SISTEMA EN LÍNEA DE PREDICCIÓN DE ACCIDENTES EN AUTOPISTA CENTRAL

Franco Basso[a] , Leonardo J. Basso[b], Francisco Bravo[c] and Raul Pezoa[b]

[a] *Escuela de Ingeniería Industrial, Universidad Diego Portales, Santiago, Chile.*
[b] *Departamento Ingeniería Civil, Universidad de Chile, Santiago Chile*
[c] *OPTILOG Consultores, Santiago de Chile*

RESUMEN

Desarrollamos modelos de predicción de accidentes en un tramo de Autopista Central utilizando información capturada a través de los pórticos de cobro, los cuales además de su baja tasa de falla tienen la ventaja de entregar información desagregada por tipo de vehículo. La metodología incluye un proceso de *random forest* para determinar los precursores de accidentes más relevantes, además de dos métodos de calibración/validación, a saber, *Support Vector Machines* y Regresiones logísticas. Descubrimos que la composición del flujo no juega un rol preponderante en el tramo estudiado de la ruta. Nuestro mejor modelo es capaz de predecir el 67.89% de los accidentes con un tasa de falsos positivos del 20.94%. Estos resultados son de los mejores en la literatura.
Palabras clave: Predicción de Accidentes, Support Vector Machines, Regresiones Logísticas.

ABSTRACT

We develop accident prediction models for a stretch of the urban expressway Autopista Central in Santiago, Chile, using disaggregate data captured by free-flow toll gates with Automatic Vehicle Identification (AVI) which, besides their low failure rate, have the advantage of providing disaggregated data per type of vehicle. The process includes a random forest procedure to identify the strongest precursors of accidents, and the calibration/estimation of two classification models, namely, Support Vector Machine and Logistic regression. We find that, for this stretch of the highway, vehicle composition does not play a first-order role. Our best model accurately predicts 67.89% of the accidents with a low false positive rate of 20.94%. These results are among the best in the literature.
Keywords: Accident Prediction, Support Vector Machines, Logistic Regression.

1. INTRODUCCIÓN

En los últimos años ha existido un creciente interés en el estudio de la seguridad vial de las autopistas urbanas, debido al aumento de la utilización de éstas y con el consiguiente aumento en las tasas de accidentes. En Chile, por ejemplo, la cantidad de accidentes en autopistas urbanas aumentó durante el 2015 un 8.3%. No es raro entonces que el trabajo de investigadores, autoridades y concesionarias se haya volcado al análisis de los precursores de los accidentes, es decir, las características que se observan previo a la ocurrencia de un accidente. Con las nuevas tecnologías y datos disponibles, diversos autores se han enfocado en definir contextos de mayor

riesgo de accidentabilidad en autopistas urbanas utilizando variables de entorno. Unos de los primeros en preocuparse de esta temática fueron Golob & Recker (2004) quienes definieron regímenes de flujo según tres tipos de condiciones climáticas. Cada uno de ellos tiene asociado un único tipo accidente con mayor probabilidad de ocurrir. Utilizando herramientas tipo *k-clustering* y con data del año 1999 en California del Sur, los autores obtuvieron una taxonomía para las condiciones previas al accidente. Abdel Aty et al. (2004) utilizaron la técnica de modelamiento *matched case-control logistic* para ajustar un modelo de predicción de accidentes que logra aciertos de un 67% pero con 76% de falsos positivos. Para una completa revisión de la literatura sobre el efecto del tráfico y condiciones climáticas en accidentes se refiere al autor a Theofilatos & Yannis (2014). Kwan & Kho (2016) desarrollan modelos estadísticos predictivos para una autopista en Korea. Según los autores, es especialmente importante considerar modelos para distintos tipos de segmentos y estados del tráfico pues los accidentes tienen características muy distintas según cada estado.

Support Vector Machines (SVM) fue recientemente aplicada en esta área por Lv et al (2009) y Yu & Abdel Aty (2013). Yu & Abdel Aty (2013) utilizaron datos de la ruta I-70 en Colorado para medir el riesgo de accidente en tiempo real. Por su parte, Lv et al (2009) ocuparon datos simulados del software TSIS para identificar condiciones de tráfico que aumenten las probabilidades de accidentes. A diferencia de nuestro estudio, Yu & Abdel Aty (2013) utilizaron solo algunas observaciones de no-accidente. Según nuestro conocimiento, este es el primer artículo que utiliza datos completos y de alta de precisión para validar los modelos. Nosotros creemos que esta virtud aumenta la probabilidad de éxito en la implementación de una herramienta computacional que funcione en línea.

Diversos medios de recolección de datos han sido utilizados para los sistemas de predicción de accidentes en tiempo real. El más ampliamente usado son las espiras. Sin embargo, según Ahmed & Abdel-Aty (2012), las espiras fallan entre 24% y 29% del tiempo por lo que su uso para sistemas en tiempo real se hace difícil. Otros medios de obtención de datos han sido desarrollados en el último tiempo. Microwaves Vehicle Detecction System (MVDS) ha sido instalado en Central Florida Expressway, que es capaz de archivar data en intervalos de 1 minuto sin interrupción. (Shi et al. 2015). También existen métodos el procesamiento de imágenes de video que pueden ayudar a estos sistemas. Estas metodologías brindan información agregada de las condiciones del tráfico.

En este estudio trabajamos con datos de Autopista Central de Santiago de Chile. Esta es una moderna autopista urbana de una longitud de 60,5 km a lo largo de la región metropolitana. Esta autopista está concesionada por una empresa que realiza el cobro de sus servicios mediante un dispositivo electrónico que identifican a los vehículos previamente inscritos cada vez que circulan por un pórtico (*Automatic Vehicle Identification*, AVI). Dado que el pago del servicio depende del sistema de detección, ésta rara vez deja de funcionar, permitiéndonos tener una base completa, con información desagregada por tipo de vehículo. Según nuestro conocimiento, todos los artículos reportados en la literatura con datos en tiempo real han trabajado con información proveniente de dispositivos con información agregada, usualmente cada 30 segundos y sin capacidad de separar según tipo de vehículo. El presente estudio es el primero que utiliza información según el tipo de vehículo (autos y camionetas, camiones y buses, motos) para medir el riesgo de accidente y a través de dispositivos más confiables que las espiras.

Esta nueva capacidad de disponer de una gran cantidad de data desagregada trae consigo el problema de selección de variables. Es sabido que la inclusión de una gran numero de variables puede causar sobreajuste del modelo (Sawalha y Sayed, 2006). Esto también puede afectar tanto la interpretación en la relación variables explicativas – accidente en casos de estudio, como la implementación en tiempo real.

El resto del artículo se organiza como sigue: en la sección 2 describimos la información, cómo fue procesada y proveemos estadística descriptiva. En la sección 3 se presenta el marco teórico de las técnicas utilizadas. En la sección 4 se calibran los modelos *Support Vector Machine* y Regresiones logísticas, para posteriormente realizar repeticiones de validaciones cruzadas que usan un 80% de la data para calibrar y 20% para validar. La sección 5 concluye.

2. PREPARACIÓN DE LOS DATOS

La base de datos de accidentes proporcionada por Autopista Central considera información relativa al periodo que va desde el 1 de noviembre del 2014 al 30 de abril 2016. De los 10 tramos de la autopista, se escogió el con mayor tasa de accidentabilidad por km, con 5.64 accidentes por km. El tramo tiene una longitud de 4.71 km y se extiende desde el rio Mapocho en el norte hasta Rondizzoni por el sur. La información de las condiciones de flujo es obtenida con 2 pórticos AC-09-s1 (PK=0+257, pórtico 14) y AC-08-s1 (PK=2+569, pórtico 12). Esto permite obtener información desagregada de cada vehículo que pasa por ellos. Estos se clasifican en 3 categorías (Autos y camionetas, Camiones, Motos), lo que nos brinda una excelente oportunidad para explicar los accidentes según la composición del flujo.

La base de datos de flujos y velocidades proporcionada por Autopista Central para el tramo seleccionado cuenta con un total de 70.037.589 datos de vehículos. Para manejar el gran número de datos disponibles, estos son discretizados en intervalos de 5 minutos, y desagregados según el tipo de vehículo. Se define una condición de accidente al intervalo donde se registra al menos un accidente. Solo un intervalo registra 2 accidentes. Para el problema de predicción, se define la condición de preaccidente como el intervalo anterior al intervalo donde ocurre el accidente.

Los ajustes realizados se hicieron considerando solo los días hábiles de la semana. Esto brinda varias ventajas. Primero permite trabajar con datos de la misma naturaleza, evitando cambios estructurales al modelo que puedan generar inestabilidades de los parámetros estimados. Segundo, dado que la tasa de accidentes por hora es mayor en días hábiles que en fin de semana, trabajar solamente con los primeros permite balancear en la base de datos el número de casos accidente con respecto al número de no accidente. También se restringió la metodología al periodo de 17:30 a 20:29, que corresponde al intervalo punta tarde, según la definición del Programa de Operación 2016 de la DTPM. Esta restricción se impone principalmente con fines exploratorios, debido a posibles diferencias en la naturaleza del flujo de vehículos al inicio de la jornada laboral con el flujo al finalizar la jornada laboral.

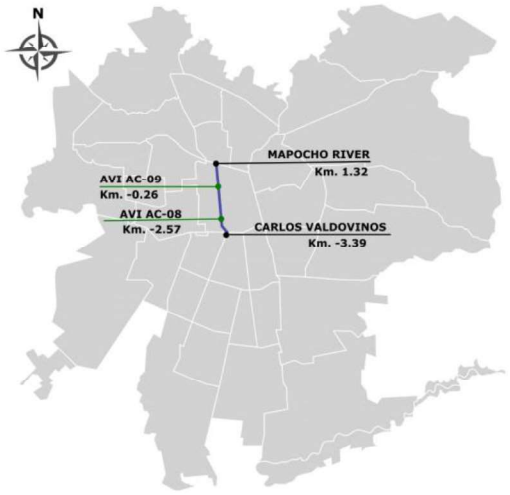

Figura 1: Sección estudiada de la autopista.

Para cada uno de los tipos de vehículo se obtienen las variables explicativas (flujo, velocidad media, desviación estándar, densidad, delta velocidad y delta densidad). La variable densidad se define como el flujo medio dividido la velocidad promedio. Las variables delta corresponden a la diferencia entre el valor de la variable y el valor del intervalo anterior.

3. METODOLOGÍA

Nuestra metodología consiste en un procedimiento de selección de variables mediante *Random Forest*, para después calibrar modelos *Support Vector Machine* y regresiones logísticas. Para el caso de *Support Vector Machine*, se utiliza además un proceso de sobremuestreo a través del algoritmo SMOTE. Asumiremos que las regresiones logísticas son conocidas por el lector, por lo que explicaremos a continuación el resto de los métodos utilizados.

3.1 *Random Forest*

Random Forest (RF) es una máquina de aprendizaje de clasificación, compuesto por una colección de árboles de decisión. El RF clasifica una entrada en la clase que ha sido más veces asignada por los árboles que la componen. (Brieman, 2001). La construcción de cada árbol del RF se realiza a través de 2 procesos aleatorios. Primero se realiza una muestra aleatoria de casos, con reemplazo, que se usara para hacer crecer el árbol. Segundo, se selecciona una muestra de entre todas las variables que son usadas para dividir los nodos. La data que no se utiliza para construir el árbol se denomina *oob* (out-of-bag). Los datos *oob* permiten obtener una estimación insesgada del error de clasificación del RF.

En este estudio, el algoritmo RF es utilizado para estimar la importancia de las variables como precursores de accidentes La importancia de una variable en un árbol de decisión se estima en su capacidad de reducir un índice de impureza (en este caso, índice de Gini) de los nodos cuando se utiliza como variable separadora.

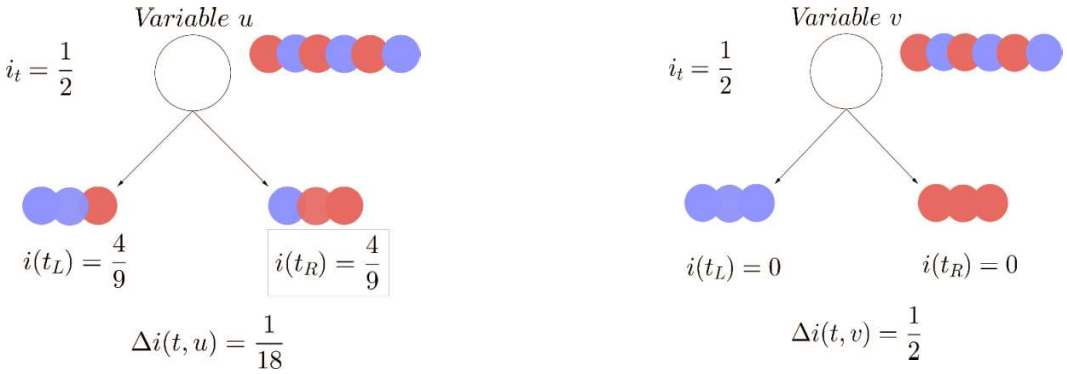

Figura 2: Ejemplo de dos separaciones, con caso 2 (variable v) preferible.

Para un índice la reducción de impureza en el nodo se calcula como cuando la *oob* es evaluada:

$$\Delta i(t,u) = i(t) - \frac{N_L}{N} i(t_L) - \frac{N_R}{N} i(t_R) \tag{1}$$

La importancia de una variable en un RF se calcula como una suma ponderada de la reducción de la impureza de los nodos donde la variable es utilizada. Este se denomina *Mean Decrease Gini*.

3.2 *Support Vector Machine*

El problema de clasificación binaria para un conjunto de datos $(x_1, y_1), \ldots, (x_n, y_n)$ $y_i \in \{-1,1\}, x_i \in \mathbb{R}^d$ $i = 1, \ldots, n$ puede ser abordado con SVM. Este algoritmo busca el hiper plano separador $f(x) = wx + b$ entre las 2 clases que maximice la distancia de las clases a la frontera de decisión.

Cuando la data de entrenamiento es linealmente separable, se cumplirá, sin pérdida de generalidad que

$$x_i \cdot w + b \geq 1 \quad \text{for } y_i = 1 \tag{2}$$

$$x_i \cdot w + b \leq -1 \quad \text{for } y_i = -1 \tag{3}$$

Combinando ambas restricciones:

$$y_i (x_i \cdot w + b) \geq 1 \quad \forall i \tag{4}$$

Es posible demostrar (Cortes y Vapnik, 1995) que el w que maximiza el margen es solución del problema de optimización

$$\min \frac{1}{2} \| w \|^2 \quad \text{s.t. } y_i (x_i \cdot w + b) \geq 1, \; i = 1, \ldots, n \tag{5}$$

Cuando los datos no son linealmente separables, se pueden agregar variables de holgura positivas ξ_i que penalicen errores de clasificación, quedando el SVM planteado como (Cortes y Vapnik, 1995):

$$\min \frac{1}{2} \| w \|^2 + C \sum_i \xi_i$$

$$\text{s.t } y_i (x_i \cdot w + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \tag{6}$$

$$\xi_i \geq 0, \; i = 1, \ldots, n$$

Introduciendo los multiplicadores de Lagrange, el problema de optimización se puede plantear como

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

$$\text{s.t } \sum_i \alpha_i y_i = 0 \tag{7}$$

$$0 \leq \alpha_i \leq C \quad \forall i$$

Cuando la función de decisión no es una función lineal de los datos, es posible mapear la data a otro espacio euclidiano H mediante una función Φ conveniente para la cual las clases sean separables. Notando que en la formulación dual la data de entrenamiento solo aparece como producto punto $x_i \cdot x_j$, el mapeo en el espacio euclidiano H es realizado computando la función kernel K que represente el producto punto en $H : K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ (Friedman et al., 2001). En este estudio se probaron los *kernels* clásicos utilizados,

- Radial Kernel: $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$
- Polinomial Kernel: $K(x_i, x_j) = (\gamma x_i \cdot x_j + 1)^q, \text{ with } q = 3$
- Sigmoid Kernel: $K(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + 1)$

3.3 *Synthetic Minority Over-sampling Technique*

Para trabajar con la base de accidentes desbalanceada, utilizamos el algoritmo SMOTE Synthetic Minority Over-sampling Technique (SMOTE) desarrollado por Chawla et al. 2002. El algoritmo consiste en submuestrear la clase mayoritaria y sobremuestrar la clase minoritaria (Figura 3). Para sobremuestrear la clase minoritaria se crean ejemplos sintéticos. Estos se introducen aleatoriamente entre los elementos de la clase desbalanceada y alguno de sus k-vecinos más cercanos. Diferentes proporciones entre clases son probadas en los ajustes, sobremuestrando la

clase minoritaria tal que sean clasificados en la etapa de entrenamiento, sin destruir la naturaleza de problema real.

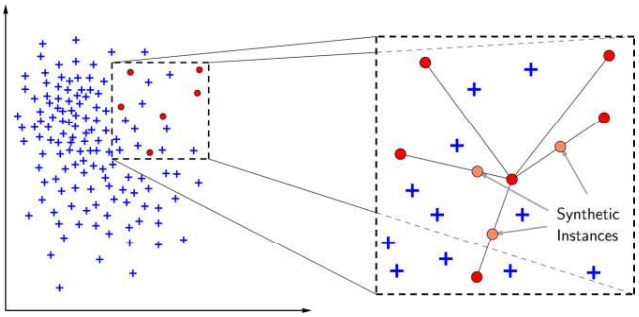

Figura 3: Sobremuestra usando SMOTE.

4. DESARROLLO DE MODELOS Y VALIDACIÓN

En la presente sección se calibrarán las metodologías descritas en la sección 3 para primero escoger las variables de modelación, y luego construir modelos de clasificación SVM y regresiones logísticas.

4.1 Selección de Variables

Se calculó el coeficiente de correlación de Pearson $\rho_{X,Y}$ para cada combinación de variables X, Y y se descartó una de ellas si $|\rho_{X,Y}| > 0.95$, para evitar efectos de multicolinealidad de variables casi perfectamente relacionadas. Esto permitió descartar las variables Den.Autos08, ya que de su definición $Den = \frac{Flujo}{Velocidad}$ se observa que, debido a un flujo que presenta pocas variaciones en el período estudiado, existe una correlación negativa casi perfecta entre esta variable y Vel.Autos08. No se descartaron más variables para evitar una subespecificación de los modelos posteriormente ajustados.

Posteriormente, se construyó un modelo RF de para ordenar las variables según su importancia. Esta metodología es similar a la aplicada en estudios anteriores por Ahmed y Abdel-Aty (2012), Abdel et al. (2008), Xu et al. (2013) y Lin et al. (2015), y consiste en 500 árboles con cuatro variables escogidas de manera aleatoria en cada división. Los resultados obtenidos, presentados en la Figura 4, muestran que las variables con mayor importancia para este indicador son el cambio en la velocidad en el pórtico AC08 y el cambio en la densidad en ambos pórticos. Además, las velocidades de los vehículos (de todas las categorías) en el pórtico AC08 muestran también una alta importancia. Estas observaciones sirvieron como base para el ajuste de modelos presentado en la sección siguiente.

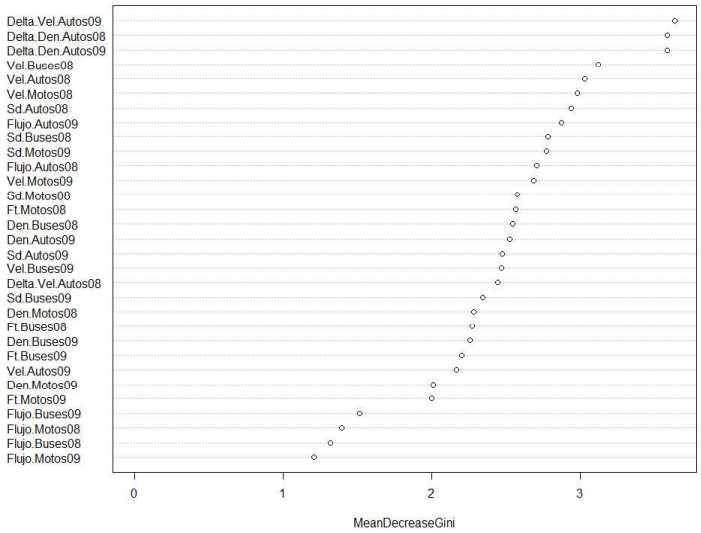

Figura 4: Cambio en índice de impureza de Gini para determinar importancia de variables.

Finalmente, se realizó un análisis gráfico de la evolución de las variables más importantes de la clasificación anterior en torno a la ocurrencia de los accidentes registrados, comparando este comportamiento con el usual de la misma variable en períodos sin accidentes, obteniéndose los gráficos mostrados a continuación, en donde es posible observar (Figura 6) que las velocidades de los autos registradas en el pórtico AC08 previo a un accidente son mucho menores que las usuales, y que en el período de 5 minutos anteriores al accidente se registra el mínimo global del período, hecho que se repite con la variable Delta.Den.Autos09 (Figura 5). Lo anterior es un buen indicador de la relevancia predictiva de estas variables, que finalmente fueron incorporadas en la modelación.

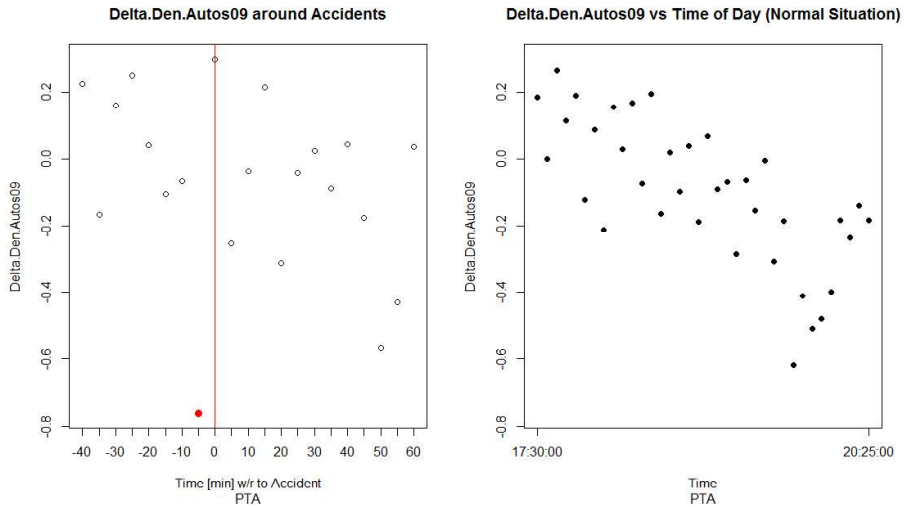

Figura 5: Comportamiento de Delta.Den.Autos09 antes y después de accidentes.

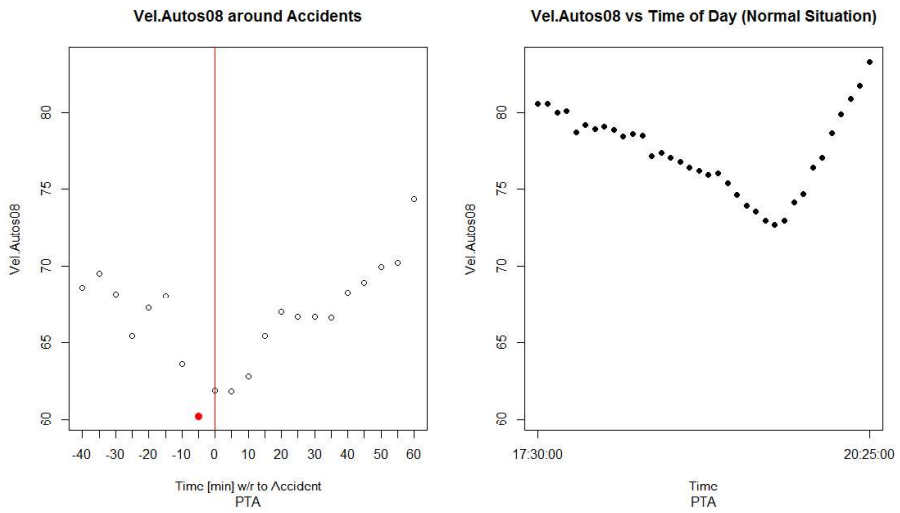

Figura 6: Comportamiento de Vel.Autos08 antes y después de accidentes.

4.2 Modelos Support Vector Machines

Utilizando las variables Vel.Autos08 y Delta.Den.Autos09 en conjunto con combinaciones del resto de variables, se ajustaron diferentes modelos SVM con múltiples combinaciones de valores para SMOTE. Se presentan a continuación los modelos que mejores resultados muestran, además de los indicadores para el ajuste sobre la base completa (sin validación):

Tabla 1: Resultados Modelos SVM sin validación.

Kernel	Radial	Sigmoid	Polinomial (Grado 3)
gamma	0,001	0,001	1
cost	10	100	1
SMOTE.perc.over	500	500	500
SMOTE.perc.under	100	100	100
Variables	Vel.Autos08 Delta.Den.Autos09	Vel.Autos08 Delta.Den.Autos09	Vel.Autos08 Vel.Autos09 Delta.Den.Autos09
Sensitivity (%)	71,79	71,79	92,31
False Positives Rate (%)	28,09	28,15	54,12

Al graficar numéricamente las zonas donde el mejor modelo SVM con kernel radial encontrado predice accidente (Figura 7 y Figura 8) se observa que la mejor frontera de decisión viene dada por una curva similar a una recta. Misma situación ocurre con el kernel sigmoidal, lo que hace intuir la no existencia de zonas no paramétricas de decisión con mejor ajuste. Lo anterior nos lleva a utilizar regresiones logísticas las cuales tienen una más fácil interpretación.

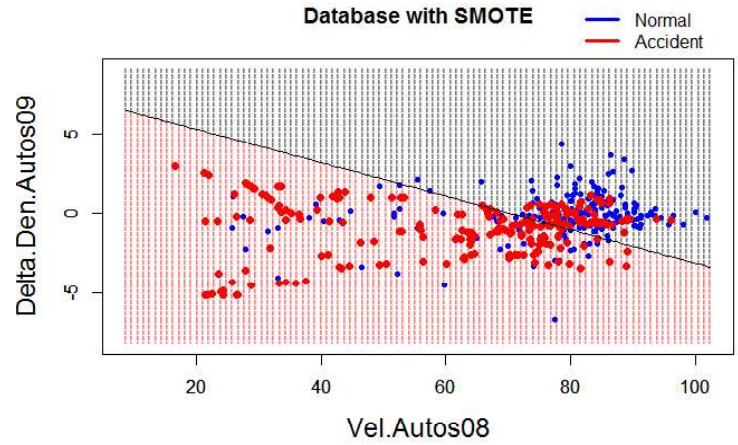

Figura 7: Zona decisión SVM con kernel radial (base con SMOTE).

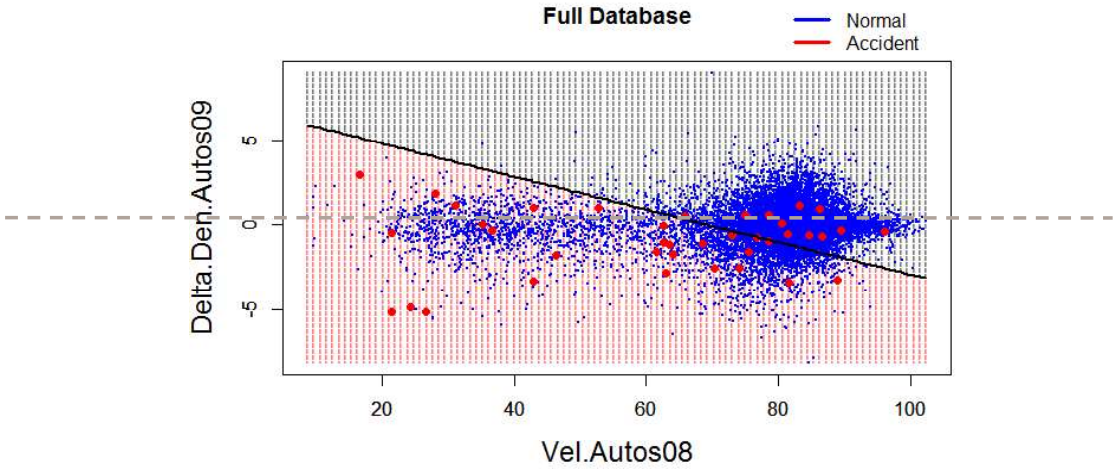

Figura 8: Zona decisión SVM con kernel radial (base completa).

4.3 Regresiones logísticas

De manera similar, se ajustaron diferentes modelos de regresión logística considerando las variables Vel.Autos08 y Delta.Den.Autos09 en conjunto con combinaciones del resto de variable, de donde se concluye que el modelo que mejores resultados presenta es:

$$\text{PreAccBin} \sim \text{Vel.Autos08} + \text{Delta.Den.Autos09} \tag{8}$$

Tabla 2: Parámetros Regresión Logística 2 variables, base completa.

Variable	Coeficiente	Error Estándar	p-valor
Intercepto	-3,277	0,481	$9{,}18 \cdot 10^{-12}$
Vel.Autos08	-0,038	0,007	$5{,}25 \cdot 10^{-8}$
Delta.Den.Autos09	-0,34	0,098	$5{,}05 \cdot 10^{-4}$

La Tabla 2 indica que ambas variables son significativas al 99%. Este modelo logra una sensibilidad de un 64,1% y una tasa de falsos positivos de 20,9% al ajustar sobre la base

completa, como se observa en la Figura 9 donde la probabilidad de *threshold* se fijó en $p = 0{,}314$. La modificación de este valor producirá únicamente una traslación de la recta mostrada. Además, con los signos de los coeficientes asociados es posible interpretar el efecto de dichas variables: velocidades bajas y disminuciones en la densidad de vehículos en categoría "Autos" aumentan la probabilidad de ocurrencia de un accidente. Esto también se aprecia graficando la distribución de accidentes versus las variables escogidas, donde se comprueba que la mayoría de accidentes se producen a velocidades menores a 80 km/h en el pórtico AC08 y con Delta.Den.Autos09 negativos, es decir con densidades de autos decrecientes en los últimos 10 minutos en este pórtico (los vehículos se están "descomprimiendo" en el pórtico AC09 y por tanto probablemente acelerando, y luego se encuentran con vehículos a bajas velocidades en el pórtico AC08).

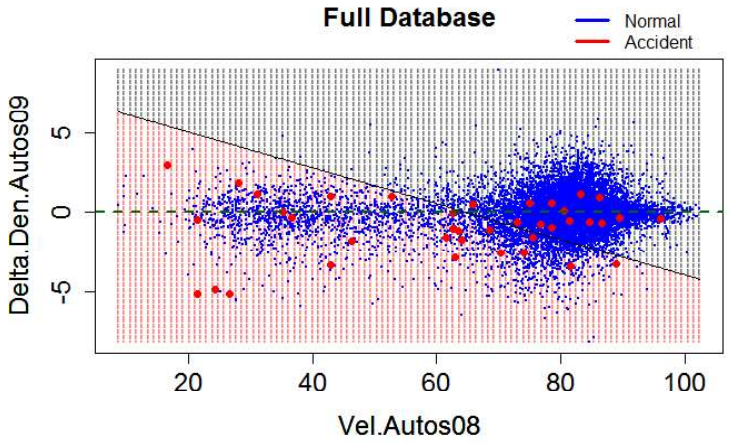

Figura 9: Frontera de Decisión Logit en 2 variables (base completa).

El comportamiento anterior induce a introducir la velocidad en el pórtico de entrada (AC09) como variable de modelación, lo que se realizó de manera no lineal, utilizando las variables artificiales $Vel.Autos08^2$ y $Delta.Den.Autos09 \cdot Vel.Autos09^2$, para posteriormente ajustar un logit sobre ellas, con los siguientes coeficientes:

Tabla 3: Parámetros Regresión Logística 3 variables, base completa.

Variable	Coeficiente	Error Estándar	p-valor
Intercepto	-4,196	0,338	$< 2 \cdot 10^{-16}$
$Vel.Autos08^2$	$-3{,}34 \cdot 10^{-4}$	$6{,}39 \cdot 10^{-5}$	$1{,}72 \cdot 10^{-7}$
$Delta.Den.Autos09 \cdot Vel.Autos09^2$,	$-7{,}69 \cdot 10^{-5}$	$2{,}00 \cdot 10^{-5}$	$1{,}19 \cdot 10^{-4}$

Se observa que ambas variables (que son función de tres variables originales) son significativas al 99%. Podemos interpretar el signo de las variables: debido a que $Vel.Autos^2 > 0$ en ambos pórticos, se sigue teniendo que velocidades bajas en el pórtico AC08 y disminuciones en la densidad de vehículos en categoría "Autos" en el pórtico AC09 aumentan la probabilidad de ocurrencia de un accidente. El efecto de la velocidad en el pórtico AC09 por contraparte depende del signo de Delta.Den.Autos09: disminuciones en la densidad (``descompresión', i.e.

Delta.Den.Autos09<0) en este pórtico hacen que velocidades altas provoquen un aumento en la probabilidad de ocurrencia de accidentes. Es decir, la situación que provoca la máxima probabilidad de ocurrencia de un accidente es:

- Velocidades Altas en Pórtico AC09
- Disminución en la densidad en el tramo que une ambos pórticos.
- Velocidades Bajas en Pórtico AC08.

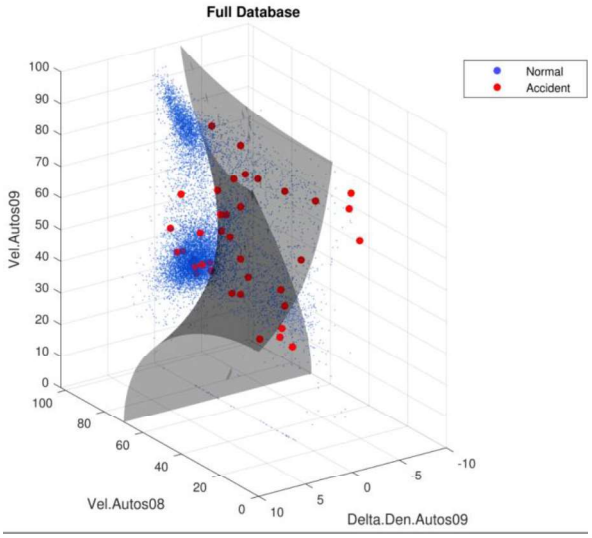

Figura 10: Frontera de Decisión Logit en 3 variables (base completa).

4.4 Validación

Para efectos de la validación de los modelos presentados se realizaron 300 repeticiones de una validación cruzada de 5 capas (5-fold *cross validation*), en donde en cada repetición se divide la información disponible de manera aleatoria en 5 conjuntos de tamaño similar, y se utilizan cuatro de ellos para entrenar (31 accidentes en promedio) y el restante para validar (8 accidentes en promedio). Esto es realizado para las cinco combinaciones posibles, y luego se calcula el promedio de los falsos positivos (*False Negatives Rate*) y de los accidentes acertados (*sensitivity*). Estos valores serán los representativos de la repetición correspondiente. En cada construcción de un modelode regresión logística, la probabilidad de *threshold* es escogida de manera que se alcance una tasa de falsos positivos cercana al 20% al realizar una predicción sobre la base de entrenamiento. Esta probabilidad (porcentual) resulta estar siempre contenida en el intervalo [0.29%,0.30%], lo que es consistente con la proporción de accidentes dentro de la base considerada (0.299% de los datos).

A continuación, se presentan los principales indicadores de los modelos anteriormente presentados:

Tabla 4: Resultados validación modelos SVM y Regresión Logistica.

Indicadores	SVM Kernel Radial	SVM Kernel Sigmoid	SVM Kernel Polinomial	Logit Lineal	Logit Extensión No Lineal
Sensibilidad Promedio (%)	68.96	68.40	77.05	62.13	67.89
Sensibilidad Máxima (%)	79.15	79.29	95.32	71.26	77.50
Sensibilidad Mínima (%)	53.50	54.64	52.06	45.63	59.17
Tasa de Falsas Alarmas Promedio (%)	28.44	27.72	59.78	20.94	20.94

De la Tabla 4, se observa que los modelos SVM presentan altos porcentajes de sensibilidad, particularmente el correspondiente a un kernel Polinomial de grado 3 que alcanza casi un 80% predicción promedio, pero por contraparte entregan altas tasas de falsos positivos al sobreestimar las zonas donde decide accidente, efecto atribuible al balanceo de la base al realizar SMOTE. Por otro lado, los modelos logit muestran tasas de falsos positivos cercanas al 20% ajustado sobre la base de entrenamiento, comportamiento esperado al realizar una validación cruzada con elección aleatoria. En esta última categoría, el modelo logit no lineal sobre las variables de modelación presenta los mejores resultados, con una sensibilidad promedio de 67.89% (similar a lo obtenido con los modelos SVM con kernel Radial y Sigmoidal), y una tasa de falsos positivos promedio de 20.94% (mucho menor a los mismos modelos SVM), lo que se compara de manera favorable a lo presentado en la literatura contemporánea, como se muestra en la Tabla 5 (generada parcialmente a partir de lo mostrado en Lin et al., 2015), especialmente si se considera la estabilidad del modelo, que presenta una sensibilidad mínima sobre las 300 repeticiones de un 59.17%.

Tabla 5: Resultados presentes en literatura.

Autores	Tipo de Selección de Variable	Tipo de Modelo de Clasificación	Sensitivity (%)	False Alarm Rate (%)
Abdel-Aty et al. (2004) [1]	N/A	Logistic regression	69	N/A
Pande and Abdel-Aty (2006) [16]	Classification tree	Neural Network	57.14	28.83
Abdel-Aty et al. (2008) [14]	Random forest	Neural Network	61	21
Hossain and Muromachi (2012) [4]	Random multinomial logit	Bayesian Network	66	20
Ahmed and Abdel-Aty (2012) [17]	Random forest	Matched case-control method	68	46
Ahmed and Abdel-Aty (2012) [18]	N/A	Bayesian logistic regression	75.71-80.09	33.59-32.31
Lin et al. (2015) [15]	Frequent Patern tree	Bayesian Network	61.11	38.16
Sun and Sun (2015) [19]	N/A	Dynamic Bayesian network	76.4*	23.7*

* Porcentajes encontrados entrenando y validando con un subconjunto de los datos.

Los estudios que logran tasas más altas de sensibilidad que lo propuesto aquí tienen asociado un porcentaje de falsos positivos mucho mayor que el 20% buscado en el presente estudio, como lo registrado en Ahmed and Abdel-Aty (2012), que proyectándolo a una tasa de falsa alarma

cercana a un 20%, presenta una sensibilidad de aproximadamente 60% según muestra la curva ROC de dicho artículo. Además, los porcentajes de sensibilidad y tasa de falsas alarmar mostrados en la Tabla 4 corresponden a repeticiones de una validación cruzada, por lo que son un excelente indicativo del poder predictivo del modelo, a diferencia de los valores obtenidos validando con un solo caso. Por otro lado, el alto porcentaje de sensibilidad encontrado en Sun & Sun (2015) tiene la desventaja metodológica de provenir de una base balanceada artificialmente, en donde para cada registro de accidente se escogieron únicamente 5 registros de situación normal, para posteriormente utilizar esta base para efectos de entrenamiento y validación, por lo que los porcentajes mostrados no reflejan el poder predictivo real del modelo, a diferencia del modelo propuesto aquí.

5. CONCLUSIONES

En el estudio mostrado se presentan técnicas basadas en aprendizaje de máquinas y en modelos de regresiones logísticas para predecir la ocurrencia de un accidente en un tramo de la Autopista Central de Santiago, las cuales se basan en los datos obtenidos en tiempo real durante los últimos 5 minutos. El presente estudio es el primero en realizar predicciones basándose en variables desagregadas por tipo de vehículo (Automóvil, Bus, Moto), lo que permite aislar la contribución de cada uno de ellos al aumento en la probabilidad de ocurrencia de un accidente. La información utilizada además es de alta precisión (comparada con sistemas de medición como espiras) debido a que de ello depende el sistema de cobro de la autopista.

La construcción de un modelo *Random Forest* usado para clasificar la importancia de las variables disponibles, sumado a la inspección visual permitió identificar las principales variables explicativas de la ocurrencia de accidentes. Usando lo anterior, se ajustaron modelos SVM y regresiones logísticas, de los cuales se concluye:

1. En el tramo estudiado las variables de modelación escogidas están relacionadas únicamente con vehículos en categoría “Autos y camionetas”, lo que tiene directa relación con la centralidad de dicho tramo, ubicado en torno a zonas de desarrollo económico y social, y por tanto de naturaleza intrínsecamente urbano, lo que se refleja en la composición del tráfico: 93,1% de los vehículos registrados en el período estudiado corresponden a la categoría “Autos y camionetas”. Al extender el presente estudio a tramos de carácter rural es esperable que variables relacionadas al resto de tipos de vehículos (principalmente a los de categoría “Buses y Camiones”) sean relevantes.

2. Los modelos SVM logran un alto porcentaje de sensibilidad, pero tienden a sobrestimar la zona de decisión “Accidente”, lo que provoca altas tasas de Falsos Positivos, muy superiores al 20% buscado a priori. Esto es atribuible al proceso de generación y modelación sobre datos sintéticos (SMOTE) requerido debido al alto nivel de desbalanceo de la base, lo que produce perturbaciones al validar sobre los datos reales.

3. El modelo de regresiones logísticas no lineal en las variables originalmente consideradas alcanza en la validación una sensibilidad promedio de 67.89% con solo un 20.94% de falsos positivos. Esta sensibilidad es comparable con los mejores resultados obtenidos en la literatura contemporánea. Más aún, los estudios que alcanzan porcentajes similares de acierto tienen asociado en general una tasa de falsos positivos mucho mayor a la aquí

encontrada. Lo anterior promueve la expansión de este estudio al resto de tramos de la autopista y otros horarios, debido a la riqueza de los datos disponibles gracias a los pórticos de cobro automático. Es esperable sin embargo que, debido a las diferencias en la geometría, largo y composición de tráfico en los diferentes tramos, se requiera ajustar distintos modelos para cada uno, utilizando eventualmente otras variables explicativas o formas funcionales.

4. A partir de los modelos de regresiones logísticas se identifica la situación de mayor probabilidad de ocurrencia de accidentes: velocidades altas aguas arriba, sumado a densidades decrecientes (reducción en la congestión), y velocidades bajas en el pórtico de salida del tramo. Lo anterior se condice con la intuición y experiencia empírica: la repentina reducción en la congestión induce conductas más agresivas en conductores que tratan de recuperar el tiempo perdido, los cuales prontamente se encuentran con velocidades bajas en el pórtico siguiente (debido a que los pórticos de los cuales proviene la información están separados por solo 2,3 km), lo que provoca maniobras de frenado que pueden terminar en colisiones.

Referencias

Abdel-Aty, Mohamed A., et al. "Real-time prediction of visibility related crashes." **Transportation research part C: emerging technologies** 24 (2012): 288-298.

Abdel-Aty, M., Pande, A., Das, A., & Knibbe, W. (2008). Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. **Transportation Research Record: Journal of the Transportation Research Board**, (2083), 153-161.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. **Transportation Research Record: Journal of the Transportation Research Board**, (1897), 88-95.

Ahmed, M. M., & Abdel-Aty, M. A. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. **IEEE Transactions on Intelligent Transportation Systems,** *13*(2), 459-468.

Ahmed, M., Abdel-Aty, M., & Yu, R. (2012). Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data. **Transportation research record: journal of the transportation research board**, (2280), 51-59.

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. **In European conference on machine learning** *(pp. 39-50). Springer Berlin Heidelberg.*

Breiman, L. (2001). Random forests. **Machine learning**, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). **Classification and regression trees**. Wadsworth & Brooks. *Monterey, CA*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, *16*, 321-357.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. **Machine learning**, *20*(3), 273-297.

Golob, T. F., & Recker, W. W. (2004). A method for relating type of crash to traffic flow characteristics on urban freeways. **Transportation Research Part A: Policy and Practice**, *38*(1), 53-80.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). Model assessment and selection. In **The elements of statistical learning** (pp. 193-224). Springer New York.
Hastie, T., & Pregibon, D. (1992) *Generalized linear models.* Chapter 6 of **Statistical Models in S** eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for realtime crash prediction on the basic freeway segments of urban expressways. **Accident; analysis and prevention***, 45, 373–81.*
Kwak, H. C., & Kho, S. (2016). Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. **Accident Analysis & Prevention**, *88*, 9-19.
Lin, L., Wang, Q., & Sadek, A. W. (2015). A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. **Transportation Research Part C: Emerging Technologies**, *55*, 444-459.
Lv, Y., Tang, S., Zhao, H., & Li, S. (2009). Real-time highway accident prediction based on support vector machines. In **Control and Decision Conference***, 2009. CCDC'09. Chinese* (pp. 4403-4407). IEEE.
Sawalha, Z., & Sayed, T. (2006). Traffic accident modeling: some statistical issues. **Canadian Journal of Civil Engineering**, *33*(9), 1115-1124.
Rizzi, L. I., & de Dios Ortúzar, J. (2003). Stated preference in the valuation of interurban road safety. **Accident Analysis & Prevention**, *35*(1), 9-22.
Shi, Q., Abdel-Aty, M., & Yu, R. (2016). Multi-level Bayesian safety analysis with unprocessed Automatic Vehicle Identification data for an urban expressway. **Accident Analysis & Prevention**, *88*, 68-76.
Shi, Q., & Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. **Transportation Research Part C: Emerging Technologies**, *58*, 380-394.
Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. **Transportation Research Part C: Emerging Technologies**, *54*, 176-186.
Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2016). Predicting road accidents: a rare-events modeling approach. **Transportation Research Procedia**, *14*, 3399-3405.
Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. **Accident Analysis & Prevention**, *72*, 244-256.
Yu, R., Abdel-Aty, M. A., Ahmed, M. M., & Wang, X. (2014). Utilizing microscopic traffic and weather data to analyze real-time crash patterns in the context of active traffic management. **IEEE Transactions on Intelligent Transportation Systems**, *15*(1), 205-213.
Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. **Accident Analysis & Prevention**, *51*, 252-259.
Xu, C., Wang, W., & Liu, P. (2013). A genetic programming model for real-time crash prediction on freeways. **IEEE Transactions on Intelligent Transportation Systems**, *14*(2), 574-586.