**BI Calidad del aire Barcelona**

Santiago Rueda M.

30/11/2017

**Fuente de datos:**

Se tomo los datos de cuatro sensores de calidad del aire en Catalunya, tres de Barcelona y uno de Girona.

**Fuente**: <http://dtes.gencat.cat/icqa/>

**Software:**

* Tratamiento de Datos con la herramienta Pentaho Community v8.0 (Spoon)
* Predicción Weka v3.8
* Cubo Olap Schema Workbench v8.1
* Servidos Apache BBDD MySQL, servidor local Xamp

**Girona:**

SO2 (µg/m³) NO (µg/m³) NO2 (µg/m³) CO (mg/m³)

**Gracia:**

SO2 (µg/m³) NO (µg/m³) NO2 (µg/m³) O3 (µg/m³) CO (mg/m³) PM10 (µg/m³)

**Eixample:**

SO2 (µg/m³) NO (µg/m³) NO2 (µg/m³) O3 (µg/m³) CO (mg/m³) PM10 (µg/m³)

**Palau Reial:**

SO2 (µg/m³) NO (µg/m³) NO2 (µg/m³) O3 (µg/m³) CO (mg/m³) PM10 (µg/m³)

Se tomaron para el tratamiento de la información los parámetros comunes :

SO2 (µg/m³) NO (µg/m³) NO2 (µg/m³) CO (mg/m³)

**Requisitos:**

Fechas de las medidas: 01/11/2016 al 31/10/2017

Medidas Automáticas

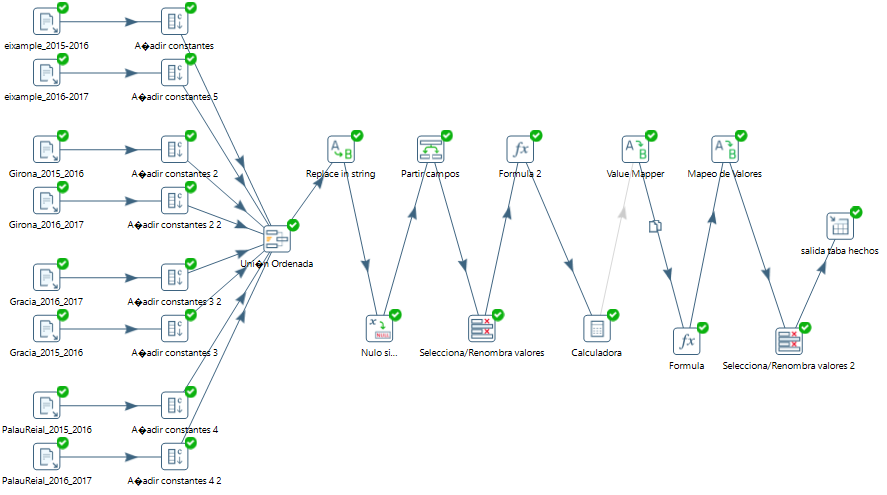
Data set de Medias diarias por horas y por dias

**PROCEDIMIENTO:**

**Tratamiento de los Datos.**

**Transformación (tranf\_estaciones)**

Extracción de datos de la fuente, normalización y creación de tabla de hechos



* Extracción de la información en archivo formato csv
* Adición de constante tipo de barrio según fuente de datos
* Unión de todas las fuentes de forma ordenada
* Normalización de puntos decimales
* Set nulos en valores inexistentes o sin datos
* Separación de campos de fecha y hora
* Cambio de tipo fecha para cálculos posteriores
* Formula para normalizar formato de hora

**IF**( **LEN**([hora])=***5***; [hora]&***":00"***; **IF**( **LEN**([hora])=***4***; ***"0"***&[hora]&***":00"***; [hora]) )

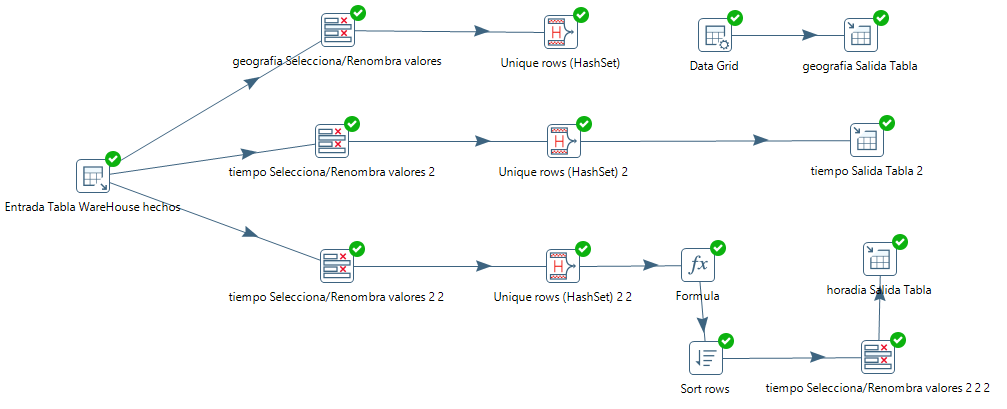
* Calculo de año, mes y día
* Calculo de día laborable
* Fórmula para cálculo de trimestres

**IF**(**[MES]** <=***3***; ***"1erTrimestre"***; **IF**(**AND**(**[MES]** >***3*** ; **[MES]** <=***6***); ***"2do Trimestre"***; **IF**(**AND**(**[MES]**>***6*** ; **[MES]** <=***9***);***"3do Trimestre"***;***"4o Trimestre"*** )))

* Calculo de nombre del día de la semana
* Cambio de formato fecha
* Salida a tabla hechos en BBDD whcalidadaire

**Transformación (tranf\_dimensiones)**

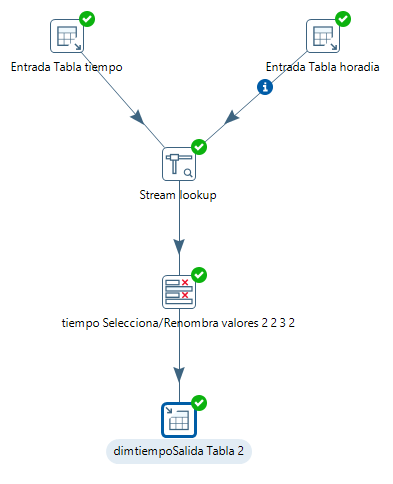
Generación de tablas de dimensiones a partir de tabla de hechos



* Entrada Tabla hechos (producto de la transformación anterior
* Selección de cada tiempo, horadía y geografía
* Remover duplicados para creación de tablas y claves primarias
* Fórmula para clasificación de la jornada del día.
* **if**(**HOUR**([hora])<=***07***; ***"madrugada"***; **if**(**AND**(**HOUR**([hora])>***07***; **HOUR**([hora])<=***12***); ***"mañana"***; **if**(**AND**(**HOUR**([hora])>***12***; **HOUR**([hora])<=***16***); ***"mediodia"***; **if**(**AND**(**HOUR**([hora])>***16***; **HOUR**([hora])<=***20***); ***"tarde"***; **if**(**HOUR**([hora])>***20***; ***"noche"***; ) ))
* Salida de tablas de geografía, tiempo y jornadas

**Transformación (tranf\_relacionadas)**

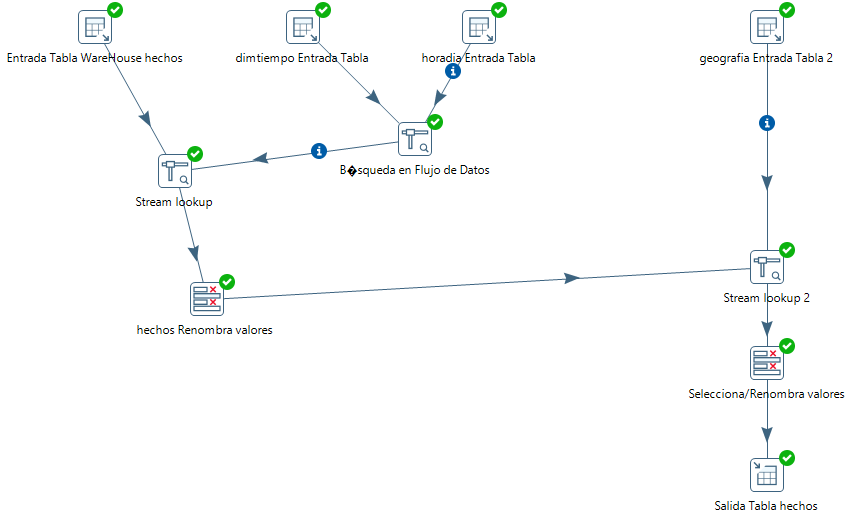
Unión de la tabla de tiempo y la de jornada por clave primaria y foránea



* Entrada de datos de tabla de tiempo y horadia
* Unión por clave común (hora)
* Cambio de tipo de dato para clave primaria
* Salida tabla dimensión tiempo.

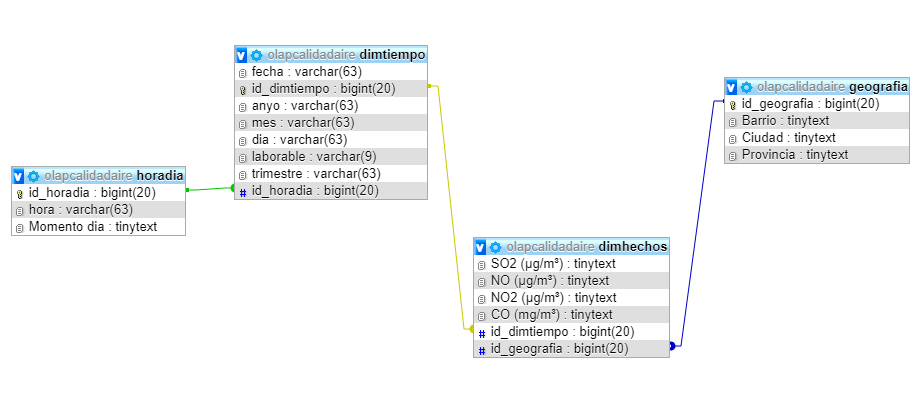
**Transformación transf\_relacionarechostiempo**

Relación de tabla de hechos con tabla de tiempo, hora día y geografía según clave primaria y foránea

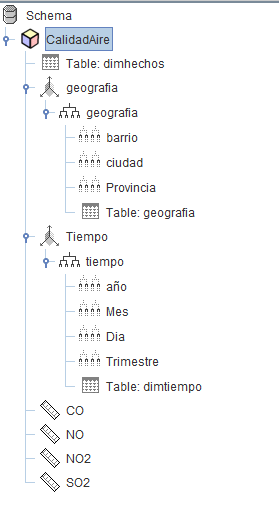


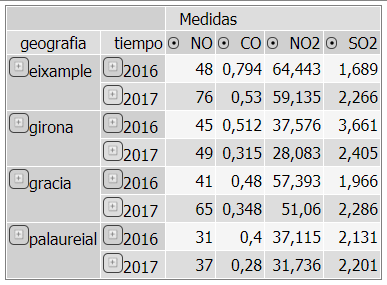
* Entrada de tablas de hechos y dimensiones.
* Unión de flujo de datos según búsqueda de coincidencias y agregación de clave primaria
* Salida de tabla de hechos

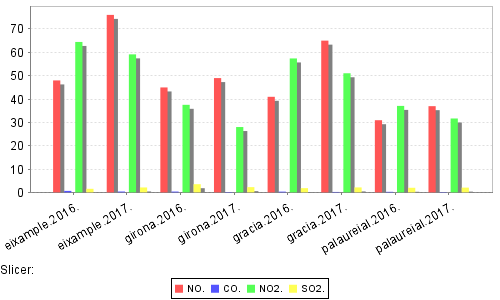
**Esquema de relación de tablas**

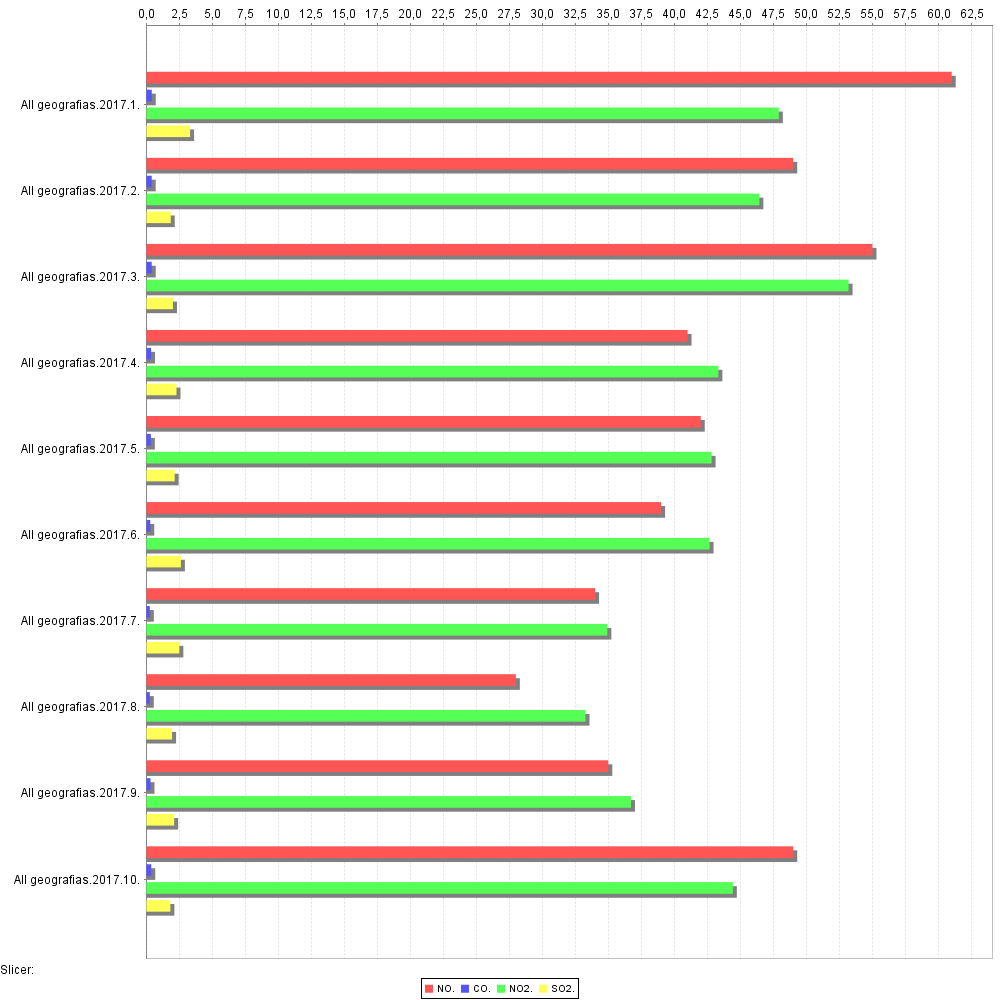
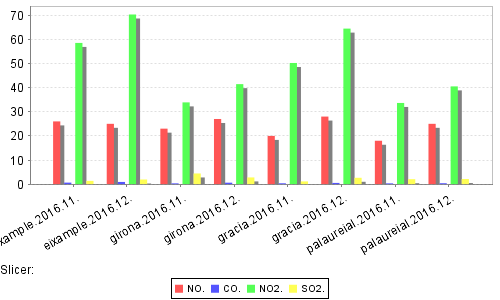


**Cubo OLAP**





****



**ANALISIS DE MODELO PREDICTIVO WEKA**

**Random Forest en base de datos por horas**

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: NewRelation

Instances: 1809

Attributes: 11

fecha

hora

SO2 (µg/m³)

NO (µg/m³)

NO2 (µg/m³)

CO (mg/m³)

barrio

anyo

mes

dia

laborable

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 2.42 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 1722 95.1907 %

Incorrectly Classified Instances 87 4.8093 %

Kappa statistic 0.8691

Mean absolute error 0.1745

Root mean squared error 0.2211

Relative absolute error 44.74 %

Root relative squared error 50.0852 %

Total Number of Instances 1809

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0,819 0,000 1,000 0,819 0,900 0,877 0,999 0,998 falso

1,000 0,181 0,939 1,000 0,968 0,877 0,999 1,000 verdadero

Weighted Avg. 0,952 0,133 0,955 0,952 0,950 0,877 0,999 0,999

=== Confusion Matrix ===

a b <-- classified as

393 87 | a = falso

0 1329 | b = verdadero

**Random forest en base de datos por días**

Ejemplo archivo arff

@relation pordiascalidadaire

@attribute fecha date 'dd-MM-yyyy'

@attribute 'SO2 (µg/m³)' numeric

@attribute 'NO (µg/m³)' numeric

@attribute 'NO2 (µg/m³)' numeric

@attribute 'CO (mg/m³)' numeric

@attribute barrio {eixample,girona,gracia,palaureial}

@attribute anyo {2016,2017}

@attribute mes {1,10,11,12,2,3,4,5,6,7,8,9}

@attribute dia {1,2,3,4,5,6,7}

@attribute laborable {falso,verdadero}

@attribute trimestre {1erTrimestre,'2do Trimestre','3do Trimestre','4o Trimestre'}

@data

01-11-2016,1.00,10.00,27.00,0.30,girona,2016,11,3,verdadero,'4o Trimestre'

01-11-2016,1.00,15.00,47.00,0.20,gracia,2016,11,3,verdadero,'4o Trimestre'

01-11-2016,1.00,43.00,66.00,0.80,eixample,2016,11,3,verdadero,'4o Trimestre'

01-11-2016,2.00,4.00,31.00,0.20,palaureial,2016,11,3,verdadero,'4o Trimestre'

…

**Resultados Modelo Predictivo**

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: pordiascalidadaire

Instances: 1460

Attributes: 11

fecha

SO2 (Âµg/mÂ³)

NO (Âµg/mÂ³)

NO2 (Âµg/mÂ³)

CO (mg/mÂ³)

barrio

anyo

mes

dia

laborable

trimestre

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.48 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.9327

Mean absolute error 5.4382

Root mean squared error 9.6476

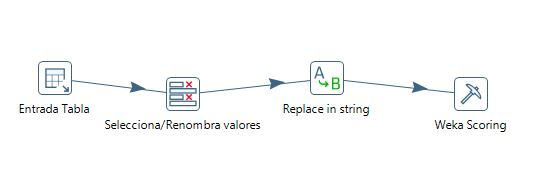
Relative absolute error 31.58 %

Root relative squared error 37.3208 %

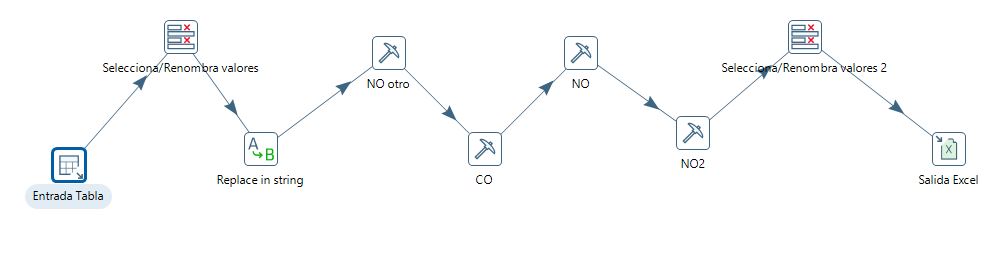
Total Number of Instances 1453

Ignored Class Unknown Instances 7

Transformación para weka scoring

****

**SALIDA ARCHIVO EXCEL**

****

**SALIDA BBDD MONGO DB**

