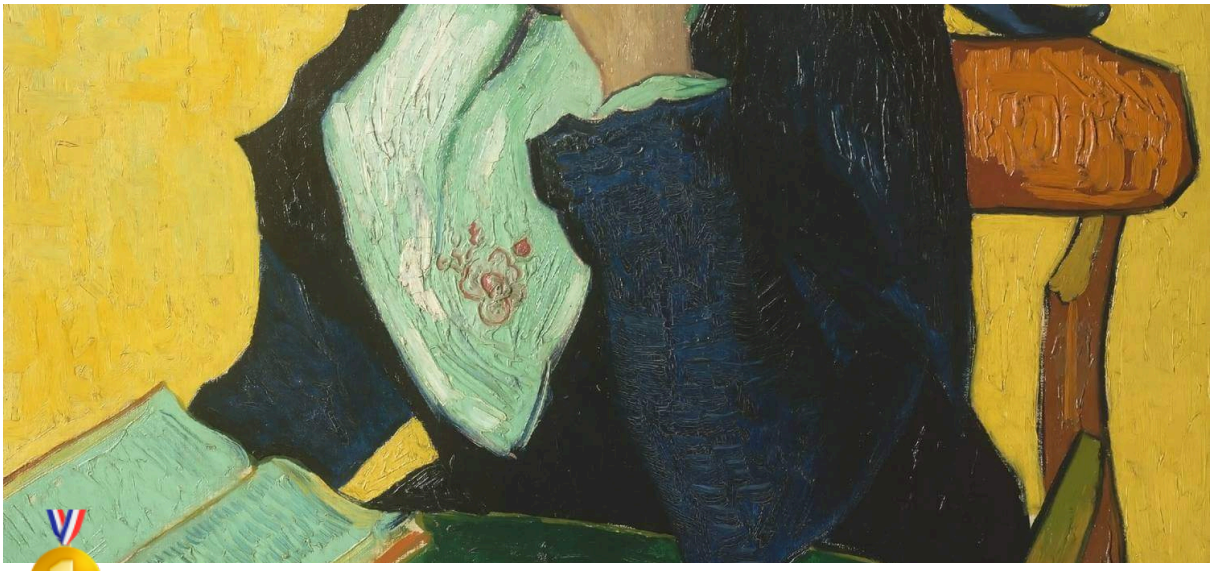
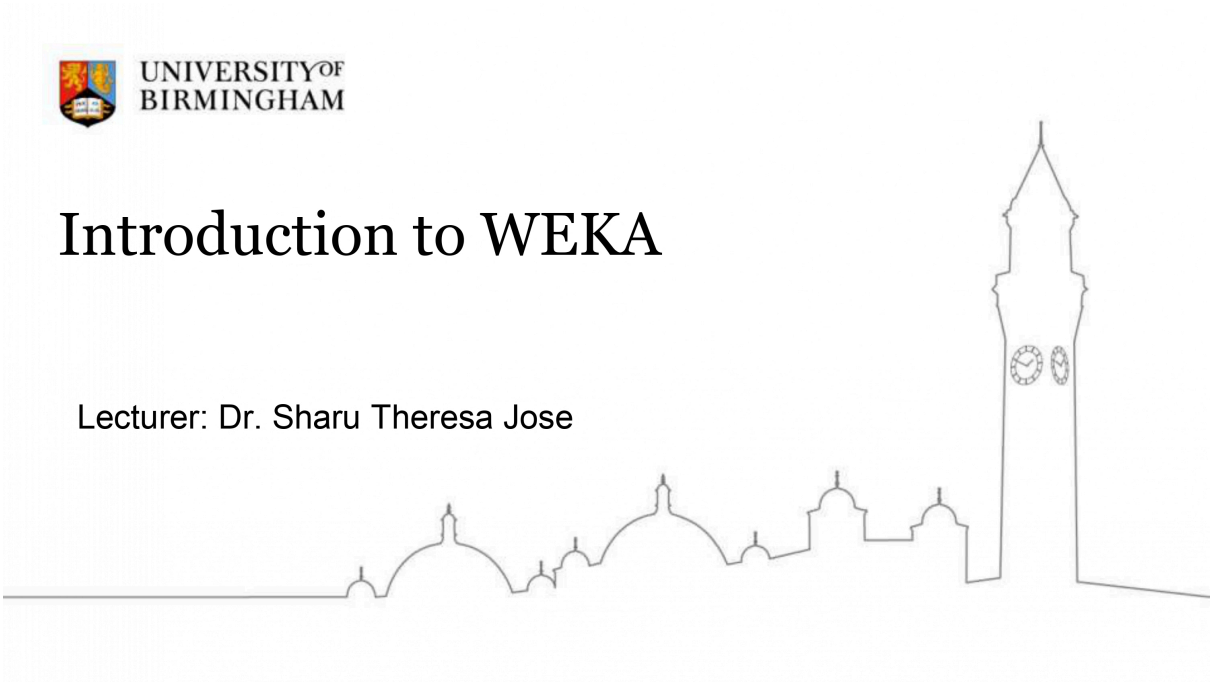


Weka


UNIVERSITY OF BIRMINGHAM
Introduction to WEKA
Lecturer: Dr. Sharu Theresa Jose

Learning Outcomes

- Familiarize with a popular ML software – WEKA
- Familiarize with different data sets
- Application of ML algorithms to real-world data sets

UNIVERSITY OF BIRMINGHAM

WEKA

- WEKA = Waikato Environment for Knowledge Analysis
- Developed at the University of Waikato in New Zealand
- Java-based open-source software
- Collection of machine learning algorithms and data pre-processing tools

UNIVERSITY OF BIRMINGHAM

What's in WEKA?

- Provides implementations of learning algorithms that can be easily applied on data set (without actually coding !)
- Includes methods for main ML problems – classification, regression, clustering, attribute selection
- How can we use it?
  - Apply a learning method to a dataset and analyze its output
  - Learn models to predict on new instances
  - Compare performances of different learning methods

UNIVERSITY OF BIRMINGHAM

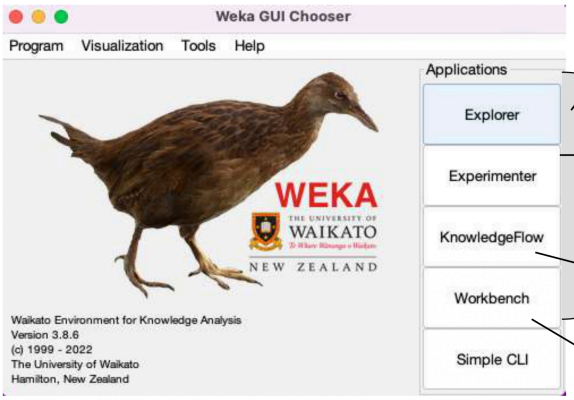
Downloading and using Weka

- Can be downloaded from https://www.cs.waikato.ac.nz/ml/weka/
- Access Weka via school's jupyter notebook server https://jupyterhub.oc1.aws.cs.bham.ac.uk. (This, however, does not support graphical user interface. Details on how to work with Jupyter notebook server will be uploaded as a separate video.)

UNIVERSITY OF BIRMINGHAM

Starting WEKA

Download from: **http://www.cs.waikato.ac.nz/ml/weka**



Graphical User Interfaces (GUIs)

- Easiest to use
- Access to all facilities via menu selection and form filling
- Applicable to small-medium sized datasets

- Easy experimentation
- Automate the choice of methods and parameter values that work best for the given problem

- Design configurations for streamed data processing
- Helps specify a data stream by connecting components

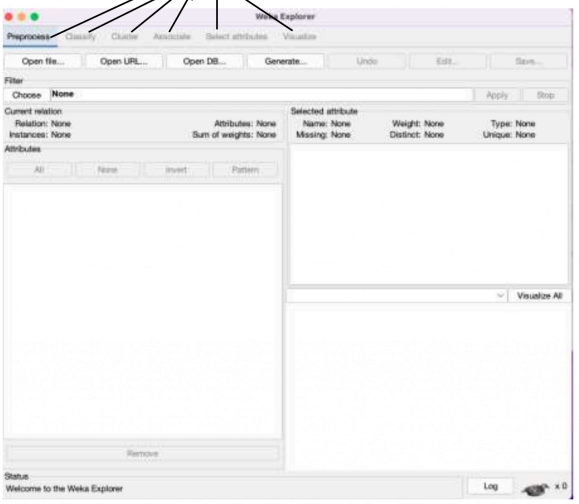
- Highly configurable, combines the other three into one application

UNIVERSITY OF BIRMINGHAM

The Explorer

- **Preprocess**: choose dataset and modify it in various ways
- **Classify**: train learning schemes that perform classification or regression and evaluate them
- **Cluster**: learn clusters from the dataset
- **Associate**: learn association rules for the data and evaluate them
- **Select Attributes**: select the most relevant aspects of the dataset
- **Visualize**: view different two-dimensional plots of the data and interact with them

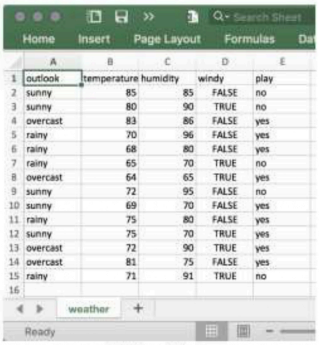
Facilities supported by the Explorer



UNIVERSITY OF BIRMINGHAM

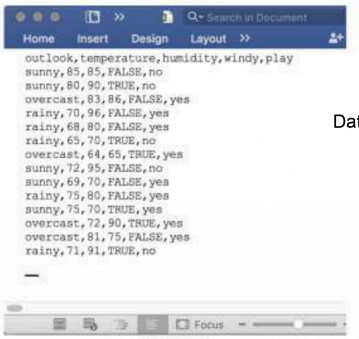
Getting the Explorer Started: Data Format

- Data storage format: ARFF (Attribute-Relation File Format)



	A	B	C	D	E
1	outlook	temperature	humidity	windy	play
2	sunny	85	85	FALSE	no
3	sunny	80	90	TRUE	no
4	overcast	83	86	FALSE	yes
5	rainy	70	96	FALSE	yes
6	rainy	68	80	FALSE	yes
7	rainy	65	70	TRUE	no
8	overcast	64	65	TRUE	yes
9	sunny	72	95	FALSE	no
10	sunny	69	70	FALSE	yes
11	rainy	75	80	FALSE	yes
12	sunny	75	70	TRUE	yes
13	overcast	72	90	TRUE	yes
14	overcast	81	75	FALSE	yes
15	rainy	71	91	TRUE	no

(a) Spreadsheet.



```
outlook,temperature,humidity,windy,play
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

(b) CSV.

(c) ARFF.

Demo

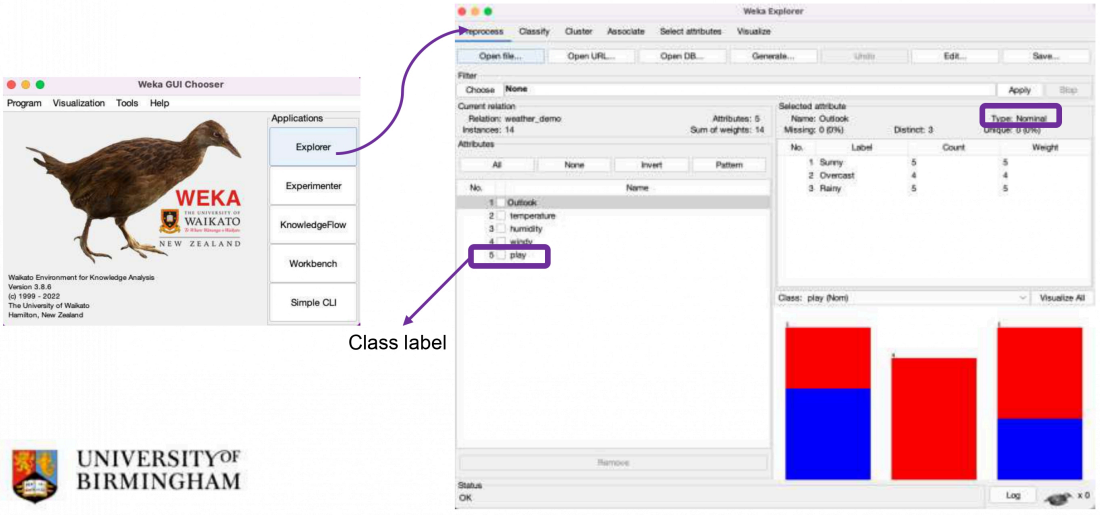
Explorer → Preprocess

- Load data files
- Pre-processing tools or ‘filters’

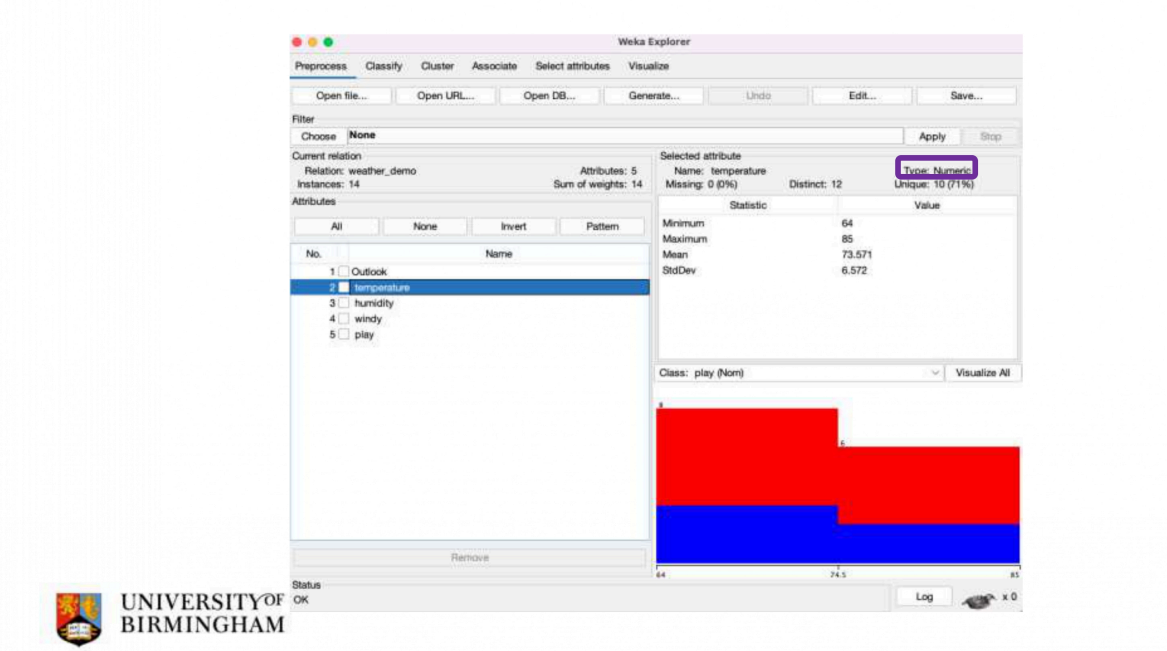
UNIVERSITY OF BIRMINGHAM

Loading data

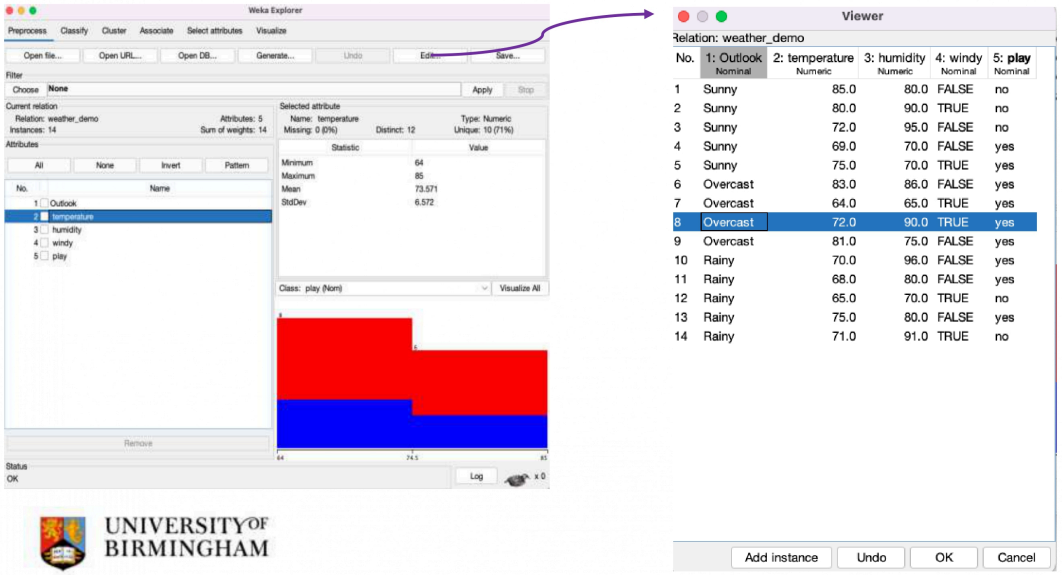
- Preprocess → Open file → select .arff file from the desired directory



UNIVERSITY OF BIRMINGHAM



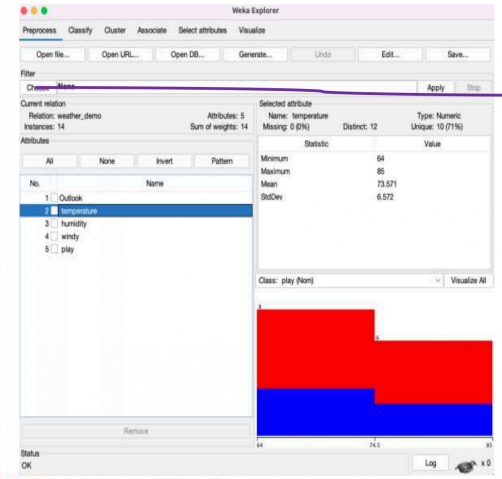
UNIVERSITY OF BIRMINGHAM

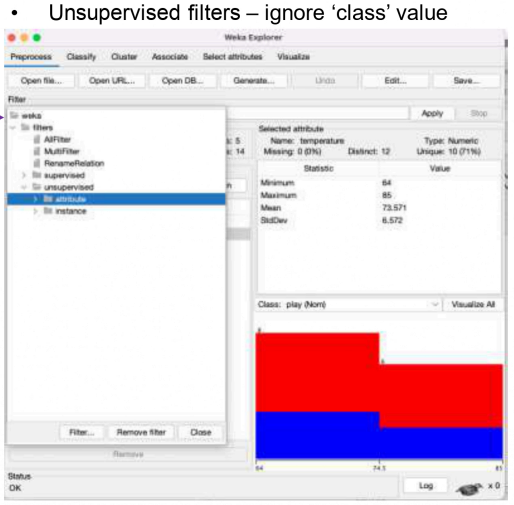


UNIVERSITY OF BIRMINGHAM

Pre-processing via Filters

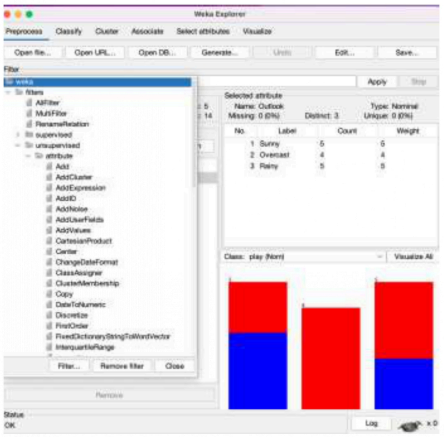
- Supervised filters – consider ‘class’ value
- Unsupervised filters – ignore ‘class’ value





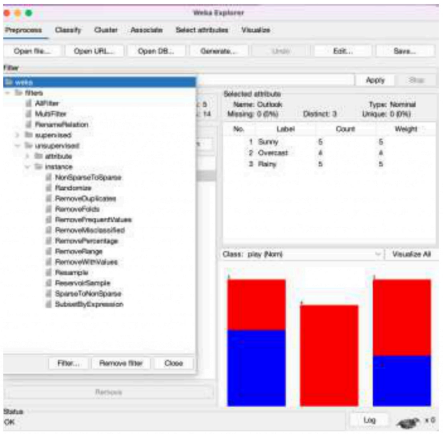
UNIVERSITY OF BIRMINGHAM

- Unsupervised attribute filters – affect specific attributes or a set of attributes
  - Example: Discretize, DatatoNumeric, StringtoWordVector



UNIVERSITY OF BIRMINGHAM

- Unsupervised instance filters – affect all instances in a data set
  - Randomize : randomizes the order of instances in the dataset
  - Resample : produce a random sample by sampling with or without replacement



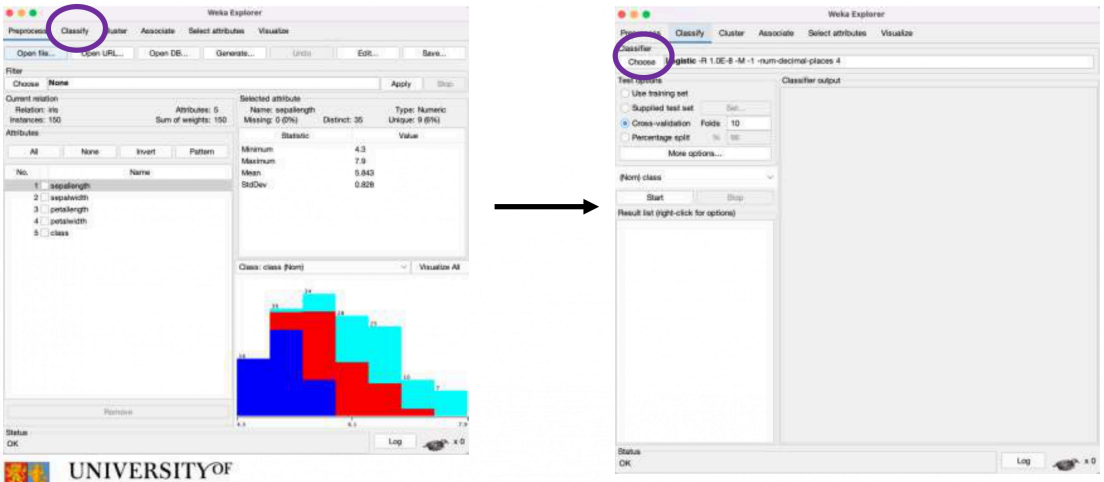
Explorer → Classify

Includes various classification algorithms:

1. Logistic Regression
2. K-Nearest Neighbors
3. Decision Trees
4. Naïve Bayes

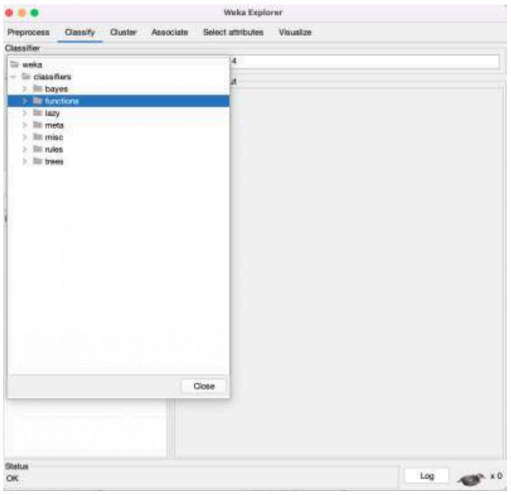
UNIVERSITY OF BIRMINGHAM

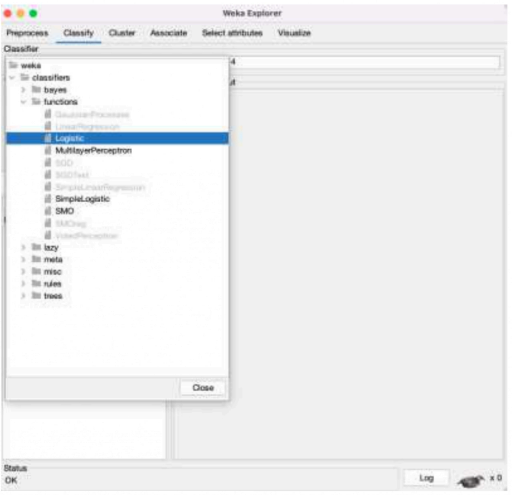
Example: Classification of Iris Data Set



UNIVERSITY OF BIRMINGHAM

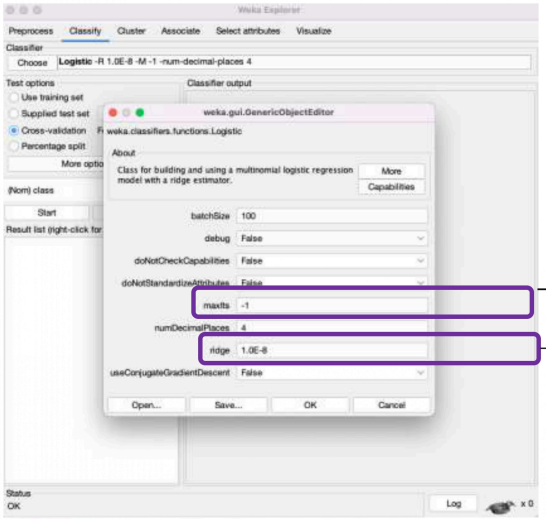
Logistic Regression





UNIVERSITY OF BIRMINGHAM

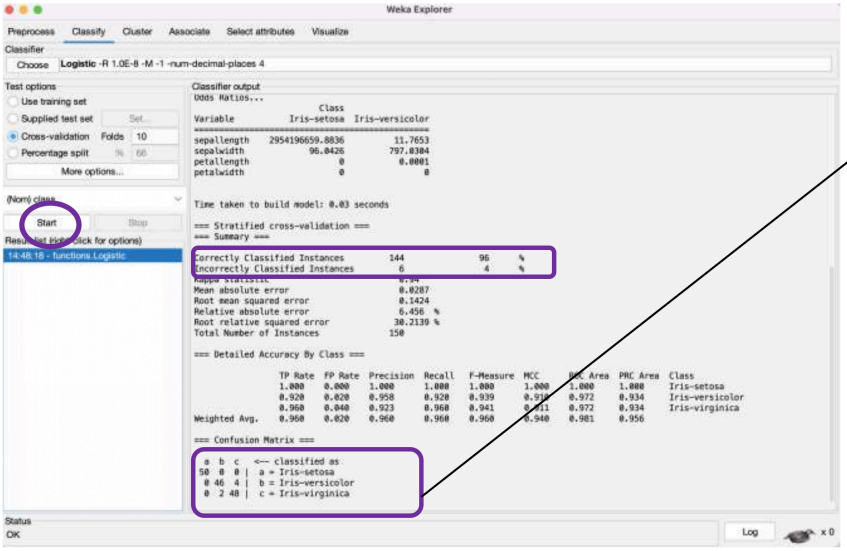
Double-clicking on chosen classifier reveals details.



Maximum number of iterations: Default – until convergence

Ridge regularization parameter: value of zero = no regularization

UNIVERSITY OF BIRMINGHAM



Confusion matrix

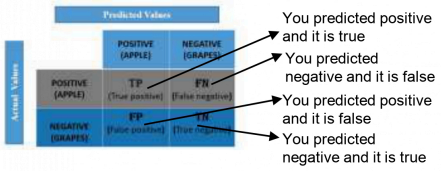
- Each (i,j)th matrix entry corresponds to number of data instances where true class 'i' is classified as class 'j'
- Row indicates true class labels, and column indicates predicted class labels

Confusion Matrix - Details

Confusion matrix is used to calculate class-specific quality metrics – accuracy, precision, recall, F-measure etc.

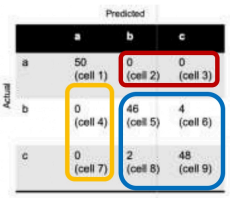
Simple binary classification (into apples and grapes)

- Corresponding to class apple (positive),



Iris classification

- Corresponding to class a,



TP = cell 1
(you predicted 'a' and it is true)

FN = cell 2+cell 3
(you predicted 'not a' and it is false)

FP = cell 4 +cell 7
(you predicted 'a' and it is false)

TN = cell 5+cell 6+ cell 8 + cell 9
(you predicted 'not a' and is true)

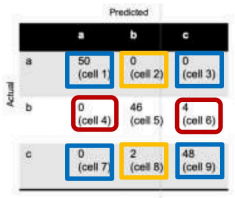
For more details: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

UNIVERSITY OF BIRMINGHAM

Confusion Matrix - Details

Iris classification

- Corresponding to class b,



TP = cell 5
FN = cell 4+cell 6
FP = cell 2 +cell 8
TN = cell 1+cell 3+ cell 7 + cell 9

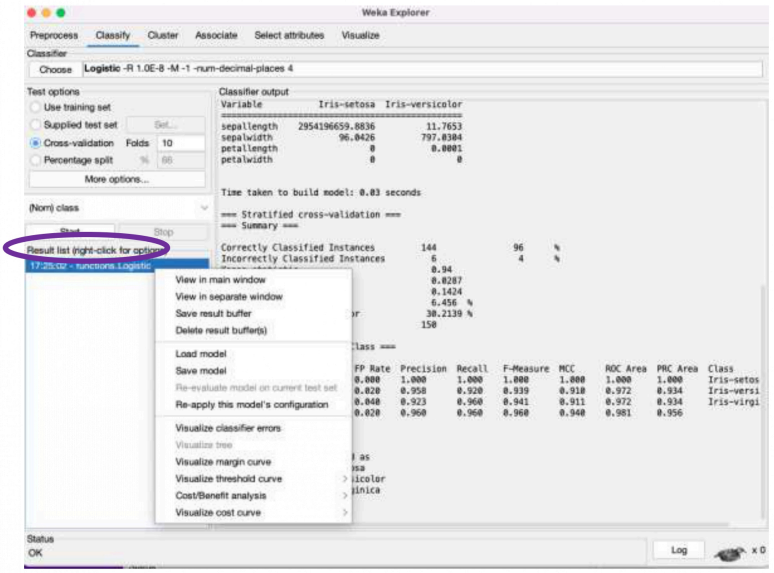
Per-class performance metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN} \qquad F_\beta = \frac{(\beta^2+1)Precision \times Recall}{\beta^2 Precision + Recall}$$

$\beta > 1$: *Favor Recall*
$\beta < 1$: *Favor Precision*

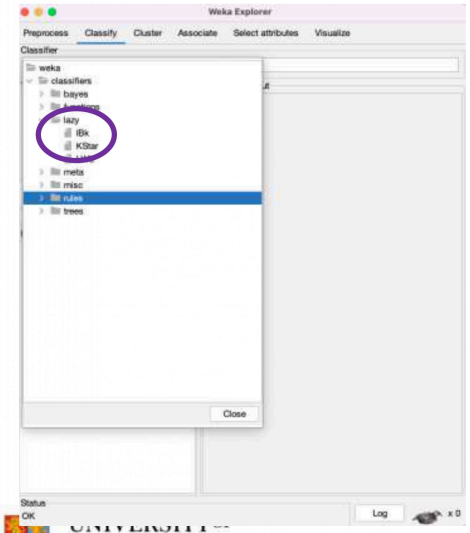
F-measure = F_1

UNIVERSITY OF BIRMINGHAM

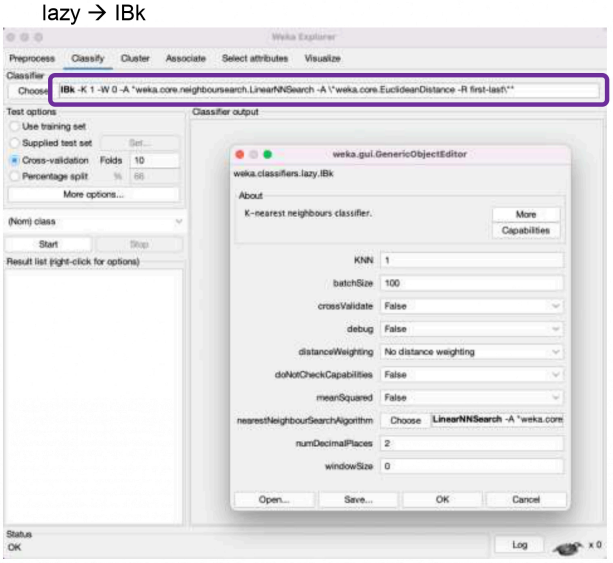


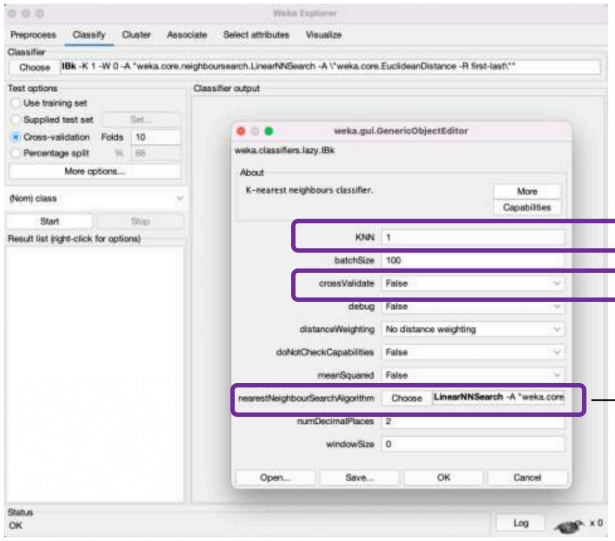
UNIVERSITY OF BIRMINGHAM

K-Nearest Neighbors



lazy → IBk

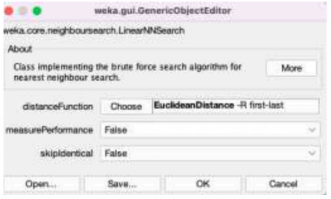


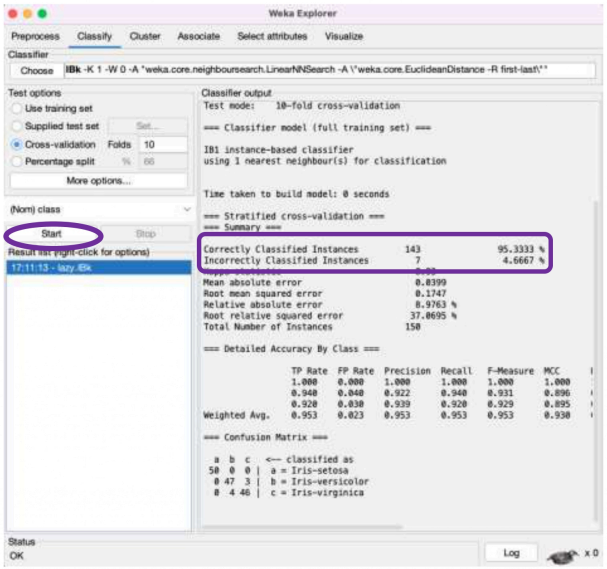


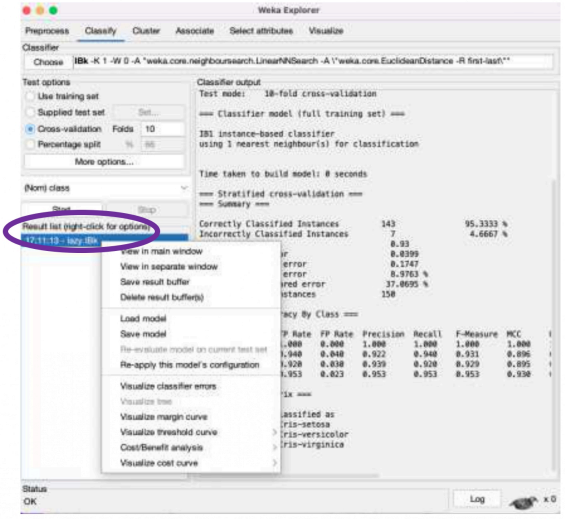
Value of 'k' – Typical values are 3,7,21 etc.

True: enables cross-validation to automatically determine 'k'

Distance measure used – by default, Euclidean distance. Clicking on the choice, reveals the distance measures used





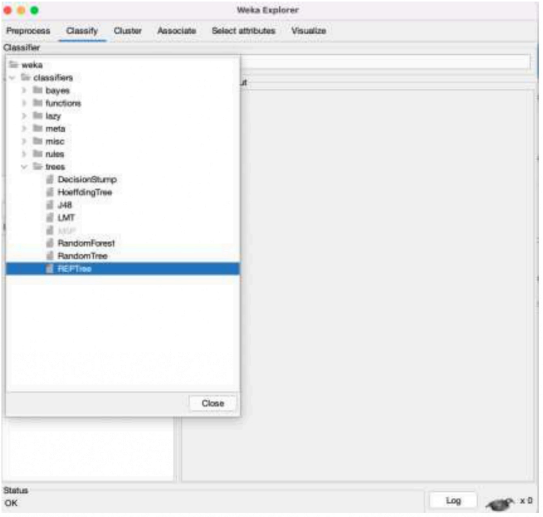


UNIVERSITY OF BIRMINGHAM

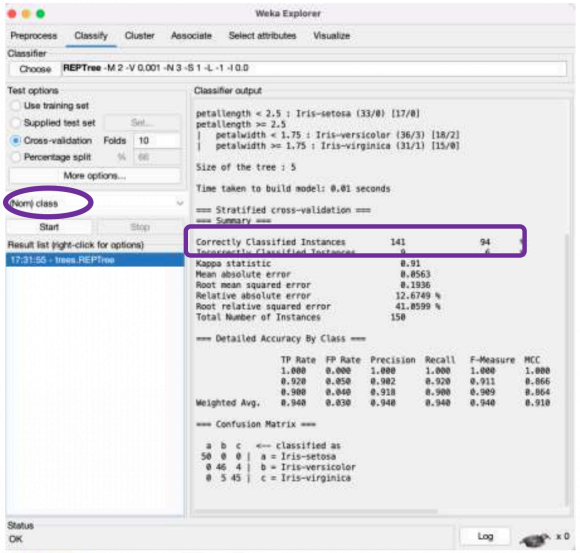
Decision Trees

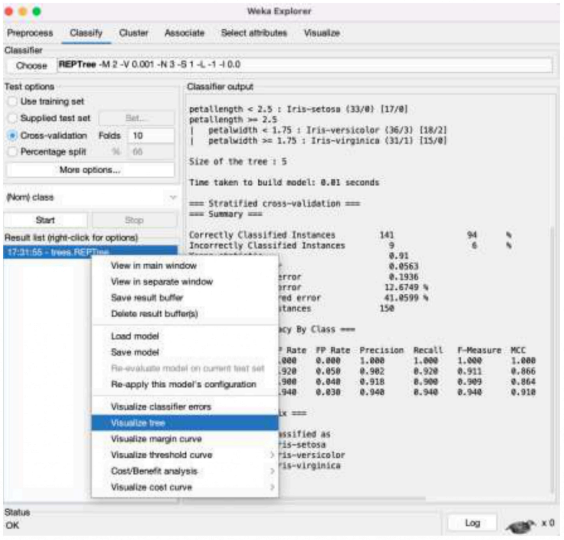
Decision Tree Algorithm:

- Tree is constructed in a top-down recursive divide-and-conquer manner
- At start, all the training examples are at the root
- Examples are partitioned recursively based on selected attributes
- Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

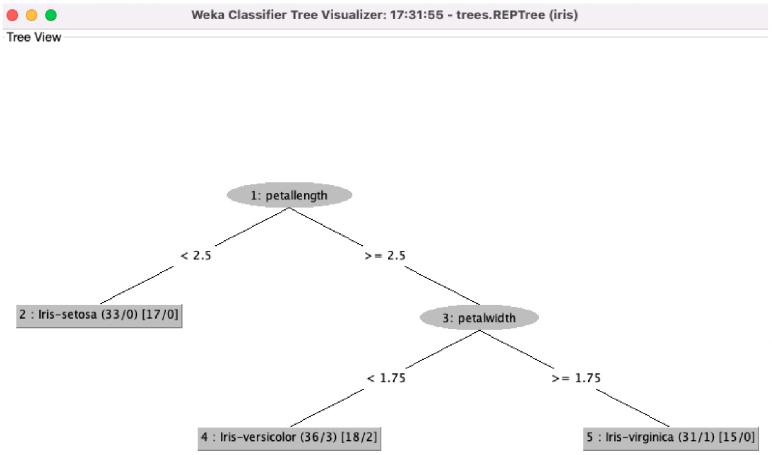


UNIVERSITY OF BIRMINGHAM





UNIVERSITY OF BIRMINGHAM



UNIVERSITY OF BIRMINGHAM

ML Application: Sentiment Analysis

UNIVERSITY OF BIRMINGHAM

Sentiment Analysis: A classification task

- Sentiments:
  - Attitudes: positive/negative/neutral?
  - Emotions: happy/sad?
  - Opinions: like/dislike?
- Analysis of feeling behind the words
- Questions that might be asked in sentiment analysis:
  - Is the product review positive or negative?
  - Peoples' response to campaign reviews?
  - Bloggers' attitude about presidential election?

UNIVERSITY OF BIRMINGHAM

Sentiment Analysis with WEKA: A Route Map

- Get text data with labels (loading the .arff file).
- Pre-processing data: Convert text strings to feature vectors via NLP Techniques (use filters).
- Train a model using a ML algorithm for classification (use Classify → Logistic regression/ Decision Trees/k-NN).
- Test the model and use it for future classification

UNIVERSITY OF BIRMINGHAM

Example : Twitter dataset

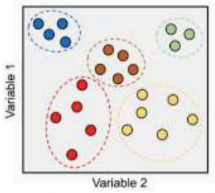
- 100 positive tweets and 100 negative tweets from Edinburgh twitter corpus

```
@relation tweets

@attribute tweet_body string
@attribute sentiment {pos,neg}

@data
'anyone feel motivated the fri afternoon prior to a holiday? wanted to get lots done... but i want
jammies and judge judy... \"SIR!\" &lt;3 her ',pos
'seriously, do you have to rub it in maggie!!!!  ',pos
'if i\'m not wrong.. Alt is when image can\'t be displayed.. Tooltip is the \'title\' ',pos
'I don\'t like social karma  much. Would rather skip it, but can\'t afford to piss my friend off
any more ',pos
'I\'d be happy to review the Iomega if EMC send me one!!! ',pos
'Something  I have wanted to make for a while now... finally done  URL',pos
'you\'re so sweet!  proving me right again...  the Dutch are the Best!  ',pos
'Look for us on the back of your Pepsi can! Our Pepsi can offer hit\nstores today! Can\'t wait to
see one ',pos
'well rob i have to admit that you have to admit that you feel cool for being on twitter and
listening to me ',pos
'big skype call at 1010! msg me if you want in ',pos
'well don\'t let it happen again ',pos
'thanks for following me and welcome to my world! ',pos
'season 3... all of it.. al is looking for season 4... i can\'t believe he didn\'t prepare... ',pos
'Yikes, it\'s windy here today; up to 50 mph NOAA says. Good thing electrician taught me generator
drill yesterday; feeling, uh, empowered. ',pos
'URL - Well, my fave bar DID give me a blue moon to-go. ',pos
'Makes at least two of us.  ',pos
'that \'slang\' doesn\'t really work for you ',pos
'yes i did THANK YOU! ',pos
```

- Input: tweets, output: positive or negative
- Try in WEKA- Demo in lecture

UNIVERSITY OF BIRMINGHAM

Explorer → Cluster

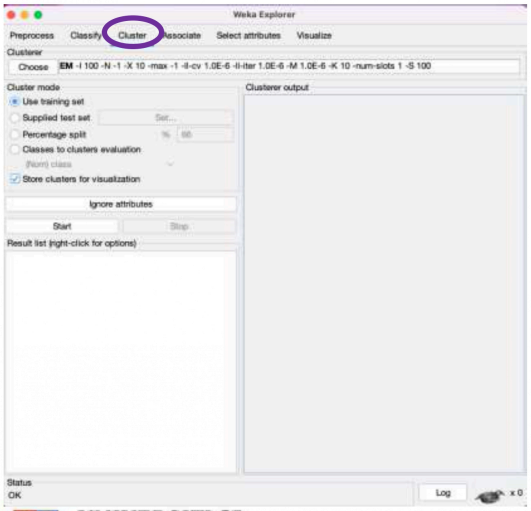


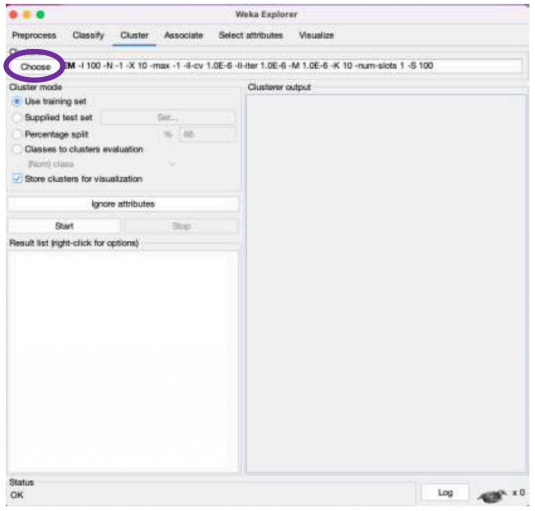
Clustering algorithms are a class of unsupervised learning algorithms (no labels) that aims to cluster the observed instances into groups with similar traits (to be covered in detail in the following lectures). WEKA supports clustering algorithms like:

1. K-Means
2. Hierarchical Clustering
3. Density based clustering

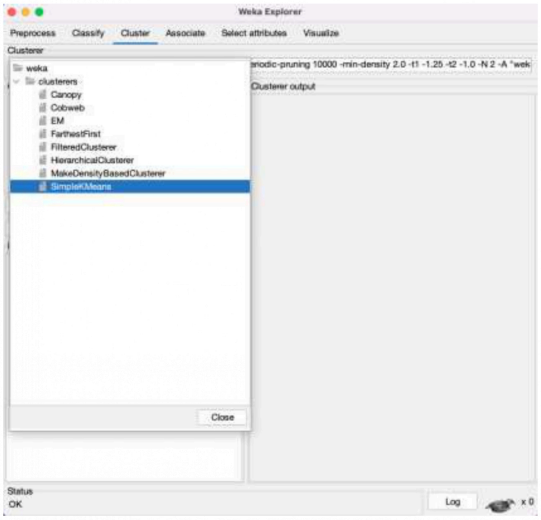
UNIVERSITY OF BIRMINGHAM

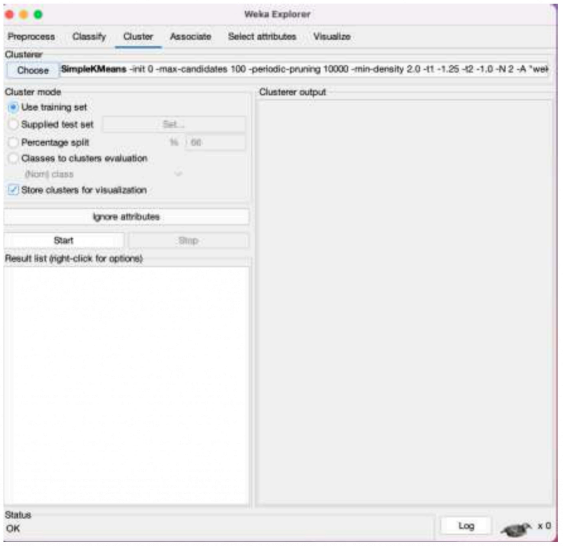
Simple K-Means





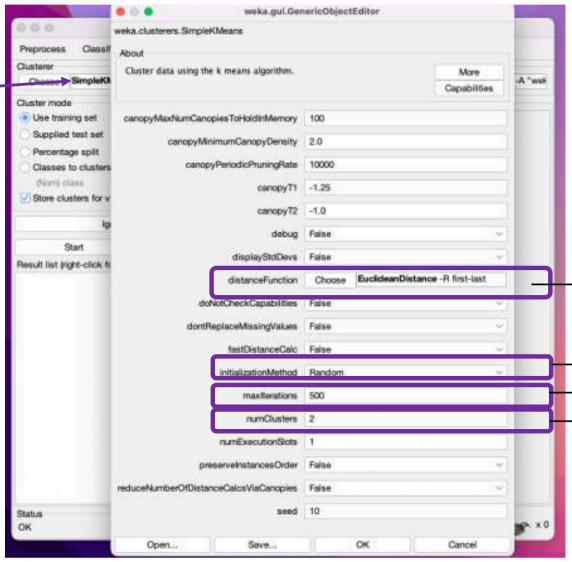
UNIVERSITY OF BIRMINGHAM





UNIVERSITY OF BIRMINGHAM

Click here to open the GenericObjectEditor



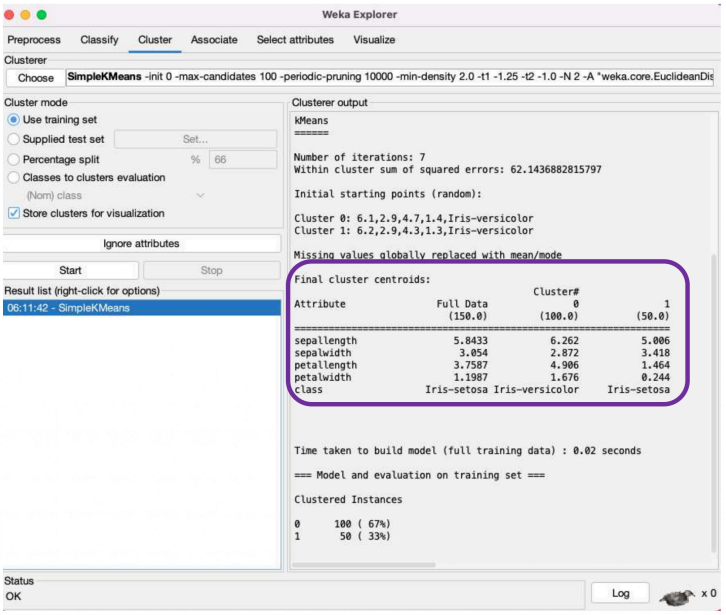
Choice of distance function: default, Euclidean distance

Choosing initial cluster centroids

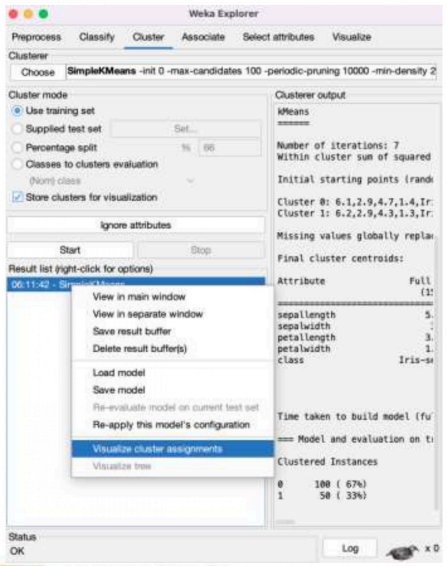
Maximum no.of iterations

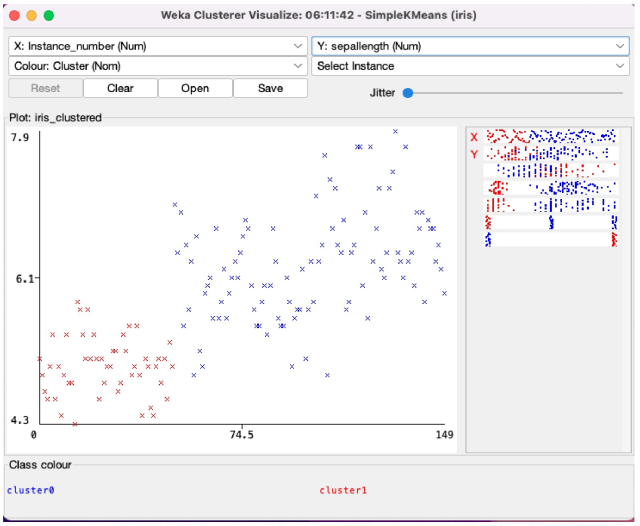
Fix 'K', the number of clusters

UNIVERSITY OF BIRMINGHAM



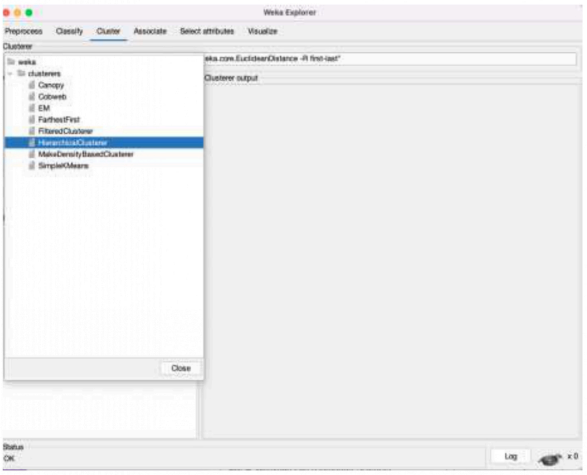
UNIVERSITY OF BIRMINGHAM

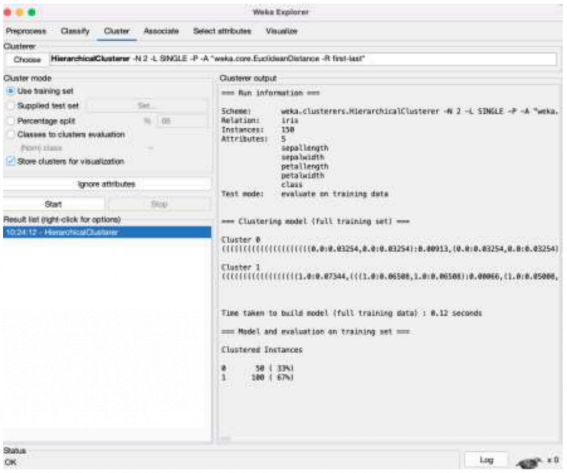




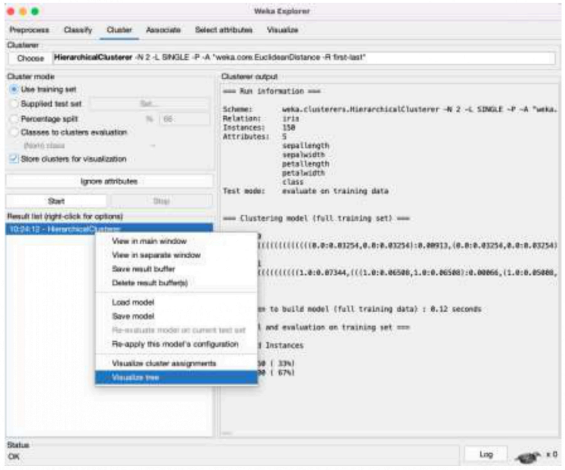
UNIVERSITY OF BIRMINGHAM

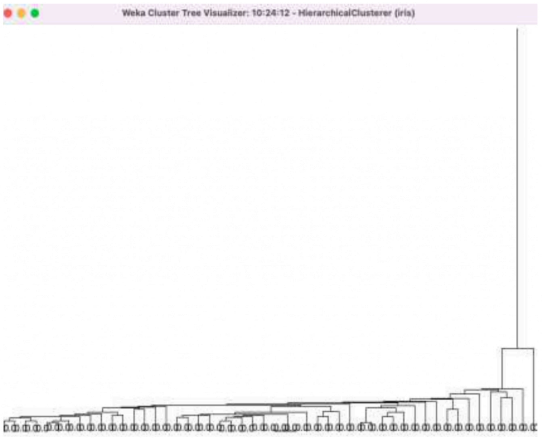
Hierarchical Clusterer





UNIVERSITY OF BIRMINGHAM





UNIVERSITY OF BIRMINGHAM

ML Application: Dominant Color Extraction

UNIVERSITY OF BIRMINGHAM

Dominant Color Extraction – A Clustering Task

- Application – extraction of color palettes from artworks
- Input data – an image

```
@relation house
@attribute red real
@attribute green real
@attribute blue real

@data
    26    20    45
    28     5    46
    28    12    44
    28    13    46
    31     4    51
    31     5    48
    31     8    44
    31     9    44
    32     5    50
    32     8    47
    32    11    47
    32    12    48
    33     9    47
    33    17    45
    33    29    48
    34     6    48
    34     9    50
    34    13    52
    34    28    45
    34    32    47
```

Each data instance: RGB value of each pixel in the image

UNIVERSITY OF BIRMINGHAM

Clustering Route Map in WEKA

- Load the .arff file corresponding to the RGB vectors of the image
- Train a model that cluster the RGB feature vectors into K groups via a clustering algorithm like K-means
- Obtain clusters and output the RGB value of each cluster centroid.

Note: cluster centroid = average of the feature vectors in a cluster (can be treated as a summary of the cluster)

Try in WEKA

UNIVERSITY OF BIRMINGHAM

References

- For documentation on WEKA, https://waikato.github.io/weka-wiki/documentation/
- WEKA book: Ian H. Witten and Eibe Frank, Data Mining : Practical Machine Learning Tools and Techniques (Second Edition)

UNIVERSITY OF BIRMINGHAM