

**BASE DE DATOS II**

*TALLER DE APLICACIÓN*

# La Aplicación del Proceso de Descubrimiento (KDD) mediante WEKA.

TEMA: ESTUDIO DEL AUSENTISMO EN UN ENTORNO DE TRABAJO

Integrantes:

- -Caro, Mariel Alejandra
- -Juárez, Pablo
- -Orieta, Leandro

## INTRODUCCIÓN:

Este taller pretende aplicar los conocimientos adquiridos del proceso de descubrimiento KDD. Cuando hablamos de grandes cantidades de datos, el Descubrimiento de Conocimiento en Bases de Datos o KDD se refiere al proceso de identificar patrones válidos, novedosos, potencialmente útiles y entendibles en un gran conjunto de datos almacenados, que constituyen a su vez una fuente de conocimiento.

Como es un proceso, para realizarlo, se deben seguir un conjunto de seis pasos fundamentales empleados en el presente taller: Integración y recopilación, Selección, limpieza y transformación, Minería de Datos, Evaluación e Interpretación, Difusión y Uso.

Se aplicaron estos pasos en datos obtenidos de una base de datos creada con registros de ausentismo en el trabajo desde julio del 2007 hasta Julio del 2010 en una compañía de delivery en Brasil. Y el propósito de este trabajo es encontrar las causas por las que los trabajadores faltan a sus puestos de trabajo, llegan tarde o se retiran de sus labores.

## DESCRIPCIÓN DEL CONJUNTO DE DATOS:

- Tipo de datos Multivariados
- Técnicas utilizadas de minería de datos: Clasificación, Clustering.
- Tipo de atributos: categóricos, numéricos enteros y reales.
- Número total de instancias: 740.
- Número total de atributos: 21.

Atributos:

1. ID de personal.
2. Razones de ausencia: divididas en 4 categorías correspondientes al International Code of Diseases.
3. Meses de ausencia.
4. Días de la semana.
5. Estaciones del año.
6. Costos de transporte.
7. Distancia desde el hogar al trabajo en kilómetros.
8. Tiempo de servicio.
9. Edad
10. Promedio de carga de trabajo por día.
11. Alcance de objetivos.
12. Fallas disciplinarias.
13. Educación alcanzada.
14. Cantidad de hijos.
15. Alcohólicos.
16. Fumadores.
17. Mascotas.
18. Peso.
19. Altura.
20. Índice de masa corporal.
21. Horas de ausentismo.



# TÉCNICAS APLICADAS:

## CLUSTERING

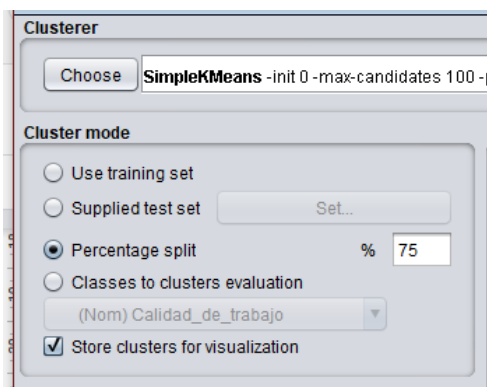
El clustering es uno de los métodos de aprendizaje no supervisado más importantes y busca caracterizar conceptos desconocidos a partir de los ejemplos disponibles. Generalmente, en un problema real se desconoce la clase y es allí donde el agrupamiento puede ayudar a identificar las características comunes entre instancias. Al no disponer de la clase utiliza una medida de similitud (distancia) para determinar el parecido entre instancias.

Permite encontrar grupos de instancias con características similares. Con este método se pudieron identificar las clases iniciales del conjunto de datos, en base a un conjunto de atributos.

### Simple K-Means

K-medias es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de  $n$  observaciones en  $k$  grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Se utilizó en todo el método de clustering con el modo división por porcentajes que utiliza el 25% para entrenamiento y el 75% para testeo.

#### Configuración inicial:



En todas las ejecuciones se utilizó la misma configuración inicial

#### PRUEBAS CLUSTER:

##### PRIMER CRITERIO DE PRUEBA - ATRIBUTOS SELECCIONADOS:

- Razones de ausencia
- Meses
- Días
- Estaciones
- Horas de ausentismo

Cantidad de Cluster: 4 (según las distintas estaciones del año)

Interpretación: Se puede observar que mayor faltas tiene la persona con la razón de ausentismo de Lesiones, intoxicaciones y otras consecuencias de causas externas, cuya mayor cantidad de faltas se

registraron en el mes mayo, los días miércoles en la estación de invierno. Este conjunto está comprendido en el cluster 1.

## Entrenamiento:

```
Number of iterations: 7
Within cluster sum of squared errors: 1169.9647399900564

Initial starting points (random):

Cluster 0: 11,6,6,3,2
Cluster 1: 19,5,6,3,64
Cluster 2: 19,12,2,2,8
Cluster 3: 23,8,5,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data      Cluster#
                   (740.0)      0          1          2          3
                   (132.0)      (153.0)      (205.0)      (250.0)
=====
Reason_for_absence      23          28          19          27          23
Month_of_absence        6.3243      7.1364      5.8693      3.2829      8.668
Day_of_the_week         2           6           4           2           3
Seasons                 4           3           3           2           1
Absenteeism_time_in_hours 6.9243      3.8939      10.4706      6.9951      6.296

Time taken to build model (full training data) : 0.05 seconds
```

## Testeo:

```
Number of iterations: 4
Within cluster sum of squared errors: 924.1976015974891

Initial starting points (random):

Cluster 0: 26,8,3,1,8
Cluster 1: 23,2,4,2,1
Cluster 2: 25,3,5,2,2
Cluster 3: 28,4,4,3,8

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data      Cluster#
                   (555.0)      0          1          2          3
                   (205.0)      (113.0)      (75.0)      (162.0)
=====
Reason_for_absence      23          23          23          27          28
Month_of_absence        6.2955      8.9122      2.9292      4.2133      6.2963
Day_of_the_week         2           3           4           5           4
Seasons                 4           1           2           2           3
Absenteeism_time_in_hours 6.8108      7.7512      6.646       3.9733      7.0494

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0      69 ( 37%)
1      32 ( 17%)
2      27 ( 15%)
3      57 ( 31%)
```

## SEGUNDO CRITERIO DE PRUEBA - ATRIBUTOS SELECCIONADOS

- Razones de ausencia
- Costo de transporte
- Distancias al trabajo
- Horas de ausentismo

Cantidad de Cluster: 3 (distintas distancias al trabajo)

Interpretación: Hemos podido observar que la persona que más lejos se encuentra del trabajo es el que menos horas de ausentismo tiene y uno de los que más gasta en el transporte (Cluster 0). Mientras que aquellas personas con mayor gasto de transporte y una distancia moderada al trabajo (24 km) son quienes más se ausentan en el trabajo (Cluster 1).

## Entrenamiento

```
Number of iterations: 10
Within cluster sum of squared errors: 530.3706998725368

Initial starting points (random):

Cluster 0: 11,179,51,2
Cluster 1: 19,300,26,64
Cluster 2: 19,289,36,8

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data      Cluster#
                   (740.0)      0          1          2
                   (258.0)    (125.0)    (357.0)
=====
Reason_for_absence      23          28          19          23
Transportation_expense  221.3297    231.5078    273.432    195.7311
Distance_from_Residence_to_Work  29.6311    42.062     24.328     22.5042
Absenteeism_time_in_hours  6.9243     4.686      15.376     5.5826

Time taken to build model (full training data) : 0.03 seconds
```

## Testeo

```
Number of iterations: 5
Within cluster sum of squared errors: 402.3500397278829

Initial starting points (random):

Cluster 0: 26,289,36,8
Cluster 1: 23,225,26,1
Cluster 2: 25,225,26,2

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data      Cluster#
                   (555.0)      0          1          2
                   (172.0)    (252.0)    (131.0)
=====
Reason_for_absence      23          28          23          28
Transportation_expense  219.582     245.9128    227.8056    169.1908
Distance_from_Residence_to_Work  29.3171     46.436     25.3095     14.5496
Absenteeism_time_in_hours  6.8108     6.2733     5.4048     10.2214

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0      59 ( 32%)
1      82 ( 44%)
2      44 ( 24%)
```

### TERCER CRITERIO DE PRUEBA - ATRIBUTOS SELECCIONADOS

- Razones de ausencia
- Tiempo de servicio
- Promedio de carga de trabajo
- Alcanzar la meta
- Fallas disciplinarias
- Horas de ausentismo

Cantidad de Cluster: 3 (calidad de trabajo)

Podemos observar que las personas con la razón de ausencia de fisioterapia son las que cumplen mayormente las metas, tienen el menor promedio de carga de trabajo y no son las más registran faltas. Mientras que aquellas con mayor carga de trabajo y mayor tiempo de servicio son las que más se ausentan. Las que menos faltas tienen son las que registran menos carga de trabajo, un tiempo de servicio menor y una tasa de alcance de objetivos menor.

### Entrenamiento

```
Number of iterations: 16
Within cluster sum of squared errors: 602.3968936674701
```

```
Initial starting points (random):
```

```
Cluster 0: 11,18,377.55,94,0,2
Cluster 1: 19,13,237.656,99,0,64
Cluster 2: 19,13,236.629,93,0,8
```

```
Missing values globally replaced with mean/mode
```

```
Final cluster centroids:
```

Attribute	Full Data (740.0)	Cluster#		
		0 (136.0)	1 (268.0)	2 (336.0)
Reason_for_absence	23	11	27	23
Service_time	12.5541	12.9559	12.7463	12.2381
Work_load_Average/day_	271.4902	326.5742	257.2229	260.5742
Hit_target	94.5878	95.1618	97.2575	92.2262
Disciplinary_failure	0	0	0	0
Absenteeism_time_in_hours	6.9243	9.6912	7.3321	5.4792

```
Time taken to build model (full training data) : 0.06 seconds
```

## Testeo

Number of iterations: 9  
Within cluster sum of squared errors: 389.6963248838995

Initial starting points (random):

Cluster 0: 26,13,249.797,93,0,8  
Cluster 1: 23,9,302.585,99,0,1  
Cluster 2: 25,9,222.196,99,0,2

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#			
	Full Data	0	1	2
	(555.0)	(187.0)	(226.0)	(142.0)
=====				
Reason_for_absence	23	28	23	27
Service_time	12.5027	12.1337	12.5442	12.9225
Work_load_Average/day_	272.2655	262.4582	292.3461	253.2215
Hit_target	94.5261	91.8984	94.8673	97.4437
Disciplinary_failure	0	0	0	0
Absenteeism_time_in_hours	6.8108	6.9519	6.5929	6.9718

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0      69 ( 37%)  
1      54 ( 29%)  
2      62 ( 34%)

## CUARTO CRITERIO DE PRUEBA - ATRIBUTOS SELECCIONADOS:

- Razones de ausentismo
- Edad
- Educación
- Hijos
- Mascotas
- Horas de ausentismo

Cantidad de Cluster: 3 (rendimiento etario)

Se puede observar que mayor faltas tiene la persona con la razón de ausentismo de Lesiones, intoxicaciones y otras consecuencias de causas externas cuya edad promedio es de 42 años (la mayor edad) la cual posee solamente el secundario completo, tiene un hijo y a lo sumo una mascota. Y quienes tienen menos faltas, no poseen ni hijos ni mascotas y una edad promedio de 36 años.

## Entrenamiento

Number of iterations: 6  
Within cluster sum of squared errors: 655.0934513360407

Initial starting points (random):

Cluster 0: 11,38,1,0,0,2  
Cluster 1: 19,43,1,2,1,64  
Cluster 2: 19,33,1,2,1,8

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#			
	Full Data	0	1	2
	(740.0)	(327.0)	(170.0)	(243.0)
=====				
Reason_for_absence	23	23	19	23
Age	36.45	36.2813	42.5471	32.4115
Education	1.2919	1.5352	1.2118	1.0206
Son	1.0189	0.1376	1.3765	1.9547
Pet	0.7459	0.104	0.8706	1.5226
Absenteeism_time_in_hours	6.9243	4.7309	11.7118	6.5267

Time taken to build model (full training data) : 0.02 seconds

## Testeo

Number of iterations: 5  
Within cluster sum of squared errors: 394.36858023615207

Initial starting points (random):

Cluster 0: 26,33,1,2,1,8  
Cluster 1: 23,28,1,1,2,1  
Cluster 2: 25,28,1,1,2,2

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#			
	Full Data	0	1	2
	(555.0)	(212.0)	(176.0)	(167.0)
=====				
Reason_for_absence	23	28	23	27
Age	36.4523	38.3538	34.8295	35.7485
Education	1.3081	1.0943	1.2159	1.6766
Son	1.0054	1.717	0.983	0.1257
Pet	0.7171	0.467	1.5739	0.1317
Absenteeism_time_in_hours	6.8108	8.9906	4.5057	6.4731

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0      80 ( 43%)  
1      52 ( 28%)  
2      53 ( 29%)



## QUINTO CRITERIO DE PRUEBA - ATRIBUTOS SELECCIONADOS:

- Razones de ausentismo
- Bebedores
- Fumadores
- Edad
- Índice de masa corporal
- Horas de ausentismo

Cluster: 2 (Factor de riesgo)

Se puede observar que mayor faltas tiene es la persona que posee mayor factor de riesgo, ya que fuma y tiene sobrepeso, con la razón de ausentismo por consulta dental y son mayores en edad. Por el contrario, tienen menos faltas las personas que llevan una vida más saludable.

## Entrenamiento

Number of iterations: 7  
Within cluster sum of squared errors: 696.6096086722433

Initial starting points (random):

Cluster 0: 11,38,1,0,31,2  
Cluster 1: 19,43,1,1,25,64

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data	0	1
	(740.0)	(405.0)	(335.0)
=====			
Reason_for_absence	23	28	23
Age	36.45	37.8889	34.7104
Social_drinker	1	1	0
Social_smoker	0	0	0
Body_mass_index	26.677	28.1531	24.8925
Absenteeism time in hours	6.9243	7.8815	5.7672

Time taken to build model (full training data) : 0 seconds

## Testeo

Number of iterations: 3

Within cluster sum of squared errors: 519.1019372793232

Initial starting points (random):

Cluster 0: 26,33,1,0,30,8

Cluster 1: 23,28,0,0,24,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#		
	Full Data	0	1
	(555.0)	(299.0)	(256.0)
=====			
Reason_for_absence	23	28	23
Age	36.4523	37.8763	34.7891
Social_drinker	1	1	0
Social_smoker	0	0	0
Body_mass_index	26.6595	28.1639	24.9023
Absenteeism_time_in_hours	6.8108	7.3043	6.2344

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0      106 ( 57%)

1      79 ( 43%)



## CLASIFICACIÓN

Las técnicas de clasificación automática buscan encontrar un modelo capaz de identificar automáticamente la clase a la cual pertenece un objeto dado. Las clases se obtuvieron mediante la técnica de clustering.

### ZeroR

ZeroR es el método de clasificación más simple que existe y depende solo en el target ignorando todos los predictores. El clasificador ZeroR simplemente predice sobre la clase o categoría principal (majority category).

#### Configuración inicial:

Choose ZeroR

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 75

More options...

#### Resultados:

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

ZeroR predicts class value: si

Time taken to build model: 0 seconds

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

Correctly Classified Instances      106          57.2973 %
Incorrectly Classified Instances    79          42.7027 %
Kappa statistic                    0
Mean absolute error                 0.4895
Root mean squared error            0.4947
Relative absolute error             100 %
Root relative squared error        100 %
Total Number of Instances         185

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	1,000	0,573	1,000	0,729	?	0,483	0,564	si
	0,000	0,000	?	0,000	?	?	0,483	0,418	no
Weighted Avg.	0,573	0,573	?	0,573	?	?	0,483	0,502	

```

=== Confusion Matrix ===
  a  b  <-- classified as
106  0 |  a = si
 79  0 |  b = no

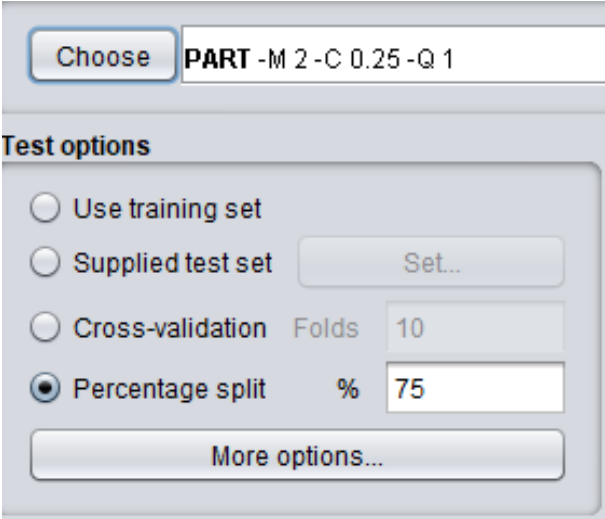
```

Este algoritmo clasificó correctamente el 57.29% de las instancias. Un 100% de las instancias de la clase "SI" (con factores de riesgo) fueron clasificadas correctamente (106). Mientras que el 100% de las instancias de la clase "NO" fueron clasificadas incorrectamente (79).

## PART

Obtiene reglas a partir de árboles de decisión construidos usando J48.

### Configuración inicial



The screenshot shows the Weka GUI's 'Test options' panel for the PART algorithm. At the top, there is a 'Choose' button and a text field containing the command 'PART -M 2 -C 0.25 -Q 1'. Below this, the 'Test options' section contains four radio buttons: 'Use training set', 'Supplied test set', 'Cross-validation', and 'Percentage split'. The 'Percentage split' option is selected. To the right of the radio buttons, there are input fields: 'Folds' is set to 10, and 'Percentage split' is set to 75. A 'Set...' button is located next to the 'Supplied test set' option. At the bottom of the panel is a 'More options...' button.

Choose **PART -M 2 -C 0.25 -Q 1**

**Test options**

☐ Use training set

☐ Supplied test set

☐ Cross-validation Folds

☒ Percentage split %

## Resultados:

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

PART decision list  
-----

Social\_drinker = 0 AND  
Reason\_for\_absence = 23: no (15.0)

Social\_drinker = 1 AND  
Reason\_for\_absence = 28: si (23.0)

Social\_drinker = 0 AND  
Reason\_for\_absence = 27: no (10.0)

Social\_drinker = 0 AND  
Reason\_for\_absence = 25: no (7.0)

Social\_drinker = 1 AND  
Reason\_for\_absence = 27: si (10.0)

Social\_drinker = 1 AND  
Reason\_for\_absence = 22: si (10.0)

Social\_drinker = 0 AND  
Reason\_for\_absence = 18: no (5.0)

Social\_drinker = 0 AND  
Reason\_for\_absence = 13: no (5.0)

Social\_drinker = 1 AND  
Reason\_for\_absence = 23 AND  
Body\_mass\_index > 28: si (9.0)

Social\_drinker = 0 AND  
Reason\_for\_absence = 1: no (4.0)

Social\_drinker = 1: si (56.0/5.0)

Reason\_for\_absence = 10: no (4.0)

```

Reason_for_absence = 22: no (4.0)

Reason_for_absence = 0: no (4.0)

Reason_for_absence = 26: no (3.0)

Reason_for_absence = 19: no (3.0)

Reason_for_absence = 28 AND
Age > 32: si (3.0)

: no (10.0)

Number of Rules :      18

Time taken to build model: 0.04 seconds

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

Correctly Classified Instances      175           94.5946 %
Incorrectly Classified Instances     10           5.4054 %
Kappa statistic                     0.8885
Mean absolute error                  0.0786
Root mean squared error              0.2167
Relative absolute error              16.0495 %
Root relative squared error          43.8125 %
Total Number of Instances           185

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                -----  -----  -
                0,981    0,101    0,929      0,981    0,954      0,890    0,952    0,943     si
                0,899    0,019    0,973      0,899    0,934      0,890    0,952    0,945     no
Weighted Avg.   0,946    0,066    0,947      0,946    0,946      0,890    0,952    0,944

=== Confusion Matrix ===

  a  b  <-- classified as
104  2 |  a = si
  8 71 |  b = no

```

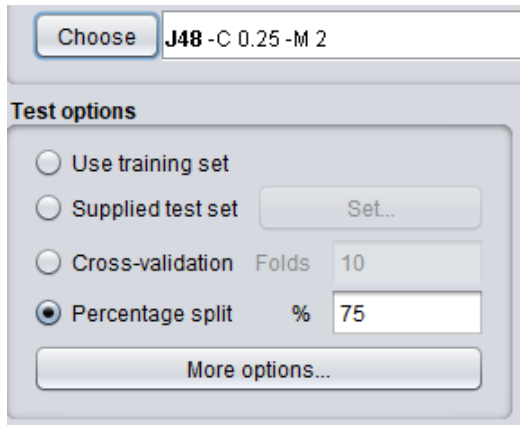
El número de reglas es 18.

Este algoritmo clasificó correctamente el 94.59% de las instancias, mejorando considerablemente el porcentaje de acierto. Un 98% de las instancias de la clase “SI” (con factores de riesgo) fueron clasificadas correctamente (104) y el resto (2) fueron clasificadas incorrectamente. Mientras que el 89% de las instancias de la clase “NO” fueron clasificadas correctamente (71) y el resto (8) clasificadas incorrectamente.

## J48

Aprende árbol de decisión.

### Configuración inicial



### Resultados:

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

```
J48 pruned tree
-----
```

```
Social_drinker = 1: si (108.0/5.0)
Social_drinker = 0: no (77.0/3.0)
```

```
Number of Leaves :    2
```

```
Size of the tree :    3
```

```
Time taken to build model: 0.01 seconds
```

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

Correctly Classified Instances	177	95.6757 %
Incorrectly Classified Instances	8	4.3243 %
Kappa statistic	0.9113	
Mean absolute error	0.0831	
Root mean squared error	0.2053	
Relative absolute error	16.976 %	
Root relative squared error	41.5086 %	
Total Number of Instances	185	

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

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,972	0,063	0,954	0,972	0,963	0,912	0,924	0,904	si
	0,937	0,028	0,961	0,937	0,949	0,912	0,924	0,895	no
Weighted Avg.	0,957	0,048	0,957	0,957	0,957	0,912	0,924	0,900	

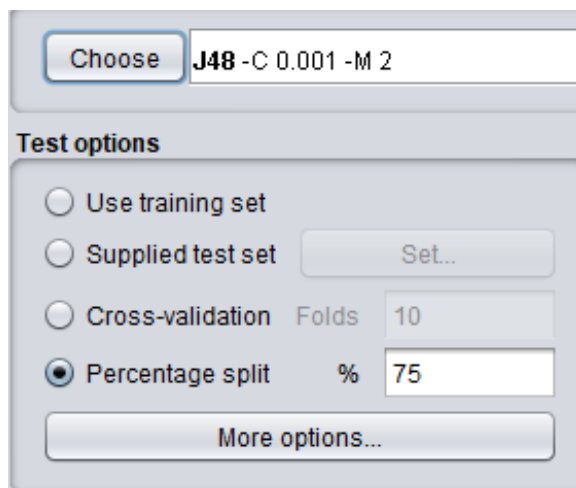
```
=== Confusion Matrix ===
```

```
  a   b  <-- classified as
103   3 |   a = si
  5  74 |   b = no
```

El árbol generado por este algoritmo tiene 2 niveles y 3 nodos con el valor de confianza en 0,25.

El 95.67% de las instancias fueron clasificadas correctamente. Un 97% de las instancias de la clase “SI” fueron clasificadas correctamente (103) y el resto (3) fueron clasificadas incorrectamente. Mientras que el 93.6% de las instancias de la clase “NO” fueron clasificadas correctamente (74) y el resto (5) fueron clasificadas incorrectamente.

A fin de generar un clasificador más simple, bajamos el valor de confianza inicial del algoritmo. Recordamos que en la primera ejecución el valor de confianza estaba en 0,25 y ahora ejecutaremos con 0,001.



## Resultados:

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

J48 pruned tree  
-----

Social\_drinker = 1: si (108.0/5.0)  
Social\_drinker = 0: no (77.0/3.0)

Number of Leaves : 2

Size of the tree : 3

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	177	95.6757 %
Incorrectly Classified Instances	8	4.3243 %
Kappa statistic	0.9113	
Mean absolute error	0.0831	
Root mean squared error	0.2053	
Relative absolute error	16.976 %	
Root relative squared error	41.5086 %	
Total Number of Instances	185	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,972	0,063	0,954	0,972	0,963	0,912	0,924	0,904	si
	0,937	0,028	0,961	0,937	0,949	0,912	0,924	0,895	no
Weighted Avg.	0,957	0,048	0,957	0,957	0,957	0,912	0,924	0,900	

=== Confusion Matrix ===

```

a  b  <-- classified as
103  3 |  a = si
  5 74 |  b = no

```



El árbol generado es igual al generado en la ejecución anterior.

## Cost-sensitive-classifier (meta clasificador)

Es un metaclassificador porque utiliza otro clasificador de base (ZeroR). Intentaremos mejorar la clasificación de las instancias. Para esto modificaremos los valores iniciales de la matriz de costos, y pondremos el valor “2” en el casillero correspondiente al valor que queremos mejorar, en este caso sería las instancias pertenecientes a “NO” que se clasificaron incorrectamente.

### Configuración inicial:

Choose	<b>CostSensitiveClassifier</b> -cost-matrix"[0.0 1.0; 2.0 0.0]" -S 1 -W weka.classifiers.rules.ZeroR
--------	--

### Resultados:

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

CostSensitiveClassifier using reweighted training instances

weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1

Classifier Model
PART decision list
-----

Social_drinker = 0 AND
Reason_for_absence = 23: no (21.02)

Social_drinker = 0 AND
Reason_for_absence = 27: no (14.02)

Social_drinker = 0 AND
Reason_for_absence = 25: no (9.81)

Social_drinker = 0 AND
Reason_for_absence = 18: no (7.01)

Social_drinker = 0 AND
Reason_for_absence = 13: no (7.01)

Social_drinker = 0 AND
Reason_for_absence = 1: no (5.61)

Social_drinker = 0 AND
Reason_for_absence = 10: no (5.61)

Social_drinker = 0 AND
Reason_for_absence = 22: no (5.61)

Social_drinker = 0 AND
Reason_for_absence = 0: no (5.61)

Social_drinker = 0 AND
Reason_for_absence = 26: no (4.2)
```

```

Social_drinker = 1 AND
Reason_for_absence = 28: si (16.12)

Social_drinker = 0 AND
Reason_for_absence = 19: no (4.2)

Social_drinker = 1 AND
Reason_for_absence = 27: si (7.01)

Social_drinker = 1 AND
Absenteeism_time_in_hours > 4: si (32.23)

Social_drinker = 0 AND
Reason_for_absence = 7: no (2.8)

Social_drinker = 0 AND
Reason_for_absence = 28 AND
Instance_number > 131: no (2.8)

Reason_for_absence = 23 AND
Body_mass_index <= 28: no (7.01)

Social_drinker = 1: si (16.82)

: no (10.51/2.1)

Number of Rules :      19

Cost Matrix
0 1
2 0

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      178           96.2162 %
Incorrectly Classified Instances      7           3.7838 %
Kappa statistic                    0.9226
Mean absolute error                  0.05
Root mean squared error              0.188
Relative absolute error              10.2165 %
Root relative squared error          37.994 %
Total Number of Instances           185

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,972    0,051    0,963     0,972    0,967      0,923    0,980    0,980     si
                0,949    0,028    0,962     0,949    0,955      0,923    0,980    0,968     no
Weighted Avg.   0,962    0,041    0,962     0,962    0,962      0,923    0,980    0,975

=== Confusion Matrix ===

  a  b  <-- classified as
103  3 |  a = si
  4 75 |  b = no

```

Como podemos observar se pudo mejorar levemente el porcentaje de instancias clasificadas correctamente en la clase “NO” que paso de un 93.6% a un 94.9% (75), mientras en la clase “SI” no se presenta variación en los resultados iniciales.

# Reglas de asociación

## A priori (Factores de riesgo)

### Resultados:

```
Apriori
=====

Minimum support: 0.55 (102 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 9

Generated sets of large itemsets:


Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 17

Size of set of large itemsets L(4): 6

Best rules found:
```



```
1. Social_drinker=1 108 ==> Factor_de_Riesgo=si 103   conf:(0.95) < lift:(1.66)> lev:(0.22) [41] conv:(7.69)
2. Factor_de_Riesgo=si 106 ==> Social_drinker=1 103   conf:(0.97) < lift:(1.66)> lev:(0.22) [41] conv:(11.03)
3. Social_drinker=1 108 ==> Instance_number='All' Factor_de_Riesgo=si 103   conf:(0.95) < lift:(1.66)> lev:(0.22) [41] conv:(7.69)
4. Instance_number='All' Social_drinker=1 108 ==> Factor_de_Riesgo=si 103   conf:(0.95) < lift:(1.66)> lev:(0.22) [41] conv:(7.69)
5. Factor_de_Riesgo=si 106 ==> Instance_number='All' Social_drinker=1 103   conf:(0.97) < lift:(1.66)> lev:(0.22) [41] conv:(11.03)
6. Instance_number='All' Factor_de_Riesgo=si 106 ==> Social_drinker=1 103   conf:(0.97) < lift:(1.66)> lev:(0.22) [41] conv:(11.03)
7. Social_drinker=1 108 ==> Absenteeism_time_in_hours='All' Factor_de_Riesgo=si 103   conf:(0.95) < lift:(1.66)> lev:(0.22) [41] conv:(7.69)
8. Social_drinker=1 Absenteeism_time_in_hours='All' 108 ==> Factor_de_Riesgo=si 103   conf:(0.95) < lift:(1.66)> lev:(0.22) [41] conv:(7.69)
9. Factor_de_Riesgo=si 106 ==> Social_drinker=1 Absenteeism_time_in_hours='All' 103   conf:(0.97) < lift:(1.66)> lev:(0.22) [41] conv:(11.03)
10. Absenteeism_time_in_hours='All' Factor_de_Riesgo=si 106 ==> Social_drinker=1 103   conf:(0.97) < lift:(1.66)> lev:(0.22) [41] conv:(11.03)
```

### Reglas de asociación identificadas:

Los bebedores sociales (alcohólicos) poseen factores de riesgo y son quienes más probabilidades tienen de ausentarse en sus trabajos.

## A priori (Calidad de trabajo)

```
Apriori
=====

Minimum support: 0.2 (37 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 36

Size of set of large itemsets L(3): 63

Size of set of large itemsets L(4): 58

Size of set of large itemsets L(5): 27

Size of set of large itemsets L(6): 5

Best rules found:
```

```
1. Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 91 ==> Cluster=cluster2 38   conf:(0.42) < lift:(1.25)> lev:(0.04) [7] conv:(1.12)
2. Cluster=cluster2 62 ==> Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 38   conf:(0.61) < lift:(1.25)> lev:(0.04) [7] conv:(1.26)
3. Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 91 ==> Instance_number='All' Cluster=cluster2 38   conf:(0.42) < lift:(1.25)> lev:(0.04) [7] conv:(1.12)
4. Instance_number='All' Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 91 ==> Cluster=cluster2 38   conf:(0.42) < lift:(1.25)> lev:(0.04) [7] conv:(1.12)
5. Cluster=cluster2 62 ==> Instance_number='All' Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 38   conf:(0.61) < lift:(1.25)> lev:(0.04) [7] conv:(1.26)
6. Instance_number='All' Cluster=cluster2 62 ==> Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 38   conf:(0.61) < lift:(1.25)> lev:(0.04) [7] conv:(1.26)
7. Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 91 ==> Service_time='All' Cluster=cluster2 38   conf:(0.42) < lift:(1.25)> lev:(0.04) [7] conv:(1.12)
8. Service_time='All' Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 91 ==> Cluster=cluster2 38   conf:(0.42) < lift:(1.25)> lev:(0.04) [7] conv:(1.12)
9. Cluster=cluster2 62 ==> Service_time='All' Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 38   conf:(0.61) < lift:(1.25)> lev:(0.04) [7] conv:(1.26)
10. Service_time='All' Cluster=cluster2 62 ==> Work_load_Average/day_='(-inf-263.572667)' Hit_target='(90.5-inf)' 38   conf:(0.61) < lift:(1.25)> lev:(0.04) [7] conv:(1.26)
```

### Reglas de asociación identificadas:

Quienes más carga tienen de trabajo y tienen un índice de alcance de objetivos promedio, pertenecen al cluster2 que es de quienes más se ausentan.

### CONCLUSIÓN:

El proceso de descubrimiento de conocimiento permite obtener información relevante y procesar una gran cantidad de datos utilizando métodos de minería como el agrupamiento o clustering y la clasificación.

Como los datos estaban en crudo, sin ningún tipo de trabajo previo más que un archivo con muchos registros, el primer método que se aplicó fue el de clustering para obtener los grupos de datos, particiones del gran conjunto. Esto permitiría obtener las clases para después usar el método de clasificación.

Pero antes de aplicar todas estas técnicas, estuvimos analizando los datos y los atributos y apartando aquellos no relevantes, como por ejemplo el ID de personal. Agrupamos los atributos que guardaban una relación significativa para cada prueba como: calidad de trabajo, momentos del año, calidad de vida o factor de riesgo, y familia o mascotas a cargo.

Cuando se decidieron estos casos de prueba, procedimos a ejecutar los algoritmos de clustering y obtuvimos los grupos para cada prueba.

Con estos grupos o clases, luego empleamos el método de clasificación para probar lo obtenido con otros datos.

Posteriormente analizamos las reglas de asociación, basándonos en las medidas de confianza y Lift. Es así como obtuvimos asociaciones entre los atributos y cómo se relacionaban con el ausentismo y las clases.

