METODOLOGÍA DE MINERÍA DE DATOS PARA EL ESTUDIO DE TABLAS DE SINIESTRALIDAD VÍAL

G. Villarino[1], D. Gómez[1], R. Cintas[1], J. T. Rodríguez[2]
[1] Dpto. de Estadística e Investigación Operativa III, Universidad Complutense de Madrid,
[2] Dpto. de Estadística e Investigación Operativa I, Universidad Complutense de Madrid
[1]gvillari@ucm.es

Resumen. La reducción de la siniestralidad en nuestras carreteras es un problema complejo, dado que incluye elementos de naturaleza muy diversa, por lo que la solución no es simple. Es en este contexto donde tienen cabida las modernas técnicas estadísticas de clasificación, que constituyen una aportación sustancial al conocimiento no sólo cuantitativo sino también cualitativo de la situación actual. El objetivo general de este trabajo es el desarrollo de una metodología de estudio que aborda aspectos como la creación de subpoblaciones de interés, el pre procesamiento y visualización de los datos, la determinación de los factores de riesgo y la creación de distintos modelos de clasificación para la gravedad de las lesiones producidas.

Palabras Clave: Clasificación supervisada, machine learning, siniestralidad vial, información bipolar.

1 Introducción

La accidentalidad en las carreteras ha sido, desde la generalización del uso de vehículos a motor, una de las principales causas de muerte en España y por ello es foco de gran preocupación para la sociedad y sus autoridades.

Han pasado muchos años desde el máximo histórico de fallecidos en accidentes de tráfico en España de 1989. Aquel aciago año, último de la década de los ochenta y en pleno aumento del parque de vehículos automóviles, 9.344 personas perdieron su vida en un accidente de tráfico. Entonces no se llegaba a 15 millones de vehículos en total.

Debido a los esfuerzos realizados y a la gran cantidad de recursos destinados, estas elevadas cifras han experimentado, afortunadamente, un descenso muy significativo. Entre las causas de esta disminución se encuentran la mayor concienciación de la sociedad en materia vial, la mejora de las infraestructuras de la red de transportes, los cambios legislativos, los grandes avances en materia de seguridad de los vehículos automóviles y de detección de infracciones mediante cinemómetros y cámaras de seguridad.
Con el principal objetivo de tratar de analizar la gravedad de los accidentes de tráfico y entender sus principales causas, en este trabajo se presenta una metodología para el estudio de datos de siniestralidad vial en España pasando por algunas de las fases clásicas de análisis de datos: pre-procesamiento, detección de factores de riesgo en la mortalidad de los accidentes en carretera y la clasificación supervisada del evento de interés, el de muerte del accidentado a 30 días del siniestro. Para finalizar este trabajo, se presentan algunas líneas de futuro a seguir como es la incorporación de información de carácter bipolar en árboles de decisión para la clasificación de variable

respuesta categórica que determina la gravedad de las lesiones en el accidentado en el momento del accidente y presenta cuatro categorías.

2 Objetivos

La fuente de información del presente estudio la constituye la base de datos de accidentes ocurridos en el año 2012 en España, proporcionada por la Dirección General de Tráfico (DGT). Dicha base de datos está integrada por 83.115 registros que se corresponden con los accidentes con víctimas notificados por los diferentes cuerpos de policía y guardia civil y 202.804 registros referentes a datos específicos de las propias víctimas. Mediante la integración de esta información se crea un único conjunto de datos cuyos registros representan a los accidentados y contiene la información tanto de éstos como de las circunstancias del siniestro en el que estuvieron implicados.

El objetivo principal de este trabajo es la creación de una metodología para el estudio de una base de datos de accidentalidad vial a partir de un procedimiento semi-automático para facilitar el tratamiento de las tablas de datos de esta naturaleza que, con periodicidad anual, son recogidas por la Dirección General de Tráfico. Esta metodología se puede dividir en las etapas secuenciales siguientes:

- Preprocesamiento de los datos
  - Proponer un procedimiento para el estudio y recodificación de las variables que históricamente se consideran de influencia en la gravedad de las lesiones producidas.
  - Determinar las subpoblaciones objetivo mediante estudio de la posible segmentación por tipo de vehículo (bicicletas, motos, camiones, turismos, autobuses...) o por tipo de víctima (peatón, conductor, pasajero...).
- Determinación de factores de riesgo
  - Identificar y evaluar los factores de riesgo influyentes en los accidentes con víctimas mortales en las distintas subpoblaciones.
  - Determinar perfiles de víctimas y escenarios de accidentalidad como resultado de la integración de distintas técnicas de clasificación.
  - Abordar el problema de clasificación de las clases poco representadas. La proporción del evento considerado, la muerte a 30 días del siniestro, es únicamente del 1,2%.
- Modelos de clasificación en minería de datos
  - Crear funciones para facilitar el ajuste de algoritmos de aprendizaje a los datos para la clasificación supervisada del evento comentado y valoración de los resultados.
  - Establecer una comparativa de bondad de ajuste entre distintos métodos de clasificación en minería de datos tales como Random Forest, Gradient Boosting y Neural Nets, en la clasificación de los fallecidos en las distintas subpoblaciones de víctimas.

En definitiva se pretende programar un procedimiento mediante el cual sea posible extraer conclusiones de los datos de forma semiautomática y con la máxima fiabilidad posible. Hay que observar en este punto que se han utilizado únicamente los datos del año 2012 por lo que quedaría abierta la línea de investigación para el tratamiento de datos de naturaleza longitudinal.
En cualquier caso, se considera que esta metodología proporciona una base sólida para obtener buenos resultados en los conjuntos de datos de accidentalidad vial recogidos de esta forma.

3 Metodología

Para alcanzar los objetivos planteados se recurre a una metodología que incluye muchos aspectos del tratamiento de datos que se detallan a continuación.

3.1 Preprocesamiento

Para un correcto análisis del problema presentado, la fase previa a un análisis estadístico es una etapa de depuración de los datos, cuyo objetivo es minimizar el 'ruido' que ciertas distribuciones de variables pueden introducir en los modelos, con la consecuente pérdida de precisión e incluso la posible obtención de conclusiones erróneas. Debido a esto se realiza una cuidadosa labor de examen y depuración de las variables que resultarán de interés a lo largo del estudio. En cuanto a los métodos de recodificación de variables, por un lado y ya que el objetivo fundamental es preparar el conjunto de datos y sus variables para un análisis de regresión logística o clasificación binaria en general, se han llevado a cabo uniones de niveles de variables de naturaleza categórica que presentaban excesivo número de niveles o categorías irrelevantes, mediante el examen de los resultados de modelos de regresión logística univariante en los que la variable respuesta es el suceso de interés de este estudio, la variable *muerte a 30 días*[1] de naturaleza dicotómica, y la variable explicativa es la variable a recodificar. De esta forma se unifican categorías no significativas por sí mismas y con parámetros estimados de igual signo, ya que su influencia en el evento se da en el mismo sentido. El otro procedimiento de recategorización de variables utilizado han sido los Arboles de clasificación CHAID[2] (CHi-squared Automatic Interaction Detection).

Tabla 1: Distribución de variables en las subpoblaciones

		Camiones	Bicis	Motos	Ciclomotores	Peatones	Turismos
muert30	*% Si*	1,23	1,28	1,41	0,74	3,1	0,64
accseg	*%No*	7,92	32,94	6,54	5,37	39,26	8,02
alcohol	*% Si*	2	1,05	1,98	2,67	2,59	5,31
infracc	*% Si*	35,97	46,03	47,78	50,46	41,08	38,59
distracc	*% Si*	46,93	33,15	28,47	30,11	37,27	42,83
velina	*% Si*	10,48	5,58	8,84	5,2	13,53	9,99
Num. Registros		5512	5535	20716	8535	11504	134878

En la tabla 1 se presentan las distribuciones de las variables[3] propias del conductor a modo de comparativa entre las distintas subpoblaciones para ilustrar la diferencia de distribuciones que justifica el estudio por separado. Se observan las variables (número de niveles), *Muerte a 30 días* (2), *Accesorios de seguridad* (2), *Alcohol* (2), *Infracción* (2), *Distracción* (2) y *Velocidad Inadecuada* (2) A las variables comentadas se une la *Edad* (4) que se recategoriza en cuatro

[1] La variable *muerte a 30 días* refleja el hecho constatado del fallecimiento de la víctima de accidente de tráfico pasados 30 días del mismo.
[2] La metodología CHAID fue desarrollada en el año 1980 en Sudáfrica por Gordon V. Kass en su tesis PhD.
[3] Para mayor información sobre las variables del estudio consúltese la versión extendida del trabajo en http://eprints.sim.ucm.es/34870/1/TFM_GuillermoVillarino_Nov_2015.pdf

tramos. Se pone de manifiesto la bajísima incidencia del evento de muerte a 30 días en la población, lo que augura una etapa de clasificación con complicaciones aseguradas.

En lo referente a los factores propios de la vía se han considerado los elementos presentes en ésta que pueden resultar peligrosos en los siniestros de vehículos de dos ruedas como *Mediana entre calzadas* (2), *Barrera de seguridad* (2), *Paneles direccionales* (2), *Hitos de arista* (2), *Captafaros* (2) o estado de la *Superficie* (4), teniendo también en cuenta variables como *Tipo* (3) *y Titularidad* (4) *de la vía, Densidad de circulación* (4) *y Zona* (4). También se consideran variables como *Factores Atmosféricos* (5), *Tipo de Accidente* (10), *Luminosidad* (4) o *Tipo de Día* (4).

3.2 Técnicas de clasificación empleadas

En lo que se refiere a modelos de clasificación, se presta especial atención a la regresión logística como la técnica clásica debido a su base estadística y a la posibilidad de cuantificar el efecto de las variables sobre la respuesta mediante los odds ratio, frente a la metodología utilizada por otros algoritmos de aprendizaje estadístico y computacional. A continuación se enumeran las distintas técnicas utilizadas:

La Regresión Logística (Hosmer & Lemeshow, 2000) cuya gran ventaja es la posibilidad de cuantificar los efectos de los predictores sobre la respuesta a través de los *Odds Ratio*. Como algoritmos de machine learning[4] se han utilizado: Neural Nets (NNet) (Ripley, 1996), Random Forest (RF) (Breiman, 2001), Gradient Boosting (GB) (Friedman, 2001), Extreme Gradient Boosting[5] (XGB), Boosted Logistic Regression (LogiBoost) (Friedman et al., 2000) y finalmente los Bayesian Generalized Linear Models (BayesGLM) (Gelman, 2009), todos ellos disponibles en el paquete Caret (Kuhn, 2008) de R.

En todas las técnicas de minería de datos se ha aplicado el método de validación cruzada repetida que consiste en dividir, mediante partición aleatoria, el archivo en n partes construyendo el modelo con n-1 de ellas y reservando la restante para la validación de los resultados obtenidos, con lo que finalmente para cada repetición del algoritmo se ajustarán n modelos validados sobre observaciones 'nuevas'. Este proceso se repite m veces consiguiendo por lo tanto m x n modelos ajustados con conjuntos de entrenamiento distinto y validado sobre observaciones no utilizadas en su construcción. Se ha elegido n = 3 y m = 4 en este estudio.

3.3 Ensamblado de modelos

Con el objetivo de mejorar la precisión alcanzada por los modelos de clasificación empleados en el estudio y reducir la varianza de los errores cometidos, se proponen distintos métodos de ensamble de clasificadores mediante la técnica de *stacking*.
Este método consiste en construir clasificadores dados por la combinación, lineal o no, de las probabilidades estimadas por los modelos ajustados, algunos de los cuales son ensambles en sí mismos (Random Forest, Gradient Boosting). Con ello se consiguen las probabilidades estimadas conjuntas y se realiza la clasificación mediante la técnica del punto de corte óptimo de la probabilidad estimada.

[4] Para mayor información sobre el ajuste de parámetros de los algoritmos consúltese la versión extendida del trabajo en http://eprints.sim.ucm.es/34870/1/TFM_GuillermoVillarino_Nov_2015.pdf
[5] http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf

Cabe destacar que para obtener mejores resultados es conveniente realizar un estudio de correlaciones entre las predicciones para descartar los ensambles de probabilidades estimadas altamente correladas que usualmente no proporcionan mejora respecto al mejor modelo (Dzeroski, 2004).

4 Factores de influencia en la mortalidad

Como resumen de este epígrafe y tras el proceso de análisis de la importancia de las variables en los distintos modelos ajustados se propone una medida de influencia de los distintos factores sobre el resultado fatal en accidentes de tráfico en las subpoblaciones de víctimas de siniestros viales en España en el año 2012.
Se construye, para cada subpoblación del estudio, una tabla que contienen las cinco variables más relevantes en cada uno de los modelos ajustados y se realiza un conteo de las frecuencias relativas de aparición de cada variable en el *top 5* de la medida de importancia a lo largo de los ocho modelos. La idea fundamental es que las variables que aparecen como importantes en los distintos modelos con mayor frecuencia han de ser los factores que mayor influencia tienen sobre el suceso de interés al haber sido seleccionados por distintos algoritmos para crear los modelos de clasificación.
En la tabla 2 se presentan los perfiles de víctimas y escenarios de accidentalidad extraídos mediante este procedimiento para cada una de las subpoblaciones de interés.

Tabla 2: Perfiles de víctimas y escenarios de accidentalidad por subpoblaciones

	Perfil	Escenario
Camiones	Edad(38,47]\Acc.Seg\Distracc	Noche\Vuelco\Sal. Izq\Barrera\Mediana\Hitos
Bicis	Edad(>57)\Distracc\Infracc\Acc.Seg	Festivo\Dens.Cir\Barrera
Motos	Edad(38,47]\Vel.Inadec\Infracc	Colis.Front\Sal.Izq\Dens.Cir\Barrera
Ciclomotores	Edad(>38)\Acc.seg\Infracc\Distracc	Festivo\Dens.Cir\Zona Urbana
Peatones	Edad(>57)\Distracc\Acc.Seg	Noche\Festivo\Dens.Cir\Zona Urb.\Hitos\Barrera
Turismos	Edad(>57)\Distracc\Acc.Seg	Dens.Cir\Colis.Frontal\Hitos\Barrera

Se han puesto de manifiesto, mediante esta metodología, los factores tanto propios de la vía como inherentes al conductor que presentan una mayor influencia en la clasificación del evento de interés en el estudio.Cabe destacar las ventajas e inconvenientes de este método para la selección de factores de influencia ya que por un lado supone un método robusto para esta tarea debido a que es un compendio de técnicas, y no un solo algoritmo, el que ha decidido extraer esa variables como influyente, evitando así posibles fallos o sesgos en la selección de variables de cada uno de ellos de manera individual.
Por otra parte, la principal desventaja de este método es la imposibilidad de cuantificar la importancia de estas variables así como el sentido de influencia en la clasificación del evento de interés debido a que se seleccionan para realizar la partición digamos en un nodo (en el caso de los métodos basados en árboles) pero es difícil saber si esa categoría de esa variable desemboca en un aumento o en un detrimento de la probabilidad estimada. Este hecho no tiene especial importancia ya que se dispone de métodos complementarios como la inspección descriptiva de la población y

la interpretación de los OR de la regresión logística, con los que se puede decidir ese sentido de influencia

5 Capacidad de clasificación

En este caso, debido a la baja incidencia del suceso de interés en la población, las probabilidades estimadas suelen ser bajas y esto hace que el punto de corte que maximiza la relación entre sensibilidad y especificidad de la clasificación no sea el 0.5.
Por ello y para una mejor clasificación, se programan funciones de predicciones que generan las matices de confusión y se recurre a la función ROC para dibujar la curva y estimar el punto de corte óptimo para la probabilidad, aquel que hace máxima la relación entre sensibilidad y especificidad para la probabilidad estimada frente a la clase real. Se comprueba que existe poca diferencia entre este valor estimado y la prevalencia a priori del evento que, sin embargo producen grandes cambios en número de mal clasificados. Este hecho es habitual en conjuntos de datos no balanceados debido a las bajas probabilidades estimadas que crean una frontera de decisión muy dispersa entre las clases a predecir y esta poca consistencia de la estimación del punto de corte es el mayor inconveniente de la utilización de este método.
En cualquier caso se considera muy superior la capacidad de clasificación del evento por medio de este procedimiento, máxime al tratarse un suceso fatal que ha de ser evitado. Es lógico actuar bajo la premisa de que las consecuencias de la mala clasificación de los registros de la clase de interés resultan de mucha mayor gravedad y por ello ha de relajarse el umbral para la especificidad incurriendo en una mayor tasa de falsos positivos.
Realizando este procedimiento para todos los modelos en todas las subpoblaciones se obtienen los valores del estadístico c o área bajo la curva ROC para el punto de corte óptimo de la probabilidad estimada. En total se han realizado ocho predicciones para cada una de las seis subpoblaciones, con lo que se han ajustado cuarentayocho modelos finales en estos datos escogidos de entre cientos probados.

Tabla 3: Comparativa de precisión (ROC) global. Punto de corte óptimo.

	Logistica	LogiBoost	NNet	RF100	RF500	GBM	XGB	Bayes	Media
Camiones	0,84	0,79	0,85	0,7	0,76	0,85	**0,91**	0,86	**0,82**
Motos	0,88	0,83	**0,93**	0,74	0,78	0,87	0,9	0,88	**0,85**
Bicis	0,9	0,84	**0,94**	0,74	0,82	0,92	0,93	0,91	**0,88**
Ciclos	0,86	0,78	**0,91**	0,76	0,79	0,87	0,89	0,86	**0,84**
Peatones	0,88	0,85	**0,96**	0,66	0,75	0,93	0,95	0,9	**0,86**
Turismos	0,89	0,86	0,94	0,72	0,77	0,93	**0,95**	0,9	**0,87**
Media	**0,88**	**0,83**	**0,92**	**0,72**	**0,78**	**0,90**	**0,92**	**0,89**	

A la vista de los resultados obtenidos, NNet es el algoritmo que mejor ajusta la clasificación en las subpoblaciones de Motos bicis Peatones y ciclomotores y XGB lo hace en Camiones con mucha diferencia, y en Turismos con no tanta. En general se considera que el ajuste de los algoritmos de

clasificación utilizados es muy bueno, alcanzando valores de area bajo la curva ROC superiores al 0.9.Sin embargo, al tratarse de una población con clases no balanceadas es importante evaluar la sensibilidad y especificidad de la clasificación.

Tabla 4: Medidas de ajuste para el mejor modelo de cada subpoblación.

	ROC	Sens.	Espec.
Camiones	0,91	**0,91**	0,78
Motos	0,93	0,87	0,87
Bicis	0,94	**0,93**	0,83
Ciclomotores	0,91	**0,95**	0,84
Peatones	0,96	0,85	0,81
Turismos	0,95	**0,93**	0,85
Media	**0,93**	**0,91**	**0,83**

La información de la tabla 3 refleja la elevada capacidad de los modelos ajustados para clasificar a los verdaderos eventos, con una sensibilidad que supera el 85% en todos los casos llegando, en el mejor de ellos (subpoblación de ciclomotores) al 95%.
La elevada capacidad de clasificación de los modelos propuestos justifica la robustez del método de extracción de factores de riesgo comentado en el anterior epígrafe.

6 Ensamblado de modelos

Como último apartado del estudio, se proponen aquí algunos métodos de ensamble de los modelos anteriormente ajustados con el fin de obtener clasificadores cuya relación entre sensibilidad y especificidad sea mayor que la proporcionada por los modelos individuales. Para ello se construye un conjunto de datos que contiene las probabilidades estimadas por los ocho modelos ajustados para cada subpoblación con el fin de combinar estas probabilidades de distintas formas. Sin ánimo de profundizar en los modelos de ensamble óptimos, se proponen varios de estos posibles clasificadores combinados y se compara su capacidad.

En primer lugar se construye un clasificador dado por la media aritmética de las probabilidades estimadas por cada uno de los modelos individuales, llamado **ensamble medio (EnsMean)**. A continuación se realizará un ajuste de regresión logística por pasos, backward, con las ocho probabilidades estimadas como predictores para la clasificación del evento y se construirá un clasificador que viene dado por la media ponderada por pesos obtenidos por los coeficientes de la regresión obtenidos, de forma relativa. Este clasificador se llamará **ensamble regresión (EnsRegw)**. Así mismo se considerará la probabilidad estimada de este modelo de regresión logística como otro posible **ensamble logístico (EnsRegPred)**.

Por último se construye un clásico ensamble dado por la media ponderada de Gradient Boosting y Random Forest que se considera interesante debido a las distintas formas de actuación de sendos algoritmos, estando el primero orientado a reducir el sesgo de las estimaciones y con la ventaja de la selección por sorteo de variables del segundo. Se consideran, tras diversas pruebas, los pesos de 0,8 y 0,2 respectivamente (**Ens2080**).

Antes de crear los ensambles se realiza un estudio de correlaciones entre las probabilidades estimadas ya que conviene integrar los resultados de los modelos que presente mayor independencia para contrarrestar errores de clasificación.

Tabla 5: Comparativa de precisión (ROC) para los modelos de ensamble.

	EnsRegPred	EnsRegw	EnsMean	Ens2080
Camiones	0,85	0,89	0,87	0,85
Motos	0,93	0,91	0,9	0,89
Bicis	0,92	0,93	0,92	0,92
Ciclomotores	**0,96**	0,94	0,93	0,93
Peatones	0,9	0,89	0,88	0,87
Turismos	0,93	0,95	0,94	0,93

Se puede observar en la Tabla 5 que el ensamblado de modelos presenta buenos ajustes a los datos de estudio, en especial en la subpoblación de ciclomotores en la que se consigue un valor del área bajo la curva ROC de 0,96, siendo de 0,91 el valor del modelo base ganador en la comparativa del epígrafe anterior.

7 Trabajos futuros: Representación bipolar del conocimiento

La modelización del conocimiento y su posterior representación es una tarea compleja donde interactúan muchos campos científicos (véase, por ejemplo, Slovic et al. 2015). Nuestro cerebro es capaz de producir conceptos bajo una representación compacta, fiable y flexible de la realidad, y esta representación es la base para un eficiente proceso de toma de decisiones, y quizás más importante, la base para un lenguaje de comunicación eficiente, cuando se pone en palabras. En este trabajo se pretende continuar con el trabajo de (Rodríguez et al 2012) asociado al problema de clasificación. En particular, aquí nos centramos en cómo la oposición (y por lo tanto las estructuras pareadas) pueden funcionar simultáneamente en dos niveles diferentes, la lógica y los de representación

En el marco de problemas de clasificación con conjuntos de datos con clases no balanceadas y sus casos extremos se evalúa la mejora de resultados con la incorporación de información bipolar, siguiendo las ideas presentadas en Rodríguez et al, 2012 en el proceso de construcción del modelo de clasificación. En particular se aborda el caso de clasificación mediante arboles de clasificación con metodología CART. La agregación de información bipolar a los árboles de decisión para la clasificación supervisada multiclase comienza con la adición de la filosofía bipolar a posteriori en la clasificación. Tomando las probabilidades estimadas de pertenencia a cada una de las clases de interés, se evalúa la información de carácter positivo y negativo asociada a cada una de ellas y se agregan mediante un operador para obtener una nueva clasificación que recoja de formas más fiel los patrones subyacentes en los datos.
La aplicación de este procedimiento se realiza para la tarea de clasificación supervisada de la variable *lesividad* que recoge la gravedad de las lesiones producidas valoradas en el momento del

siniestro con cuatro niveles, Ileso (I), Herido Leve (HL), Herido Grave (HG) y Muerto (M) y cuyas clases están altamente desbalanceadas. En la tabla 6 se presentan los resultados preliminares de la aplicación a los datos con dos matrices de disimilaridad D1, considerada ad-hoc para el caso particular y D2 la propuesta en (Rodríguez et al. 2012).

Tabla 6: Comparativa de precisión de arboles y arboles con información bipolar.

Tipo	Accuracy	Kappa	TVP	Data	Disim.
Arbol	0,4464	0,2618	0,5405	Balan	D1
ArbolBip	0,4453	0,2605	**0,5445**	Balan	D1
Arbol	0,4464	0,2618	0,5405	Balan	D2
ArbolBip	0,4317	0,2422	**0,7267**	Balan	D2
Arbol	0,5787	0,1874	0,0455	NoBalan	D1
ArbolBip	**0,5792**	**0,1914**	**0,0682**	NoBalan	D1
Arbol	0,5787	0,1874	0,0455	NoBalan	D2
ArbolBip	0,5751	0,1784	**0,1212**	NoBalan	D2

Se aplica a conjunto de datos balanceado y no balanceado y se computan medidas clásicas de ajuste y la tasa de verdaderos positivos (TVP) considerada como la suma de Muertos y Heridos Graves. Se pueden apreciar ligeras mejoras.
Como trabajo a futuro, se procederá a la adición de la información bipolar, con el aprendizaje de la matriz de disimilaridades y los operadores de agregación adecuados, en la construcción del árbol de decisión, creando de esta forma un algoritmo de árbol bipolar que tenga en cuenta esta filosofía como criterio a la hora de realizar la particiones recurrentes propias de los arboles de decisión. Con ello se pretende enriquecer el conocimiento del algoritmo de clasificación introduciendo la información de carácter negativo, es decir, la probabilidad de pertenencia a las clases disimilares.

8 Conclusiones

A la vista de los resultados obtenidos en este estudio se pueden extraer las siguientes conclusiones. La metodología de estudio propuesta responde correctamente a los objetivos planteados en cuanto al estudio descriptivo de la población y el preprocesamiento de los datos disponibles, así como a los dos grandes puntos de interés del estudio. La determinación de los factores, ya sean propios del conductor, de la vía o aquellos circunstanciales, que elevan la probabilidad de resultar fallecido en accidente de tráfico en esta población, con resultados que confirman lo ya obtenido en otros muchos estudios sobre accidentalidad. Por otra parte la comparativa de bondad de ajuste de modelos de machine learning para la clasificación supervisada de la variable de interés *muerte a 30 días*, que arroja buenos resultados siendo Extreme Gradient Boosting y Neural Nets los algoritmos con mayor capacidad de clasificación.
En cuanto al apartado de representación bipolar de conocimiento aplicada a la clasificación de la variable categórica de carácter ordinal *lesividad*, que determina la gravedad del accidente, se observa en esta etapa inicial la ligera mejora que aporta su consideración a posteriori respecto al árbol de clasificación clásico, por lo que se considera una interesante línea de investigación a futuro.

Agradecimientos. A la Dirección General de Tráfico por la cesión de los datos de estudio. Este estudio se enmarca dentro del proyecto TIN2015-66471-P recientemente concedido y del grupo de investigación FORAID.

Referencias

Amo, A., Gómez, D., Montero, J., Biging, G. (2001). Relevance and redundancy in fuzzy classification systems. Mathw. Soft Comput. 8, 203–216.

Amo, A., Montero, J., Biging, G., Cutello, V. (2004). Fuzzy classification systems. Eur. J. Oper. Res. 156, 495–507.

Aguero-Valverde, J., (2013). Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: comparing the precision of crash frequency estimates. *Accident Analysis and Prevention* 50, 289-297.

Alemany, R., Ayuso, M., Guillén, M., (2013). Impact of road traffic injuries on disability rates and long-term care costs in Spain. *Accident Analysis and Prevention* 60, 95-102.

Brijs, T., Karlis, D., Van den Boussche and Geert Wets, F., (2007). A Bayesian model for ranking hazardous road sites. *Journal of the Royal Statistical Society: Series A* 170(4), 1001-1017.

Dzeroski, S. & Zenko, B. (2004). "Is combining classifiers with stacking better than selecting the best one?". Machine learning, 54, pp. 225-273. Kluwer Academic Publishers.

Friedman, H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist*. 29, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. Annals of Statistics 28(2). 337–407

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. (2009). A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models. The Annals of Applied Statistics, 2(4), 1360-1383.

Hosmer, D.W. & Lemeshow, S. (2000). "Applied Logistic Regression". John Wiley and Sons.

Kim, D-G., Lee, Y., Washinton, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. *Accident Analysis and Prevention* 39, 125-134.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software. Volume 28.

Rodríguez, JT., Vitoriano, B., Montero, J. (2012) A general methodology for data-based rule building and its application to natural disaster management. Computers & Operations Research 39 (4), 863-873

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.*Cambridge