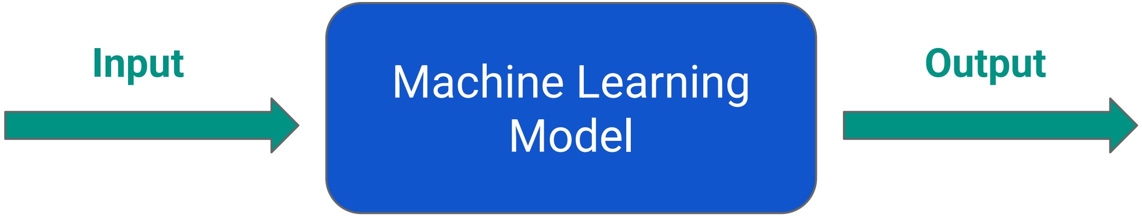
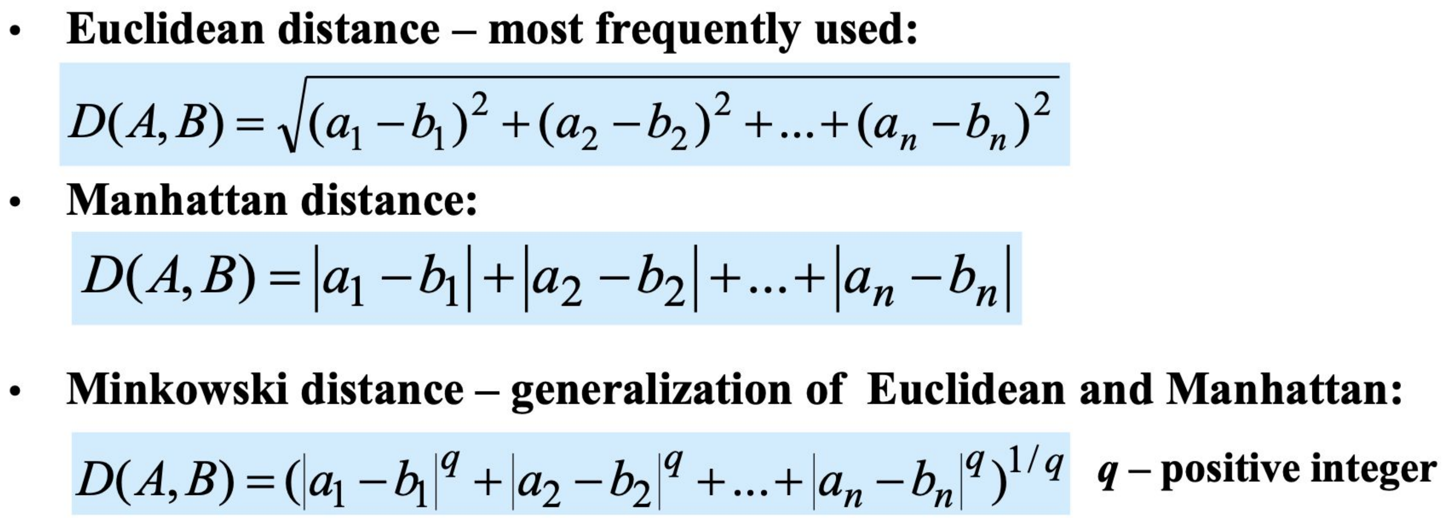
Machine Learning



K-Nearest Neighbour - Numeric Attributes



- **Euclidean distance – most frequently used:**

$$D(A,B)=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\ldots+(a_n-b_n)^2}$$

- **Manhattan distance:**

$$D(A,B)=|a_1-b_1|+|a_2-b_2|+\ldots+|a_n-b_n|$$

- **Minkowski distance – generalization of Euclidean and Manhattan:**

$$D(A,B)=(|a_1-b_1|^q+|a_2-b_2|^q+\ldots+|a_n-b_n|^q)^{1/q}$$ ***q* – positive integer**

K-Nearest Neighbour - Numeric Attributes

- **Euclidean distance – most frequently used:**

$$D(A,B)=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\ldots+(a_n-b_n)^2}$$

- **Manhattan distance:**

$$D(A,B)=|a_1-b_1|+|a_2-b_2|+\ldots+|a_n-b_n|$$

Homework

- **Minkowski distance – generalization of Euclidean and Manhattan:**

$$D(A,B)=(|a_1-b_1|^q+|a_2-b_2|^q+\ldots+|a_n-b_n|^q)^{1/q}$$

***q* – positive integer**

Exercise 1 – Homework

Student	Assignment 1	Assignment 2	Exam Grade
1	60	85	HD
2	50	40	P
3	40	40	F
4	20	30	F
5	55	75	CR
6	50	90	D

Exercise 1 – Homework

Step 1. Calculate the distances

Student	Assignment 1	Assignment 2	Exam Grade	D(Isabella, student)
1	60	85	HD	
2	50	40	P	
3	40	40	F	
4	20	30	F	
5	55	75	CR	
6	50	90	D	

Exercise 1 – Homework

Step 1. Calculate the distances

Student	Assignment 1	Assignment 2	Exam Grade	D(Isabella, student)
1	60	85	HD	**10**
2	50	40	P	**55**
3	40	40	F	**85**
4	20	30	F	**95**
5	55	75	CR	**15**
6	50	90	D	**25**

Exercise 1 – Homework

Step 2. Identify the 1 Nearest Neighbour (i.e. shortest distance)

Student	Assignment 1	Assignment 2	Exam Grade	D(Isabella, student)
1	60	85	HD	**10**
2	50	40	P	**55**
3	40	40	F	**85**
4	20	30	F	**95**
5	55	75	CR	**15**
6	50	90	D	**25**

Exercise 1 – Homework

Step 3. Return the class associated with the 1 Nearest Neighbour

Student	Assignment 1	Assignment 2	Exam Grade	D(Isabella, student)
1	60	85	HD	**10**
2	50	40	P	**55**
3	40	40	F	**85**
4	20	30	F	**95**
5	55	75	CR	**15**
6	50	90	D	**25**

Exercise 1 – Homework

=> 1NN will predict HD for Isabella’s exam grade.

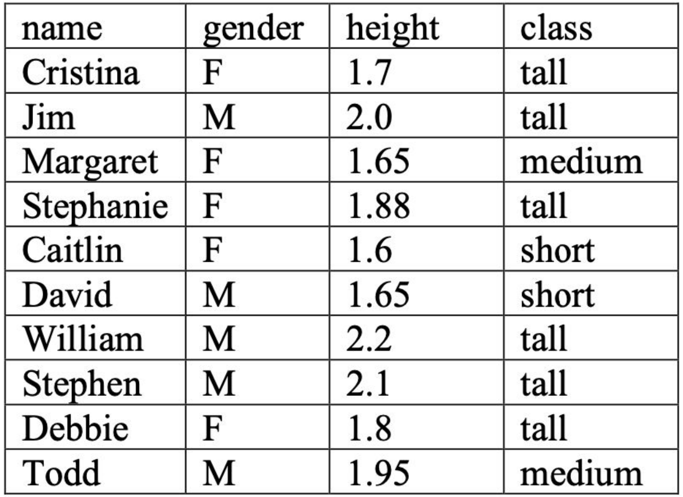
Student	Assignment 1	Assignment 2	Exam Grade	D(Isabella, student)
1	60	85	HD	**10**
2	50	40	P	**55**
3	40	40	F	**85**
4	20	30	F	**95**
5	55	75	CR	**15**
6	50	90	D	**25**

K-Nearest Neighbour - Nominal Attributes

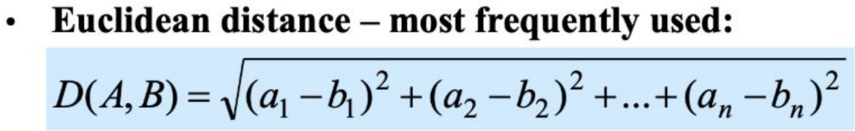
Distance between attribute values is:

- **1** - if the two values are not the same
- **0** - if the two values are the same

Exercise 2



name	gender	height	class
Cristina	F	1.7	tall
Jim	M	2.0	tall
Margaret	F	1.65	medium
Stephanie	F	1.88	tall
Caitlin	F	1.6	short
David	M	1.65	short
William	M	2.2	tall
Stephen	M	2.1	tall
Debbie	F	1.8	tall
Todd	M	1.95	medium



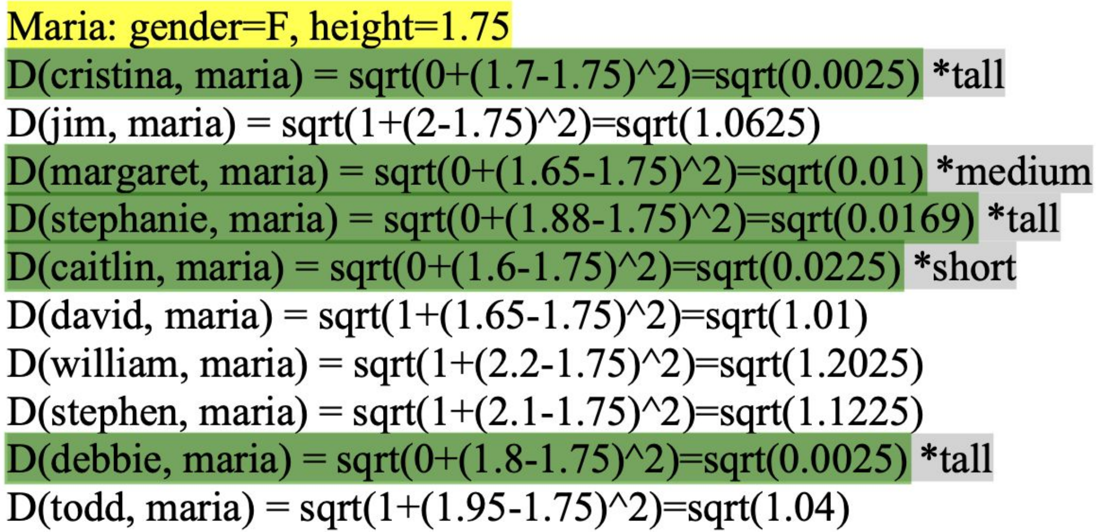
- **Euclidean distance – most frequently used:**

$$D(A,B)=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+...+(a_n-b_n)^2}$$

What would be the prediction of 5-Nearest-Neighbor using Euclidian distance for Maria who is (F, 1.75)? Show your calculations. Do not apply normalization for this exercise (but note that in practical situations you need to do this for the numeric attributes). In case of ties, make random selection.

Exercise 2

The 5 Nearest Neighbours are their distances are highlighted in green



Maria: gender=F, height=1.75

D(cristina, maria) = sqrt(0+(1.7-1.75)^2)=sqrt(0.0025) *tall

D(jim, maria) = sqrt(1+(2-1.75)^2)=sqrt(1.0625)

D(margaret, maria) = sqrt(0+(1.65-1.75)^2)=sqrt(0.01) *medium

D(stephanie, maria) = sqrt(0+(1.88-1.75)^2)=sqrt(0.0169) *tall

D(caitlin, maria) = sqrt(0+(1.6-1.75)^2)=sqrt(0.0225) *short

D(david, maria) = sqrt(1+(1.65-1.75)^2)=sqrt(1.01)

D(william, maria) = sqrt(1+(2.2-1.75)^2)=sqrt(1.2025)

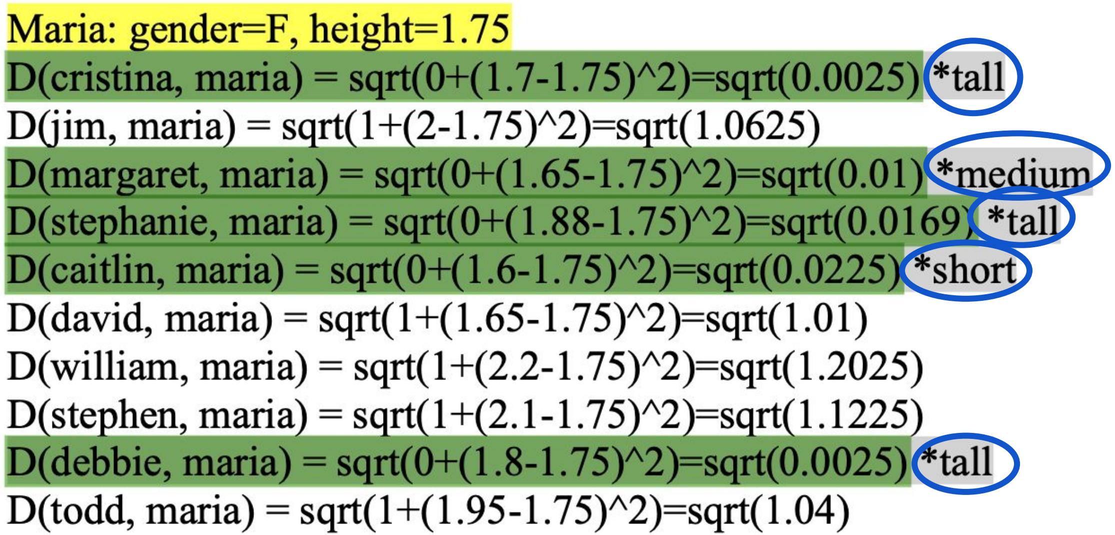
D(stephen, maria) = sqrt(1+(2.1-1.75)^2)=sqrt(1.1225)

D(debbie, maria) = sqrt(0+(1.8-1.75)^2)=sqrt(0.0025) *tall

D(todd, maria) = sqrt(1+(1.95-1.75)^2)=sqrt(1.04)

Exercise 2

Out of these 5NN: 3 *tall*, 1 *medium*, 1 *short*



Maria: gender=F, height=1.75

D(cristina, maria) = sqrt(0+(1.7-1.75)^2)=sqrt(0.0025) *tall

D(jim, maria) = sqrt(1+(2-1.75)^2)=sqrt(1.0625)

D(margaret, maria) = sqrt(0+(1.65-1.75)^2)=sqrt(0.01) *medium

D(stephanie, maria) = sqrt(0+(1.88-1.75)^2)=sqrt(0.0169) *tall

D(caitlin, maria) = sqrt(0+(1.6-1.75)^2)=sqrt(0.0225) *short

D(david, maria) = sqrt(1+(1.65-1.75)^2)=sqrt(1.01)

D(william, maria) = sqrt(1+(2.2-1.75)^2)=sqrt(1.2025)

D(stephen, maria) = sqrt(1+(2.1-1.75)^2)=sqrt(1.1225)

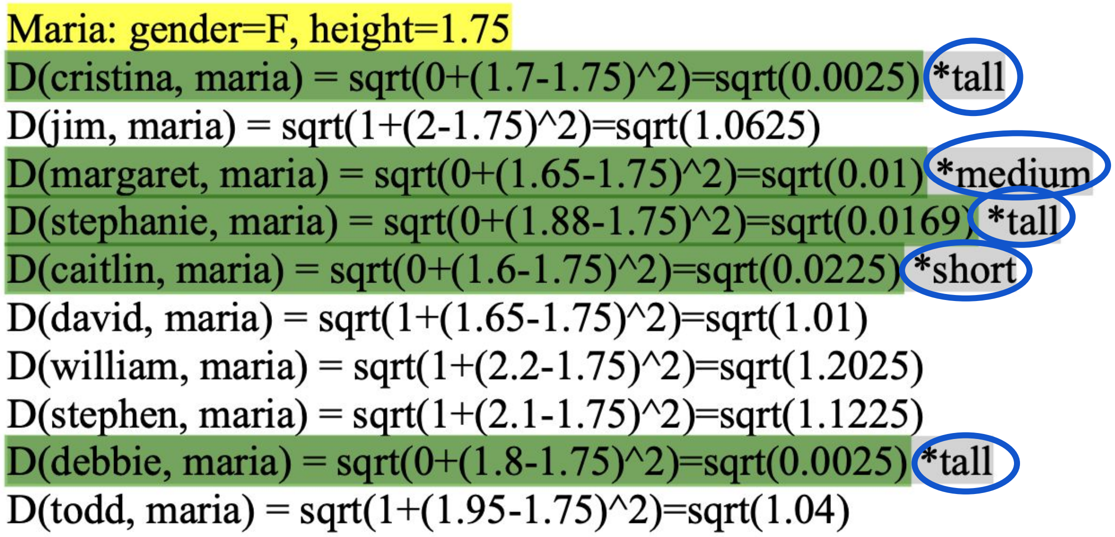
D(debbie, maria) = sqrt(0+(1.8-1.75)^2)=sqrt(0.0025) *tall

D(todd, maria) = sqrt(1+(1.95-1.75)^2)=sqrt(1.04)

Exercise 2

Out of these 5NN: 3 *tall*, 1 *medium*, 1 *short*

=> 5NN classified Maria as *tall*



Maria: gender=F, height=1.75

D(cristina, maria) = sqrt(0+(1.7-1.75)^2)=sqrt(0.0025) *tall

D(jim, maria) = sqrt(1+(2-1.75)^2)=sqrt(1.0625)

D(margaret, maria) = sqrt(0+(1.65-1.75)^2)=sqrt(0.01) *medium

D(stephanie, maria) = sqrt(0+(1.88-1.75)^2)=sqrt(0.0169) *tall

D(caitlin, maria) = sqrt(0+(1.6-1.75)^2)=sqrt(0.0225) *short

D(david, maria) = sqrt(1+(1.65-1.75)^2)=sqrt(1.01)

D(william, maria) = sqrt(1+(2.2-1.75)^2)=sqrt(1.2025)

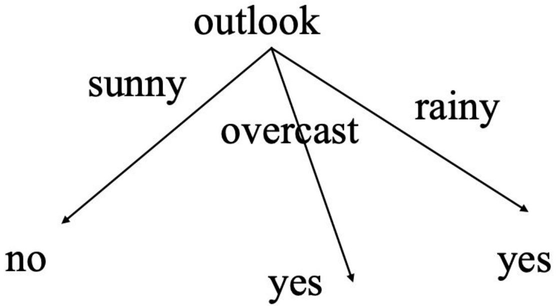
D(stephen, maria) = sqrt(1+(2.1-1.75)^2)=sqrt(1.1225)

D(debbie, maria) = sqrt(0+(1.8-1.75)^2)=sqrt(0.0025) *tall

D(todd, maria) = sqrt(1+(1.95-1.75)^2)=sqrt(1.04)

1R

- 1R stands for “1-rule”
- Generate 1 rule that tests the value of a single attribute
- The rule can be represented as 1-level decision tree (decision stump)
  - At the root: test the attribute value
  - Each branch corresponds to a value
  - Each leaf corresponds to a class



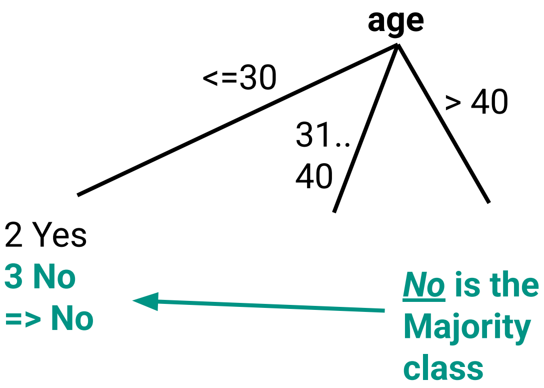
1R

- 1R stands for “1-rule”
- Generate 1 rule that tests the value of a single attribute

There are many attributes. How to select the best attribute?

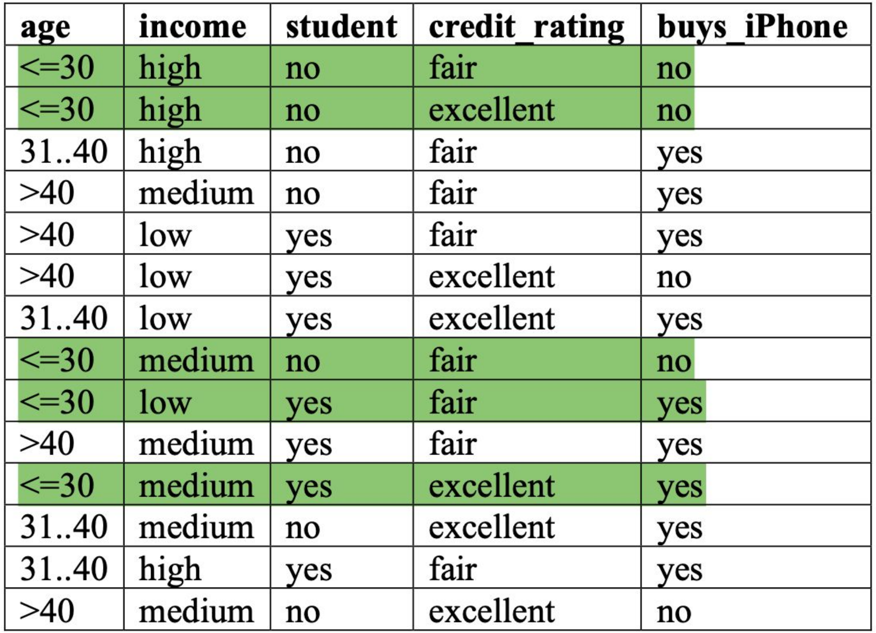
The best attribute minimise the number of training examples being misclassified (i.e. number of errors).

Exercise 3 – 1R



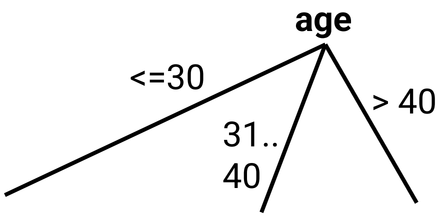
2 Yes
3 No
=> No

No **is the Majority class**

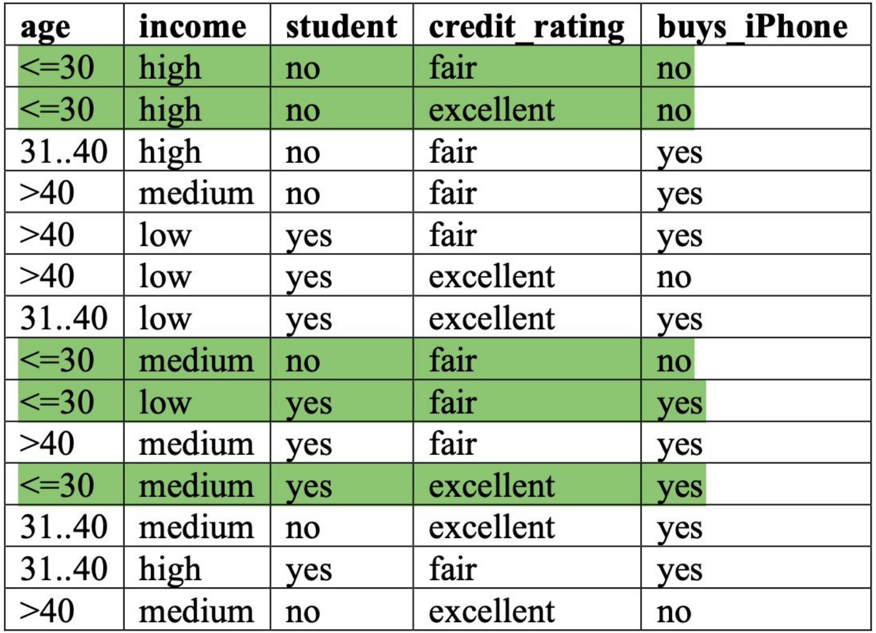


age	income	student	credit_rating	buys_iPhone
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise 3 – 1R

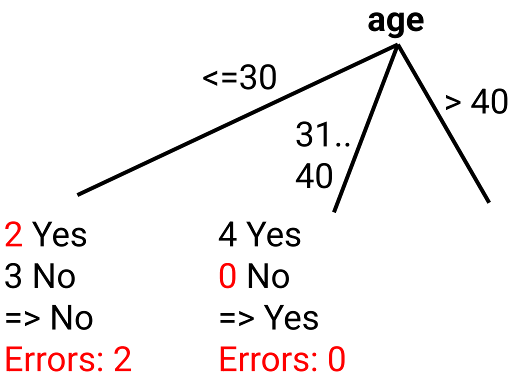


2 Yes
3 No
=> No
Errors: 2



age	income	student	credit_rating	buys_iPhone
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise 3 – 1R



age

<=30

31..
40

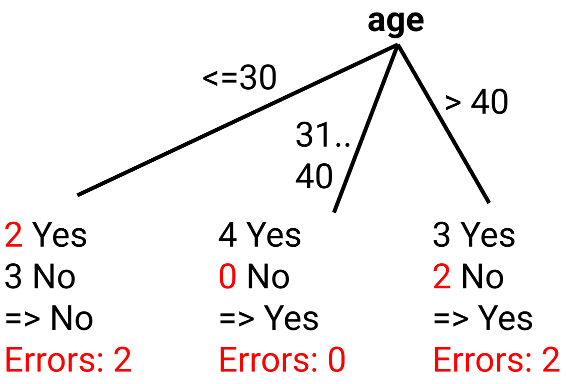
> 40

2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

age	income	student	credit_rating	buys_iPhone
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise 3 – 1R

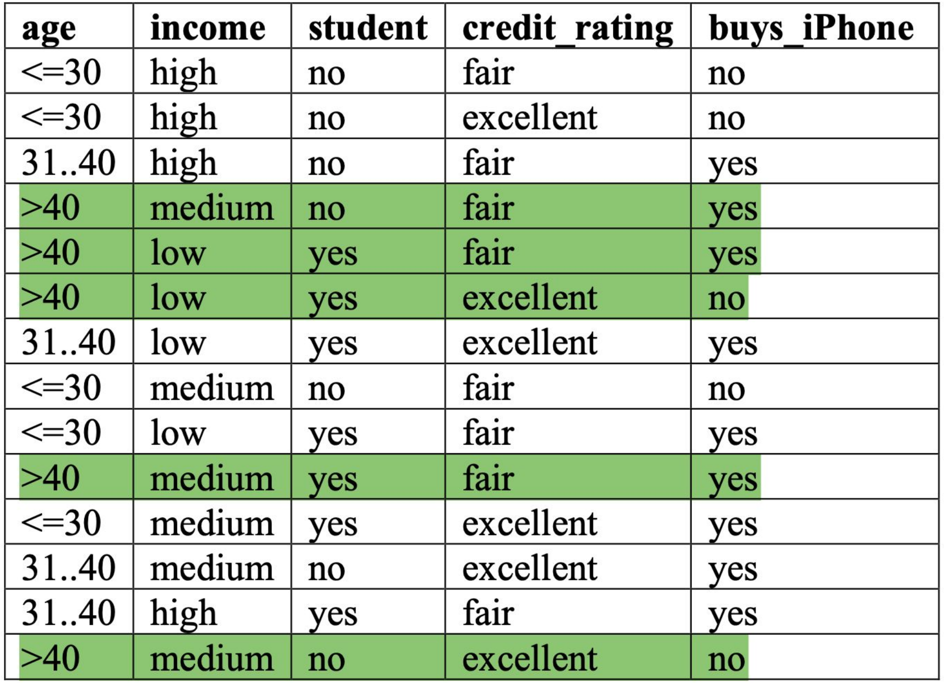


2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

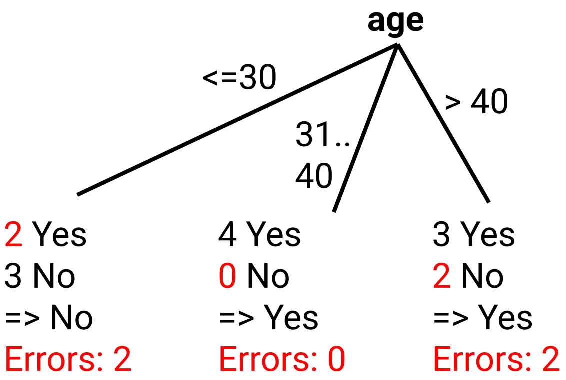
3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14



age	income	student	credit_rating	buys_iPhone
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise 3 – 1R



age

<=30

31..
40

> 40

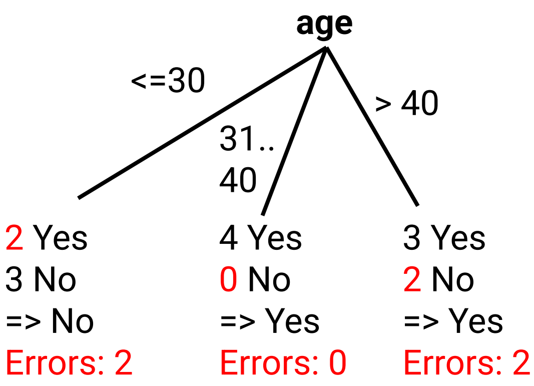
2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14

Your turn to do the same for *income, student* and *credit_rating*!



age

<=30

31..
40

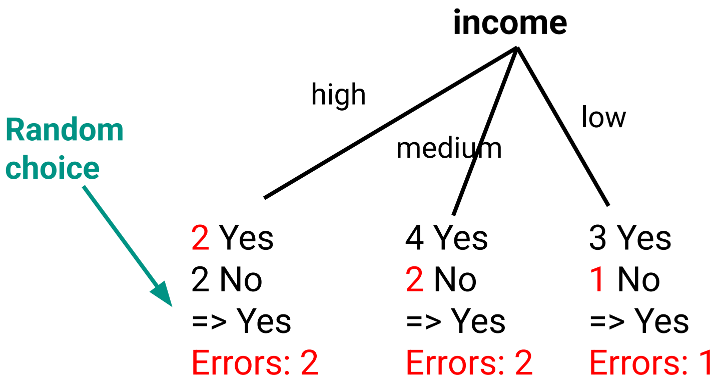
> 40

2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14



income

high

medium

low

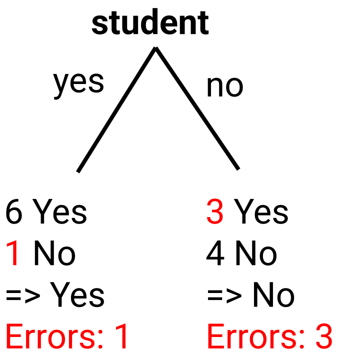
Random
choice

2 Yes
2 No
=> Yes
Errors: 2

4 Yes
2 No
=> Yes
Errors: 2

3 Yes
1 No
=> Yes
Errors: 1

Total errors: 5/14



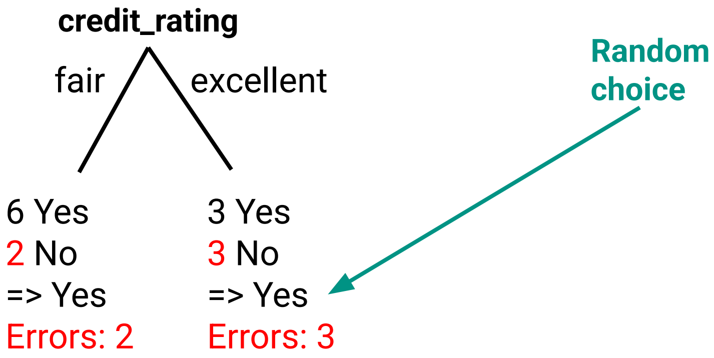
student

yes

no

6 Yes
1 No
=> Yes
Errors: 1

3 Yes
4 No
=> No
Errors: 3



credit_rating

fair

excellent

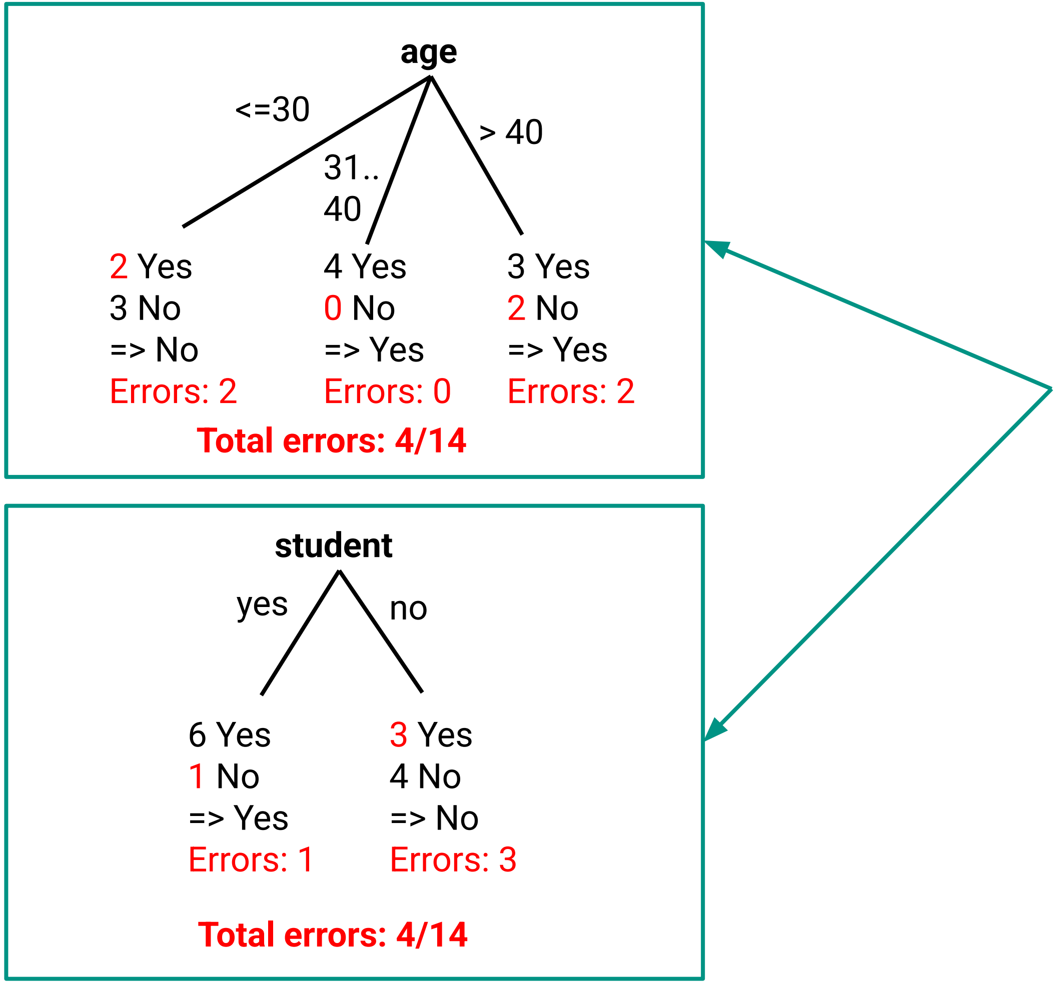
Random
choice

6 Yes
2 No
=> Yes
Errors: 2

3 Yes
3 No
=> Yes
Errors: 3

Total errors: 4/14

Total errors: 5/14



age

<=30

31..
40

> 40

2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14

student

yes

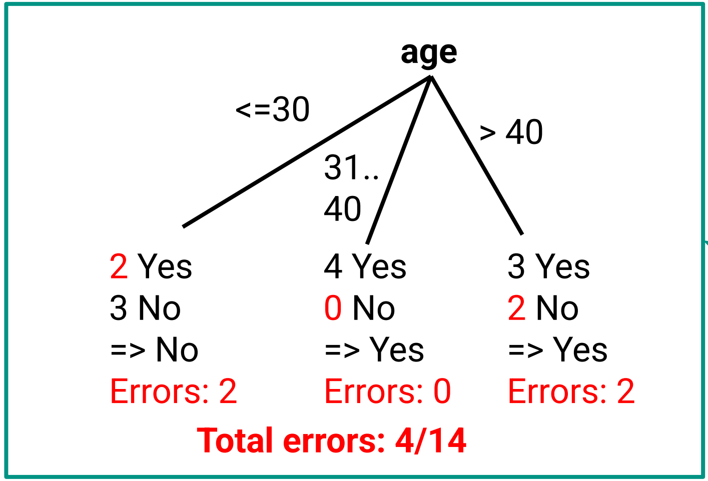
no

6 Yes
1 No
=> Yes
Errors: 1

3 Yes
4 No
=> No
Errors: 3

Total errors: 4/14

Both features give the minimum number of errors



age

<=30

31..
40

> 40

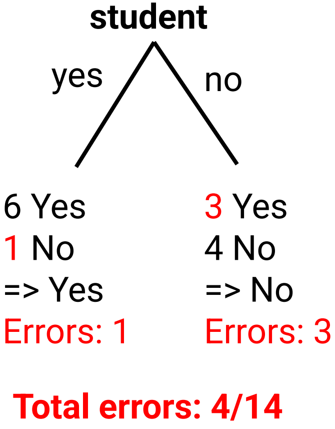
2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14

Let’s randomly pick *age* for 1R



student

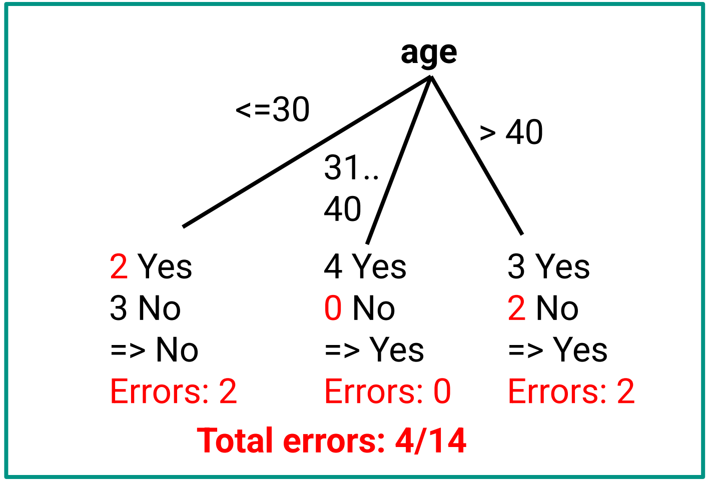
yes

no

6 Yes
1 No
=> Yes
Errors: 1

3 Yes
4 No
=> No
Errors: 3

Total errors: 4/14



age

<=30

31..
40

> 40

2 Yes
3 No
=> No
Errors: 2

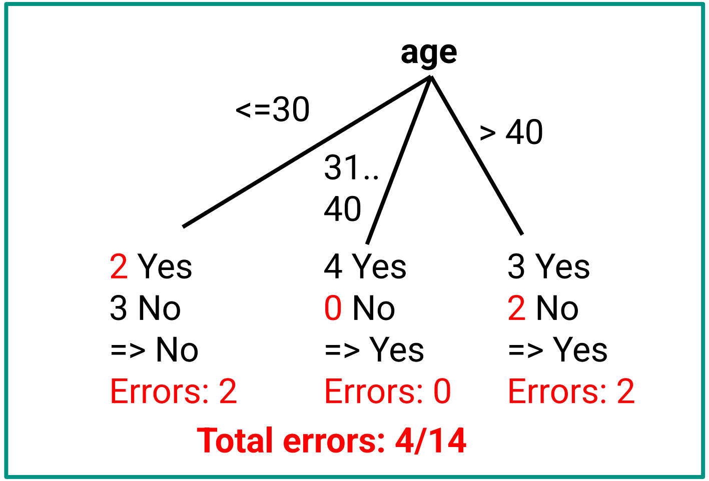
4 Yes
0 No
=> Yes
Errors: 0

3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14

1R produces the following rule:

if age <= 30 then buys_iPhone = No
Elif age = 31..40 then buys_iPhone = Yes
Elif age > 40 then buys_iPhone = Yes



age

<=30

31..
40

> 40

2 Yes
3 No
=> No
Errors: 2

4 Yes
0 No
=> Yes
Errors: 0

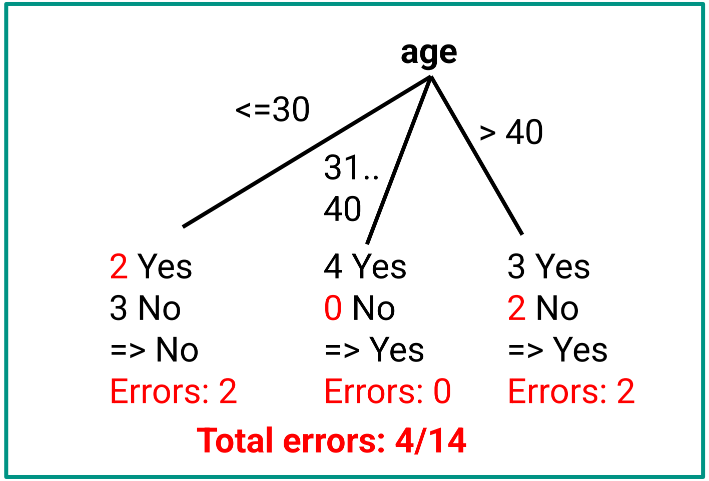
3 Yes
2 No
=> Yes
Errors: 2

Total errors: 4/14

1R produces the following rule:

if age <= 30 then buys_iPhone = No
Elif age = 31..40 then buys_iPhone = Yes
Elif age > 40 then buys_iPhone = Yes

New example:
age<=30, income=medium, student=yes, credit-rating=fair



1R produces the following rule:

if age <= 30 then buys_iPhone = No
Elif age = 31..40 then buys_iPhone = Yes
Elif age > 40 then buys_iPhone = Yes

New example:
age<=30, income=medium, student=yes, credit-rating=fair

=> Classified as buys_iPhone = No

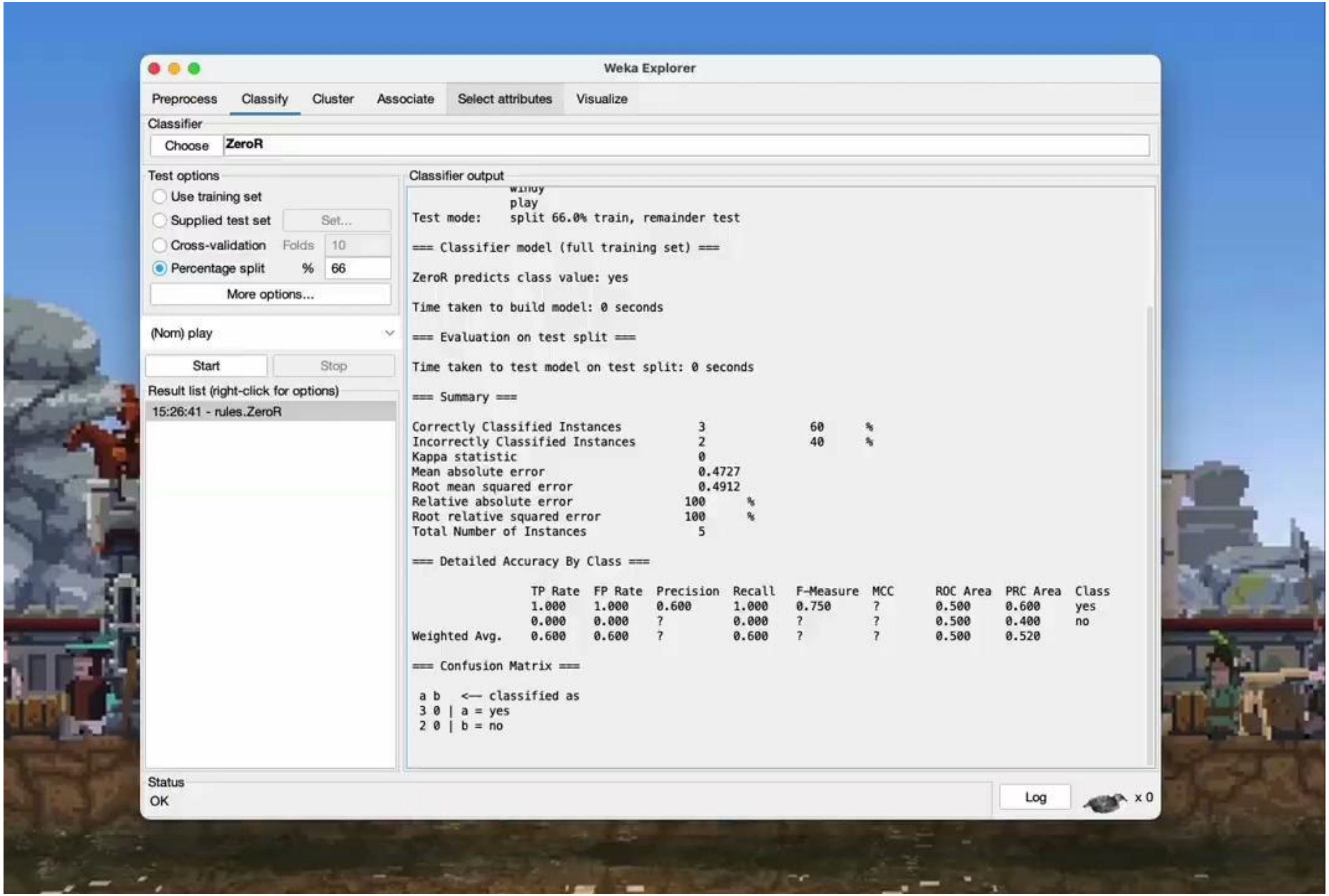
Exercise 4

Step 1-3

Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

Exercise 4

Step 4



Classifier: ZeroR

Test options:

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) play

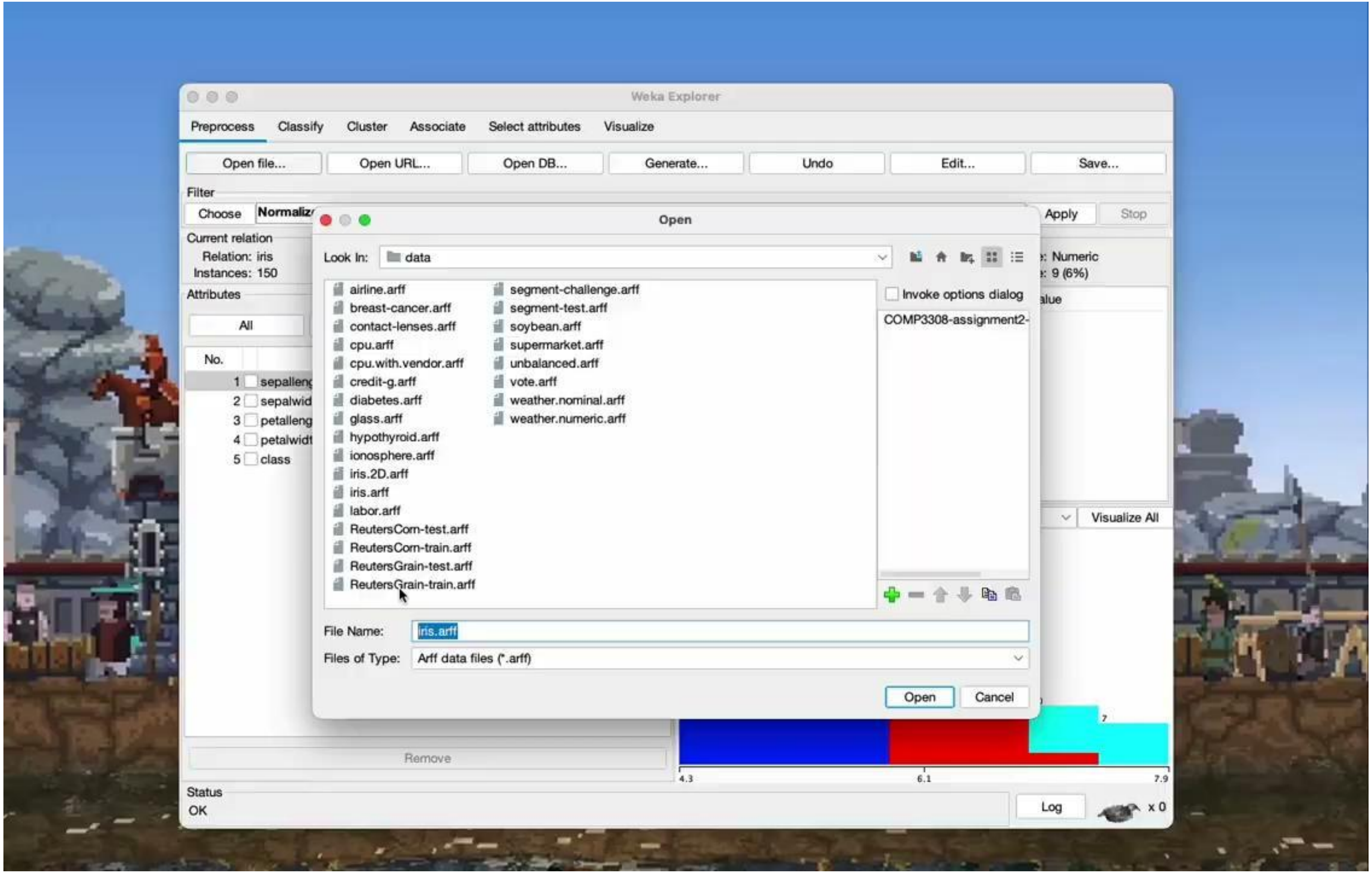
15:26:41 - rules.ZeroR

```
              windy
              play
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

ZeroR predicts class value: yes

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances           3               60      %
Incorrectly Classified Instances         2               40      %
Kappa statistic                          0
Mean absolute error                      0.4727
Root mean squared error                  0.4912
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.600      1.000    0.750      ?        0.500     0.600     yes
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.400     no
Weighted Avg.    0.600    0.600    ?          0.600    ?          ?        0.500     0.520

=== Confusion Matrix ===

 a b   <-- classified as
 3 0 | a = yes
 2 0 | b = no
```

Status: OK

Exercise 4

Step 6 (Normalisation)



Relation: iris
Instances: 150

: Numeric
: 9 (6%)
alue

No.	
1	□ sepalleng
2	□ sepalwid
3	□ petalleng
4	□ petalwidt
5	□ class

Open

Look In: data

airline.arff
breast-cancer.arff
contact-lenses.arff
cpu.arff
cpu.with.vendor.arff
credit-g.arff
diabetes.arff
glass.arff
hypothyroid.arff
ionosphere.arff
iris.2D.arff
iris.arff
labor.arff
ReutersCorn-test.arff
ReutersCorn-train.arff
ReutersGrain-test.arff
ReutersGrain-train.arff
segment-challenge.arff
segment-test.arff
soybean.arff
supermarket.arff
unbalanced.arff
vote.arff
weather.nominal.arff
weather.numeric.arff

□ Invoke options dialog

COMP3308-assignment2-

File Name: iris.arff

Files of Type: Arff data files (*.arff)