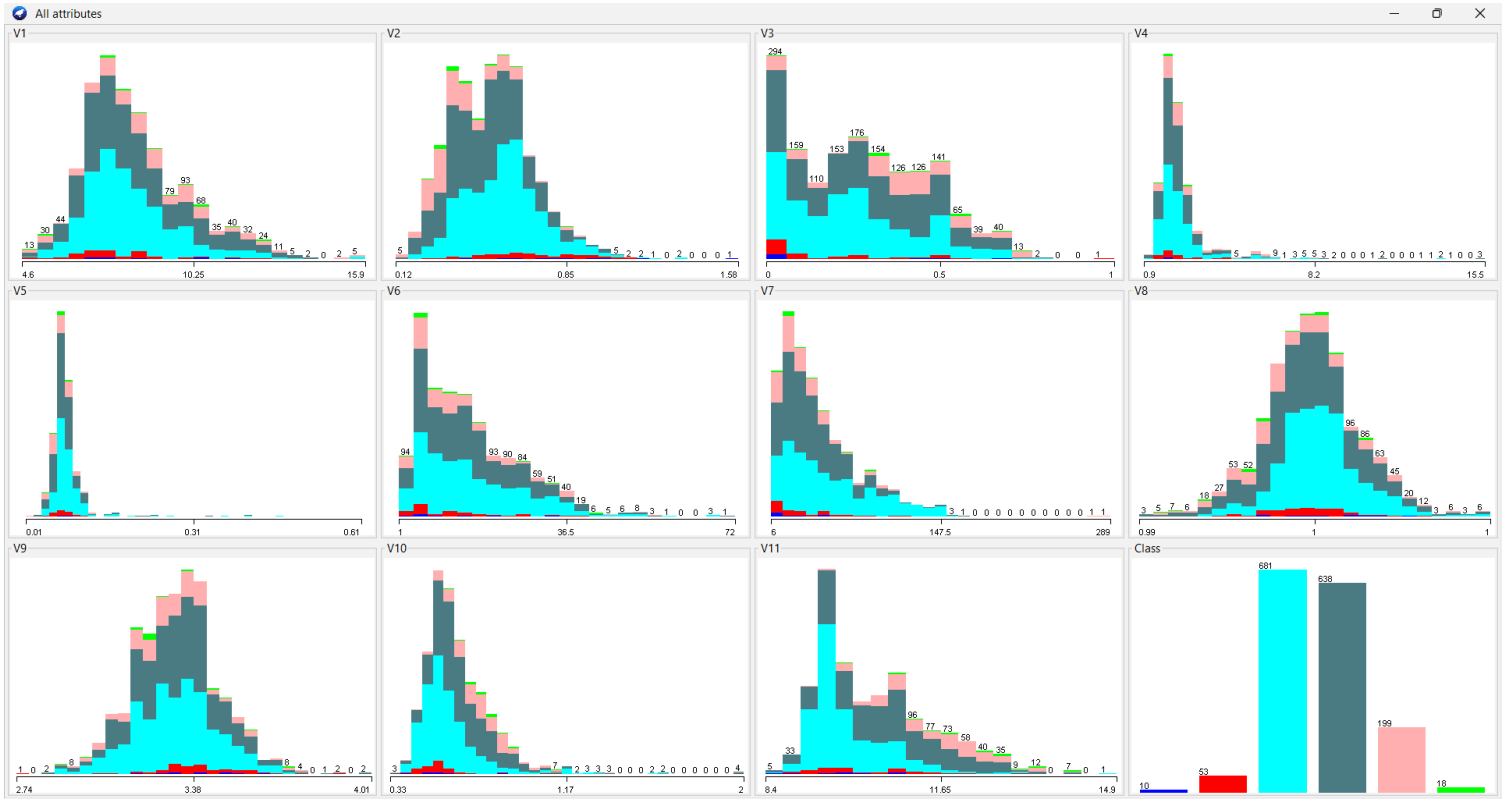

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: R_data_frame
Instances: 1599
Attributes: 12
Sum of weights: 1599
Attributes
All
None
Invert
Pattern
No.
Name
1 V1
2 V2
3 V3
4 V4
5 V5
6 V6
7 V7
8 V8
9 V9
10 V10
11 V11
12 Class
Remove
Selected attribute
Name: V1
Missing: 0 (0%)
Distinct: 96
Type: Numeric
Unique: 11 (1%)
Statistic
Value
Minimum 4.6
Maximum 15.9
Mean 8.32
StdDev 1.741
Class: Class (Nom)
Visualize All
253
219
212
181
138
113
93
79
68
44
40
35
32
30
24
13
11
5
5
2
2
0
4.6
10.25
15.9
Status
OK
Log
x 0

Wine Quality Dataset

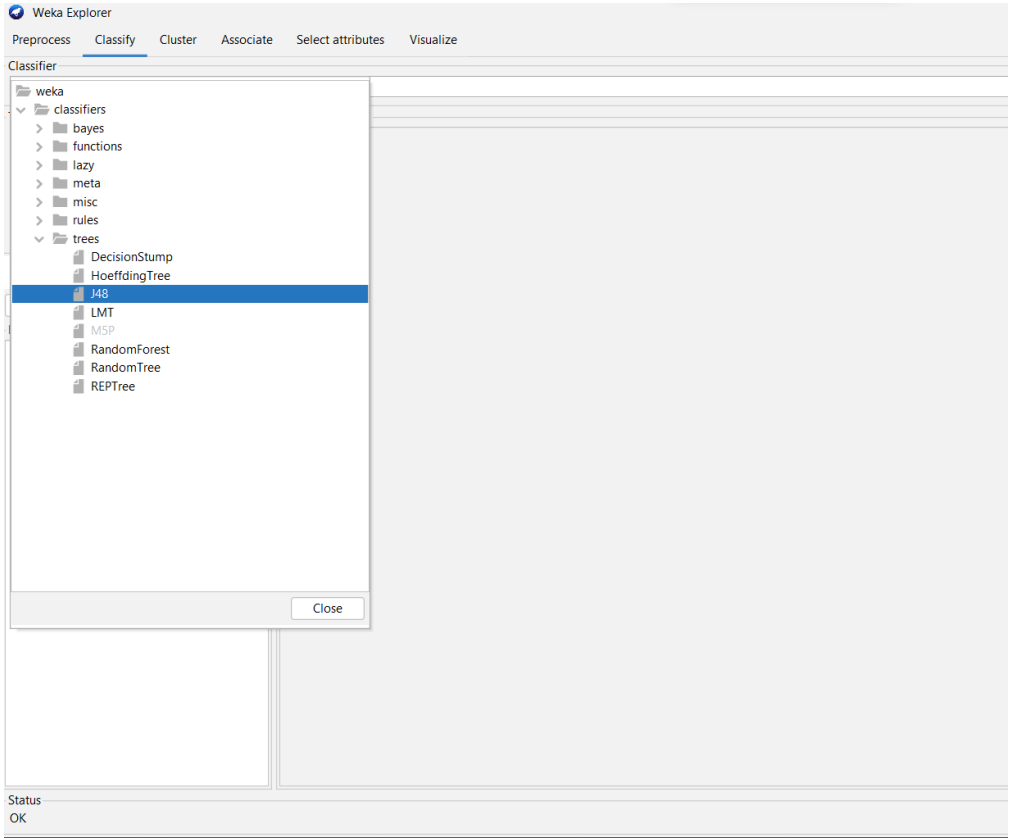
.: Wine Quality Dataset :.
****

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	Class
0	7.400000	0.700000	0.000000	1.900000	0.076000	11.000000	34.000000	0.997800	3.510000	0.560000	9.400000	b'3'
1	7.800000	0.880000	0.000000	2.600000	0.098000	25.000000	67.000000	0.996800	3.200000	0.680000	9.800000	b'3'
2	7.800000	0.760000	0.040000	2.300000	0.092000	15.000000	54.000000	0.997000	3.260000	0.650000	9.800000	b'3'
3	11.200000	0.280000	0.560000	1.900000	0.075000	17.000000	60.000000	0.998000	3.160000	0.580000	9.800000	b'4'
4	7.400000	0.700000	0.000000	1.900000	0.076000	11.000000	34.000000	0.997800	3.510000	0.560000	9.400000	b'3'
5	7.400000	0.660000	0.000000	1.800000	0.075000	13.000000	40.000000	0.997800	3.510000	0.560000	9.400000	b'3'
6	7.900000	0.600000	0.060000	1.600000	0.069000	15.000000	59.000000	0.996400	3.300000	0.460000	9.400000	b'3'
7	7.300000	0.650000	0.000000	1.200000	0.065000	15.000000	21.000000	0.994600	3.390000	0.470000	10.000000	b'5'
8	7.800000	0.580000	0.020000	2.000000	0.073000	9.000000	18.000000	0.996800	3.360000	0.570000	9.500000	b'5'
9	7.500000	0.500000	0.360000	6.100000	0.071000	17.000000	102.000000	0.997800	3.350000	0.800000	10.500000	b'3'

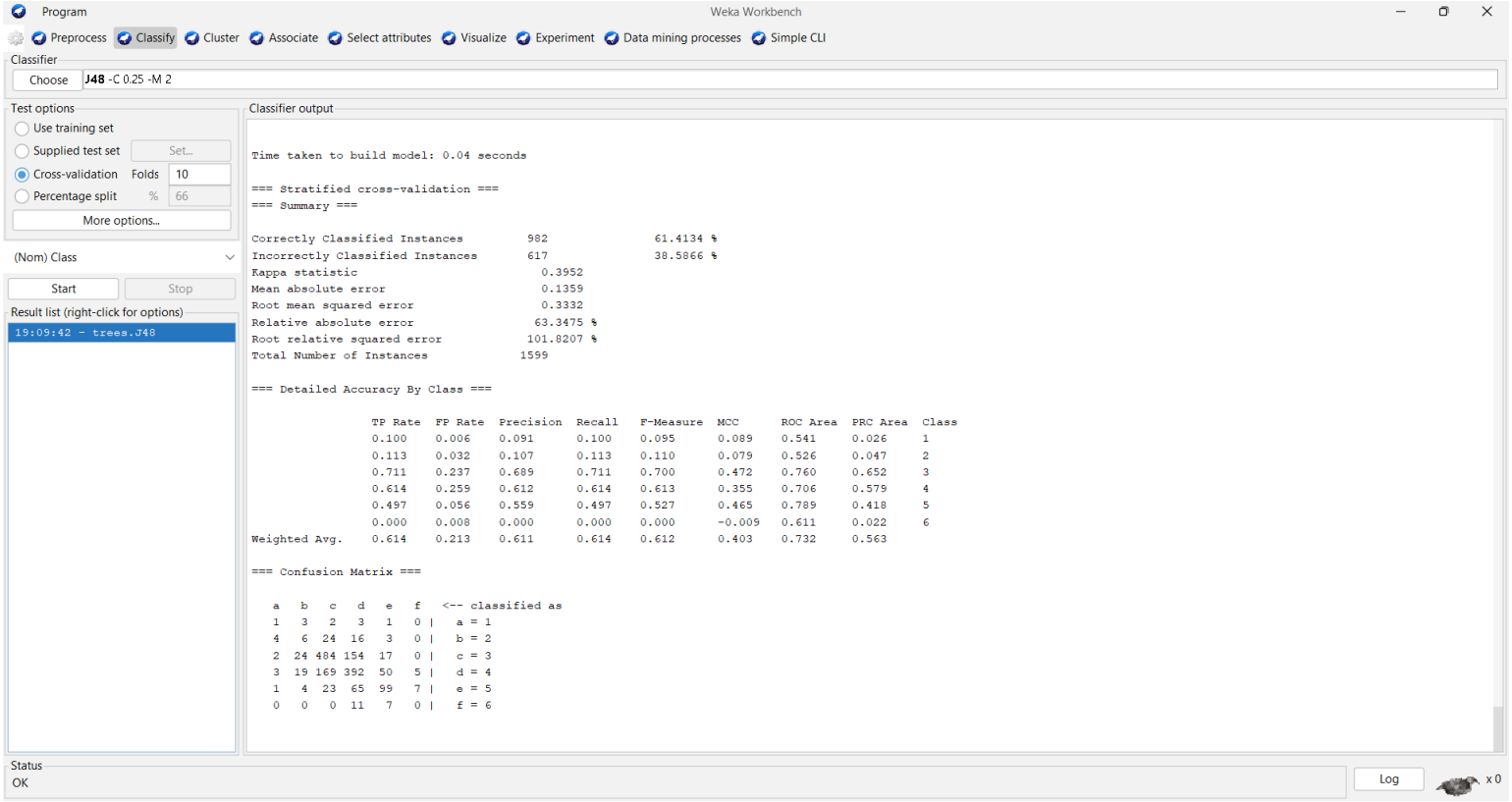
3. Visualize



4. Select J48 Classifier



5. Evaluate Result



Summary

Classification

Kappa statistic	0.3952
Mean absolute error	0.1359
Root mean squared error	0.3332
Relative absolute error	63.3475 %
Root relative squared error	101.8207 %
Total Number of Instances	1599

Corretly Classified

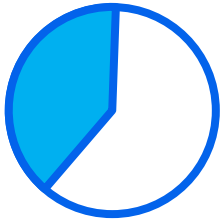
61.4%

982 Instances

Incorrectly Classified

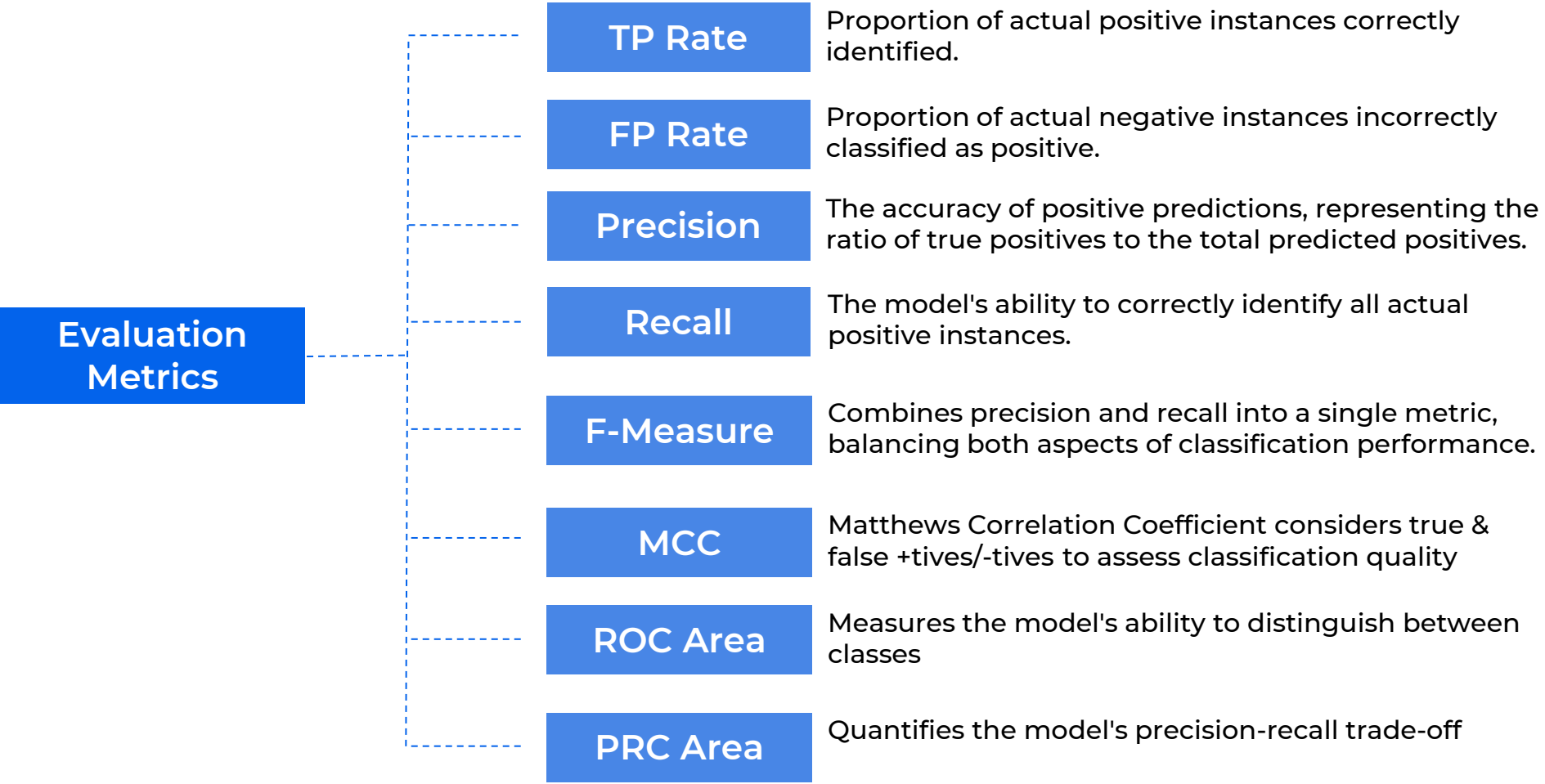
38.5%

617 Instances



Classifier - J48

- Decision tree classifier
- Recursively splits based on attribute
- Select features to create decision nodes and branches.
- Emphasizes information gain.
- Effective for categorical data.


Evaluation Metrics
TP Rate
Proportion of actual positive instances correctly identified.
FP Rate
Proportion of actual negative instances incorrectly classified as positive.
Precision
The accuracy of positive predictions, representing the ratio of true positives to the total predicted positives.
Recall
The model's ability to correctly identify all actual positive instances.
F-Measure
Combines precision and recall into a single metric, balancing both aspects of classification performance.
MCC
Matthews Correlation Coefficient considers true & false +tives/-tives to assess classification quality
ROC Area
Measures the model's ability to distinguish between classes
PRC Area
Quantifies the model's precision-recall trade-off

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.1	0.006	0.091	0.1	0.095	0.089	0.541	0.026	1
0.113	0.032	0.107	0.113	0.11	0.079	0.526	0.047	2
0.711	0.237	0.689	0.711	0.7	0.472	0.76	0.652	3
0.614	0.259	0.612	0.614	0.613	0.355	0.706	0.579	4
0.497	0.056	0.559	0.497	0.527	0.465	0.789	0.418	5
0	0.008	0	0	0	-0.009	0.611	0.022	6
0.614	0.213	0.611	0.614	0.612	0.403	0.732	0.563	Weighted Avg.

Confusion Matrix

	a	b	c	d	e	f	
a	1	3	2	3	1	0	a = 1
b	4	6	24	16	3	0	b = 2
c	2	24	484	154	17	0	c = 3
d	3	19	169	392	50	5	d = 4
e	1	4	23	65	99	7	e = 5
f	0	0	0	11	7	0	f = 6

A confusion matrix is a compact table summarizing the performance of a classification model, detailing true positive, true negative, false positive, and false negative predictions for each class, aiding in model evaluation and error analysis.

03

Conclusion

This presentation on wine quality classification using the J48 algorithm in Weka highlights the significance of classification in machine learning, specifically exploring the Weka software and its application to the Wine Quality dataset. The detailed accuracy metrics and confusion matrix provide valuable insights into model performance, aiding in informed decision-making using classification in Data Mining

Thank You