WEKA
The University
of Waikato

WEKA
The University
of Waikato

Data Jam

Workshop 2: Weka

Presenter: Mark Voortman
https://datajam.it.pointpark.edu/
(all materials downloadable)

data
jam

ROWLAND
SCHOOL OF BUSINESS
POINT PARK UNIVERSITY

Introduction – What is this Data Jam About?

(Big) Data

HR Attrition

who leaves voluntarily

Insights

Logistics – Important Dates

- Workshops
  - ~~February 20th (today) – Tableau (data visualization and exploration)~~
  - March 6th – Weka (predictive analytics software)
- Poster Competition
  - April 3rd – Poster Presentations (present your results!)
  - More details ~~next~~ *this* workshop

Logistics – Random Notes

- Team formation
- You can use any tool you want
  - We teach you Tableau and Weka
  - But feel free to use any other tool (Excel, Python, etc.)
- Judges
  - Industry professionals
  - Very experienced with data and modeling
  - Names, titles, and affiliations to be announced
- The Data Jam is co-organized with ITSO: http://itso.pointpark.edu/
  - Join ITSO if you like this kind of stuff

Slack – A Tool for Communication

- Slack is a popular **communication** tool used by many **tech** companies
- Go to https://pointparkuniversity.slack.com/ and join
- Use for
  - Reaching out to mentors with questions
  - Team collaboration
- Apps available for iOS, Android, etc.
- See next slides for steps and screen shots

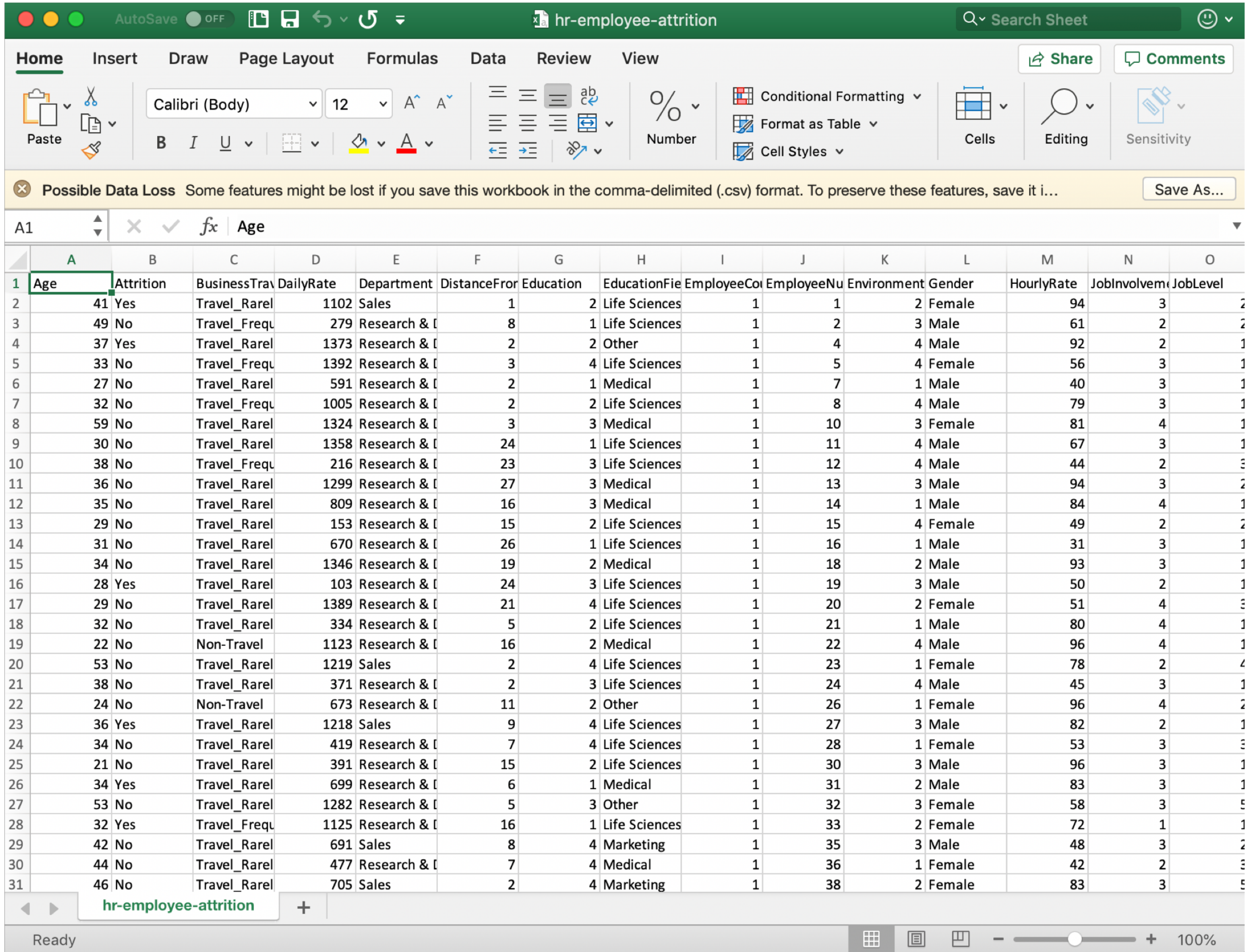
The Data – How Do I Obtain It?

Download from https://datajam.it.pointpark.edu/

hr-employee-attrition.csv

CSV format

What does that mean?



Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it i...

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Attrition	BusinessTrav	DailyRate	Department	DistanceFror	Education	EducationFie	EmployeeCo	EmployeeNu	Environment	Gender	HourlyRate	JobInvolveme	JobLevel
2	41	Yes	Travel_Rarel	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2
3	49	No	Travel_Frequ	279	Research & D	8	1	Life Sciences	1	2	3	Male	61	2	2
4	37	Yes	Travel_Rarel	1373	Research & D	2	2	Other	1	4	4	Male	92	2	1
5	33	No	Travel_Frequ	1392	Research & D	3	4	Life Sciences	1	5	4	Female	56	3	1
6	27	No	Travel_Rarel	591	Research & D	2	1	Medical	1	7	1	Male	40	3	1
7	32	No	Travel_Frequ	1005	Research & D	2	2	Life Sciences	1	8	4	Male	79	3	1
8	59	No	Travel_Rarel	1324	Research & D	3	3	Medical	1	10	3	Female	81	4	1
9	30	No	Travel_Rarel	1358	Research & D	24	1	Life Sciences	1	11	4	Male	67	3	1
10	38	No	Travel_Frequ	216	Research & D	23	3	Life Sciences	1	12	4	Male	44	2	3
11	36	No	Travel_Rarel	1299	Research & D	27	3	Medical	1	13	3	Male	94	3	2
12	35	No	Travel_Rarel	809	Research & D	16	3	Medical	1	14	1	Male	84	4	1
13	29	No	Travel_Rarel	153	Research & D	15	2	Life Sciences	1	15	4	Female	49	2	2
14	31	No	Travel_Rarel	670	Research & D	26	1	Life Sciences	1	16	1	Male	31	3	1
15	34	No	Travel_Rarel	1346	Research & D	19	2	Medical	1	18	2	Male	93	3	1
16	28	Yes	Travel_Rarel	103	Research & D	24	3	Life Sciences	1	19	3	Male	50	2	1
17	29	No	Travel_Rarel	1389	Research & D	21	4	Life Sciences	1	20	2	Female	51	4	3
18	32	No	Travel_Rarel	334	Research & D	5	2	Life Sciences	1	21	1	Male	80	4	1
19	22	No	Non-Travel	1123	Research & D	16	2	Medical	1	22	4	Male	96	4	1
20	53	No	Travel_Rarel	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	4
21	38	No	Travel_Rarel	371	Research & D	2	3	Life Sciences	1	24	4	Male	45	3	1
22	24	No	Non-Travel	673	Research & D	11	2	Other	1	26	1	Female	96	4	2
23	36	Yes	Travel_Rarel	1218	Sales	9	4	Life Sciences	1	27	3	Male	82	2	1
24	34	No	Travel_Rarel	419	Research & D	7	4	Life Sciences	1	28	1	Female	53	3	3
25	21	No	Travel_Rarel	391	Research & D	15	2	Life Sciences	1	30	3	Male	96	3	1
26	34	Yes	Travel_Rarel	699	Research & D	6	1	Medical	1	31	2	Male	83	3	1
27	53	No	Travel_Rarel	1282	Research & D	5	3	Other	1	32	3	Female	58	3	5
28	32	Yes	Travel_Frequ	1125	Research & D	16	1	Life Sciences	1	33	2	Female	72	1	1
29	42	No	Travel_Rarel	691	Sales	8	4	Marketing	1	35	3	Male	48	3	2
30	44	No	Travel_Rarel	477	Research & D	7	4	Medical	1	36	1	Female	42	2	3
31	46	No	Travel_Rarel	705	Sales	2	4	Marketing	1	38	2	Female	83	3	5

The Data – What is the Problem/Goal?

1. How well can you predict attrition based on other characteristics (e.g., age)?

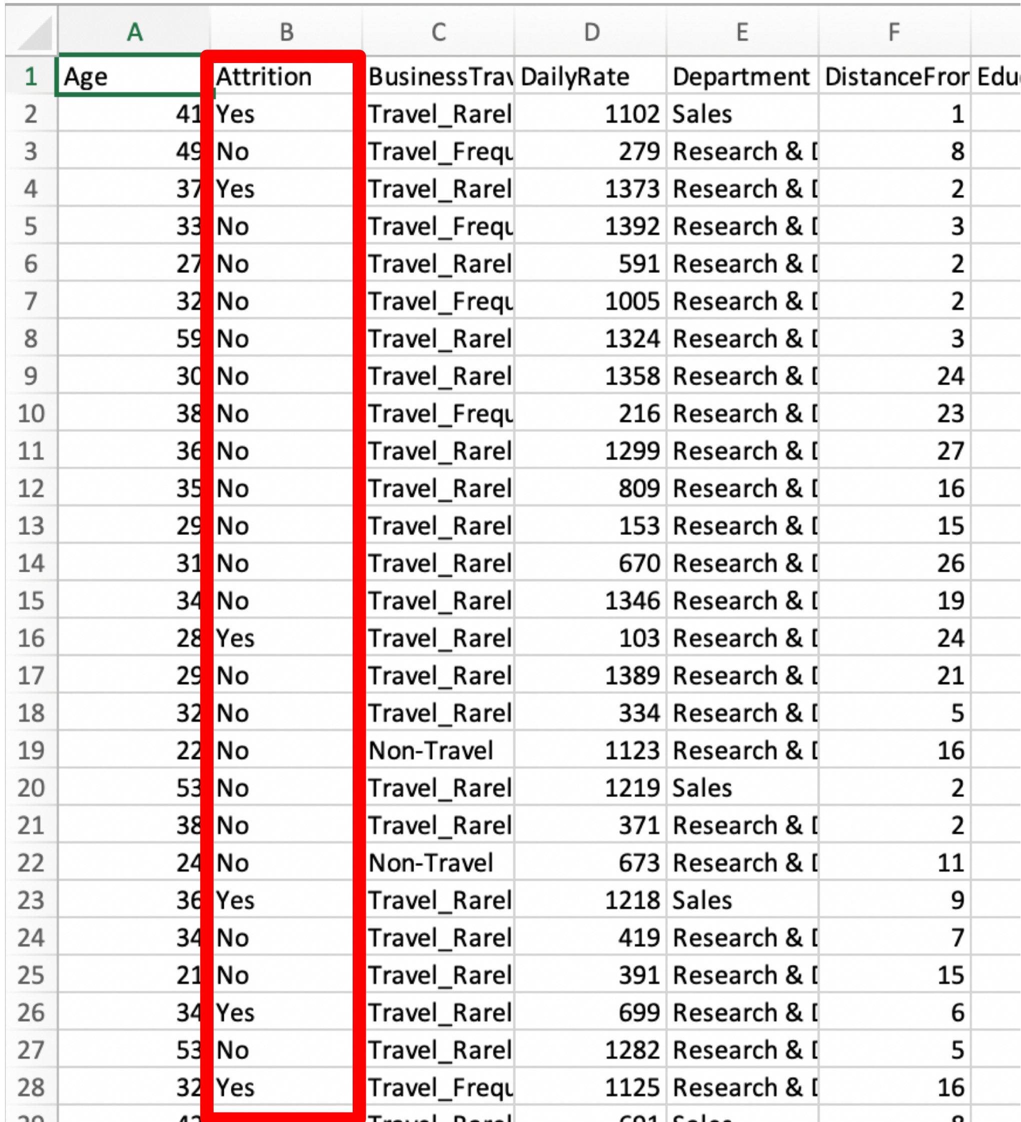
Build a model, e.g., if *age >= 65 => attrition=yes*

2. What drives attrition?

For example, age

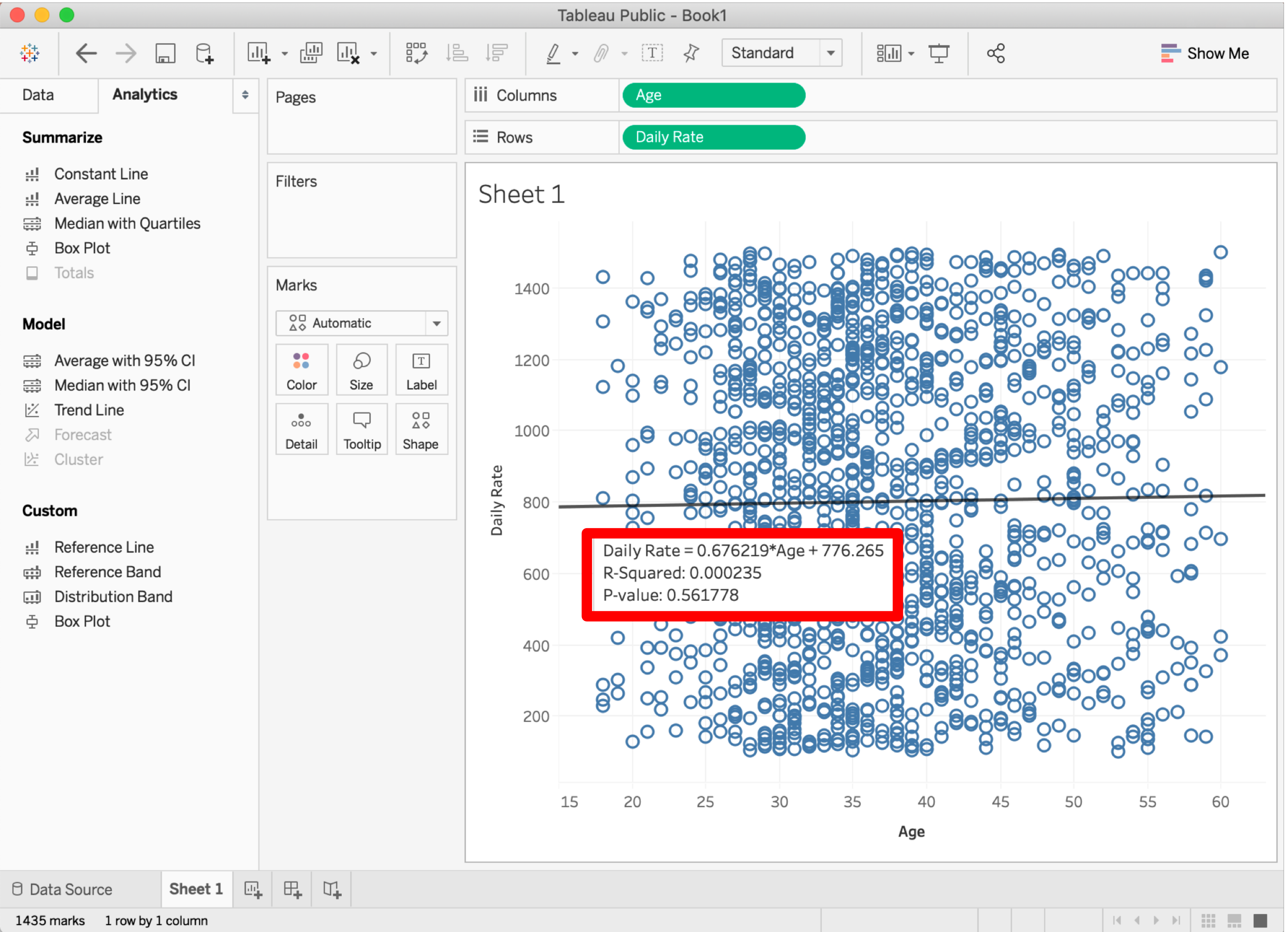
3. What other general insights can you obtain from the data?

E.g., what distinguishes high performers?



	A	B	C	D	E	F	
1	Age	Attrition	BusinessTrav	DailyRate	Department	DistanceFror	Edu
2	41	Yes	Travel_Rarel	1102	Sales	1	
3	49	No	Travel_Frequ	279	Research & [	8	
4	37	Yes	Travel_Rarel	1373	Research & [	2	
5	33	No	Travel_Frequ	1392	Research & [	3	
6	27	No	Travel_Rarel	591	Research & [	2	
7	32	No	Travel_Frequ	1005	Research & [	2	
8	59	No	Travel_Rarel	1324	Research & [	3	
9	30	No	Travel_Rarel	1358	Research & [	24	
10	38	No	Travel_Frequ	216	Research & [	23	
11	36	No	Travel_Rarel	1299	Research & [	27	
12	35	No	Travel_Rarel	809	Research & [	16	
13	29	No	Travel_Rarel	153	Research & [	15	
14	31	No	Travel_Rarel	670	Research & [	26	
15	34	No	Travel_Rarel	1346	Research & [	19	
16	28	Yes	Travel_Rarel	103	Research & [	24	
17	29	No	Travel_Rarel	1389	Research & [	21	
18	32	No	Travel_Rarel	334	Research & [	5	
19	22	No	Non-Travel	1123	Research & [	16	
20	53	No	Travel_Rarel	1219	Sales	2	
21	38	No	Travel_Rarel	371	Research & [	2	
22	24	No	Non-Travel	673	Research & [	11	
23	36	Yes	Travel_Rarel	1218	Sales	9	
24	34	No	Travel_Rarel	419	Research & [	7	
25	21	No	Travel_Rarel	391	Research & [	15	
26	34	Yes	Travel_Rarel	699	Research & [	6	
27	53	No	Travel_Rarel	1282	Research & [	5	
28	32	Yes	Travel_Frequ	1125	Research & [	16	
29	42		Travel_Rarel	691	Sales	8	

Tableau – A Scatter Plot



Daily Rate = 0.676219*Age + 776.265
R-Squared: 0.000235
P-value: 0.561778

Click on Trend Line, drag to the right, and drop on Linear

What do you think?

Any Questions at This Point?

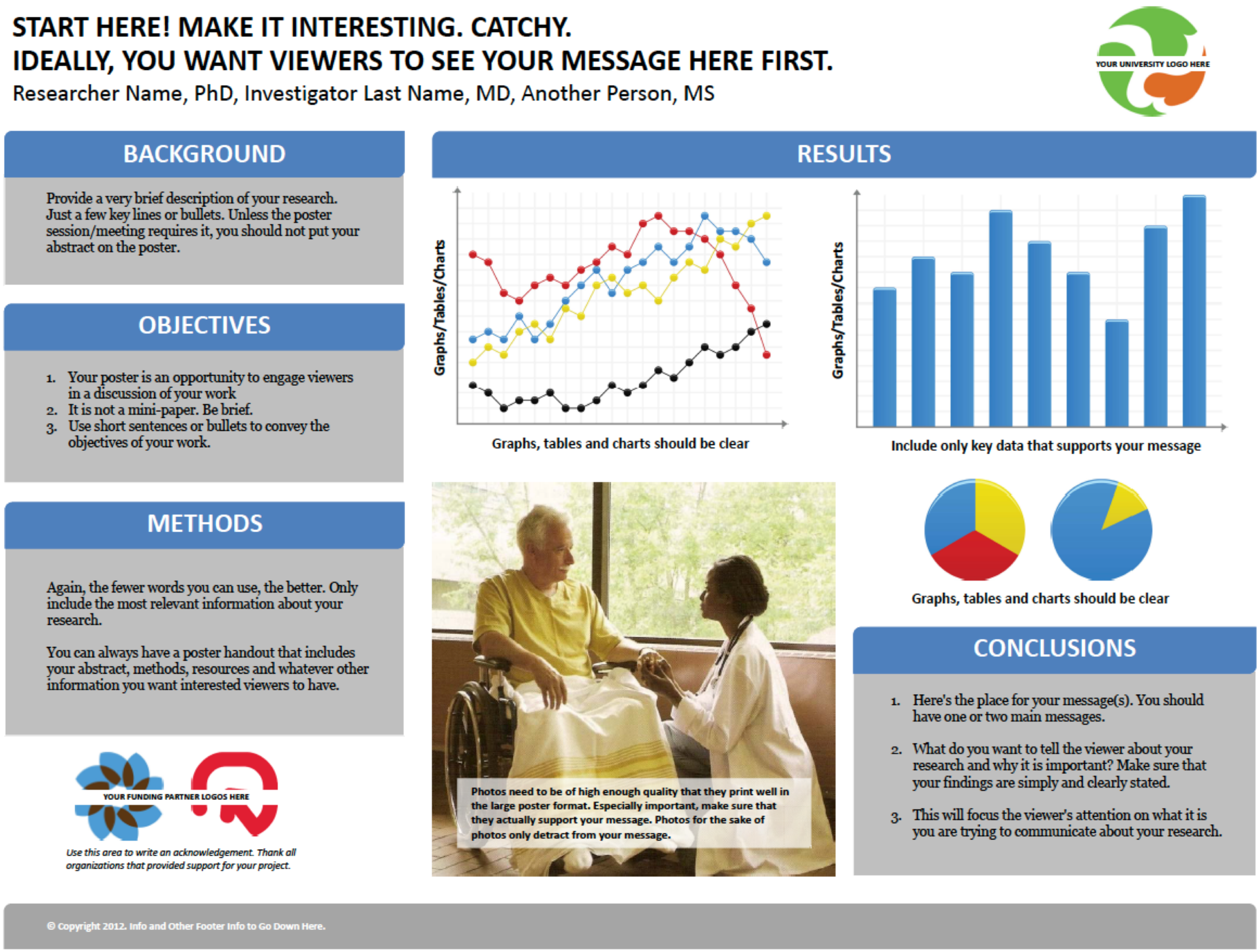
- Do you need help with anything?
- Were you able to create some visualizations with Tableau?
- Were you able to obtain some insights?
- …

Posters – Instructions

- Full instructions available at: https://datajam.it.pointpark.edu/data-jam-poster-guidelines.pdf
- Poster template available at: https://datajam.it.pointpark.edu/data-jam-poster-template.pptx
- Poster size: 24”x36”. All posters will be printed on foam board and displayed on easels (Data Jam team will coordinate printing).
- Email your PPT poster file to Jaime Ballesteros at jballesteros@pointpark.edu
- Submit by **Sunday, March 31 at 11:59 p.m.**

Posters – Guidelines

- **Project title**
- **Full names of all team members**
- **Include information and visuals that address the following:**
  - **Introduction** – State your team's key research question(s)… what are you trying to solve/uncover from the data and why is this relevant?
  - **Method(s) for data analysis** – Tactics to approach… how did you analyze the data?
  - **Results** – Include graphical visualizations of data and key findings. Add legends, captions, or BRIEF explanations if necessary.
  - **Analysis to Insights** – Clearly and concisely explain your findings (what you uncovered through your analysis).
  - **Conclusion** – Link back to your key research question(s) and summarize your impactful findings. Include your team's perspective on the impact of your findings and any recommendations. Also, share problems you encountered.



START HERE! MAKE IT INTERESTING. CATCHY. IDEALLY, YOU WANT VIEWERS TO SEE YOUR MESSAGE HERE FIRST.

Researcher Name, PhD, Investigator Last Name, MD, Another Person, MS

BACKGROUND

Provide a very brief description of your research. Just a few key lines or bullets. Unless the poster session/meeting requires it, you should not put your abstract on the poster.

OBJECTIVES

1. Your poster is an opportunity to engage viewers in a discussion of your work
2. It is not a mini-paper. Be brief.
3. Use short sentences or bullets to convey the objectives of your work.

METHODS

Again, the fewer words you can use, the better. Only include the most relevant information about your research.

You can always have a poster handout that includes your abstract, methods, resources and whatever other information you want interested viewers to have.

Use this area to write an acknowledgement. Thank all organizations that provided support for your project.

RESULTS

Graphs, tables and charts should be clear

Include only key data that supports your message

Photos need to be of high enough quality that they print well in the large poster format. Especially important, make sure that they actually support your message. Photos for the sake of photos only detract from your message.

Graphs, tables and charts should be clear

CONCLUSIONS

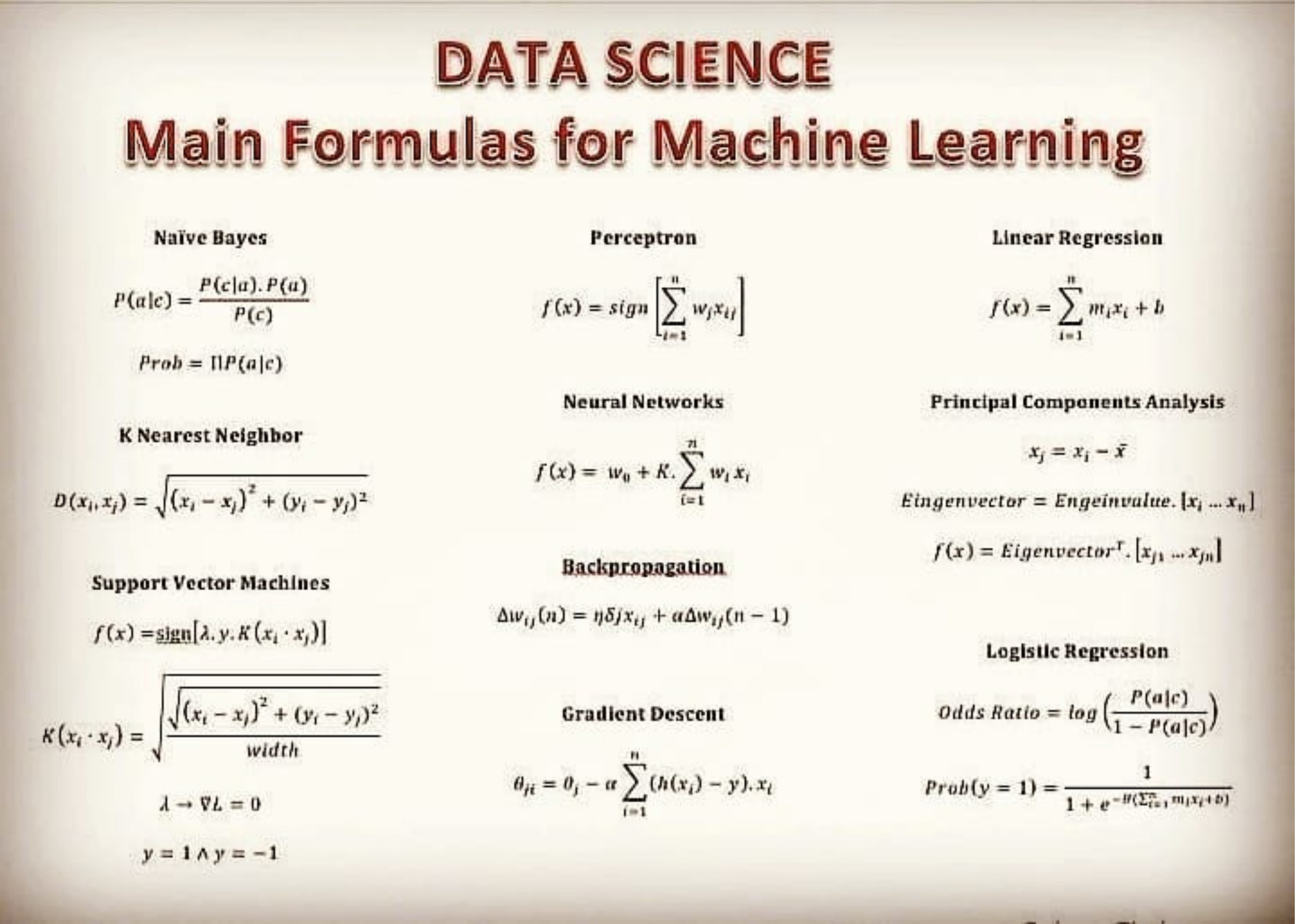
1. Here's the place for your message(s). You should have one or two main messages.
2. What do you want to tell the viewer about your research and why it is important? Make sure that your findings are simply and clearly stated.
3. This will focus the viewer's attention on what it is you are trying to communicate about your research.

© Copyright 2012. Info and Other Footer Info to Go Down Here.

Example Poster

Posters – Tips

- **Design your poster as a stand-alone artifact**. Be sure that you "tell the story" of your analysis and findings on the poster. Does it make stand-alone sense without someone there to explain it?
- **Include a brief but descriptive title**. People DO judge a book by its cover… the first thing people will read is your title, so consider your title an invitation to the audience. Your title should let the audience know what your poster is about in a brief sentence or phrase.
- **Emphasize graphics**. Convert information into graphical representations… charts, graphs, and images will capture attention and can effectively communicate data relationships.
- **Keep it clean**. Improve audience engagement and readability… avoid "chart junk" (information not required to understand the graphic), stick to a simple color palette (two to three colors max that don't detract from your content), use dark colors against a light background for better readability when lighting isn't ideal, and leave space between poster elements.



DATA SCIENCE
Main Formulas for Machine Learning

Naïve Bayes

$$P(a|c) = \frac{P(c|a).P(a)}{P(c)}$$

$$Prob = \Pi P(a|c)$$

K Nearest Neighbor

$$D(x_i, x_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Support Vector Machines

$$f(x) = \underline{sign}[\lambda. y. K(x_i \cdot x_j)]$$

$$K(x_i \cdot x_j) = \sqrt{\frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{width}}$$

$$\lambda \to \nabla L = 0$$

$$y = 1 \wedge y = -1$$

Perceptron

$$f(x) = sign\left[\sum_{i=1}^{n} w_j x_{ij}\right]$$

Neural Networks

$$f(x) = w_0 + K.\sum_{i=1}^{n} w_i x_i$$

Backpropagation

$$\Delta w_{ij}(n) = \eta \delta j x_{ij} + \alpha \Delta w_{ij}(n-1)$$

Gradient Descent

$$\theta_{ji} = \theta_j - \alpha \sum_{i=1}^{n} (h(x_i) - y).x_i$$

Linear Regression

$$f(x) = \sum_{i=1}^{n} m_i x_i + b$$

Principal Components Analysis

$$x_j = x_i - \bar{x}$$

$$Eingenvector = Engeinvalue.\,[x_i \dots x_n]$$

$$f(x) = Eigenvector^T.\,[x_{j1} \dots x_{jn}]$$

Logistic Regression

$$Odds\ Ratio = log\left(\frac{P(a|c)}{1 - P(a|c)}\right)$$

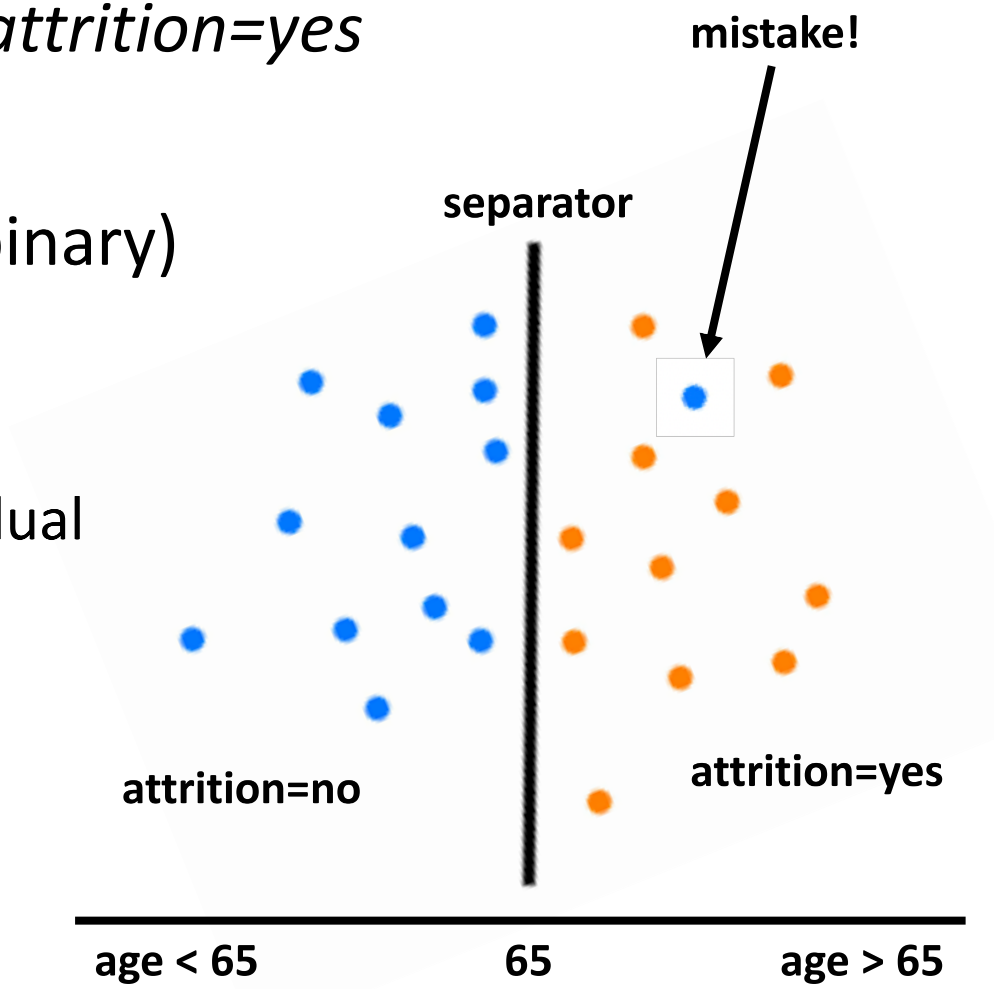
$$Prob(y = 1) = \frac{1}{1 + e^{-\theta(\sum_{i=1}^{n} m_i x_i + b)}}$$

Are you ready for some data mining?

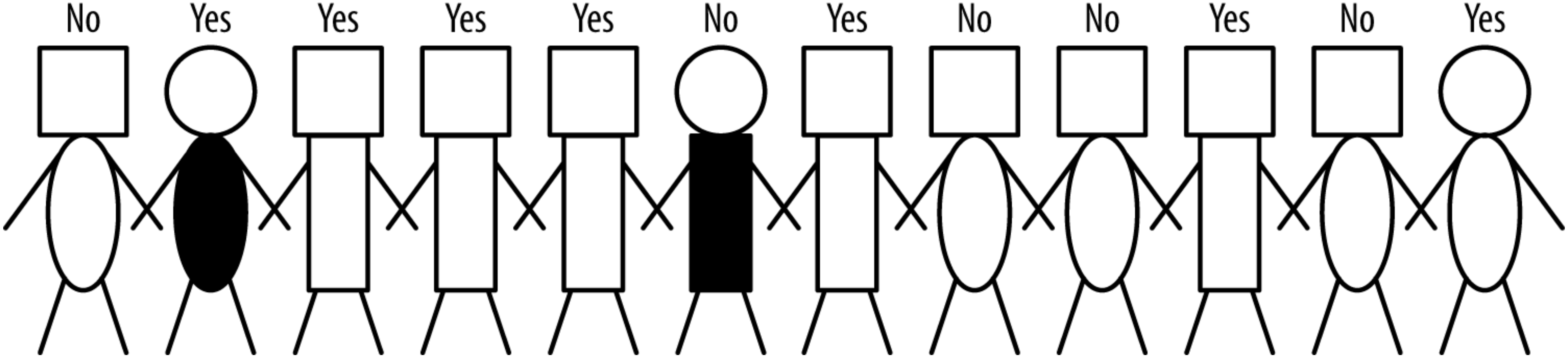
Data Mining – Introduction

- Remember this example? *age >= 65 => attrition=yes*
- This is known as a **classification** problem
- Classes: **attrition=yes** and **attrition=no** (binary)
  - Or **yes** and **no** for short
- Prediction is based on **features**
  - Capture the **key** characteristics of the individual
  - In the example above: **age**

Note: this slide and the next few are all relevant for Weka

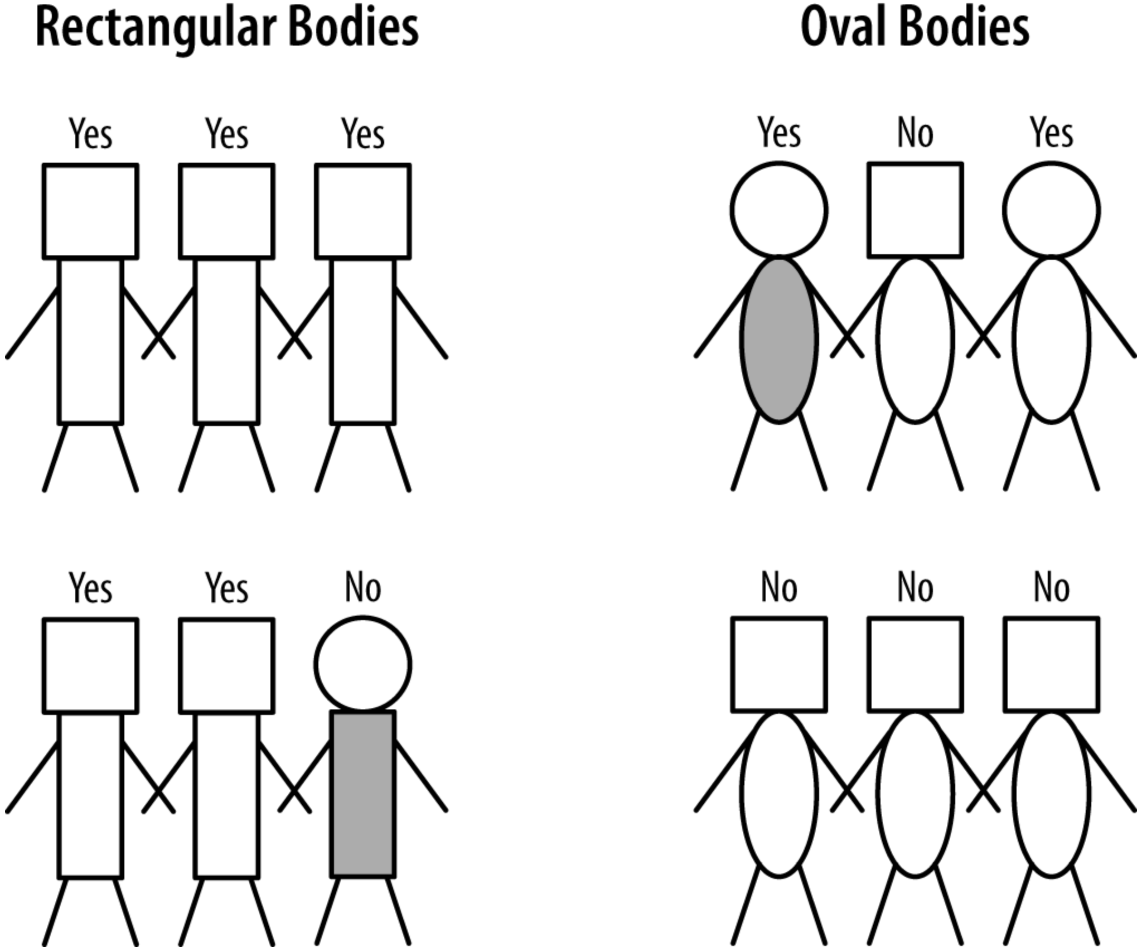


Data Mining – How Does It Work?



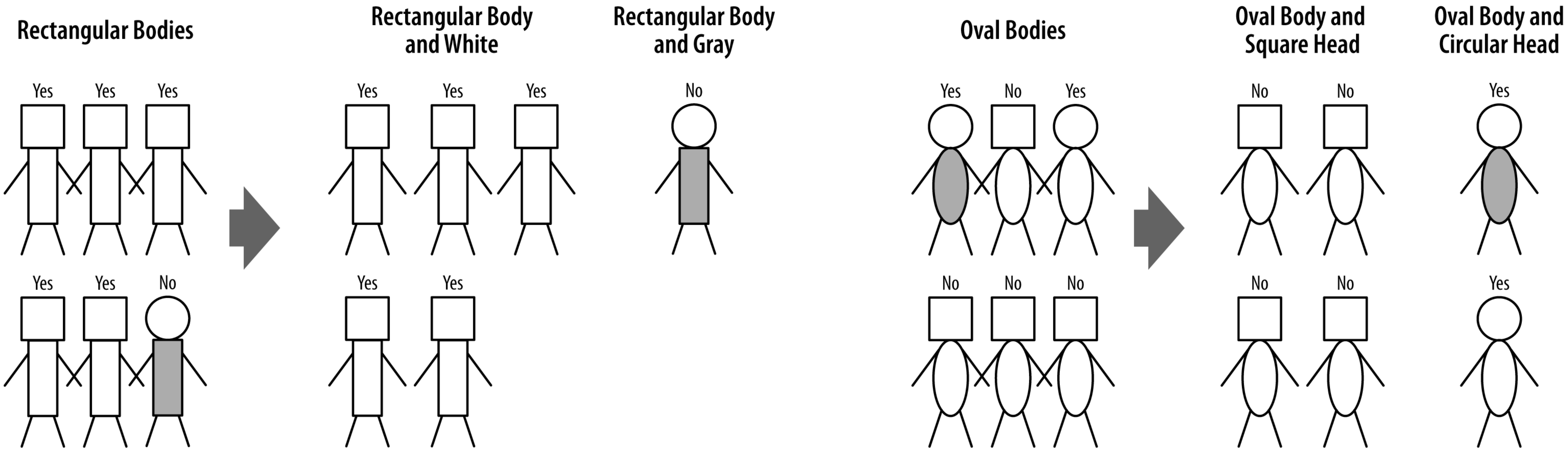
Which characteristics make someone likely to leave?

Data Mining – How Does It Work?



Rectangular bodies are much more likely to leave than oval bodies!

Data Mining – How Does It Work?

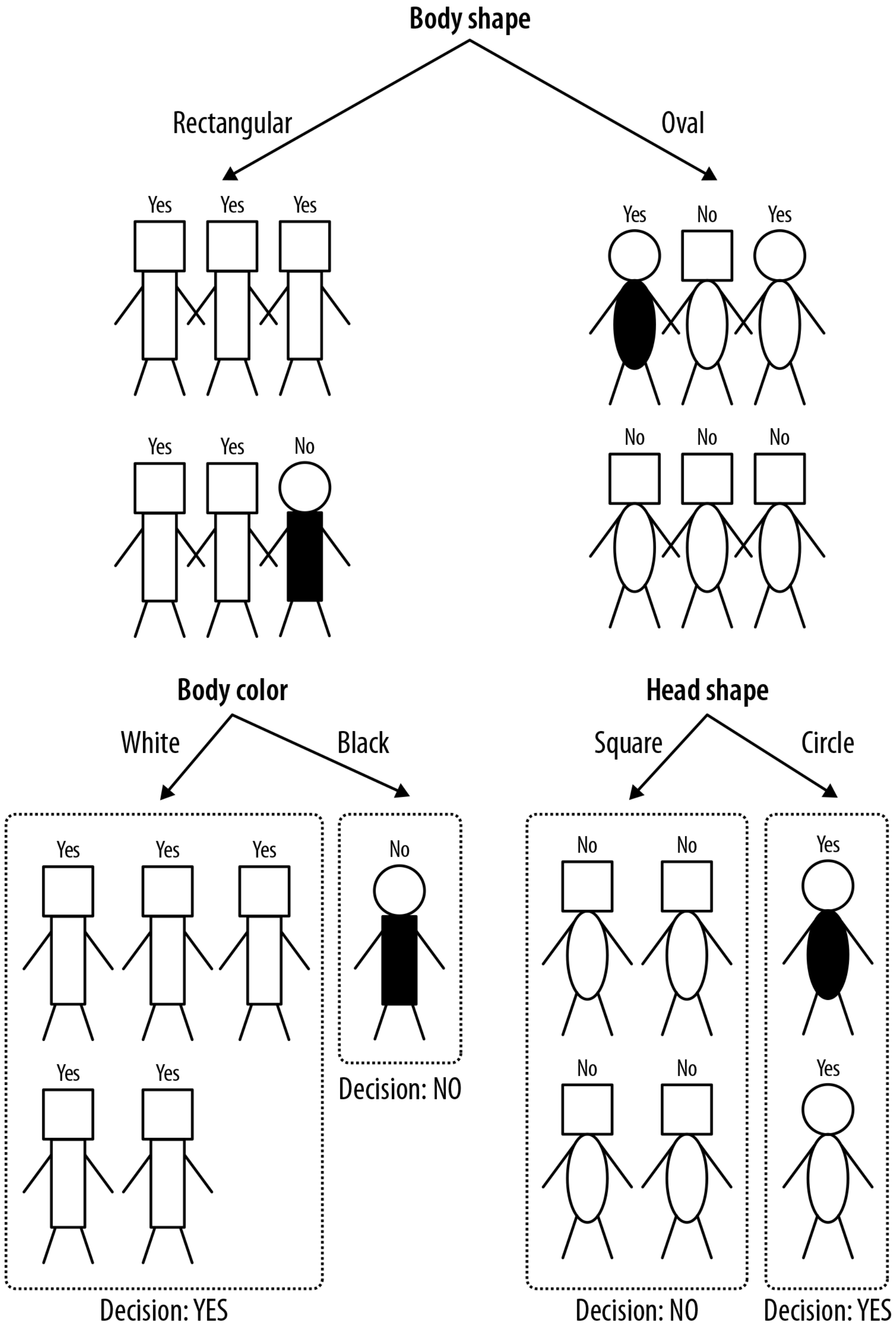

Rectangular Bodies
Yes
Yes
Yes
Yes
Yes
No
Rectangular Body and White
Yes
Yes
Yes
Yes
Yes
Rectangular Body and Gray
No
Oval Bodies
Yes
No
Yes
No
No
No
Oval Body and Square Head
No
No
No
No
Oval Body and Circular Head
Yes
Yes

Repeating the process!

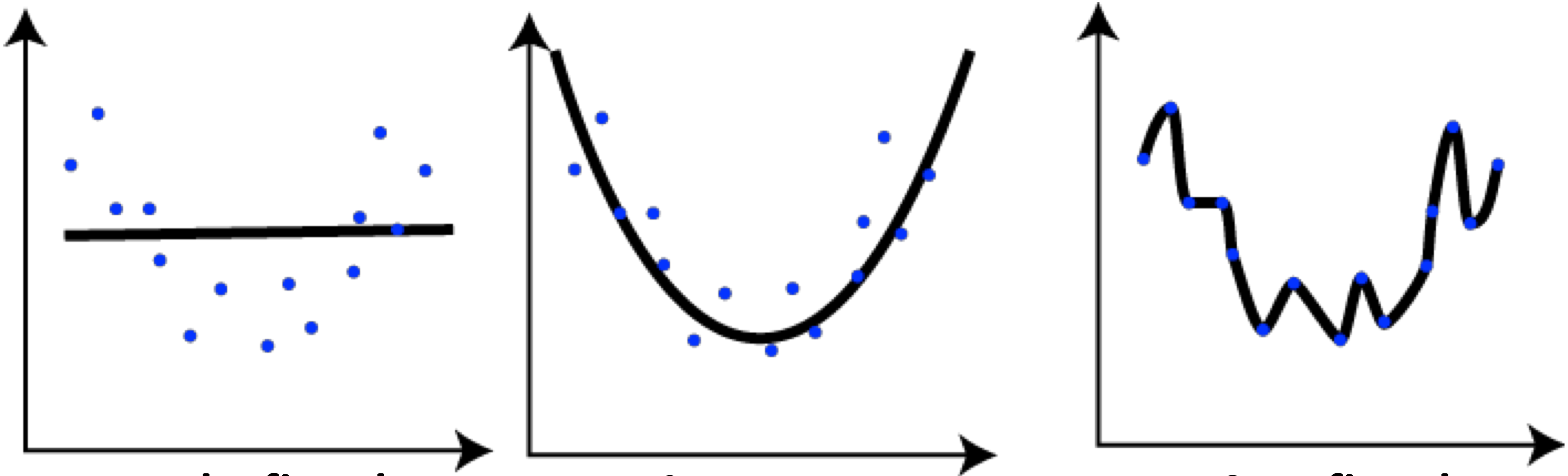
Data Mining – The Result

This is known as
a **decision tree** or,
more generally,
(supervised) segmentation

It can be created **automatically!**



Data Mining – Overfitting



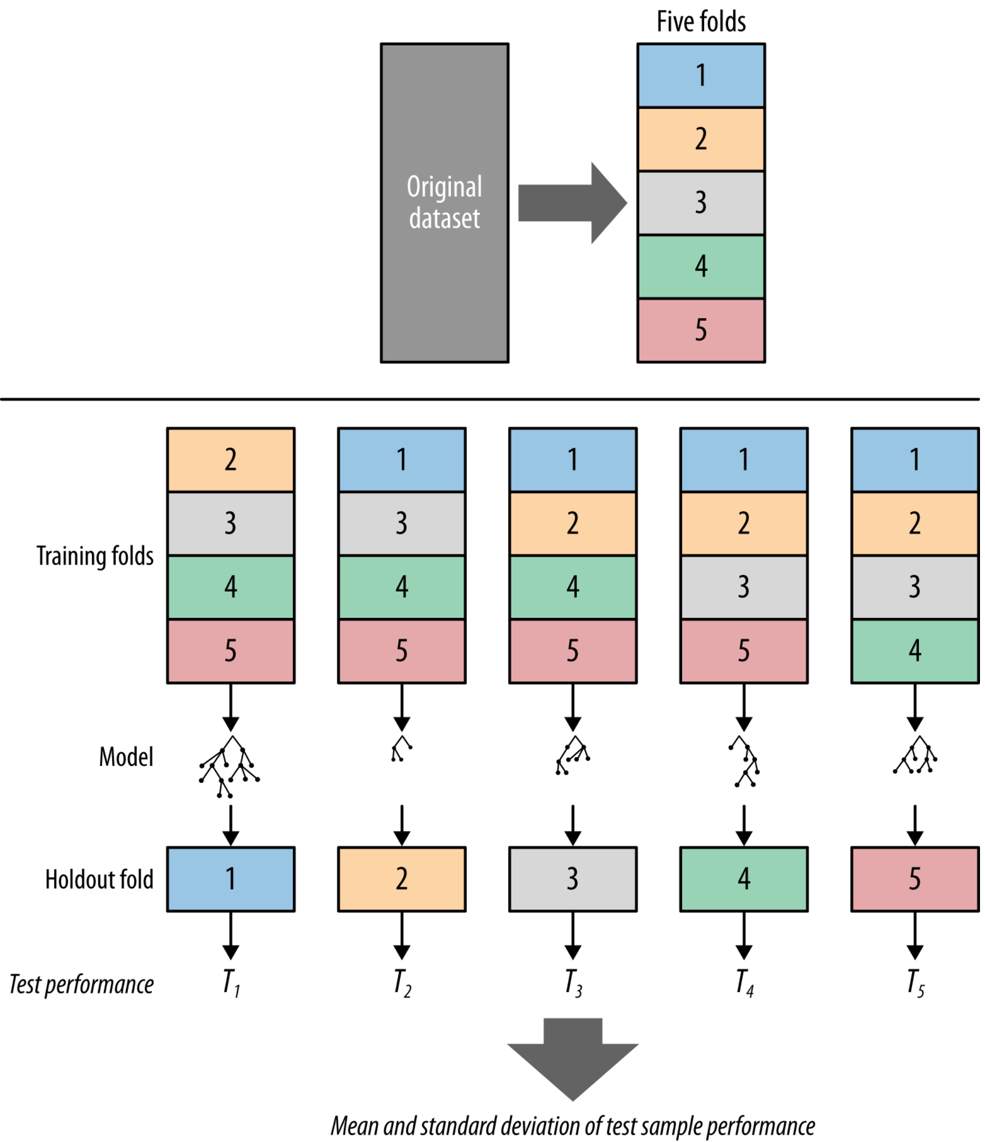
Underfitted

Sweet spot

Overfitted

We want a model that is not too general and not too specific!

Data Mining – Cross-Validation



Cross-validation prevents overfitting by using independent data to assess performance

Use 80% to train a model and use 20% to test and repeat 5 times

Weka does this for you ☺

Data Mining – Assessing Performance

Accuracy = #CorrectPredictions / #TotalPredictions

What should be the baseline to which you compare? 25%? 50%? 75%
What would you consider to be good accuracy?

Data Mining – Assessing Performance

Accuracy = #CorrectPredictions / #TotalPredictions

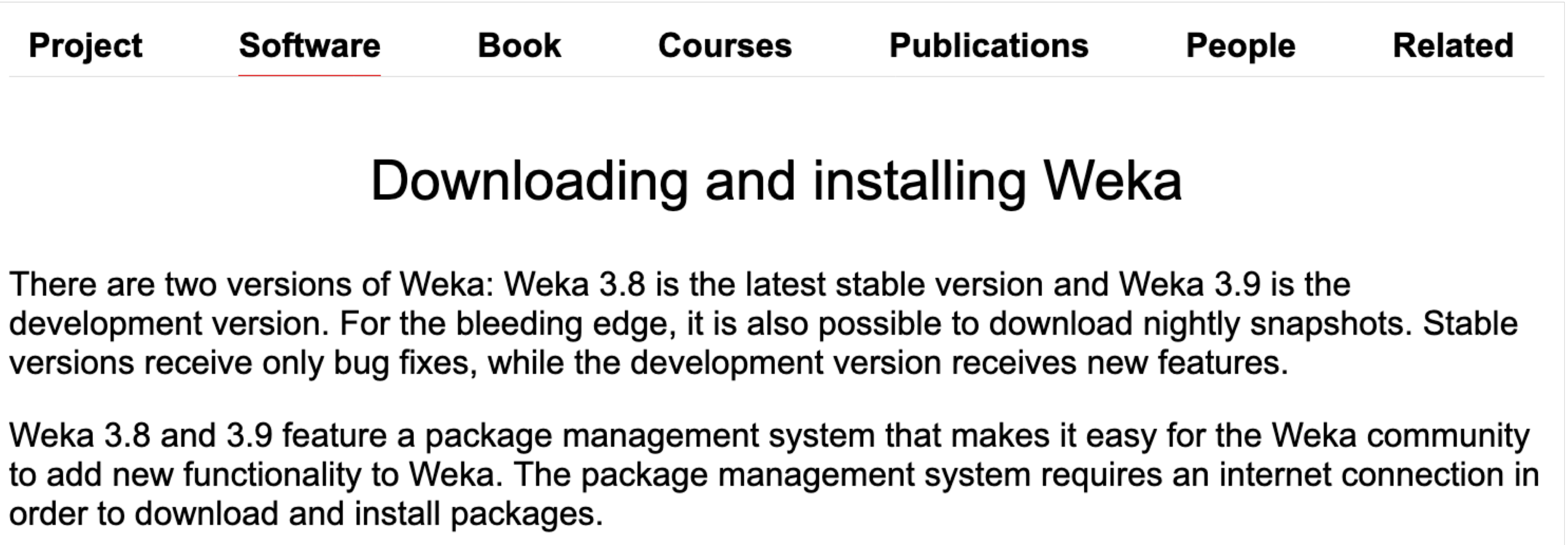
What should be the baseline to which you compare? 25%? 50%? 75%
What would you consider to be good accuracy?

Baseline:
1233/(1233+237)
= 83.9%

Weka – Downloading

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Get the version with the Java VM!

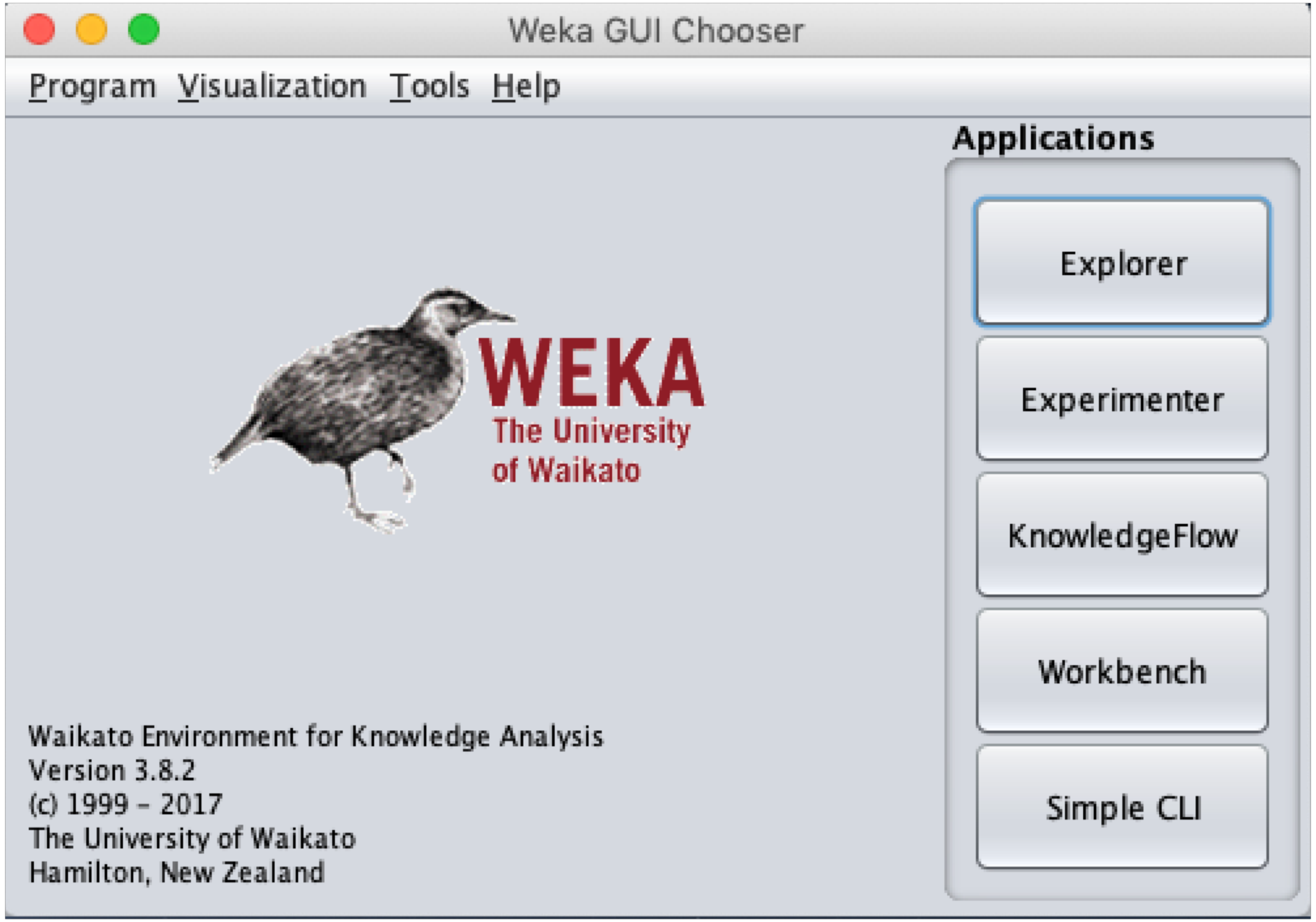


Downloading and installing Weka

There are two versions of Weka: Weka 3.8 is the latest stable version and Weka 3.9 is the development version. For the bleeding edge, it is also possible to download nightly snapshots. Stable versions receive only bug fixes, while the development version receives new features.

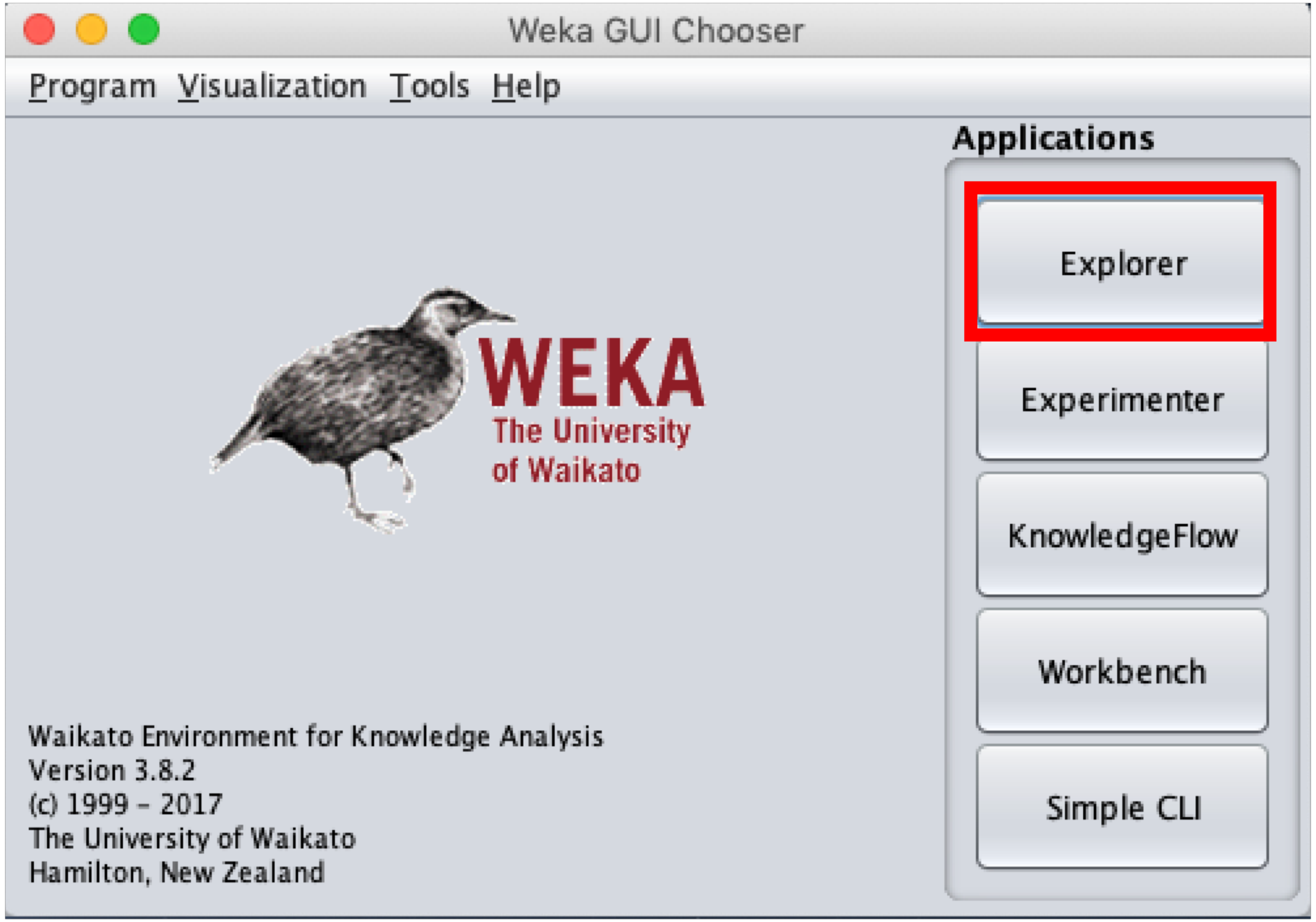
Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

Weka – Install and Run



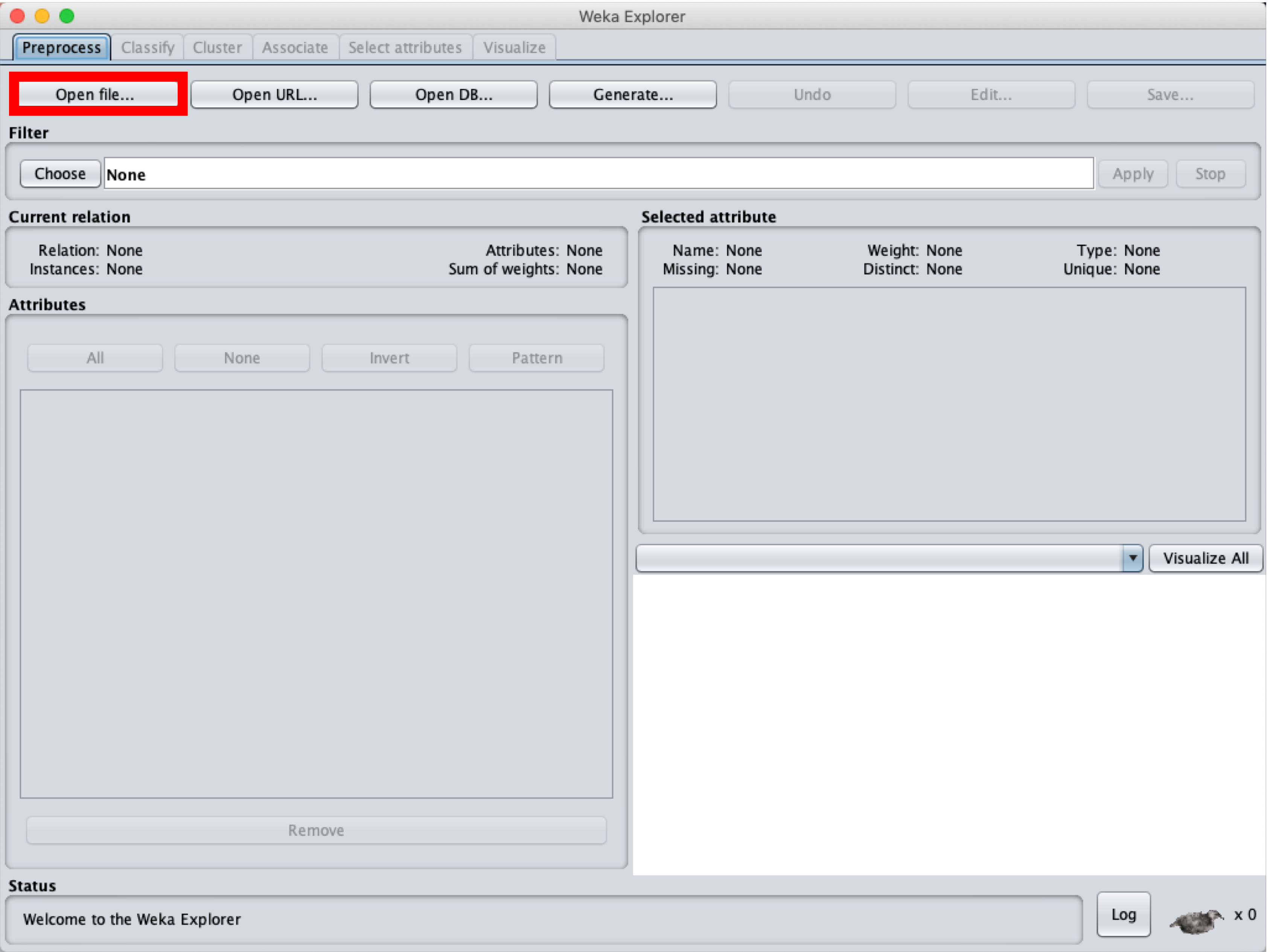
Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Weka – Click Explorer



Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Weka – Explorer



Filter

None

Current relation

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Attributes

Selected attribute

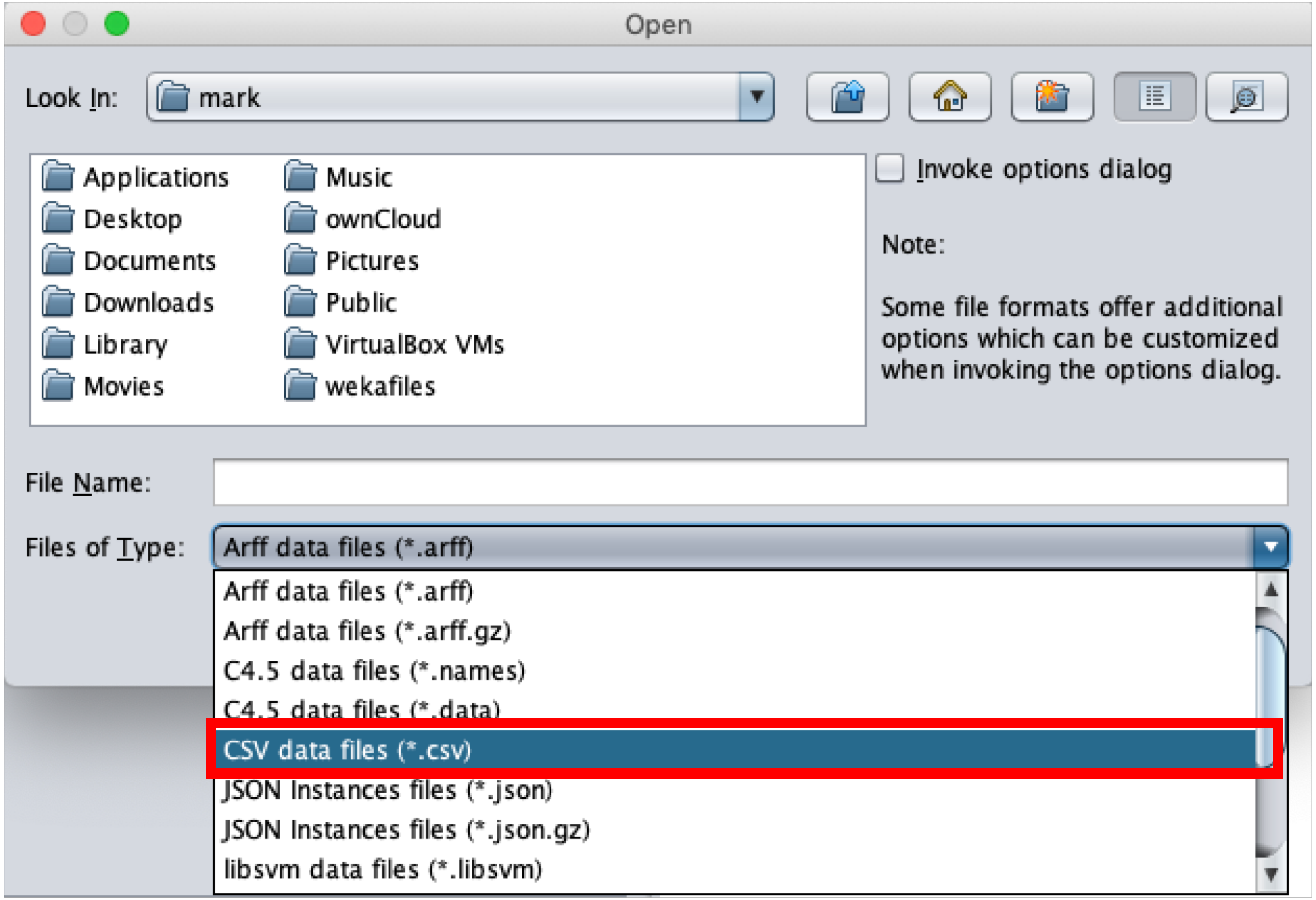
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Status

Welcome to the Weka Explorer

And now we load the data …

Weka – Loading Data



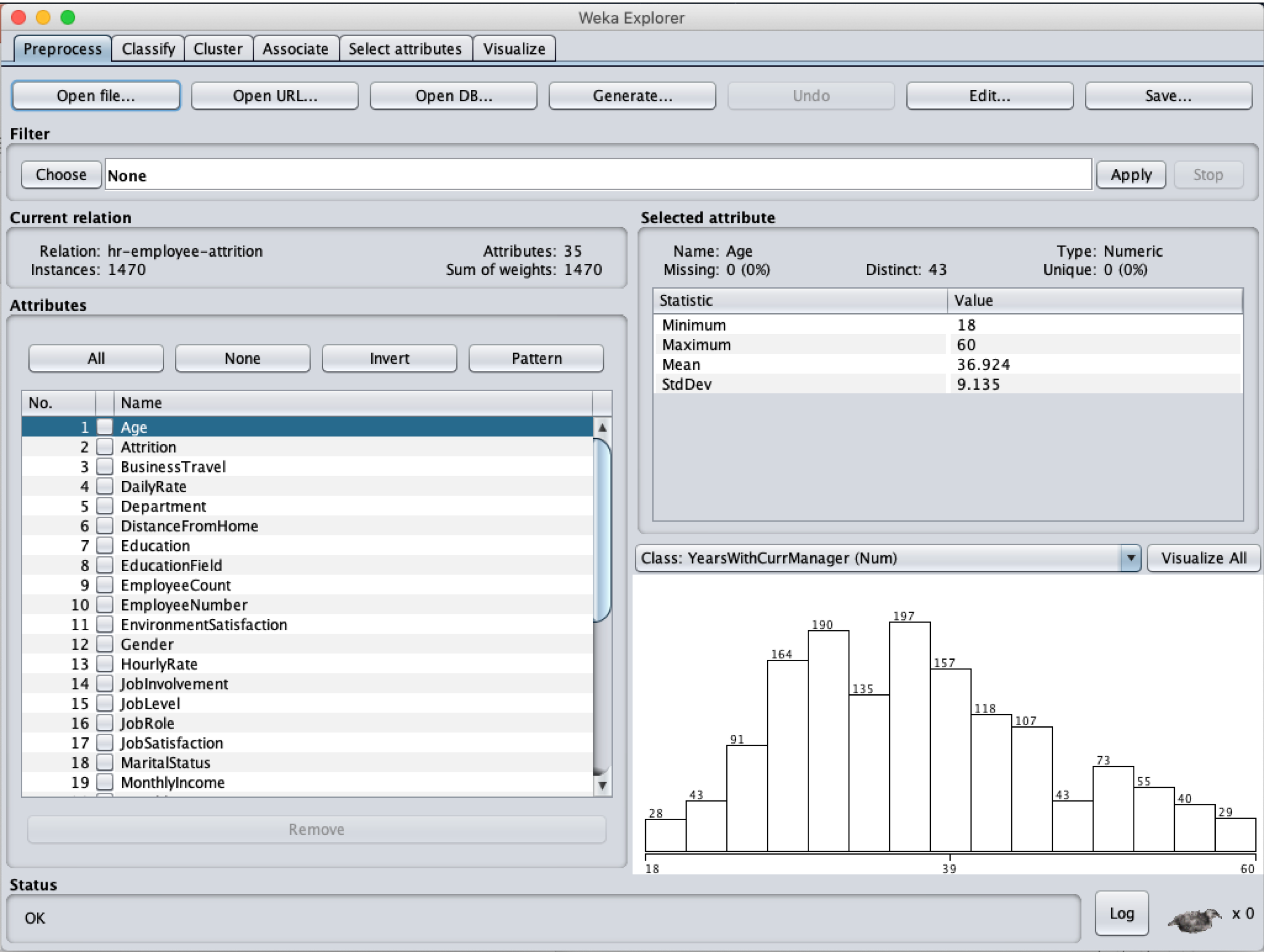
□ Invoke options dialog

Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

Make sure to select *.csv

Weka – Loading Data



Filter

None

Current relation

Relation: hr-employee-attrition
Instances: 1470
Attributes: 35
Sum of weights: 1470

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	60
Mean	36.924
StdDev	9.135

Attributes

No.	Name
1	☐ Age
2	☐ Attrition
3	☐ BusinessTravel
4	☐ DailyRate
5	☐ Department
6	☐ DistanceFromHome
7	☐ Education
8	☐ EducationField
9	☐ EmployeeCount
10	☐ EmployeeNumber
11	☐ EnvironmentSatisfaction
12	☐ Gender
13	☐ HourlyRate
14	☐ JobInvolvement
15	☐ JobLevel
16	☐ JobRole
17	☐ JobSatisfaction
18	☐ MaritalStatus
19	☐ MonthlyIncome

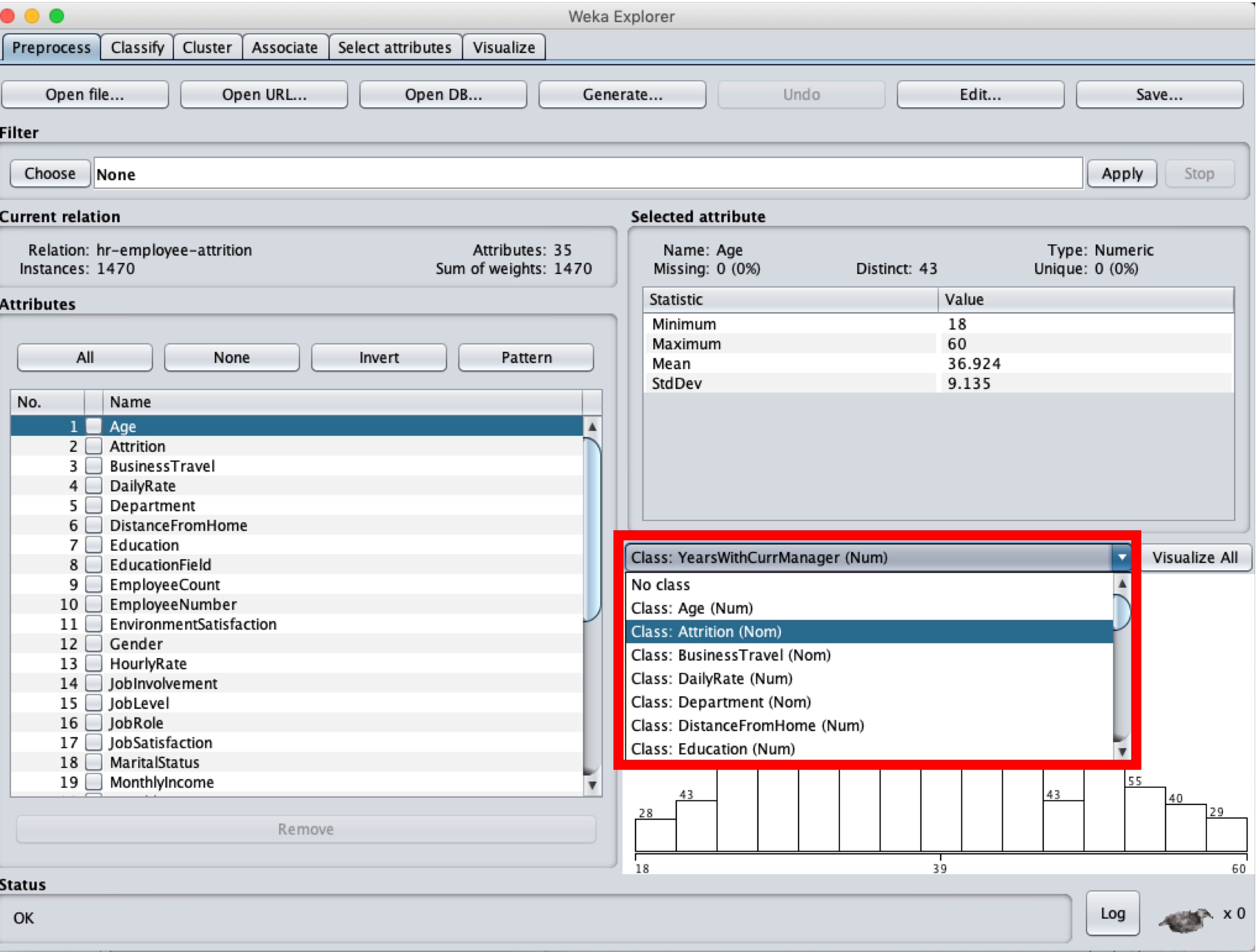
Class: YearsWithCurrManager (Num)

Status

OK

We now have our data loaded

Weka – Change Class to Attrition



Filter: None

Current relation

Relation: hr-employee-attrition
Instances: 1470
Attributes: 35
Sum of weights: 1470

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	60
Mean	36.924
StdDev	9.135

Attributes

No.		Name
1	□	Age
2	□	Attrition
3	□	BusinessTravel
4	□	DailyRate
5	□	Department
6	□	DistanceFromHome
7	□	Education
8	□	EducationField
9	□	EmployeeCount
10	□	EmployeeNumber
11	□	EnvironmentSatisfaction
12	□	Gender
13	□	HourlyRate
14	□	JobInvolvement
15	□	JobLevel
16	□	JobRole
17	□	JobSatisfaction
18	□	MaritalStatus
19	□	MonthlyIncome

Class: YearsWithCurrManager (Num)

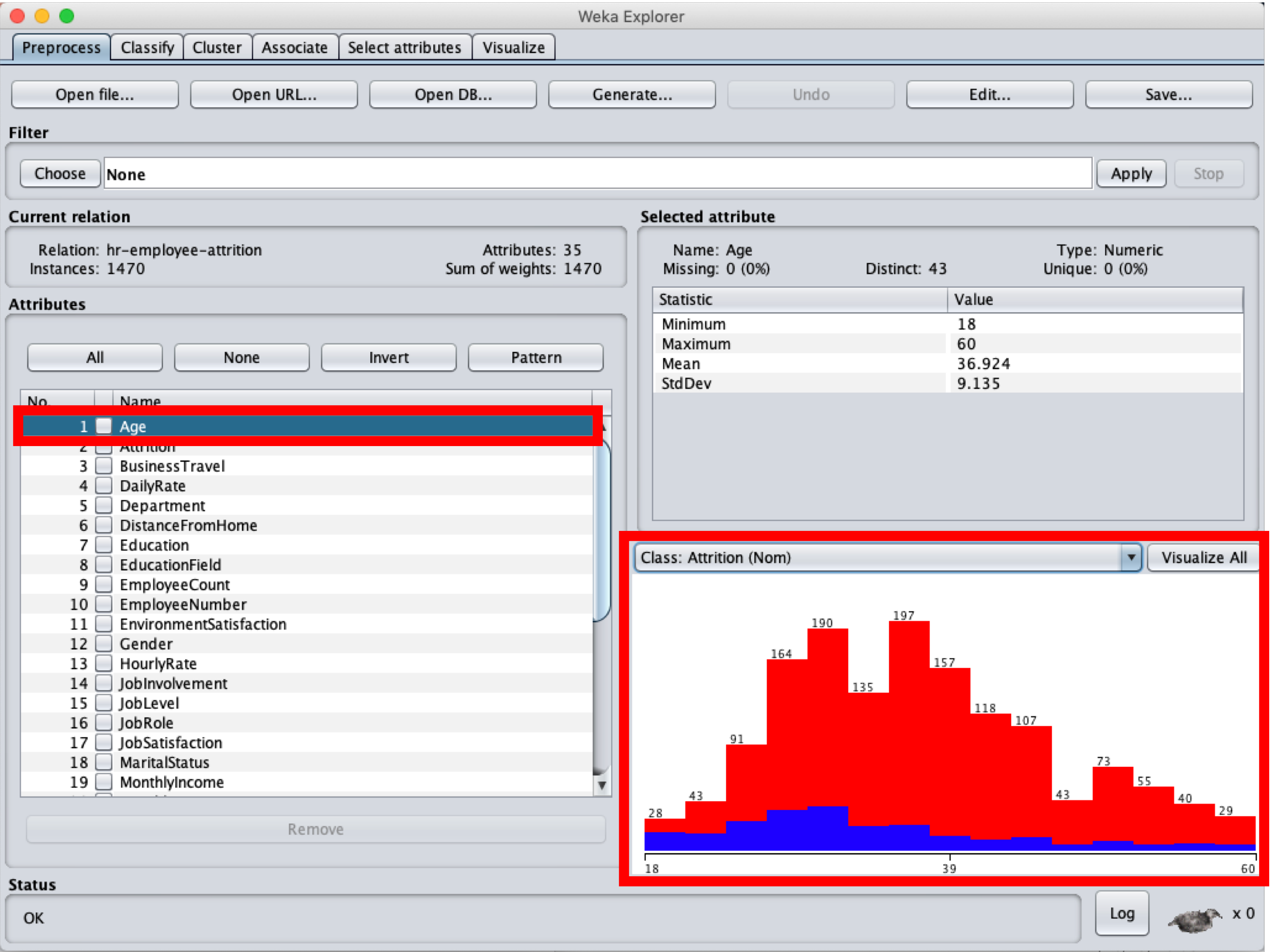
- No class
- Class: Age (Num)
- Class: Attrition (Nom)
- Class: BusinessTravel (Nom)
- Class: DailyRate (Num)
- Class: Department (Nom)
- Class: DistanceFromHome (Num)
- Class: Education (Num)

Status

OK

Click on the dropdown and change to Attrition

Weka – (Again) Young People Leave!



Filter

None

Current relation

Relation: hr-employee-attrition
Instances: 1470
Attributes: 35
Sum of weights: 1470

Attributes

No	Name
1	Age
2	Attrition
3	BusinessTravel
4	DailyRate
5	Department
6	DistanceFromHome
7	Education
8	EducationField
9	EmployeeCount
10	EmployeeNumber
11	EnvironmentSatisfaction
12	Gender
13	HourlyRate
14	JobInvolvement
15	JobLevel
16	JobRole
17	JobSatisfaction
18	MaritalStatus
19	MonthlyIncome

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	60
Mean	36.924
StdDev	9.135

Class: Attrition (Nom)

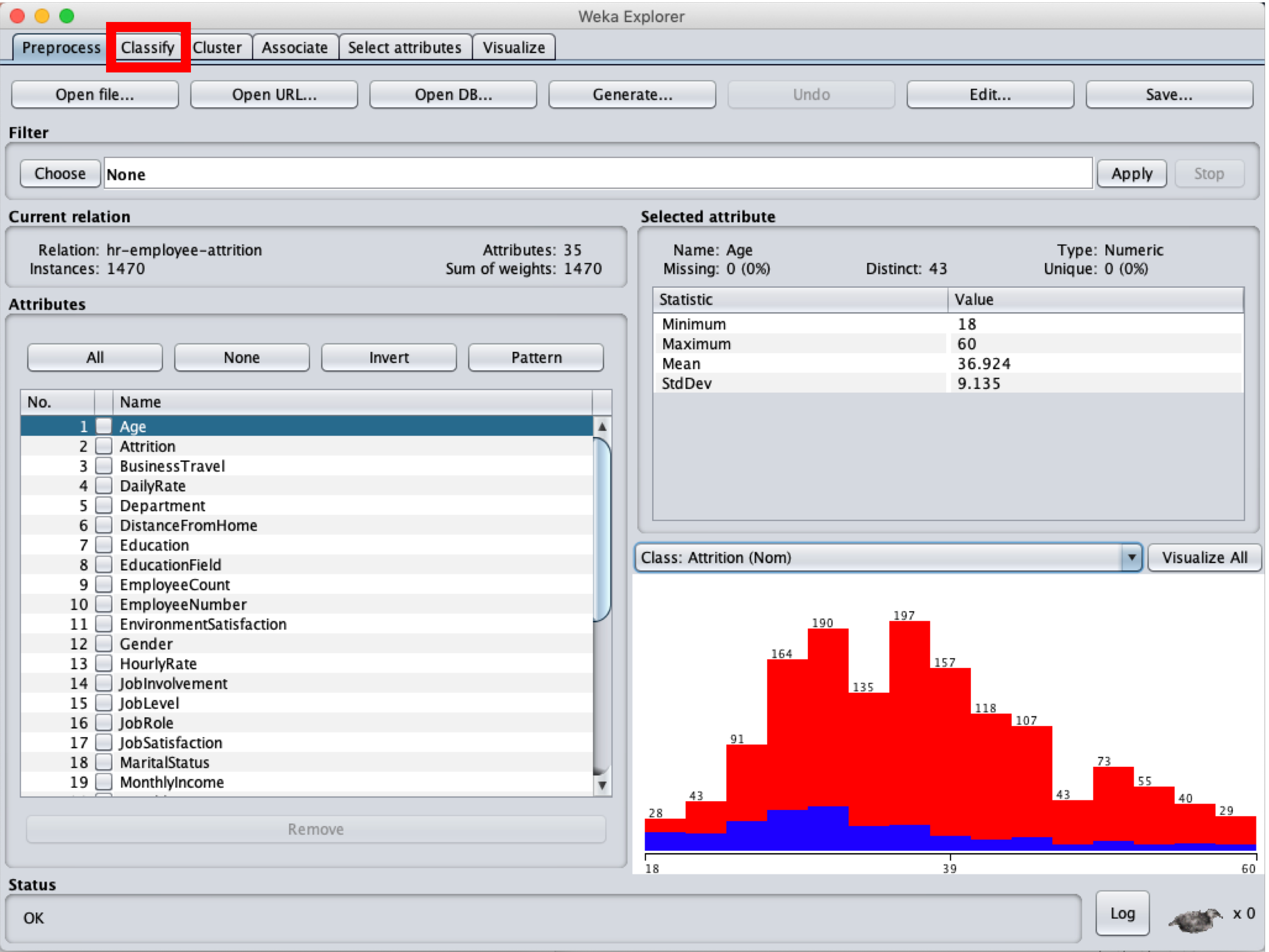
Status

OK

Red = stay

Blue = leave

Weka – Let's Build a Model!



Filter

None

Current relation

Relation: hr-employee-attrition
Instances: 1470
Attributes: 35
Sum of weights: 1470

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 43
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	60
Mean	36.924
StdDev	9.135

Attributes

No.	Name
1	Age
2	Attrition
3	BusinessTravel
4	DailyRate
5	Department
6	DistanceFromHome
7	Education
8	EducationField
9	EmployeeCount
10	EmployeeNumber
11	EnvironmentSatisfaction
12	Gender
13	HourlyRate
14	JobInvolvement
15	JobLevel
16	JobRole
17	JobSatisfaction
18	MaritalStatus
19	MonthlyIncome

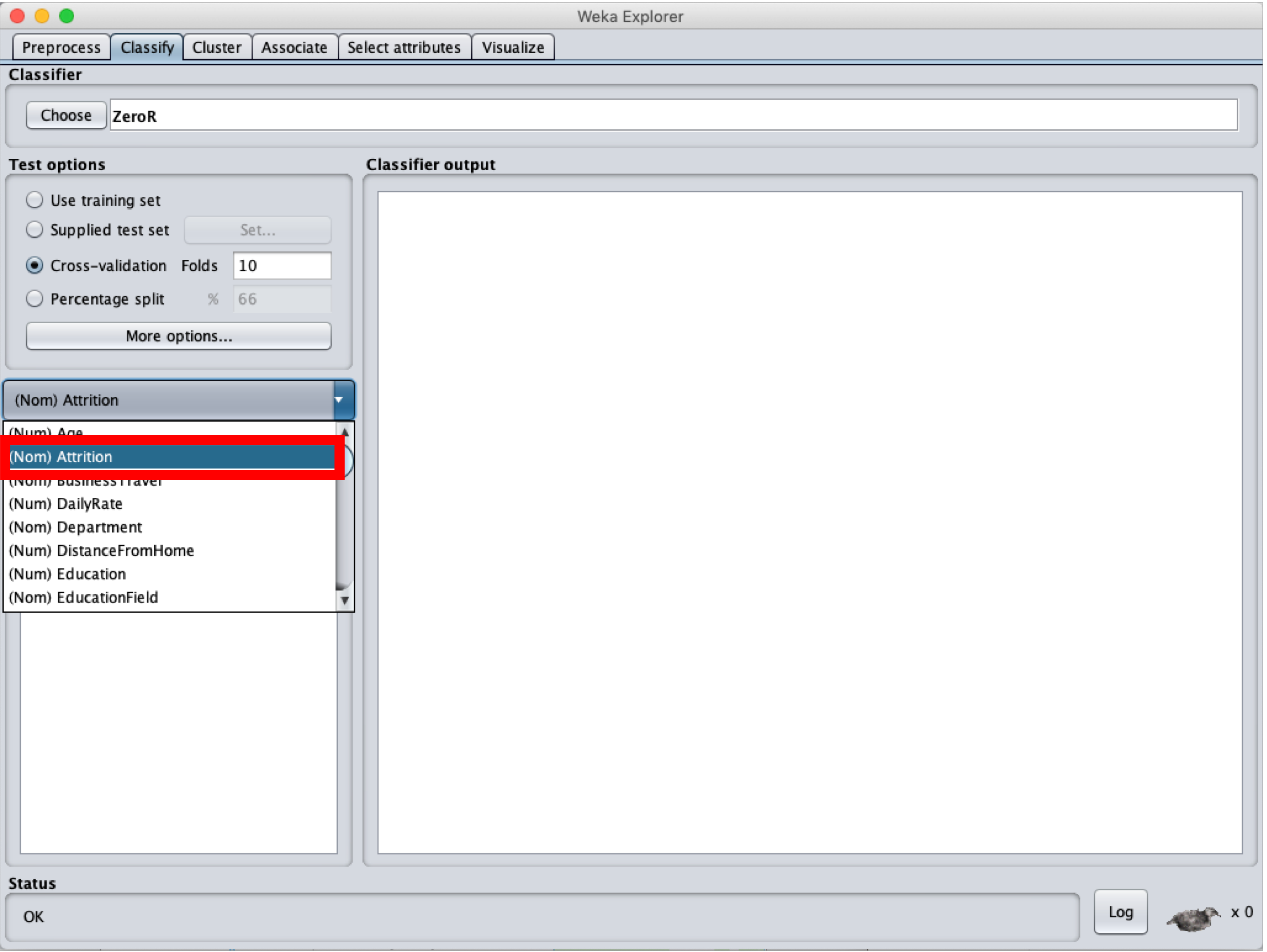
Class: Attrition (Nom)

Status

OK

Click on Classify

Weka – Select the Right Target Variable



ZeroR

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

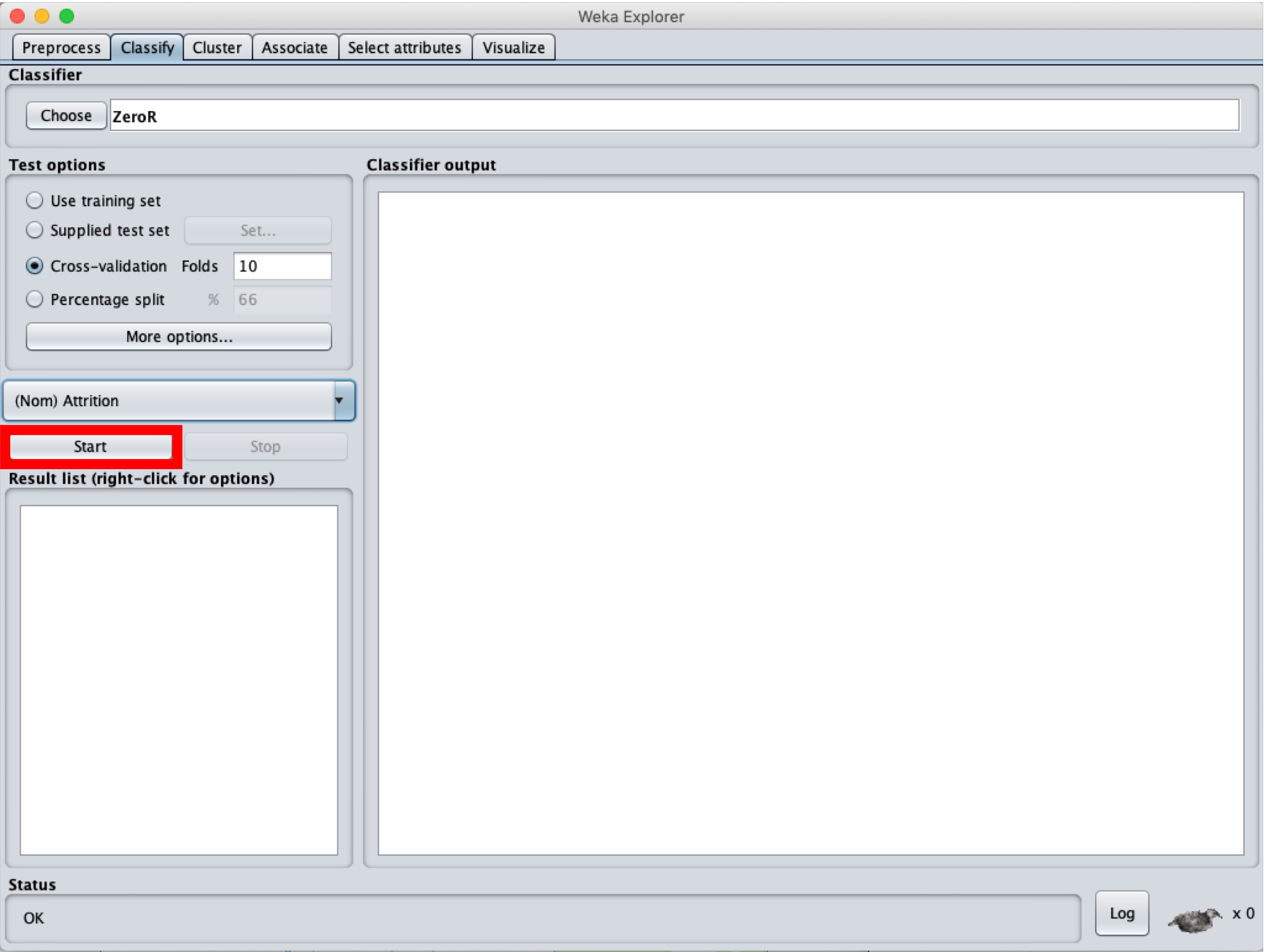
(Nom) Attrition

(Num) Age
(Nom) Attrition
(Nom) BusinessTravel
(Num) DailyRate
(Nom) Department
(Num) DistanceFromHome
(Num) Education
(Nom) EducationField

OK

Select Attrition as the target variable

Weka – Other Things to Note



Weka Explorer

Classifier

Choose ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Attrition

Status

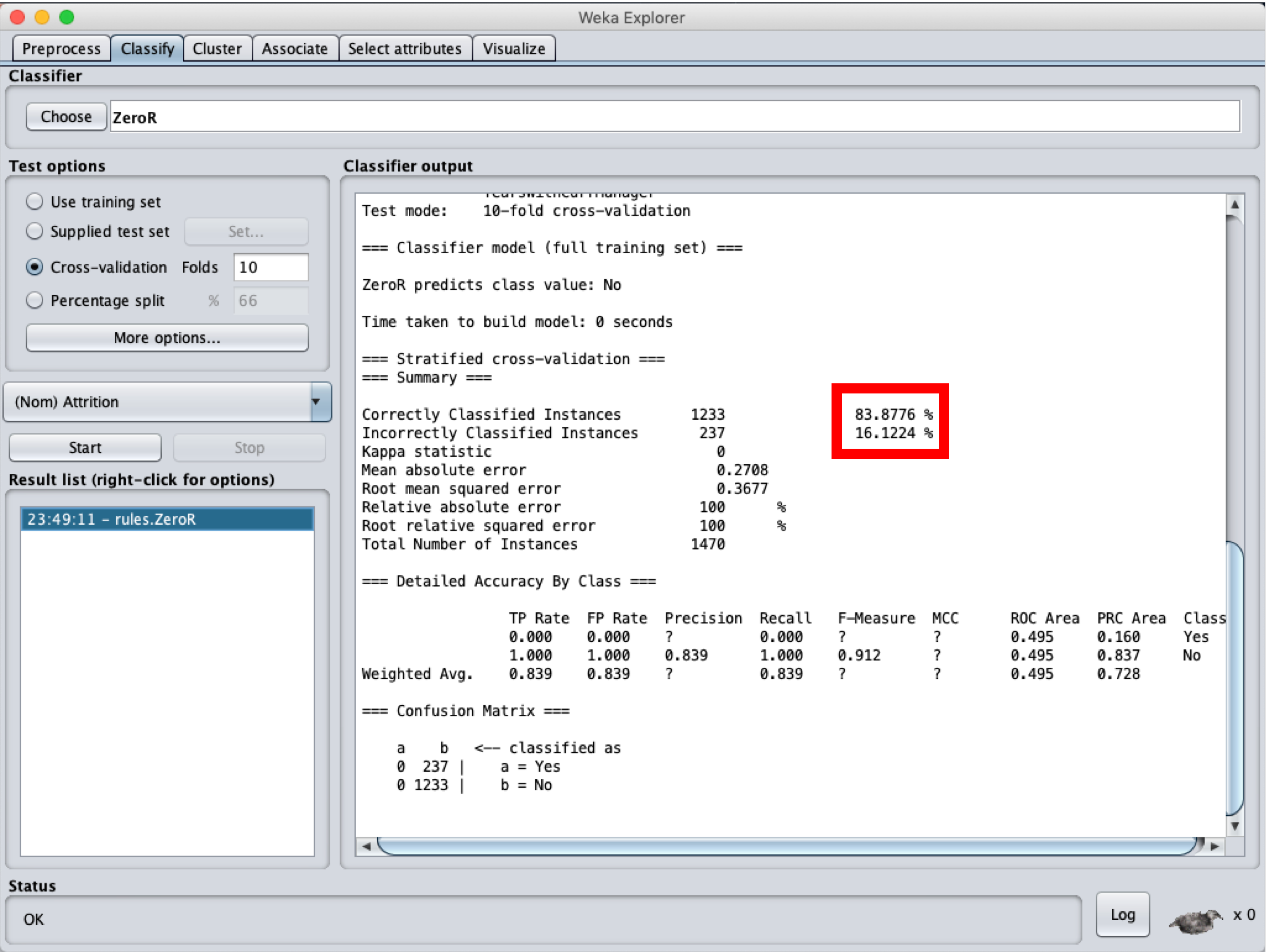
OK

ZeroR is the algorithm we run – this happens to be the same as the baseline we discussed

Also note that cross-validation is selected by default – no need to worry about it!

Now click Start

Weka – Baseline Results



Classifier

Choose ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Attrition

Result list (right-click for options)

23:49:11 - rules.ZeroR

Status

OK

Classifier output

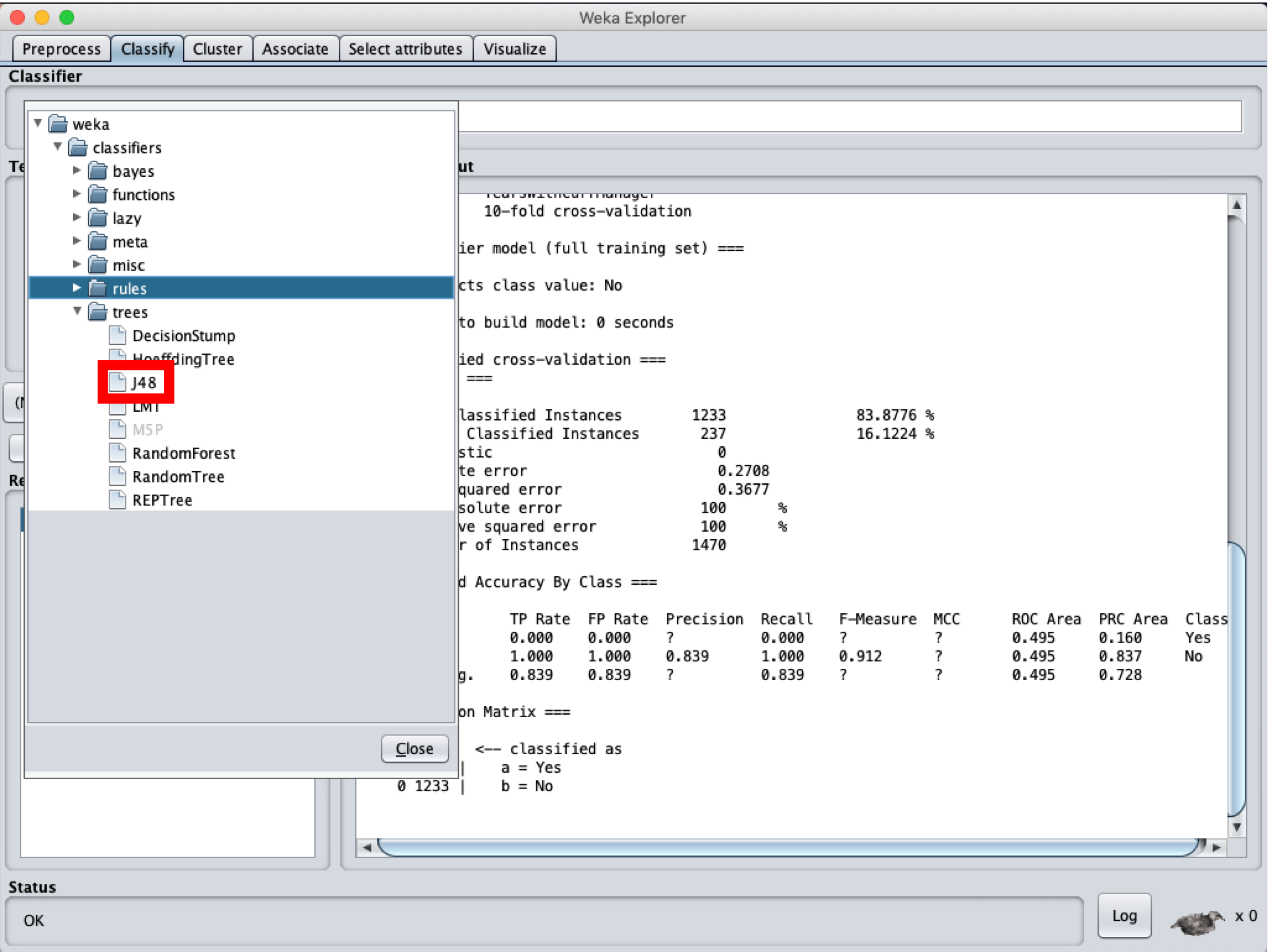
```
Test mode:    10-fold cross-validation

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

ZeroR predicts class value: No

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1233               83.8776 %
Incorrectly Classified Instances       237               16.1224 %
Kappa statistic                          0
Mean absolute error                      0.2708
Root mean squared error                  0.3677
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.495     0.160     Yes
                 1.000    1.000    0.839      1.000    0.912      ?        0.495     0.837     No
Weighted Avg.    0.839    0.839    ?          0.839    ?          ?        0.495     0.728

=== Confusion Matrix ===

    a    b   <-- classified as
    0  237 |    a = Yes
    0 1233 |    b = No
```

The baseline is 83.9% just like we discussed

Many other performance metrics are displayed

Try to understand the confusion matrix (yes, it is confusing ☺)

Weka – Creating a Decision Tree

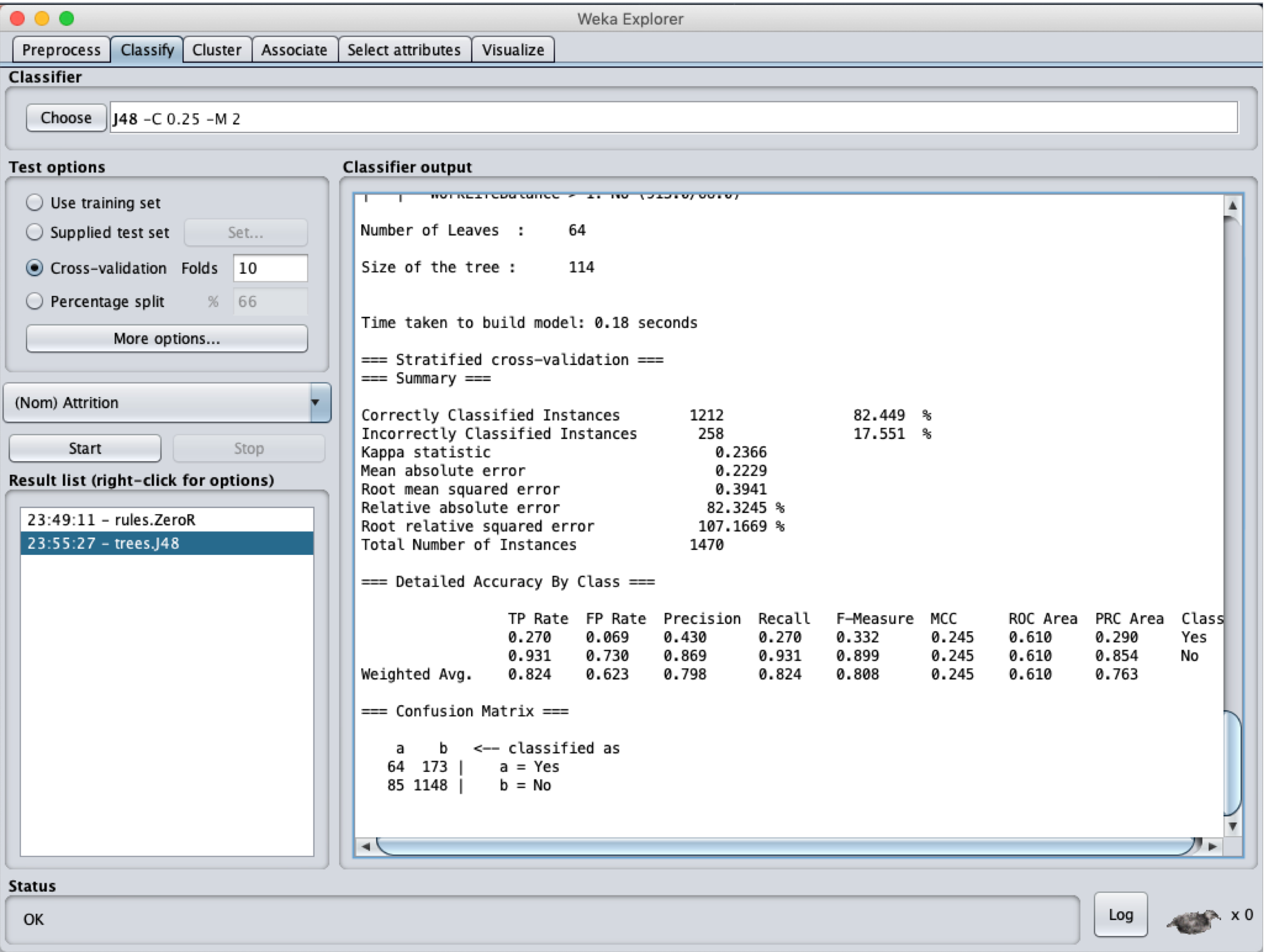


```
   10-fold cross-validation

ier model (full training set) ===

cts class value: No

to build model: 0 seconds

ied cross-validation ===
 ===

lassified Instances        1233               83.8776 %
 Classified Instances       237               16.1224 %
stic                          0
te error                      0.2708
quared error                  0.3677
solute error                100      %
ve squared error            100      %
r of Instances             1470

d Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.000    0.000    ?          0.000    ?          ?        0.495     0.160     Yes
      1.000    1.000    0.839      1.000    0.912      ?        0.495     0.837     No
g.    0.839    0.839    ?          0.839    ?          ?        0.495     0.728

on Matrix ===

   <-- classified as
 |    a = Yes
0 1233 |    b = No
```

Select J48 under trees

Then click Start again

Weka – Decision Tree Results



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Attrition

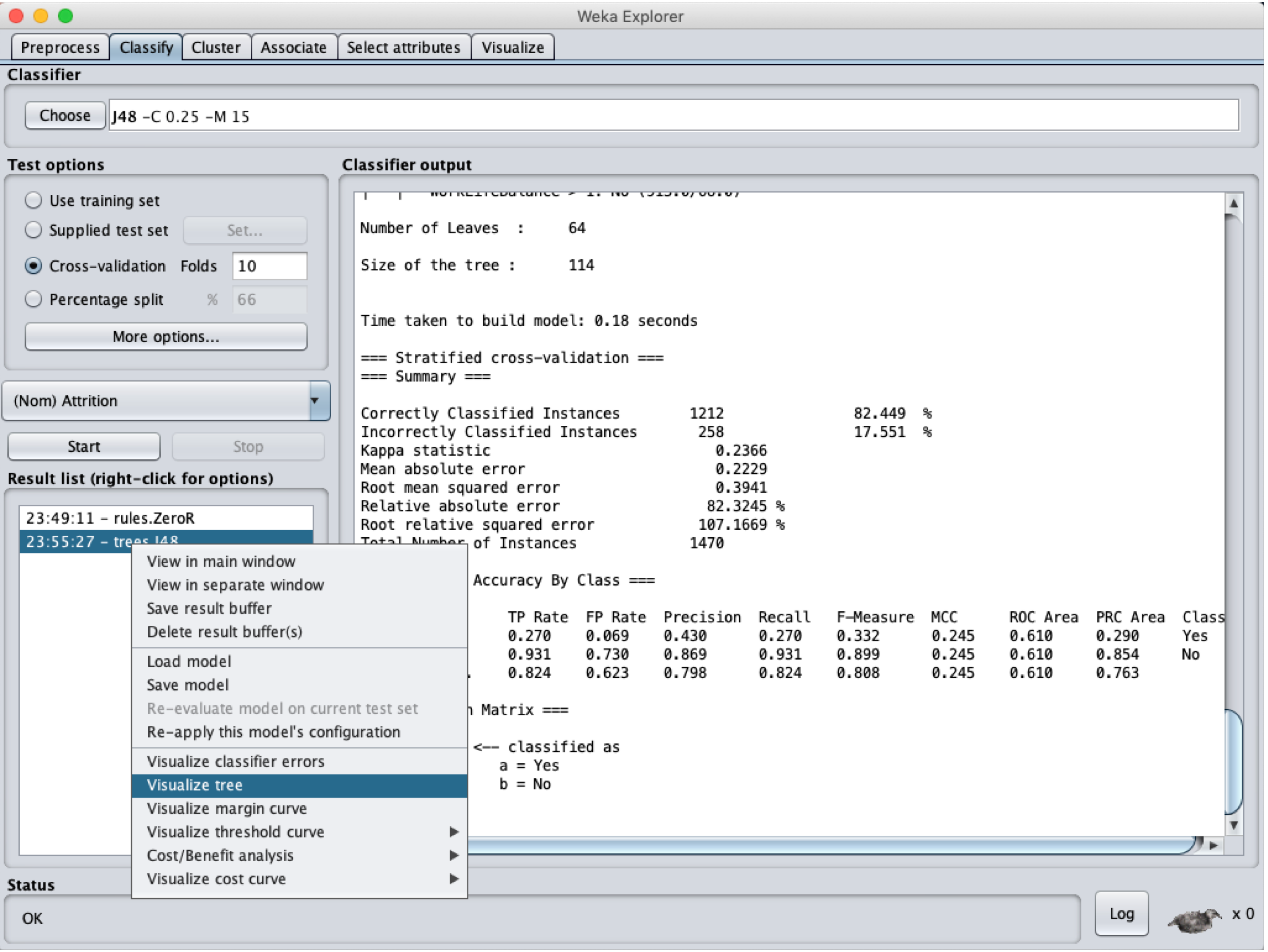
23:49:11 - rules.ZeroR
23:55:27 - trees.J48

```
|   |   WorkLifeBalance > 1: No (313.0/60.0)

Number of Leaves  :     64

Size of the tree :      114


Time taken to build model: 0.18 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1212               82.449  %
Incorrectly Classified Instances       258               17.551  %
Kappa statistic                          0.2366
Mean absolute error                      0.2229
Root mean squared error                  0.3941
Relative absolute error                 82.3245 %
Root relative squared error            107.1669 %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.270    0.069    0.430      0.270    0.332      0.245    0.610     0.290     Yes
                 0.931    0.730    0.869      0.931    0.899      0.245    0.610     0.854     No
Weighted Avg.    0.824    0.623    0.798      0.824    0.808      0.245    0.610     0.763

=== Confusion Matrix ===

    a    b   <-- classified as
   64  173 |    a = Yes
   85 1148 |    b = No
```

OK

What do you think?

Weka – Visualizing the Tree



J48 -C 0.25 -M 15

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Attrition

23:49:11 - rules.ZeroR

23:55:27 - trees.J48

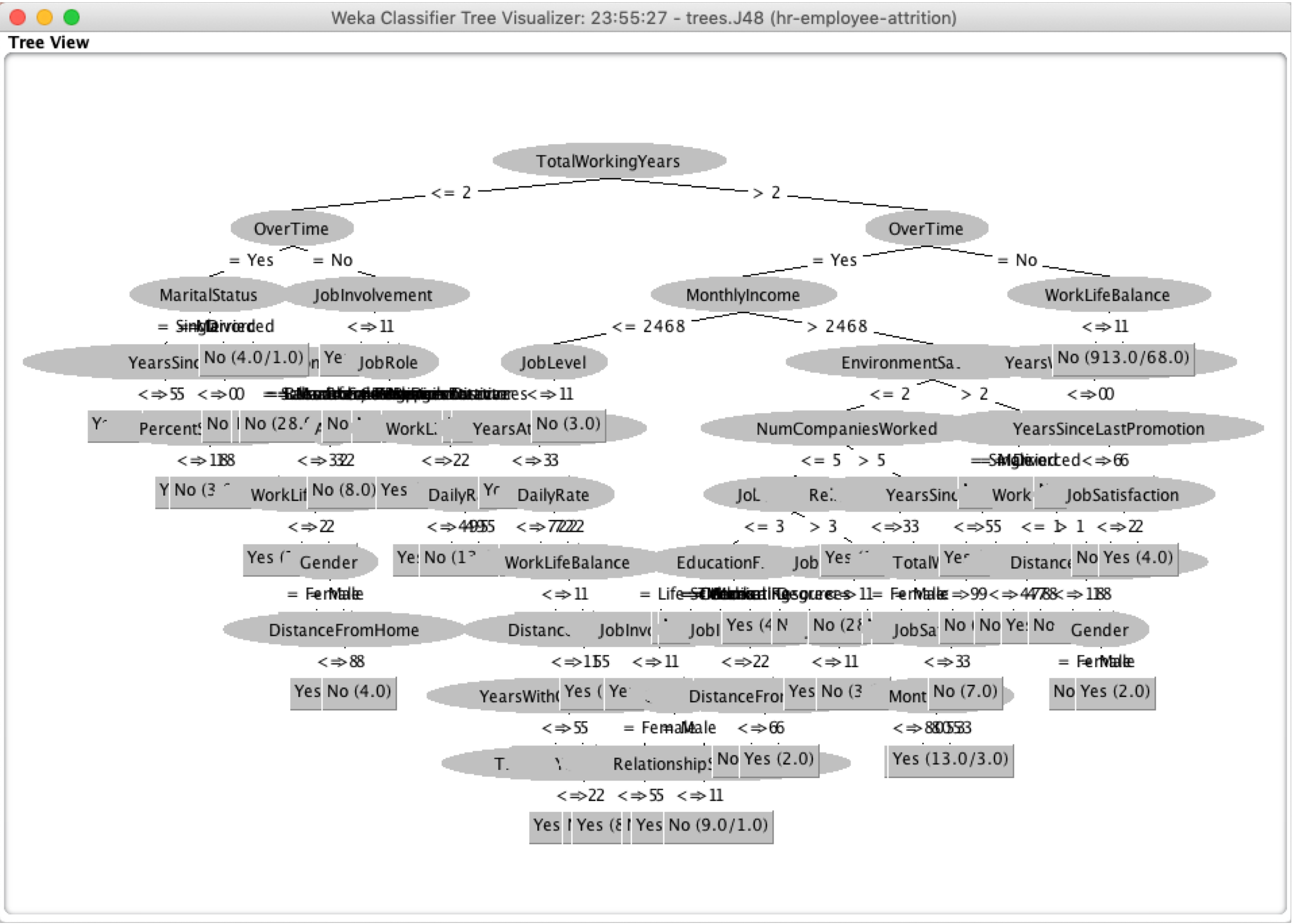
```
|   |   WorkLifeBalance > 1: No (313.0/68.0)

Number of Leaves  :     64

Size of the tree :      114


Time taken to build model: 0.18 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1212               82.449  %
Incorrectly Classified Instances       258               17.551  %
Kappa statistic                          0.2366
Mean absolute error                      0.2229
Root mean squared error                  0.3941
Relative absolute error                 82.3245 %
Root relative squared error            107.1669 %
Total Number of Instances             1470

             Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.270    0.069    0.430      0.270    0.332      0.245    0.610     0.290     Yes
                0.931    0.730    0.869      0.931    0.899      0.245    0.610     0.854     No
             .  0.824    0.623    0.798      0.824    0.808      0.245    0.610     0.763

            n Matrix ===

             <-- classified as
                a = Yes
                b = No
```

Status

OK

Right click and select Visualize tree

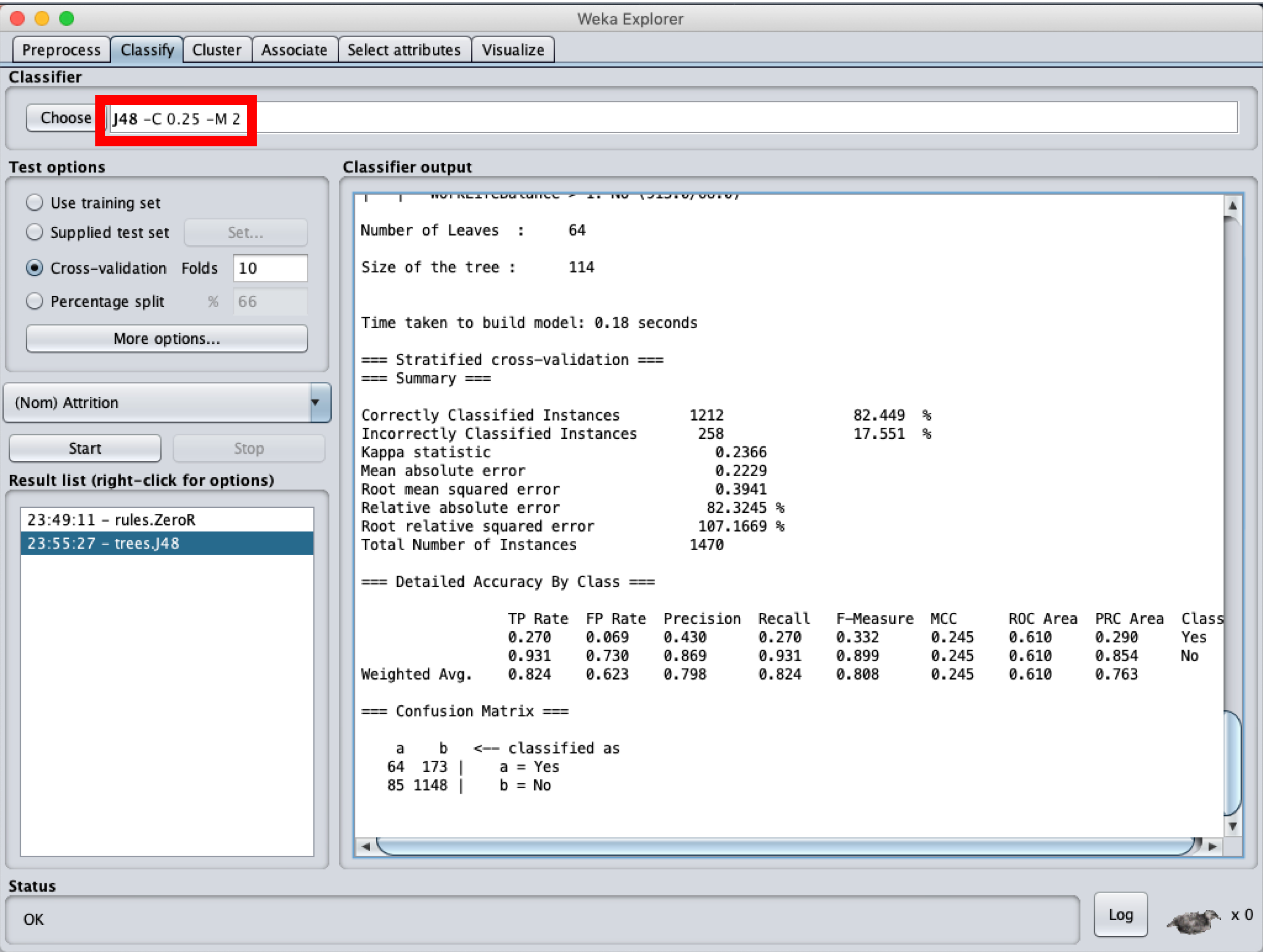
Weka – Visualizing the Tree



A bit messy!

But we gain insights: TotalWorkingYears is important!

Weka – Tweaking



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Attrition

Result list (right-click for options)

23:49:11 - rules.ZeroR

23:55:27 - trees.J48

Classifier output

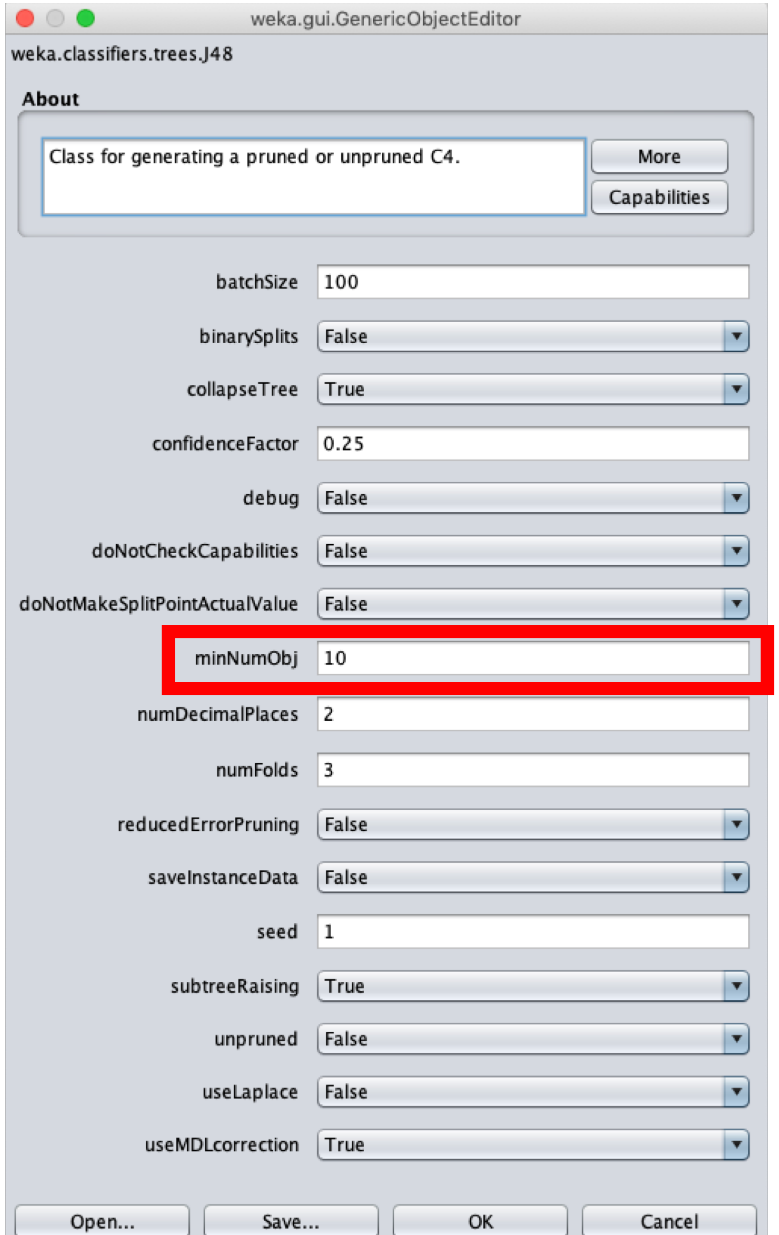
```
Number of Leaves  :     64

Size of the tree :      114


Time taken to build model: 0.18 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1212               82.449  %
Incorrectly Classified Instances       258               17.551  %
Kappa statistic                          0.2366
Mean absolute error                      0.2229
Root mean squared error                  0.3941
Relative absolute error                 82.3245 %
Root relative squared error            107.1669 %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.270    0.069    0.430      0.270    0.332      0.245    0.610     0.290     Yes
                 0.931    0.730    0.869      0.931    0.899      0.245    0.610     0.854     No
Weighted Avg.    0.824    0.623    0.798      0.824    0.808      0.245    0.610     0.763

=== Confusion Matrix ===

    a    b   <-- classified as
   64  173 |    a = Yes
   85 1148 |    b = No
```

Status

OK

Click on the algorithm name

Weka – Changing minNumObj



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

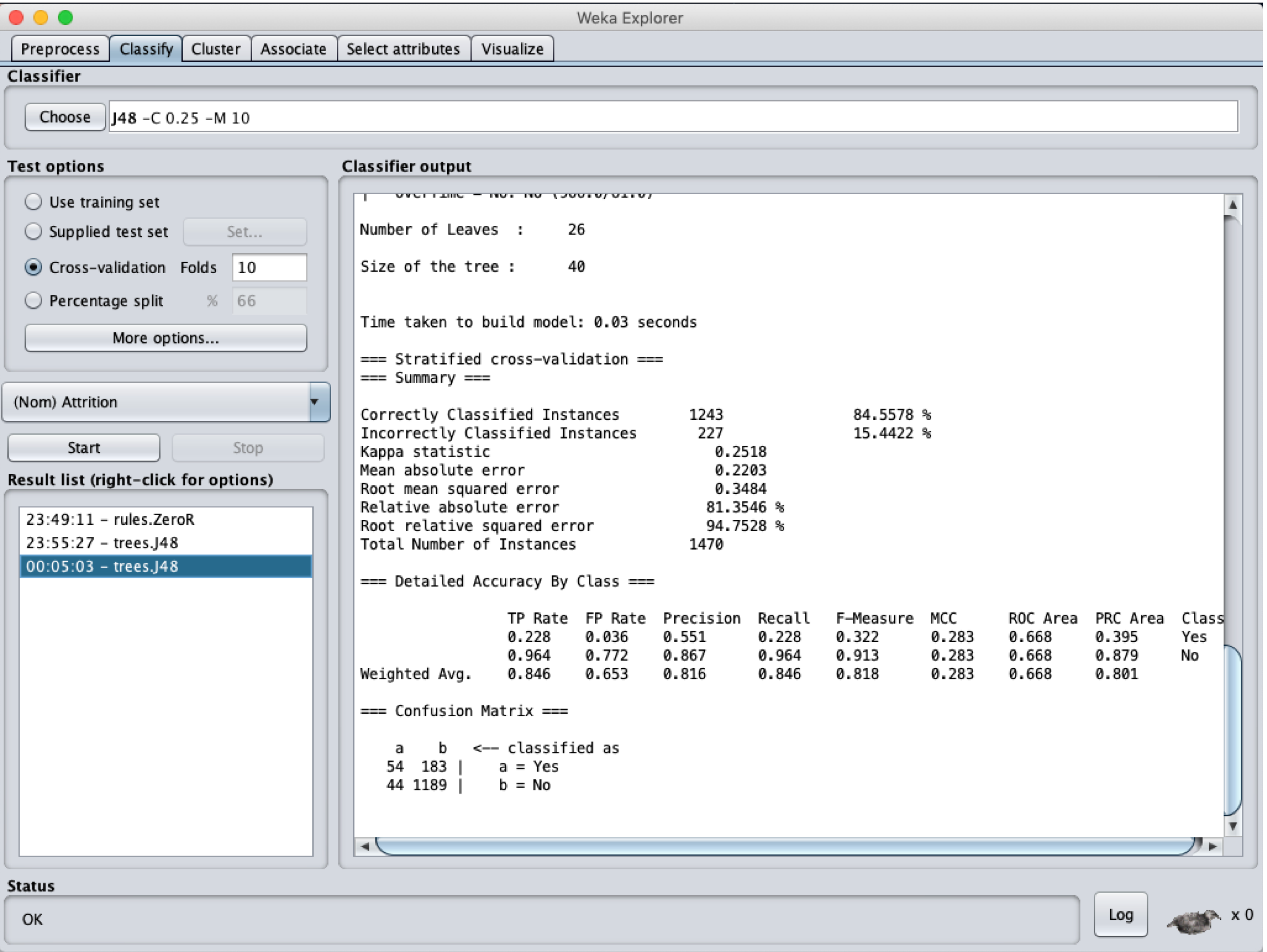
About

Class for generating a pruned or unpruned C4.

batchSize	100
binarySplits	False
collapseTree	True
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
doNotMakeSplitPointActualValue	False
minNumObj	10
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False
useMDLcorrection	True

Change MinNumObj to 10 and run again

Weka – Updated Results



J48 -C 0.25 -M 10

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Attrition

23:49:11 - rules.ZeroR
23:55:27 - trees.J48
00:05:03 - trees.J48

```
Number of Leaves  :     26

Size of the tree :      40


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1243               84.5578 %
Incorrectly Classified Instances       227               15.4422 %
Kappa statistic                          0.2518
Mean absolute error                      0.2203
Root mean squared error                  0.3484
Relative absolute error                 81.3546 %
Root relative squared error             94.7528 %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.228    0.036    0.551      0.228    0.322      0.283    0.668     0.395     Yes
                 0.964    0.772    0.867      0.964    0.913      0.283    0.668     0.879     No
Weighted Avg.    0.846    0.653    0.816      0.846    0.818      0.283    0.668     0.801

=== Confusion Matrix ===

    a    b   <-- classified as
   54  183 |    a = Yes
   44 1189 |    b = No
```

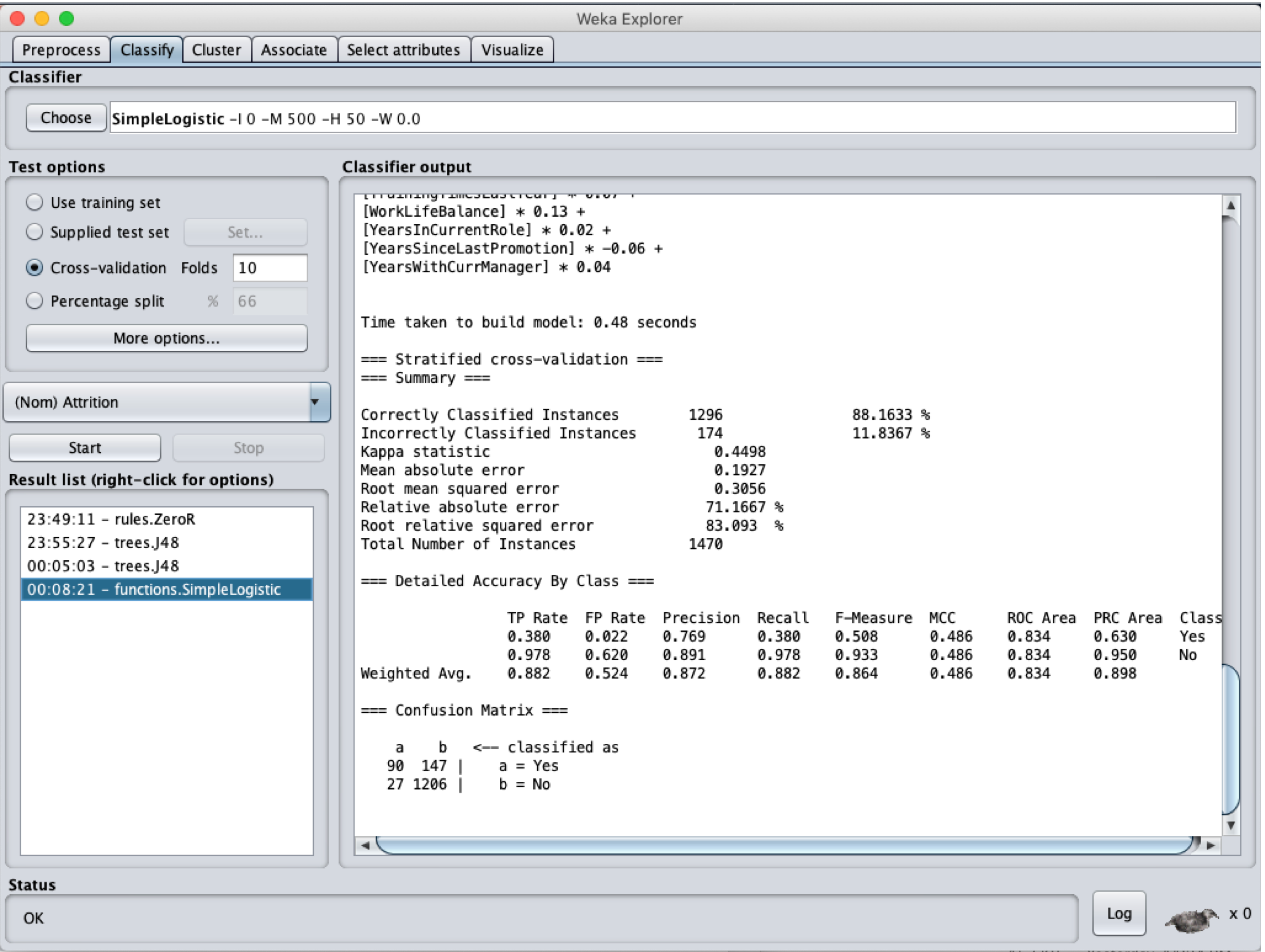
Status: OK

What do you think?

Visualize the tree again: it is less overfitted!

(simpler, but better performance)

Weka – Logistic Regression (= Classification)



Classifier

SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Attrition

Result list (right-click for options)

23:49:11 - rules.ZeroR
23:55:27 - trees.J48
00:05:03 - trees.J48
00:08:21 - functions.SimpleLogistic

Classifier output

```
[WorkLifeBalance] * 0.13 +
[YearsInCurrentRole] * 0.02 +
[YearsSinceLastPromotion] * -0.06 +
[YearsWithCurrManager] * 0.04


Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1296               88.1633 %
Incorrectly Classified Instances       174               11.8367 %
Kappa statistic                          0.4498
Mean absolute error                      0.1927
Root mean squared error                  0.3056
Relative absolute error                 71.1667 %
Root relative squared error             83.093  %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.380    0.022    0.769      0.380    0.508      0.486    0.834     0.630     Yes
                 0.978    0.620    0.891      0.978    0.933      0.486    0.834     0.950     No
Weighted Avg.    0.882    0.524    0.872      0.882    0.864      0.486    0.834     0.898

=== Confusion Matrix ===

    a    b   <-- classified as
   90  147 |    a = Yes
   27 1206 |    b = No
```

Status

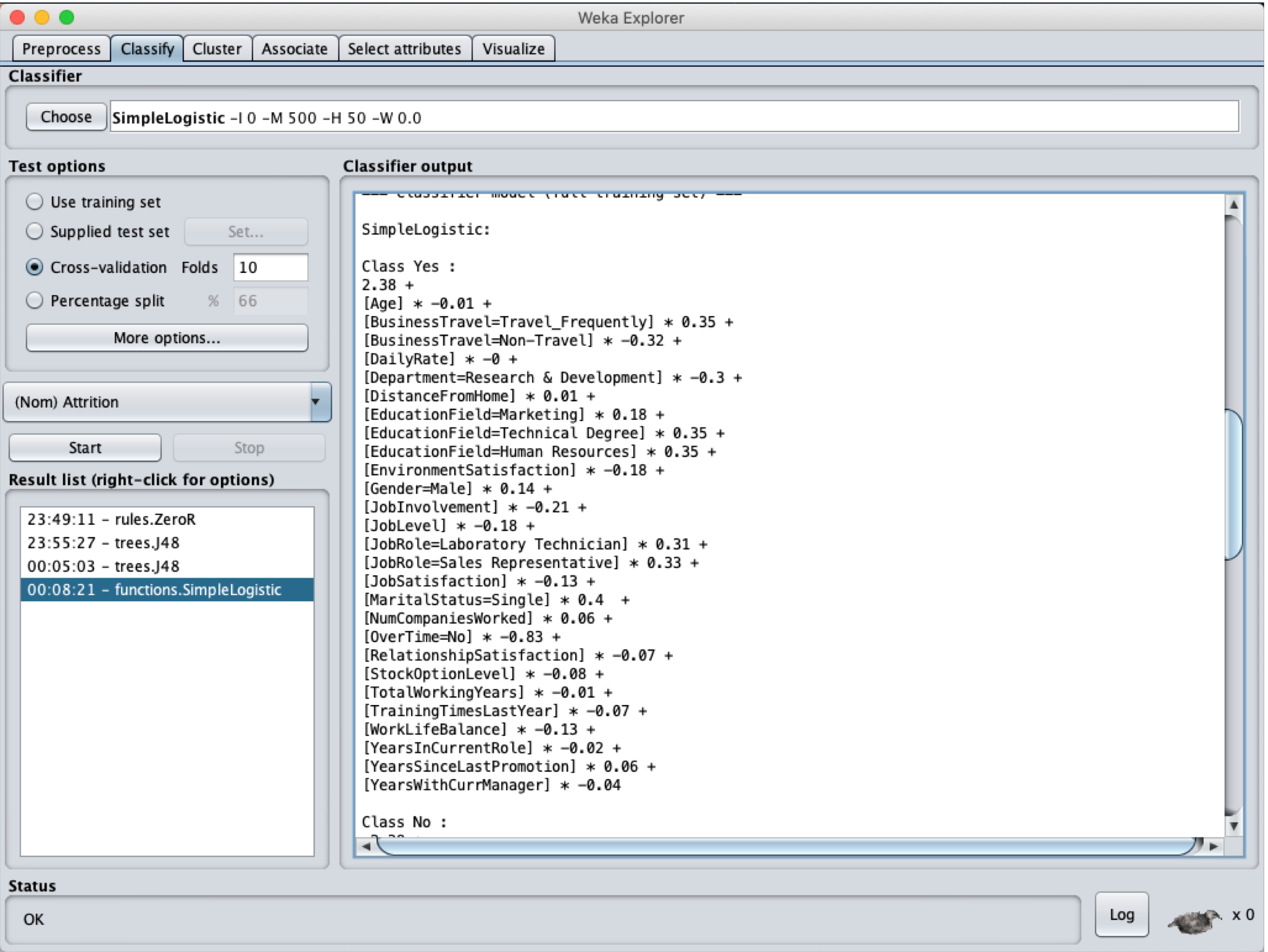
OK

Click Choose and select Simple Logistic under functions

Then run it

How about these results?

Weka – Logistic Regression (= Classification)



Classifier

SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

Test options

- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) Attrition

Result list (right-click for options)

- 23:49:11 - rules.ZeroR
- 23:55:27 - trees.J48
- 00:05:03 - trees.J48
- 00:08:21 - functions.SimpleLogistic

Classifier output

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

SimpleLogistic:

Class Yes :
2.38 +
[Age] * -0.01 +
[BusinessTravel=Travel_Frequently] * 0.35 +
[BusinessTravel=Non-Travel] * -0.32 +
[DailyRate] * -0 +
[Department=Research & Development] * -0.3 +
[DistanceFromHome] * 0.01 +
[EducationField=Marketing] * 0.18 +
[EducationField=Technical Degree] * 0.35 +
[EducationField=Human Resources] * 0.35 +
[EnvironmentSatisfaction] * -0.18 +
[Gender=Male] * 0.14 +
[JobInvolvement] * -0.21 +
[JobLevel] * -0.18 +
[JobRole=Laboratory Technician] * 0.31 +
[JobRole=Sales Representative] * 0.33 +
[JobSatisfaction] * -0.13 +
[MaritalStatus=Single] * 0.4  +
[NumCompaniesWorked] * 0.06 +
[OverTime=No] * -0.83 +
[RelationshipSatisfaction] * -0.07 +
[StockOptionLevel] * -0.08 +
[TotalWorkingYears] * -0.01 +
[TrainingTimesLastYear] * -0.07 +
[WorkLifeBalance] * -0.13 +
[YearsInCurrentRole] * -0.02 +
[YearsSinceLastPromotion] * 0.06 +
[YearsWithCurrManager] * -0.04

Class No :
```

Status

OK

Scroll up to the first equation

The coefficients determine impact (just like with linear regression)

For example, traveling frequently is correlated with leaving

Next Steps

- Try it out!
- A **lot** of **trial** and **error** (try to be somewhat systematic)
- Use Google
- Try different algorithms and see how well they work
    - Are all of them telling a consistent story?
- Ask questions