Evaluación Data Science & Business Intelligence

Pentaho / Weka

Inmaculada Perea Fernández

julio 2017

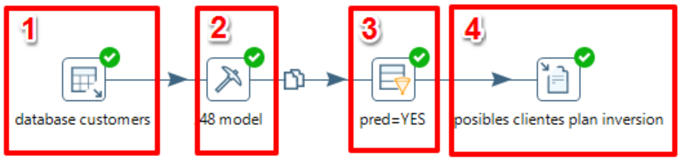
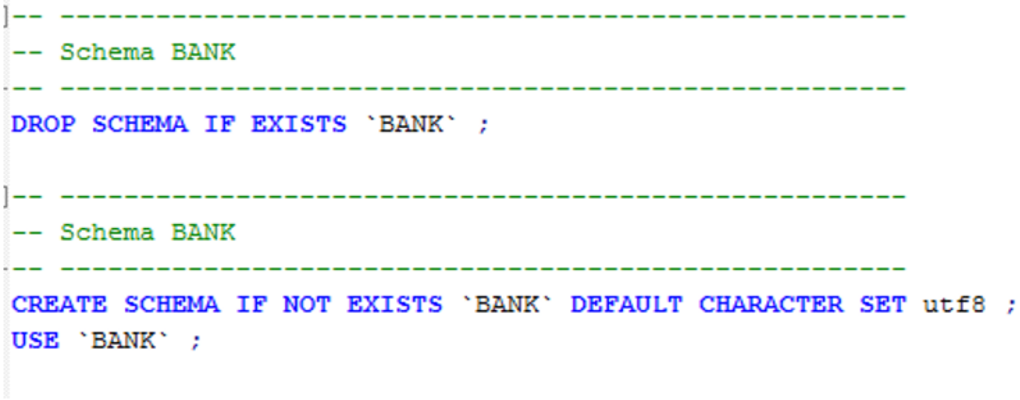


Figure 1: Transformación kettle

Bloque 1: Lectura base de datos

Antes de realizar la lectura de la base de datos es necesario ejecutar el script customes.sql, este script realiza las siguientes acciones:

- Crea schema *BANK*



```
-- -----------------------------------------------------
-- Schema BANK
-- -----------------------------------------------------
DROP SCHEMA IF EXISTS `BANK` ;

-- -----------------------------------------------------
-- Schema BANK
-- -----------------------------------------------------
CREATE SCHEMA IF NOT EXISTS `BANK` DEFAULT CHARACTER SET utf8 ;
USE `BANK` ;
```

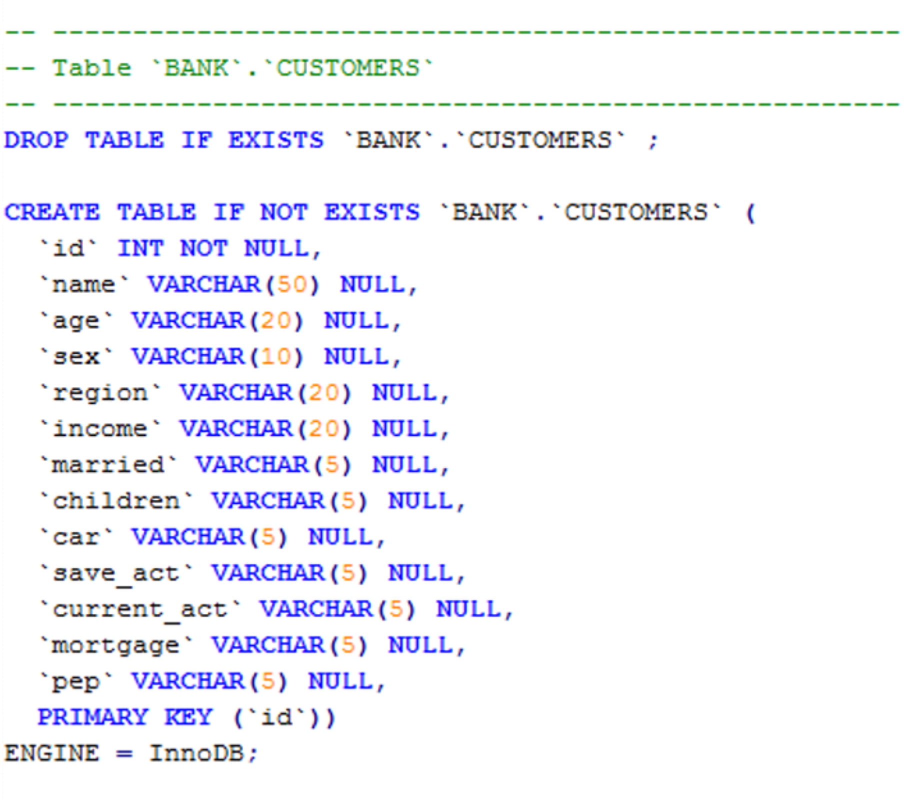
Figure 2: Creación del schema BANK

- Crea la tabla *CUSTOMERS*
- Inserta valores en los clientes en la base de datos

Una vez poblada la base de datos se podrá consultar usando kettle y el componente "Entrada tabla", paramétrizado como se muestra en la figura a continuación:

Bloque 2: Modelo predictivo

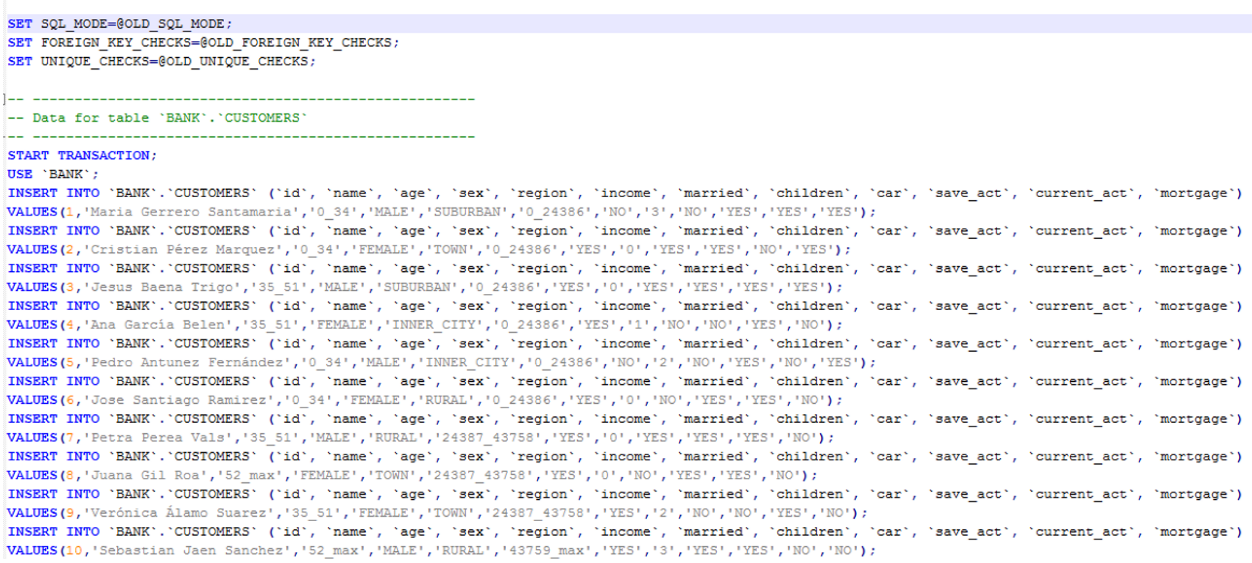
El conjunto de datos data-bank consta de 600 observaciones y 12 variables, a continuación el significado de cada una:



```
-- -----------------------------------------------------
-- Table `BANK`.`CUSTOMERS`
-- -----------------------------------------------------
DROP TABLE IF EXISTS `BANK`.`CUSTOMERS` ;

CREATE TABLE IF NOT EXISTS `BANK`.`CUSTOMERS` (
  `id` INT NOT NULL,
  `name` VARCHAR(50) NULL,
  `age` VARCHAR(20) NULL,
  `sex` VARCHAR(10) NULL,
  `region` VARCHAR(20) NULL,
  `income` VARCHAR(20) NULL,
  `married` VARCHAR(5) NULL,
  `children` VARCHAR(5) NULL,
  `car` VARCHAR(5) NULL,
  `save_act` VARCHAR(5) NULL,
  `current_act` VARCHAR(5) NULL,
  `mortgage` VARCHAR(5) NULL,
  `pep` VARCHAR(5) NULL,
  PRIMARY KEY (`id`))
ENGINE = InnoDB;
```

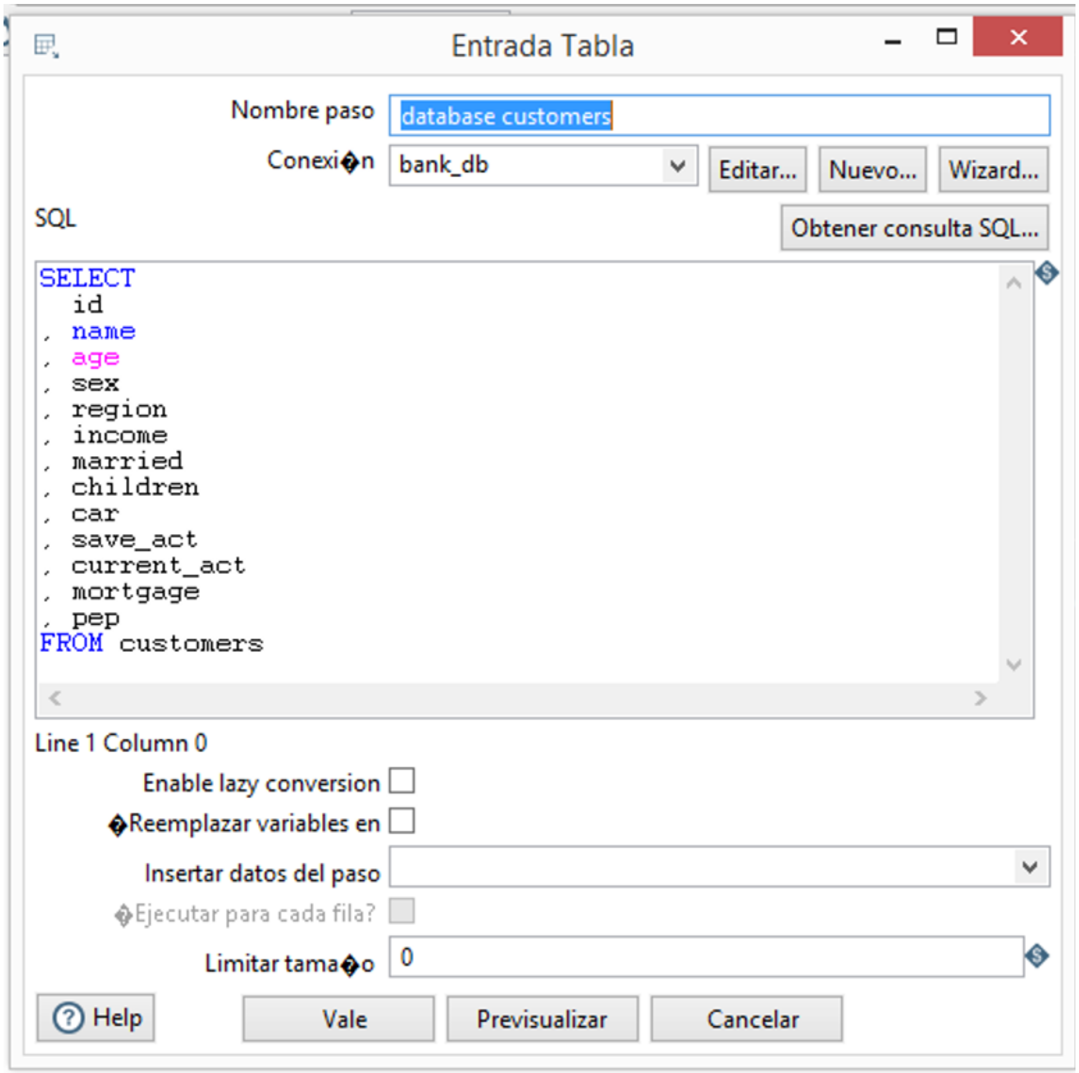
Figure 3: Creación tabla CUSTOMERS



```
SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;

-- -----------------------------------------------------
-- Data for table `BANK`.`CUSTOMERS`
-- -----------------------------------------------------
START TRANSACTION;
USE `BANK`;
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(1,'Maria Gerrero Santamaria','0_34','MALE','SUBURBAN','0_24386','NO','3','NO','YES','YES','YES');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(2,'Cristian Pérez Marquez','0_34','FEMALE','TOWN','0_24386','YES','0','YES','YES','NO','YES');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(3,'Jesus Baena Trigo','35_51','MALE','SUBURBAN','0_24386','YES','0','YES','YES','YES','YES');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(4,'Ana García Belen','35_51','FEMALE','INNER_CITY','0_24386','YES','1','NO','NO','YES','NO');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(5,'Pedro Antunez Fernández','0_34','MALE','INNER_CITY','0_24386','NO','2','NO','YES','NO','YES');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(6,'Jose Santiago Ramirez','0_34','FEMALE','RURAL','0_24386','YES','0','NO','YES','YES','NO');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(7,'Petra Perea Vals','35_51','MALE','RURAL','24387_43758','YES','0','YES','YES','YES','NO');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(8,'Juana Gil Roa','52_max','FEMALE','TOWN','24387_43758','YES','0','NO','YES','YES','NO');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(9,'Verónica Álamo Suarez','35_51','FEMALE','TOWN','24387_43758','YES','2','NO','NO','YES','NO');
INSERT INTO `BANK`.`CUSTOMERS` (`id`, `name`, `age`, `sex`, `region`, `income`, `married`, `children`, `car`, `save_act`, `current_act`, `mortgage`)
VALUES(10,'Sebastian Jaen Sanchez','52_max','MALE','RURAL','43759_max','YES','3','YES','YES','NO','NO');
```

Figure 4: Inserción de clientes



Nombre paso: database customers

Conexi�n: bank_db

SQL

```
SELECT
  id
, name
, age
, sex
, region
, income
, married
, children
, car
, save_act
, current_act
, mortgage
, pep
FROM customers
```

Line 1 Column 0

Enable lazy conversion □

�Reemplazar variables en □

Insertar datos del paso

�Ejecutar para cada fila? □

Limitar tama�o 0

Figure 5: Conexion DB

- **id**: identificador único
- **age**: edad del cliente en años
- **sex**: sexo (MALE / FEMALE)
- **region**: inner_city/rural/suburban/town
- **income**: sueldo del cliente
- **married**: está casado el cliente (YES/NO)
- **children**: número de hijos del cliente
- **car**: tiene el cliente coche propio (YES/NO)
- **save_acct**: tiene el cliente cuenta de ahorro (YES/NO)
- **current_acct**: tiene el cliente una cuenta corriente (YES/NO)
- **mortgage**: tiene hipoteca el cliente (YES/NO)
- **pep**: contratará el cliente un plan de inversión (YES/NO)

La variable objetivo es *pep*

```
credit_approval = read.table(file="data/bank-data.csv", header=TRUE, sep=",", dec=".")
dim(credit_approval)
```

```
## [1] 600  12
```

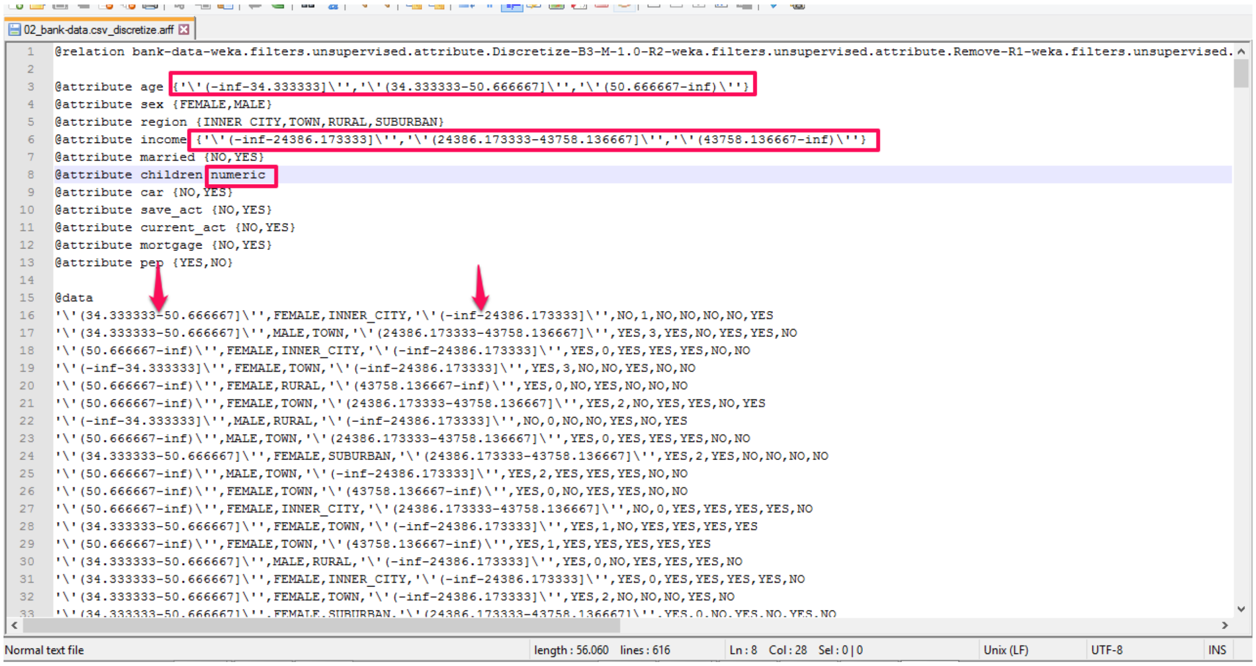
```
summary(credit_approval)
```

```
##        id           age            sex             region   
##  ID12101:  1   Min.   :18.00   FEMALE:300   INNER_CITY:269  
##  ID12102:  1   1st Qu.:30.00   MALE  :300   RURAL     : 96  
##  ID12103:  1   Median :42.00                SUBURBAN  : 62  
##  ID12104:  1   Mean   :42.40                TOWN      :173  
##  ID12105:  1   3rd Qu.:55.25                                
##  ID12106:  1   Max.   :67.00                                
##  (Other):594                                                
##      income      married      children      car      save_act  current_act
##  Min.   : 5014   NO :204   Min.   :0.000   NO :304   NO :186   NO :145    
##  1st Qu.:17265   YES:396   1st Qu.:0.000   YES:296   YES:414   YES:455    
##  Median :24925             Median :1.000                                  
##  Mean   :27524             Mean   :1.012                                  
##  3rd Qu.:36173             3rd Qu.:2.000                                  
##  Max.   :63130             Max.   :3.000                                  
##                                                                           
##  mortgage   pep     
##  NO :391   NO :326  
##  YES:209   YES:274  
##                     
##                     
##                     
##                     
## 
```

Para construir el modelo se han realizado las siguientes transformaciones:

Filtrado de atributos

El atributo *id* no es de interés para el estudio, por ello lo eliminaremos

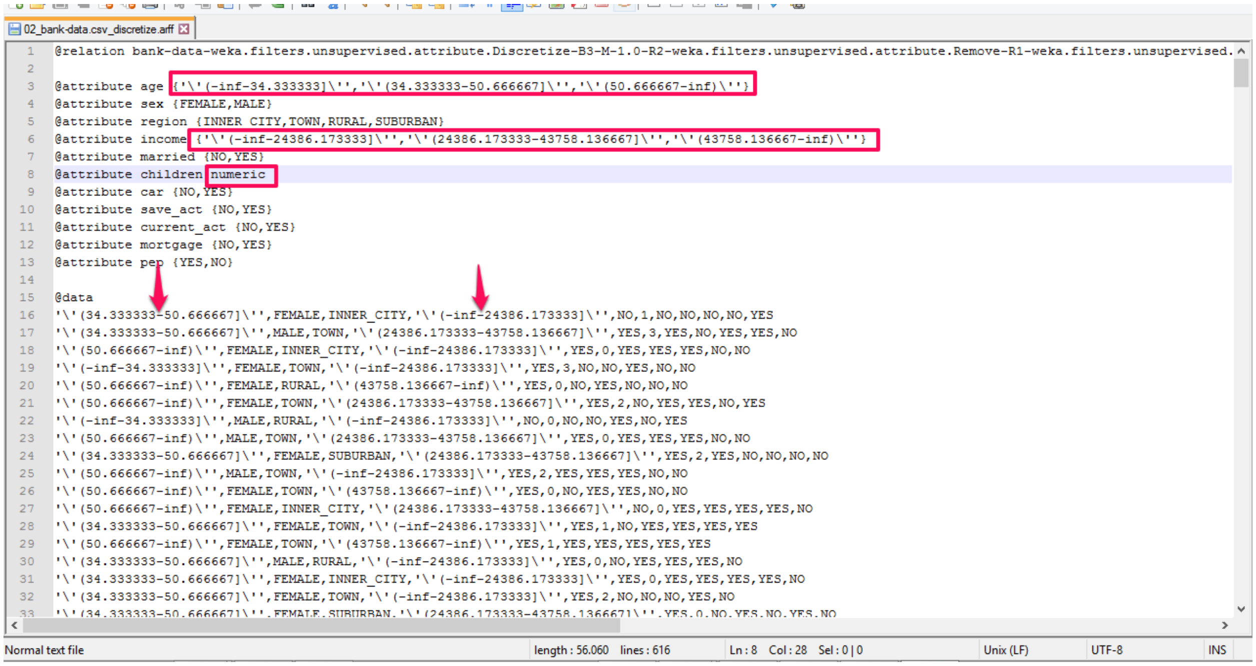

02_bank-data.csv_discretize.arff
Normal text file
length : 56.060 lines : 616
Ln : 8 Col : 28 Sel : 0 | 0
Unix (LF)
UTF-8
INS

```
@relation bank-data-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.

@attribute age {'\'(-inf-34.333333]\'','\'(34.333333-50.666667]\'','\'(50.666667-inf)\''}
@attribute sex {FEMALE,MALE}
@attribute region {INNER CITY,TOWN,RURAL,SUBURBAN}
@attribute income {'\'(-inf-24386.173333]\'','\'(24386.173333-43758.136667]\'','\'(43758.136667-inf)\''}
@attribute married {NO,YES}
@attribute children numeric
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

@data
'\'(34.333333-50.666667]\'',FEMALE,INNER_CITY,'\'(-inf-24386.173333]\'',NO,1,NO,NO,NO,NO,YES
'\'(34.333333-50.666667]\'',MALE,TOWN,'\'(24386.173333-43758.136667]\'',YES,3,YES,NO,YES,YES,NO
'\'(50.666667-inf)\'',FEMALE,INNER_CITY,'\'(-inf-24386.173333]\'',YES,0,YES,YES,YES,NO,NO
'\'(-inf-34.333333]\'',FEMALE,TOWN,'\'(-inf-24386.173333]\'',YES,3,NO,NO,YES,NO,NO
'\'(50.666667-inf)\'',FEMALE,RURAL,'\'(43758.136667-inf)\'',YES,0,NO,YES,NO,NO,NO
'\'(50.666667-inf)\'',FEMALE,TOWN,'\'(24386.173333-43758.136667]\'',YES,2,NO,YES,YES,NO,YES
'\'(-inf-34.333333]\'',MALE,RURAL,'\'(-inf-24386.173333]\'',NO,0,NO,NO,YES,NO,YES
'\'(50.666667-inf)\'',MALE,TOWN,'\'(24386.173333-43758.136667]\'',YES,0,YES,YES,YES,NO,NO
'\'(34.333333-50.666667]\'',FEMALE,SUBURBAN,'\'(24386.173333-43758.136667]\'',YES,2,YES,NO,NO,NO,NO
'\'(50.666667-inf)\'',MALE,TOWN,'\'(-inf-24386.173333]\'',YES,2,YES,YES,YES,NO,NO
'\'(50.666667-inf)\'',FEMALE,TOWN,'\'(43758.136667-inf)\'',YES,0,NO,YES,YES,NO,NO
'\'(50.666667-inf)\'',FEMALE,INNER_CITY,'\'(24386.173333-43758.136667]\'',NO,0,YES,YES,YES,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,TOWN,'\'(-inf-24386.173333]\'',YES,1,NO,YES,YES,YES,YES
'\'(50.666667-inf)\'',FEMALE,TOWN,'\'(43758.136667-inf)\'',YES,1,YES,YES,YES,YES,YES
'\'(34.333333-50.666667]\'',MALE,RURAL,'\'(-inf-24386.173333]\'',YES,0,NO,YES,YES,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,INNER_CITY,'\'(-inf-24386.173333]\'',YES,0,YES,YES,YES,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,TOWN,'\'(-inf-24386.173333]\'',YES,2,NO,NO,NO,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,SUBURBAN,'\'(24386.173333-43758.136667]\'',YES,0,NO,YES,NO,YES,NO
```

Figure 6: Transformación kettle


02_bank-data.csv_discretize.arff
Normal text file
length : 56.060 lines : 616
Ln : 8 Col : 28 Sel : 0 | 0
Unix (LF)
UTF-8
INS

```
@relation bank-data-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.

@attribute age {'\'(-inf-34.333333]\'','\'(34.333333-50.666667]\'','\'(50.666667-inf)\''}
@attribute sex {FEMALE,MALE}
@attribute region {INNER CITY,TOWN,RURAL,SUBURBAN}
@attribute income {'\'(-inf-24386.173333]\'','\'(24386.173333-43758.136667]\'','\'(43758.136667-inf)\''}
@attribute married {NO,YES}
@attribute children numeric
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

@data
'\'(34.333333-50.666667]\'',FEMALE,INNER_CITY,'\'(-inf-24386.173333]\'',NO,1,NO,NO,NO,NO,YES
'\'(34.333333-50.666667]\'',MALE,TOWN,'\'(24386.173333-43758.136667]\'',YES,3,YES,NO,YES,YES,NO
'\'(50.666667-inf)\'',FEMALE,INNER_CITY,'\'(-inf-24386.173333]\'',YES,0,YES,YES,YES,NO,NO
'\'(-inf-34.333333]\'',FEMALE,TOWN,'\'(-inf-24386.173333]\'',YES,3,NO,NO,YES,NO,NO
'\'(50.666667-inf)\'',FEMALE,RURAL,'\'(43758.136667-inf)\'',YES,0,NO,YES,NO,NO,NO
'\'(50.666667-inf)\'',FEMALE,TOWN,'\'(24386.173333-43758.136667]\'',YES,2,NO,YES,YES,NO,YES
'\'(-inf-34.333333]\'',MALE,RURAL,'\'(-inf-24386.173333]\'',NO,0,NO,NO,YES,NO,YES
'\'(50.666667-inf)\'',MALE,TOWN,'\'(24386.173333-43758.136667]\'',YES,0,YES,YES,YES,NO,NO
'\'(34.333333-50.666667]\'',FEMALE,SUBURBAN,'\'(24386.173333-43758.136667]\'',YES,2,YES,NO,NO,NO,NO
'\'(50.666667-inf)\'',MALE,TOWN,'\'(-inf-24386.173333]\'',YES,2,YES,YES,YES,NO,NO
'\'(50.666667-inf)\'',FEMALE,TOWN,'\'(43758.136667-inf)\'',YES,0,NO,YES,YES,NO,NO
'\'(50.666667-inf)\'',FEMALE,INNER_CITY,'\'(24386.173333-43758.136667]\'',NO,0,YES,YES,YES,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,TOWN,'\'(-inf-24386.173333]\'',YES,1,NO,YES,YES,YES,YES
'\'(50.666667-inf)\'',FEMALE,TOWN,'\'(43758.136667-inf)\'',YES,1,YES,YES,YES,YES,YES
'\'(34.333333-50.666667]\'',MALE,RURAL,'\'(-inf-24386.173333]\'',YES,0,NO,YES,YES,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,INNER_CITY,'\'(-inf-24386.173333]\'',YES,0,YES,YES,YES,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,TOWN,'\'(-inf-24386.173333]\'',YES,2,NO,NO,NO,YES,NO
'\'(34.333333-50.666667]\'',FEMALE,SUBURBAN,'\'(24386.173333-43758.136667]\'',YES,0,NO,YES,NO,YES,NO
```

Figure 7: Transformación kettle

Discretización

Transformaciones editando el fichero:

‘’(-inf-34.333333]” por 0_34

‘’(34.333333-50.666667]” por 35_51

‘’(50.666667-inf)” por 52_max

‘’(-inf-24386.173333]” por 0_24386

Módulo 3: Filtrado

Módulo 4: Salida