Programa de Formación: Aplicación del Big Data en el sector del calzado para la mejora de la competitividad y de los procesos productivos

Octubre 2018

Sesión: Predicción de la rotura de stock (WEKA)
(Parte II: ejercicio rotura de stock)

Profesores:
Alex Rabasa
Carolina Belso

ROTURA DE STOCK

- **¿ Cuando se rompe stock ?**
- **¿ A qué es debido ? Almacén ? Tendencias ? Compra ? Venta ? Precio ?…**

BASE DE DATOS " BD_marca"

Nuestra base de datos se compone de información para realizar la compra de los próximos meses en el calzado de moda en diferentes marcas internacionales.

La base de datos que vamos a trabajar se llama "BD_marca", compuesta por los siguientes 15 atributos:

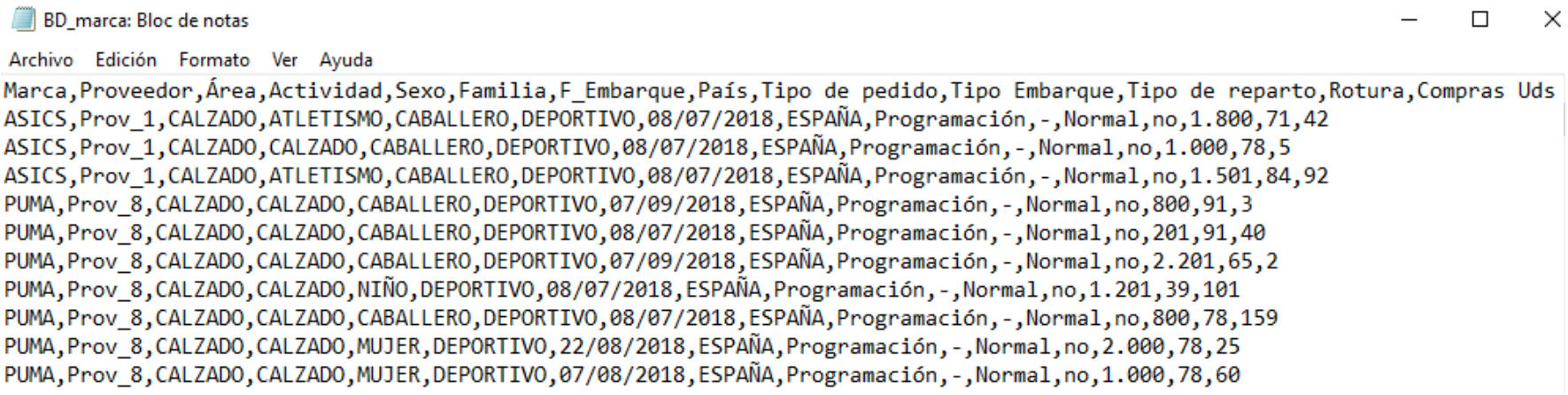
- **3 Variables Numéricas (Compra, Precio Uni., Venta).**
- **12 Variables Nominal.**

Marca	Proveedor	Área	Actividad	Sexo	Familia	F_Embarque	País	Tipo de pedido	Tipo Embarque	Tipo de reparto	Rotura	Compras Uds	Precio unitario	Venta 2017
ASICS	Prov_1	CALZADO	ATLETISMO	CABALLERO	DEPORTIVO	08/07/2018	ESPAÑA	Programación	-	Normal	no	1.800	71	42
ASICS	Prov_1	CALZADO	CALZADO	CABALLERO	DEPORTIVO	08/07/2018	ESPAÑA	Programación	-	Normal	no	1.000	78	5
ASICS	Prov_1	CALZADO	ATLETISMO	CABALLERO	DEPORTIVO	08/07/2018	ESPAÑA	Programación	-	Normal	no	1.501	84	92
PUMA	Prov_8	CALZADO	CALZADO	CABALLERO	DEPORTIVO	07/09/2018	ESPAÑA	Programación	-	Normal	no	800	91	3
PUMA	Prov_8	CALZADO	CALZADO	CABALLERO	DEPORTIVO	08/07/2018	ESPAÑA	Programación	-	Normal	no	201	91	40
PUMA	Prov_8	CALZADO	CALZADO	CABALLERO	DEPORTIVO	07/09/2018	ESPAÑA	Programación	-	Normal	no	2.201	65	2
PUMA	Prov_8	CALZADO	CALZADO	NIÑO	DEPORTIVO	08/07/2018	ESPAÑA	Programación	-	Normal	no	1.201	39	101
PUMA	Prov_8	CALZADO	CALZADO	CABALLERO	DEPORTIVO	08/07/2018	ESPAÑA	Programación	-	Normal	no	800	78	159
PUMA	Prov_8	CALZADO	CALZADO	MUJER	DEPORTIVO	22/08/2018	ESPAÑA	Programación	-	Normal	no	2.000	78	25
PUMA	Prov_8	CALZADO	CALZADO	MUJER	DEPORTIVO	07/08/2018	ESPAÑA	Programación	-	Normal	no	1.000	78	60
FILA	Prov_18	CALZADO	ATLETISMO	MUJER	DEPORTIVO	07/09/2018	CHINA	Programación	Marìtimo	-	no	2.500	45	108
CONVERSE	Prov_24	CALZADO	CALZADO	MUJER	LONA	08/07/2018	ESTADOS UNIDOS DE AMERICA	Programación	-	Normal	no	1.800	65	65
NIKE	Prov_31	CALZADO	CALZADO	CABALLERO	DEPORTIVO	20/06/2018	ESTADOS UNIDOS DE AMERICA	Reposición	-	Packing List	no	1.134	78	216

BASE DE DATOS " BD_marca" para WEKA

Para trabajar con la herramienta WEKA necesitamos modificar los archivos de la siguiente manera:

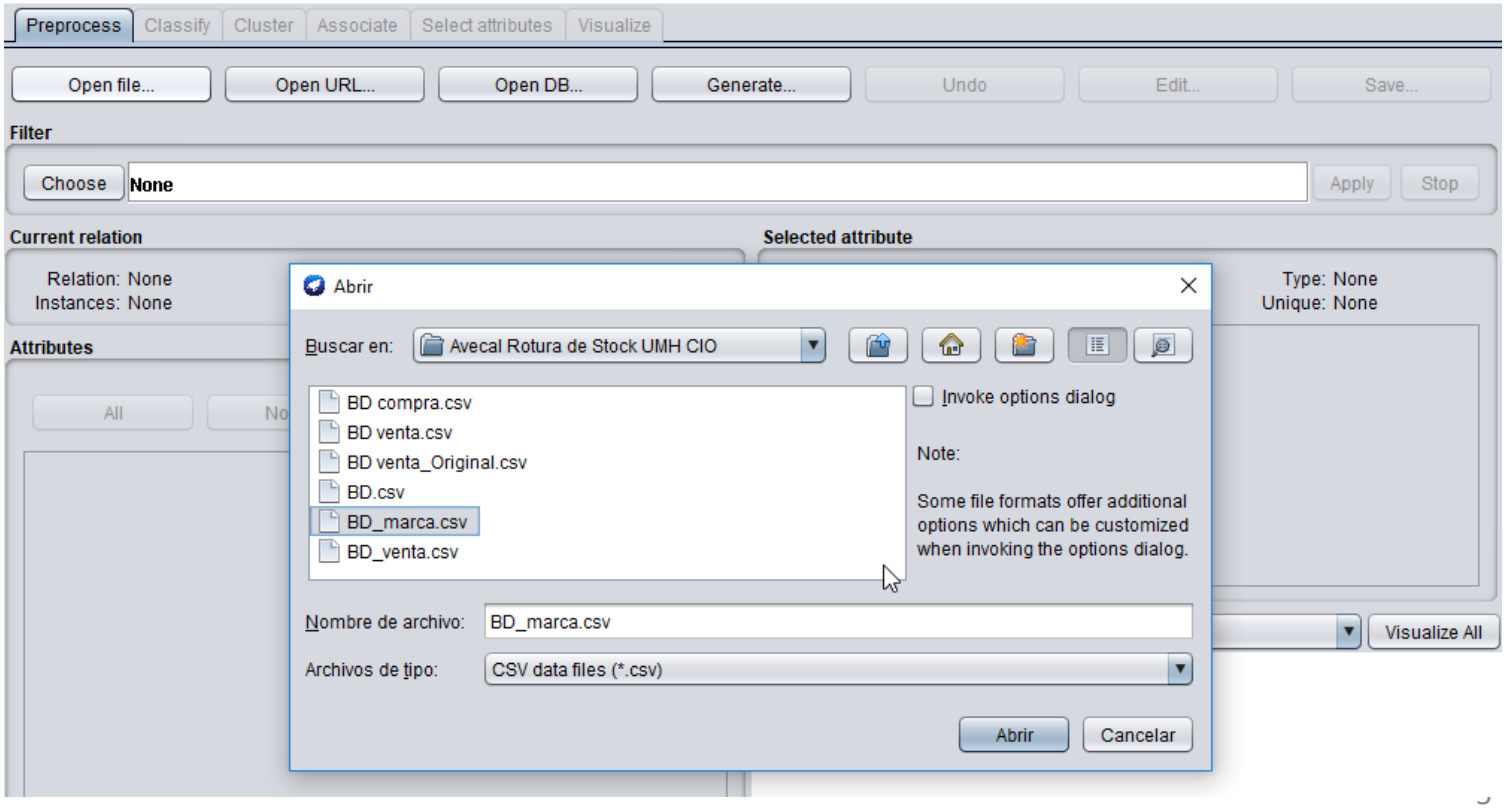
1. **Cambiar las comas de la base de datos a puntos.**
2. **Guardar el documento Excel en .csv (de .xlsx a .csv)**
3. **Editar el nuevo .csv reemplazando los puntos y comas (;) por comas (,)**



WEKA

Después de preparar nuestra base de datos tenemos que utilizar la parte de "Explorer" para realizar nuestro análisis.

Abrimos el archivo de tipo: *.csv

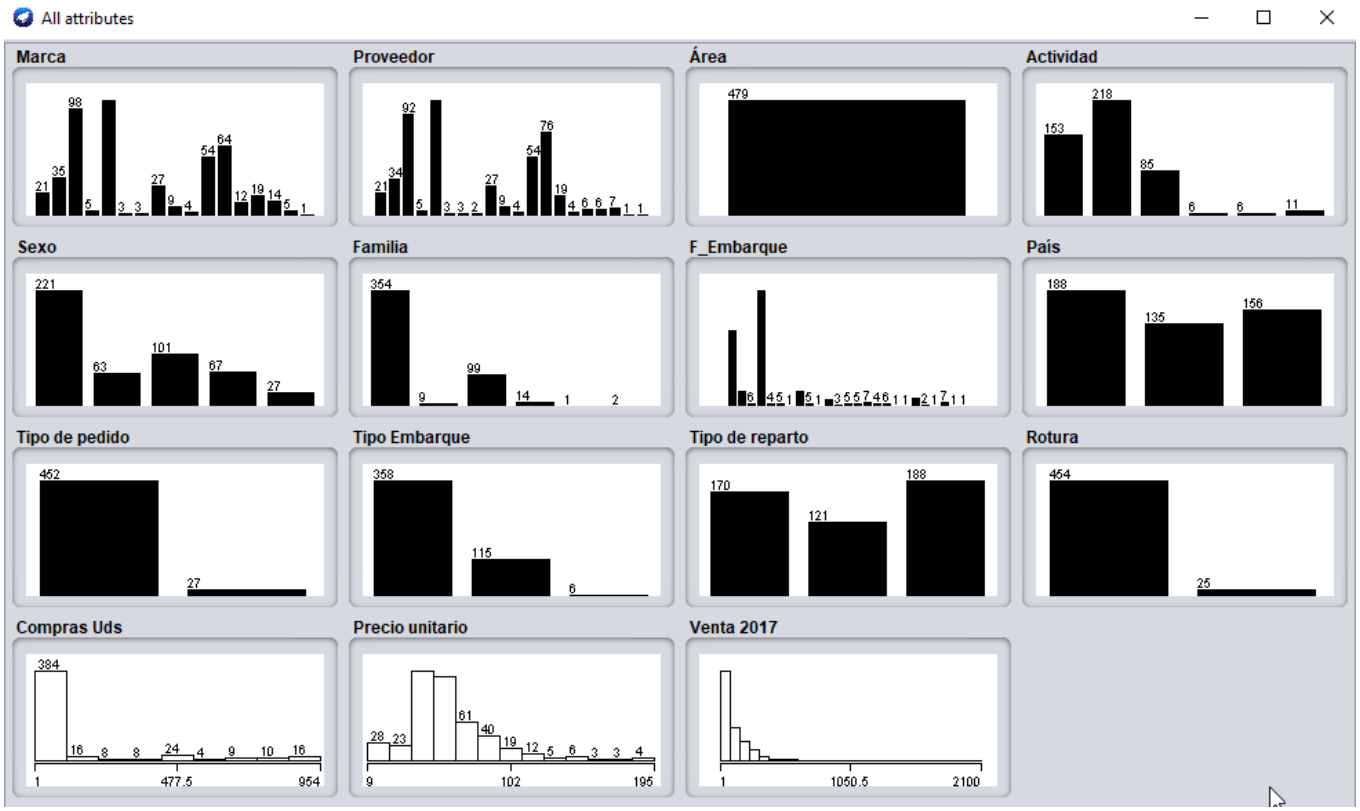


WEKA

Análisis de atributos

En la primera pantalla tenemos un pequeño resumen de cómo son los atributos que vamos a trabajar.

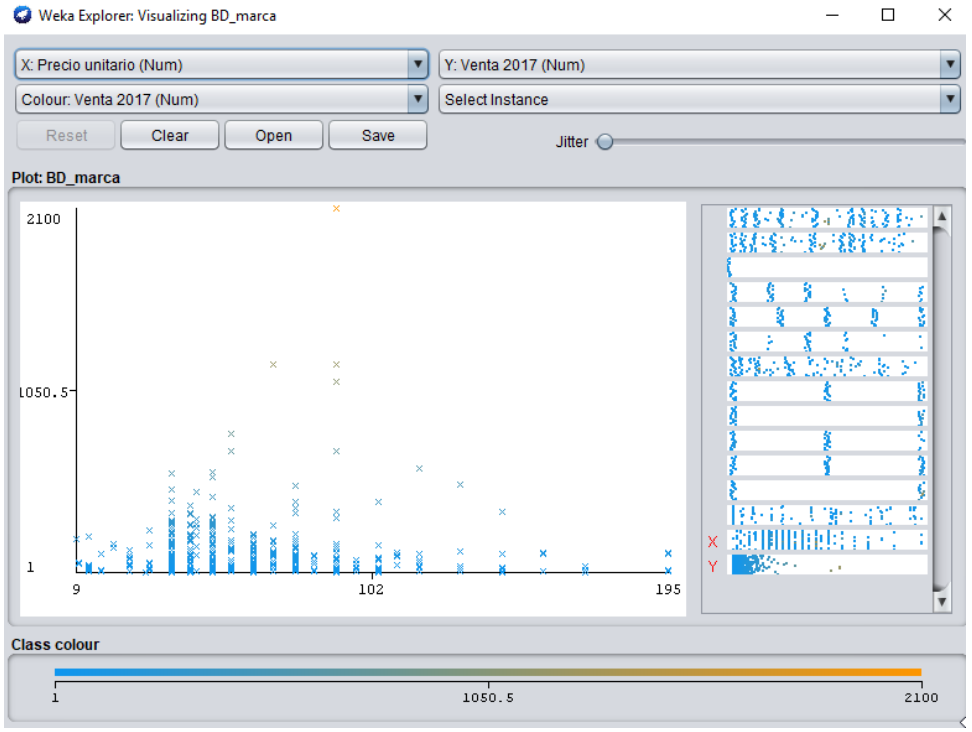
En "Visualize All" encontramos de manera gráfica cuantas categorizaciones hay por cada variable:

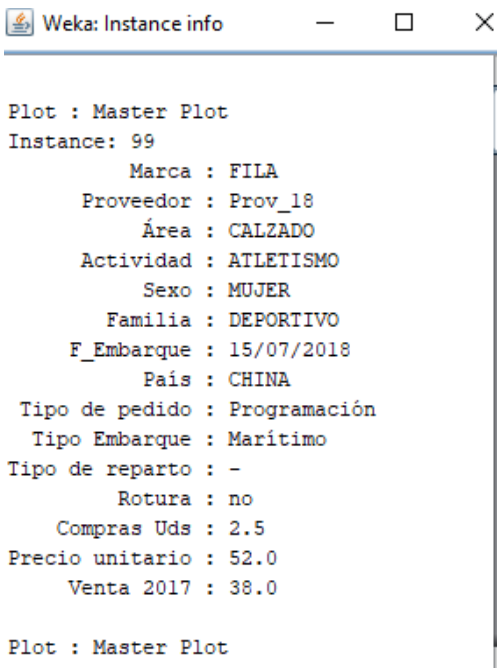


WEKA

RELACIÓN DE ATRIBUTOS

Antes de comenzar a aplicar métodos de minería de datos, analizaremos las gráficas de relación obtenidas para cada uno de los atributos que hemos seleccionado como válidos para comprobar cuales son influyentes para la rotura de stock o compra.





Precio unitario/Venta 2017

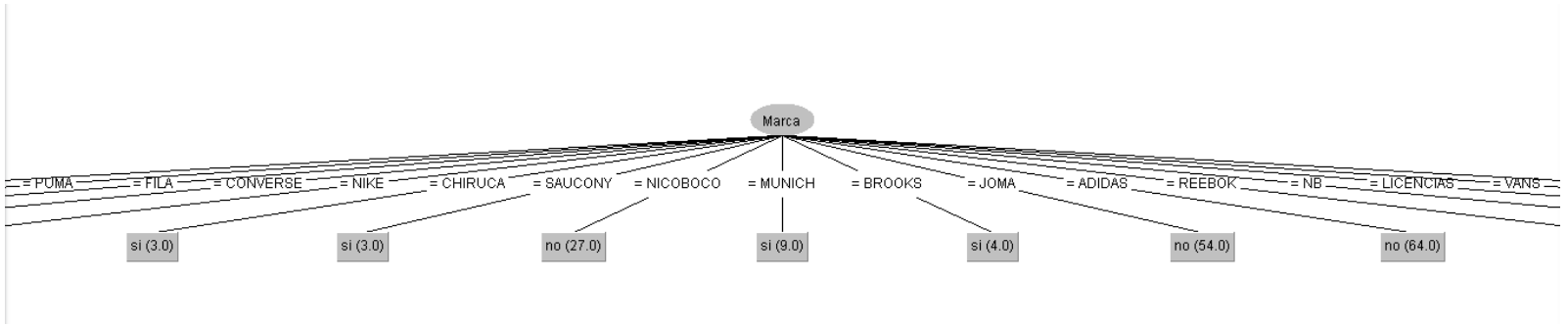
ÁRBOL DE CLASIFICACIÓN POR ROTURA DE STOCK "Choose"/tree/J48

El algoritmo J48 es uno de los algoritmos de minería de datos más utilizado. Aporta información sobre el atributo seleccionado para ver que realización hay con los demás atributos.
Para ello vamos a la pestaña de "Classify" y elegimos la variable ROTURA ya que para saber si hemos roto stock necesitamos saber que marca ha sido.

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

Marca = ASICS: no (21.0)
Marca = PUMA: no (35.0)
Marca = FILA: no (98.0)
Marca = CONVERSE: no (5.0)
Marca = NIKE: no (105.0)
Marca = CHIRUCA: si (3.0)
Marca = SAUCONY: si (3.0)
Marca = NICOBOCO: no (27.0)
Marca = MUNICH: si (9.0)
Marca = BROOKS: si (4.0)
Marca = JOMA: no (54.0)
Marca = ADIDAS: no (64.0)
Marca = REEBOK: no (12.0)
Marca = NB: no (19.0)
Marca = LICENCIAS: no (14.0)
Marca = VANS: si (5.0)
Marca = NORTH FACE: si (1.0)

Number of Leaves  :     17
```

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

Correctly Classified Instances         478               99.7912 %
Incorrectly Classified Instances         1                0.2088 %
Kappa statistic                          0.9785
Mean absolute error                      0.002
Root mean squared error                  0.0434
Relative absolute error                  1.9652 %
Root relative squared error             19.4889 %
Total Number of Instances              479
```

ÁRBOL DE CLASIFICACIÓN POR ROTURA DE STOCK



ÁRBOLES DE REGRESIÓN

IBK

A pesar de que este algoritmo no crea ningún tipo de modelo, merece la pena aplicarlo a nuestro conjunto de datos y observar los resultados. Este algoritmo es de la familia de algoritmos incluidos en "lazy learning". Este algoritmo se basa en instancias, por lo que únicamente almacena los datos presentados.

El concepto principal que fundamental de este algoritmo, es que cada instancia encontrada se va a clasificar en la clase más frecuente a la que pertenezcan sus K vecinos más cercanos.

Para un KNM = 5 y el atributo Precio Unitario:

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8286
Mean absolute error                     11.1936
Root mean squared error                 17.3126
Relative absolute error                 48.2456 %
Root relative squared error             55.9131 %
Total Number of Instances              479
```

Resultado del modelo bastante bueno, el coeficiente de correlación es de 0,82 y el error absoluto medio no es muy elevado 11,19€.

No obstante, este método, no crea un modelo para poder implementarlo ni una serie de reglas a aplicar, tan sólo clasifica las instancias.

ÁRBOLES DE REGRESIÓN

Regresión Lineal

Aplicaremos ahora el método de regresión lineal implementado en Weka. Teniendo en cuenta como afectan en mayor o menor medida el valor de los atributos para el precio unitario.

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8088
Mean absolute error                     12.3312
Root mean squared error                 18.2374
Relative absolute error                 53.1488 %
Root relative squared error             58.9    %
Total Number of Instances              479
```

Resultado del modelo bastante bueno, el coeficiente de correlación es de 0,81 y el error absoluto medio no es muy elevado 12,33€.

Weka nos muestra el modelo construido a partir de los datos y el resumen de resultados.

```
Precio unitario =

      7.7014 * Marca=JOMA,FILA,PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS +
      3.146  * Marca=REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS +
      6.188  * Marca=CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS +
      8.6846 * Marca=VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS +
      7.179  * Marca=MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS +
     -7.1222 * Marca=ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS +
     15.3897 * Marca=SAUCONY,BROOKS +
      7.7013 * Proveedor=Prov_16,Prov_9,Prov_26,Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
    -16.1624 * Proveedor=Prov_26,Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
     15.0655 * Proveedor=Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
      4.1384 * Proveedor=Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
     -3.3592 * Proveedor=Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
      3.3922 * Proveedor=Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
      7.179  * Proveedor=Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 +
     20.0003 * Proveedor=Prov_3,Prov_2,Prov_17,Prov_28 +
     15.3896 * Proveedor=Prov_17,Prov_28 +
     -5.9234 * Actividad=EQUIPO,GIMNASIO,RAQUETA,ATLETISMO,MONTAÑA +
      4.8688 * Actividad=GIMNASIO,RAQUETA,ATLETISMO,MONTAÑA +
     -4.0162 * Actividad=RAQUETA,ATLETISMO,MONTAÑA +
     13.5045 * Actividad=ATLETISMO,MONTAÑA +
     15.3459 * Sexo=MUJER,CABALLERO +
     10.1893 * Sexo=CABALLERO +
     16.5493 * Familia=SANDALIA,ZAPATO,BOTA,DEPORTIVO,LONA +
      8.7542 * Familia=ZAPATO,BOTA,DEPORTIVO,LONA +
    -14.5519 * F_Embarque=08/04/2018,22/07/2018,21/05/2018,14/07/2018,15/07/2018,07/06/2018,09/07/2018,22/08/2018,25/06/2018,17/07/2018,06/08/2018,31
    -19.3147 * F_Embarque=21/05/2018,14/07/2018,15/07/2018,07/06/2018,09/07/2018,22/08/2018,25/06/2018,17/07/2018,06/08/2018,31/08/2018,07/12/2018,07
     26.5572 * F_Embarque=14/07/2018,15/07/2018,07/06/2018,09/07/2018,22/08/2018,25/06/2018,17/07/2018,06/08/2018,31/08/2018,07/12/2018,07/09/2018,07
```

ÁRBOLES DE REGRESIÓN

M5P

Este algoritmo combina un árbol de decisión normal con funciones de regresión lineal en los nodos.

Se obtienen 10 reglas y los siguientes valores:

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.884
Mean absolute error                      9.2523
Root mean squared error                 14.4486
Relative absolute error                 39.8785 %
Root relative squared error             46.6638 %
Total Number of Instances              479
```

Resultado del modelo bastante bueno, el coeficiente de correlación es de 0,88 y el error absoluto medio no es muy elevado 9,25€.

Se genera el siguiente árbol de decisión:

```
M5 pruned model tree:
(using smoothed linear models)

Marca=PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS <= 0.5 :
|   Proveedor=Prov_9,Prov_26,Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 <= 0.5 :
|   |   Familia=SANDALIA,ZAPATO,BOTA,DEPORTIVO,LONA <= 0.5 : LM1 (13/4.603%)
|   |   Familia=SANDALIA,ZAPATO,BOTA,DEPORTIVO,LONA >  0.5 : LM2 (32/3.269%)
|   Proveedor=Prov_9,Prov_26,Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28 >  0.5 :
|   |   Sexo=MUJER,CABALLERO <= 0.5 :
|   |   |   Actividad=GIMNASIO,RAQUETA,ATLETISMO,MONTAÑA <= 0.5 : LM3 (26/4.751%)
|   |   |   Actividad=GIMNASIO,RAQUETA,ATLETISMO,MONTAÑA >  0.5 : LM4 (28/0%)
|   |   Sexo=MUJER,CABALLERO >  0.5 :
|   |   |   Marca=FILA,PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS <= 0.5 :
|   |   |   |   Compras Uds <= 2 : LM5 (14/35.909%)
|   |   |   |   Compras Uds >  2 : LM6 (24/12.33%)
|   |   |   Marca=FILA,PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS >  0.5 : LM7 (56/14.775%)
Marca=PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS >  0.5 :
|   Sexo=MUJER,CABALLERO <= 0.5 : LM8 (77/27.811%)
|   Sexo=MUJER,CABALLERO >  0.5 :
|   |   Compras Uds <= 78 : LM9 (134/37.296%)
|   |   Compras Uds >  78 : LM10 (75/70.771%)
```

ÁRBOLES DE REGRESIÓN

M5P

Una de las reglas generadas son las siguientes: **:**

```
LM num: 1
Precio unitario =
        3.3553 * Marca=NICOBOCO,JOMA,FILA,PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS
        + 6.5626 * Marca=JOMA,FILA,PUMA,REEBOK,CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS
        + 0.4924 * Marca=CONVERSE,VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS
        + 0.8443 * Marca=VANS,ADIDAS,NIKE,MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS
        + 0.9271 * Marca=MUNICH,NB,ASICS,NORTH FACE,CHIRUCA,SAUCONY,BROOKS
        + 2.7908 * Marca=NORTH FACE,CHIRUCA,SAUCONY,BROOKS
        + 8.3582 * Proveedor=Prov_25,Prov_13,Prov_16,Prov_9,Prov_26,Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov
        + 1.6368 * Proveedor=Prov_9,Prov_26,Prov_18,Prov_8,Prov_32,Prov_24,Prov_10,Prov_31,Prov_11,Prov_14,Prov_1,Prov_3,Prov_2,Prov_17,Prov_28
        - 9.8522 * Actividad=EQUIPO,GIMNASIO,RAQUETA,ATLETISMO,MONTAÑA
        + 2.2835 * Actividad=GIMNASIO,RAQUETA,ATLETISMO,MONTAÑA
        + 0.9255 * Actividad=ATLETISMO,MONTAÑA
        + 3.714 * Sexo=MUJER,CABALLERO
        + 0.7246 * Sexo=CABALLERO
        + 16.7143 * Familia=SANDALIA,ZAPATO,BOTA,DEPORTIVO,LONA
        - 8.2285 * Familia=DEPORTIVO,LONA
        - 1.1664 * F_Embarque=22/08/2018,25/06/2018,17/07/2018,06/08/2018,31/08/2018,07/12/2018,07/09/2018,07/10/2018,07/08/2018,08/07/2018,23/07/201
        + 1.3897 * F_Embarque=25/06/2018,17/07/2018,06/08/2018,31/08/2018,07/12/2018,07/09/2018,07/10/2018,07/08/2018,08/07/2018,23/07/2018,20/06/201
        - 0.6394 * F_Embarque=07/10/2018,07/08/2018,08/07/2018,23/07/2018,20/06/2018,27/07/2018,20/12/2018,20/07/2018,20/08/2018
        + 0.6388 * F_Embarque=07/08/2018,08/07/2018,23/07/2018,20/06/2018,27/07/2018,20/12/2018,20/07/2018,20/08/2018
        + 0.8005 * F_Embarque=23/07/2018,20/06/2018,27/07/2018,20/12/2018,20/07/2018,20/08/2018
        - 1.0098 * Tipo Embarque=Aéreo,-
        + 0.6003 * Tipo de reparto=Normal
        + 0.0069 * Compras Uds
        - 0.0011 * Venta 2017
```

ÁRBOLES DE REGRESIÓN

Obtención del modelo óptimo

En la siguiente tabla, podemos ver un pequeño resumen con los datos más importantes y relevantes para tomar la decisión sobre que método ha obtenido mejores resultados y, por tanto, qué modelo implementaremos en nuestra aplicación.

Algoritmo	Coeficiente de Correlación	Error de la media absoluta	Error absoluto relativo
IBK	0.82	11,19 €	48.24%
Regresión Lineal	0.81	12,33 €	53.14%
M5P	0.88	9,25 €	39.87%

Queda bastante claro a simple vista, que el modelo que deberíamos implementar es el construido por el algoritmo M5P, pues es el que ha alcanzado un coeficiente de correlación mejor, conjuntamente con una media de error absoluto y relativo.

GRACIAS