
# Trabajo Final de Minería de Datos

**Grado en Ingeniería Informática**

**Curso 2023/24**

**Manuel Ramírez Ballesteros**

# ÍNDICE

# 1. Abstract.

En este proyecto vamos a desarrollar un proyecto KDD de extracción de conocimiento completo, concretamente de clasificación para el *dataset cmc.arff.* En él detallaremos cada una de las etapas del proceso, desde la descripción del problema y la preparación de los datos hasta finalizar con un análisis de los resultados sobre los conjuntos de entrenamiento y de test, con el fin de ver que algoritmos nos ofrecen un mejor rendimiento.

# 2. Descripción del problema y del dataset.

El conjunto de datos seleccionado es un subconjunto de los datos registrados en la Encuesta Nacional de Prevalencia de Anticonceptivos de Indonesia de 1987. En él encontramos 1473 instancias que representan a mujeres casadas que, o bien que no estaban embarazadas, o que no lo sabían en el momento de la entrevista. Este *dataset* lo dividiremos en dos: Uno con 1179 instancias para entrenamiento y otro con 294 para test. El problema que se aborda es predecir el método anticonceptivo elegido actualmente de una mujer en función de sus características demográficas y socioeconómicas.

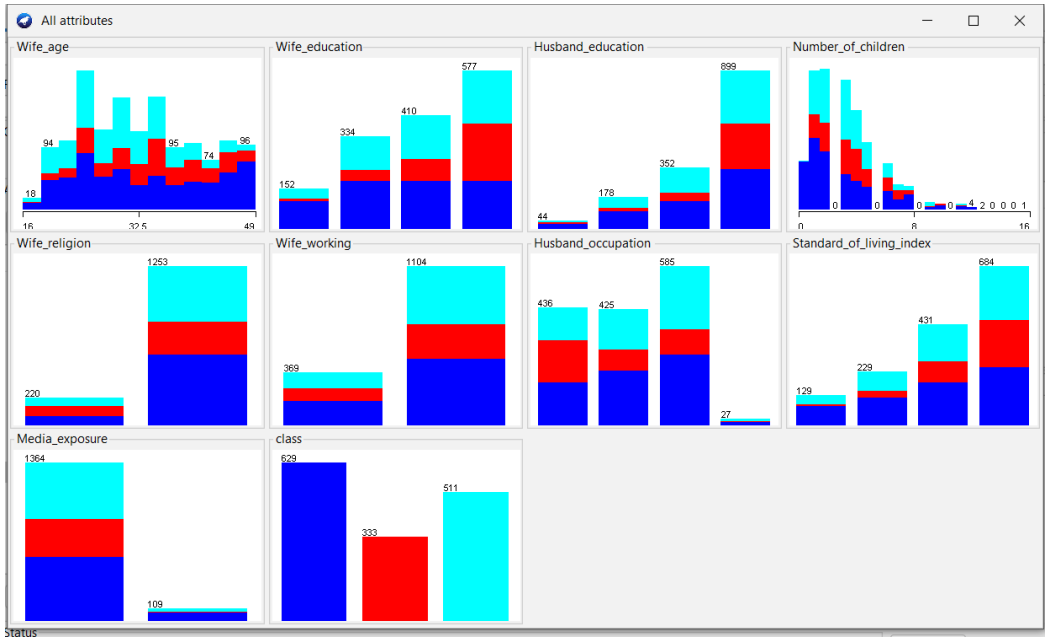
Cada una de las instancias se representará con 10 atributos, incluido el de clase y sin ningún valor ausente, que son:

- **Wife_age**: Edad de la mujer (entero).
- **Wife_education**: Nivel educativo de la mujer (1, 2, 3, 4).
- **Husband_education**: Nivel educativo del esposo (1, 2, 3, 4).
- **Number_of_children**: Número de hijos (entero).
- **Wife_religion**: Religión de la mujer (0, 1).
- **Wife_working**: Trabajo de la mujer (0, 1).
- **Husband_occupation**: Ocupación del marido (1, 2, 3, 4).
- **Standard_of_living_index**: Nivel de vida medio (1, 2, 3, 4).
- **Media_exposure**: Exposición media (0, 1).
- **class**: Atributo de clase (1, 2, 3).

Tampoco se observa que exista demasiado desbalance entre las distintas clases.

# 3. Preparación de los datos.

A continuación, analizaremos más en detalle el conjunto de datos para prepararlos. Al no tener valores ausentes, no será necesario aplicar técnicas como el descarte o la imputación. En cuanto a los *outliers*, podemos visualizar todos los atributos:




La mayoría de los datos a primera vista parecen correctos, donde lo más destacable puede ser la instancia 974, cuyo número de hijos es 16:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 973 | 30.0 | 2 | 2 | 4.0 | 1 | 1 | 2 | 4 | 0 | 3 |
| 974 | 48.0 | 4 | 4 | 16.0 | 1 | 1 | 1 | 4 | 0 | 3 |
| 975 | 29.0 | 2 | 4 | 3.0 | 1 | 0 | 3 | 3 | 0 | 3 |

De momento, no eliminaremos esta instancia por si fuese relevante para la clasificación.

# 4. Descripción de los atributos seleccionados.

Para la selección de atributos, se han empleado los métodos de evaluación *CorrelationAttributeEval* e *InfoGainAttributeEval* y, como método de búsqueda, *Ranker*, obteniendo los siguientes resultados sobre todo el conjunto de datos:

Correlación:

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 class):
        Correlation Ranking Filter
Ranked attributes:
 0.1377    1 Wife_age
 0.1062    9 Media_exposure
 0.0975    2 Wife_education
 0.0889    3 Husband_education
 0.0855    4 Number_of_children
 0.0734    7 Husband_occupation
 0.0683    5 Wife_religion
 0.0639    8 Standard_of_living_index
 0.0421    6 Wife_working

Selected attributes: 1,9,2,3,4,7,5,8,6 : 9
```

Ganancia:

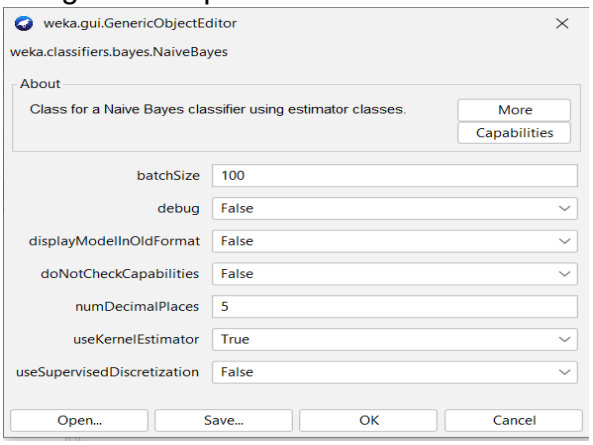
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.10194    4 Number_of_children
 0.07091    2 Wife_education
 0.0615     1 Wife_age
 0.04014    3 Husband_education
 0.03251    8 Standard_of_living_index
 0.03047    7 Husband_occupation
 0.01579    9 Media_exposure
 0.00982    5 Wife_religion
 0.00258    6 Wife_working

Selected attributes: 4,2,1,3,8,7,9,5,6 : 9
```

En ambos casos, el atributo que menos aporta es *Wife_working,* pero al estudiar los resultados obtenidos en los distintos algoritmos con y sin este atributo, vemos que no se obtienen resultados mejores, sino que en algunos casos incluso se empeoran. Por lo tanto, los modelos que presentaremos a continuación se obtendrán con el dataset original, sin llevar a cabo ninguna modificación.

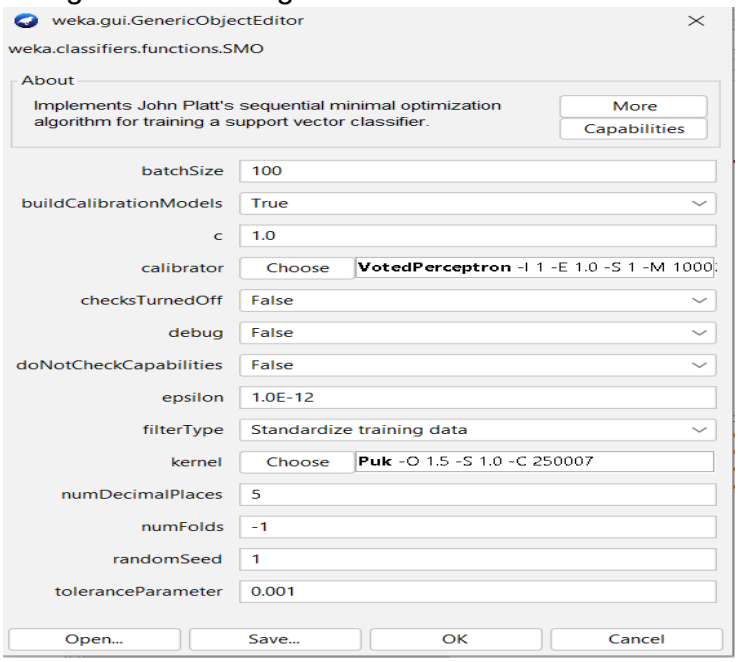
# 5. Descripción de los algoritmos de clasificación escogidos.

En cuanto a los algoritmos de clasificación escogidos en este proyecto, se ha tratado de emplear cierta variedad:

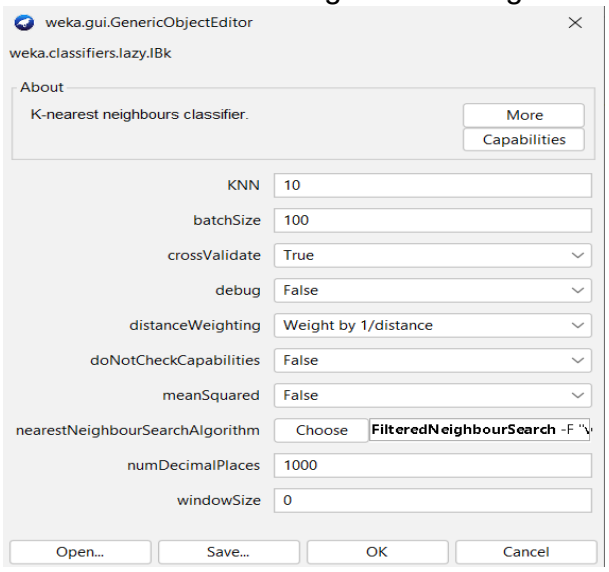
- ***Naive Bayes***: La configuración que se ha establecido es:



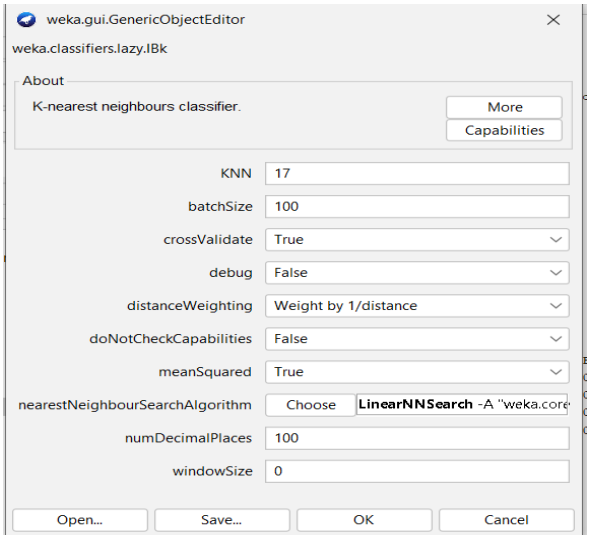

- ***SMO***: Se ha configurado de la siguiente forma:



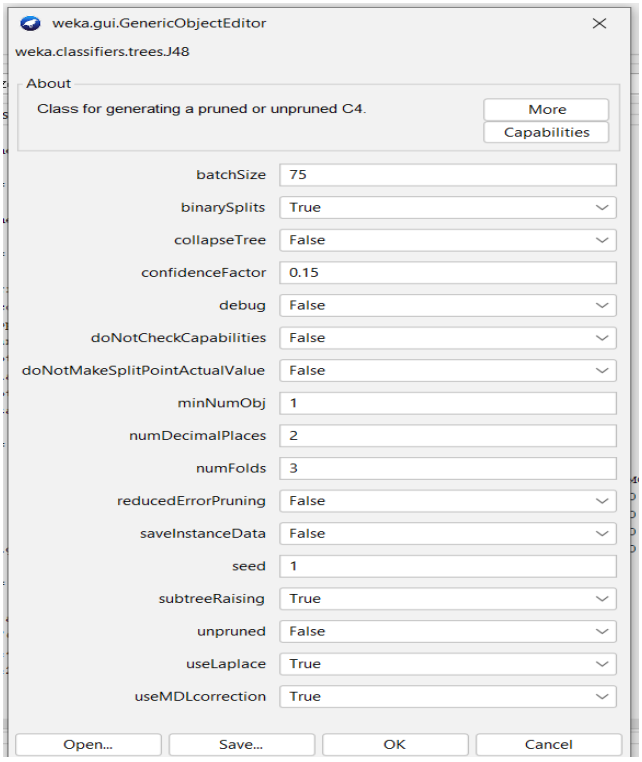

- ***IBk (con k = 10)***: Se ha establecido la siguiente configuración:




- ***IBk (con k = 17)***: Configurado de la siguiente manera:




- ***J48***: Se ha establecido la siguiente configuración:




# 6. Evaluación de los resultados obtenidos con el conjunto de entrenamiento.

A continuación, se mostrarán los resultados obtenidos al ejecutar cada uno de los algoritmos anteriores y así comparar los distintos resultados que nos ofrecen dichos modelos sobre el conjunto de entrenamiento *cmc.arff_tra.*

- ***Naive Bayes:*** Obtiene un 53.3325% de acierto al clasificar:

```
Correctly Classified Instances         617               52.3325 %
Incorrectly Classified Instances       562               47.6675 %
Kappa statistic                          0.2768
Mean absolute error                      0.3593
Root mean squared error                  0.4412
Relative absolute error                 83.4114 %
Root relative squared error             95.065  %
Total Number of Instances             1179

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,553    0,220    0,651      0,553    0,598      0,342    0,733     0,696
                 0,504    0,228    0,392      0,504    0,441      0,254    0,731     0,441
                 0,500    0,267    0,500      0,500    0,500      0,233    0,686     0,540
Weighted Avg.    0,523    0,238    0,540      0,523    0,528      0,284    0,716     0,584

=== Confusion Matrix ===

   a   b   c   <-- classified as
 278 101 124 |   a = 1
  51 134  81 |   b = 2
  98 107 205 |   c = 3
```

- ***SMO:*** Obtiene un 86.514% de acierto al clasificar:

```
Correctly Classified Instances        1020               86.514  %
Incorrectly Classified Instances       159               13.486  %
Kappa statistic                          0.7899
Mean absolute error                      0.0908
Root mean squared error                  0.2993
Relative absolute error                 21.0871 %
Root relative squared error             64.496  %
Total Number of Instances             1179

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,920    0,099    0,874      0,920    0,896      0,817    0,923     0,874
                 0,763    0,044    0,835      0,763    0,798      0,743    0,925     0,802
                 0,863    0,068    0,872      0,863    0,868      0,798    0,927     0,844
Weighted Avg.    0,865    0,076    0,864      0,865    0,864      0,793    0,925     0,847

=== Confusion Matrix ===

   a   b   c   <-- classified as
 463  18  22 |   a = 1
  33 203  30 |   b = 2
  34  22 354 |   c = 3
```

- ***IBk (con k = 10):*** Obtiene un 95.4198% de acierto al clasificar.

```
=== Summary ===

Correctly Classified Instances        1125               95.4198 %
Incorrectly Classified Instances        54                4.5802 %
Kappa statistic                          0.9291
Mean absolute error                      0.0313
Root mean squared error                  0.1251
Relative absolute error                  7.263  %
Root relative squared error             26.9507 %
Total Number of Instances             1179

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,996    0,041    0,947      0,996    0,971      0,949    0,999     0,998
                 0,959    0,026    0,914      0,959    0,936      0,917    0,997     0,986
                 0,900    0,003    0,995      0,900    0,945      0,920    0,997     0,991
Weighted Avg.    0,954    0,025    0,956      0,954    0,954      0,932    0,998     0,993

=== Confusion Matrix ===

   a   b   c   <-- classified as
 501   1   1 |   a = 1
  10 255   1 |   b = 2
  18  23 369 |   c = 3
```

- ***IBk (con k = 17):*** Obtiene un 95.4198% de acierto al clasificar. Aunque falla más al clasificar la clase *a* que el IBk anterior, mejora la clasificación en el resto de las clases.

```
=== Summary ===

Correctly Classified Instances        1125               95.4198 %
Incorrectly Classified Instances        54                4.5802 %
Kappa statistic                          0.9292
Mean absolute error                      0.0883
Root mean squared error                  0.1451
Relative absolute error                 20.4877 %
Root relative squared error             31.2711 %
Total Number of Instances             1179

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Ar
                 0,960    0,010    0,986      0,960    0,973      0,953    0,999     0,998
                 0,925    0,018    0,939      0,925    0,932      0,912    0,997     0,991
                 0,966    0,040    0,927      0,966    0,946      0,917    0,997     0,994
Weighted Avg.    0,954    0,022    0,955      0,954    0,954      0,931    0,998     0,995

=== Confusion Matrix ===

   a   b   c   <-- classified as
 483   5  15 |   a = 1
   4 246  16 |   b = 2
   3  11 396 |   c = 3
```

- ***J48:*** Obtiene un 65.8185% de acierto al clasificar

```
=== Summary ===

Correctly Classified Instances         776               65.8185 %
Incorrectly Classified Instances       403               34.1815 %
Kappa statistic                          0.4677
Mean absolute error                      0.3104
Root mean squared error                  0.394
Relative absolute error                 72.0598 %
Root relative squared error             84.8915 %
Total Number of Instances             1179

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Ar
                 0,708    0,186    0,739      0,708    0,723      0,525    0,825     0,793
                 0,481    0,094    0,598      0,481    0,533      0,420    0,809     0,544
                 0,712    0,248    0,605      0,712    0,654      0,449    0,798     0,626
Weighted Avg.    0,658    0,187    0,660      0,658    0,656      0,475    0,812     0,679

=== Confusion Matrix ===

   a   b   c   <-- classified as
 356  33 114 |   a = 1
  61 128  77 |   b = 2
  65  53 292 |   c = 3
```

Podemos observar los resultados obtenidos de forma conjunta en la siguiente tabla:

| | **Tasa de aciertos (%)** | **F-Measure** | **Area ROC** |
|---|---|---|---|
| **Naive Bayes** | 53.3325 | 0.528 | 0.716 |
| **SMO** | 86.514 | 0.864 | 0.847 |
| **IBk (k=10)** | 95.4198 | 0.954 | 0.993 |
| **IBk (k=17)** | 95.4198 | 0.954 | 0.995 |
| **J48** | 65.8185 | 0.656 | 0.679 |

Con estos resultados parece que los modelos generados mediante *IBk* nos proporcionan un mejor rendimiento y equilibrio entre precisión y exhaustividad que el resto de los modelos sobre el conjunto de entrenamiento. El modelo generado mediante *Naive Bayes* es el que peores resultados nos ofrece.

# 7. Evaluación de los resultados obtenidos con el conjunto de test.

Por último, vamos a analizar los resultados que nos ofrecen estos modelos sobre un conjunto de datos nuevos, utilizando el conjunto de test *cmc.arff_tst:*

- ***Naive Bayes:*** Obtiene un 53.7415% de acierto al clasificar:

```
=== Summary ===

Correctly Classified Instances         158               53.7415 %
Incorrectly Classified Instances       136               46.2585 %
Kappa statistic                          0.3003
Mean absolute error                      0.3536
Root mean squared error                  0.4387
Relative absolute error                 82.0527 %
Root relative squared error             94.5069 %
Total Number of Instances              294

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,571    0,220    0,661      0,571    0,613      0,360    0,753     0,665
                 0,597    0,229    0,435      0,597    0,503      0,333    0,756     0,484
                 0,455    0,244    0,495      0,455    0,474      0,216    0,675     0,503
Weighted Avg.    0,537    0,230    0,552      0,537    0,540      0,304    0,727     0,568

=== Confusion Matrix ===

  a  b  c   <-- classified as
 72 22 32 |  a = 1
 12 40 15 |  b = 2
 25 30 46 |  c = 3
```

- ***SMO:*** Obtiene un 52.7211% de acierto al clasificar:

```
=== Summary ===

Correctly Classified Instances         155               52.7211 %
Incorrectly Classified Instances       139               47.2789 %
Kappa statistic                          0.2322
Mean absolute error                      0.3155
Root mean squared error                  0.5606
Relative absolute error                 73.2186 %
Root relative squared error            120.7651 %
Total Number of Instances              294

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,810    0,554    0,523      0,810    0,636      0,268    0,649     0,526
                 0,328    0,079    0,550      0,328    0,411      0,305    0,637     0,357
                 0,307    0,145    0,525      0,307    0,388      0,192    0,584     0,402
Weighted Avg.    0,527    0,305    0,530      0,527    0,499      0,250    0,624     0,445

=== Confusion Matrix ===

   a   b   c   <-- classified as
 102   7  17 |   a = 1
  34  22  11 |   b = 2
  59  11  31 |   c = 3
```

- ***IBk (con k = 10):*** Obtiene un 44.2177% de acierto al clasificar.

```
Correctly Classified Instances         130               44.2177 %
Incorrectly Classified Instances       164               55.7823 %
Kappa statistic                          0.1491
Mean absolute error                      0.377
Root mean squared error                  0.6017
Relative absolute error                 87.4949 %
Root relative squared error            129.6323 %
Total Number of Instances              294

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,492    0,345    0,517      0,492    0,504      0,148    0,623     0,501
                 0,507    0,229    0,395      0,507    0,444      0,257    0,611     0,354
                 0,337    0,280    0,386      0,337    0,360      0,059    0,510     0,385
Weighted Avg.    0,442    0,296    0,444      0,442    0,441      0,142    0,582     0,428

=== Confusion Matrix ===

  a  b  c   <-- classified as
 62 26 38 |  a = 1
 17 34 16 |  b = 2
 41 26 34 |  c = 3
```

- ***IBk (con k = 17):*** Obtiene un 53.4014% de acierto al clasificar.

```
=== Summary ===

Correctly Classified Instances         157               53.4014 %
Incorrectly Classified Instances       137               46.5986 %
Kappa statistic                          0.2759
Mean absolute error                      0.3707
Root mean squared error                  0.4433
Relative absolute error                 86.0427 %
Root relative squared error             95.4919 %
Total Number of Instances              294

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Are
                 0,611     0,286     0,616       0,611   0,614      0,326    0,734     0,646
                 0,403     0,141     0,458       0,403   0,429      0,274    0,722     0,429
                 0,525     0,295     0,482       0,525   0,502      0,225    0,653     0,483
Weighted Avg.    0,534     0,256     0,534       0,534   0,533      0,279    0,703     0,541

=== Confusion Matrix ===

  a  b  c   <-- classified as
 77 14 35 |  a = 1
 18 27 22 |  b = 2
 30 18 53 |  c = 3
```

- ***J48:*** Obtiene un 58.8435% de acierto al clasificar

```
=== Summary ===

Correctly Classified Instances         173               58.8435 %
Incorrectly Classified Instances       121               41.1565 %
Kappa statistic                          0.3587
Mean absolute error                      0.3427
Root mean squared error                  0.4333
Relative absolute error                 79.5366 %
Root relative squared error             93.354  %
Total Number of Instances              294

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Are
                 0,643     0,250     0,659       0,643   0,651      0,394    0,751     0,718
                 0,478     0,093     0,604       0,478   0,533      0,420    0,742     0,467
                 0,594     0,301     0,508       0,594   0,548      0,284    0,669     0,480
Weighted Avg.    0,588     0,231     0,595       0,588   0,589      0,362    0,721     0,579

=== Confusion Matrix ===

  a  b  c   <-- classified as
 81  5 40 |  a = 1
 17 32 18 |  b = 2
 25 16 60 |  c = 3
```

Podemos observar los resultados obtenidos de forma conjunta en la siguiente tabla:

| | **Tasa de aciertos (%)** | **F-Measure** | **Area ROC** |
|---|---|---|---|
| **Naive Bayes** | 53.7415 | 0.540 | 0.560 |
| **SMO** | 52.7211 | 0.499 | 0.445 |
| **IBk (k=10)** | 44.2177 | 0.441 | 0.428 |
| **IBk (k=17)** | 53.4014 | 0.533 | 0.541 |
| **J48** | 58.8435 | 0.589 | 0.579 |

Esta tabla nos ofrece resultados interesantes sobre un conjunto de datos desconocidos, donde modelos como *IBk* con k=10, que antes nos daba muy buen rendimiento, ahora pasa a ser el peor de los modelos de clasificación. El mejor ahora es *J48*, que antes no nos ofrecía buen rendimiento. Esto puede indicar que el modelo generaliza mejor nuevos datos.
Pese a esto, ninguno de los modelos nos ofrece un rendimiento aceptable para clasificar todos los datos del conjunto. Esto puede ser porque los datos seleccionados en el conjunto de entrenamiento no son representativos y algunas características importantes sobre el conjunto de prueba no se consideraron durante

el entrenamiento. También puede deberse a que modelos que obtuvieron grandes resultados en la fase de entrenamiento hayan sobreaprendido los datos y no sepan clasificar datos nuevos. Otra opción es que haya una falta de correlación o ruido entre los datos como para poder clasificarlos correctamente, aunque no se logró observar esto en la etapa de preprocesamiento.

# 8. Conclusiones y valoración del proyecto.

En este proyecto hemos aprendido a realizar un proceso completo de extracción de conocimiento sobre un conjunto de datos, recorriendo cada una de sus etapas y generando diversos modelos para poder compararlos y extraer conclusiones de los mismos. También nos ha ayudado, junto a las prácticas, a profundizar más en entornos de Minería de Datos como *Weka,* y en los distintos algoritmos que se emplean en tareas de clasificación, que es en lo que se ha centrado este proyecto.

Al configurar los algoritmos, he tenido en cuenta los resultados que observaba en test para escoger las configuraciones presentadas. Al ver que en todos los casos eran bajas (incluso inferiores al 50% de aciertos) me he conformado con los resultados expuestos, permitiendo que sea *J48* quien, sorprendentemente, encabece el ranking de rendimiento obtenido en test pese a los resultados en entrenamiento.

También se han elaborado algunas pruebas modificando los conjuntos de datos y eliminando atributos que no han llegado a plasmarse en este proyecto debido a que no aportaban mejores resultados, ya sea por el preprocesamiento o por los datos.

# 9. Recursos.

https://datahub.io/machine-learning/cmc

https://www.openml.org/search?type=data&sort=runs&id=23&status=active

Prácticas de la asignatura.