



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Escola Tècnica
Superior d'Enginyeria
Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Estudio sobre patrones de accidentes en la ciudad de València

Trabajo Fin de Máster
Máster Universitario en Gestión de la Información

Autor: Joaquín Montesinos Muñoz
Tutor: César Ferri Ramírez
2017-2018

Resumen

Los accidentes de tráfico son una de las causas más frecuentes de lesión y muerte en la población adulta en el mundo. Existen factores que pueden afectar a las condiciones de la conducción como puede ser el tráfico intenso o la meteorología adversa, y de una manera indirecta provocar que aumenten los accidentes de tráfico.

Este proyecto se elabora con el fin de detectar patrones y situaciones propensas a accidentes de tráfico y de este modo poder reducir el número de muertes por siniestralidad.

Para la realización del proyecto se llevará a cabo la implementación de un sistema de análisis, predicción y visualización que utiliza tecnologías de procesado de datos y tecnologías de visualización web.

Los diferentes análisis y visualizaciones podrán ser utilizados para encontrar patrones de accidentes a nivel nacional, provincial y local, estudiando en estos la relación entre accidentes, clima y tráfico recogidos a través de los datos históricos de diferentes portales de datos abiertos.

Además, se generará una aplicación que sea capaz de estimar el número de accidentes que sucederá en las calles de Valencia y en sus carreteras periféricas. Para dicha tarea se utiliza un modelo de aprendizaje automático basado en regresión lineal que se nutre de datos históricos y datos a tiempo real, todo esto integrado mediante una herramienta de visualización interactiva que permite la simulación de diferentes escenarios.

Consideramos que la herramienta resultado del Trabajo de Fin de Máster puede servir para conocer qué factores son más propensos a producir problemas en el tráfico, concienciar a los conductores, y de este modo reducir el número de accidentes y de víctimas mediante técnicas de prevención.

Palabras clave: Accidentes, Tráfico, Meteorología, Datos abiertos, Aprendizaje automático, Predicción, Visualización, Análisis de datos, Dashboard.

Abstract

Traffic accidents are one of the most frequent causes of lesions and death in the adult population worldwide. There are some factors that can affect to conduction conditions such as intense traffic or adverse meteorology, and in an indirect way they can increase the number of traffic accidents.

This project has been developed with the purpose of detecting patterns and situations liable to traffic accidents and, in this way, decrease the number of deaths.

To achieve the main purpose of the project, a system based on analysis, prediction and visualization is going to be implemented using data processing and web visualization technologies.

The different analysis and visualizations could be used to find patterns of accidents at a national, provincial or local level, by studying the relationship between accidents, weather and traffic collected thanks to the historical data available in many open data portals.

Moreover, it is going to be developed an application able to predict the number of accidents that are going to happen on the streets of Valencia and on its peripheral roads. For this task, it is going to be developed an automatic machine learning model based on linear regression which has been trained by using historical data and that uses real time data to predict the number of accidents in real time. All these features are going to be integrated through an interactive visualization tool that allows the simulation of different scenarios.

It is considered that the tool developed on the End Master Thesis allows to know which factors are more liable to produce traffic problems, to raise awareness among drivers and, in this way, to reduce the number of accidents and deaths through prevention techniques.

Keywords: Accident, Traffic, Meteorology, Open data, Machine learning, Prediction, Visualization, Data analysis, Dashboard.

Índice

1. Introducción	10
1.1. Contexto y motivación	11
1.2. Objetivos	12
1.2.1. Objetivo general	12
1.2.2. Objetivos específicos.....	13
2. Estado del arte	14
3. Tecnologías y metodologías.....	16
3.1. Metodología	16
3.2. Herramientas y tecnologías	17
4. Fuentes de datos	20
5. Preparación de Datos.....	26
5.1. Relación entre tablas	26
5.2. Día y fecha de accidentes	27
5.3. Municipios y Cartografía	28
5.4. Accidentes en España, comunidades autónomas y Valencia	29
5.5. Accidentes en la provincia de Valencia	31
5.6. Tráfico en Valencia	31
5.6.1. Análisis del flujo de tráfico a la entrada y salida de Valencia.....	32
5.7. Predicción de accidentes en Valencia	34
5.7.1. Modelos de regresión lineal	35
5.7.2. Análisis del modelo de predicción de accidentes en la periferia de Valencia.....	35
5.7.3. Análisis del modelo de predicción de accidentes en Valencia	38
6. Visualización de datos.....	42
6.1. Accidentes en España, en las comunidades autónomas y en Valencia	42
6.2. Como actúan los accidentes en la provincia de Valencia.....	45
6.3. Flujo de Tráfico a la entrada y la salida de Valencia	46
6.4. Predicción de accidentes en Valencia y periferia.....	48
7. Conclusiones	50
7.1. Futuras mejoras	52
7.2. Crítica constructiva	52
8. Bibliografía	53
9. Anexos.....	55
9.1. Descripción de columnas	55



Índice de figuras

Figura 1. Evolución del número de fallecidos en vías 1960-2016	11
Figura 2. Ciclo de la metodología CRISP-DM.	16
Figura 3. Logo R.	17
Figura 4. Logo PHP.....	18
Figura 5. Logo JavaScript.	18
Figura 6. Logo JQuery.	18
Figura 7. Logo AJAX.....	18
Figura 8. Logo D3.js.	19
Figura 9. Logo CARTO.	19
Figura 10. Logo Fusion Tables.	19
Figura 11. Logo Dirección General de Tráfico.	20
Figura 12. Logo Agencia Estatal de Meteorología.....	22
Figura 13. Logo Portal de Transparencia y Datos Abiertos del Ayto. de Valencia.	22
Figura 14. Logo Instituto Nacional de Estadística.	23
Figura 15. Logo del Instituto Geográfico Nacional.	23
Figura 16. Logo policía local de Valencia.	25
Figura 17. Relación general entre tablas.	26
Figura 18. Relación entre tablas para el estudio en Valencia capital.	27
Figura 18. Algoritmo de fecha en accidentes.	28
Figura 19. Estructura de datos de la leyenda, el gráfico circular y el gráfico de barras.	30
Figura 20. Estructura de datos de la leyenda de puntos negros.....	30
Figura 21. Estructura de datos del radar horario.	30
Figura 22. Estructura de datos del radar diario.	30
Figura 23. Muestra de datos procesado añadidos a Carto.	31
Figura 25. Visualización de espiras en Valencia. Muestra la espira 2232 de la V-31.	32
Figura 25. Muestra de datos que resultan del procesado.....	34
Figura 27. Número de accidentes entre 2011 y 2015 por carretera.....	35
Figura 28. Valores de la regresión lineal.....	37
Figura 29. Modelo de regresión lineal generado para predicción de accidentes en la periferia de valencia.....	38
Figura 30. Calles de Valencia con mayor número de accidentes durante el año 2016.	39
Figura 31. Valores de la regresión lineal.....	40
Figura 32. Modelo de regresión lineal generado para la predicción de accidentes en la ciudad de valencia.....	41
Figura 33. Dashboard de accidentes en España.	43
Figura 34. Dashboard de accidentes en la comunidad Valenciana.	43

Figura 35. Dashboard de accidentes en Valencia	44
Figura 30. Mapa interactivo del histórico de accidentes en la provincia de Valencia.	46
Figura 31. Flujo de salida de tráfico en Valencia	47
Figura 32. Flujo de entrada de tráfico en Valencia.	47
Figura 39. Herramienta para predicción de accidentes en la periferia de Valencia.	49
Figura 40. Geometría de las calles de Valencia usadas para la herramienta de predicción de accidentes.	50

Índice de Tablas

Tabla 1. Columnas de la TABLA_ACCVICT	21
Tabla 2. Columnas de la TABLA_PERS.	21
Tabla 3. Columnas de la TABLA_VEHIC.	21
Tabla 4. Columnas de la tabla de Meteorología.	22
Tabla 5. Columnas de la tabla de tráfico.	23
Tabla 6. Columnas de la tabla de datos de municipales.	23
Tabla 7. Columnas de las tablas cartográficas en Carto.	24
Tabla 8. Columnas de la tabla de accidentes de tráfico en Valencia.	25
Tabla 9. Espiras por carretera para entradas y salidas.	33
Tabla 10. Generando variable predictora.	37

1. Introducción

Según la organización mundial de la salud, los traumatismos debidos a accidentes de tráfico se encuentran dentro de las principales causas de muerte en el mundo (OMS, 2012). Dentro de los países de europeos, España se sitúa por debajo de la tasa europea de muertes por accidente, posicionándose como uno de los mejores países del mundo en cuanto a seguridad vial (DGT, 2017). A pesar de estos positivos datos, la cifra de accidentes y muertes por año dentro del territorio español asciende a 89.000 accidentes y 1.300 muertes entre los años 2011 a 2015. Estas cifras no pueden ser despreciadas y más cuando hablamos de vidas humanas. Esta situación invita a buscar urgentemente una solución que permita reducir el número de víctimas de accidente viales.

Son numerosos los factores que contribuyen a la presencia de accidentes. Sería interesante analizar estos factores con el objetivo de reducir el número de accidentes de tráfico. Estos factores abarcan campos como patrones de conducción, agentes meteorológicos, tráfico en carreteras, calendario, infraestructuras viales, disposición y geometría de las carreteras, vehículos, factores humanos, etc.

La gran cantidad de factores que influyen en los accidentes de tráfico hace imposible su análisis a través de las técnicas tradicionales. En los últimos años técnicas como el Big Data han permitido la realización de estos análisis de una forma eficiente y empleando un tiempo mucho menor que el utilizado por otras técnicas obteniendo, además, consiguiendo unos resultados más precisos.

Para cumplir con este objetivo, este trabajo se propone el desarrollo de una plataforma web que permita una visualización interactiva donde elaborar un estudio de los accidentes de tráfico ocurridos tanto a nivel nacional como provincial. Esta plataforma se focalizará finalmente en la ciudad de Valencia y en su periferia, zona de la cual se han conseguido registros de datos de los factores prioritarios en los que se centra el estudio: accidentes, meteorología y tráfico.

La web presentara diferentes tipos de visualización permitiendo al usuario comprender y estudiar la información a conveniencia, generando así una experiencia de estudio de datos más amena al lector. Adicionalmente, la plataforma web también dispone de un apartado de predicción, el cual estima el número de accidentes que sucederán en las carreteras de la ciudad de Valencia y su periferia. Con todo esto se pretende prevenir y conocer cómo se desarrolla el tráfico en la ciudad y en sus alrededores y de este modo conseguir una gestión más eficiente y rápida los sistemas de alerta y prevención.

Para el estudio, se procesan conjuntos masivos de datos procedentes de portales de datos abiertos, tanto históricos como en tiempo real, además de un conjunto de datos referente a accidentes producidos en la ciudad de Valencia, el cual no se encuentra en el catálogo de datos abiertos de la ciudad y que ha sido solicitado exclusivamente para este proyecto.

Este proyecto busca abordar problemas que afectan al desarrollo sostenible de la sociedad mediante análisis de datos, con el fin de impactar positivamente en la sociedad.

Gracias a estos estudios, se puede mejorar la atención en carreteras y la logística de cuerpos de policía y de centros médicos encargados de atender a los heridos y actuar ante los accidentes. Además, mediante estos estudios se busca mejorar las

infraestructuras viales y su correcta señalización, adaptar campañas de concienciación a conductores y poner foco en los principales factores de riesgo, entre otros.

1.1. Contexto y motivación

En las últimas dos décadas, el número de muertos en las carreteras españolas ha experimentado un significativo descenso de cerca del 60%. Las cifras obtenidas en el año 2000 apuntan este año como un punto de inflexión en cuanto a la seguridad vial se refiere. El número de muertes se situaba en torno a 4.000 con cerca de 23 millones de vehículos en circulación. Estas cifras continuaron disminuyendo, 15 años después el número de muertes había descendido hasta casi un cuarto de las víctimas que se produjeron en el año 2000, produciéndose 1.126 muertos en 2015 a pesar de que el número de vehículos había aumentado a 31 millones de vehículos.

La principal causa del descenso de este tipo de siniestros críticos tiene origen en las diversas políticas ejercidas por el gobierno tales como la creación de un carnet por puntos, limitación de velocidades, radares, mejoras de infraestructuras y renovación de la flota de vehículos.

Estas medidas han conseguido reducir de forma eficaz la mortalidad en los últimos años, sin embargo, actualmente nos encontramos ante un punto de estancamiento. Las medidas aplicadas hasta hace un par de años no son capaces de seguir reduciendo el número de siniestros y de muertes en carretera. Este hecho pone de manifiesto la clara necesidad de diseñar nuevos métodos que hagan frente a este problema (DGT, 2017).

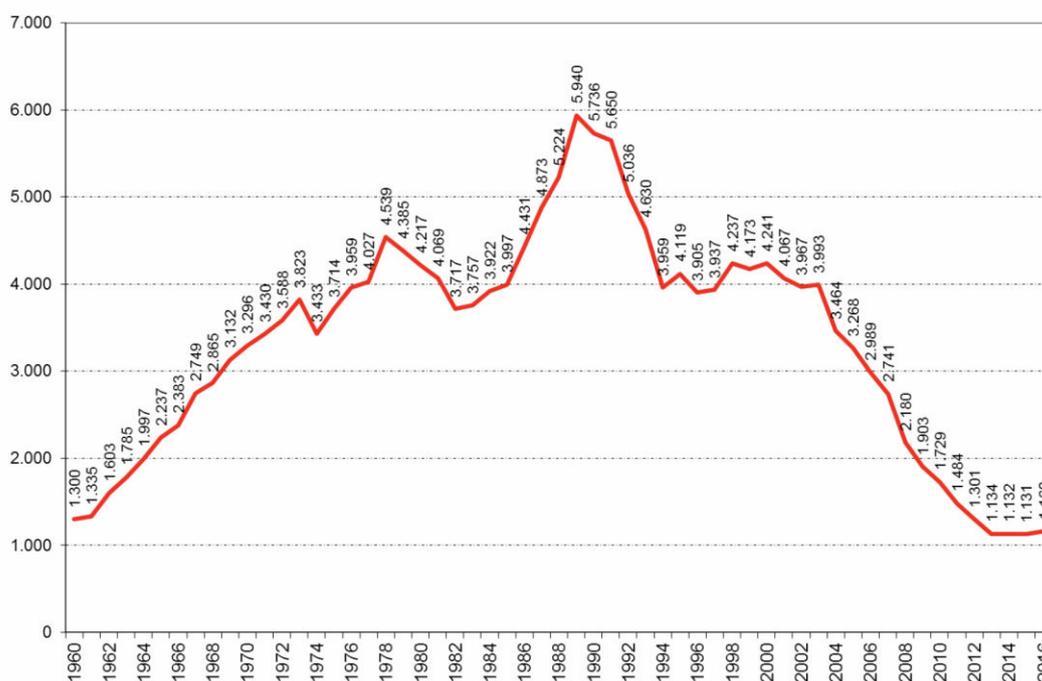


Figura 1. Evolución del número de fallecidos en vías 1960-2016

Consciente del contexto actual, la propuesta de este proyecto es realizar un estudio con el fin de detectar puntos y situaciones propensas a accidentes consiguiendo, de este modo prevenir accidentes adaptando las medidas a los verdaderos causantes y así conseguir reducir aún más el número de siniestros en carretera.

Este proyecto tendrá dos focos de estudio. Por un lado, se estudiarán los siniestros ocurridos a nivel nacional y a nivel de comunidad autónoma, siendo este estudio más descriptivo. Por otro lado, se estudiarán los accidentes de tráfico en la ciudad de Valencia, puesto que tras una intensa búsqueda no se han encontrado estudios relacionado para dicha ciudad, lo cual es un incentivo adicional ya que es la ciudad donde el autor de este proyecto cursa sus estudios de máster.

Teniendo en cuenta los conjuntos de datos disponibles en el momento de la realización del trabajo, otro de los alicientes para la llevar a cabo este proyecto es el reto de adaptación de los datos para conseguir el resultado más cercano al foco principal del estudio.

En la Comunidad Valenciana entre los años 2011 a 2015 se produjeron una media de 7.350 accidentes por año, de los cuales cerca de 2.500 accidente se produjeron en zonas próximas a Valencia capital, es por ello que el estudio principal no solo se encierra en la misma ciudad, sino que busca de una manera más amplia contextualizar los accidentes en la capital.

Como la información de accidentes de tráfico en la ciudad de Valencia no está disponible desde el inicio del proyecto, el estudio se plantea centrándose en la zona periférica de Valencia. Tras la entrega conjuntos de datos referentes a accidentes de tráfico en la ciudad de Valencia, se adaptan los estudios realizados para conseguir entender mejor los accidentes de tráfico en la ciudad.

El método propuesto en este trabajo de fin de máster para reducir el número de muertes por siniestralidad es un modelo predictivo basado en técnicas estadísticas. Este modelo permite predecir con cierto margen los accidentes que se producen en determinadas carreteras, mostrando la influencia de diferentes factores a tiempo real y facilitando la actuación de cuerpos del estado, ya sea policía o servicios médicos, además de arrojar información que facilite la modificación de las normas ante ciertos agentes externos y condiciones ambientes.

1.2. Objetivos

1.2.1. Objetivo general

Con el fin de impactar positivamente a la sociedad y contribuir a disminuir la cifra de accidentes de tráfico y mejorar la actuación y trabajo de los cuerpos oficiales del estado, este proyecto tiene como objetivo la elaboración de una plataforma web que permita entender de forma sencilla y visual patrones y situaciones propensas a accidentes.

Esta plataforma web constará de diferentes análisis y visualizaciones a nivel nacional y autonómico, así como para la ciudad de Valencia, estudiando en estos la relación entre accidentes, clima y tráfico recogidos a través de los datos históricos de diferentes portales datos abiertos y de datos personalizados a petición que deberían encontrarse en portales de datos abiertos.

Además, la plataforma web constará de una aplicación que prediga el número de accidentes en las calles de valencia y en las carreteras periféricas de la ciudad, ya sea a tiempo real o para un caso hipotético, mediante modelos de aprendizaje automático de regresión lineal nutridos por distintas fuentes de datos.

1.2.2. Objetivos específicos

Es totalmente necesario dividir el objetivo principal en diferentes objetivos más pequeños y específico con el fin de que la unión de estos objetivos complete el objetivo principal.

- Captura y limpieza del histórico de datos. Objetivo: Recolectar conjuntos de datos de los diferentes portales de datos abiertos, definir la calidad de los mismo y prepararlos para los diferentes análisis.
- Estudio a nivel nacional, autonómico y de Valencia. Objetivo: Crear y nutrir de información los *dashboard* de visualización de datos con la que estudiar los patrones de accidentes a partir del histórico de datos.
- Estudio a nivel provincial para la provincia de Valencia. Objetivo: Generar una visualización interactiva para encontrar patrones a través de los datos y entender cómo se desarrollan en la línea del tiempo los accidentes en la provincia.
- Estudio del flujo de tráfico a nivel local. Objetivo: Generar una visualización interactiva para conocer el comportamiento del flujo de tráfico de entrada y de salida a través de los datos.
- Crear un modelo predictivo de accidentes en las calles de Valencia y en las carreteras periféricas de Valencia. Para ello se genera un modelo de aprendizaje automático de regresión lineal para ambos focos, que se nutra de las diferentes fuentes de datos y que predigan el número de accidentes que habrán dependiendo de unas entradas.
- Herramienta interactiva que visualice los modelos predictivos para las calles de Valencia y para las carreteras periféricas de la ciudad. Objetivo: Generar una herramienta interactiva para cada caso que muestre y permita conocer la predicción de accidentes generada por los modelos predictivos, ya sea para predicciones a tiempo real o para casos hipotéticos. Para la obtención de datos a tiempo real, las herramientas deben conectarse a los servicios web necesarios para obtener esa información.
- Creación de una página web para que el proyecto tenga visibilidad y pueda ser utilizado por usuarios interesados en el estudio.

2. Estado del arte

Según datos aportados por la UE (European union, 2015), la carga económica anual en Europa que tiene como causa los accidentes de tráfico, tiene un coste aproximado de 10.000 a 14,000 millones de euros. Esto motiva a los gobiernos a la mejora de la seguridad vial y se marca la reducción de accidentes de tráfico como uno de los principales retos en el Horizonte 2020, en el cual se ha incrementado la inversión en seguridad vial.

La gran cantidad de datos abiertos disponibles ha posibilitado el inicio de diferentes estudios para conseguir satisfactoriamente la reducción de accidentes de tráfico.

España, aún se encuentra alejada del nivel de transparencia de datos necesaria para este tipo de estudio, sin embargo, se ha avanzado con la creación de portales como el portal estadístico de datos de la DGT u otros portales municipales donde ya se ofrecen datos de accidentes en la ciudad, como es el caso de Madrid.

En España, el organismo encargado de la seguridad vial es la DGT. Este organismo genera anualmente multitud de informes relacionados con el tráfico. Para el estudio en materia de accidentes de tráfico, esta entidad genera desde 2007 informes anuales estadísticos en los que detalla cuantitativamente las cifras de accidentes dependiendo de su casuística (ANU, 2016).

Existe un gran número de estudios de proyectos de investigación que buscan mediante diferentes técnicas predecir el número de accidentes en las carreteras. Algunos ejemplos tanto a nivel nacional como internacional:

- Predicción de la severidad de accidentes de tráfico en la Red de Carreteras de España y Reino Unido mediante modelos estadísticos basados en Random Forest y Regresión Logística (González, D. Ú.).
- Modelos predictivos de accidentes de tráfico en Madrid (Cruz-Bellas, L.).
- Tratamiento de outliers en los modelos de predicción de accidentes de tráfico (Flores, F. S., Mayora, J. P., & Piña, R. J.).
- *Predicting Traffic Accidents Through Heterogeneous Urban* (Yuan, Z., Zhou, X., Yang, T., Tamerius, J., & Mantilla, R.).

También existen proyectos en los que se generan herramientas interactivas de predicción de accidentes. Uno de los proyectos pioneros que más llaman la atención en el ámbito de predicción de accidentes es el proyecto *Crash Risk Map* del gobierno de Indiana (Estados Unidos). Este proyecto propone una herramienta que predice donde es probable que se produzca un accidente en el estado de Indiana. Esta herramienta es generada por la policía y por el portal de transparencia del estado de Indiana.

La herramienta está disponible en el portal del estado de Indiana como una aplicación que suma al día a día de sus ciudadanos. La web del proyecto se puede visitar en <https://www.in.gov/isp/ispCrashApp/main.html>.

A pesar de la multitud de estudios y proyectos realizados en materia de predicción de accidentes de tráfico, se concluye que aún queda mucho camino por recorrer tanto de estudio de predicción como de puesta en marcha de herramientas para la ciudadanía.

En España existen herramientas que permiten al conductor conocer el flujo de vehículos a tiempo real y dan a conocer puntos donde se ha producido un accidente, sin embargo, no existe ninguna plataforma dirigida a predecir qué puntos o zonas son propensas a accidentes y que de este modo se puedan tomar medidas previas al accidente y se mejore la logística del lugar y la logística de los profesionales del estado que trabajan para salvar vidas y mantener el orden en accidentes.

Los proyectos científicos que se comentan previamente, consiguen unos resultados de predicción de accidentes mejorables. Es más que probable que los resultados de los estudios mejoren a la par que mejoren los catálogos de datos abiertos.

De igual modo, exceptuando el proyecto de Indiana, el resto cortan su proyecto en la evaluación del mismo y no despliegan una herramienta de uso.

En este proyecto se desarrollan diferentes herramientas que permiten ir un paso más allá en materia de estudio de accidentes de tráfico, ya que aportan la posibilidad de interacción y aprendizaje de patrones de accidentes mediante herramientas de visualización de datos.

Aunque a nivel nacional se han realizado diferentes estudios de predicción de accidentes de tráfico, para Valencia que es la ciudad que se tiene como foco este estudio, no se ha realizado previamente un estudio similar.

La herramienta de predicción de accidentes tiene como referente la plataforma generada en la web del estado de Indiana para el mismo cometido. Para intentar mejorarla, la aplicación web de predicción de accidentes que se genera en el proyecto permitirá al usuario generar posibles escenarios que ayude a entender mejor el entorno.

Como punto diferenciador con respecto a otros estudios, en este proyecto se parte de la hipótesis de que factores climatológicos y factores de tráfico tienen gran relación con los accidentes de tráfico. Es por ello que los diferentes estudios centran el valor en estos factores.

3. Tecnologías y metodologías

3.1. Metodología

Para llevar a cabo el proyecto se ha seguido la metodología de trabajo CRISP-DM (*Cross Industry Standard Process for Data Mining*) (MODELER, I. S., 2011). Esta metodología es una de las más utilizadas en proyectos de minería de datos.

Las fases en las que se divide el ciclo vital de la metodología CRISP-DM son las siguientes:

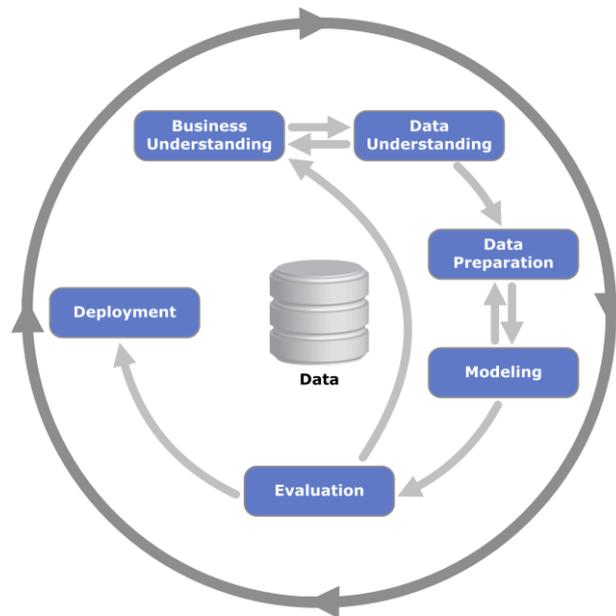


Figura 2. Ciclo de la metodología CRISP-DM.

El trabajo comienza a la vez que nos hacemos la pregunta de si hay relación entre los accidentes y los diferentes factores meteorológicos o factores de tráfico. Partiendo de la hipótesis de que estos pueden ser los factores causantes de un gran número de accidentes, se valora la posibilidad de generar una plataforma web con visualizaciones interactivas que ayude a predecir patrones y situaciones propensas a accidentes y que genere un impacto en los conductores y en las personas que toman decisiones y modelan en cuanto a tráfico nacional. Para ello se pone el foco del estudio en la ciudad de Valencia.

El primer paso es comenzar buscando datos relacionados con accidentes, clima y tráfico. Para ello se buscan en diferentes portales de datos abiertos a nivel nacional. Una vez localizados estos datos y habiendo hecho una lectura de los mismos, se valora su calidad y la calidad de las estructuras que tienen los *datasets*.

El siguiente paso es focalizar hacia donde se quiere centrar el estudio.

Una vez se han definido la fase de *Business Understanding* y de *Data Understanding* el siguiente paso es la fase de *Data preparation*. En esta fase, una vez ya se tienen los datos recopilados y se sabe que se va a estudiar, se procede a la preparación de estos datos.

Para cada uno de los estudios se hace un tratamiento de datos específico. Para ello se seleccionan los datos deseados, se limpian y se modifica la estructura para que posteriormente puedan integrar. En esta fase también se generan datos y variables adicionales dependientes de los datos ya existentes, los cuales no están por defecto en los *dataset* y que se consiguen cruzando los diferentes conjuntos de datos.

Para la fase de *Data preparation* en este proyecto se usan herramientas de análisis como hojas de cálculo o R.

La siguiente fase es la de *Modeling*. En este proyecto se realizan modelos de regresión lineal. Es en este punto donde se decide qué modelo elegir. Para el correcto procesado de los datos, en continuas ocasiones se requiere volver a la preparación de datos para obtener mejores resultados. Además, en esta fase se ejecuta un plan de prueba del modelo y se evalúa.

La siguiente fase es la fase de evaluación. Es aquí donde se verifica que el modelo sea factible y que cubra con el objetivo de estudio o que no lo cubra y se deba revisar el proceso.

Por último, la fase de *deployment*. Como plan de implantación, se desarrolla una web donde mostrar los diferentes estudios y las diferentes visualizaciones. Es aquí donde se desarrollan las diferentes herramientas de estudio de accidentes de tráfico.

Para finalizar el proyecto, se evalúan los resultados, se obtienen conclusiones y se establece una planificación con posibilidades a futuro.

3.2. Herramientas y tecnologías

Para el desarrollo de este proyecto se han utilizado diferentes lenguajes de programación y herramientas que se detallan a lo largo de este punto. Para la realización del proyecto se utiliza una maquina con sistema operativo macOS 10.13.6.

A modo resumen, se utiliza el *software* R para generar modelos estadísticos y procesar los datos procedentes de los diferentes portales. El proyecto se desenvuelve en un entorno web, por lo que se requiere de un dominio y un servidor, además del conocimiento de diversos lenguajes de programación web.

Para enriquecer el proyecto con visualizaciones cartográficas, se utiliza *software* de mapeo web como Carto o Fusion Tables.

➤ Lenguaje R



Figura 3. Logo R.

La principal herramienta de análisis de los datos ha sido R (Team R., 2015). El IDE de desarrollo en el que se programa para R es RStudio.

R es un lenguaje y un entorno de programación orientado a análisis estadístico. Este language permite trabajar con gran cantidad de datos de una manera eficiente además de ser usada para visualización y limpieza de datos.

R es de software libre bajo la licencia de *Free Software Foundation's GNU General Public Licence* (License, G. G. P, 1991), para la mayoría de sistemas UNIX, Windows y Mac OS.

Además, R ofrece una gran variedad de paquetes y librerías. En este proyecto se han usado las librerías “*dplyr*”, “*tidyverse*”, “*pbapply*”, “*clickR*”, “*sqlf*”, “*DAAG*” y “*ggplot2*”.

➤ Servidor y Dominio

Para el alojamiento web se ha utilizado un servidor NAS Synology, que hace de unidad de almacenamiento compartida.

Para el dominio de la web se ha utilizado el servicio NO-IP, donde se ha creado un *Hostname* gratuito cuyo nombre es “<http://thesischallenge.sytes.net/>”.

Durante el desarrollo del proyecto se ha utilizado la aplicación MAMP que permite trabajar en local.

Para la web se ha usado la plantilla “*Timber template*”.

➤ Lenguaje PHP



Figura 4. Logo PHP.

Como lenguaje web del lado del servidor se ha utilizado PHP.

PHP es un lenguaje totalmente libre y de código abierto, multiplataforma y con una sintaxis simple. Su entorno de desarrollo es rápido y de fácil configuración. Tiene un fácil acceso a bases de datos y tiene una gran comunidad detrás.

➤ Lenguaje JavaScript



JavaScript

Figura 5. Logo JavaScript.

Como lenguaje web en el lado del cliente se ha utilizado JavaScript, haciendo uso de las librerías JQuery, AJAX y D3.js. Gracias a estas herramientas, se han desarrollado las visualizaciones de datos interactivas y otros efectos dinámicos de la web.



Figura 6. Logo JQuery.

JQuery permite simplificar la manera de interactuar con el HTML, manejar eventos, desarrollar animaciones e interactuar con el AJAX en las páginas web.



Figura 7. Logo AJAX.

AJAX o *Asynchronous JavaScript And XML* es una técnica de desarrollo web para crear aplicaciones interactivas. Estas aplicaciones se ejecutan en el cliente mientras se mantiene una comunicación asíncrona con el servidor en segundo plano. De esta forma es posible realizar cambios sobre las páginas sin necesidad de recargarlas, mejorando la interactividad, velocidad y usabilidad en las aplicaciones.



Figura 8. Logo D3.js.

D3.js o *Data-Driven Documents* se utiliza para producir, a partir de conjuntos de datos, infogramas dinámicos e interactivos en navegadores web. Utiliza las tecnologías SVG, HTML5 y CSS entre otras.

➤ **Herramientas de mapeo**

Para visualizaciones geográficas se ha utilizado la tecnología de CARTO y Fusion Tables de Google, además del API Maps JavaScript de Google.



Figura 9. Logo CARTO.

CARTO es una plataforma Software de computación en nube que proporciona herramientas SIG y de mapeo web para generar visualizaciones. CARTO se ofrece como un servicio *freemium*, con un límite de tamaño.

En este proyecto se ha usado CARTO Builder que ofrece a los usuarios gestionar datos, realizar análisis en el lado del usuario y diseñar mapas personalizados. Para la manipulación de datos en ciertas ocasiones se ha utilizado lenguaje SQL, igual que para algunos diseños, se ha usado CartoCSS.



Figura 10. Logo Fusion Tables.

Fusion Tables es un servicio web de Google para gestión y visualización de datos que permite el almacenamiento de datos en tablas visibles y descargables entre los usuarios de Internet. Además, visualiza mapas geográficos basados en Google Maps.

Maps JavaScript API permite la interacción con los mapas de Google Maps. Esta herramienta ha sido utilizada en este proyecto para añadir marcadores y secciones interactivas sobre los mapas.

➤ **Herramientas de edición y formato de texto**

Para gran parte de pre-procesado y post-procesado de datos se han usado hojas de cálculo Excel.

Se ha trabajado con formatos CSV, JSON, GeoJSON, XML, GML y KML a la hora de captar los datos que ofrecían los diferentes portales.

Para el estructurado y estilo de la página web se ha utilizado HTML, CSS y Bootstrap que son comunes en desarrollo web. El editor de texto que se ha utilizado ha sido Sublime Text 3 y como control de versiones Git se ha utilizado GitHub.

4. Fuentes de datos

El proyecto centra su valor en torno a los datos. Es completamente necesario obtener registros de los diferentes portales de datos abiertos. Es por ello que en este apartado se detallan los portales de donde se han obtenido los diferentes conjuntos de datos.

Para los diferentes estudios se utiliza información referente a accidentes de tráfico, intensidad de tráfico, meteorología, cartografía y datos demográficos. Es por ello que se accede al portal de la Dirección General de Tráfico, al portal de datos abiertos de Valencia, al portal de la Agencia Estatal de Meteorología, al portal del Instituto Geográfico Nacional y al portal del Instituto Nacional de Estadística. Para las fuentes de datos no encontradas y necesarias para los diferentes estudios se ha recurrido a la vía administrativa y se han solicitado a la unidad que genera dichos conjuntos de datos. Es el caso de datos de accidentes de tráfico para las diferentes vías de la ciudad de Valencia.

A continuación, se nombran los diferentes portales y conjuntos de datos que se utilizan para el proyecto:

➤ Datos de accidentes



Figura 11. Logo Dirección General de Tráfico.

Dirección General de Tráfico (DGT) pone a la disposición de la ciudadanía a través de su portal estadístico (P.E. DGT, s.f.). los datos anuales relativos a los accidentes de tráfico. Estos datos se encuentran en el apartado de Microdatos y son los datos estadísticos desagregados relativos a víctimas, vehículos y personas para cada accidente. Estos datos además están anonimizados, por lo que no incumplen la ley de Reglamento General de Protección de Datos (RGPD).

Se dispone de datos estadísticos entre los años 2008 y 2015, aunque en el proyecto se usan datos entre 2011 y 2015, ya que los ficheros anteriores y posteriores a 2010 tienen diseños de registros diferentes.

Para cada año se descargan varios ficheros que se relacionan entre sí por una columna llamada “ID_ACCIDENTE”. Los ficheros que se descargan se llaman: víctimas (TABLA_ACCVICT_XXX), vehículos (TABLA_VEHIC_XXXX) y personas (TABLA_PERS_XXXX), siendo “XXXX” el año de ocurrencia de los accidentes. Además, el fichero vehículo y personas se relaciona adicionalmente por otra columna llamada “ID_VEHICULO”.

Para cada registro de la tabla Accidentes puede haber 1 o más vehículos implicados. De igual forma, para cada vehículo implicado puede haber 1 o más ocupantes del vehículo.

Columnas de los accidentes de tráfico		
ID_ACCIDENTE	TOT_MUERTOS30D	PRIORIDAD_AGENTE
ANIO	TOT_HERIDOS_GRAVES	PRIORIDAD_SEMAFORO
MES	TOT_HERIDOS_GRAVES30D	PRIORIDAD_STOP
HORA	TOT_HERIDOS_LEVES	PRIORIDAD_CEDA
DIASEMANA	TOT_HERIDOS_LEVES30D	PRIORIDAD_MARCAS
PROVINCIA	TOT_VEHICULOS_IMPLICADOS	PRIORIDAD_PASO
COMUNIDAD_AUTONOMA	ZONA	PRIORIDAD_OTRA
ISLA	ZONA_AGRUPADA	SUPERFICIE_CALZADA
COD_MUNICIPIO	CARRETERA	LUMINOSIDAD
MUNICIPIO	RED_CARRETERA	FACTORES_ATMOSFERICOS
TOT_VICTIMAS	TIPO_VIA	VISIBILIDAD_RESTRINGIDA
TOT_VICTIMAS30D	TRAZADO_NO_INTERSEC	FACTORES_ATMOSFERICOS
TOT_MUERTOS	TIPO_INTERSEC	TIPO_ACCIDENTE

Tabla 1. Columnas de la TABLA_ACCVICT

Columnas de personas involucradas	
ID_ACCIDENTE	HERIDO_GRAVE30D
ID_VEHICULO	HERIDO_LEVE_24H
ID_PERSONA	HERIDO_LEVE30D
ID_CONDUCTOR	DICCIONARIO_MANIOBRAS
ID_PASAJERO	MANIOBRAS
ID_PEATON	INFRACC_VELOCIDAD
EDAD	INFRACC_COND
SEXO	INFRACC_APERTURA
ANIO_PERMISO	INFRACC_ALUMBRADO
POSICION	INFRACC_CARGA_VEHICULO
USO_CINTURON	INFRACC_RESUMEN
USO_SRI	INFRACC_PEATON
USO_CASCO	DICCIONARIO_ACCION_PEATON
MUERTO_24H	ACCION_PEATON
MUERTO_30D	ANIO
HERIDO_GRAVE_24H	

Tabla 2. Columnas de la TABLA_PERS.

Columnas de vehículos involucrados	
ID_ACCIDENTE	ANOMALIA_REVENTON
ID_VEHICULO	ANOMALIA_DIRECCION
ANIO_MATRICULA_VEHICULO	ANOMALIA_FRENOS
MES_MATRICULA_VEHICULO	NUMERO_OCUPANTES_VEH
TIPO_VEHICULO	MERCANCIAS_PELIGROSAS
ANOMALIA_NINGUNA	VEHICULO_INCENDIADO
ANOMALIA_NEUMATICO	

Tabla 3. Columnas de la TABLA_VEHIC.

➤ Datos meteorológicos



Figura 12. Logo Agencia Estatal de Meteorología

La entidad encargada de los datos de meteorología a nivel nacional es la Agencia Estatal de Meteorología o AEMET (AEMET, s.f).

AEMET OpenData es una API REST desarrollada por AEMET que permite la difusión y la reutilización de la información meteorológica y climatológica (AEMET O.D., s.f).

En este portal se utilizan dos de sus servicios. Por un lado, el histórico de datos climatológicos de valencia y, por otro lado, la predicción del tiempo para obtener datos a tiempo real para la herramienta de predicción. Para estos servicios es necesaria una API Key personal que no se adjunta en los repositorios.

La asociación DataUPV ofrece en su catálogo el histórico de la observación meteorológica en Valencia (DataUPV Meteo, 2016). Dado que estos datos ya están trabajados son los que se van a usar. Se dispone de datos entre los años 2012 y 2016.

DataUPV es un grupo estudiantil de la Universitat Politècnica de Valencia que promueve el uso de los datos abiertos entre la Comunidad Universitaria.

Velocidad del viento: m/s
Dirección del viento: grados
Temperatura: °C
Humedad relativa: %
Presión: mb
Precipitación: l/m2

Tabla 4. Columnas de la tabla de Meteorología.

➤ Datos de tráfico



Figura 13. Logo Portal de Transparencia y Datos Abiertos del Ayto. de Valencia.

En la ciudad de valencia, la entidad encargada de recopilar los datos referentes a la ciudad es el Portal de Transparencia y Datos Abiertos (Trafico PTDAAV, s.f).

El objetivo de Valencia Datos Abiertos es poner a disposición de la ciudadanía, una serie de datos elaborados por el propio Ayuntamiento que puedan dar lugar a nuevos desarrollos de aplicaciones y estudios (Trafico PTDAAV, s.f).

Para acceder a los datos de referentes a tráfico, se debe buscar el dataset “Intensidad de los Puntos de Medida de Tráfico (Espiras electromagnéticas)” (Trafico PTDAAV, s.f). Este servicio ofrece a tiempo real en diferentes formatos según se prefiera el tráfico en diferentes puntos de la ciudad. Siendo más concretos, actualmente hay 996 puntos de recogida de datos referente al tráfico en funcionamiento.

Estos datos son útiles para la plataforma de predicción. Sin embargo, para el estudio de tráfico es de interés el histórico de estos datos y en este portal no se muestra un histórico. De nuevo se acude a la asociación DataUPV ya que ofrecen en su catálogo el histórico de tráfico en Valencia (Tráfico DataUPV, 2016). Se dispone de datos entre los años 2012 y 2014.

Ángulo: Ángulo representación del sentido de la circulación en el punto de medida.
Fecha de actualización.
Hora de actualización.
Idpm: Identificador de punto de medida.
Ih: Intensidad hora de vehículos (vehículos/hora)

Tabla 5. Columnas de la tabla de tráfico.

➤ Datos de carácter municipal



Figura 14. Logo Instituto Nacional de Estadística.

El Instituto Nacional de Estadística (INE) es el organismo encargado de la coordinación general de los servicios estadísticos de la Administración General del Estado y la vigilancia, control y supervisión de los procedimientos técnicos de los mismos (INE, s.f).

Los datos que se obtienen en el INE son muy útiles para enlazar el código municipal con los que se cataloga cada accidente con los datos cartográficos. Además, son completamente necesario para que las visualizaciones y los estudios sean completos.

Como dato importante a tener en cuenta y que no es obvio, el código municipal de un municipio es diferente al código postal. Todos los municipios tienen dos códigos asociados a ellos. Esta confusión en el análisis de datos puede llevar a análisis erróneos.

CodProvincia
CodMunicipio
CodigoPostal
Municipio

Tabla 6. Columnas de la tabla de datos de municipales.

➤ Datos de carácter cartográficos



Figura 15. Logo del Instituto Geográfico Nacional.

Para conseguir datos cartográficos se ha acudido a la plataforma del Instituto Geográfico Nacional (IGN) (IGN, s.f). El Instituto Geográfico Nacional es el organismo encargado del desarrollo de la instrumentación e infraestructuras para la realización de



trabajos y estudios en los campos de la astronomía, la geodesia y la geofísica.

En este proyecto se necesita la cartografía de comunidades, provincias y municipios de España, para ello accedemos al catálogo de datos que tiene el portal (IGN Data, s.f).

Cabe comentar que Carto tiene pre-cargado diferentes dataset cartográficos. Estos dataset se llaman “spanish_adm1_ccaa” para cartografía de las comunidades autónomas de España, “ign_spanish_adm2_provinces” para cartografía de las provincias, “ejes_calle” para la cartografía de los ejes de las calles de Valencia y “ign_spanish_adm3_municipalities” para cartografía de los municipios españoles. La existencia de estos dataset cartográficos tiene lógica en esta plataforma ya que son datos comúnmente usados y actualizados por parte de la comunidad de usuarios.

Estos serán los datos que posteriormente usaremos para las visualizaciones cartográficas de los diferentes mapas.

La columna “the_geom” es la que porta la información de la geometría de las calles de valencia, los municipios, provincias y comunidades autónomas. “the_geom” aparece como un listado de coordenadas, vértices o nodos que delimitan el polígono de la geometría. Para algunos mapas, estas coordenadas pueden ser modificadas para crear polígonos a placer y así conseguir mapas más adaptados al resultado buscado.

the_geom
cod_ccaa / cod_mun / cod_prov
noml_ccaa / noml_mun / noml_prov
nom_ccaa / nom_mun / nom_prov
created_at
updated_at

Tabla 7. Columnas de las tablas cartográficas en Carto.

➤ Datos de accidentes de tráfico en la ciudad de Valencia

Para el llevar a cabo el desarrollo de este proyecto se necesitan datos relativos a accidentes de tráfico producidos en las vías de la ciudad de valencia. Los datos que aporta la Dirección General de Tráfico en su portal de estadística se refieren a accidentes de tráfico de carreteras, zonas urbanas y travesías, no entrando en las correspondencias locales de los municipios. Los encargados de almacenar la información referente a los accidentes de tráfico son las unidades de tráfico locales, que posteriormente se encargan de transmitir esa información a la DGT y almacenar los conjuntos de datos generados.

Tras una búsqueda en los diferentes portales de datos abiertos de España, solo se encuentran datos de accidentes de tráfico locales en el portal de Madrid con datos referentes a Madrid. Como en el Portal de Transparencia y Datos Abiertos de Valencia no se encuentran estos datos, se les hace una petición, derivando la búsqueda al portal estadístico del ayuntamiento de Valencia (Portal estadístico, s.f), entidad que tiene informes agregados de accidentes de tráfico. Tras una petición de datos a esta entidad, se informa que no tienen permiso de entregar esos datos de forma desagregada y que la entidad que almacena y tiene posibilidad de distribuir el conjunto de datos de accidentes de tráfico es el departamento de atestados de la policía local de Valencia (Policía Local, s.f) mediante previa solicitud administrativa vía sede electrónica del ayuntamiento de Valencia (Sede electrónica , s.f) mediante la opción “otras solicitudes no catalogadas” (Solicitud , s.f).



Figura 16. Logo policía local de Valencia.

Se dispone de datos estadísticos entre los años 2010 y 2017. Para cada año existe un fichero llamado “red-atestado-YYYY”. Los datos generados son desagregados y han sido previamente anonimizados por la unidad de atestados cumpliendo así la ley de protección de datos y evitando exponer a las personas implicadas.

Las columnas que tienen los ficheros son iguales en todos los ficheros, a excepción de la columna de la ubicación donde se produce el accidente que hasta el año 2010 tiene un campo para calle, número y/o cruce, en 2011 tiene dos campos, uno para la calle y el número y otro para el cruce y a partir de 2012 tiene tres campos, uno para la calle, otro para el número y otro para el cruce.

Cada fichero contiene los registros con un identificador de registro por año. El nombre de las columnas es descriptivo y en algunas columnas se tiene un comentario extra, aunque existen otras columnas cuyo significado de contenido es difícil de interpretar.

Columnas de los accidentes de tráfico en Valencia		
ID_REGISTRO	J.R. PERMISO	CRUCE
DIA	J.R. COOPERADOR	TIPO DE ACCIDENTE
MES	J.R. DROGAS	OCUPANTES
HORA	J.R. SE NIEGA	PEATON
FUGA	J.R. ALCOHOLEMIA	DAÑOS MUNICIPALES
INMOVILIZACION	ALCOHOL ADMTVA.	UNIDAD 1
FECHA INMOVILIZACION	ATTDO. X DELITO+ACCTE. (penal)	DELITO CONTRA LA SEGURIDAD VIAL
FECHA DESINMOVILIZACION	ACCTE. SIN VICTIMAS	UNIDAD 2
DETENIDO	DILIG. PREVENCIÓN	HERIDOS
PUESTA LIBERTAD	ATTDO. ACCTE. CON VICTIMAS	MUERTOS
TRASLADO CEDRO	INTERV. CON MENORES	RADAR Km/Hora
J.R. RADAR	CALLE	ALCOHOLEMIA
J.R. TEMERARIA	Nº	1ª PRUEBA
LESIONES LEVES	LESIONES GRAVES	2ª PRUEBA
SAMU - SVB	UNIDAD	PAIS DE ORIGEN
HOSPITAL	TURNOS	SENTENCIAS

Tabla 8. Columnas de la tabla de accidentes de tráfico en Valencia.

5. Preparación de Datos

Uno de los pasos más importantes para llevar a cabo los distintos análisis del proyecto es la preparación de los datos. Para esta preparación es imprescindible conocer la relación que tienen las diferentes tablas de datos.

Para simplificar las relaciones y evitar posibles errores en nombres de columnas, se estandarizan cambiándolos al mismo formato. Posteriormente se ha de conocer que parámetros no se van a usar en los estudios y eliminar aquellas características que no se emplean en los análisis, para así reducir el vector y reducir tiempos de computación.

De igual modo, es importante saber con qué datos se cuenta, es decir, si están completos todos los registros. Una vez se conoce esto, se eliminan nulos, duplicados y espurios. Además, como se cuentan con datos de diferentes sitios, es importante que los valores de las diferentes variables tengan un mismo formato y también es algo que hay que estandarizar. Por ejemplo, formato de la variable fecha “dd-mm-yyyy” y el formato “yyyy-mm-dd”.

Como cada estudio es diferente, una vez se hayan resuelto los anteriores pasos, se generan los diferentes *data frame* según se requiera. Para ello, es común la creación de columnas a partir de otras columnas ya existentes, mediante técnicas de cruzado de tablas, que aportan un nuevo valor o significado que no se tenía o que simplifica el posterior uso de los datos.

5.1. Relación entre tablas

A continuación, se muestra la relación entre tablas y las columnas imprescindibles para los diferentes estudios. Ya aparecen los nombres estandarizados.

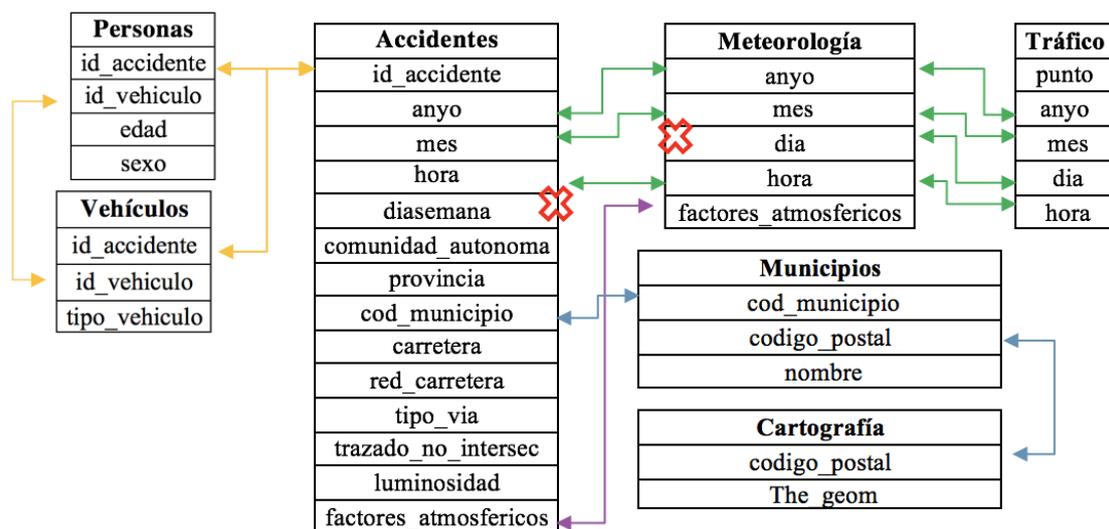


Figura 17. Relación general entre tablas.

Las diferentes tablas tienen relación casi completa, exceptuando el día del mes en que se produjo el accidente. Para los accidentes en vez de tener el día del mes o la semana del mes del accidente se tiene el día de la semana (lunes a domingo) en el que se produjo el accidente. El año, mes y hora sí que se tiene en las diferentes tablas. Esto es

algo muy a tener en cuenta ya que dificulta la conexión de los datos e implica utilizar metodologías con las que se pierden información. De igual modo, el hecho de no tener las fechas completas implica que no se puede utilizar el calendario como un factor para los diferentes estudios del proyecto.

Usando los datos de accidentes de la ciudad de Valencia conseguidos se consigue relación directa entre el tráfico, la meteorología y el registro del accidente, permitiendo así un análisis de mayor calidad.

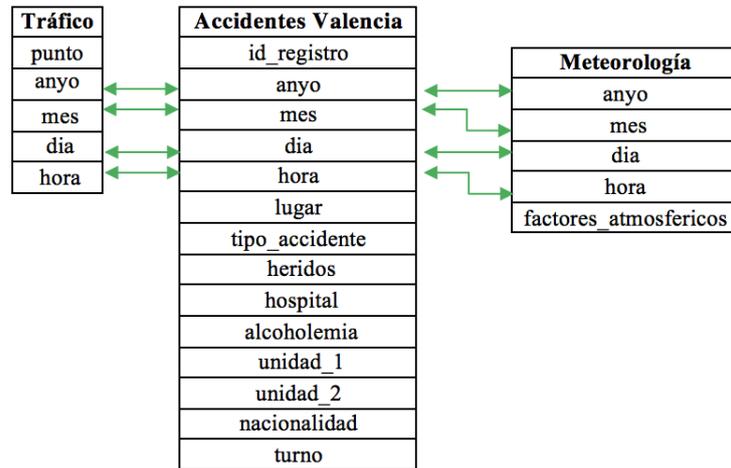


Figura 18. Relación entre tablas para el estudio en Valencia capital.

5.2. Día y fecha de accidentes

Gracias a los datos obtenidos de accidentes en Valencia, relacionar las tablas para posteriormente nutrir cada registro de accidentes con mayor cantidad de información y de este modo conseguir un estudio de mayor calidad es relativamente sencillo, sin embargo, ¿cómo conseguir el día y la fecha de los accidentes procedentes del conjunto de datos de la DGT y de este modo poder relacionarlos con el resto de tablas y conseguir una mayor cantidad de información con mejor calidad de datos?

En un intento por conseguir el día del mes en que se produjo el accidente (1 al 31) y por consecuente la fecha completa del siniestro, se propone un método que se acerca a ese resultado.

En primer lugar, este método se basa en la hipótesis de que todos los registros de accidentes están ordenados secuencialmente según sucedieron. Esto unido al calendario anual y a la columna “DiaSemana” (1 a 7, siendo 1 lunes y 7 domingo), se intentaba crear una columna con la semana en la que producía cada accidente. El número de semana iría de la 1 a la 52, dependiendo del año. Una vez se tiene el número de la semana y el día de la semana, sabiendo cual es el primer día de una semana, se consigue llegar hasta la fecha del siniestro.

Teóricamente funciona, pero a la hora de la verdad hay registros cuya semana de pertenencia es difícil de catalogar, ya que no siguen un patrón. Este patrón lo dicta el cambio de varios 6 y 7 seguidos (sábados y domingos) y varios 1 y 2 seguidos (lunes y martes). Cuando en la cadena aparece una secuencia similar el algoritmo lo asimila como cambio de semana.

Tras varios intentos fallidos de programar un algoritmo para llegar a la fecha, se dedujo que la única manera de entender un patrón era asignar a mano la semana a cada registro. Esto es poco eficiente y tedioso, además de estar sujeto a errores humanos y errores de catalogación. Es por ello que no se continuó con esta opción.

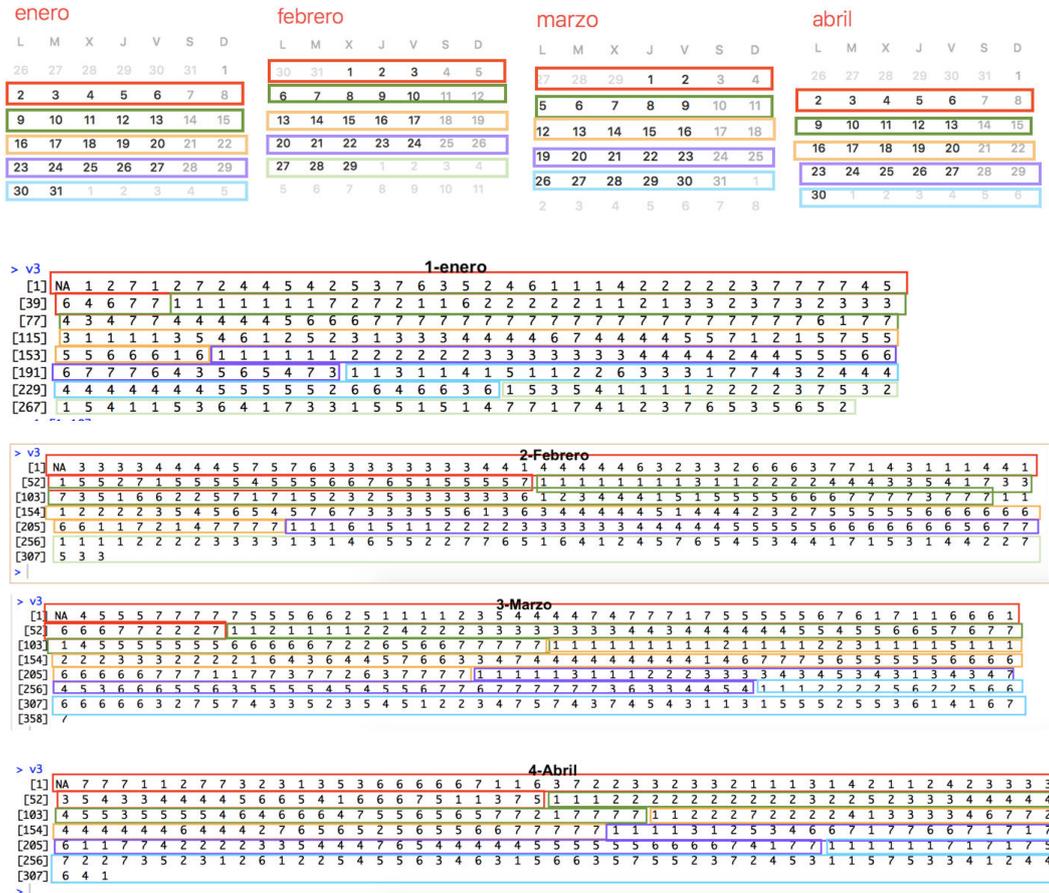


Figura 18. Algoritmo de fecha en accidentes.

5.3. Municipios y Cartografía

Para que los mapeados de datos tengan sentido y puedan utilizarse de manera correcta, es necesaria relacionar la geometría de un municipio con el código del municipio.

Como se comenta previamente, todos los municipios tienen asociados 2 códigos diferentes: código postal y código de municipio.

Los registros de los accidentes tienen asignados el código de municipio, mientras que los registros de la geometría tienen asignados el código postal. Es necesario la unión de estos mediante una tercera tabla que relacione el código de municipio con el código postal.

5.4. Accidentes en España, comunidades autónomas y Valencia

Se pretende realizar un análisis cuantitativo de España, de las comunidades autónomas y de Valencia mediante un *dashboard* personalizado en cada uno de los casos. Para nutrir de datos estos *dashboard* interactivos, en primer lugar, se ha de pensar cómo está estructurada la plataforma. Como hablamos de una herramienta web que se ha diseñado exclusivamente para el proyecto, se tiene cierta libertad de cómo tratar los datos.

No interesa que todos los datos estén pre cargados en la plataforma, sino que se carguen a medida que el usuario acceda a una determinada vista de las herramientas, así se consigue que la web sea más ligera. Partiendo de esta base se decide usar AJAX. Gracias a AJAX es posible realizar peticiones al servidor y obtener respuesta de este en segundo plano sin necesidad de recargar la página web completa.

Una vez se sabe que tecnología se va a utilizar, lo siguiente es saber cómo estructurar la información. No es lógico que todos los datos referentes a los diferentes estudios estén almacenados en un mismo archivo muy grande. Es más eficiente generar archivos pequeños para cada estudio y cada gráfico interactivo. Este es el motivo por el que se piensa en usar rutas de almacenaje de la información.

Las herramientas web permiten mediante botoneras, selectores y mapas interactivos navegar por los diferentes análisis de datos. De este modo se acceden a los datos almacenados en las diferentes rutas, consiguiendo así una página mas ligera.

En las visualizaciones de España y de las comunidades autónomas se cuentan con datos de los años 2011 al 2015. En esta aplicación se estudian los accidentes de tráfico dependiendo de los factores a los que se veían sometidos o a las características del conductor. Se realiza un estudio dependiendo de factores atmosféricos, factores lumínicos, tipo de superficie, zona, tipo de vía, tipo de intersección, sexo o edad del conductor o años de experiencia del conductor. La herramienta desarrollada tiene como objetivo poder adaptarse a tantos estudios como la imaginación lo permita y que el desarrollador no encuentre gran dificultad a la hora de generar la información que posteriormente muestra.

En las visualizaciones, en cuanto a Valencia se cuentan con datos de los años 2010 al 2017. En esta aplicación se estudian los accidentes de tráfico dependiendo de la intensidad de tráfico, del clima, país de procedencia, motivo del accidente, hospitales a los que se lleva a los involucrados, grado de alcohol en sangre, vehículos implicados, delitos contra la seguridad vial, unidades de la policía desplazadas, entre otros.

Los datos almacenados en las rutas tienen diferentes estructuras. Estas estructuras dependen del gráfico que los lee. En la visualización hay dos leyendas, una para puntos negros dependiendo del factor de estudio y otra donde se muestran los parámetros de estudio, un gráfico circular que representa los porcentajes en relación a los parámetros estudiados, un diagrama de barras que muestra la cantidad mensual para cada parámetro estudiado y dos gráficos de radar que determinan el número de accidentes dependiendo la hora y por día de la semana. A continuación, se muestran las diferentes estructuras de datos que lee la aplicación.

```
[
  {month: 'ENE',
    freq: {
      Pleno_Día: 6,
      Crepúsculo: 1,
      Noche_Iluminación_Suficiente: 3,
      Noche_Iluminación_Insuficiente: 1}},
  {month: 'FEB',
    freq: {
      Pleno_Día: 3,
      Crepúsculo: 6,
      Noche_Iluminación_Suficiente: 9,
      Noche_Iluminación_Insuficiente: 1}},
  {month: 'MAR',
    freq: {
      Pleno_Día: 9,
      Crepúsculo: 1
```

Figura 19. Estructura de datos de la leyenda, el gráfico circular y el gráfico de barras.

```
[
  {info: 'carretera',
    carretera: {
      1: 'CV-500',
      2: 'V-21',
      3: 'V-31',
      4: 'A-3',
      5: 'CV-36'}},
  {info: 'municipio',
    carretera: {
      1: 'Valencia',
      2: 'Alboralla',
      3: 'Lliria',
      4: 'Torrent',
      5: 'Albal'}},
  {info: 'latitud',
    carretera: {
      1: '39.47726',
      2: '39.232848',
      3: '39.667533',
      4: '37.468545',
      5: '42.19623'}},
  {info: 'longitud',
    carretera: {
      1: '-0.347329',
      2: '-1.623102',
      3: '-4.540085',
      4: '-6.256298',
      5: '-2.971794'}}}]
```

Figura 20. Estructura de datos de la leyenda de puntos negros.

```
[
  {axis: '00-02', value: 4},
  {axis: '02-04', value: 0},
  {axis: '04-06', value: 1},
  {axis: '06-08', value: 121},
  {axis: '08-10', value: 1525},
  {axis: '10-12', value: 3808},
  {axis: '12-14', value: 4938},
  {axis: '14-16', value: 922},
  {axis: '16-18', value: 4229},
  {axis: '18-20', value: 6565},
  {axis: '20-22', value: 1012},
  {axis: '22-24', value: 16}],
```

Figura 21. Estructura de datos del radar horario.

```
[
  {axis: '1', value: 4},
  {axis: '2', value: 0},
  {axis: '3', value: 1},
  {axis: '4', value: 121},
  {axis: '5', value: 1525},
  {axis: '6', value: 3808},
  {axis: '7', value: 4938}],
```

Figura 22. Estructura de datos del radar diario.

Como se puede deducir, ha de generarse un gran número de ficheros para las diferentes visualizaciones. Estos ficheros no se generan a mano, sino que se generan

mediante programación de *scripts* específicos que generan y almacenan los datos a leer dependiendo del mapa (España, comunidad autónoma o Valencia), año o estudio seleccionado, todo esto en la ruta que posteriormente leerá la aplicación web.

Los scripts para generar los datos en crudo se generan mediante R, que filtra y exporta los datos, guardando estos en la ruta correcta. La información es guardada en formato “.csv”. Posteriormente otro script en PHP lee los archivos y genera las estructuras de datos mostradas en las figuras 19, 20 21 y 22 dependiendo del gráfico al que se destine la información y exporta estas estructuras a un fichero “.js” con un nombre específico dependiendo de los datos leídos y que será el fichero que lea la aplicación web.

5.5. Accidentes en la provincia de Valencia

Para llegar hasta la visualización deseada de como transcurren los accidentes en la provincia de Valencia y encontrar patrones en los datos, es necesario procesar previamente estos datos. En esta visualización se pretende utilizar Carto, por lo que hay que conocer que formatos de datos acepta esa herramienta y que datos queremos visualizar.

La intención es conseguir una serie temporal durante el año 2012 de los accidentes que se producen en la provincia de Valencia. Para ello se filtran los registros de los accidentes por “provincia” y por “año”. Para esta visualización se usan las fechas conseguidas a través del algoritmo mencionado en el apartado 5.2 para obtener el día del mes. Además, para posicionar en el mapa cada accidente se utiliza la columna “cod_municipio” que hace referencia a la geometría como se comenta en el punto 5.3.

Las fechas en Carto tienen un formato diferente al que se tienen en los conjuntos de datos, es por ello que también es necesario modificar estos. Carto tiene el formato “yyyy-mm-ddThh:mm:ssZ” mientras que nuestros datos siguen el formato “yyyy-mm-dd”.

cartodb_id ↑ number	the_geom geometry	category string	value number
1	-0.38855923, 39.47329...	2012-8-28 4:00:00+00	1
2	-0.38855923, 39.47329...	2012-1-19 14:00:00+00	1
3	-0.38855923, 39.47329...	2012-7-14 17:00:00+00	1
4	-0.10965884, 38.90746...	2012-2-12 19:00:00+00	2

Figura 23. Muestra de datos procesado añadidos a Carto.

5.6. Tráfico en Valencia

El tráfico en Valencia no es un valor genérico para toda la ciudad. Para conocer la intensidad de tráfico se necesita una espira electromagnética que es la encargada de

medir dicha intensidad. Estas espiras se encuentran dispersas por toda la ciudad. Actualmente se encuentran 996 espiras en funcionamiento, de las cuales se puede obtener información de la mayoría.

La intensidad de tráfico equivale al número de vehículos que pasa a través de una sección fija de una carretera por unidad de tiempo. El catálogo de datos usado para el estudio utiliza vehículos/hora. La intensidad de tráfico es una variable básica en el análisis del tráfico ya que permite caracterizar el tipo de circulación en un tramo viario.

En muchas ocasiones, la intensidad de tráfico no solo viene condicionada por la demanda, sino que también se condiciona por la capacidad de las vías.

También, se debe de tener en cuenta que el valor de intensidad de tráfico de una espira no solo equivale a el tráfico de la calle donde está situada esta. Existen calles que tienen bifurcaciones, separadores o cruces que dificultan saber la intensidad de tráfico real que hay en dicha vía, ya que el tráfico real puede equivaler a la suma o resta del valor de intensidad de tráfico de 2 o más espiras.

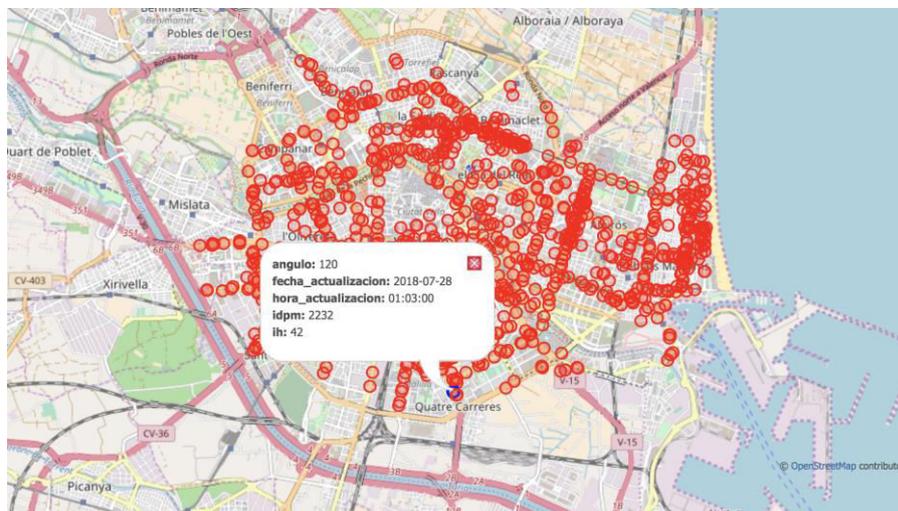


Figura 25. Visualización de espiras en Valencia. Muestra la espira 2232 de la V-31.

Para estudiar los accidentes en las calles Valencia se le aplica el valor medio de intensidad de las espiras contiguas al punto donde se produce el accidente. Esto puede generar cierto error en el valor de la intensidad de tráfico real que había en el momento de los accidentes, pero simplifica en gran manera la relación entre espiras y accidente y facilita su análisis.

5.6.1. Análisis del flujo de tráfico a la entrada y salida de Valencia

Para estudiar cómo es el flujo de entrada y de salida de coches de la ciudad de Valencia, es necesario seleccionar las espiras adecuadas. Estas espiras son las que se utilizarán en el estudio de predicción de accidentes en la periferia de Valencia y que además permiten conocer el comportamiento del tráfico que sigue la ciudad.

Para localizar las espiras que se quieren utilizar, es necesario analizar que carreteras son las más conflictivas en cuanto a accidentes. Esto se estudia mediante un análisis cuantitativo usando R.

Tras definir que carreteras son las más conflictivas, se buscan los puntos o espiras que más se acercan a esa carretera y que miden los niveles de intensidad de entrada y salida en Valencia.

Las carreteras que se estudian son A-3, CV-35, CV-36, CV-315, CV-400, V-21, V-30 y V-31. Hay carreteras que son conflictivas, pero no se puede estimar el tráfico en ellas puesto que no hay espiras que delimiten su intensidad de flujo, como es el caso de la V-30, aunque podría intuirse si hay o no tráfico dependiendo de la intensidad del resto de vías.

Los puntos o espiras que se eligen para este estudio son los que se muestran en la tabla 9. Para elegir esos puntos, se usa la visualización de datos GML que ofrece el portal de datos abiertos del ayuntamiento de Valencia. Se eligen así las espiras que contienen información de flujo de tráfico de entrada y salida, dependiendo de donde estén situadas. Algunas de estas vías tienen varias espiras por lo que la intensidad de flujo total se calcula sumando las intensidades de tráfico.

	in	out
V-31	3407	2232
	2237	3406
V-21	2719	2753
	2754	2752
		2751
		2757
CV-36	2411	2404
A-3	930	909
	907	
	904	
CV-315	1212	1223
CV-35	3632	1019
CV-400	1909	1908

Tabla 9. Espiras por carretera para entradas y salidas.

Una vez localizadas las espiras que se usarán, es necesario conocer tipo de visualización se pretende hacer y cómo adaptar los datos.

La intención es conseguir una serie temporal para el mes de abril del año 2012, en la que se muestre el flujo de tráfico por horas en las entradas y salidas de las carreteras de Valencia citadas en la tabla 9.

Es necesario saber si Carto que es la herramienta que se usa para esta visualización permite generar series temporales en las que el grosor del punto que se mueve temporalmente varía dependiendo del valor de la variable intensidad para cada hora. La herramienta no tiene generada esa función por lo que hay que tratar los datos de una forma más elaborada para conseguir la visualización buscada.

La preparación de los datos, se realiza carretera por carretera para entrada y salida del tráfico. Poniendo foco en una única carretera y en la entrada o salida de tráfico, lo primero que se hace es seleccionar la media mensual por horas. Así se obtienen 24 registros de 0 a 23 y en cada registro la intensidad de tráfico media. En caso de utilizar varias espiras para medir la intensidad de tráfico, se suman las intensidades obteniendo el mismo resultado que se explicaba anteriormente.

Los valores de intensidad de tráfico varían entre un mínimo de 0 y un máximo de 4200 v/h dependiendo de la carretera y la hora de medida. Como queremos que la visualización de intensidad de tráfico esté normalizada para todas las carreteras, se decide visualizar con un punto rojo cada 400 v/h. Así se pasa a una visualización de mínimo 0 y máximo 20 puntos rojos para representar la intensidad de tráfico.

Una vez se sabe el número de puntos que tiene de cada registro, se generan tantos registros nuevos como número de puntos tenga.

Se busca que, para cada hora, las bolas se muevan a lo largo de la carretera a la que pertenezca dependiendo si es la visualización de entrada o salida de tráfico, es por ello que se juega con la columna date y se multiplica por 4 cada registro, asignando a cada uno de ellos el minuto cambiado.

Por ejemplo:

- 1- yyyy-mm-ddThh:00:ssZ
- 2- yyyy-mm-ddThh:18:ssZ
- 3- yyyy-mm-ddThh:37:ssZ
- 4- yyyy-mm-ddThh:55:ssZ

Además, como estamos en una herramienta cartográfica, a la vez que se generan los registros temporales, a las bolas se les asigna una latitud y longitud en torno a una circunferencia y así se consigue en la visualización la impresión de cantidad de tráfico por carretera y de movimiento.

Una vez procesados todas las carreteras se unen los registros para la entrada y para la salida de todas las carreteras.

Punto	HORA	AVG(Intensidad)	date_time	latitud	longitud
2237	2	311.033.333.333.333	2012-4-01T2:55:00Z	39.440.127	-0.534166
2237	2	311.033.333.333.333	2012-4-01T2:37:00Z	39.435.946	-0.370108
2237	2	311.033.333.333.333	2012-4-01T2:19:00Z	39.389.839	-0.393275
2237	2	311.033.333.333.333	2012-4-01T2:00:00Z	39.349.157	-0.429306
2237	3	221.233.333.333.333	2012-4-01T3:55:00Z	39.440.127	-0.534166
2237	3	221.233.333.333.333	2012-4-01T3:37:00Z	39.435.946	-0.370108
2237	3	221.233.333.333.333	2012-4-01T3:19:00Z	39.389.839	-0.393275
2237	3	221.233.333.333.333	2012-4-01T3:00:00Z	39.349.157	-0.429306
2237	4	218.4	2012-4-01T4:55:00Z	39.440.127	-0.534166
2237	4	218.4	2012-4-01T4:37:00Z	39.435.946	-0.370108
2237	4	218.4	2012-4-01T4:19:00Z	39.389.839	-0.393275
2237	4	218.4	2012-4-01T4:00:00Z	39.349.157	-0.429306

Figura 25. Muestra de datos que resultan del procesado.

5.7. Predicción de accidentes en Valencia

Para completar este apartado, es necesario entender la teoría de los modelos de regresión lineal. Una vez se conoce como funciona, se comenta como se ha procesado el modelo de regresión lineal generado para la predicción de accidentes en la periferia de Valencia y el modelo generado para predicción de accidentes en la ciudad de Valencia. Además, se hablará a cerca de los resultados que se obtiene del modelo generado.

5.7.1. Modelos de regresión lineal

Estadísticamente hablando, regresión es un proceso estadístico que estima las relaciones entre diferentes variables. Cuando las variables están relacionadas, es posible predecir un valor de respuesta a partir de un valor predictor.

Los modelos basados en regresión lineal siguen la siguiente ecuación:

$$y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e$$

donde:

- y representa la componente aleatoria o variable dependiente.
- X_k representa las variables explicativas de dependencia lineal.
- b_0 representa el *intercept* estimado.
- b_k representan los coeficientes predictores.
- e representa el término de perturbación o error

El objetivo del modelo es buscar una estimación de valores a los diferentes parámetros para que la suma de los errores al cuadrado se minimice.

Un modelo lineal es sencillo de visualizar en dos dimensiones, donde equivalen a una recta que estima el valor lineal de los datos existentes.

5.7.2. Análisis del modelo de predicción de accidentes en la periferia de Valencia

El objetivo final que se busca es estimar del número de accidentes que habrá en las carreteras periféricas de Valencia.

Para poder llevar a cabo la predicción, previamente es necesario transformar los datos. En primer lugar, se seleccionan los registros de accidentes de la provincia de Valencia que pertenezcan a las carreteras A-3, CV-35, CV-36, CV-315, CV-400, V-21, V-30 y V-31, que son las carreteras periféricas a la ciudad de Valencia y de las cuales se tiene el histórico de datos.

	acc\$carretera	n
1	A-3	562
2	CV-315	60
3	CV-35	459
4	CV-36	91
5	CV-400	94
6	V-21	161
7	V-30	352
8	V-31	312

Figura 27. Número de accidentes entre 2011 y 2015 por carretera.

Como los registros de accidentes no tienen un parámetro de intensidad de tráfico por defecto, es necesario generarlos mediante los datos de tráfico. Para ello se le asigna a

cada registro la media de intensidad histórica mensual, dependiendo de la hora y de la carretera en la que sucediese el accidente.

Para realizar el estudio de regresión lineal se cuentan con varios parámetros que a priori pueden ser variables que estén directamente relacionadas con la causa de accidente. Estas son el clima, el tráfico, la curvatura de la carretera, la luminosidad, la hora, el mes y día de la semana.

Para poder usar en el modelo estas variables, es necesario generar nuevas columnas en el *data frame* que simplifique la información para cada registros. A continuación, se habla de cómo se trató cada variable:

- Variable clima: se divide en buen tiempo o mal tiempo. Es por ello que se aplican valores *dummy* (0,1). Previamente esta variable contaba con diferentes valores: soleado, lluvia suave, lluvia intensa, nieve, granizo, viento. El motivo por el que se divide en mal tiempo y buen tiempo se debe a que se cuenta con poca cantidad de datos y los registros asociados a mal tiempo como son lluvia, nieve, viento o granizo eran insignificantes por separado, comparando con buen tiempo o soleado.
- Variable luminosidad: se divide en día, noche y crepúsculo. Para ello se generan dos columnas con variables *dummy*, una llamada día (0 equivale a crepúsculo y 1 a día) y otra llamada noche (0 equivale a crepúsculo y 1 a día).
- Variable curvatura: se divide en curva o no curva. Para ello se genera una columna con variables *dummy* llamada curva (0 curva y 1 no curva).
- Variable mes: Se genera una columna con 4 valores. 0 primavera, 1 verano, 2 otoño, 3 invierno.
- Variable hora: Se genera una columna *dummy* (0 para [23h - 6h] y 1 para [7h - 22h]).
- Variable día de la semana: valores de 1 a 7.
- Variable intensidad: Se genera una columna con 3 valores. 0 para intensidad baja, 1 para intensidad media y 2 para intensidad alta. Para considerar si una intensidad es alta, baja o media se utiliza un percentil de 0 a 25% para intensidad baja, 25 a 70% para intensidad media y 70% a 100% para intensidad alta. Los niveles equivalen a [0 – 700]v/h, [700 – 2000]v/h y [2000 – 4300]v/h. Para este estudio se estiman esos percentiles, aunque no sea correcto medir flujo de tráfico sin tener en cuenta el número de carriles, velocidad de circulación y densidad de tráfico que son los parámetros de los que depende la intensidad de tráfico.

El modelo que se usa es un modelo de regresión lineal. Para realizar una predicción mediante regresión lineal es necesaria una variable predictora. Partiendo de que se quiere conseguir el número de accidentes dependiendo de unos parámetros de entrada, se genera un nuevo *data frame* que contempla todas las posible causas de accidente y a cada causa de accidente se le añade una nueva columna con el número de accidentes que hubieron dependiendo de esas causas de accidente.

tiempo	dia	Nº accidentes en esas circunstancias
0	0	50
0	1	13
1	0	30
1	1	20

Tabla 10. Generando variable predictora.

Una vez procesada toda la información referente a las variables que usa el modelo, el siguiente paso es generar el modelo de regresión lineal. Para ello se usa la función `lm()` que tiene R para generar regresiones. En esta función el primer argumento `y ~ x` especifica cuál es la variable respuesta o dependiente (y) y cuál es la variable regresora o independiente (X). El segundo argumento, especifica el fichero en el que se encuentran las variables. A continuación, se muestra la función con los parámetros finales del modelo de regresión lineal.

```
lm(n ~ dia + noche + diasemana + intensidad + tiempo, data = data_resultados)
```

Para llegar hasta un modelo utilizando las variables que, a priori eran las que tenían relación con el suceso de accidentes, se utiliza la técnica de prueba y error. Para conseguir mejores resultados se combinan las diferentes variables, deduciendo de cada una de las combinaciones si el modelo es mejor o peor. Esto se deduce mediante el valor del coeficiente de determinación o correlación al cuadrado. Este parámetro mide la bondad del ajuste de la recta a los datos.

Finalmente, se deduce que las variables que tienen mayor relación con la predicción de accidentes es la luminosidad (dia, noche, crepúsculo), el día de la semana (lunes a domingo), intensidad de tráfico (nivel bajo, medio o alto) y clima (buen o mal tiempo).

```
> report(lm1)
              Estimate Std. Error Lower 95% Upper 95% P-value
(Intercept)    0.775      0.804   -0.812    2.362    0.336
dia             1.712      0.521    0.682    2.742    0.001
noche          -3.674      0.521   -4.704   -2.644   <0.001
diasemana      -0.021      0.13   -0.279    0.236    0.87
intensidad_0baja_1media_2alta  0.932      0.319    0.302    1.563    0.004
buen_mal_tiempo_0_1  3.698      0.521    2.668    4.727   <0.001
R Squared      0.424
Adj.R Squared  0.4062
```

Figura 28. Valores de la regresión lineal.

El coeficiente de determinación toma valores entre 0 y 1, y cuanto más se aproxime a 1 mejor es el ajuste y la fiabilidad de las predicciones que con él realicemos. El valor del coeficiente de determinación de la regresión generada es de 0.424 por lo que no es perfecto, pero tampoco es malo del todo.

La ecuación generada que se usará en la visualización de datos para definir el número de accidentes para cada carretera será el siguiente:

$$y = 0.775 + X_1 * 1.712 + X_2 * (-3.674) + X_3 * (-0.021) + X_4 * 0.932 + X_5 * 3.698$$

Finalmente se comparan los valores observados (negro) con los valores obtenidos de la regresión (rojo). El eje X representa las diferentes observaciones y en el eje Y el número de accidentes.

Este modelo no predice de manera fiable el número de accidentes, pero sí que tiene sentido si se tienen en cuenta las variables de entrada que se han tomado como factores de causa de accidentes. Analizando los resultados se aprecia que hay un balanceo entre datos.



El modelo generado es más que mejorable. Como primer paso para mejorar el modelo, es necesario un mayor número de registro de accidentes. Además, es imprescindible probar con más variables de entrada o factores causantes de accidentes para ajustar la predicción a lo que sucede en la realidad.

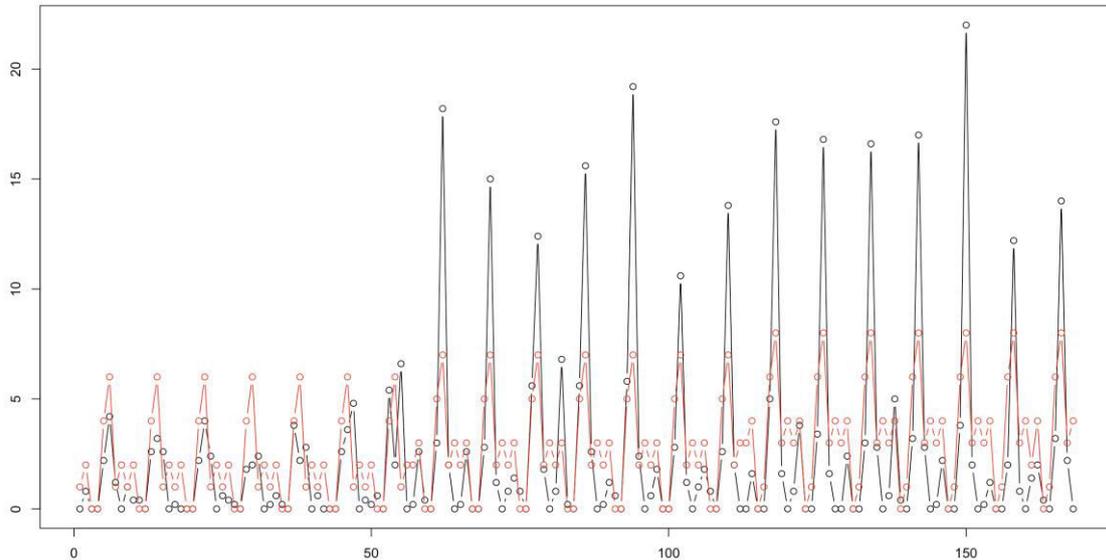


Figura 29. Modelo de regresión lineal generado para predicción de accidentes en la periferia de valencia.

5.7.3. Análisis del modelo de predicción de accidentes en Valencia

El objetivo final que se busca es estimar del número de accidentes que habrá en las carreteras de la ciudad de Valencia.

Para poder llevar a cabo la predicción, previamente es necesario transformar los datos. En primer lugar, se analizan los puntos negros de las calles de valencia. Para ello es necesario tratar el campo referente a la localización puesto que los nombres que aparecen no son siempre iguales, tienen abreviaciones, faltas ortográficas, nombres en valenciano o castellano, calle y numero unido, columnas de ubicación unidas en unas tablas y en otras separadas, puntos y comas, caracteres, etc.

Una vez tratados se analizan los puntos con mayor número de accidentes en la ciudad.

En el análisis se diferencia entre calles, cruces de avenidas y carretera de entrada o salida a autovía. Los puntos más críticos en cuanto a accidentes de tráfico se refieren se sitúan en las calles mostradas en la figura 30. Estas son las calles que más interés de estudio tienen puesto que sería interesante reducir ese número de siniestralidad mediante modelos de predicción como los propuestos en este apartado.

	AC2016\$CALLE	n
2158	BLASCO IBAÑEZ	258
2157	CID	250
2156	HERMANOS MACHADO	207
2155	AUSIAS MARCH	168
2154	TRES CRUCES	154
2153	SAN VICENTE MARTIR	151
2152	ALAMEDA	135
2151	PEREZ GALDOS	134
2150	MARQUES DEL TURIA	129
2149	PERIS Y VALERO	127
2148	MAESTRO RODRIGO	117
2147	PUERTO	107
2146	CORTES VALENCIANAS	96
2145	PRIMADO REIG	92
2144	FERNANDO EL CATOLICO	90
2143	PIO XII	86
2142	ARAGON	84
2141	FERNANDO ABRIL MARTORELL	81
2140	GENERAL AVILES	74
2139	CATALUÑA	70
2138	V-30	69
2137	PAVIA	68
2136	RAMON Y CAJAL	66
2135	DOCTOR MANUEL CANDELA	63
2132	CONSTITUCION	61
2133	GASPAR AGUILAR	61
2134	NUEVE DE OCTUBRE	61
2131	EUGENIA VIÑES	59

Figura 30. Calles de Valencia con mayor número de accidentes durante el año 2016.

Como los registros de accidentes no tienen un parámetro de intensidad de tráfico por defecto, es necesario generarlos mediante los datos de tráfico. Para poder referenciar la intensidad de tráfico de las espiras a cada registro es necesario generar un campo que asocie la ubicación del accidente a la espira más cercana. Una vez generado el campo de ubicación con espira, gracias a que en el *dataset* de accidentes de valencia tiene los campos hora, día, mes y año, se puede asociar con la intensidad de tráfico de la espira.

De igual modo, conociendo la hora, día, mes y año en el que se produjo el accidente, se generara una columna que añade las características meteorológicas que se daban en el momento de los hechos.

Teniendo en cuenta la fecha y el calendario, se generan columnas que indican si un accidente se produjo en festividad o no festividad, si es fin de semana o no es fin de semana, el periodo del año, operación salida u operación retorno, etc.

Para llevar a cabo el estudio de regresión lineal se cuentan con parámetros que a priori pueden ser variables que estén directamente relacionadas con la causa de accidente. Estas variables a tener en cuenta son: el clima, el tráfico, cruce o calle, la luminosidad, la hora, el mes, el día de la semana, la festividad, etc.

Para poder usar en el modelo estas variables, es necesario generar nuevas columnas en el *data frame* que simplifique la información para cada registros. A continuación, se habla de cómo se trató cada variable:

- Variable clima: se divide en buen tiempo o mal tiempo. Es por ello que se aplican valores *dummy* (0,1). La división es en mal tiempo y buen tiempo ya que se cuenta con poca cantidad de datos y de registros asociados a mal tiempo como son la lluvia o el viento y se busca que los datos sean significativos juntos, ya que por separado no lo son.

- Variable luminosidad: se divide en día, noche y crepúsculo. Para ello se generan dos columnas con variables *dummy*, una llamada día (0 equivale a crepúsculo y 1 a día) y otra llamada noche (0 equivale a crepúsculo y 1 a día).
- Variable intensidad: Se genera una columna con 3 valores. 0 para intensidad baja, 1 para intensidad media y 2 para intensidad alta. Para considerar si una intensidad es alta, baja o media se utiliza un percentil de 0 a 25% para intensidad baja, 25 a 70% para intensidad media y 70% a 100% para intensidad alta.
- Variable calle, intersección o carretera: se genera una columna con variables *dummy* llamada para cada tipo.
- Variable festividad: Se genera una columna *dummy* siendo 1 festivo y 0 no festivo.
- Variable día de la semana: valores de 1 a 7.

Al igual que en el modelo generado para la periferia de Valencia, el modelo que se usa es un modelo de regresión lineal. Para conseguir la variable predictora se genera un *data frame* que contemple las causas de accidente.

Tras procesar las variables que usa el modelo, se genera el modelo de regresión lineal. A continuación, se muestra la función con los parámetros finales del modelo de regresión lineal.

$$lm(n \sim dia + noche + diasemana + intensidad + tiempo + festividad + tipo_calle, data = data_resultados)$$

Mediante prueba y error se deduce el modelo con el que a priori mejor resultado de predicción se obtiene. Esto se deduce mediante el valor del coeficiente de determinación o correlación al cuadrado.

Finalmente, se deduce que las variables que tienen mayor relación con la predicción de accidentes son la luminosidad (día, noche, crepúsculo), el día de la semana (lunes a domingo), intensidad de tráfico (nivel bajo, medio o alto), clima (buen o mal tiempo), tipo punto negro (carretera, calle o intersección) y festividad.

```
> report(lmi)
```

	Estimate	Std. Error	Lower 95%	Upper 95%	P-value
(Intercept)	0.235	0.093	0.053	0.417	0.012
día	0.317	0.055	0.21	0.425	<0.001
noche	-0.482	0.055	-0.589	-0.374	<0.001
diasemana	0.005	0.014	-0.022	0.031	0.736
intensidad	-0.283	0.033	-0.349	-0.218	<0.001
tiempo	0.502	0.055	0.395	0.609	<0.001
festividad	0.105	0.055	-0.002	0.213	0.054
tipo_calle	0.216	0.055	0.109	0.323	<0.001
R Squared	0.3018				
Adj.R Squared	0.2944				

Figura 31. Valores de la regresión lineal.

La ecuación generada que se usará en la visualización de datos para definir el número de accidentes para cada carretera será el siguiente:

$$y = 0.234673 + X1 * 0.317262 + X2 * (-0.481548) + X3 * 0.004613 + X4 * (-0.283482) + X5 * 0.501786 + X6 * 0.105357 + X7 * 0.216071$$

Finalmente se comparan los valores observados (negro) con los valores obtenidos de la regresión (rojo).

El valor del coeficiente de determinación de la regresión generada es de 0,3018. A primera vista, parece un modelo aceptable, sin embargo, a la hora de representar los datos se obtiene como resultado un gran desbalanceo de datos, por lo que se deduce que no es un modelo fiable.

El modelo generado es más que mejorable. Como primer paso para mejorar el modelo, es necesario un mayor número de registro de accidentes. Además, es imprescindible probar con más variables de entrada o factores causantes de accidentes para ajustar la predicción a lo que sucede en la realidad y de este modo reducir el balanceo de datos. Otra posible solución es reducir el número de

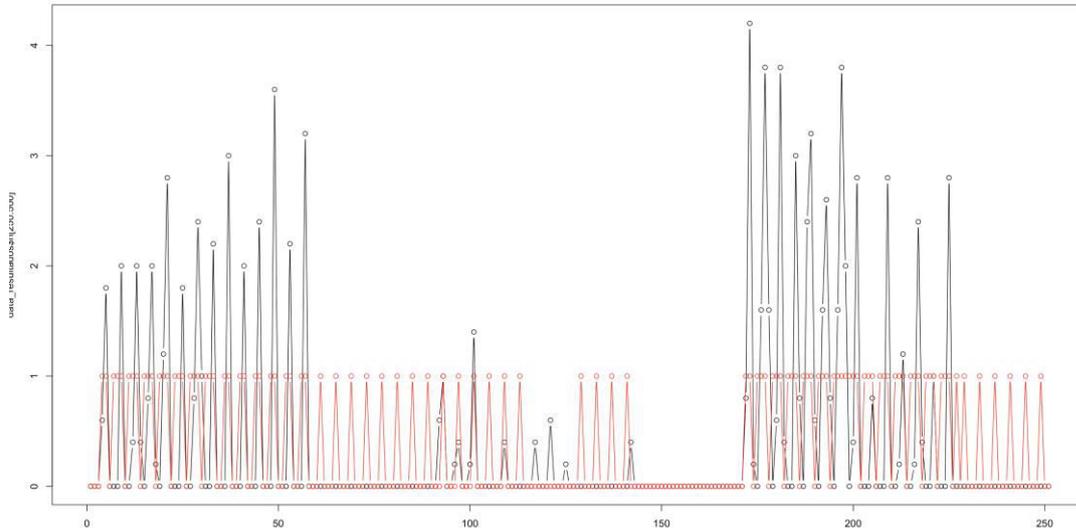


Figura 32. Modelo de regresión lineal generado para la predicción de accidentes en la ciudad de valencia.

6. Visualización de datos

Los datos por sí mismos son invisibles. Para poder verlos y encontrarles sentido, necesitamos visualizarlos. Es por ello que se buscan mecanismos que ayuden a tomar decisiones y a entender los datos de una manera fácil, entendible y valiosa.

Con el fin de conseguir tener un impacto en el usuario se presentan las diferentes visualizaciones del análisis. Se pretende así que los usuarios, independientemente del conocimiento que tengan, sean capaces de interpretar los resultados.

En este apartado se muestra la funcionalidad de cada herramienta generada, al igual que algún aspecto más técnico de cómo se han creado dichas herramientas.

En las visualizaciones valencia se cuentan con datos de los años 2010 al 2017. En esta aplicación se estudian los accidentes de tráfico dependiendo de la intensidad de tráfico, del clima, país de procedencia, motivo del accidente, hospitales a los que se lleva a los involucrados, grado de alcohol en sangre, vehículos implicados, delitos contra la seguridad vial, unidades de la policía desplazadas, entre otros.

6.1. Accidentes en España, en las comunidades autónomas y en Valencia

Con el fin de entender el contexto a nivel nacional, autonómico y local, se realiza un análisis descriptivo y cuantitativo de los datos relativos a accidentes de tráfico. Para estudiar y entender los datos que aportan los diferentes *dataset*, se ha elaborado una herramienta web o *dashboard* que permite al usuario interactuar y entender mejor los datos de estudio.

Al ser una herramienta desarrollada exclusivamente para el proyecto, se cuenta con total libertad a la hora de enlazar las vistas que muestran la información. Es por ello que se han utilizado diferentes gráficos estadísticos, apoyados de un mapa y botoneras que permiten al usuario comprender la información mientras interactúan con esta. La herramienta tiene como objetivo ser sencilla de adaptar a nuevos factores o casuísticas de condiciones que se dieron en los accidentes. De este modo se facilita al desarrollador la generación de tantos nuevos estudios como este imagine.

El *dashboard* de España y las comunidades autónomas tiene como cometido estudiar los accidentes acontecidos entre los años 2011 y 2015 que dependen o se relacionan con características del conductor o factores acontecidos en el lugar y momento del accidente. Se analizan los accidentes dependiendo de factores atmosféricos, factores lumínicos, tipo de superficie, zona, tipo de vía, tipo de intersección, sexo o edad del conductor y años de experiencia del conductor. Estos estudios se muestran como la media para el territorio nacional o como la media para cada comunidad autónoma, consiguiendo de este modo focalizar y diferenciar entre diferentes escenarios.

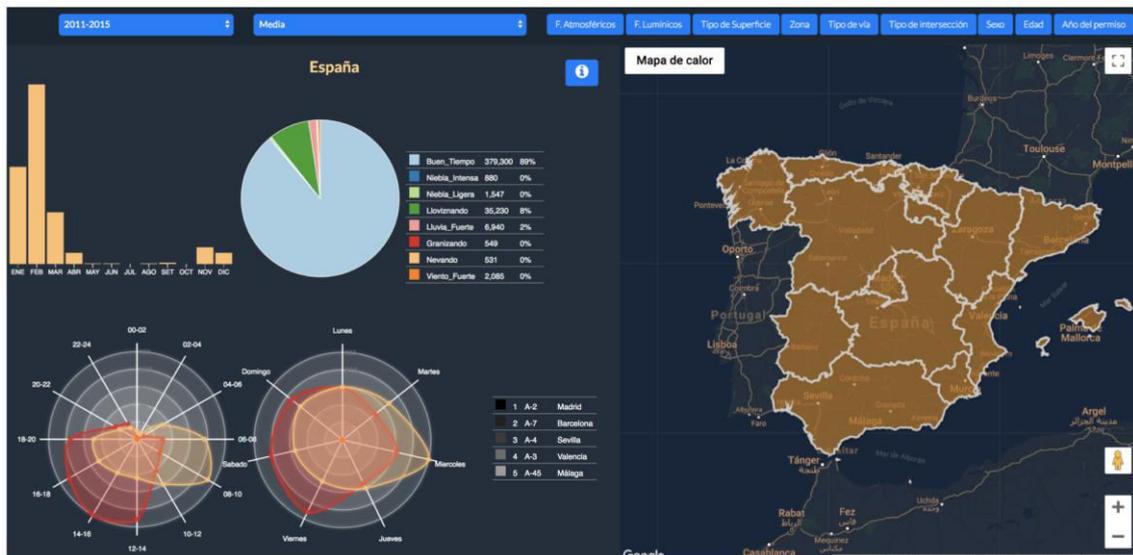


Figura 33. Dashboard de accidentes en España.

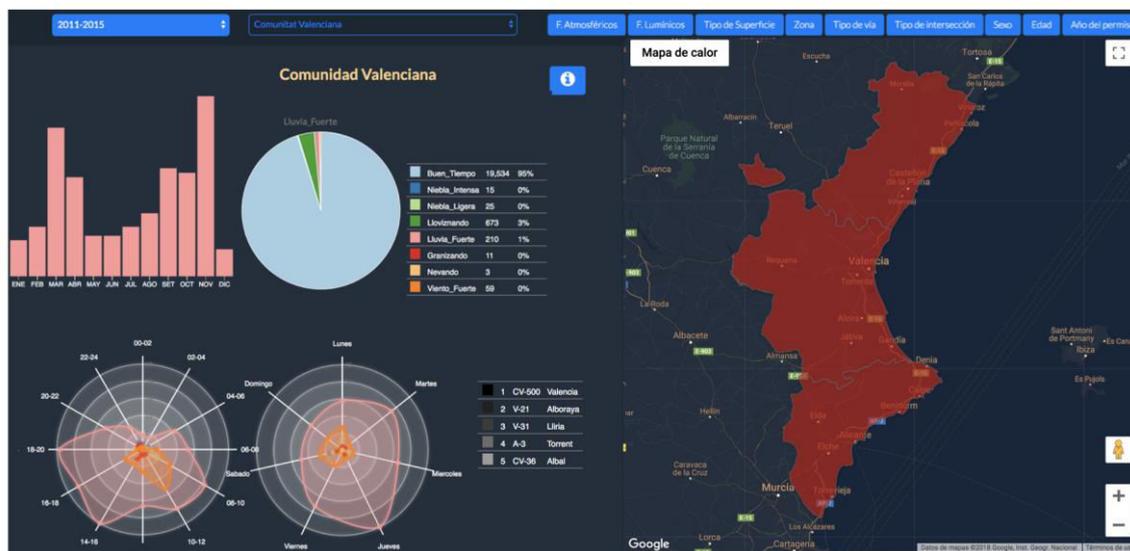


Figura 34. Dashboard de accidentes en la comunidad Valenciana.

El *dashboard* de Valencia tiene como cometido estudiar los accidentes acontecidos entre los años 2010 y 2017. Los factores a estudiar son diferentes a los factores del anterior dashboard puesto que los datos que se tienen para el estudio de accidentes de tráfico en la ciudad son diferentes a los que proporciona la DGT. Se analizan los accidentes dependiendo de factores atmosféricos, factores lumínicos, país de la persona que comete la infracción, tráfico, alcohol en sangre, motivos del accidente, calles más conflictivas, tipo de vehículo involucrado, hospitales en los que se ingresa a las personas implicadas en los accidentes, calendario, etc.

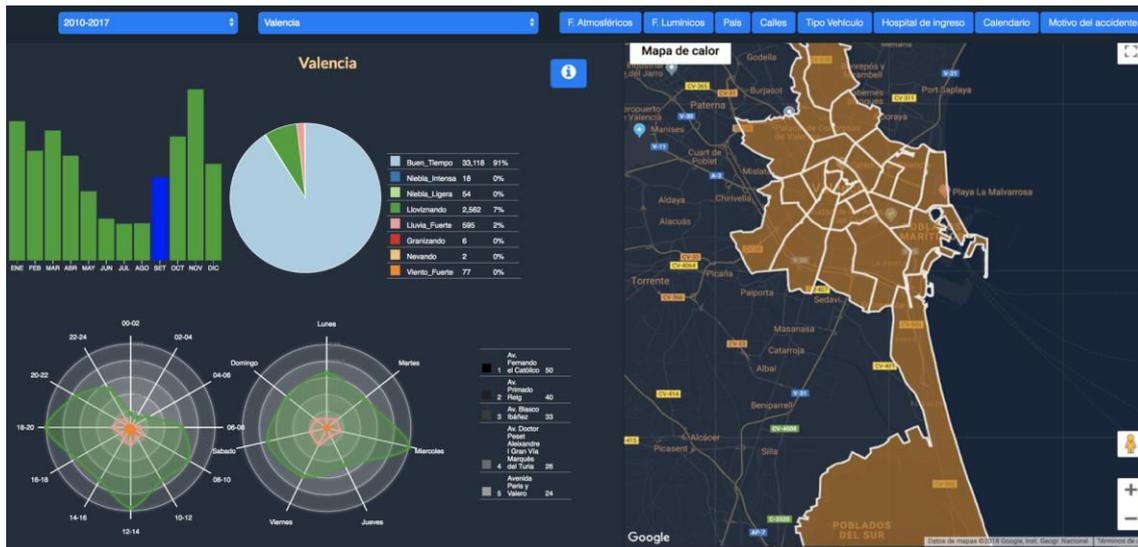


Figura 35. Dashboard de accidentes en Valencia.

Desde un primer momento se ha diseñado un *dashboard* que contenga toda la información a vista de ojo, sin necesidad de cambiar de página o hacer *scroll* para poder interactuar.

El *dashboard* contiene varios elementos que se detallan a continuación:

- Leyenda superior. Esta leyenda muestra los parámetros que se estudian para cada vista. Representa el número de accidentes para cada factor y el porcentaje total con respecto a los otros factores.
- Gráfico de barras. Muestra los accidentes producidos por mes y por factor.
- Gráfico circular. Muestra el porcentaje de accidentes correspondiente a cada factor.
- Gráfico de radar horario. Muestra el volumen de accidentes que se produjeron respecto a la franja horaria para cada factor de estudio.
- Gráfico de radar diario. Muestra el volumen de accidentes que se produjeron respecto a los diferentes días de la semana para cada factor de estudio.
- Leyenda de puntos negros. Muestra los puntos más conflictivos donde se produjeron accidentes.
- Mapa. Muestra la zona de estudio de los datos para cada vista.
- Botones. Permite seleccionar que vista o estudio mostrar.
- Selectores. Permite navegar entre años y comunidades autónomas.

Los diferentes elementos del *dashboards* tienen relación en sí:

- Los valores de la leyenda superior varían dependiendo de la condición seleccionada.
- Cada factor de la leyenda superior tiene un botón que permite seleccionar que gráfico o que factor estudiar en los gráficos de radar inferiores.
- Al pasar el puntero sobre el gráfico circular se muestra el nombre del factor al que pertenece esa porción.

- Al hacer *click* en una porción determinada del gráfico circular, la información del gráfico de barras cambia, mostrando la información de factor seleccionado mes a mes.
- El valor del gráfico de barras se muestra cuando se desliza el puntero sobre el mes deseado. El valor se muestra en la leyenda.
- Al deslizar el puntero sobre el mes de estudio del gráfico de barras, la información de los radares inferiores varía.
- Al deslizar el puntero sobre los vértices de los gráficos dibujados en cada radar, se muestra el valor de dicho vértice.
- Al clicar en cada botón de la leyenda de puntos negros, se actualiza el mapa y muestra la localización del punto.
- El mapa varía dependiendo de la comunidad seleccionada.

6.2. Como actúan los accidentes en la provincia de Valencia

Entender cómo transcurren los accidentes en la provincia de Valencia durante un año determinado y encontrar patrones en estos datos es la intención con la que se crea esta vista.

Para entender donde se producen estos accidentes a través de datos se crea una visualización a mediante la herramienta Carto. Esta herramienta muestra cronológicamente los accidentes que se produjeron, permitiendo filtrar por franja de mes, día y hora. El filtrado que permite esta herramienta es muy interesante ya que adapta y representa el resultado a visualizar dependiendo lo que se selecciona y mostrando de este modo la información generada.

Además, se añade una capa a la visualización con información referente a los municipios. De este modo se muestra la separación de los diferentes municipios mediante polígonos. Al hacer *click* en un municipio que interese, se muestra la información relativa a su nombre y código postal.

En la visualización se aprecia que la mayoría de accidentes se centran en las zonas costeras de la provincia, teniendo mayor número de accidentes cuanto más cerca se está de ciudades grandes como Valencia y Gandía.

De igual modo, se puede apreciar existencia de puntos negros donde se produce un gran número de accidentes como por ejemplo Paterna que es el lugar donde está el parque tecnológico y al que acude gran multitud de personas a trabajar, Manises donde se encuentra el aeropuerto o la pista de Silla que es la zona de acceso de la ciudad por el sur.

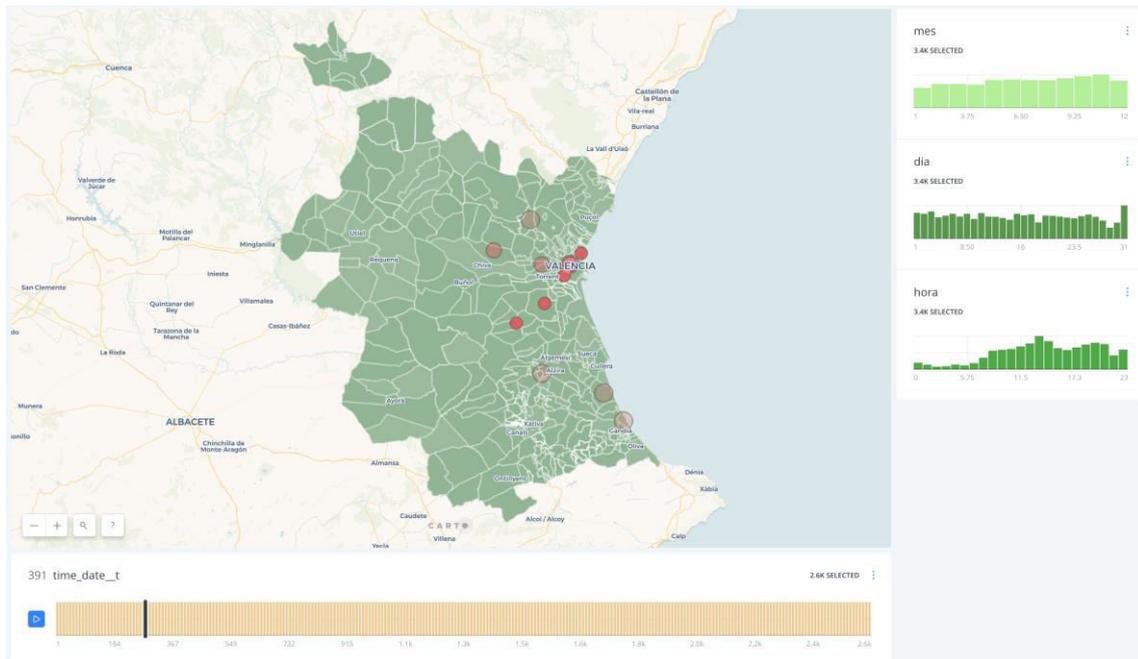


Figura 30. Mapa interactivo del histórico de accidentes en la provincia de Valencia.

6.3. Flujo de Tráfico a la entrada y la salida de Valencia

Para estudiar cómo es el flujo de entrada y de salida de coches de la ciudad de Valencia, es necesario identificar como varía la intensidad de tráfico en la ciudad.

Las series temporales son de gran ayuda para entender datos de flujos, es por ello que se desarrolla una visualización mediante Carto, para así entender mejor los datos.

Como se comenta anteriormente, uno de los análisis de predicción de accidentes toma como foco la periferia de Valencia. Esta es la zona más conflictiva en cuanto a accidentes, es por ello que interesa generar una herramienta que ayude a entender el tráfico en las carreteras A-3, CV-35, CV-36, CV-315, CV-400, V-21, V-30 y V-31.

La herramienta generada tiene dibujada la geometría de las diferentes carreteras desde la salida de Valencia, que es donde se tiene la información de la intensidad de tráfico, hasta la A-7 donde es difícil identificar hacia donde se dirige el tráfico.

Se tiene en cuenta que cuanto más próximo a la periferia de Valencia, la información de tráfico en la carretera es más exacta que para la zona limítrofe con la A-7, sin embargo, se parte de la hipótesis que, aunque el tráfico en un extremo u otro de la zona seleccionada de la carretera no tiene la misma intensidad de tráfico, si cumple con una proporcionalidad relativa.

Para que la visualización evite confusión al ojo humano, no se representa el flujo en la carretera V-30 ya que sigue una geometría no radial como el resto de las carreteras.

Para generar la impresión de volumen de tráfico al ojo humano, se muestran bolas parpadeantes. A mayor número de bolas, mayor es el tráfico de la carretera.

Además, para generar la impresión de entrada o salida de tráfico las bolas se mueven a lo largo de la carretera.

Se han generado dos vistas diferentes que representan la entrada y la salida de flujo de tráfico. Esta herramienta muestra temporalmente la intensidad y permite al usuario entender e interactuar con la herramienta para navegar por horas e interpretar los niveles de intensidad de tráfico.

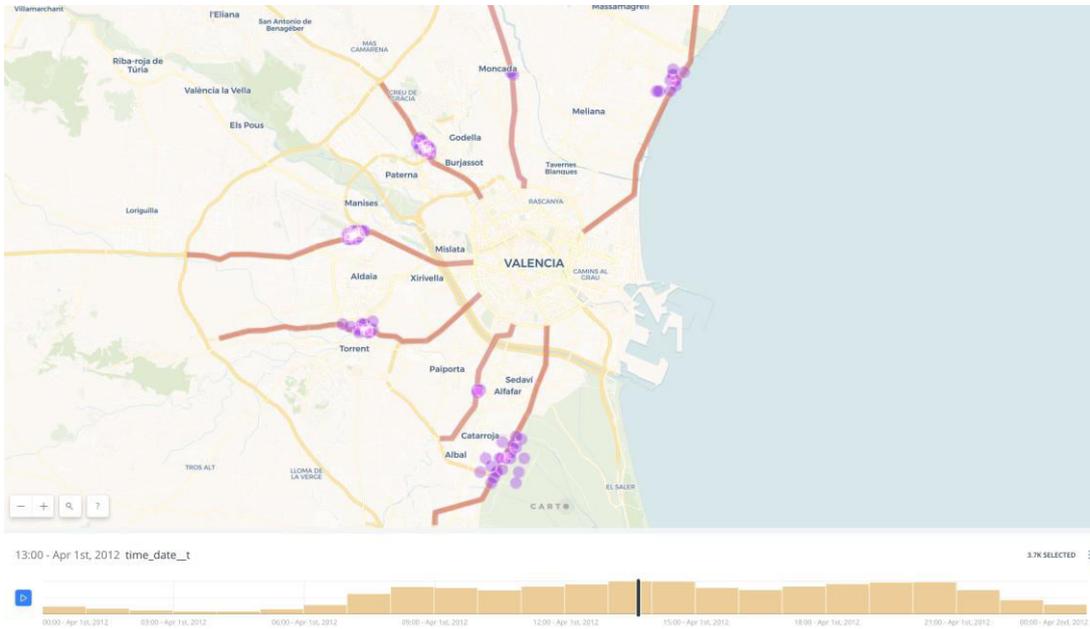


Figura 31. Flujo de salida de tráfico en Valencia.

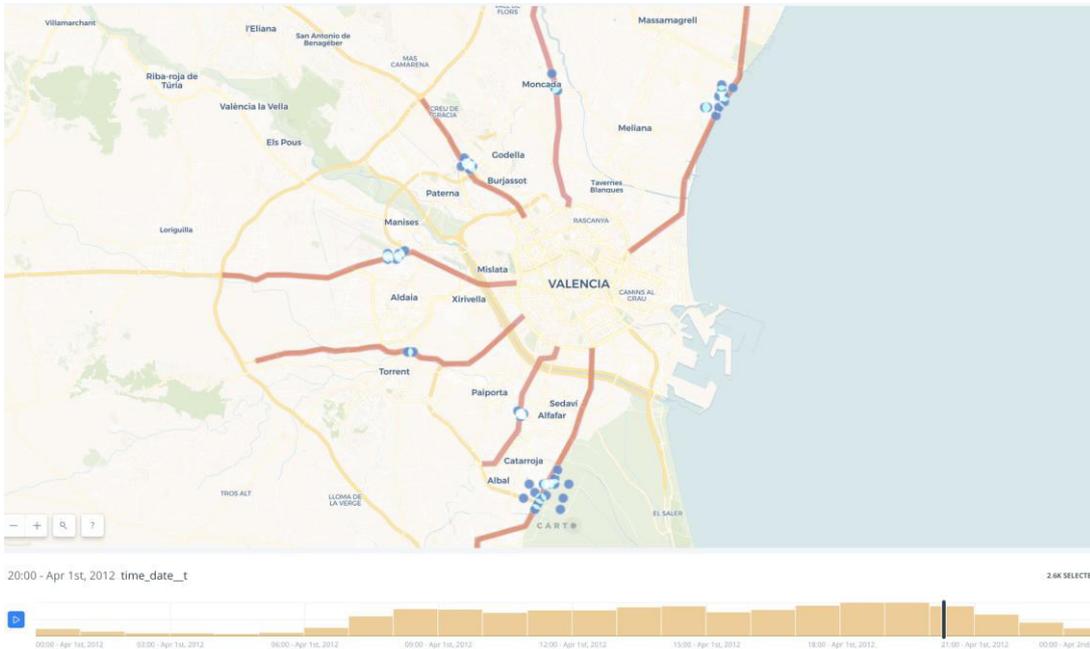


Figura 32. Flujo de entrada de tráfico en Valencia.

El tráfico en días laborables (de lunes a viernes), es diferente al tráfico para fines de semana (sábados y domingos). Si se tiene en cuenta la barra que muestra la serie temporal, se aprecia que el tráfico sigue un patrón de entrada y salida a la ciudad.

A primera hora de la mañana se presenta una intensidad punta muy alta. Esta intensidad disminuye hasta un valor medio que se mantiene relativamente constante hasta alcanzar otro periodo de máxima intensidad a última hora. Tras esto, la intensidad de tráfico registradas desciende considerablemente hasta llegar a mínimos niveles.



Las horas en las que más tráfico hay son las horas de entrada y salida laboral. La intensidad punta se da de 8 a 10h que es la hora de entrada a colegios y al trabajo. A partir de ese momento la evolución depende del tipo de vía y de su emplazamiento, sin embargo, se aprecia un aumento de 2 a 3h que es la hora de la comida y de 18 a 21h que son las horas de salida de trabajo y son las horas en las que las personas tienen tiempo libre para moverse. A partir de las 23h hasta las 6 de la mañana del día siguiente, la intensidad de tráfico es muy baja o casi nula, dándose el valor mínimo entre las 3h y las 5h.

Haciendo un estudio para cada día de la semana, se aprecia que para días laborales el patrón que sigue la intensidad de tráfico es muy similar. Los viernes la intensidad de tráfico de salida es mayor de 18h a 21h y los domingos la intensidad de tráfico de entrada aumenta de 16 a 21h. Durante los fines de semana, la intensidad de tráfico se reduce considerablemente. En las vías urbanas los domingos suelen ser el día de menor tráfico. Sin embargo, en las vías interurbanas el tráfico en sábados y domingos es variable.

Las cifras de intensidades de tráfico en los días festivos son mucho más variables que las que corresponden a días laborables. Las intensidades de tráfico están sujetas a la estación del año (por ejemplo, horario laboral de verano) o a la meteorología, entre otros factores.

6.4. Predicción de accidentes en Valencia y periferia

Con el fin de predecir, gracias al histórico de datos, el número de accidentes, que se producirán en las carreteras periféricas a Valencia, se genera una herramienta que ayude a entender los resultados.

La intención es generar un panel que informe y de opción al usuario a interactuar con este. Se busca un diseño sencillo al ojo humano, que no de opción a mal entendimiento.

Como se busca estudiar la predicción en las calles de Valencia y en su periferia, se generan dos paneles con un mismo formato.

El estudio de las carreteras periféricas de Valencia se centra en las carreteras A-3, CV-35, CV-36, CV-315, CV-400, V-21, V-30 y V-31, ya que son las carreteras más conflictivas y de las que más datos históricos de accidentes se puede conseguir.

La predicción de accidentes en las calles de Valencia se centra en las calles y avenidas en las que se detecta un mayor número de accidentes, dejando de lado las calles que no presentaban un número de accidentes significativo y que permitiese realizar la predicción.

Para generar los paneles, se ha pensado en la información que puede ser útil transmitir al usuario. Es por este motivo que la herramienta permite conocer la predicción de accidentes a tiempo real, dependiendo de las condiciones atmosféricas y los niveles de tráfico. También permite generar una predicción que simule otros escenarios.

Mediante selectores se muestran parámetros de entrada como son el día, hora y meteorología, seleccionando por defecto el caso actual. El tráfico no se ha considerado como un parámetro seleccionable pues es habitual que siga un patrón y no tiene lógica

simular escenarios de tráfico irreales. Para el caso por defecto, la herramienta hace una petición al servicio de predicción meteorológica para conocer el tiempo y de igual modo, la intensidad de tráfico se consigue a tiempo real haciendo una petición a su API.

Como los modelos ya están generados, sobra con sustituir en las formula los parámetros de entrada y la aplicación se encarga de generar y mostrar al usuario la información de salida.

El resultado de la predicción de accidentes se muestra actualizada para cada carretera o calle mediante un código de colores. Verde implica baja posibilidad de accidente, amarillo implica un peligro medio y rojo implica una alta probabilidad de accidentes. Haciendo *click* en la carretera que se desea obtener información se muestra el nombre de dicha carretera o calle y el numero previsto de accidentes para las condiciones de entrada seleccionadas.

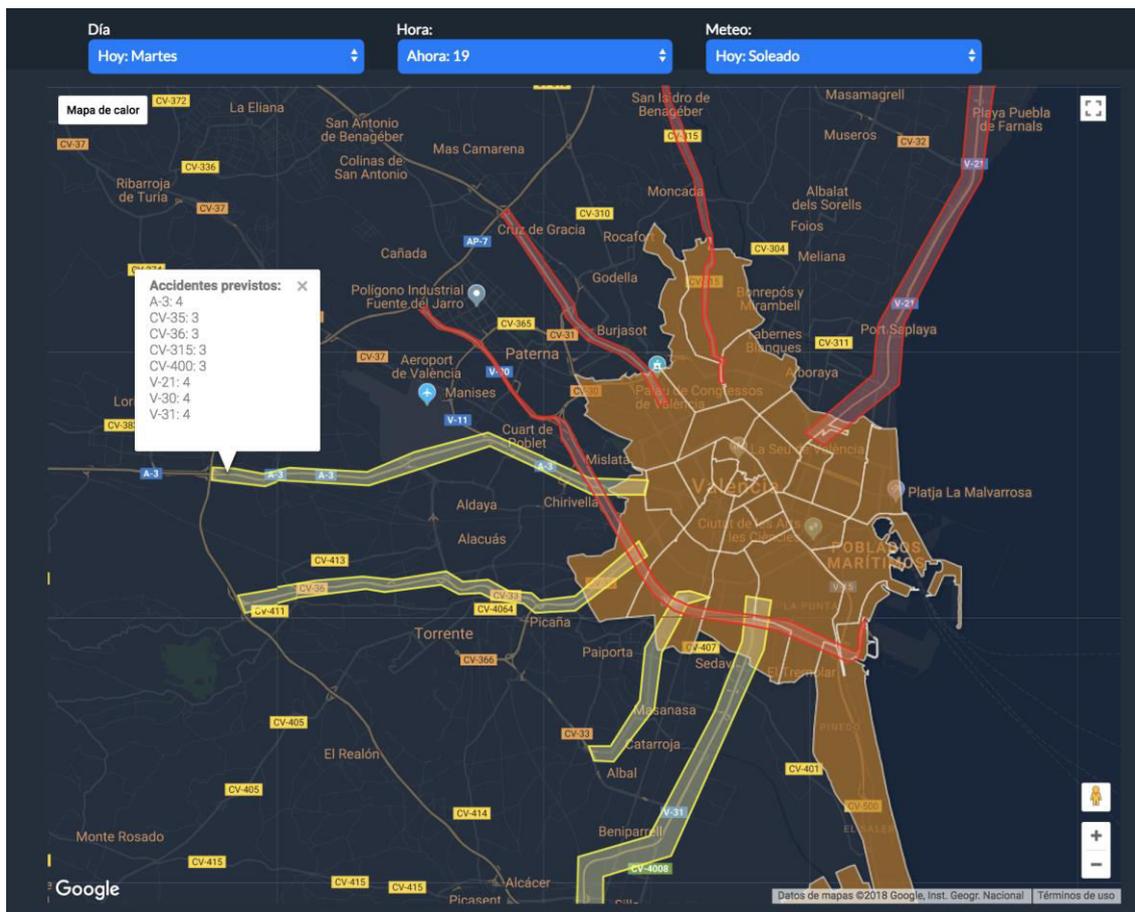


Figura 39. Herramienta para predicción de accidentes en la periferia de Valencia.

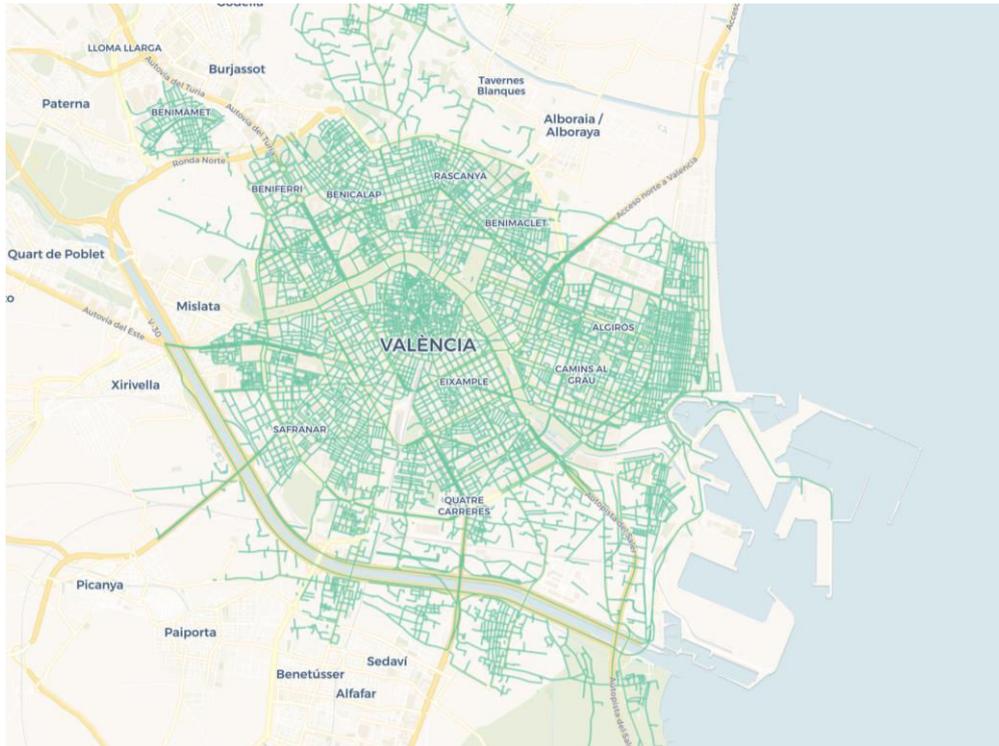


Figura 40. Geometría de las calles de Valencia usadas para la herramienta de predicción de accidentes.

7. Conclusiones

El objetivo principal de este proyecto es la elaboración de una plataforma web que permita al usuario entender de forma sencilla y visual patrones y situaciones propensas a accidentes. La plataforma web generada consta de diferentes análisis y visualizaciones a nivel nacional, autonómico y local, además de una aplicación de predicción del número de accidentes que habrá en las calles de Valencia y en la carreteras periferia de la ciudad, a tiempo real o para un caso hipotético.

Este es un objetivo ambicioso, debido a la gran cantidad de información que se maneja y la dificultad de conseguir herramientas que aporte *feedback* útil y de calidad al usuario final. Para llegar a este objetivo se tiene infinidad de soluciones. Las soluciones elegías, aunque mejorables, cubren el cometido del proyecto.

Tras hacer frente a la recolección de datos de diferentes portales de datos abiertos, se ha encontrado que la calidad de los diferentes conjuntos de datos deja mucho que desear. Aunque en España se haya avanzado en el nivel de transparencia de datos, aún se necesita mucho trabajo que recorrer, al igual que la agilización de trámites electrónicos de solicitud de datos. A esto se le une a la dificultad de encontrar ciertos históricos de datos como los de meteorología que se catalogan como datos abiertos y la mala calidad de los datos de los accidentes de tráfico obtenidos de la DGT que es más que mejorable. La falta de una columna con la información básica del día del mes en que se produjo un accidente, repercute en una alta dificultad de relación entre diferentes fuentes como tráfico y meteorología. Esto obliga a realizar estudios menos precisos, además de mayor complejidad de síntesis para el análisis de los datos.

Para proyectos en los que se depende de disposición de los datos, es necesario saber que el escenario de estudio puede variar dependiendo de estos, por lo que no hay que

cerrar la posibilidad de pivotar el foco del proyecto. En este proyecto se tienen dos fases. En la fase 1 el foco de estudio cambia de un estudio de predicción de accidentes en Valencia a un estudio de predicción de accidentes las carreteras periféricas de Valencia, debido a la falta de datos de accidentes en la ciudad de Valencia. En el momento que se obtienen los datos de registros de accidentes de Valencia se pasa a la fase 2, en la cual hay si se tiene la fecha del accidente y es más sencillo relacionar las diferentes tablas, facilitando así el tratamiento de datos.

La solicitud de datos que deberían ser de carácter abierto pero que no se encuentran disponibles en los catálogos de datos abiertos, requieren de una larga espera tras la solicitud. Además, pueden ser datos no catalogados, en los cuales, para llegar al organismo que los almacena es necesario recorrer varios organismos públicos hasta llegar a esos datos.

Para el estudio de accidentes a nivel nacional, autonómico y local se ha creado un dashboard que permite al usuario conocer de forma cuantitativa mediante gráficos interactivos, el número de accidentes ante un determinado factor externo. Gracias a la interacción con la herramienta generada, el usuario puede deducir si existen patrones de accidentes ante una causa o factor.

Haciendo foco en el estudio de accidentes en Valencia, en un primer momento era necesaria una herramienta para entender cómo se desarrollan los accidentes en la provincia de Valencia. Gracias a la herramienta generada, se muestra que la mayoría de accidentes se centran en la zona costera y en los alrededores de las grandes ciudades.

Para entender como es el flujo de tráfico en la ciudad de Valencia y si tiene relación directa con los accidentes se crea una herramienta que visibiliza el flujo para la entrada y la salida de Valencia, concluyendo que el flujo varía dependiendo de la hora del día.

Además, se deduce que los accidentes siguen una lógica según el día de la semana, el periodo del año en el que nos encontremos y dependiendo de factores externos como pueden ser los factores meteorológicos.

Para generar la herramienta de predicción de accidentes, previamente se generan dos modelos de aprendizaje automático basado en regresión lineal. Estos modelos se basan en diferentes parámetros. El modelo que estudia las carreteras periféricas de Valencia se basa en el día de la semana, la luminosidad, la meteorología y los niveles de tráfico. El modelo generado a partir de estos parámetros tiene un coeficiente de determinación de 0.424. El modelo que estudia las calles de Valencia se basa en el día de la semana, la luminosidad, la meteorología, los niveles de tráfico, los días festivos y el tipo de calle. El modelo generado a partir de estos parámetros tiene un coeficiente de determinación de 0.3.

Estos modelos no predicen cien por ciento de manera fiable el número de accidentes, pero sí que tienen sentido los resultados obtenidos. Los modelos generados son más que mejorables. El resultado de la predicción de accidentes en las calles de Valencia es bastante malo en comparación al modelo obtenido para accidentes en la periferia de Valencia puesto que en los datos se encuentra un desbalanceo de datos.

Como primer paso para mejora de los modelos, es necesario un mayor número de registro de accidentes. Además, es imprescindible probar con más variables de entrada o factores causantes de accidentes para ajustar la predicción a lo que sucede en la realidad.

Para que los modelos sean aplicados, y generen valor al usuario final, se desarrolla una herramienta interactiva que permite conocer la previsión de accidentes en las calles de Valencia y en las carreteras periféricas de la ciudad de Valencia. En esta aplicación

se utilizan datos a tiempo real de meteorología y de tráfico. Además, la aplicación permite conocer predicciones para escenarios hipotéticos.

Finalmente, para que todo el contenido generado tenga visibilidad y pueda ser utilizado, se ha generado una web en la aparecen todos los estudios.

7.1. Futuras mejoras

Para seguir trabajando en este campo se proponen posibles mejoras a realizar con el fin de mejorar los resultados obtenidos y conseguir que este proyecto tenga un impacto positivo en la sociedad.

- **Mejoras del modelo de predicción:** con la intención de conseguir una predicción de accidentes más cercana a lo que ocurre en la realidad, es interesante estudiar si otros factores como curvatura de la carretera, calendario laboral, edad del conductor, el tipo de vía u otros factores que puedan mejorar la predicción de accidentes. Además, puede ser interesante predecir accidentes de tráfico aplicando otros modelos basados en aprendizaje automático.
- **Mejoras en las herramientas:** aunque las herramientas web que se han generado para este estudio cubren con el objetivo del mismo, puede ser interesante añadir nuevas funcionalidades a las mismas, como puede ser mayor interacción entre gráficos o un mayor número de parámetros de estudio dentro de las diferentes herramientas.
- **Análisis más profundo de los datos a través de las herramientas generadas:** las herramientas de visualización generadas en este proyecto muestran de forma sencilla al usuario la interpretación de grandes conjuntos de datos. Parece obvio el significado que tienen las gráficas y los datos, sin embargo, puede ser interesante realizar un estudio más profundo sobre las herramientas generadas.

7.2. Crítica constructiva

Con el fin de ayudar al colectivo de personas encargadas del desarrollo de informes, noticias y aplicaciones relacionadas con los datos, es totalmente necesario el avance en materia de datos abiertos por parte de los diferentes portales de datos abiertos.

Tras la solicitud por vía electrónica de conjuntos de datos de accidentes de la ciudad de Valencia y la larga espera hasta conseguir los datos solicitados, he llegado a la conclusión de que los trámites para solicitar datos que deberían ser abiertos son un lastre a la hora de desarrollar todo tipo de proyectos relacionados con los datos abiertos.

Es necesario concienciar y enseñar a las administraciones y entidades públicas a generar y ofrecer conjuntos de datos, ya sean con carácter nacional, provincial o local y hacer estos conjuntos de datos públicos, permitiendo el uso de estos.

De igual modo, no entiendo la privatización de históricos de datos con carácter público, como son los históricos de datos de meteorología que ofrece en su catálogo la Agencia Estatal de Meteorología. Es necesario que los organismos encargados de la regulación de datos abiertos actúen en la mejora y el control de datos que pertenecen a todos los ciudadanos.

8. Bibliografía

- OMS (2012). ¿Cuál es la enfermedad que causa más muertes en el mundo?. <http://www.who.int/features/qa/18/es/>. [En línea; Recuperado el 20 de Julio de 2018].
- DGT (2017). <http://www.dgt.es/es/prensa/notas-de-prensa/2017/20170103-Los-accidentes-de-trafico-se-cobran-la-vida-de-1160-personas-durante-el-anio-pasado.shtml>. [En línea; Recuperado el 21 de Julio de 2018].
- European union (2015). Smart seat and seatbelt to help sleepy drivers stay alert. *Research*eu Results Magazine*, 42:6–7.
- ANU (2016). <http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/publicaciones/anuario-estadistico-accidentes/>. [En línea; Recuperado el 21 de Julio de 2018].
- González, D. Ú. (2017). Predicción de la severidad de accidentes de tráfico en la red de carreteras de España y Reino Unido mediante modelos estadísticos basados en random forest y regresión logística (Doctoral dissertation, Universidad Miguel Hernández de Elche).
- Cruz-Bellas, L. (2018). Modelos predictivos de accidentes de tráfico en Madrid (Master's thesis).
- Flores, F. S., Mayora, J. P., & Piña, R. J. (2008). Tratamiento de outliers en los modelos de predicción de accidentes de tráfico. In VIII Congreso de Ingeniería del Transporte, España.
- Yuan, Z., Zhou, X., Yang, T., Tamerius, J., & Mantilla, R. (2017, August). Predicting traffic accidents through heterogeneous urban data: A case study. In Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017), Halifax, NS, Canada (Vol. 14).
- MODELER, I. S. (2011). CRISP-DM Guide. Dostupné z: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf. [En línea; Recuperado el 21 de Julio de 2018].
- License, G. G. P. (1991). Free software foundation. License used by the Free Software Foundation for the GNU Project. <http://www.fsf.org/copyleft/gpl.html>. [En línea; Recuperado el 21 de Julio de 2018].
- Team, R. (2015). RStudio: integrated development for R. RStudio, Inc. <http://www.rstudio.com>. [En línea; Recuperado el 21 de Julio de 2018].
- P.E. DGT (s.f.). Portal estadístico de la DGT. https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/. [En línea; Recuperado el 21 de Julio de 2018].
- AEMET (s.f.). Agencia Estatal de Meteorología. <http://www.aemet.es/es/portada>. [En línea; Recuperado el 21 de Julio de 2018].
- AEMET O.D. (s.f.). AEMET OpenData. <https://opendata.aemet.es/centrodedescargas/inicio>. [En línea; Recuperado el 21 de Julio de 2018].
- DataUPV Meteo. (2016). DataUPV. Histórico Meteorológico. <http://dataupv.webs.upv.es/datos-historicos-de-la-observacion-meteorologica-en-valencia/>. [En línea; Recuperado el 21 de Julio de 2018].
- PTDAAV (s.f.). Portal de Transparencia y Datos Abiertos del Ayuntamiento de Valencia. <http://gobiernoabierto.valencia.es/es/>. [En línea; Recuperado el 21 de Julio de 2018].



- Trafico PTDAAV (s.f.). Valencia Datos Abiertos. Intensidad de los Puntos de Medida de Tráfico (Espiras electromagnéticas). <http://gobiernoabierto.valencia.es/es/resource/?ds=puntos-medida-trafico-espiras-electromagneticas&id=23bd3e1e-1d80-44d4-81cc-c33746d1f445>. [En línea; Recuperado el 21 de Julio de 2018].
- Trafico DataUPV (2016). DataUPV. Histórico de tráfico. <http://dataupv.webs.upv.es/datos-historicos-de-trafico-en-valencia/>. [En línea; Recuperado el 21 de Julio de 2018].
- INE (s.f.). Instituto Nacional de Estadística - INE. <http://www.ine.es/welcome.shtml>. [En línea; Recuperado el 21 de Julio de 2018].
- IGN (s.f.). Instituto Geográfico Nacional - IGN. <http://www.ign.es/>. [En línea; Recuperado el 21 de Julio de 2018].
- IGN Data (s.f.). Conjunto de datos IGN. <http://www.ign.es/csw-inspire/srv/eng/conjuntodatos>. [En línea; Recuperado el 21 de Julio de 2018].
- Portal estadístico (s.f.). Portal de estadística del ayuntamiento de Valencia. <https://www.valencia.es/ayuntamiento/estadistica.nsf>. [En línea; Recuperado el 10 de septiembre de 2018].
- Policía Local (s.f.). Web de la policía local de Valencia. http://www.valencia.es/ayuntamiento/Infocidad_accesible.nsf/vDocumentsWebListado/A1CF2CEC108A67A7C12572C20023E378. [En línea; Recuperado el 10 de septiembre de 2018].
- Sede Electronica (s.f.). Conjunto de datos la Sede electrónica del ayuntamiento de Valencia. <https://sede.valencia.es>. [En línea; Recuperado el 10 de septiembre de 2018].
- Solucitud (s.f.). Conjunto de datos la Sede electrónica del ayuntamiento de Valencia. https://sede.valencia.es/sede/registro/procedimiento/WEB_ASSET_3_2012000028268. [En línea; Recuperado el 10 de septiembre de 2018].

9. Anexos

9.1. Descripción de columnas

A continuación, se detallan los campos que se han utilizado en el desarrollo de este proyecto. Las columnas que no se han utilizado no se listan en este apartado, aunque tengan calidad, ya que no se han usado para los diferentes estudios.

Se detalla la tabla y las diferentes columnas de las tablas. Para cada columna se detallan las diferentes variables que albergan.

Tabla “*TABLA_ACCVICT*”:

- ID_ACIDENTE: numérico, no nulo y clave.
- ANIO: Formato ‘YYYY’.
- MES: 1 (enero), 2 (febrero), 3 (marzo), 4 (abril), 5 (mayo), 6 (junio), 7 (julio), 8 (agosto), 9 (septiembre), 10 (octubre), 11 (noviembre), 12 (diciembre).
- HORA: valor numérico comprendido entre 0 y 23.
- DIASEMANA: 1 (lunes), 2 (martes), 3 (miércoles), 4 (jueves), 5 (viernes), 6 (sábado), 7 (domingo).
- PROVINCIA: 1 (Araba/Álava), 2 (Albacete), 3 (Alicante/Alacant), 4 (Almería), 5 (Ávila), 6 (Badajoz), 7 (Balears, Illes), 8 (Barcelona), 9 (Burgos), 10 (Cáceres), 11 (Cádiz), 12 (Castellón/Castelló), 13 (Ciudad Real), 14 (Córdoba), 15 (Coruña, A), 16 (Cuenca), 17 (Girona), 18 (Granada), 19 (Guadalajara), 20 (Gipuzkoa), 21 (Huelva), 22 (Huesca), 23 (Jaén), 24 (León), 25 (Lleida), 26 (Rioja, La), 27 (Lugo), 28 (Madrid), 29 (Málaga), 30 (Murcia), 31 (Navarra), 32 (Ourense), 33 (Asturias), 34 (Palencia), 35 (Palmas, Las), 36 (Pontevedra), 37 (Salamanca), 38 (S.C.Tenerife), 39 (Cantabria), 40 (Segovia), 41 (Sevilla), 42 (Soria), 43 (Tarragona), 44 (Teruel), 45 (Toledo), 46 (Valencia/València), 47 (Valladolid), 48 (Bizkaia), 49 (Zamora), 50 (Zaragoza), 51 (Ceuta), 52 (Melilla).
- COMUNIDAD_AUTONOMA: 1 (Andalucía), 2 (Aragón), 3 (Asturias, Principado de), 4 (Balears, Illes), 5 (Canarias), 6 (Cantabria), 7 (Castilla y León), 8 (Castilla-La Mancha), 9 (Cataluña), 10 (Comunitat Valenciana), 11 (Extremadura), 12 (Galicia), 13 (Madrid, Comunidad de), 14 (Murcia, Región de), 15 (Navarra, Comunidad Foral de), 16 (Rioja, La), 17 (País Vasco), 18 (Ceuta y Melilla),
- COD_MUNICIPIO: Hace referencia a el código municipal, diferente al código postal. Todos los municipios de España tienen 2 códigos diferentes de 5 dígitos.
- ZONA: 1 (carretera), 2 (zona urbana), 3 (travesía).
- ZONA_AGRUPADA: 1 (vías interurbanas), 2 (vías urbanas).
- CARRETERA: nombre de la carretera. ejemplo: a-7, cv-35, etc.
- TIPO_VIA: 1 (autopista), 2 (autovía), 3 (vía para automóviles), 4 (vía convencional con carril lento), 5 (vía convencional), 6 (camino vecinal), 7 (vía de servicio), 8 (ramal de enlace).
- TRAZADO_NO_INTERSEC: 1 (recta), 2 (curva suave), 3 (curva fuerte sin señalizar), 4 (curva fuerte con señal y sin velocidad señalizada), 5 (curva fuerte con señal y velocidad señalizada).

Estudio sobre patrones de accidentes en la ciudad de València

- TIPO_INTERSEC: 1 (n T ó Y), 2 (n X ó +), 3 (enlace de entrada), 4 (enlace de salida), 5 (giratoria).
- SUPERFICIE_CALZADA: 1 (seca y limpia), 2 (umbría), 3 (mojada), 4 (helada), 5 (nevada), 6 (barrillo), 7 (gravilla suelta), 8 (aceite).
- LUMINOSIDAD: 1 (pleno día), 2 (crepúsculo), 3 (noche: iluminación suficiente), 4 (noche: iluminación insuficiente), 5 (noche: sin iluminación).
- FACTORES_ATMOSFERICOS: 1 (buen tiempo), 2 (niebla intensa), 3 (niebla ligera), 4 (lloviznando), 5 (lluvia fuerte), 6 (granizando), 7 (nevando), 8 (viento fuerte).

Tabla “TABLA_PER.”:

- ID_ACIDENTE: numérico, no nulo y clave.
- ID_VEHICULO: numérico, no nulo y clave.
- ID_PERSONA: numérico, no nulo y clave.
- EDAD: numérico entre 18 y 99.
- SEXO: 1 (mujer), 2 (hombre).
- ANIO_PERMISO: numérico entre 0 y 99

Tabla “TABLA_VEHIC”:

- ID_ACIDENTE: numérico, no nulo y clave.
- ID_VEHICULO: numérico, no nulo y clave.
- ANIO_MATRICULA_VEHICULO: Formato ‘YYYY’.

Tabla “Tabla_Meteorología”:

- ANIO: Formato ‘YYYY’.
- MES: Formato ‘MM’.
- DIA: Formato ‘DD’.
- HORA: valor numérico comprendido entre 0 y 23.
- Weekday: 1 (Monday), 2 (Tuesday), 3 (Wednesday), 4 (Thursday), 5 (Friday), 6 (Saturday), 7 (Sunday).
- VelMax.m.s: Velocidad de viento máximas. Campo numérico, con 1 decimal.
- Vel.m.s: Velocidad viento. Campo numérico, con 1 decimal.
- Temp.c: Temperatura: °C. Campo numérico, con 1 decimal.
- HRel.HR: Humedad Relativa. Campo numérico, con 1 decimal.
- Pres.mb: Presión atmosférica. Campo numérico, con 1 decimal.
- Precip.l.m2: Precipitaciones. Campo numérico, con 1 decimal.

Tabla “Tabla_tráfico”:

- Punto: Idpm. Identificador de punto de medida.
- ANIO: Formato ‘YYYY’.
- MES: Formato ‘MM’.
- DIA: Formato ‘DD’.
- HORA: valor numérico comprendido entre 0 y 23.
- Intensidad: Ih. Intensidad hora de vehículos (vehículos/hora).

Tabla “Tabla_datos_municipales”:

- CodProvincia: Campo numérico comprendido entre 01 y 52.
- CodMunicipio: Campo numérico. Específico para cada código de provincia.
- CodigoPostal: Campo numérico referente al código postal del municipio.

- Municipio: Nombre del municipio.
- Cod_Municipio: Campo numérico generado a través del CodProvincia y CodMunicipio. Es diferente al código postal y es el código que anota tráfico para cada accidente.

Tabla “*Tabla_cartográfica*”:

- the_geom: Geometría referente. Listado de coordenadas, vértices o nodos que delimitan un polígono.
- cod_ccaa / cod_mun / cod_prov: Campo numérico. Campo numérico referente al código postal del municipio, código de provincia o código de comunidad autónoma.
- nom_ccaa / nom_mun / nom_prov: Nombre del municipio, comunidad autónoma o provincia.

Tabla “*Tabla_Red_Attestados_Valencia*”:

- ID_ACIDENTE: numérico, no nulo y clave.
- ANIO: Formato ‘YYYY’.
- MES: 1 (enero), 2 (febrero), 3 (marzo), 4 (abril), 5 (mayo), 6 (junio), 7 (julio), 8 (agosto), 9 (septiembre), 10 (octubre), 11 (noviembre), 12 (diciembre).
- HORA: valor numérico comprendido entre 0 y 23.
- DIASEMANA: 1 (lunes), 2 (martes), 3 (miércoles), 4 (jueves), 5 (viernes), 6 (sábado), 7 (domingo).
- LUGAR: Nombre de la calle, avenida, plaza, cruce o carretera. Incluye el número en algunos registros.
- MOTIVO: Separa los registros dependiendo de si se deben a choques, atropellos, colisiones, salida de vía, caídas, etc.
- UNIDAD: Define si el accidente es debido por un vehículo, ciclomotor, bicicleta, peatón, mobiliario, etc.
- OCUPANTES: define el número de personas implicadas en el accidente.
- HOSPITAL: Centro médico al que se traslada el herido tras el accidente.
- TURNO: Define si es por la mañana, tarde o noche.
- ALCOHOLEMIA: Define el grado de alcoholemia en sangre.
- PAÍS: Define el país de origen.
- HOSPITAL: Centro médico al que se traslada el herido tras el accidente.
- TURNO: Define si es por la mañana, tarde o noche.