



Universitat Oberta
de Catalunya

www.uoc.edu

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
2. Enfoque de la solución
3. Solución BigData
4. Solución Business Intelligence
5. Conclusiones

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:

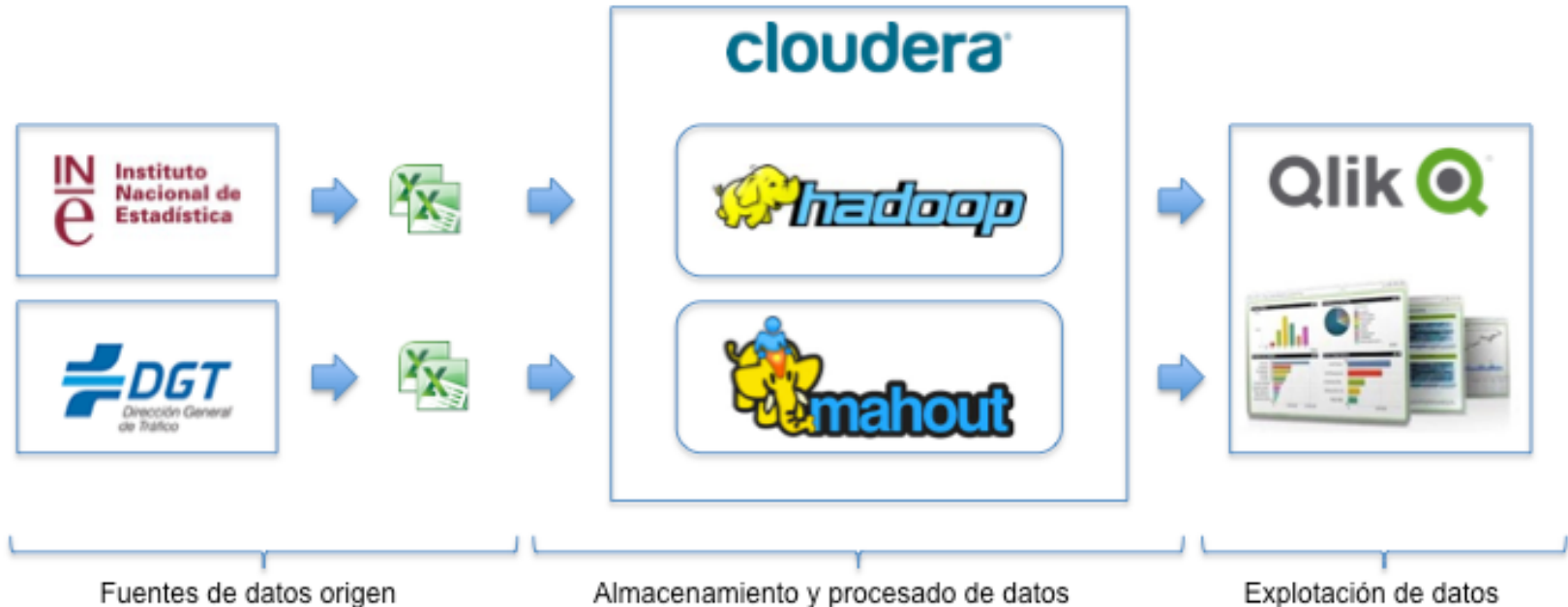
- Disponer de un entorno integrado para las tecnologías **BigData** y **Business Intelligence**
- Encontrar **relaciones causa–efecto** sobre los datos climatológicos (Instituto Nacional de Estadística) y sobre accidentes de tráfico (Dirección General de Tráfico)
- Disponer de una herramienta para poder **analizar la información** pasada, presente y futura (predicción)

Índice

1. Objetivos del proyecto
2. Enfoque de la solución
3. Solución BigData
4. Solución Business Intelligence
5. Conclusiones

2. Enfoque de la solución

Solución propuesta en tres grandes 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 los datos predictivos.

Índice

1. Objetivos del proyecto
2. Enfoque de la solución
- 3. Solución BigData**
4. Solución Business Intelligence
5. Conclusiones

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.

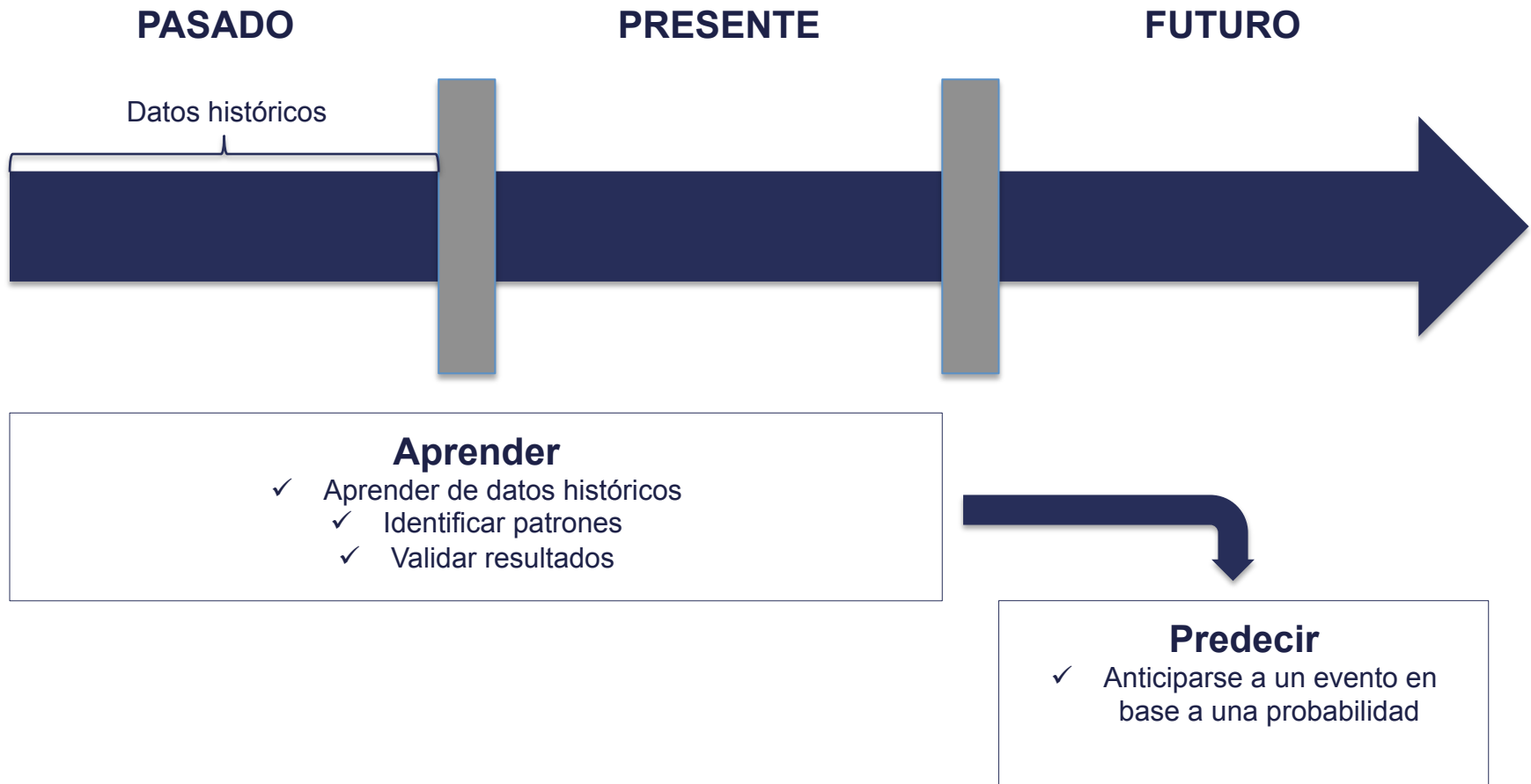
Las necesidades del mercado requieren poder almacenar y acceder a los datos en tiempos cercanos al real-time, con la finalidad de poder realizar un óptima toma de decisiones.

Volumen de la información.

En los últimos años el tamaño de los datos ha 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.

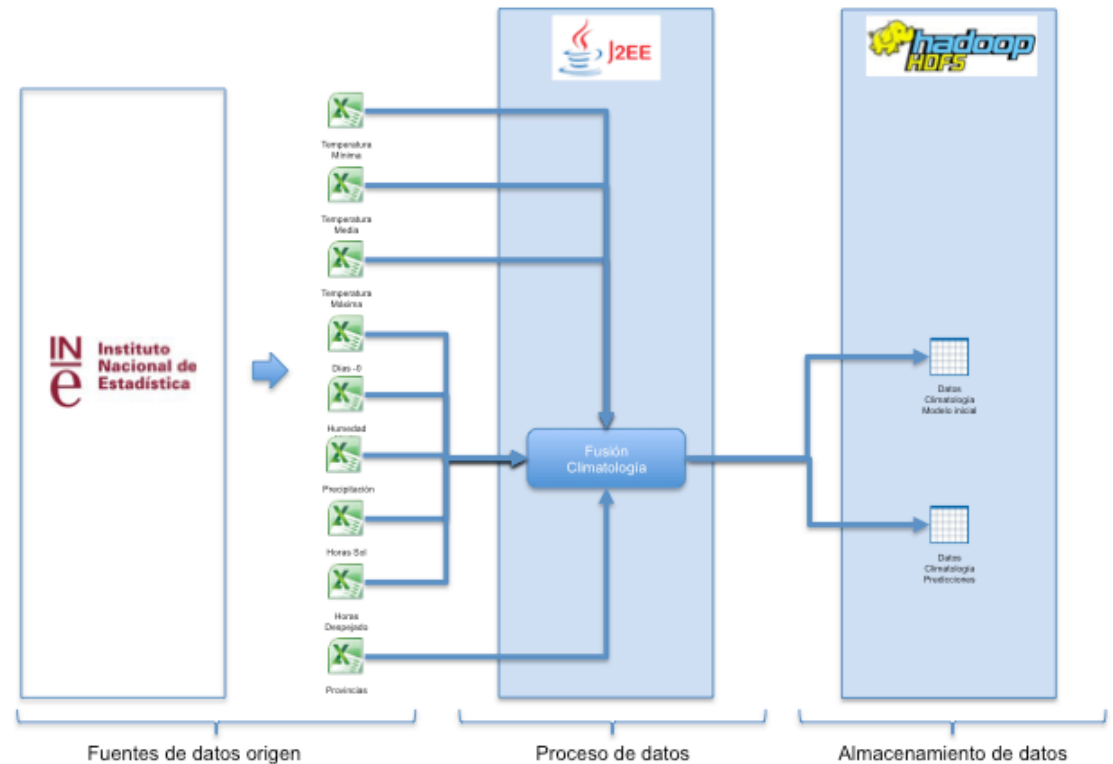


3. Solución BigData

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>

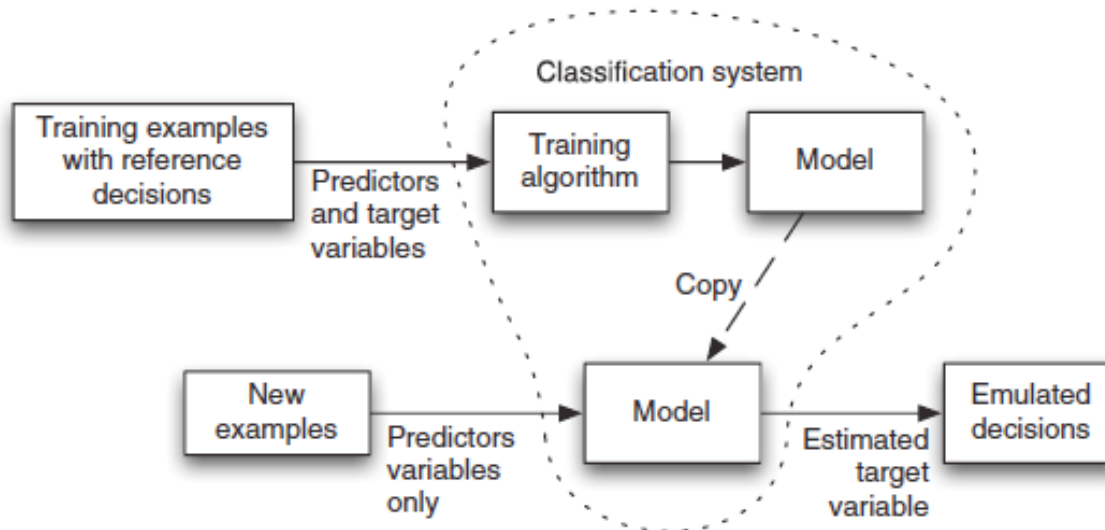
3. Solución BigData

Logistic Regression

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:

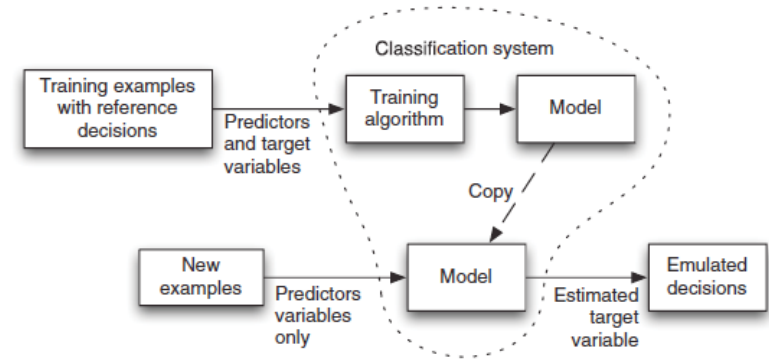
- Tmedia: Temperatura media
- Tmin: Temperatura mínima
- Tmax: Temperatura máxima
- Hmedia: Humedad media
- Ptotal: Precipitación total
- Dias-0: Número de días con temperaturas inferiores a cero grados
- Ddes: Días con el cielo destapado
- Hsol: Horas de sol

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: T_{media} , T_{min} , T_{max}
- 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

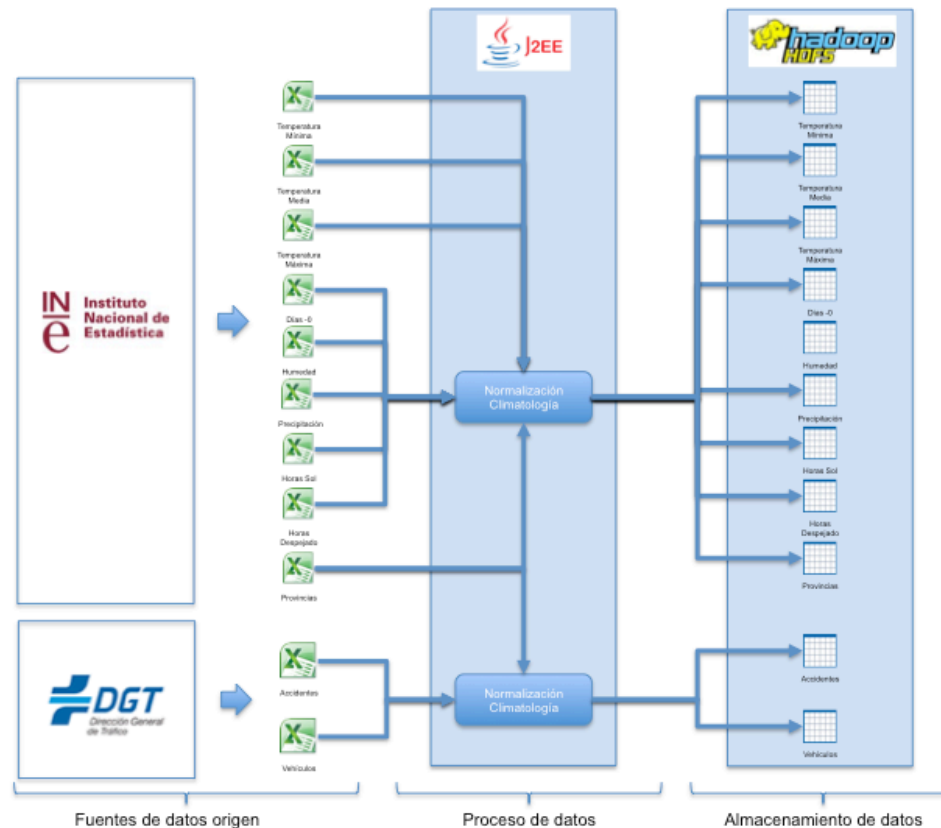
- Variables predictivas utilizadas: *Tmedia*, *Tmin*, *Tmax*, *Hmedia*, *Ptotal*
- Resultados trainlogistic:
 - *AUC: 0,82*
 - *Matriz de confusión: [[32.0, 3.0], [3.0, 7.0]]*
- Resultados runlogistic:
 - *AUC: 0,65*
 - *Matriz de confusión: [[248.0, 57.0], [16.0, 21.0]]*
- Valoración: Utilizado cinco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

1. Objetivos del proyecto
2. Enfoque de la solución
3. Solución BigData
- 4. Solución Business Intelligence**
5. Conclusiones

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.

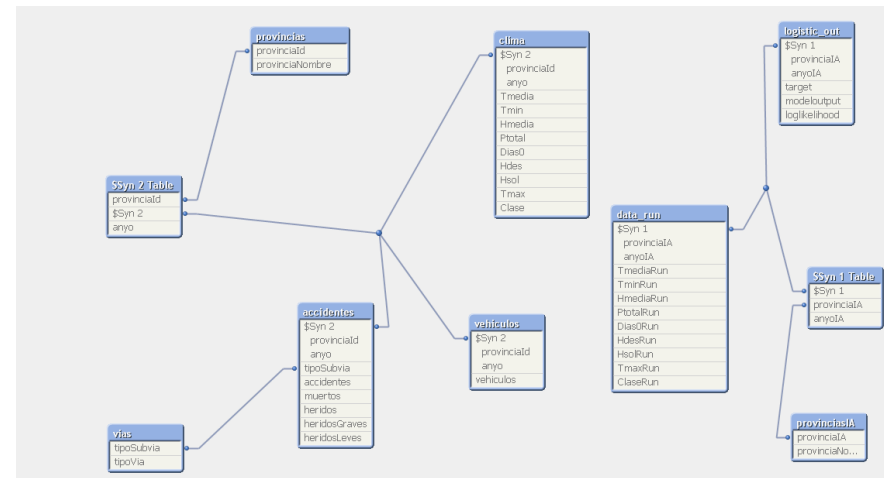


Para la incorporación de los datos se utiliza un proceso desarrollado mediante el lenguaje de programación Java

4. Solución Business Intelligence

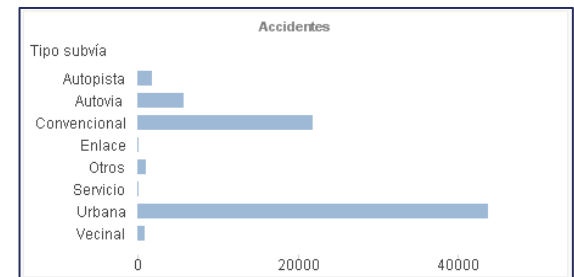
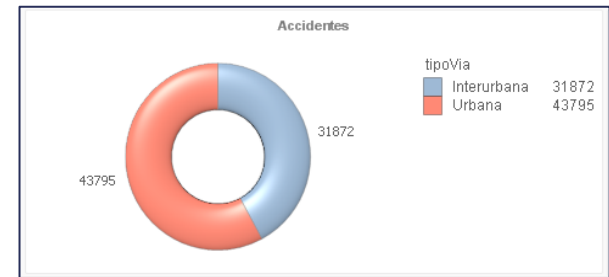
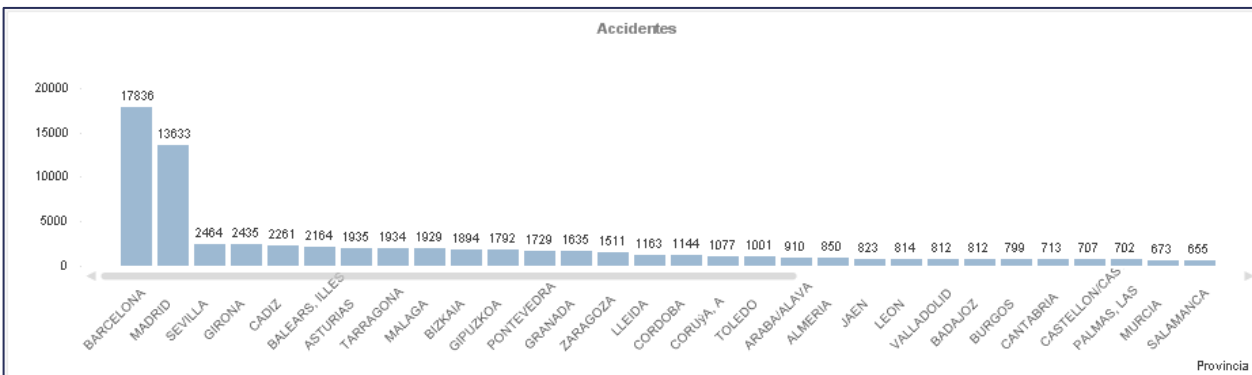
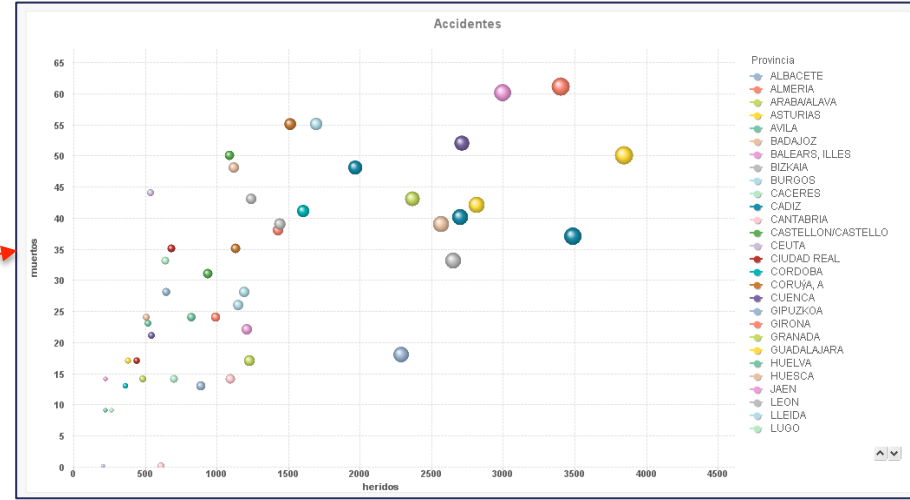
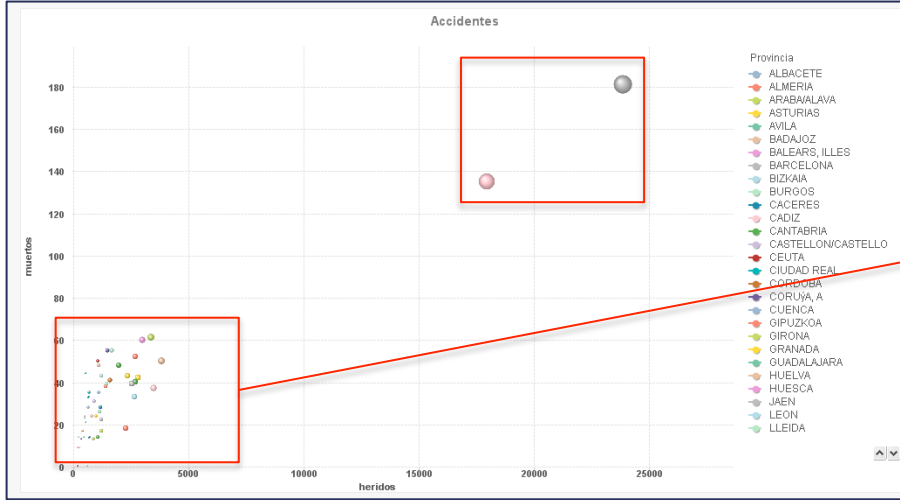
El modelo está constituido por la siguientes dimensiones y hechos:

- **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



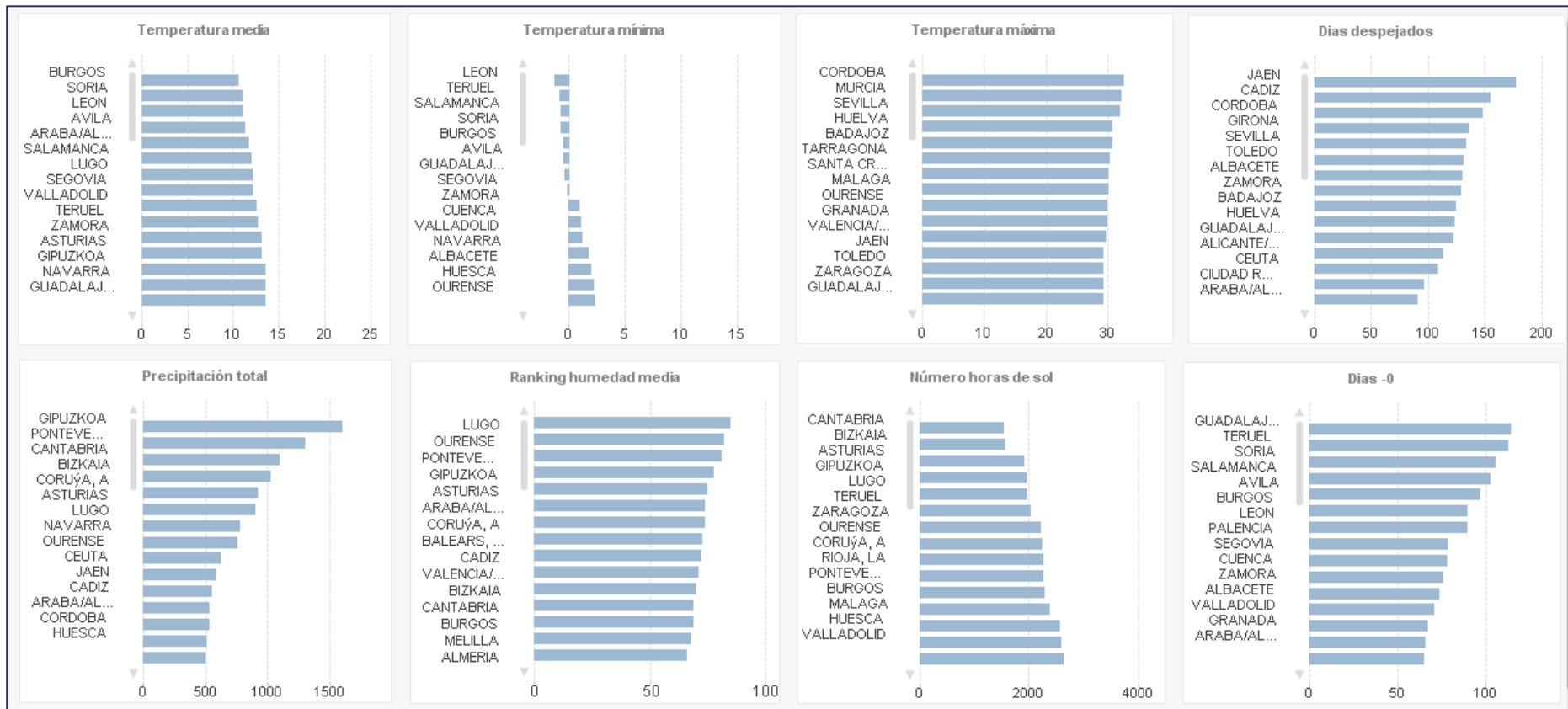
4. Solución Business Intelligence

Accidentes de tráfico



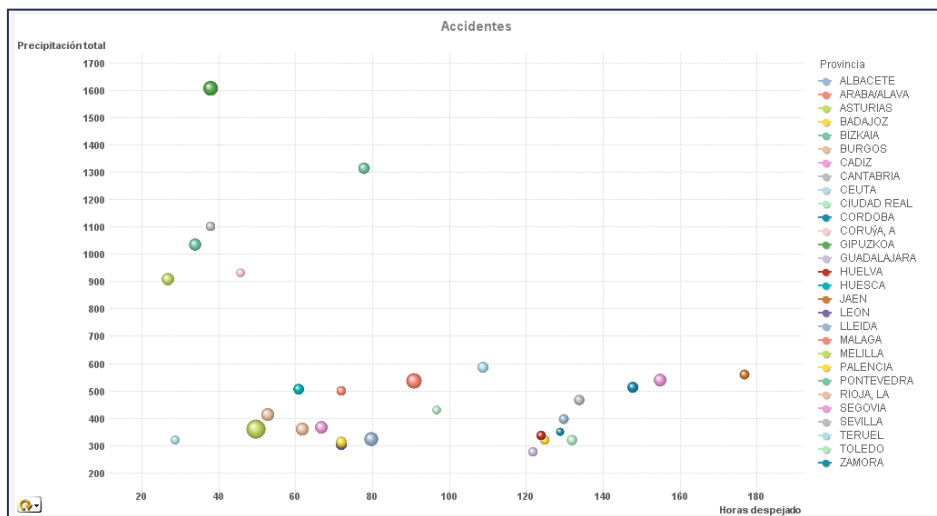
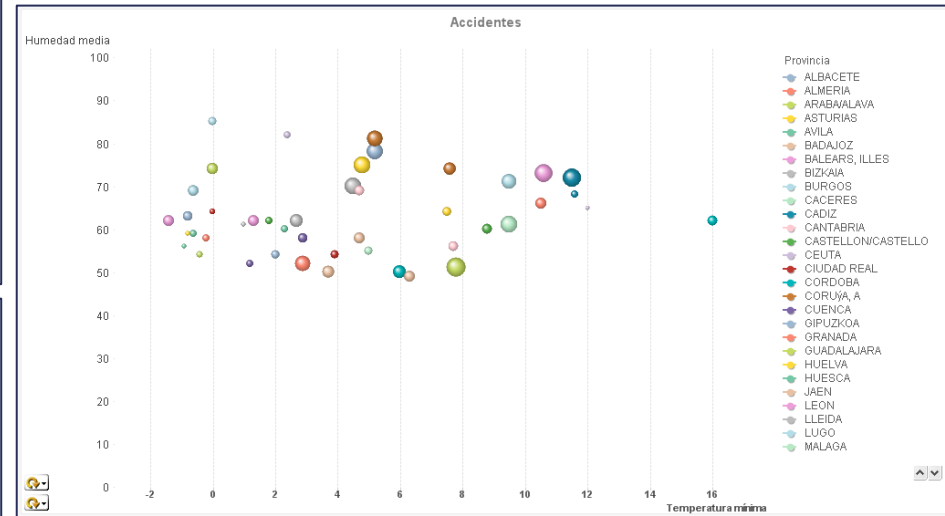
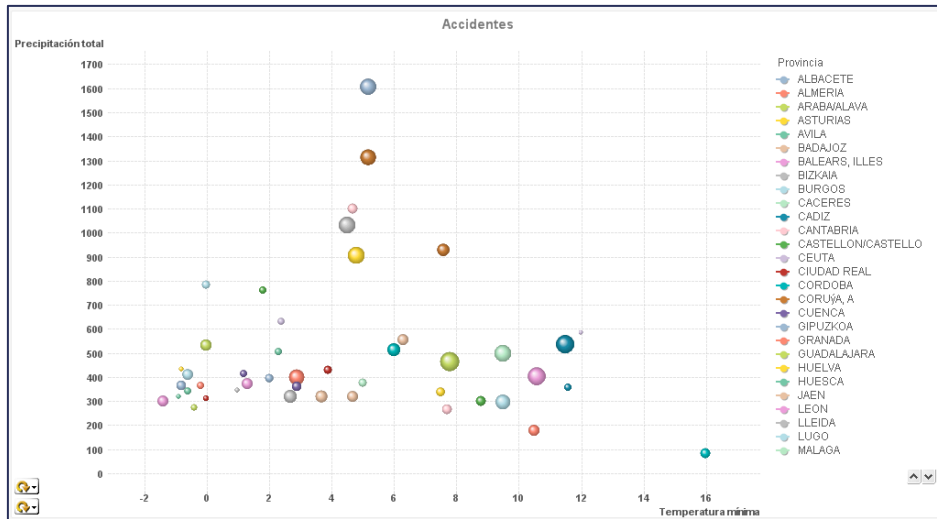
4. Solución Business Intelligence

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

Selecciones

anyoIA 2011

Año

- 2007
- 2008
- 2009
- 2010
- 2011
- 2012

Provincia

- ALBACETE
- ALICANTE/ALACANT
- ALMERIA
- ARABA/ALAVA
- ASTURIAS
- AVILA
- BADAJOS
- BALEARS, ILLES
- BARCELONA
- BIZKAIA
- BURGOS
- CACERES
- CADIZ
- CANTABRIA
- CASTELLON/CASTELLO
- CELTA

anyoIA	provinciaIA	TmediaRun	TminRun	HmediaRun	TmaxRun	PtotalRun	Clas0Run	HdesRun	HsolRun	ClaseRun
2011	1	12,6	0,3	74	26,8	503,6	42	78	0	A
2011	2	15,2	2,8	60	27,9	288,8	47	0	2743	B
2011	3	18,9	9,5	69	30,1	300,2	1	0	0	B
2011	4	19,3	11,9	65	28,1	145,4	0	128	3018	B
2011	5	12,4	0,4	61	25,1	356,7	63	68	2443	A
2011	6	17,5	5,3	62	30,5	476,6	20	94	2974	B
2011	7	17,6	9,7	72	25,5	562,9	0	0	0	B
2011	7	18,8	11,4	75	26,4	512,9	0	60	2558	B
2011	8	17,2	8,5	64	26,1	700,2	5	129	2884	B
2011	9	11,5	-0,2	70	25,1	504,7	63	97	2840	A
2011	10	17	5,6	59	28,7	515,6	9	79	2568	B
2011	11	19,5	12,1	73	29	443	0	143	3055	B
2011	12	18,4	9,7	62	27,6	513,1	3	0	2967	B
2011	13	16,6	5,3	60	29,2	369,9	24	0	0	B
2011	14	18,9	7,7	54	31,2	432,8	4	0	0	B
2011	15	13,5	4,2	76	25,4	1020,5	4	103	2575	B
2011	15	15,6	8,7	75	24,7	772,1	0	0	0	B
2011	16	14,6	2,6	55	27,3	441,8	44	135	0	A
2011	17	16,2	4,5	65	28,1	630	14	70	2864	B
2011	18	16,1	4,6	57	29	364,1	30	164	3137	B
2011	19	14,6	2,4	59	26,8	360,1	59	0	3123	B
2011	20	14,5	6,8	78	26,5	1400,9	4	48	1634	A
2011	21	18,9	8,5	66	30,5	529,7	0	72	2571	B
2011	22	15,4	3,3	62	28,4	371,8	18	68	2316	B
2011	23	17,8	7,9	53	28,6	375,7	1	137	3129	B
2011	24	11,8	-0,5	66	24,2	502,4	61	79	2960	B
2011	25	16,2	4,5	65	29	291,6	30	0	0	B
2011	26	14,9	3,4	60	28,9	317,6	27	0	0	B
2011	27	13	0,9	76	26,6	903,3	34	0	0	B
2011	28	16	6	57	27,1	380,4	11	137	2997	B
2011	29	19,3	10,5	64	30,2	454,2	0	0	2521	B
2011	30	19,6	6,8	58	32,4	227,2	1	166	3023	B
2011	31	14,3	2,6	63	28,1	511,6	29	28	1970	B
2011	32	15,9	2,8	68	29,8	770	25	93	2354	A
2011	33	14	5	76	26,6	792,8	2	43	1942	A
2011	33	15,4	8,6	75	25	649,7	0	34	1866	A
2011	34	11,9	1	66	24,3	326,3	54	98	2832	B
2011	35	21,4	15,9	63	27,7	81,7	0	0	0	B
2011	36	15,3	6,8	73	26	1418,3	0	38	1621	A
2011	36	15,4	6,4	73	27,1	1367	1	62	1977	A
2011	37	13	-0,1	60	27	246,8	67	0	0	B
2011	38	10,2	0,6	39	19,7	207,1	57	0	2797	B
2011	38	21,6	16,4	61	28,9	175,4	0	93	2722	B

an...	provinciaIA	target	modeloutput	loglikelihood
2011	1	1	1000	1000
2011	2	0	0,000	0,000
2011	3	0	0,000	0,000
2011	4	0	0,000	0,000
2011	5	1	0,000	0,000
2011	6	0	0,000	0,000
2011	7	0	0,000	0,000
2011	8	0	0,000	0,000
2011	9	1	1000	1000
2011	10	0	0,000	0,000
2011	11	0	0,000	0,000
2011	12	0	0,000	0,000
2011	13	0	0,000	0,000
2011	14	0	0,000	0,000
2011	15	0	0,000	0,000
2011	15	0	1000	1000
2011	16	1	0,000	0,000
2011	17	0	0,000	0,000
2011	18	0	0,000	0,000
2011	19	0	0,000	0,000
2011	20	1	1000	1000
2011	21	0	0,000	0,000
2011	22	0	0,000	0,000
2011	23	0	0,000	0,000
2011	24	0	1000	1000
2011	25	0	0,000	0,000
2011	26	0	0,000	0,000
2011	27	0	1000	1000
2011	28	0	0,000	0,000
2011	29	0	0,000	0,000
2011	30	0	0,000	0,000
2011	31	0	0,000	0,000
2011	32	1	0,000	0,000
2011	33	1	0,000	0,000
2011	33	1	1000	1000
2011	34	0	1000	1000
2011	35	0	0,000	0,000
2011	36	1	0,000	0,000
2011	37	0	0,000	0,000
2011	38	0	0,000	0,000
2011	39	1	0,000	0,000
2011	40	1	0,000	0,000
2011	41	0	0,000	0,000

Índice

1. Objetivos del proyecto
2. Enfoque de la solución
3. Solución BigData
4. Solución Business Intelligence
- 5. Conclusiones**

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.



**Universitat Oberta
de Catalunya**

www.uoc.edu