Proyecto OLAP

Hotel

Héctor Rodríguez Salgado & Marta Loriente Nieves

Objetivo

- Construcción y comparación de modelos de minería de datos
- Análisis de datos de opiniones sobre hoteles
- Ayudar a la elección de hoteles

Obtención de datos (I)

- Comentarios sobre hoteles obtenido de Tripadvisor

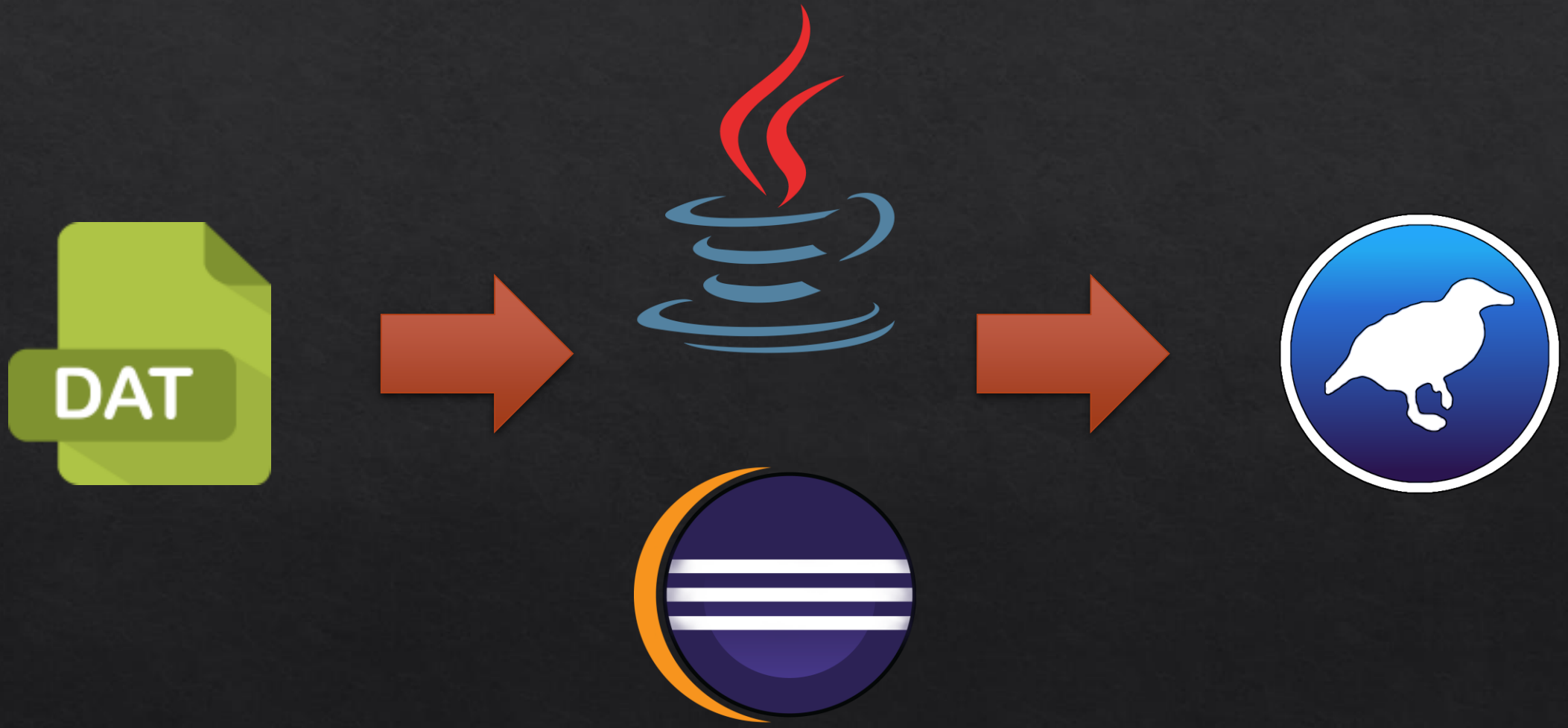
Obtención de datos (II)

Datos recogidos

- Opinión del cliente
- Valoración

```
<Author>KGBT
<Content>Wow, what charm! As a Travel Agent, I've stayed at quite a few hotel, but this is the only Historic
hotel so far...  I loved it! Had to go back for a personal stay. The decor is beautiful, the lobby furniture
fits the time period  is still comfy. The city view rooms are great - love the little balconies. Great
breakfast, nice people, great location - The Seattle Underground Tours is a 1/2 block away. I've aready sent
my folk there for a stay  have told others.
<Date>Dec 14, 2008
<No. Reader>-1
<No. Helpful>-1
<Overall>5
<Value>4
<Rooms>5
<Location>5
<Cleanliness>5
<Check in / front desk>4
<Service>-1
<Business service>4
```

Extracción, transformación y carga de datos



Aplicación de algoritmos

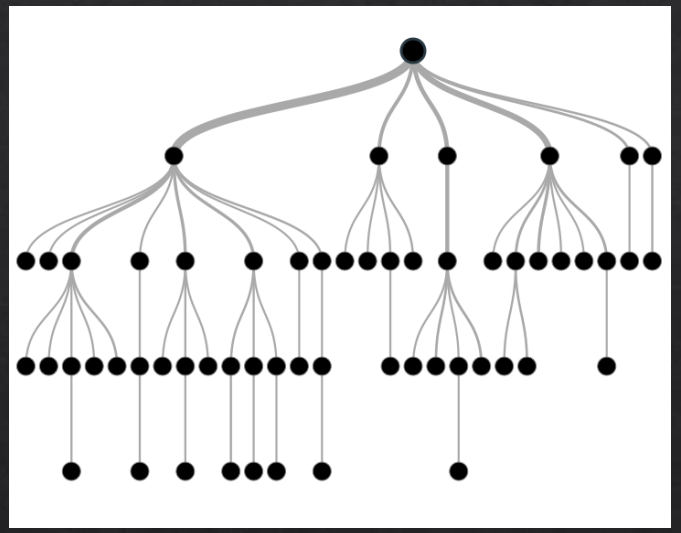
- Análisis de predicción de atributos
- Agrupación
- Asociación

Análisis de predicción de atributos (I)

Lazy learning

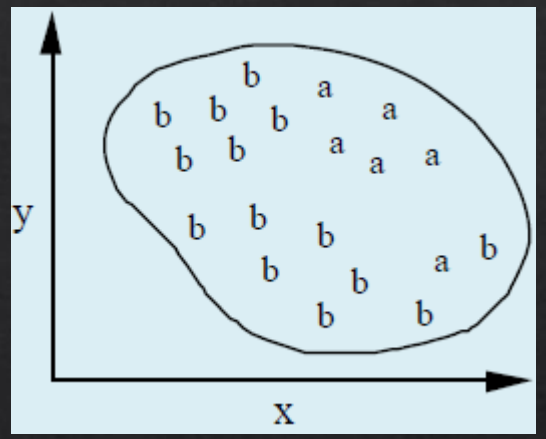
- IBk

Decision trees



- J48

Rules



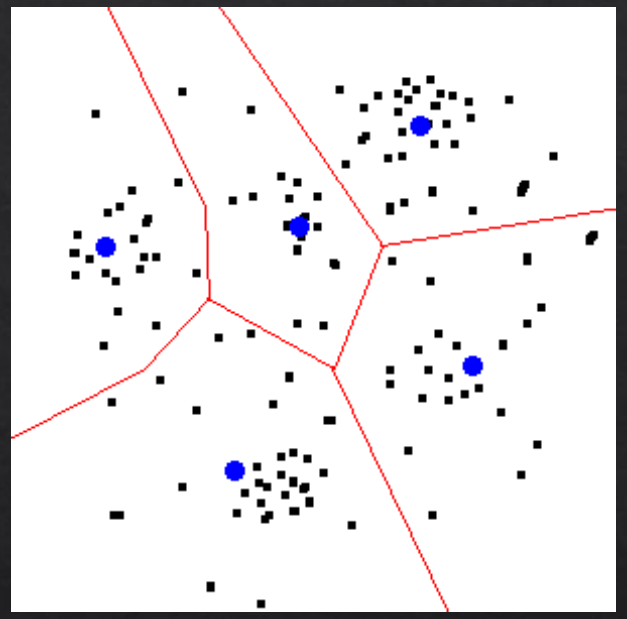
- JRIP
- ZeroR
- PART

Análisis de predicción de atributos (II)

```
=== Summary ===

Correctly Classified Instances        8305               67.9401 %
Incorrectly Classified Instances      3919               32.0599 %
Kappa statistic                          0.3049
Mean absolute error                      0.4183
Root mean squared error                  0.4627
Relative absolute error                 86.4136 %
Root relative squared error             94.0598 %
Total Number of Instances            12224

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,454     0,164     0,659       0,454   0,538      0,317    0,659     0,604     0
               0,836     0,546     0,687       0,836   0,755      0,317    0,659     0,684     1
Weighted Avg.  0,679     0,389     0,676       0,679   0,666      0,317    0,659     0,651

=== Confusion Matrix ===

    a    b   <-- classified as
 2280 2741 |    a = 0
 1178 6025 |    b = 1
```

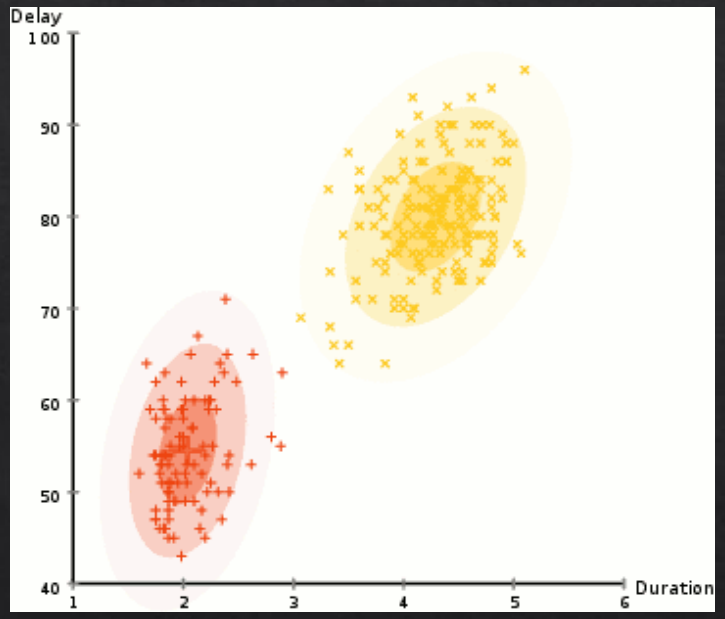
Agrupación (I)

K-Means



- SimpleKMeans

EM



- EM

Agrupación (II)

```
=== Model and evaluation on training set ===

Clustered Instances

0       2813 ( 23%)
1       9411 ( 77%)


Log likelihood: -22.38216


Class attribute: valuation
Classes to Clusters:

    0    1  <-- assigned to cluster
 1605 3416 | 0
 1208 5995 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       4624.0   37.8272 %
```

Asociación

- A priori

```
Apriori
=======

Minimum support: 0.95 (11613 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

 1. didnt=0 dont=0 12029 ==> seattle=0 11986    <conf:(1)> lift:(1) lev:(0) [5] conv:(1.1)
 2. dont=0 12115 ==> seattle=0 12070    <conf:(1)> lift:(1) lev:(0) [3] conv:(1.06)
 3. didnt=0 12113 ==> seattle=0 12067    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.03)
 4. seattle=0 didnt=0 12067 ==> dont=0 11986    <conf:(0.99)> lift:(1) lev:(0) [26] conv:(1.31)
 5. didnt=0 12113 ==> dont=0 12029    <conf:(0.99)> lift:(1) lev:(0) [24] conv:(1.27)
 6. seattle=0 dont=0 12070 ==> didnt=0 11986    <conf:(0.99)> lift:(1) lev:(0) [25] conv:(1.29)
 7. dont=0 12115 ==> didnt=0 12029    <conf:(0.99)> lift:(1) lev:(0) [24] conv:(1.26)
 8. seattle=0 12175 ==> dont=0 12070    <conf:(0.99)> lift:(1) lev:(0) [3] conv:(1.02)
 9. seattle=0 12175 ==> didnt=0 12067    <conf:(0.99)> lift:(1) lev:(0) [2] conv:(1.01)
10. didnt=0 12113 ==> seattle=0 dont=0 11986    <conf:(0.99)> lift:(1) lev:(0) [25] conv:(1.19)
```