Análisis de datos de accidentes de tráfico mediante soluciones BigData y Business Intelligence

Marc Alvarez Brotons
Ingeniería Informática

David Isern Alarcón

27/12/2014

Índice

1. Objetivos del proyecto

En la actualidad, las diferentes entidades públicas y privadas disponen de grandes cantidades de datos tanto de hechos actuales como históricos, los cuales es necesario poder utilizarlos para distintas finalidades: seguridad, prevención del fraude, gestión de riesgos, mejora de procesos, acciones comerciales, etc.

Mediante nuevas tecnologías como BigData y Business Intelligence es posible la gestión y análisis de grandes volúmenes de información.

BigData

- Volumen de información
- Velocidad de procesamiento
- Variedad de información

Business Intelligence

- Informes y cuadro de mandos
- Análisis estadístico
- Pronósticos y modelos predictivos

1. Objetivos del proyecto

Los objetivos del actual proyecto de final de carrera son:

- Disponer de un entorno integrado para las tecnologías **BigData** y **Business Intelligence**

- Encontrar **relaciones causa–efecto** sobre los datos climatológicos (Instituto Nacional de Estadística) y sobre accidentes de tráfico (Dirección General de Tráfico)

- Disponer de una herramienta para poder **analizar la información** pasada, presente y futura (predicción)

Índice

2. Enfoque de la solución

Solución propuesta en tres granes pilares tecnológicos:

Fuentes origen
Datos estadísticos referentes a accidentes de tráfico (Dirección General de Tráfico) así como de climatología (Instituto Nacional de Estadística) de las diferentes regiones del territorio español.

Big Data
Elementos que permitirán almacenar y procesar los datos:
Mediante Hadoop Distributed File System se almacenarán todos los datos.
Mediante los algoritmos de Mahout se procesarán los datos para poder detectar posibles patrones.

Business Intelligence
Mediante la herramienta QlikView se analizará mediante gráficos e informes, los datos climatológicos y de accidentes, así como lo datos predictivos.

Índice

3. Solución BigData

El concepto BigData gira alrededor de los siguientes tres ejes, más conocidos como las tres Vs:

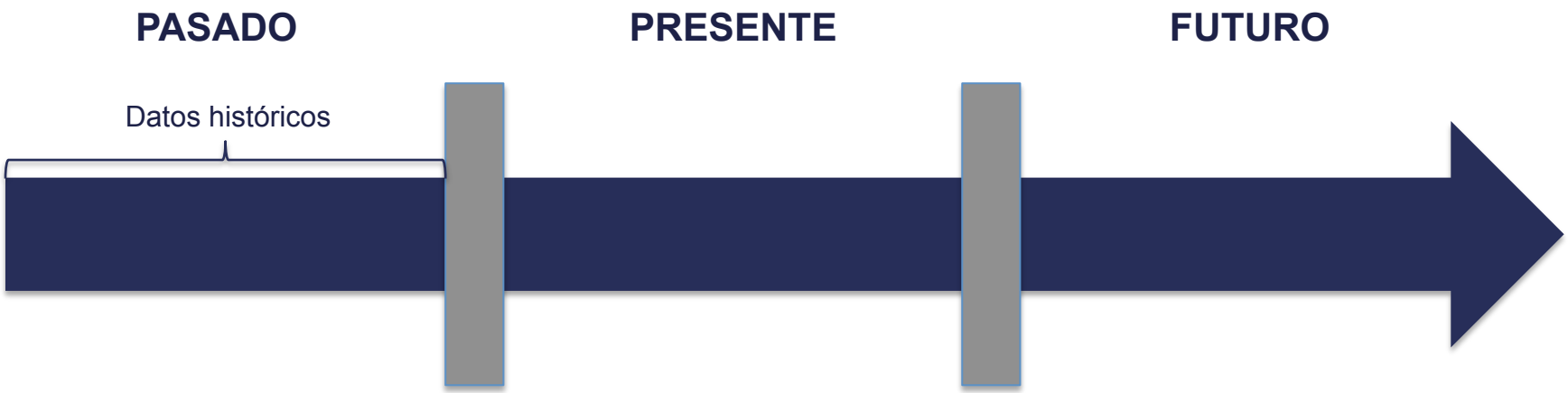
Variedad de la información.
La gran variedad de fuentes de origen de información hace que la información estructurada esté perdiendo su estructura y se esté convirtiendo en cientos de formatos: texto plano, fotografías, audio, video, web, datos GPS, de sensores, posts, documentos, etc.

Velocidad de acceso.
Las necesidades del mercado requieren poder almacenar y acceder a los datos en tiempos cercanos al real-time, con la finalidad de poder realizar un óptima toma de decisiones.

Volumen de la información.
En los últimos años el tamaño de los datos has estado aumentando a un ritmo creciente. Este hecho requiere poder almacenar y gestionar grandes volúmenes de información de forma ágil.

2. Enfoque de la solución

La solución propuesta pretende optimizar la eficiencia operativa por medio del análisis de datos, mediante la utilización de técnicas y modelos analíticos que provienen de las ciencias de la computación y más concretamente de la Inteligencia Artificial.



Aprender

- ✓ Aprender de datos históricos
- ✓ Identificar patrones
- ✓ Validar resultados

Predecir

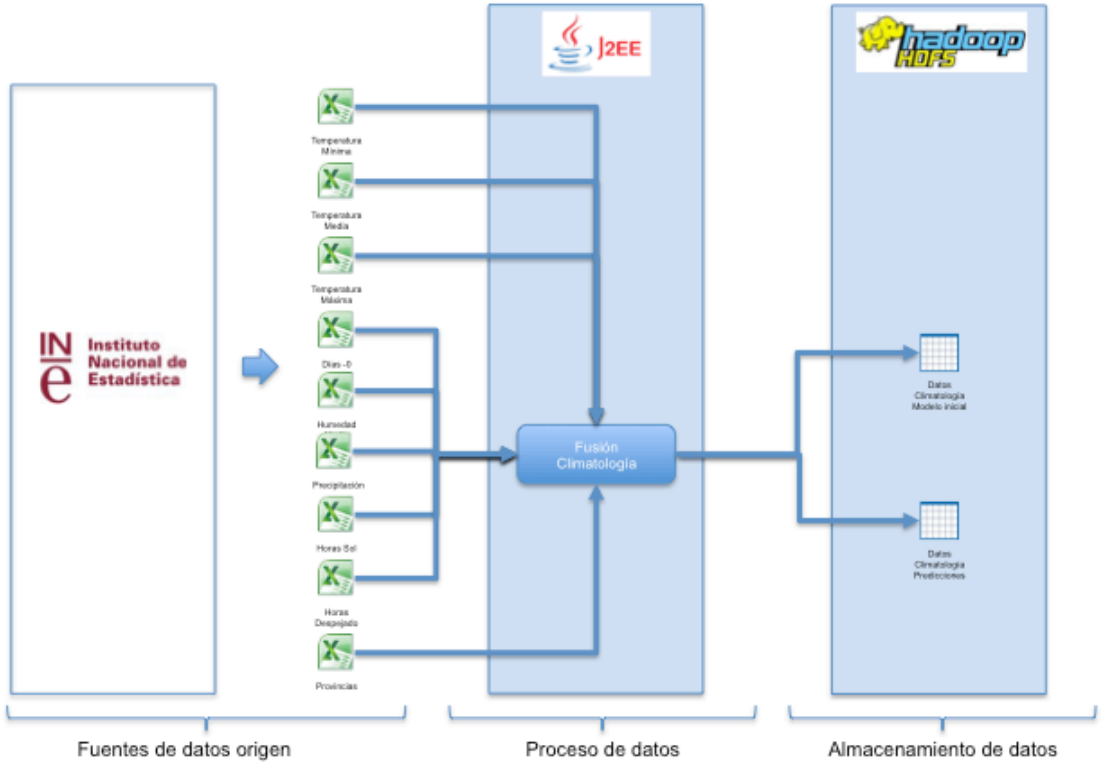
- ✓ Anticiparse a un evento en base a una probabilidad

3. Solución BigData

Flujo de datos en la incorporación de los datos procedentes del INE referentes a la climatología con la finalidad de ser incorporados en el repositorio HDFS con la finalidad de ser utilizados por los algoritmos de predicción

Variables:

- Identificador de provincia
- Nombre de la provincia
- Año correspondiente a los datos del fichero
- Temperatura máxima
- Temperatura media
- Temperatura mínima
- Número de días con temperatura igual o menor a cero grados
- Precipitación total
- Humedad media
- Número de días con el cielo despejado
- Horas de sol
- Clase que clasifica en riesgo alto o bajo de accidente a causa del clima



Instituto Nacional de Estadística: http://www.ine.es/inebmenu/mnu_entornofis.htm

Dirección General de Tráfico: http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores

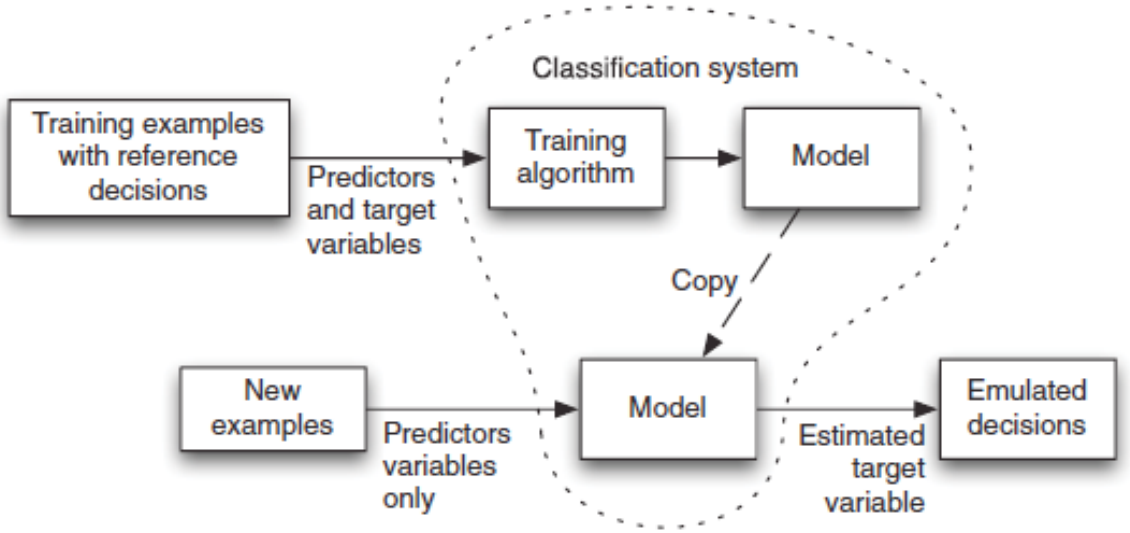
3. Solución BigData

Logistic Regresion

Algoritmo destinado a la predicción de la probabilidad de la ocurrencia de un evento concreto, mediante la utilización de variables predictivas

El algoritmo ofrece dos posibilidades:

- Determinar numéricamente la relación entre las variables predictivas y la variable a predecir, así como el nivel de confusión entre dichas variables.
- Clasificar datos según la probabilidad de pertenecer a la variable objetivo en base a sus propias variable predictivas.



3. Solución BigData

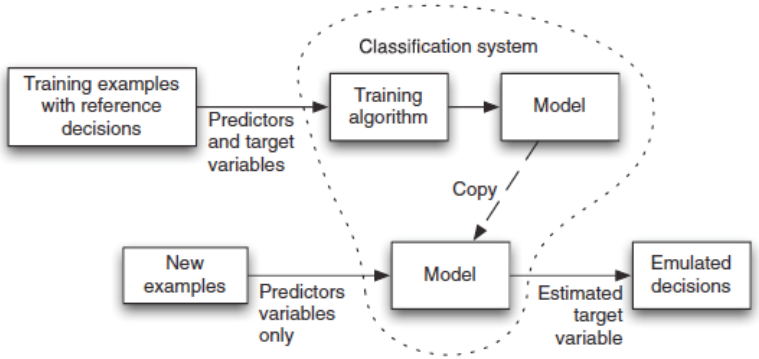
Los datos a utilizar para el aprendizaje y la predicción será información procedente del Instituto Nacional de Estadística referente a la climatología de las provincias de España:

- Tmedia: Temperatura media
- Tmin: Temperatura mínima
- Tmax: Temperatura máxima
- Hmedia: Humedad media
- Ptotal: Precipitación total
- Dias-0: Número de días con temperaturas inferiores a cero grados
- Ddes: Días con el cielo destapado
- Hsol: Horas de sol

Adicionalmente existirá una variable marcada como objetivo (target) y se le asignará asignarle un valor coherente para los datos iniciales de aprendizaje (nivel de posibles accidentes: alto o bajo).

No todas las provincias estarán en este subset de datos inicial ya que algunas pueden distorsionar la causa de accidentes debido que son ajenas a la climatología, como el alto volumen de vehículos, ser un lugar de paso, zona de alta afluencia turística, etc.

3. Solución BigData



Ejecución del algoritmo:

1. Aprendizaje

   $mahout trainlogistic --input /home/cloudera/ia/clima_train --output /home/cloudera/ia/model --target Clase --categories 2 --predictors Provincia Tmedia Tmin Hmedia Hdes Ptotal Dias-0 Hsol --types numeric --features 114 --passes 1000 --rate 80

2. Validación modelo

   $mahout runlogistic --input /home/cloudera/ia/clima_train --model /home/cloudera/ia/model --auc –confusion

3. Ejecución/validación predicción

   $mahout runlogistic --input /home/cloudera/ia/clima_run --model /home/cloudera/ia/model --auc –confusion

4. Obtención predicción

   $mahout runlogistic --input /home/cloudera/ia/clima_run --model /home/cloudera/ia/model --scores

3. Solución BigData

Resultados obtenidos:

Modelo 1

- Variables predictivas utilizadas: *Tmedia, Tmin, Tmax, Hmedia, Ptotal, Dias-0, Ddes, Hsol*
- Resultados trainlogistic:
  - *AUC: 0,61*
  - *Matriz de confusión: [[29.0, 6.0], [6.0, 4.0]]*
- Resultados runlogistic:
  - *AUC: 0,50*
  - *Matriz de confusión: [[225.0, 49.0], [39.0, 29.0]]*

- Valoración: Hay variables que distorsionan el modelo y generan una predicción con una probabilidad aleatoria.

3. Solución BigData

Resultados obtenidos:

Modelo 2

- Variables predictivas utilizadas: *Tmedia, Tmin, Tmax*
- Resultados trainlogistic:
  - *AUC: 0,70*
  - *Matriz de confusión: [[30.0, 5.0], [5.0, 5.0]]*
- Resultados runlogistic:
  - *AUC: 0,63*
  - *Matriz de confusión: [[141.0, 24.0], [123.0, 54.0]]*

- Valoración: La selección de las variables predictivas han dado lugar a una mejora respecto al modelo anterior, pero los resultados siguen siendo bajos, todo y que se alejan de las predicciones aleatorias.

3. Solución BigData

Resultados obtenidos:

Modelo 3

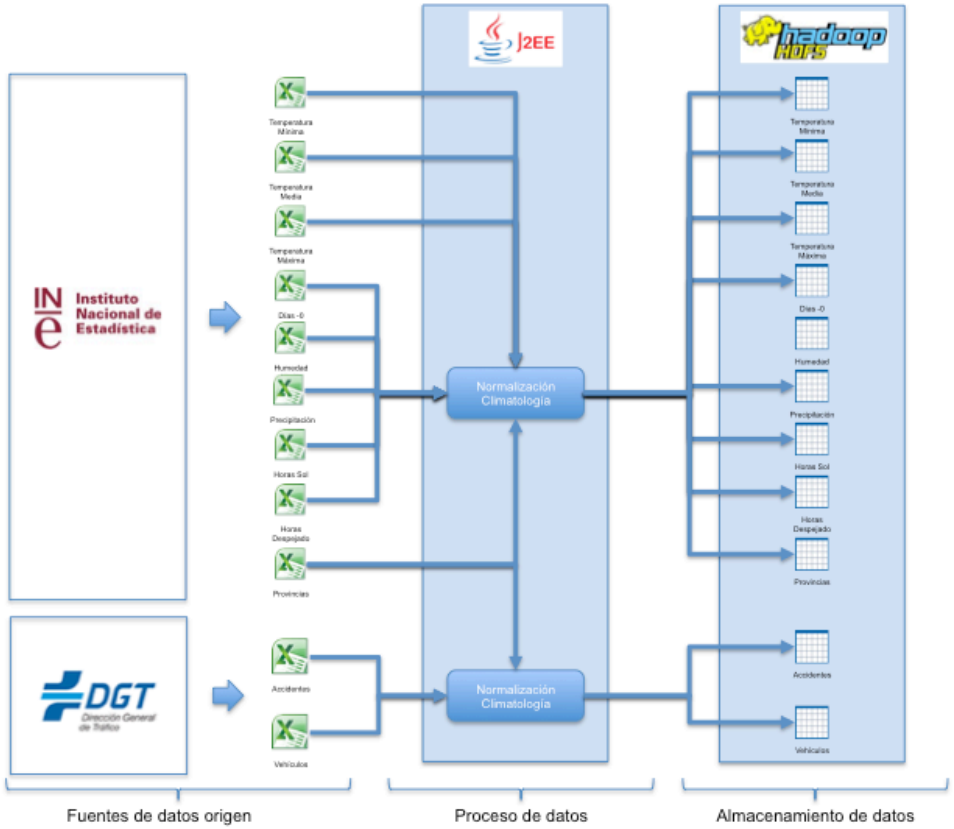
- Variables predictivas utilizadas: *Tmedia, Tmin, Tmax, Hmedia, Ptotal*
- Resultados trainlogistic:
  - *AUC: 0,82*
  - *Matriz de confusión: [[32.0, 3.0], [3.0, 7.0]]*
- Resultados runlogistic:
  - *AUC: 0,65*
  - *Matriz de confusión: [[248.0, 57.0], [16.0, 21.0]]*

- Valoración: Utilizado cinco de las variables disponibles, relacionadas con la temperatura y la humedad o precipitación, los resultados han mejorado tanto a nivel de entrenamiento como de predicción. Dichas variables que tienen un alto grado de influencia en el modelo de aprendizaje.

Índice

4. Solución Business Intelligence

Los datos procedentes del Instituto Nacional de Estadística y de la Dirección General de Tráfico son incorporados al repositorio HDFS, para ser explotados mediante la herramienta de Business Intelligence QlikView.



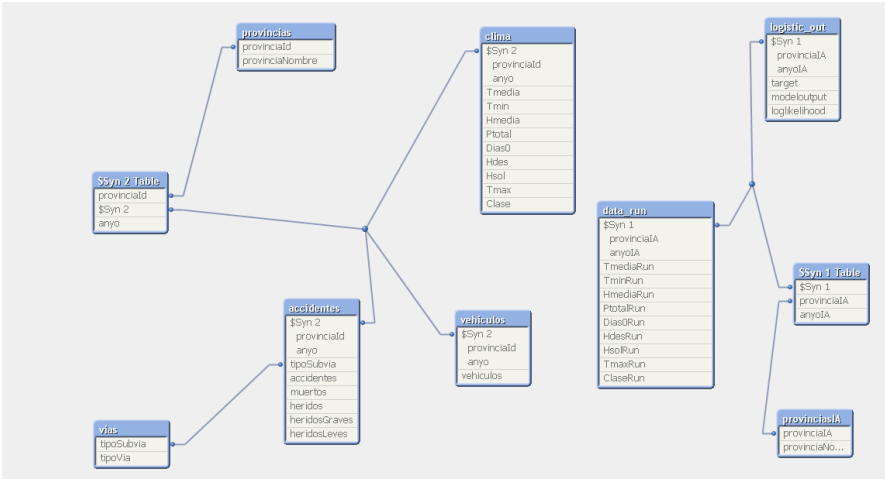
Para la incorporación de los datos se utiliza un procesos desarrollado mediante el lenguaje de programación Java

4. Solución Business Intelligence

El modelo está constituido por la siguientes dimensiones y hechos:

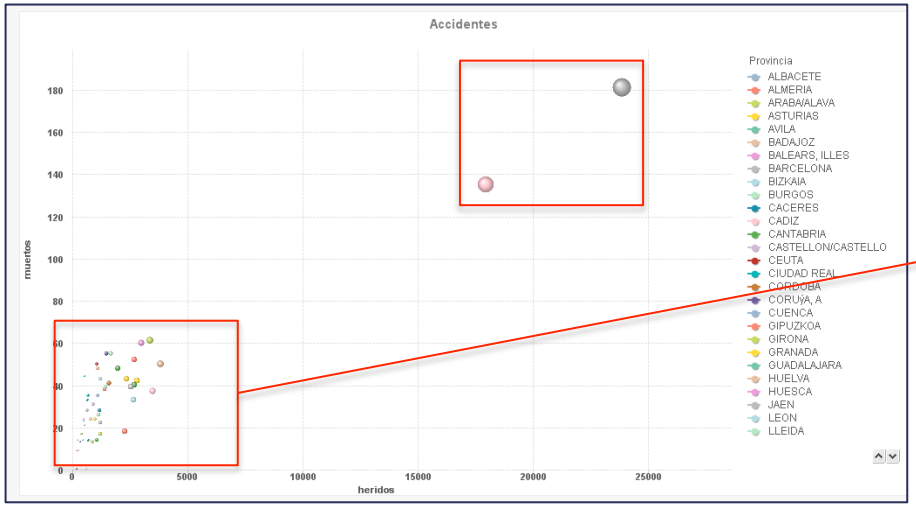
- **Dimensión temporal.** Reflejará los distintos periodos de tiempo en los que se darán lugar los hechos climatológicos y de tráfico
- **Dimensión territorial.** Reflejará los distintos espacios geográficos en los que se dará lugar los hechos climatológicos y de tráfico.
- **Dimensión vía.** Reflejará los distintos tipos de vía en los que se dará lugar los hechos de tráfico. Dicha dimensión estará constituida por los siguientes atributos:

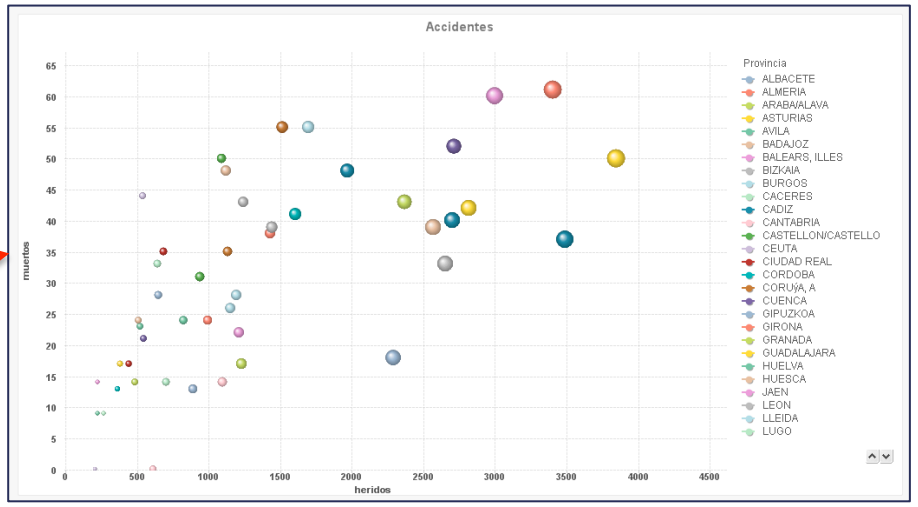
- **Hechos de tráfico.** Datos disponibles a nivel anual, por provincia y tipo de subvía: accidentes, muertos, heridos, heridos graves, heridos leves y parque de vehículos.
- **Hechos climatológicos.** Datos disponibles a nivel anual y por provincia: temperatura media, temperatura mínima, temperatura máxima, humedad media, precipitación total, número de días con temperatura inferior a 0 grados centígrados

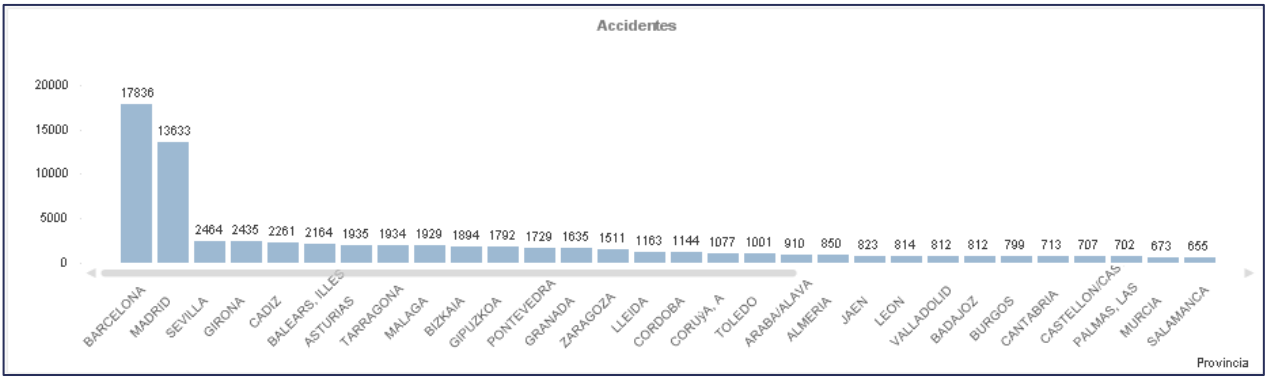


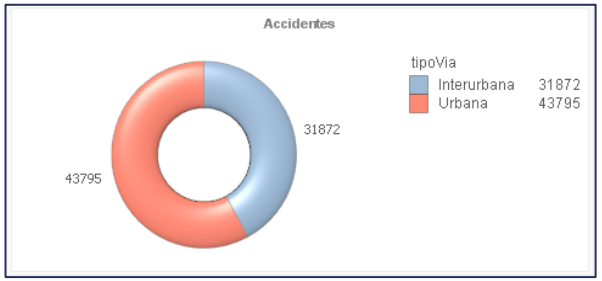
4. Solución Business Intelligence

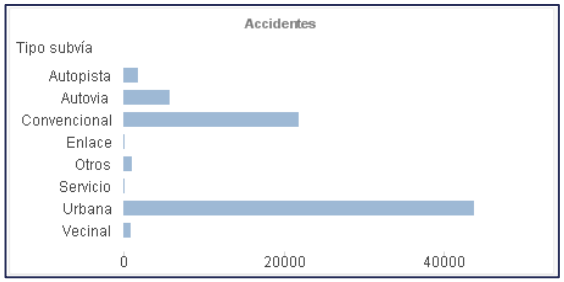
Accidentes de tráfico





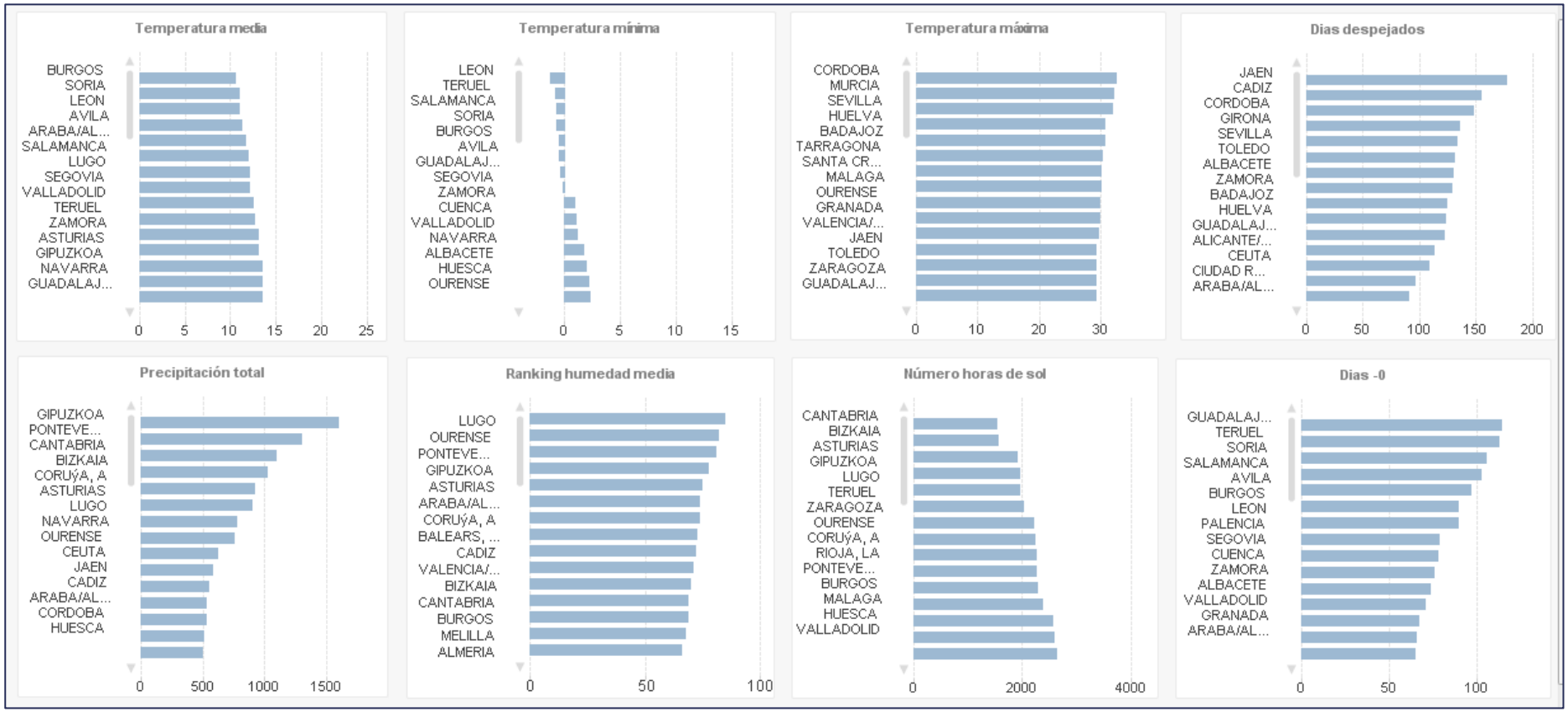






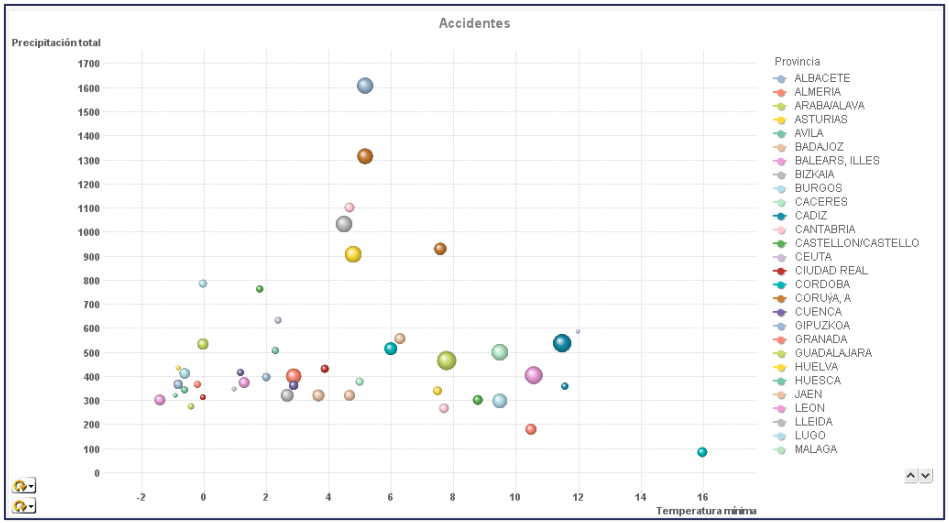
4. Solución Business Intelligence

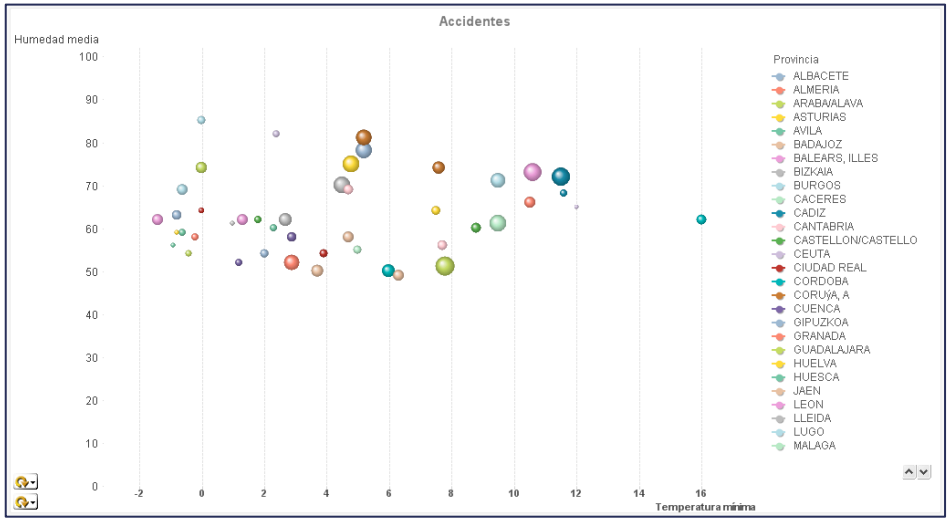
Datos climatológicos utilizados para la analítica y la selección de las provincias candidatas para el aprendizaje.

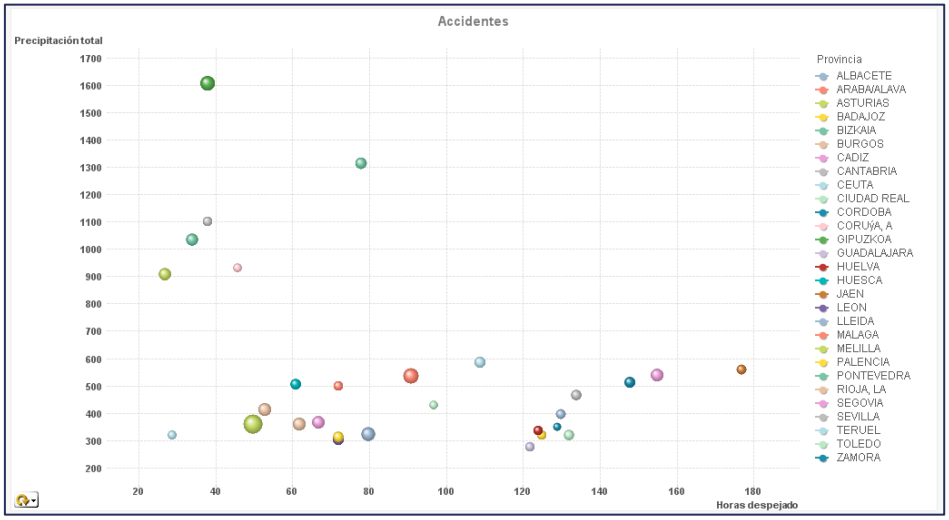


4. Solución Business Intelligence

Datos cruzados: clima y accidentes

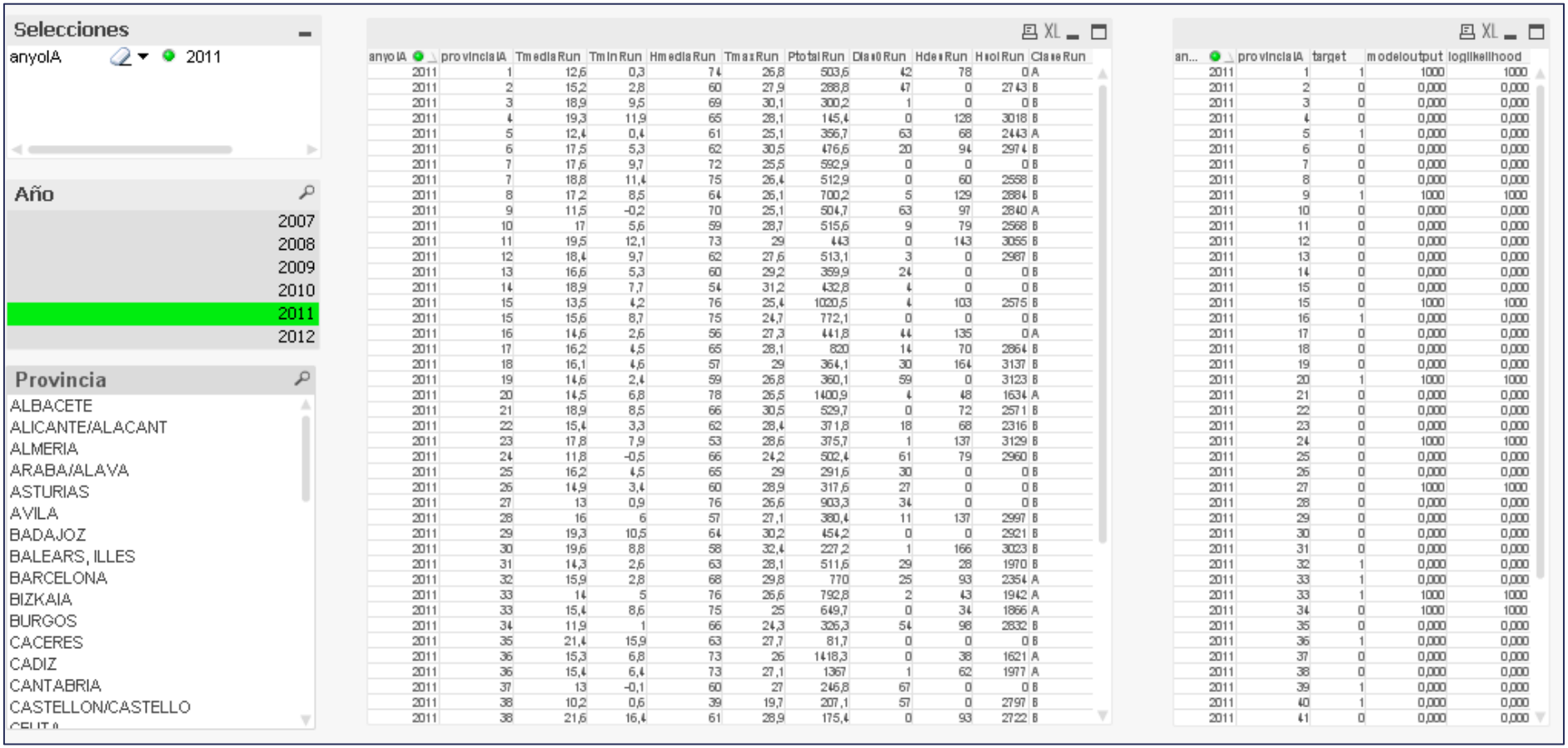






4. Solución Business Intelligence

Datos predictivos



Índice

5. Conclusiones

En este proyecto se dispone de dos tecnologías que permiten realizar un ejercicio de relación causa efecto:

- Mediante los algoritmos predictivos e información histórica, se crea un modelo predictivo de la afectación de accidentes que podría tener una provincia en base a la climatología.

- Mediante una herramienta de análisis se exploran los datos de accidentes y para poder detectar patrones de comportamiento, así como analizar los resultados de los algoritmos predictivos.

5. Conclusiones

En base a la pruebas llevadas a cabo con los datos disponibles se ha llegado a las siguientes conclusiones:

- Para llegar a un modelo predictivo con un nivel de fiabilidad alto de relación causa-efecto, no todas la variables predictivas pueden ayudar a tal fin, sino que algunas de ellas pueden llevar a resultados erróneos. Para ello, se debe trabajar detenidamente en su elección.
- El actual modelo predictivo, al disponer de los datos de accidentes y climatológicos a nivel de anual y provincia, sólo permite realizar predicciones muy genéricas. En caso de disponer de una información más detallada (nivel diario), las predicciones podrían ser mucho más ajustadas.
- Este modelo predictivo abre las puertas a la incorporación de nuevas variables predictivas que permitan analizar y predecir la probabilidad de accidentes en base a ratios económicos, antigüedad del parque de vehículos, niveles de alcoholemia, así como el estado de las vías, etc.


UOC
Universitat Oberta
de Catalunya
www.uoc.edu