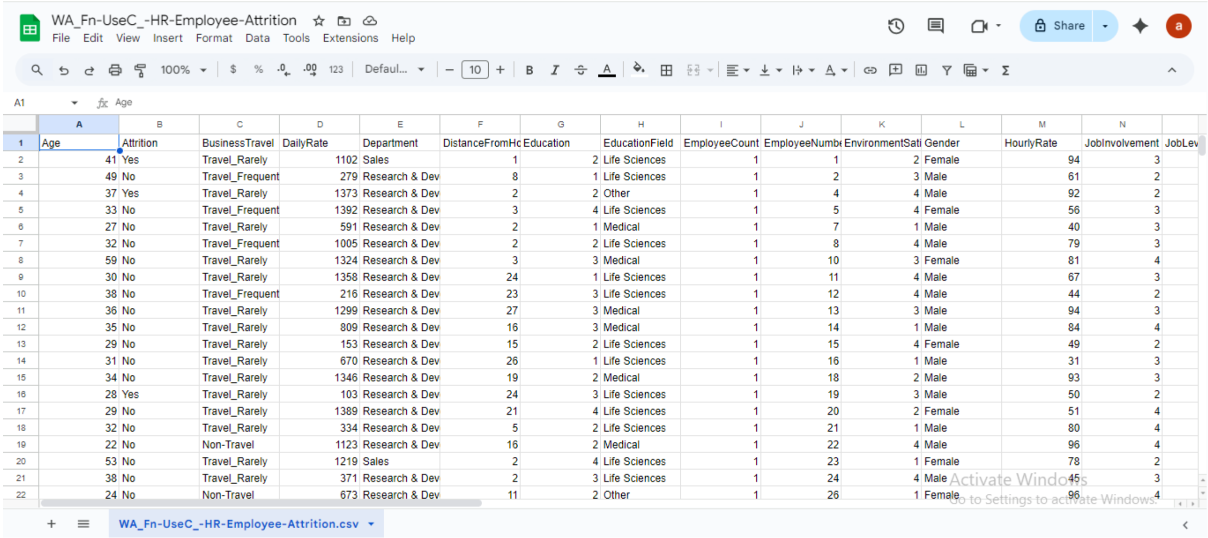
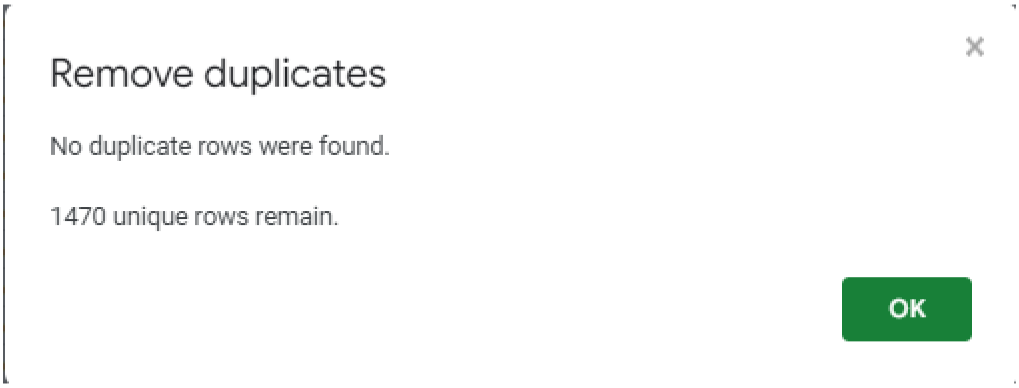
STEP 1: Load the Dataset


WA_Fn-UseC_-HR-Employee-Attrition
File
Edit
View
Insert
Format
Data
Tools
Extensions
Help
Share
100%
Defaul...
10
A1
fx Age
WA_Fn-UseC_-HR-Employee-Attrition.csv
Activate Windows
Go to Settings to activate Windows.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHc	Education	EducationField	EmployeeCount	EmployeeNumbe	EnvironmentSati	Gender	HourlyRate	JobInvolvement	JobLev
2	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	
3	49	No	Travel_Frequent	279	Research & Dev	8	1	Life Sciences	1	2	3	Male	61	2	
4	37	Yes	Travel_Rarely	1373	Research & Dev	2	2	Other	1	4	4	Male	92	2	
5	33	No	Travel_Frequent	1392	Research & Dev	3	4	Life Sciences	1	5	4	Female	56	3	
6	27	No	Travel_Rarely	591	Research & Dev	2	1	Medical	1	7	1	Male	40	3	
7	32	No	Travel_Frequent	1005	Research & Dev	2	2	Life Sciences	1	8	4	Male	79	3	
8	59	No	Travel_Rarely	1324	Research & Dev	3	3	Medical	1	10	3	Female	81	4	
9	30	No	Travel_Rarely	1358	Research & Dev	24	1	Life Sciences	1	11	4	Male	67	3	
10	38	No	Travel_Frequent	216	Research & Dev	23	3	Life Sciences	1	12	4	Male	44	2	
11	36	No	Travel_Rarely	1299	Research & Dev	27	3	Medical	1	13	3	Male	94	3	
12	35	No	Travel_Rarely	809	Research & Dev	16	3	Medical	1	14	1	Male	84	4	
13	29	No	Travel_Rarely	153	Research & Dev	15	2	Life Sciences	1	15	4	Female	49	2	
14	31	No	Travel_Rarely	670	Research & Dev	26	1	Life Sciences	1	16	1	Male	31	3	
15	34	No	Travel_Rarely	1346	Research & Dev	19	2	Medical	1	18	2	Male	93	3	
16	28	Yes	Travel_Rarely	103	Research & Dev	24	3	Life Sciences	1	19	3	Male	50	2	
17	29	No	Travel_Rarely	1389	Research & Dev	21	4	Life Sciences	1	20	2	Female	51	4	
18	32	No	Travel_Rarely	334	Research & Dev	5	2	Life Sciences	1	21	1	Male	80	4	
19	22	No	Non-Travel	1123	Research & Dev	16	2	Medical	1	22	4	Male	96	4	
20	53	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	
21	38	No	Travel_Rarely	371	Research & Dev	2	3	Life Sciences	1	24	4	Male	45	3	
22	24	No	Non-Travel	673	Research & Dev	11	2	Other	1	26	1	Female	96	4	

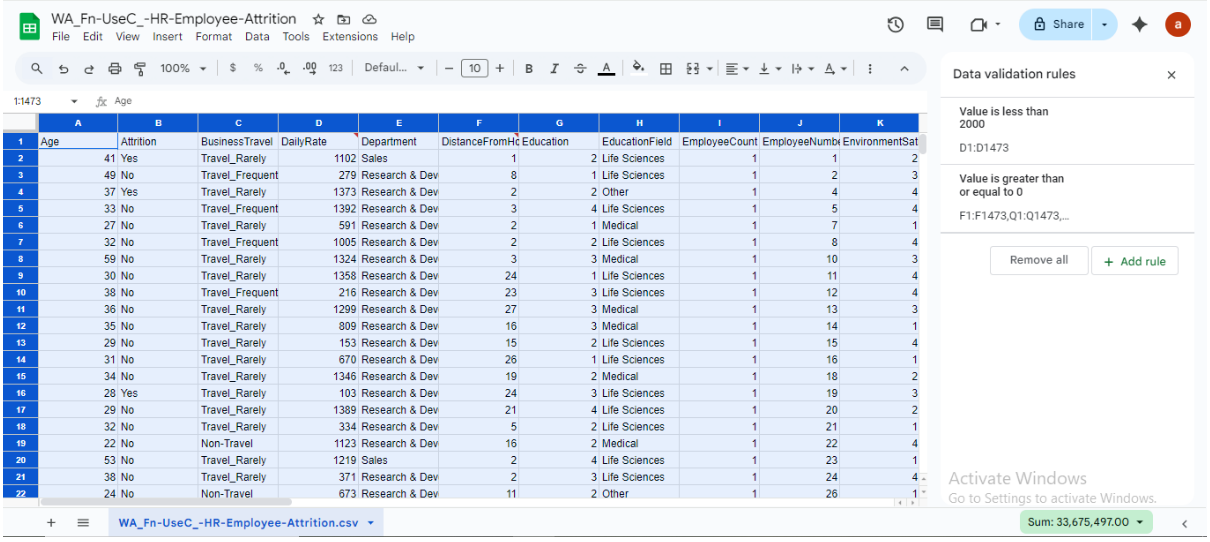
STEP 2: Clean & Prepare the Dataset



Remove duplicates

No duplicate rows were found.

1470 unique rows remain.


WA_Fn-UseC_-HR-Employee-Attrition
File
Edit
View
Insert
Format
Data
Tools
Extensions
Help
Share
100%
Defaul...
10
1:1473
fx Age
WA_Fn-UseC_-HR-Employee-Attrition.csv
Activate Windows
Go to Settings to activate Windows.

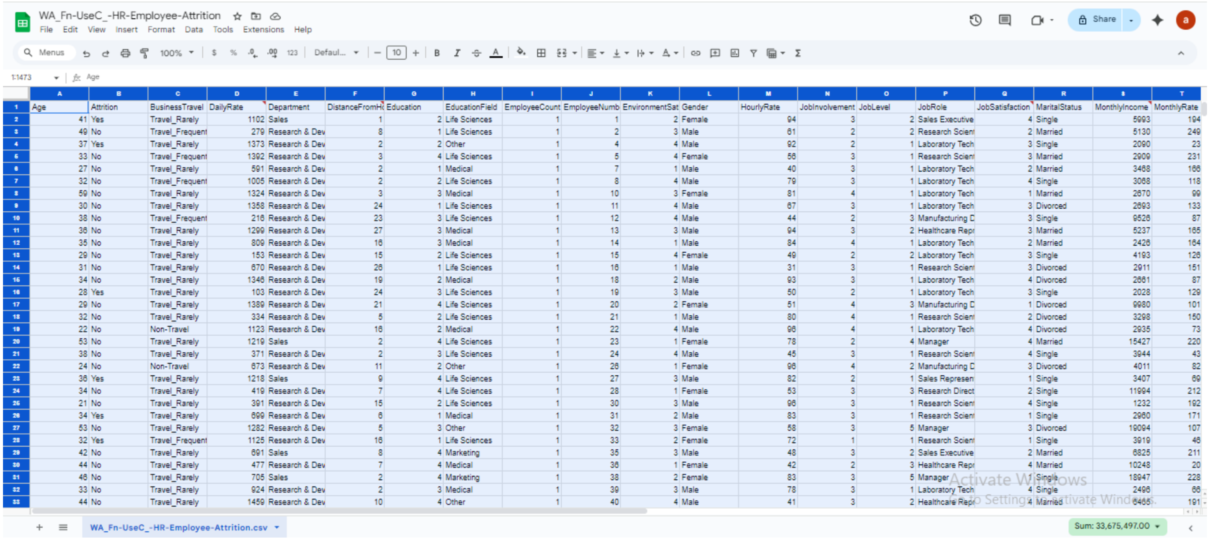
	A	B	C	D	E	F	G	H	I	J	K
1	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHc	Education	EducationField	EmployeeCount	EmployeeNumbe	EnvironmentSat
2	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2
3	49	No	Travel_Frequent	279	Research & Dev	8	1	Life Sciences	1	2	3
4	37	Yes	Travel_Rarely	1373	Research & Dev	2	2	Other	1	4	4
5	33	No	Travel_Frequent	1392	Research & Dev	3	4	Life Sciences	1	5	4
6	27	No	Travel_Rarely	591	Research & Dev	2	1	Medical	1	7	1
7	32	No	Travel_Frequent	1005	Research & Dev	2	2	Life Sciences	1	8	4
8	59	No	Travel_Rarely	1324	Research & Dev	3	3	Medical	1	10	3
9	30	No	Travel_Rarely	1358	Research & Dev	24	1	Life Sciences	1	11	4
10	38	No	Travel_Frequent	216	Research & Dev	23	3	Life Sciences	1	12	4
11	36	No	Travel_Rarely	1299	Research & Dev	27	3	Medical	1	13	3
12	35	No	Travel_Rarely	809	Research & Dev	16	3	Medical	1	14	1
13	29	No	Travel_Rarely	153	Research & Dev	15	2	Life Sciences	1	15	4
14	31	No	Travel_Rarely	670	Research & Dev	26	1	Life Sciences	1	16	1
15	34	No	Travel_Rarely	1346	Research & Dev	19	2	Medical	1	18	2
16	28	Yes	Travel_Rarely	103	Research & Dev	24	3	Life Sciences	1	19	3
17	29	No	Travel_Rarely	1389	Research & Dev	21	4	Life Sciences	1	20	2
18	32	No	Travel_Rarely	334	Research & Dev	5	2	Life Sciences	1	21	1
19	22	No	Non-Travel	1123	Research & Dev	16	2	Medical	1	22	4
20	53	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1
21	38	No	Travel_Rarely	371	Research & Dev	2	3	Life Sciences	1	24	4
22	24	No	Non-Travel	673	Research & Dev	11	2	Other	1	26	1

Data validation rules

Value is less than 2000
D1:D1473

Value is greater than or equal to 0
F1:F1473,Q1:Q1473,...

Sum: 33,675,497.00

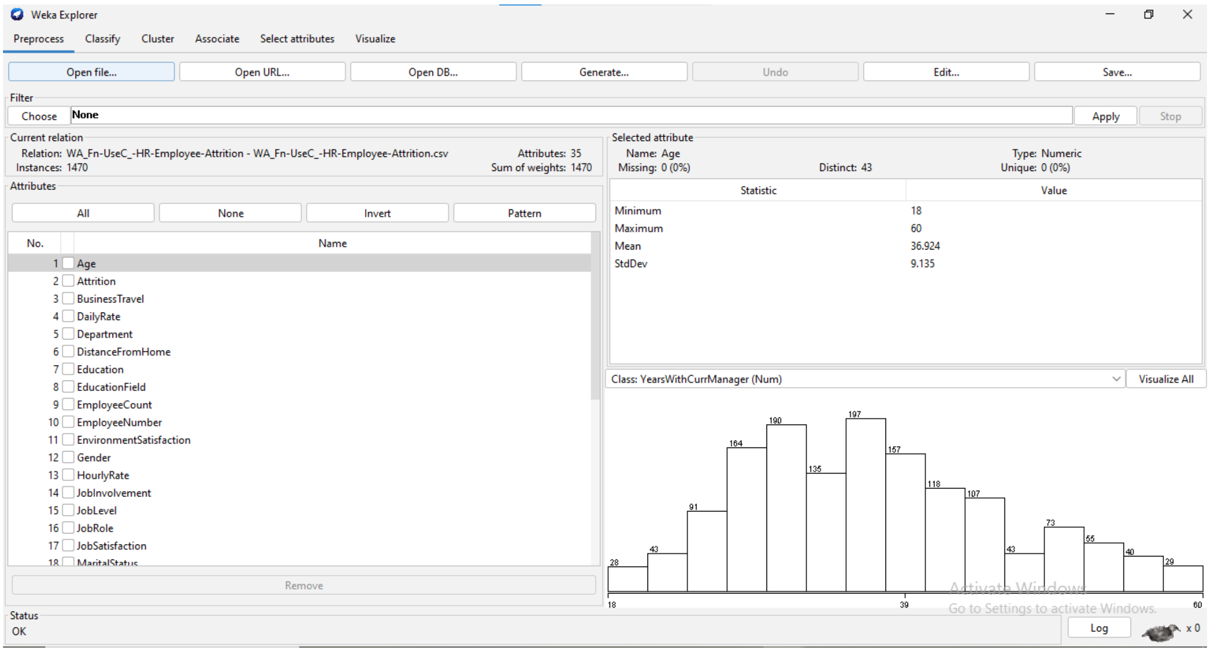

WA_Fn-UseC_-HR-Employee-Attrition
File Edit View Insert Format Data Tools Extensions Help
Menus
100%
Defaul...
10
Share
1:1473
Age
WA_Fn-UseC_-HR-Employee-Attrition.csv
Sum: 33,675,497.00

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromH	Education	EducationField	EmployeeCount	EmployeeNumb	EnvironmentSat	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate
2	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2	Sales Executive	4	Single	5993	194
3	49	No	Travel_Frequent	279	Research & Dev	8	1	Life Sciences	1	2	3	Male	61	2	2	Research Scient	2	Married	5130	249
4	37	Yes	Travel_Rarely	1373	Research & Dev	2	2	Other	1	4	4	Male	92	2	1	Laboratory Tech	3	Single	2090	23
5	33	No	Travel_Frequent	1392	Research & Dev	3	4	Life Sciences	1	5	4	Female	56	3	1	Research Scient	3	Married	2909	231
6	27	No	Travel_Rarely	591	Research & Dev	2	1	Medical	1	7	1	Male	40	3	1	Laboratory Tech	2	Married	3468	166
7	32	No	Travel_Frequent	1005	Research & Dev	2	2	Life Sciences	1	8	4	Male	79	3	1	Laboratory Tech	4	Single	3068	118
8	59	No	Travel_Rarely	1324	Research & Dev	3	3	Medical	1	10	3	Female	81	4	1	Laboratory Tech	1	Married	2670	99
9	30	No	Travel_Rarely	1358	Research & Dev	24	1	Life Sciences	1	11	4	Male	67	3	1	Laboratory Tech	3	Divorced	2693	133
10	38	No	Travel_Frequent	216	Research & Dev	23	3	Life Sciences	1	12	4	Male	44	2	3	Manufacturing D	3	Single	9526	87
11	36	No	Travel_Rarely	1299	Research & Dev	27	3	Medical	1	13	3	Male	94	3	2	Healthcare Repr	3	Married	5237	165
12	35	No	Travel_Rarely	809	Research & Dev	16	3	Medical	1	14	1	Male	84	4	1	Laboratory Tech	2	Married	2426	164
13	29	No	Travel_Rarely	153	Research & Dev	15	2	Life Sciences	1	15	4	Female	49	2	2	Laboratory Tech	3	Single	4193	126
14	31	No	Travel_Rarely	670	Research & Dev	26	1	Life Sciences	1	16	1	Male	31	3	1	Research Scient	3	Divorced	2911	151
15	34	No	Travel_Rarely	1346	Research & Dev	19	2	Medical	1	18	2	Male	93	3	1	Laboratory Tech	4	Divorced	2661	87
16	28	Yes	Travel_Rarely	103	Research & Dev	24	3	Life Sciences	1	19	3	Male	50	2	1	Laboratory Tech	3	Single	2028	129
17	29	No	Travel_Rarely	1389	Research & Dev	21	4	Life Sciences	1	20	2	Female	51	4	3	Manufacturing D	1	Divorced	9980	101
18	32	No	Travel_Rarely	334	Research & Dev	5	2	Life Sciences	1	21	1	Male	80	4	1	Research Scient	2	Divorced	3298	150
19	22	No	Non-Travel	1123	Research & Dev	16	2	Medical	1	22	4	Male	96	4	1	Laboratory Tech	4	Divorced	2935	73
20	53	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	4	Manager	4	Married	15427	220
21	38	No	Travel_Rarely	371	Research & Dev	2	3	Life Sciences	1	24	4	Male	45	3	1	Research Scient	4	Single	3944	43
22	24	No	Non-Travel	673	Research & Dev	11	2	Other	1	26	1	Female	96	4	2	Manufacturing D	3	Divorced	4011	82
23	36	Yes	Travel_Rarely	1218	Sales	9	4	Life Sciences	1	27	3	Male	82	2	1	Sales Represen	1	Single	3407	69
24	34	No	Travel_Rarely	419	Research & Dev	7	4	Life Sciences	1	28	1	Female	53	3	3	Research Direct	2	Single	11994	212
25	21	No	Travel_Rarely	391	Research & Dev	15	2	Life Sciences	1	30	3	Male	96	3	1	Research Scient	4	Single	1232	192
26	34	Yes	Travel_Rarely	699	Research & Dev	6	1	Medical	1	31	2	Male	83	3	1	Research Scient	1	Single	2960	171
27	53	No	Travel_Rarely	1282	Research & Dev	5	3	Other	1	32	3	Female	58	3	5	Manager	3	Divorced	19094	107
28	32	Yes	Travel_Frequent	1125	Research & Dev	16	1	Life Sciences	1	33	2	Female	72	1	1	Research Scient	1	Single	3919	46
29	42	No	Travel_Rarely	691	Sales	8	4	Marketing	1	35	3	Male	48	3	2	Sales Executive	2	Married	6825	211
30	44	No	Travel_Rarely	477	Research & Dev	7	4	Medical	1	36	1	Female	42	2	3	Healthcare Repr	4	Married	10248	20
31	46	No	Travel_Rarely	705	Sales	2	4	Marketing	1	38	2	Female	83	3	5	Manager	1	Single	18947	228
32	33	No	Travel_Rarely	924	Research & Dev	2	3	Medical	1	39	3	Male	78	3	1	Laboratory Tech	4	Single	2496	66
33	44	No	Travel_Rarely	1459	Research & Dev	10	4	Other	1	40	4	Male	41	3	2	Healthcare Repr	4	Married	6465	191

STEP 3: Convert CSV to ARFF

(actually not needed)

STEP 4: Load Dataset in Weka Explorer



Filter: None

Current relation
Relation: WA_Fn-UseC_-HR-Employee-Attrition - WA_Fn-UseC_-HR-Employee-Attrition.csv
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

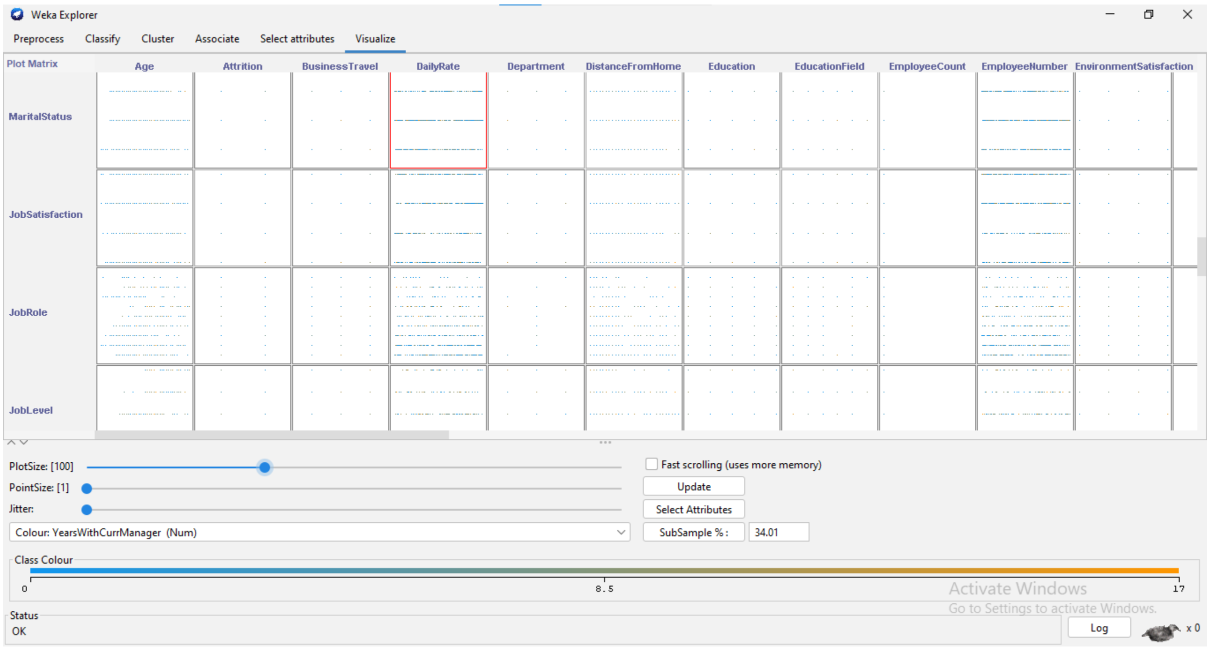
Class: YearsWithCurrManager (Num)

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

Status
OK

STEP 5: Explore Attributes & Visualize Patterns



PlotSize: [100]

PointSize: [1]

Jitter:

Colour: YearsWithCurrManager (Num)

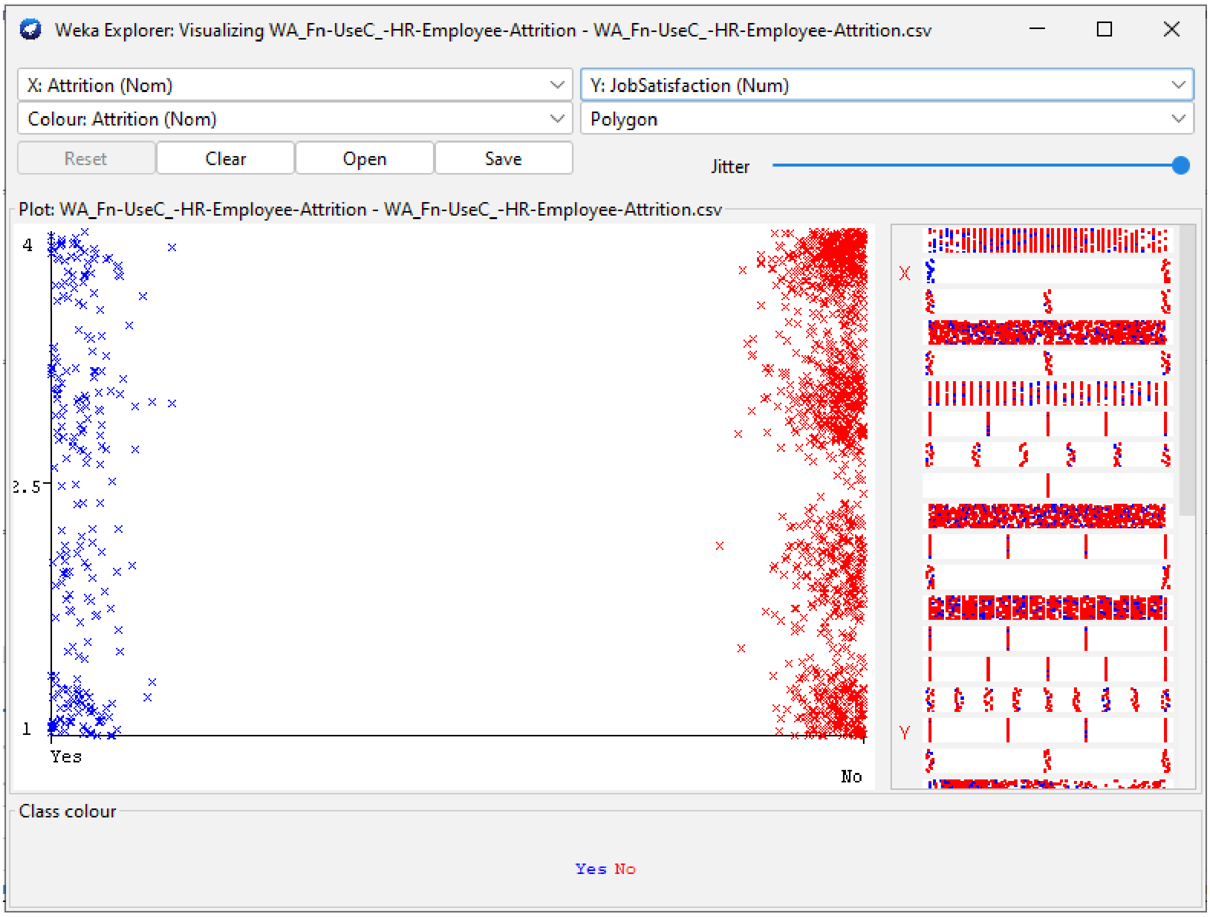
☐ Fast scrolling (uses more memory)

SubSample % : 34.01

Status

OK

- Attrition vs Job Satisfaction


Weka Explorer: Visualizing WA_Fn-UseC_-HR-Employee-Attrition - WA_Fn-UseC_-HR-Employee-Attrition.csv
Reset
Clear
Open
Save
Jitter
Plot: WA_Fn-UseC_-HR-Employee-Attrition - WA_Fn-UseC_-HR-Employee-Attrition.csv
4
2.5
1
Yes
No
X
Y
Class colour
Yes No

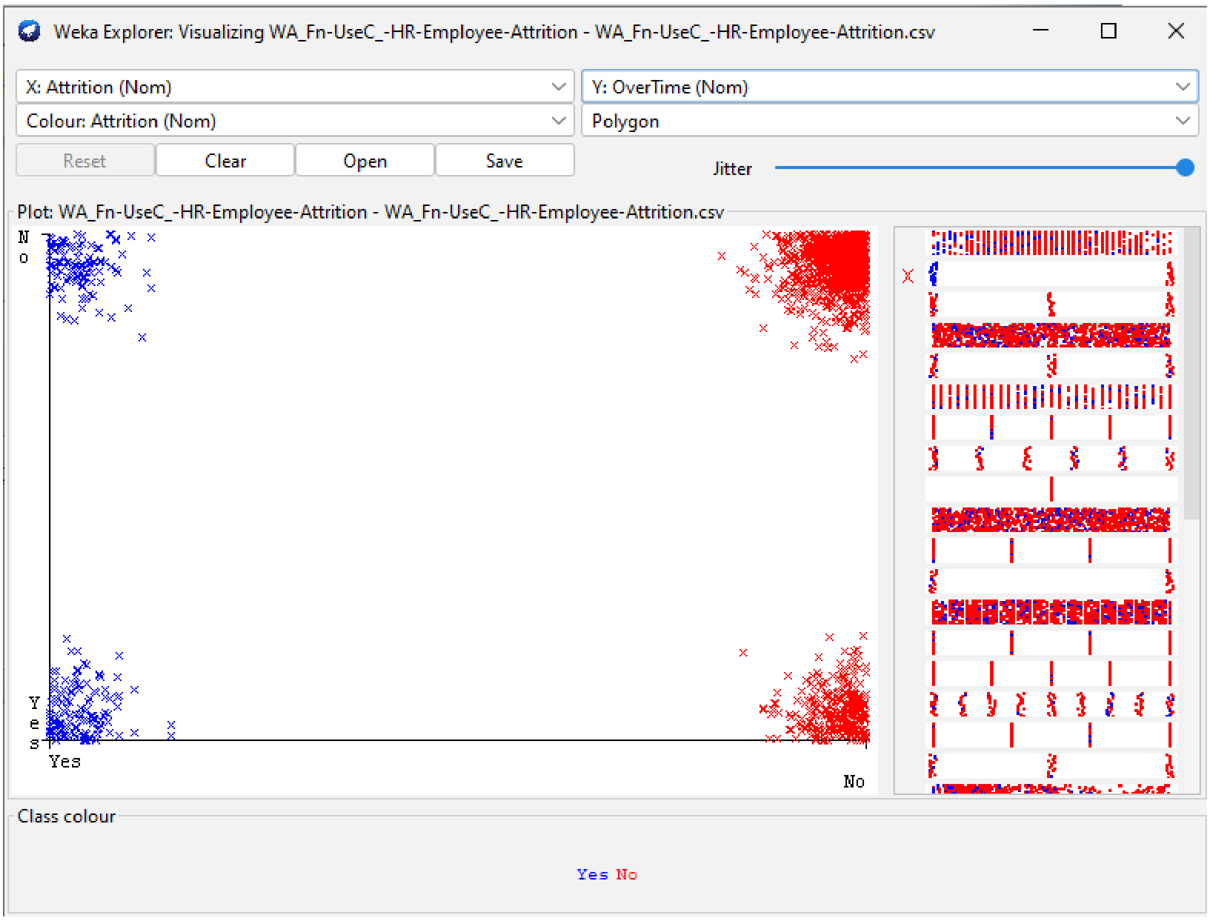
X: Attrition (Nom)

Y: JobSatisfaction (Num)

Colour: Attrition (Nom)

Polygon

- Attrition vs OverTime



Weka Explorer: Visualizing WA_Fn-UseC_-HR-Employee-Attrition - WA_Fn-UseC_-HR-Employee-Attrition.csv

X: Attrition (Nom)

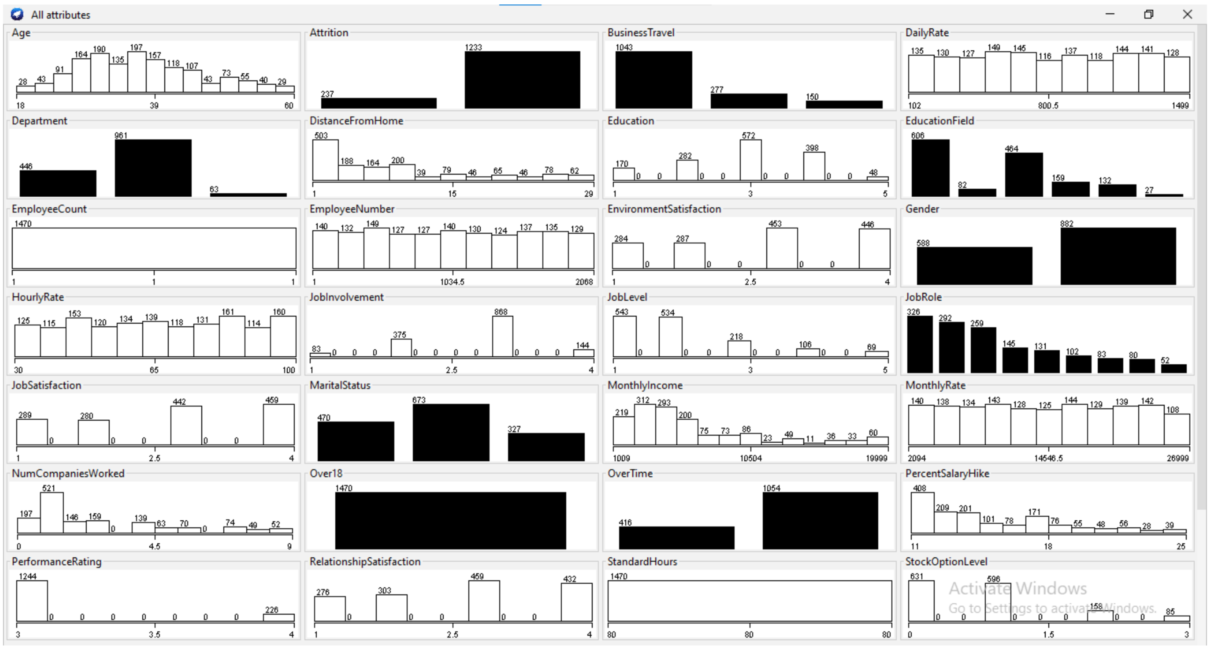
Y: OverTime (Nom)

Colour: Attrition (Nom)

Polygon

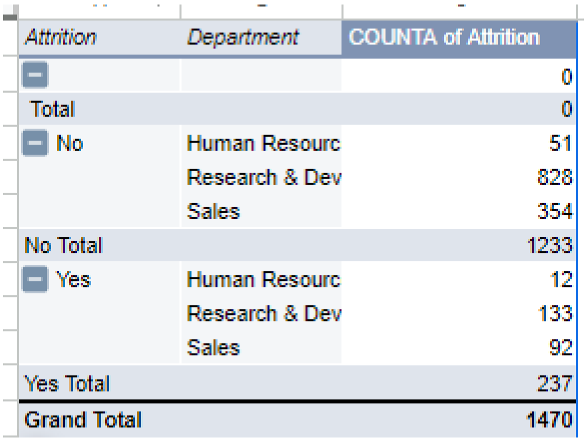
Plot: WA_Fn-UseC_-HR-Employee-Attrition - WA_Fn-UseC_-HR-Employee-Attrition.csv

- Histograms

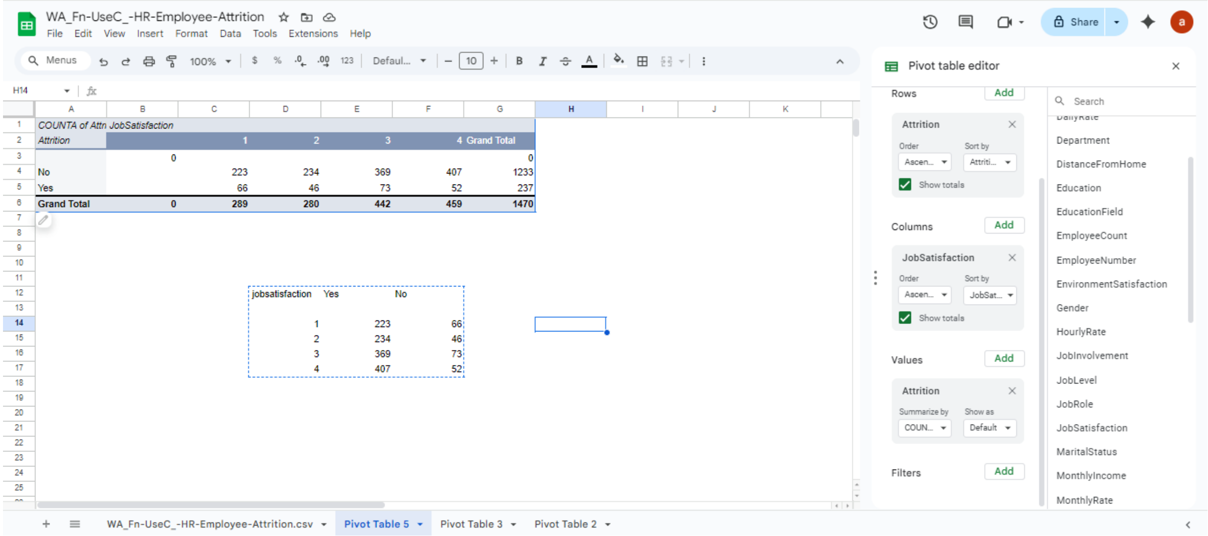


All attributes

STEP 6: Create Visual Narratives in Flourish



Attrition	Department	COUNTA of Attrition
⊟		0
Total		0
⊟ No	Human Resourc	51
	Research & Dev	828
	Sales	354
No Total		1233
⊟ Yes	Human Resourc	12
	Research & Dev	133
	Sales	92
Yes Total		237
Grand Total		1470



WA_Fn-UseC_-HR-Employee-Attrition

H14

COUNTA of Attn	JobSatisfaction					
Attrition		1	2	3	4	Grand Total
	0					0
No		223	234	369	407	1233
Yes		66	46	73	52	237
Grand Total	0	289	280	442	459	1470

jobsatisfaction	Yes	No
1	223	66
2	234	46
3	369	73
4	407	52

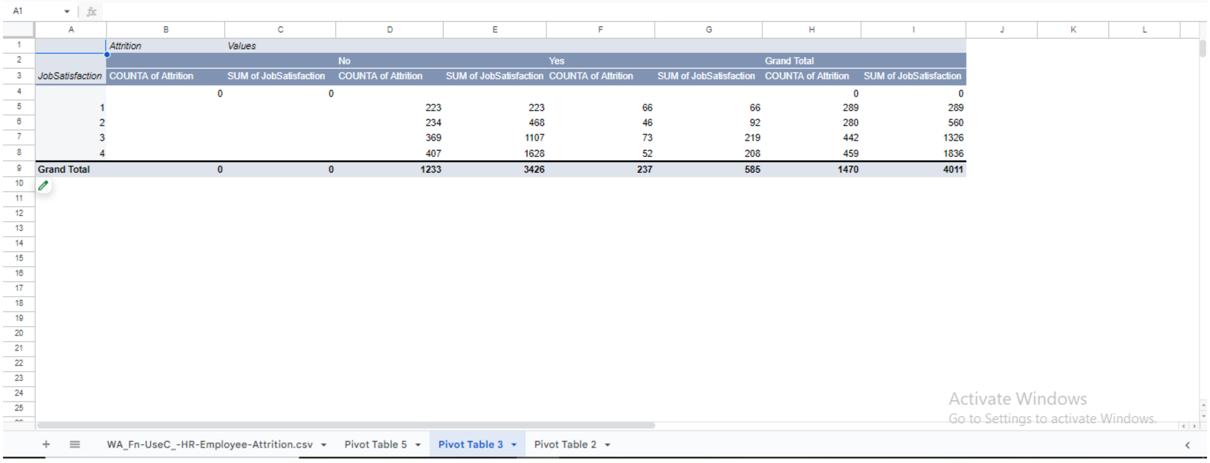
Pivot table editor

Rows: Attrition — Order: Ascen... — Sort by: Attriti... — ☑ Show totals

Columns: JobSatisfaction — Order: Ascen... — Sort by: JobSat... — ☑ Show totals

Values: Attrition — Summarize by: COUN... — Show as: Default

Filters



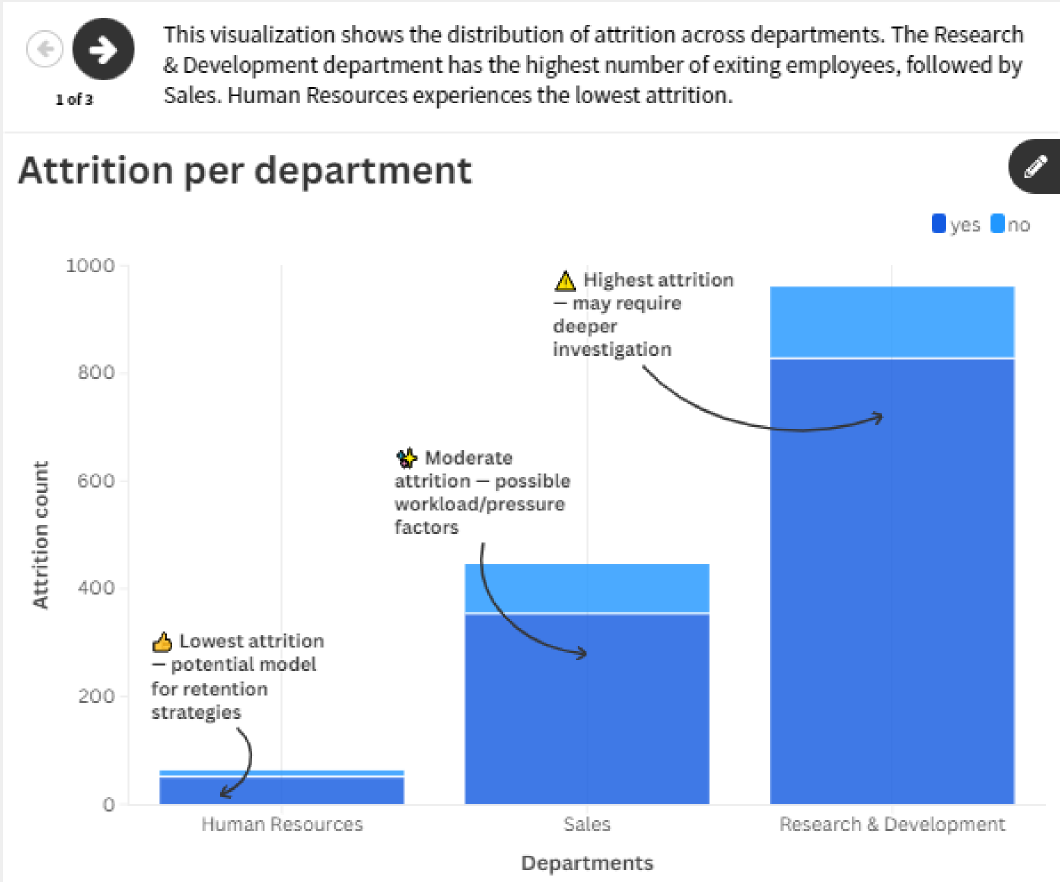
A1

	Attrition	Values						
			No		Yes		Grand Total	
JobSatisfaction	COUNTA of Attrition	SUM of JobSatisfaction	COUNTA of Attrition	SUM of JobSatisfaction	COUNTA of Attrition	SUM of JobSatisfaction	COUNTA of Attrition	SUM of JobSatisfaction
	0	0					0	0
1			223	223	66	66	289	289
2			234	468	46	92	280	560
3			369	1107	73	219	442	1326
4			407	1628	52	208	459	1836
Grand Total	0	0	1233	3426	237	585	1470	4011

Activate Windows
Go to Settings to activate Windows.
WA_Fn-UseC_-HR-Employee-Attrition.csv
Pivot Table 5
Pivot Table 3
Pivot Table 2

Click for the story : https://public.flourish.studio/story/3459871/

Screen shots :

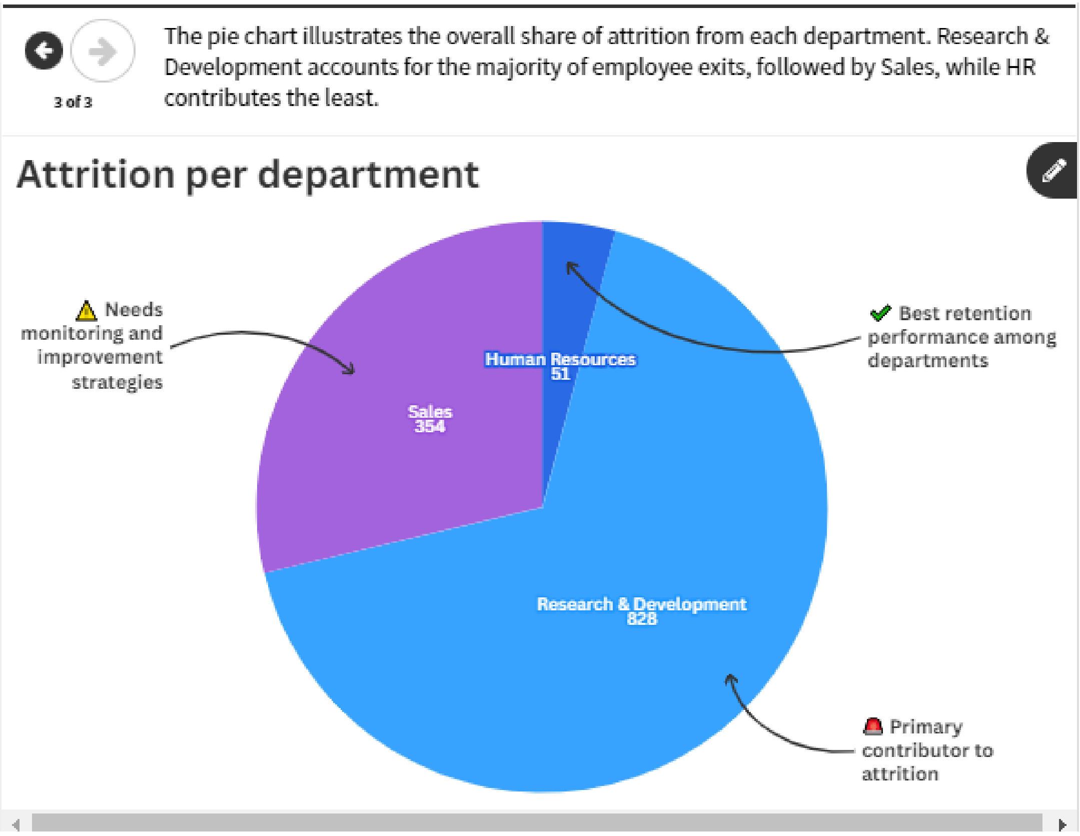


1 of 3

This visualization shows the distribution of attrition across departments. The Research & Development department has the highest number of exiting employees, followed by Sales. Human Resources experiences the lowest attrition.

Attrition per department

yes
no
Attrition count
1000
800
600
400
200
0
Human Resources
Sales
Research & Development
Departments
⚠ Highest attrition – may require deeper investigation
Moderate attrition – possible workload/pressure factors
Lowest attrition – potential model for retention strategies



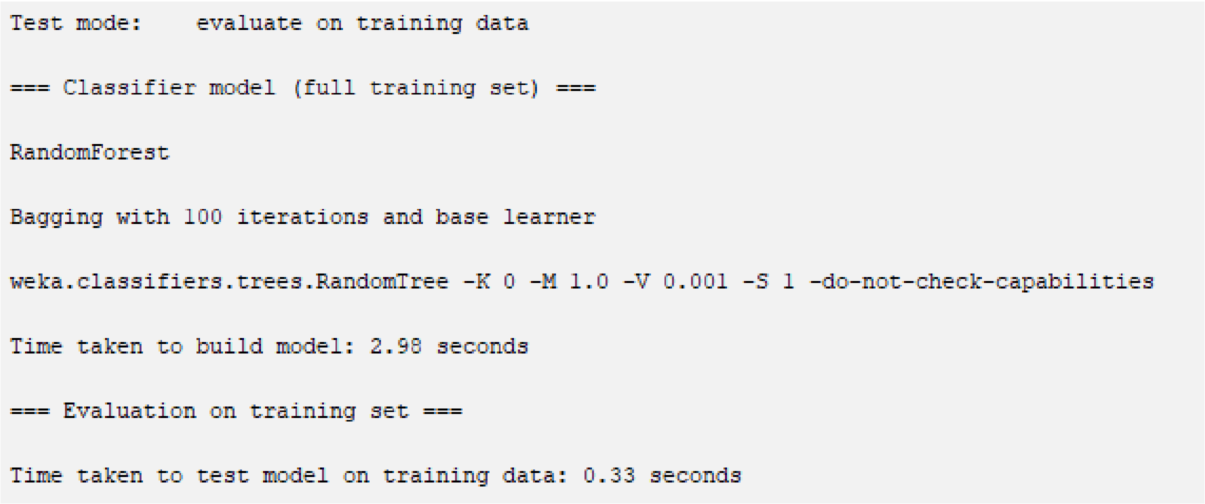
3 of 3

The pie chart illustrates the overall share of attrition from each department. Research & Development accounts for the majority of employee exits, followed by Sales, while HR contributes the least.

Attrition per department

Human Resources
51
Sales
354
Research & Development
828
⚠ Needs monitoring and improvement strategies
✓ Best retention performance among departments
Primary contributor to attrition

Step 7: Train Baseline Model



```
Test mode:    evaluate on training data

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 2.98 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.33 seconds
```

=== Summary ===

```
Correlation coefficient              0.97981
Mean absolute error                   0.6278
Root mean squared error                 0.8194
Relative absolute error              30.4805 %
Root relative squared error            32.8113 %
Total Number of Instances            1470
```

```
Number of Leaves  :       2225
Size of the tree :        2247
```

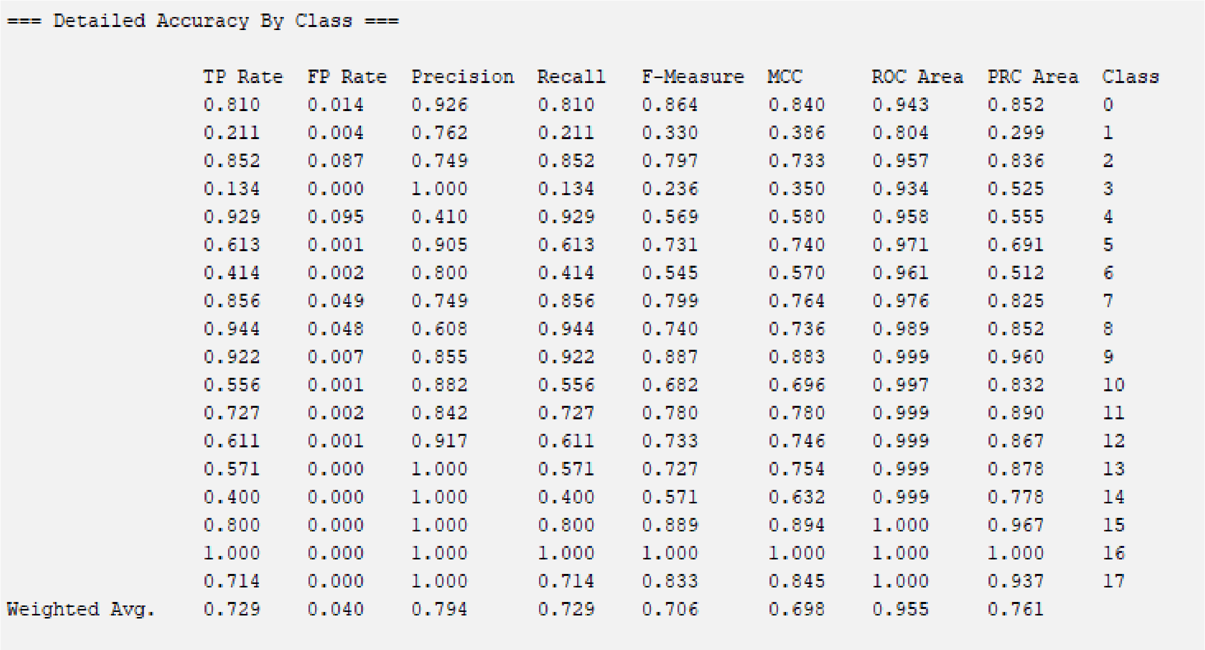
Time taken to build model: 0.05 seconds

=== Evaluation on training set ===

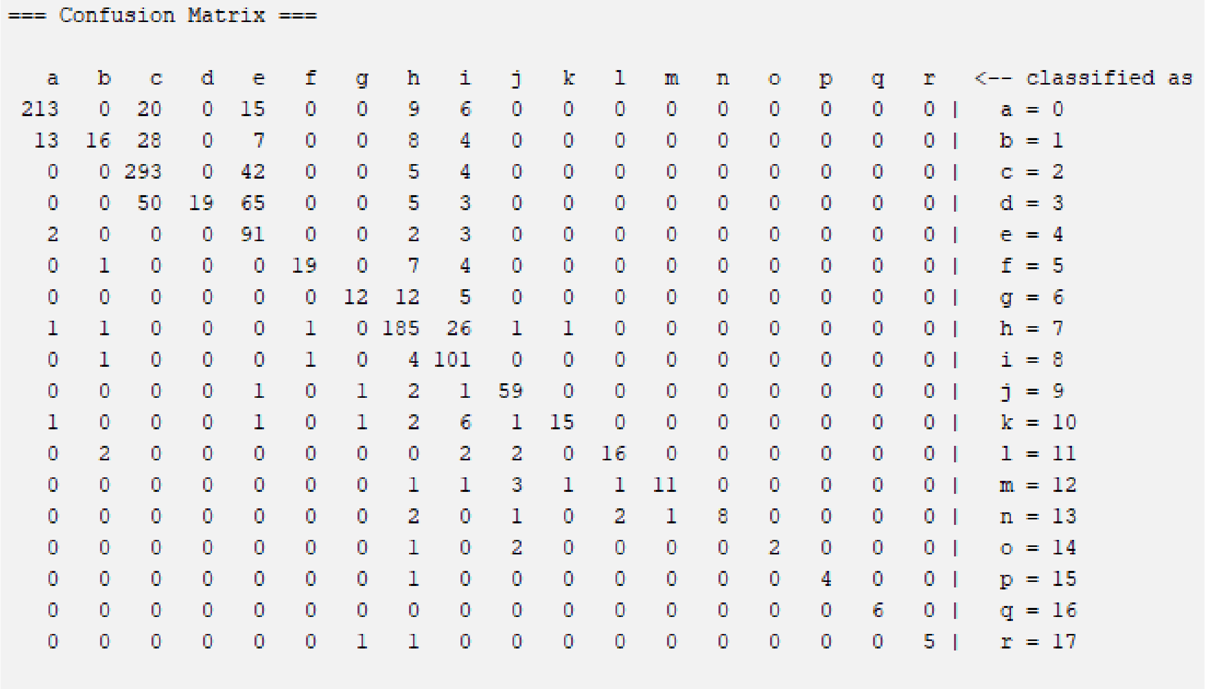
Time taken to test model on training data: 0.07 seconds

=== Summary ===

```
Correctly Classified Instances       1075             72.9308 %
Incorrectly Classified Instances       399             27.0692 %
Kappa statistic                    0.6861
Mean absolute error                   0.0377
Root mean squared error                 0.1373
Relative absolute error              39.0965 %
Root relative squared error            62.5605 %
Total Number of Instances            1474
```



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.810    0.014    0.926      0.810    0.864      0.840    0.943     0.852     0
                 0.211    0.004    0.762      0.211    0.330      0.386    0.804     0.299     1
                 0.852    0.087    0.749      0.852    0.797      0.733    0.957     0.836     2
                 0.134    0.000    1.000      0.134    0.236      0.350    0.934     0.525     3
                 0.929    0.095    0.410      0.929    0.569      0.580    0.958     0.555     4
                 0.613    0.001    0.905      0.613    0.731      0.740    0.971     0.691     5
                 0.414    0.002    0.800      0.414    0.545      0.570    0.961     0.512     6
                 0.856    0.049    0.749      0.856    0.799      0.764    0.976     0.825     7
                 0.944    0.048    0.608      0.944    0.740      0.736    0.989     0.852     8
                 0.922    0.007    0.855      0.922    0.887      0.883    0.999     0.960     9
                 0.556    0.001    0.882      0.556    0.682      0.696    0.997     0.832     10
                 0.727    0.002    0.842      0.727    0.780      0.780    0.999     0.890     11
                 0.611    0.001    0.917      0.611    0.733      0.746    0.999     0.867     12
                 0.571    0.000    1.000      0.571    0.727      0.754    0.999     0.878     13
                 0.400    0.000    1.000      0.400    0.571      0.632    0.999     0.778     14
                 0.800    0.000    1.000      0.800    0.889      0.894    1.000     0.967     15
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     16
                 0.714    0.000    1.000      0.714    0.833      0.845    1.000     0.937     17
Weighted Avg.    0.729    0.040    0.794      0.729    0.706      0.698    0.955     0.761
```



```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   <-- classified as
 213   0  20   0  15   0   0   9   6   0   0   0   0   0   0   0   0   0 |   a = 0
  13  16  28   0   7   0   0   8   4   0   0   0   0   0   0   0   0   0 |   b = 1
   0   0 293   0  42   0   0   5   4   0   0   0   0   0   0   0   0   0 |   c = 2
   0   0  50  19  65   0   0   5   3   0   0   0   0   0   0   0   0   0 |   d = 3
   2   0   0   0  91   0   0   2   3   0   0   0   0   0   0   0   0   0 |   e = 4
   0   1   0   0   0  19   0   7   4   0   0   0   0   0   0   0   0   0 |   f = 5
   0   0   0   0   0   0  12  12   5   0   0   0   0   0   0   0   0   0 |   g = 6
   1   1   0   0   0   1   0 185  26   1   1   0   0   0   0   0   0   0 |   h = 7
   0   1   0   0   0   1   0   4 101   0   0   0   0   0   0   0   0   0 |   i = 8
   0   0   0   0   1   0   1   2   1  59   0   0   0   0   0   0   0   0 |   j = 9
   1   0   0   0   1   0   1   2   6   1  15   0   0   0   0   0   0   0 |   k = 10
   0   2   0   0   0   0   0   0   2   2   0  16   0   0   0   0   0   0 |   l = 11
   0   0   0   0   0   0   0   1   1   3   1   1  11   0   0   0   0   0 |   m = 12
   0   0   0   0   0   0   0   2   0   1   0   2   1   8   0   0   0   0 |   n = 13
   0   0   0   0   0   0   0   1   0   2   0   0   0   0   2   0   0   0 |   o = 14
   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   4   0   0 |   p = 15
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   6   0 |   q = 16
   0   0   0   0   0   0   1   1   0   0   0   0   0   0   0   0   0   5 |   r = 17
```

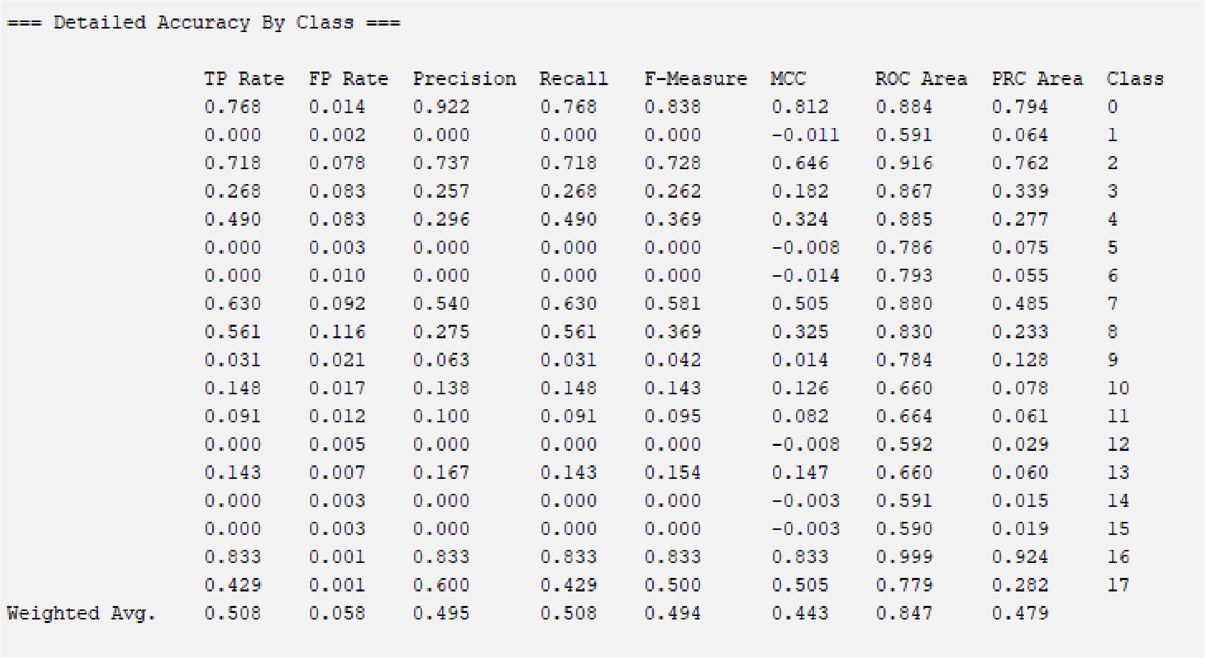
**Number of Leaves :	2225**
**Size of the tree :	2247**

Time taken to build model: 0.12 seconds

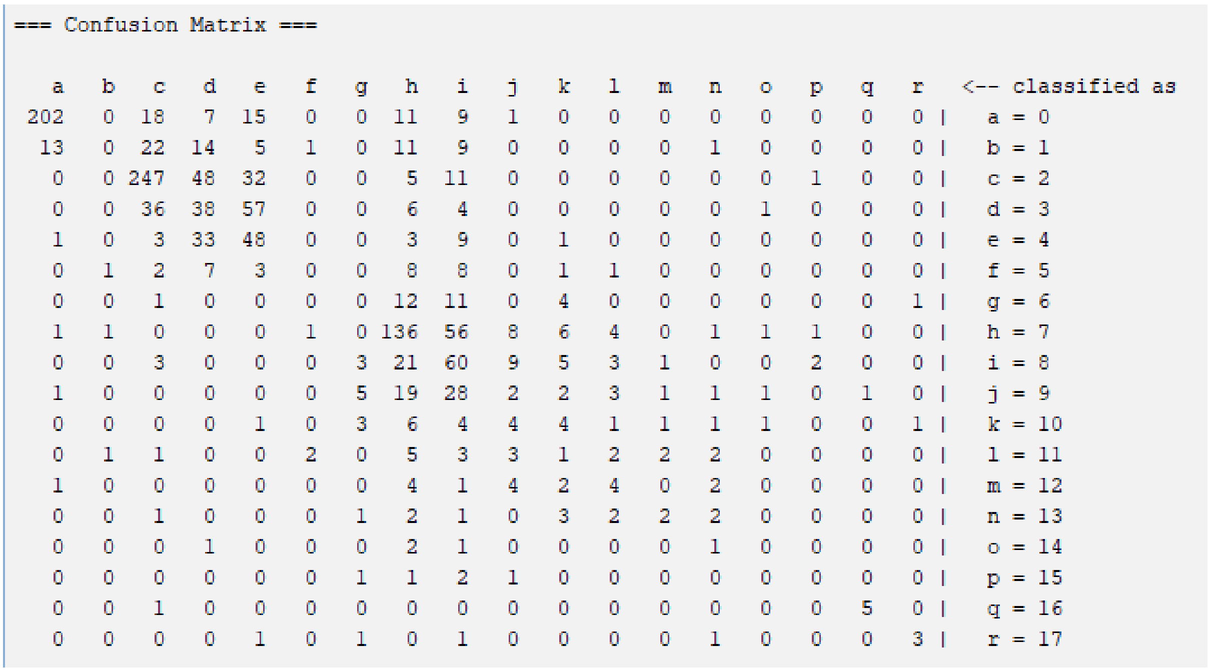
=== Stratified cross-validation ===

=== Summary ===

```
Correctly Classified Instances        749            50.8141 %
Incorrectly Classified Instances       725            49.1859 %
Kappa statistic                    0.4318
Mean absolute error                    0.0609
Root mean squared error                  0.1874
Relative absolute error              63.1553 %
Root relative squared error            85.3806 %
Total Number of Instances            1474
```



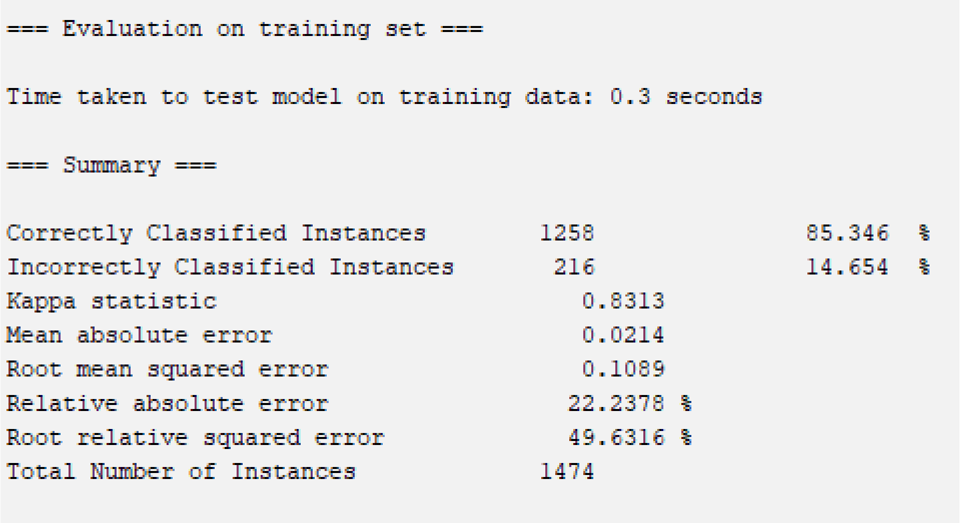
```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.768    0.014    0.922      0.768    0.838      0.812    0.884     0.794     0
                 0.000    0.002    0.000      0.000    0.000      -0.011   0.591     0.064     1
                 0.718    0.078    0.737      0.718    0.728      0.646    0.916     0.762     2
                 0.268    0.083    0.257      0.268    0.262      0.182    0.867     0.339     3
                 0.490    0.083    0.296      0.490    0.369      0.324    0.885     0.277     4
                 0.000    0.003    0.000      0.000    0.000      -0.008   0.786     0.075     5
                 0.000    0.010    0.000      0.000    0.000      -0.014   0.793     0.055     6
                 0.630    0.092    0.540      0.630    0.581      0.505    0.880     0.485     7
                 0.561    0.116    0.275      0.561    0.369      0.325    0.830     0.233     8
                 0.031    0.021    0.063      0.031    0.042      0.014    0.784     0.128     9
                 0.148    0.017    0.138      0.148    0.143      0.126    0.660     0.078     10
                 0.091    0.012    0.100      0.091    0.095      0.082    0.664     0.061     11
                 0.000    0.005    0.000      0.000    0.000      -0.008   0.592     0.029     12
                 0.143    0.007    0.167      0.143    0.154      0.147    0.660     0.060     13
                 0.000    0.003    0.000      0.000    0.000      -0.003   0.591     0.015     14
                 0.000    0.003    0.000      0.000    0.000      -0.003   0.590     0.019     15
                 0.833    0.001    0.833      0.833    0.833      0.833    0.999     0.924     16
                 0.429    0.001    0.600      0.429    0.500      0.505    0.779     0.282     17
Weighted Avg.    0.508    0.058    0.495      0.508    0.494      0.443    0.847     0.479
```



```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   <-- classified as
 202   0  18   7  15   0   0  11   9   1   0   0   0   0   0   0   0   0 |   a = 0
  13   0  22  14   5   1   0  11   9   0   0   0   0   1   0   0   0   0 |   b = 1
   0   0 247  48  32   0   0   5  11   0   0   0   0   0   0   1   0   0 |   c = 2
   0   0  36  38  57   0   0   6   4   0   0   0   0   0   1   0   0   0 |   d = 3
   1   0   3  33  48   0   0   3   9   0   1   0   0   0   0   0   0   0 |   e = 4
   0   1   2   7   3   0   0   8   8   0   1   1   0   0   0   0   0   0 |   f = 5
   0   0   1   0   0   0   0  12  11   0   4   0   0   0   0   0   0   1 |   g = 6
   1   1   0   0   0   1   0 136  56   8   6   4   0   1   1   1   0   0 |   h = 7
   0   0   3   0   0   0   3  21  60   9   5   3   1   0   0   2   0   0 |   i = 8
   1   0   0   0   0   0   5  19  28   2   2   3   1   1   1   0   1   0 |   j = 9
   0   0   0   0   1   0   3   6   4   4   4   1   1   1   1   0   0   1 |   k = 10
   0   1   1   0   0   2   0   5   3   3   1   2   2   2   0   0   0   0 |   l = 11
   1   0   0   0   0   0   0   4   1   4   2   4   0   2   0   0   0   0 |   m = 12
   0   0   1   0   0   0   1   2   1   0   3   2   2   2   0   0   0   0 |   n = 13
   0   0   0   1   0   0   0   2   1   0   0   0   0   1   0   0   0   0 |   o = 14
   0   0   0   0   0   0   1   1   2   1   0   0   0   0   0   0   0   0 |   p = 15
   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   5   0 |   q = 16
   0   0   0   0   1   0   1   0   1   0   0   0   0   1   0   0   0   3 |   r = 17
```

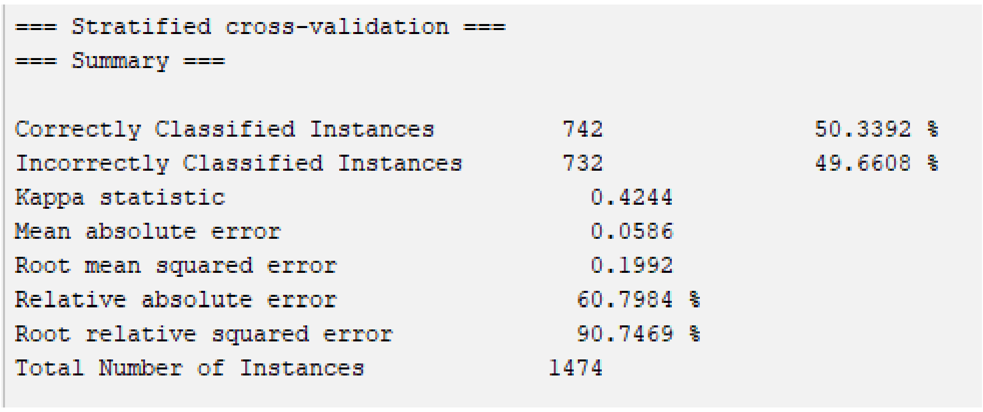
Overfitting

Using naïve bayes as training results



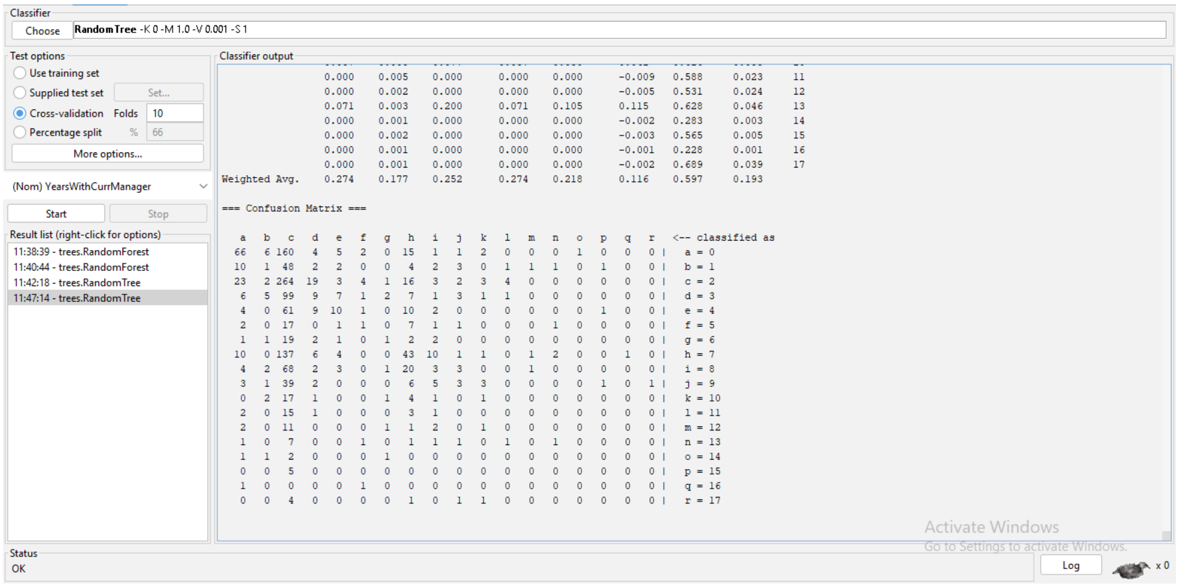
```
=== Evaluation on training set ===

Time taken to test model on training data: 0.3 seconds

=== Summary ===

Correctly Classified Instances        1258               85.346  %
Incorrectly Classified Instances       216               14.654  %
Kappa statistic                          0.8313
Mean absolute error                      0.0214
Root mean squared error                  0.1089
Relative absolute error                 22.2378 %
Root relative squared error             49.6316 %
Total Number of Instances             1474
```

Using naïve bayes for cross validation



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

Correctly Classified Instances         742               50.3392 %
Incorrectly Classified Instances       732               49.6608 %
Kappa statistic                          0.4244
Mean absolute error                      0.0586
Root mean squared error                  0.1992
Relative absolute error                 60.7984 %
Root relative squared error             90.7469 %
Total Number of Instances             1474
```

By using random forest without converting all data to nominal and removing unrelated rows



Classifier

Choose RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) YearsWithCurrManager

Result list (right-click for options)

11:38:39 - trees.RandomForest
11:40:44 - trees.RandomForest
11:42:18 - trees.RandomTree
11:47:14 - trees.RandomTree

Classifier output

```
                 0.000    0.005    0.000      0.000    0.000      -0.009   0.588     0.023     11
                 0.000    0.002    0.000      0.000    0.000      -0.005   0.531     0.024     12
                 0.071    0.003    0.200      0.071    0.105      0.115    0.628     0.046     13
                 0.000    0.001    0.000      0.000    0.000      -0.002   0.283     0.003     14
                 0.000    0.002    0.000      0.000    0.000      -0.003   0.565     0.005     15
                 0.000    0.001    0.000      0.000    0.000      -0.001   0.228     0.001     16
                 0.000    0.001    0.000      0.000    0.000      -0.002   0.689     0.039     17
Weighted Avg.    0.274    0.177    0.252      0.274    0.218      0.116    0.597     0.193

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   <-- classified as
  66   6 160   4   5   2   0  15   1   1   2   0   0   0   1   0   0   0 |   a = 0
  10   1  48   2   2   0   0   4   2   3   0   1   1   1   0   1   0   0 |   b = 1
  23   2 264  19   3   4   1  16   3   2   3   4   0   0   0   0   0   0 |   c = 2
   6   5  99   9   7   1   2   7   1   3   1   1   0   0   0   0   0   0 |   d = 3
   4   0  61   9  10   1   0  10   2   0   0   0   0   0   0   1   0   0 |   e = 4
   2   0  17   0   1   1   0   7   1   1   0   0   0   1   0   0   0   0 |   f = 5
   1   1  19   2   1   0   1   2   2   0   0   0   0   0   0   0   0   0 |   g = 6
  10   0 137   6   4   0   0  43  10   1   1   0   1   2   0   0   1   0 |   h = 7
   4   2  68   2   3   0   1  20   3   3   0   0   1   0   0   0   0   0 |   i = 8
   3   1  39   2   0   0   0   6   5   3   3   0   0   0   0   1   0   1 |   j = 9
   0   2  17   1   0   0   1   4   1   0   1   0   0   0   0   0   0   0 |   k = 10
   2   0  15   1   0   0   0   3   1   0   0   0   0   0   0   0   0   0 |   l = 11
   2   0  11   0   0   0   1   1   2   0   1   0   0   0   0   0   0   0 |   m = 12
   1   0   7   0   0   1   0   1   1   1   0   1   0   1   0   0   0   0 |   n = 13
   1   1   2   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0 |   o = 14
   0   0   5   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   p = 15
   1   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0 |   q = 16
   0   0   4   0   0   0   0   1   0   1   1   0   0   0   0   0   0   0 |   r = 17
```

Status

OK

=== Detailed Accuracy By Class ===

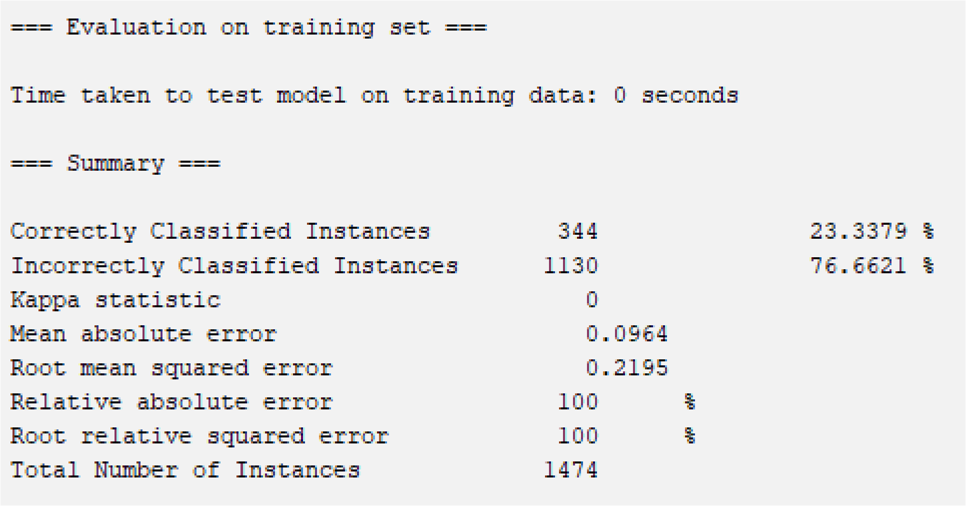
```
           TP Rate  FP Rate  Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Class

           0.251    0.058    0.485      0.251    0.331      0.255    0.629    0.299     0
           0.013    0.014    0.048      0.013    0.021      -0.002   0.449    0.048     1
           0.767    0.630    0.271      0.767    0.401      0.123    0.614    0.310     2
           0.063    0.036    0.158      0.063    0.090      0.042    0.572    0.132     3
           0.102    0.019    0.278      0.102    0.149      0.134    0.595    0.109     4
           0.032    0.007    0.091      0.032    0.048      0.042    0.544    0.027     5
           0.034    0.005    0.125      0.034    0.054      0.056    0.579    0.030     6
           0.199    0.077    0.307      0.199    0.242      0.147    0.622    0.215     7
           0.028    0.023    0.086      0.028    0.042      0.008    0.571    0.085     8
           0.047    0.011    0.158      0.047    0.072      0.064    0.629    0.077     9
           0.037    0.008    0.077      0.037    0.050      0.041    0.616    0.053     10
           0.000    0.005    0.000      0.000    0.000      -0.009   0.588    0.023     11
           0.000    0.002    0.000      0.000    0.000      -0.005   0.531    0.024     12
           0.071    0.003    0.200      0.071    0.105      0.115    0.628    0.046     13
           0.000    0.001    0.000      0.000    0.000      -0.002   0.283    0.003     14
           0.000    0.002    0.000      0.000    0.000      -0.003   0.565    0.005     15
           0.000    0.001    0.000      0.000    0.000      -0.001   0.228    0.001     16
           0.000    0.001    0.000      0.000    0.000      -0.002   0.689    0.039     17

Weighted Avg.  0.274  0.177  0.252     0.274   0.218      0.116    0.597    0.193
```

=== Summary ===

```
Correctly Classified Instances        403          27.415  %
Incorrectly Classified Instances      1067          72.585  %
Kappa statistic                   0.099
Mean absolute error                  0.089
Root mean squared error                0.231
Relative absolute error              92.3523 %
Root relative squared error           105.2777 %
Total Number of Instances            1470
```

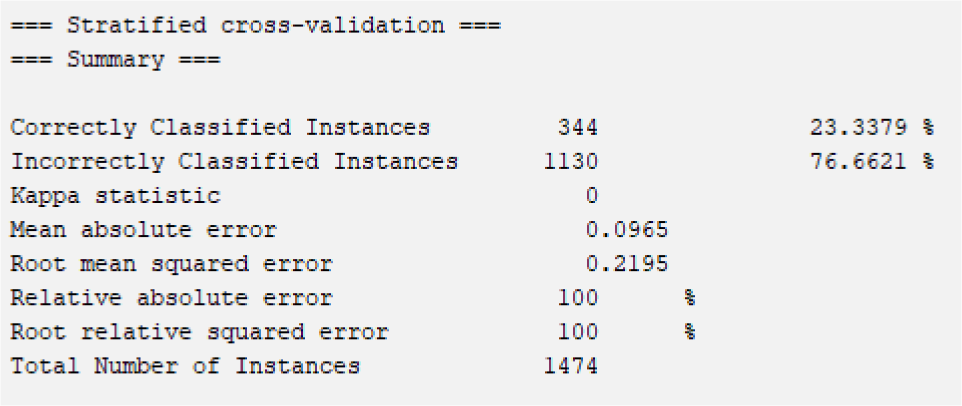
Underfitting

Using zeroR classifiervto test results
As a training set



```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         344               23.3379 %
Incorrectly Classified Instances      1130               76.6621 %
Kappa statistic                          0
Mean absolute error                      0.0964
Root mean squared error                  0.2195
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1474
```

Cross validating



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

Correctly Classified Instances         344               23.3379 %
Incorrectly Classified Instances      1130               76.6621 %
Kappa statistic                          0
Mean absolute error                      0.0965
Root mean squared error                  0.2195
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1474
```

Step 10: Summarize Insights : using napkin

Napkin ai summary link :

https://app.napkin.ai/page/CgoiCHByb2Qtb25lEiwKBFBhZ2UaJDc5ZDE5MDc3LWMxMjUtNGEyMi1hNWZhLWY3NDQ2ODEyMmVhNA?s=1

summary main points

- The evaluation of models (J48, Naive Bayes, Random Forest, ZeroR) showed mixed performance, with J48 suffering from **overfitting** and Random Forest and ZeroR showing **underfitting**, indicating improper learning of meaningful patterns.

- A major challenge identified across models was **class imbalance**, where "no attrition" cases dominate the dataset, causing models to bias predictions toward the majority class and reduce detection accuracy of actual attrition cases.
- The inclusion of irrelevant features like **EmployeeNumber** negatively impacted learning by creating meaningless model splits; removing such non-predictive features is essential to improve model interpretability and accuracy.
- The summary highlighted the need for techniques like **resampling, cost-sensitive learning, proper feature selection, and hyperparameter tuning** to improve prediction performance and handle imbalanced data effectively.
- Potential key drivers of attrition include factors such as **job satisfaction, salary, work-life balance, tenure, promotion history, and workload (overtime)**, although further feature importance analysis is needed once the model is improved.
- Based on insights, HR can leverage refined predictive models to identify high-risk employees and implement targeted retention strategies (e.g., salary review, career development programs, flexible working policies, and department-specific interventions).

Reflection Questions

Answer any 3:

1. Which factors contributed most to attrition?
2. Did your model overfit or underfit? Why?
3. How can dataset quality be improved?
4. How should HR teams use this model?

Ans 1 - From the patterns observed, employees with low job satisfaction, poor work-life balance, lower salaries, and limited growth opportunities seemed more likely to leave. Tenure and overtime also appeared to influence the likelihood of attrition.

Ans2 - Yes — one model overfitted and another underfitted. The J48 model overfitted because it learned the training data too perfectly, including noise. Meanwhile, Random Forest and ZeroR underfitted due to class imbalance and irrelevant features affecting learning.

Ans3 - The dataset can be improved by removing non-useful columns like EmployeeNumber, balancing the classes, and cleaning or encoding categorical values properly. Doing this will help the model learn real patterns instead of random or misleading ones.

Overall, the mini-project demonstrated how different machine learning models behave when applied to HR attrition data. The experiment helped me understand how data imbalance, irrelevant features, and model settings affect prediction accuracy. After improving feature selection and handling imbalance, the model can become more useful for HR teams to identify employees at risk and take preventive actions.

Final dataset :

https://docs.google.com/spreadsheets/d/1FvJYP03OPrZUGc3dTIgn6uXQWM7MtOs21azyTC5QoO0/edit?usp=sharing

---

Submitted by Abduttaiyeb Huseni Matcheswala (b24bs1015)