IX Simposio de Teoría y Aplicaciones de la Minería de Datos (IX TAMIDA)

TAMIDA 3:

APLICACIONES

Modeling the navigation on enrolment web information area of a university using machine learning techniques*

*Note: The full contents of this paper have been published in the volume *Lecture Notes in Artificial Intelligence 11160* (LNAI 11160)

Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, Javier Muguerza
Faculty of Informatics
University of the Basque Country
Donostia, Spain
{ainhoa.yera, inigo.perona, olatz.arbelaitz, j.muguerza}@ehu.eus

***Abstract*—This work analyses the navigation in the enrolment web information area of the University of the Basque Country. A complete data mining process shows that successful and failure navigation behaviors can be modeled using machine learning techniques. Unsupervised learning algorithms have been applied on two different domains: URLs visited by the users in each session (navigation sequence) and some interaction parameters extracted from the recorded click-stream (navigation style). Both domains have been used satisfactorily to model the behavior of success and failure navigation sessions achieving more than 78 % of accuracy predicting success or failure sessions. Furthermore, the clustering based on the navigation style was able to identify the main characteristics of each type of session and to build a subsystem that enables to detect failure type sessions with high precision.**

***Index Terms*—Web Usage Mining, Navigation Models, Web Interaction Characterization**

Risk factors for development of antibiotic resistance of *Enterococcus faecium* to Vancomycin. A subgroup discovery approach*

*Note: The full contents of this paper have been published in the volume *Lecture Notes in Artificial Intelligence 11160* (LNAI 11160)

A. Fajfar
Faculty of Health Sciences
University of Maribor
Maribor, Slovenia

M. Campos
Faculty of Computer Science
University of Murcia
Murcia, Spain
manuelcampos@um.es

F. Palacios
University Hospital of Getafe
Madrid, Spain

B. Canovas-Segura
Faculty of Computer Science
University of Murcia
Murcia, Spain

G. Stiglic
Faculty of Health Sciences
University of Maribor
Maribor, Slovenia

R. Marin
Faculty of Computer Science
University of Murcia
Murcia, Spain

***Abstract*—Health-care associated infections (HAI) are infections that are not present or incubated at the time of admission to hospital. HAI are one of the major causes of morbidity and mortality among immunocompromised patients and have an important economic impact. The bacteria isolated in microbiology cultures can be treated with a limited combination of antibiotics owing to their resistance to many groups of antibiotics, which represents a major challenge. This paper focuses on the problem of vancomycin resistant *Enterococcus faecium* (VREfm) due to its high prevalence in HAI, multidrug resistance and ability to survive under intense selective pressure. We use the subgroup discovery technique to identify target populations with high risk clinical factors for VREfm infections that we shall be able to incorporate into a clinical decision support system for antimicrobial stewardship program. The dataset used contained 201 susceptibility tests with *Enterococcus faecium* from a University Hospital in years 2014 and 2015. The clinician evaluated and discussed the knowledge reported by the most interesting subgroups based on their positive predictive value and sensitivity.**

***Index Terms*—**

Algoritmos de aprendizaje automático para predicción de niveles de niebla usando ventanas estáticas y dinámicas.

M. Díaz-Lozano[a], D. Guijo-Rubio[a], P. A. Gutiérrez[a], C. Casanova-Mateo[b,c], S. Salcedo-Sanz[d], C. Hervás-Martínez[a]

***Resumen*—Los eventos de muy baja visibilidad producidos por niebla son un problema recurrente en ciertas zonas cercanas a ríos y grandes montañas, que afectan fuertemente a la actividad humana en diferentes aspectos. Este tipo de eventos pueden llegar a suponer costes materiales e incluso humanos muy importantes. Uno de los sectores más influenciados por las condiciones de muy baja visibilidad son los medios de transporte, fundamentalmente el transporte aéreo, cuya actividad se ve seriamente mermada, provocando retrasos, cancelaciones y, en el peor de los casos, terribles accidentes. En el aeropuerto de Valladolid son muy frecuentes las situaciones de baja visibilidad por niebla, especialmente en los meses considerados de invierno (noviembre, diciembre, enero y febrero). Esto afecta de forma directa a la manera en la que operan los vuelos de este aeropuerto. De esta forma, es muy importante conocer las posibles condiciones de niebla a corto plazo para aplicar procedimientos de seguridad y organización dentro del aeropuerto. En el presente artículo se propone el uso de diferentes modelos de ventanas dinámicas y estáticas junto con clasificadores de aprendizaje automático, para la predicción de niveles de niebla. En lugar de abordar el problema como una tarea de regresión, la variable de interés para la caracterización del nivel de visibilidad en el aeropuerto (Rango Visual de Pista, RVR) se discretiza en 3 categorías, lo que aporta mayor robustez a los modelos de clasificación obtenidos. Los resultados indican que una combinación de ventana dinámica con ventana estática, junto con modelos de clasificación basados en *Gradient Boosted Trees* es la metodología que proporciona los mejores resultados.**

***Palabras clave*—Series temporales, Eventos de baja visibilidad, modelos autorregresivos, predicción.**

I. Introducción

La niebla es un fenómeno meteorológico que consiste en la aparición de gotas de agua en suspensión en forma de gotas, lo suficientemente pequeñas como para que la gravedad terrestre no las atraiga hacia la superficie. Este fenómeno se manifiesta a nivel de suelo, pudiendo considerarse una nube a muy baja altura. Su aparición puede deberse a diferentes causas, como la evaporación de la humedad del suelo o a la expedición de vapor por parte de vegetación o de grandes masas de agua [1]. En cualquier caso, su aparición está íntimamente ligada a la disminución de las condiciones de visibilidad en la superficie. Estos eventos de baja visibilidad suponen un riesgo particular para el tráfico aéreo, marítimo y terrestre, provocando interrupciones y problemas cuyos costes humanos se han llegado a comparar a los causados por tormentas y tornados [2]. En las operaciones llevadas a cabo en los aeropuertos, el trafico de vuelos se ve fuertemente afectado en estas condiciones [3], [4], debiendo ampliar el tiempo entre aterrizajes y despegues y pudiendo provocar retrasos y cancelaciones. Por esta razón, el personal de los aeropuertos necesita conocer con cierta precisión y antelación si en un futuro cercano tendrán que trabajar con condiciones de baja visibilidad por niebla, para activar los protocolos necesarios en su caso, y tratar de mitigar este tipo de situaciones problemáticas.

La predicción de eventos futuros tiene interés en la mayoría de campos de estudio, creándose multitud de modelos destinados para este cometido [5], [6], aplicados con éxito a problemas de predicción reales. Todos estos modelos suelen estar basados en análisis de series temporales, normalmente sobre codificación real. Los eventos de niebla que provocan baja visibilidad son un problema recurrente en multitud de aeropuertos en todo el mundo, y su predicción se ha enfocado mediante diversas técnicas: en Perth, el aeropuerto más grande de la costa sudoeste de Australia, se desarrolló un modelo que aplicaba lógica difusa para conseguir una predicción precisa [7]; en el aeropuerto de Calcuta, India, se abordó el problema utilizando árboles de decisión para identificar los parámetros más importantes que influyen en la visibilidad, realizando predicciones mediante una red neuronal [8]; en el aeropuerto de Valladolid, situado al noroeste de España, se afrontó este problema haciendo uso de diversos algoritmos de aprendizaje automático para regresión como máquinas de vectores soporte, preprocesando la serie antes de aplicarlos [9].

El objetivo de este artículo es proponer un modelo de

[a]: Dpto. de Informática y Análisis Numérico, Universidad de Córdoba, Córdoba, España. E-mail: {i42dilom, dguijo, pagutierrez, chervas}@uco.es

[b]: LATUV: Laboratorio de Teledetección, Universidad de Valladolid, Valladolid, España.

[c]: Dpto. de Ingeniería Civil: Construcción, Infraestructura y Transporte, Universidad Politécnica de Madrid, Madrid, España.

[d]: Dpto. de Teoría de la Señal y Comunicaciones, Universidad de Alcalá, Alcalá de Henares, Madrid, España. E-mail: sancho.salcedo@uah.es

Este trabajo ha sido desarrollado con la financiación de los proyectos TIN2017-85887-C2-1-P, TIN2017-85887-C2-2-P y TIN2017-90567-REDT del Ministerio de Economía y Competitividad de España (MINECO) y fondos FEDER. La investigación de David Guijo Rubio ha sido subvencionada por el proyecto PI15/01570 de la Fundación de Investigación Biomédica (FIBICO) y por el Programa Predoctoral FPU (Ministerio de Educación y Ciencia de España), referencia de beca FPU16/02128.

predicción horario basado en el preprocesamiento de series temporales categóricas, donde los eventos temporales son tres diferentes condiciones atmosféricas relacionadas con la aparición de eventos de baja visibilidad por niebla: no-niebla, neblina y niebla. Este preprocesamiento se lleva a cabo mediante el uso de diversos tipos de ventanas de valores pasados de las series temporales consideradas, de forma que la información obtenida a partir de cada serie pueda ser usada en el entrenamiento de cualquier modelo de aprendizaje automático.

En la siguiente sección (sección II), se presenta la base de datos considerada. En la sección III, se detallan los modelos de ventana propuestos para preprocesar los datos y para realizar las predicciones. En la sección IV, se detalla la configuración de la experimentación y los resultados obtenidos. Por último, la sección V incluye las conclusiones obtenidas a partir de los resultados conseguidos.

II. Base de datos

II-A. Origen de los datos

Los datos utilizados en este artículo han sido recopilados mediante un sistema localizado en el aeropuerto de Valladolid y perteneciente a la Agencia Estatal de Meteorología Española (AEMET). Dicho sistema provee información relevante sobre las condiciones meteorológicas al personal de los aeropuertos (pilotos, controladores y personal de tierra). De forma horaria, el sistema recoge información sobre diferentes factores meteorológicos, tales como la temperatura o la humedad. Dichos datos son recogidos por sensores y, por tanto, no existen datos perdidos, pudiéndose considerar cada variable meteorológica obtenida como una serie temporal. Las variables recogidas por este sistema son las siguientes: Temperatura (grados Celsius), Humedad relativa (%), Velocidad del viento (m/s), Dirección del viento (grados), Presión reducida al nivel del mar (QNH, hPa) y Rango Visual de Pista (*Runaway Visual Range, RVR*), que es la variable objetivo a predecir y se mide en metros.

El RVR se obtiene a partir de la media ponderada de tres sensores de visibilidad (visibilímetros), colocados a diferentes alturas de la pista (zona de toma de tierra, media pista y zona de parada). Además, las condiciones de niebla pueden modelarse mediante la combinación del resto de variables meteorológicas medidas, por estar ambas intrínsecamente relacionadas. Los experimentos llevados a cabo en este artículo se han realizado con datos horarios de 8 años completos, concretamente desde el 1 de noviembre de 2009 hasta el 31 de diciembre de 2016. De la totalidad de datos, el 70 % inicial ha sido usado para la fase de entrenamiento y el último 30 % para la fase de generalización.

II-B. Umbralización del RVR

Los visibilímetros utilizados en el aeropuerto de Valladolid solo obtienen medida hasta 2000m, considerando situaciones de visibilidad óptima por encima de esta medida (es decir marcan 2000m incluso cuando el valor real de visibilidad es más alto). Esto es debido a que los valores de visibilidad por encima de 2000m son considerados óptimos, y por tanto no son relevantes para la gestión de situaciones de baja visibilidad en el aeropuerto. En la experimentación diseñada en este artículo se propone un enfoque de predicción categórico mediante 3 clases, en el que a cada hora se le asigna una de las 3 posibles clases (alta, media o baja en relación al valor de RVR, asociado a las condiciones de niebla). Los umbrales utilizados para la discretización de los valores de RVR son los siguientes:

$$\text{Clase} = \begin{cases} \text{niebla}, & \text{si } RVR < 1000, \\ \text{neblina}, & \text{si } 1000 <= RVR < 1990, \\ \text{no niebla}, & \text{si } RVR >= 1990. \end{cases}$$

De esta forma, se obtiene un problema de clasificación con 3 clases, altamente desequilibrado, debido a que las condiciones de baja visibilidad son, afortunadamente, muy minoritarias respecto a situaciones de visibilidad óptima. La Tabla I muestra la proporción de las diferentes clases en los conjuntos de entrenamiento y generalización considerados.

Tabla I
Proporción de clases

Clase	Entrenamiento	Generalización
niebla	856 (6 %)	772 (12 %)
neblina	912 (6 %)	514 (8 %)
no niebla	12294 (88 %)	4740 (80 %)
Total	14062 (70 %)	6026 (30 %)

III. Modelos utilizados

Una serie temporal se define como un conjunto de datos cronológicamente ordenados y muestreados con una frecuencia constante. Formalmente, una serie temporal unidimensional se define como:

$$Y = \{y_0, y_1, y_2, ..., y_N\},$$

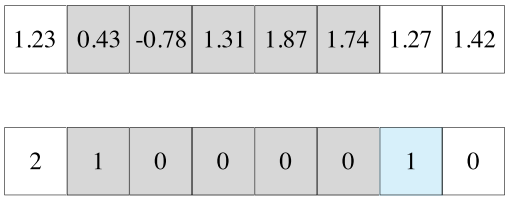
donde N es la longitud de la serie temporal.

La experimentación desarrollada en este artículo se ha llevado a cabo mediante el análisis de las series temporales descritas en la sección II-A, utilizando diferentes ventanas de valores pasados. Inicialmente, se crea un conjunto de patrones en los que la variable dependiente es el valor de la serie objetivo en el instante de tiempo a predecir y las variables independientes están formadas por la información extraída a partir de las ventanas. Con este conjunto de patrones, entrenamos cualquier modelo de aprendizaje automático, por lo que podríamos considerar que las ventanas actúan como un método de preprocesamiento.

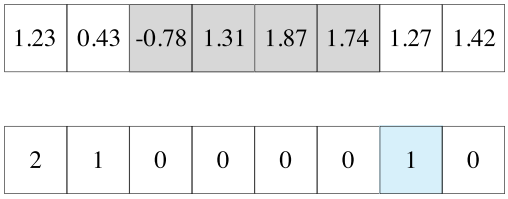
III-A. Extracción de características basada en ventanas

El análisis de series temporales mediante métodos autorregresivos permite modelar una variable en función de valores pasados, tanto de ella misma como de otras variables independientes relacionadas con el problema. En este artículo se propone el uso de 3 métodos, cada uno de los cuales limita de distinta forma la ventana de valores pasados utilizada para predecir el siguiente. Estos métodos pueden ser utilizados de forma individual, aplicando un único tipo de análisis,

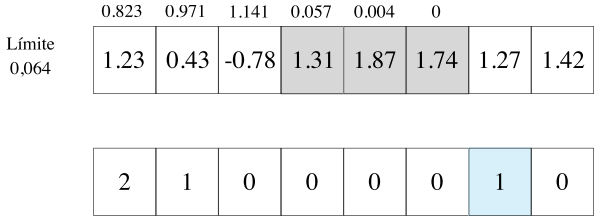
o combinada, de forma que se pueda aunar la información obtenida mediante varios tipos de ventanas. La Figura 1 muestra gráficamente el funcionamiento de estos 3 métodos en un problema sintético: una serie temporal categórica asociada a una variable dependiente a predecir y una serie temporal de valores reales asociada a una variable independiente, correlada con la primera. En dicha figura, el valor de la serie dependiente a predecir se encuentra sombreado en azul, mientras que las ventanas utilizadas para ello se encuentran sombreadas en gris.



(a) Ventana fija con *p=5*



(b) Ventana dinámica basada en cambio de clase



(c) Ventana dinámica basada en varianza con *r=0.1*

Figura 1. Distintos tipos de ventana propuestos.

III-A1. Modelo de ventana fija (VF): utiliza un número de instantes pasados constante para la predicción de cada valor de la serie temporal. Esta ventana es la que emplean los modelos autorregresivos (AR) clásicos, comúnmente utilizados en Estadística. El número de instantes es un parámetro del modelo, p (o orden del modelo AR), y los resultados obtenidos mediante este preprocesamiento son muy sensibles al valor de p. El modelo AR de orden p puede definirse como:

$$X_t = c + \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t \quad (1)$$

donde c representa una constante, α_i el coeficiente correspondiente al valor de la serie temporal en el instante $t-i$ y ε_t ruido blanco, esto es, una variable aleatoria Normal de media 0 y de varianza a determinar. En la Figura 1(a) se muestra un ejemplo de este tipo de preprocesamiento haciendo uso de una ventana fija de tamaño 5. Como se puede ver en dicha figura, para prededecir el valor sombreado en azul hace uso de los 5 valores anteriores de todas las series involucradas en el problema, tanto la dependiente como las independientes.

III-A2. Modelo de ventana dinámica basada en cambio de clase (VDCC): en series temporales categóricas, este modelo crea ventanas de tamaño variable limitándolas en base a los cambios de clase. De esta forma, para predecir un determinado valor, se identifica la clase inmediatamente anterior y se añaden valores anteriores mientras la clase no cambie, creándose ventanas más grandes cuando la clase es estable (existe una racha). En este tipo de preprocesamiento, los valores pasados de la serie dependiente son utilizados de forma indirecta dado que, aunque no se utilizan sus valores, son estos los que determinan el tamaño de las ventanas que se crean. En la Figura 1(b) se representa gráficamente la ventana creada para un determinado valor. En este ejemplo, la clase inmediatamente anterior al valor a analizar es 0, por lo que la ventana se extenderá hacia detrás, hasta que se encuentre una categoría distinta de 0, creando en este caso una ventana de 4 elementos. Tras ello, la información de las muestras de cada serie independiente que caigan dentro de la ventana se resume mediante las métricas detalladas en la sección IV-A.

III-A3. Modelo de ventana dinámica basada en varianza (VDV): en series temporales de valores reales, proponemos este modelo que crea ventanas de tamaño variable en función de la dinámica de la serie. Se analiza, de forma individual para cada serie temporal, la varianza de los valores incluidos en la ventana. Se añaden valores previos a la ventana hasta que se alcanza un determinado límite de varianza, que se establece como un porcentaje de la varianza total de la serie. Dicho porcentaje es un parámetro del modelo, r. La idea subyacente es que, cuando la varianza es demasiado alta, puede no tener sentido intentar resumir la información incluida en la misma. Además, dado que cada serie temporal se analiza de forma independiente, se pueden obtener ventanas de distinto tamaño para cada variable de entrada, cuando haya muchas variables independientes (como sucede en el problema considerado). La Figura 1(c) muestra un ejemplo de este procesamiento con un porcentaje de varianza total del 10 %. Este procesamiento no puede aplicarse sobre la serie categórica. En la serie de valores reales, se muestra, sobre cada valor, la varianza parcial, de forma tal que la ventana utilizada para predecir el valor sombreado en azul crecerá mientras que la varianza parcial sea menor que el límite (10 % de la total). Las ventanas serán resumidas mediante las métricas de la sección IV-A.

III-B. Clasificadores

La información obtenida tras el uso de las diferentes combinaciones de ventanas será utilizada para entrenar clasificadores que realicen la predicción de la categoría. En una primera aproximación, hemos considerado algunos de los modelos más clásicos de los incluidos en *scikit-learn* [15], una librería de código abierto implementada en *Python*. De esta forma, los clasificadores considerados son: Regresión logística (RL) [11],

Árboles de decisión (AD) [12] y conjuntos de clasificadores (concretamente *RandomForest (RF)* [13] y *GradientBoosting-Classifier (GB)* [14]).

IV. Experimentación y resultados

IV-A. Configuración

Como ya se especificó en la sección II-A, un 70 % de los datos se ha empleado como conjunto de entrenamiento, mientras que el 30 % formó el conjunto de *test*.

Para ambas ventanas dinámicas (tanto la basada en varianza como en el cambio de clase), se ha elegido resumir la información de las ventanas creadas utilizando dos estadísticos:

- La media aritmética ($\overline{W_s}$) de la ventana, definida como:

$$\overline{W_s} = \frac{1}{s} \sum_{y \in W_s} y$$

  donde s es el número de valores de la ventana y W_s la ventana creada.
- La varianza ($S^2_{W_s}$), utilizada para obtener una medida de dispersión de los valores contenidos en la ventana y definida como:

$$S^2_{W_s} = \frac{1}{s-1} \sum_{y \in W_s} (y - \overline{W_s})^2.$$

Los parámetros de los métodos de preprocesamiento y de los clasificadores han sido optimizados mediante una validación cruzada anidada de tipo 5-*fold*. A continuación se indican los valores explorados. Para el análisis de ventana fija, se ha utilizado un numero de muestras previas $p \in \{1, 2, 3, \ldots, 6\}$. Para el análisis de ventana dinámica basado en varianza, se ha optado por optimizar la proporción de varianza total utilizada para limitar las ventanas según $r \in \{0,1; 0{,}2; 0{,}3; \ldots; 0{,}6\}$.

Los parámetros de los modelos de clasificación también han sido optimizados mediante validación cruzada. Para RL, se ha optimizado la intensidad de la regularización mediante el parámetro $C \in \{10^{-3}, 10^{-2}, \ldots, 10^2, 10^3\}$. Para los AD se ha optimizado la función para medir la calidad de una división, utilizando el coeficiente *Gini* y la entropía. En el caso de RF, se ha optimizado el número de árboles en cada bosque $n \in \{10, 50, 100\}$, y, para GB, el número de árboles utilizado en el algoritmo ha sido $n \in \{50, 100, 150, 200\}$ y el ratio de aprendizaje $l \in \{0{,}05; 0{,}1; 0{,}2; 0{,}3\}$.

IV-B. Métricas de rendimiento

Existen multitud de métricas para evaluar la calidad de una clasificación obtenida. Una de las más comunes es el porcentaje de patrones bien clasificados (*Correct Classified Ratio, CCR*) con la que se obtiene una medida de rendimiento global. No obstante, el problema a tratar en este artículo es altamente desequilibrado, haciendo que cualquier clasificador trivial que clasifique todos los patrones en la clase mayoritaria obtenga buena puntuación en *CCR*. Por ello, el rendimiento ha de medirse mediante otras métricas. Hemos elegido la Media Geométrica de Sensibilidades (*Geometric Mean of the Sensitivities, GMS*), una medida de rendimiento que tiene en cuenta la precisión de la clasificación en todas las clases. El GMS se define como:

$$GMS = \sqrt[n]{\prod_{i=1}^{n} S_i}$$

donde n es el número de clases y S_i la sensibilidad de la clase i-ésima, es decir:

$$S_i = \frac{n_i}{N_i}$$

siendo n_i el número de instancias de la clase *i* correctamente clasificadas y N_i el número total de instancias de la clase *i*. Es una medida a maximizar.

Por otro lado, la naturaleza ordinal de nuestro problema (las categorías están organizadas en una escala ordinal) hace que sea necesario obtener una medida de error que penalice más algunos tipos de errores (por ejemplo, la confusión de "no niebla" con "niebla alta"), para lo que se utilizará la Media de Errores Absolutos Medios (*Average Mean Absolute Error, AMAE*). El $AMAE$ se define como:

$$AMAE = \frac{1}{n} \sum_{i=1}^{n} MAE_i$$

donde n es el número de clases y MAE_i es la desviación media producida con respecto a la clase real considerando solo la clase *i*. Se define como:

$$MAE_i = \frac{1}{N_i} \sum_{i=1}^{N_i} |C(y_i) - C(\hat{y_i})|$$

siendo N_i el número de instancias de la clase *i*, $C(y_i)$ la clase real de la instancia i y $C(\hat{y_i})$ la clase predicha para la instancia *i*, representadas mediante valores numéricos (es decir, $C(y_i) = 0$ para no niebla, $C(y_i) = 1$ para no neblina y $C(y_i) = 2$ para niebla). Al tratarse de una medida de error, es una métrica a minimizar.

Tabla II
Resultados de *test* en GMS

Tipos de ventanas	Clasificadores			
	GB	RF	RL	AD
VF	*0.624*	*0.453*	*0.567*	*0.518*
VDCC	0.286	0.375	0.297	0.405
VDV	0.0	0.072	0.278	0.286
VF+VDCC	0.617	**0.492**	0.547	**0.540**
VF+VDV	**0.656**	0.355	0.533	0.467
VDCC+VDV	0.475	0.368	0.442	0.425
VF+VDCC+VDV	0.620	0.447	**0.591**	0.469

IV-C. Resultados

Las Tablas II y III incluyen todos los resultados obtenidos, para GMS y $AMAE$, respectivamente. Por filas, se muestran las combinaciones de ventanas (ventana fija, VF, ventana dinámica basada en cambio de clase, VCCC, y ventana dinámica basada en varianza, VDV, más todas las combinaciones posibles, VF+VDCC, VF+VDV, VDCC+VDV y VF+VDCC+VDV) y por columnas cada uno de los 4

Tabla III
Resultados de *test* en $AMAE$

Tipos de ventanas	Clasificadores			
	GB	RF	RL	AD
VF	0.317	**0.373**	*0.314*	*0.437*
VDCC	0.538	0.592	0.419	0.660
VDV	*0.312*	0.406	0.333	0.442
VF+VDCC	0.314	0.385	0.318	**0.426**
VF+VDV	**0.304**	0.396	0.323	0.457
VDCC+VDV	0.349	0.412	**0.306**	0.474
VF+VDCC+VDV	*0.312*	*0.384*	0.316	0.461

clasificadores considerados. Los mejores resultados por cada clasificador se muestran en negrita, los segundos mejores en cursiva. El mejor resultado global se encuentra doblemente subrayado, mientras que el segundo mejor está marcado con subrayado simple.

Atendiendo a dichas tablas, se pueden obtener diversas conclusiones:

- En términos de *GMS*, los mejores resultados para cada modelo se obtienen siempre mediante un preprocesamiento en el que intervienen dos o más ventanas. En términos de *AMAE*, a excepción del modelo *RandomForest*, ocurre lo mismo, lo que demuestra que el uso combinado de distintas ventanas mejora ampliamente los resultados obtenidos, en comparación con los obtenidos mediante el uso individual de los mismos.
- Al optimizar el *AMAE*, conviene utilizar la combinación de ventanas que produce el segundo mejor resultado, debido a que el tamaño del patrón que genera el uso de VDCC+VDV será mucho más pequeño que usar VF+VDV. Esto se debe a que el uso de ventanas dinámicas resume las muestras de las ventanas en dos métricas (media y varianza), haciendo que como máximo los patrones cuenten con 20 características. Atendiendo al *grid* de parámetros utilizados para ventana fija, los patrones generados con este método junto con la ventana dinámica pueden ser de hasta 40 características, ralentizando bastante el proceso de entrenamiento.
- La combinación VF+VDV consigue los mejores resultados, tanto en *GMS* como *AMAE*, mediante el uso del clasificador *GradientBoosting*. La ventana basada en varianza provee de una capacidad dinámica en el análisis de las series independientes, adaptándose a cada serie de forma individual, lo que combinado con una ventana fija que utilice muestras previas devuelva el mejor resultado.
- La VF siempre obtiene los segundos mejores resultados en *GMS*. Este aspecto era esperable, dado que existe una alta persistencia en la serie categórica dependiente a predecir. Así, en la optimización del tamaño de ventana fija, se incluye la posibilidad de crear ventanas de tamaño 1, haciendo que el modelo tienda a predecir la salida únicamente en función de un instante pasado. Aunque esto devuelva unos resultados aceptables, no es deseable, dado que este tipo de modelos nunca serán capaces de detectar un cambio de clase en la serie.
- El uso de un preprocesamiento VDV aislado devuelve unos resultados pobres. Esto es debido a que el uso de este tipo de ventanas está orientado a series reales, y al tratar con una serie dependiente de naturaleza categórica, la serie a predecir se modela únicamente en función de series independientes, perdiendo una parte importante de la información.
- Los resultados de *GMS* obtenidos con el uso de VDCC aislado, a pesar de utilizar solamente los valores de las series independientes, son mejores que los de VDV. La razón de este resultado es que, a pesar de que VDCC no usa los valores de la serie dependiente de forma directa, son éstos los que determinan el tamaño de ventana para cada muestra, por lo que de forma indirecta se está utilizando una información inherente a la serie dependiente.

Por último, la Figura 2 muestra una comparación de un fragmento de 96 horas de la serie real y la serie predicha por *GradientBoosting* con la mejor combinación de ventanas. Como puede observarse, el modelo funciona de forma aceptable, prediciendo correctamente los cambios producidos en las etiquetas reales. Tan solo en la última parte del gráfico se observan algunos artificios introducidos por el método.

V. Conclusiones

Este artículo evalúa el uso de diferentes clasificadores de aprendizaje automático para predecir eventos de baja visibilidad por niebla en el aeropuerto de Valladolid. Se realiza una predicción categórica basada en tres posibles tipos de situaciones (no niebla, neblina y niebla), utilizando como entrada valores pasados de un conjunto de variables meteorológicas medidas en el aeropuerto. Proponemos considerar distintos tipos de ventanas a la hora de analizar los valores pasados: ventanas fijas junto con dos tipos de ventanas dinámicas (una basada en cambios de la categoría de los días pasados analizados y otra basada en varianza de la variable independiente examinada). La ventaja fundamental de estos métodos dinámicos es que evitan prefijar el tamaño de la ventana, pudiéndose adaptar de forma dinámica a la serie temporal considerada.

Los resultados obtenidos indican que la combinación de ventanas fijas y dinámicas (especialmente aquella basada en varianza), junto con el conjunto de clasificadores *GradientBoosting* obtiene los mejores resultados, con valores de $AMAE$ cercanos a $0{,}3$ (es decir, la predicción difiere en $0{,}3$ categorías, en media, con respecto al valor real) y un GMS mayor que el $65\,\%$. Teniendo en cuenta estos resultados, se puede concluir que los modelos obtenidos pueden mejorar la seguridad y la eficiencia de las operaciones aeronáuticas que se llevan a cabo en aeropuertos bajo condiciones de baja visibilidad por niebla. Como trabajo futuro, planteamos el uso de clasificadores ordinales, dada la naturaleza ordinal de la clases consideradas.

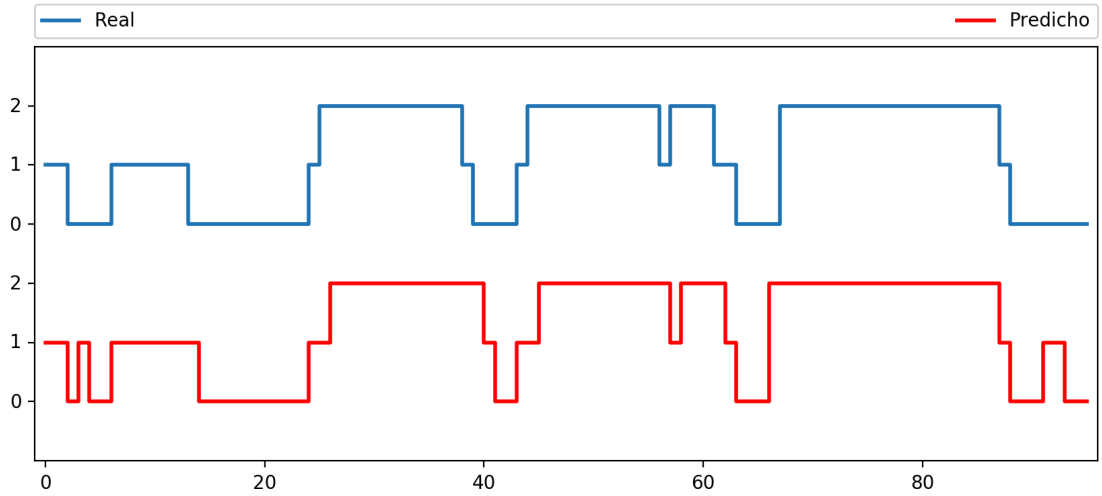


Figura 2. Comparación de etiquetas reales con las predichas por *GradientBoosting* con VF +VDV durante 96 horas. La clase 1 corresponde a no niebla, la clase 2 a neblina y la clase 3 a niebla.

Referencias

[1] D. Koracin, C. E. Dorman, J. M. Lewis, J. G. Hudson, E. M. Wilcox and A. Torregrosa, "Marine fog: A review," Atmospheric Research, vol. 143, pp. 142-175, 2014.

[2] I. Gultepe, et al., "Fog Research: A Review of Past Achievements and Future Perspectives," Pure and Applied Geophysics, vol. 164, pp. 1121-1159, 2007.

[3] H. Huang and C. Chen, "Climatological aspects of dense fog at Urumqi Diwopu International Airport and its impacts on flight on-time performance," Natural Hazards, vol. 81, pp. 1091-1106, 2016.

[4] N. Fedorova et al. "Fog Events at Maceio Airport on the Northern Coast of Brazil During 2002–2005 and 2007," Pure and Applied Geophysics, vol. 172, pp. 2727-2749, 2015.

[5] L. Zhenling et al., "Novel forecasting model based on improved wavelet transform, informative feature selection, and hybrid support vector machine on wind power forecasting," Journal of Ambient Intelligence and Humanized Computing, pp. 1-13, 2018.

[6] G. Noradin, A. Adel, S. Hossein and A. Oveis, "A new prediction model based on multi-block forecast engine in smart grid," Journal of Ambient Intelligence and Humanized Computing, pp. 1-16, 2017.

[7] Miao, Y. and Potts, R. and Huang, X., and Elliott, G. and Rivett, R., "A Fuzzy Logic Fog Forecasting Model for Perth Airport," Pure and Applied Geophysics, vol. 169, pp. 1107-1119, 2012.

[8] D. Duta and S. Chaudhuri, "Nowcasting visibility during wintertime fog over the airport of a metropolis of India: decision tree algorithm and artificial neural network approach," Natural Hazards, vol. 75, pp. 1349-1368, 2015.

[9] L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo, E. Cerro-Prada and S. Salcedo-Sanz, "Efficient Prediction of Low-Visibility Events at Airports Using Machine-Learning Regression," Boundary-Layer Meteorology, vol. 165, pp. 349-370, 2017.

[10] J. Yan-jie, G. Liang-peng, C. Xiao-shi and G. Wei-hong, "Strategies for multi-step-ahead available parking spaces forecasting based on wavelet transform," Journal of Central South University, vol. 24, pp. 1503-1512, 2017.

[11] J. P. Chao-Ying, L. L. Kuk and M. I. Gary, "An Introduction to Logistic Regression Analysis and Reporting," The Journal of Educational Research, vol. 96, pp. 3–14, 2002.

[12] L. Wei-Yin, "Fifty Years of Classification and Regression Trees," International Statistical Review, pp. 329-348, 2002.

[13] L. Breiman, "Random forests," Machine Learning, vol. 45(1), pp. 5-32, 2001.

[14] J. F. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, pp. 1189-1232, 2001.

[15] F. Pedregosa, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

Intelligent Management of Measurement Units Equivalences in Food Databases*

*Note: The full contents of this paper have been published in the volume *Lecture Notes in Artificial Intelligence 11160* (LNAI 11160)

Beatriz Sevilla-Villanueva, Karina Gibert
Department of Statistics and Operations Research
Universitat Politècnica de Catalunya-BarcelonaTech (UPC)
Barcelona, Spain

Miquel Sànchez-Marrè
Department of Computer Science
Universitat Politècnica de Catalunya-BarcelonaTech (UPC)
Barcelona, Spain

***Abstract*—It is currently well-known that diet plays an important role in the promotion of healthy lifestyle and the prevention of chronic diseases. The Diet4You project is conceived to support the creation of an intelligent decision support system that provides personalized menus fitting a nutritional plan and taking into account the characteristics, needs and preferences of the person. The system involves a background food database, recording a collection of foods and prepared dishes with their standard portions as well as their nutritional decomposition in different food families.This DB is used to search the best combination of dishes approaching the total intake of different nutrients specified in the prescribed nutritional plan. The available background databases, specify the quantities of standard portions of several foods based on different measurement units which are not standardized, and it happens that the weight specified by one cup of melon is different from that of one cup of berries, among others. This arises the need of applying variable conversion factors to the dish description, before assessing whereas the total quantities of a certain menu fit well to the prescription. In this paper, a knowledge based approach is presented to the automatically management. An annotated reference food ontology is built on the basis of additional documentation. However the granularity of the information provided is heterogeneous and non exhaustive. The ontology-based missing values imputation is presented to overcome this limitations.**

***Index Terms*—Ontology, Missing Imputation, Database**

Predicción de delincuencia con datos públicos

1st Roberto Cuesta Calvo
Servicios Técnicos. Servicio Informática
Dirección General Guardia Civil
Madrid, España
rcuesta@guardiacivil.es

2nd Jesús Maudes Raedo
Departamento de Ingeniería Civil
Universidad de Burgos
Burgos, España
jmaudes@ubu.es

3rd José-Francisco, Díez-Pastor
Departamento de Ingeniería Civil
Universidad de Burgos
Burgos, España
jfdpastor@ubu.es

4th Ivan Arjona
Departamento de Ingeniería Civil
Universidad de Burgos
Burgos, España
email iaa0037@alu.ubu.es

***Resumen*—La concentración de recursos policiales en lugares considerados conflictivos contribuye a la reducción de la criminalidad en los mismos y a la optimización de esos recursos. En este artículo se presenta la utilización de técnicas de regresión para predecir el número de hechos delictivos en los municipios españoles. Para ello, se ha generado un conjunto de datos que fusiona los datos de la Guardia Civil con datos públicos sobre la estructura demográfica y las tendencias de voto en los municipios. El mejor regresor obtenido (i.e., Random Forests) alcanza con estos datos un RRSE (ráiz del error cuadrático relativo) del 41,23 %, y abre el camino para seguir incorporando datos públicos de otro tipo que tengan un mayor poder predictivo. Asimismo, se han utilizado reglas M5Rules para interpretar en lo posible los resultados.**

***Index Terms*—datos públicos, minería de datos, predicción de hechos**

I. Introducción

Para toda Fuerza y Cuerpo de Seguridad del Estado encargado de velar por los derechos y libertades de los ciudadanos se establece el concepto de territorialidad como un concepto clave en la búsqueda de maximizar la eficacia de sus recursos.

Es por ello que es constante el estudio estadístico de la criminalidad para optimizar la disposición de la fuerza sobre el terreno. Este estudio, constante en el tiempo, se acrecenta, aún mucho más, en épocas de crisis económica, pues es aquí donde existe una tendencia marcada en reducir recursos precisamente cuando, por diferentes y obvios motivos, existe riesgo de aumentar la criminalidad.

Si se deja a un lado la investigación en ciberdelincuencia y fraudes, la utilización de técnicas de aprendizaje automático en la predicción de delitos comunes ha sido muy escasa. Sólo en los últimos tiempos empiezan a aparecer algunos trabajos prometedores en esta línea. El enfoque del presente trabajo consiste en relacionar datos demográficos y de tendencia de voto en los municipios españoles, con los hechos delictivos cometidos durante un año; para así predecir la criminalidad de los municipios en función de dichas variables.

Trabajo parcialmente financiado por el proyecto TIN2015-67534-P del Ministerio de Economía Industria y Competitividad.

En este sentido, el estudio que se asemeja más al presentado en este artículo es quizás el de Alves y otros [1], que aplican técnicas de regresión con Random Forests [5] para predecir la cantidad de homicidios urbanos en Brasil a través de los datos sociológicos y demográficos de las ciudades. Las predicciones alcanzan un coeficiente de determinación de 0.97. Los autores apuntan al desempleo y el analfabetismo como las principales variables que utiliza su modelo predictivo.

Existen otros estudios en ámbitos geográficos más reducidos. En [17] se estudian cuales son las zonas conflictivas de un distrito policial al objeto de planificar la acción de las patrullas de calle. Se trabaja con los datos de Los Angeles (EEUU) y Kent (GB), y utilizan una aproximación basada en series temporales para estudiar la evolución de esas zonas.

En [15] se utilizan SVMs para predecir si una zona es conflictiva o no, en Columbus (Ohio) y St. Luis (Missouri).

En [7] se analizan las zonas de riesgo de crímenes sexuales en el campus de Charlottesville de la Universidad de Virginia. Se utilizan los clasificadores regresión logística y Random Forests [5] para clasificar un punto como conflictivo o no. Establece como variables mas importantes la proximidad de personas con antecedentes en violencia sexual y de residencias con fraternidades estudiantiles. También utilizan una aproximación de series temporales para analizar el intervalo horario, dia de la semana y época del año con mayor riesgo; así como la influencia de las condiciones meteorológicas, hallando la temperatura como factor climatológico más determinante. Utilizan *Kernel Density Estimation* (KDE) para comparar las probabilidades de riesgo en los distintos periodos de tiempo.

También se han utilizado técnicas de *Deep Learning* en la predicción de crímenes [14]. En este caso, las redes profundas predicen localizaciones y momentos probables de criminalidad en la ciudad de Chicago. Este estudio fusiona datos socioeconómicos con los datos policiales, además de datos climáticos. Otro trabajo en la misma línea para la ciudad de Manila es el de Báculo y otros [3]; en esta ocasión son las Redes Bayesianas el algoritmo de clasificación de entre los testados que mejores resultados ofrece.

El estudio que se presenta en el presente trabajo abarca todo tipo de delitos en la geografía Española. A diferencia de buena parte de los trabajos anteriormente revisados, no considera la evolución temporal de los hechos delictivos; ya que se centra en predicciones para todo el año; y además del uso de datos demográficos, presenta como novedad el uso de datos de preferencias políticas obtenidos a partir de los resultados electorales. Los resultados obtenidos son un primer paso dentro un proyecto que pretende integrar más datos públicos para mejorar las predicciones, pero que en el estado actual ya ofrece unos resultados interesantes.

El artículo se estructura como sigue; en la sección II se describen los datos y su procedencia, en la sección III se describen los experimentos con distintas técnicas de regresión para predecir los hechos delictivos, la sección IV trata de interpretar los resultados obtenidos, y finalmente en V se muestran las conclusiones y líneas futuras.

II. Obtención y descripción de los datos

Para este trabajo se han cruzado datos públicos de organismos oficiales con estadísticas de la Secretaría de Estado de seguridad a través de la Dirección General de la Guardia Civil.

Se entiende como datos públicos o abiertos aquellos que deben estar disponibles de manera libre, para acceder, utilizar, modificar y publicar sin restricciones de *copyright* [19].

Este trabajo se aprovecha de propuestas como la 'Iniciativa Aporta' [2] que promueve la apertura de información en el sector público en España. Esta iniciativa tiene el objetivo de favorecer el desarrollo de la reutilización de la información del sector público y ayudar a las administraciones para que publiquen sus datos de acuerdo al marco legislativo vigente.

Los gobiernos tienen la capacidad de obtener grandes cantidad de información sobre la población a través de varios organismos (como podría ser el *Instituto Nacional de Estadística*).

En este trabajo cada instancia del conjunto de datos representa un municipio. El número de registros del conjunto es 8.125, con un total de 124 atributos sin contar la clase (i.e., número de hechos delictivos). Se han utilizado dos fuentes públicas:

1. Instituto Nacional de Estadística (INE). Estadísticas del año 2016, correspondientes a lugar de nacimiento y rangos de edad. Un total de 114 atributos. Lugar de nacimiento (51 atributos), Rangos de edad (63 atributos).
2. Ministerio de Interior. Datos electorales, elecciones al Congreso (Junio 2016). Un total de 10 atributos.

Los atributos de lugar de nacimiento se distribuyen en 17 categorías, cada una de ellas desglosada en 3 sub-categorías (mujeres y hombres, solo mujeres, solo hombres), las categorías son: 1) Total, 2) Españoles, 3) Nacidos en España, 4) En la misma Comunidad Autónoma, 5) Misma Comunidad Autónoma. Misma Provincia, 6) Misma Comunidad Autónoma. Misma Provincia. Mismo Municipio, 7) Misma Comunidad Autónoma.Misma Provincia. Distinto Municipio, 8) Misma Comunidad Autónoma. Distinta Provincia, 9) En distinta Comunidad Autónoma, 10) Nacidos en el Extranjero, 11) Nacionalidad extranjera, 12) Europa, 13) Unión Europea, 14) África, 15) América, 16) Asia, 17) Oceanía, Apátridas y Resto.

Los atributos de rangos de edad se distribuyen en 21 categorías (0-4 años, 5-9 años, ..., 90-94 años, 95-99 años, más de 100 años) cada una de ellas desglosada en 3 sub-categorías (mujeres y hombres, solo mujeres, solo hombres).

El número de votos de cada formación se ha agrupado en 10 categorías (Extrema Izquierda, Izquierda, Centro Izquierda, Centro, Centro Derecha, Derecha, Extrema Derecha, Otros, En blanco y Nulos.). En búsqueda optimizar la objetividad de las conclusiones finales estas categorías fueron seleccionadas y categorizadas por fuentes abiertas y externas al personal de este estudio, con el objeto de no introducir subjetividades que pudiesen introducir errores o inducir a conclusiones erróneas.

En cuanto a la clase, en los 8.125 municipios evaluados se registraron un total de 36.806.873 hechos, de los cuales, los más comunes fueron: Delito de hurto (11,5 %), Infracción por el consumo o la tenencia de drogas en lugares públicos (9,4 %), Robo con fuerza (8,8 %), Infracciones al reglamento de vehículos (5,4 %), Infracciones al reglamento de circulación (3,9 %) y Alcoholemia (3,7 %).

Es importante recalcar la naturaleza pública y externa a la Guardia Civil de los datos utilizados, y que dicha institución está, por tanto, al margen de cómo se han categorizado los mismos tanto en general, como en particular en lo concerniente a cómo se han agrupado los partidos políticos y a cómo se han agrupado los inmigrantes por su procedencia.

III. Análisis de los datos con métodos de regresión

El conjunto de datos de la sección anterior sufrió dos transformaciones antes de ser utilizado. En primer lugar, se normalizaron todos los atributos, excepto la variable a predecir en el intervalo [0,1]. En segundo lugar, dado que la variable a predecir representa un recuento (i.e., *Nº de hechos delictivos*), se asume que sigue una distribución de Poisson, por lo que se ha aplicado la raíz cuadrada a dicha variable.

Una vez transformados los datos se procedió a experimentar en WEKA [12] mediante validación cruzada 10×10 diversas técnicas de regresión, para así conocer la más idónea. En principio, en todos los regresores se ha utilizado la configuración por defecto de WEKA, salvo en los casos en los que a continuación se indique lo contrario.

Los regresores utilizados en el experimento se agrupan en dos familias: por un lado regresores en solitario o *singletons*, y por otro multiregresores o *ensembles*.

Entre los *singletons* se probaron:

- Árbol de decisión M5P [20]. Los M5P son árboles de decisión de la familia de los *model trees*. Estos árboles contienen una regresión lineal en los nodos hoja.
- M5Rules [13], se trata de un método que obtiene reglas de decisión a partir de árboles M5P, por lo que no se espera que den unos resultados muy distintos que el propio M5P. La razón de incluirlos es justificar su fiabilidad cara a utilizarlos en la sección IV como herramienta para interpretar los resultados.

- Regresión Lineal, optimizando el parámetro *ridge* para cada una de las cuatro versiones que se probaron, y que surgieronn de activar/desactivar la selección de variables y la eliminación de atributos colineales. Se tomo como mejor versión la que no hacía selección de variables pero si eliminaba atributos colineales.
- SVM para regresión (SVM-Reg) [21] utilizando la implementación LIBLINEAR [9] con kernel lineal y el parámetro C optimizado.
- *k*-NN. Debido al elevado número de características en el conjunto de datos, se probaron dos versiones. La primera sin selección de atributos, la segunda con selección de atributos mediante *Correlation-based Feature Subset Selection* [11]. En ambas versiones se optimizó el número de vecinos. La mejor versión resultó ser la que no hace selección de atributos, y es la que se reporta en el artículo.

Los *ensembles* probados fueron los siguientes:

- Random Forest [5].
- AdaBoost.R2 [8], se han probado tres configuraciones con función de pérdida lineal, cuadrática y exponencial. Se seleccionó la configuración con mejores resultados en la validación cruzada 10×10 (i.e., pérdida cuadrática)
- Additive Regression, que es una implementación de *Stochastic Gradient Boosting* [10]
- Bagging [4]
- Iterated Bagging [6]. En este caso, para simular 100 árboles se utilizan 20 iteraciones Bagging de 5 árboles cada una.

Todos los *ensembles* utilizan 100 árboles M5P como regresores base, excepto Random Forest, que obviamente utiliza 100 Random Trees.

La métrica utilizada para evaluar los regresores es el RRSE o ráiz del error cuadrático relativo, que para θ_i el valor verdadero a estimar para la instancia $i-$ésima, $\hat{\theta}_i$ el valor resultado de la estimación de esa instancia, y $\bar{\theta}$ el valor medio de las θ_i, estimado a través de las instancias del conjunto de entrenamiento, se define como:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta} - \theta_i)^2}}$$

Los resultados se muestran en el Cuadro I. Como se puede apreciar, los *ensembles* están a la cabeza, mientras que los métodos lineales quedan en la parte inferior. Esto puede deberse a que, como se indica en [16] la relación entre el tamaño de las ciudades y su criminalidad obedece mas a una ley potencial que a una relación lineal. De hecho, los coeficientes de correlación en la tabla muestran una correlación muy discreta cuando se utilizan modelos lineales, que mejoran considerablemente en el caso de los árboles M5P y las reglas M5Rules, los cuales también usan modelos lineales en las hojas, y alcanzan un valor aproximado de 0.9 *ensembles* utilizados.

El mejor método, tanto por RRSE, como por coeficiente de correlación, es *Random Forests*. El RRSE alcanzado, 41.23,

Método	RRSE	Coef. Corr.
Random Forests	41.23	0.91
AdaBoost R2	41.63	0.91
Bagging	42.20	0.91
Iterated Bagging	43.42	0.90
Additive Regression	44.62	0.90
M5P	47.10	0.88
M5Rules	49.41	0.87
k-NN	52.27•	0.85•
SVM-Reg	74.59•	0.67•
Regresión Lineal	82.15•	0.62•

Cuadro I
RESULTADOS DE LOS DIFERENTES MÉTODOS ORDENADOS POR RRSE. LOS • INDICAN DIFERENCIAS SIGNIFICATIVAS CON EL MEJOR MÉTODO.

parece mostrar que los datos demográficos y de intención de voto sirven para explican aceptablemente la concentración de hechos delictivos, pero quizás aún hay margen de mejora incorporando en el futuro nuevas variables al modelo.

En la tabla se ha marcado con • aquellos valores que son estadísticamente peores, con un nivel de confianza del 95 %, al compararlos con el mejor método. El test estadístico utilizado es el *corrected resampled t*-test [18], debido a su idoneidad en el caso de utilizar validación cruzada. Se aprecia que no hay diferencias entre los métodos del grupo de los *ensembles*.

IV. Interpretación de los Resultados del Análisis y Líneas de Mejora

Para descubrir las posibles líneas de mejora del modelo actual se han seguido dos caminos:

- Investigar los municipios en los que peor se comporta el modelo.
- Generar reglas con el algoritmo *M5Rules* [13], para poder interpretar el conjunto de datos.

IV-A. Municipios que peor responden al modelo

Para tener una lista ordenada de los municipios que peor responden al modelo se halló el valor absoluto de la diferencia entre el número de hechos real y el número de hechos predichos por un *Random Forest* entrenado con todos los datos del conjunto. Cierto es que esta diferencia arroja unos valores muy optimistas, en tanto las diferencias se obtienen a partir de predicciones sobre los propios datos de entrenamiento, pero se asume que esa ventaja la van a tener todos los municipios. Una vez obtenida esa diferencia en valor absoluto, se divide por el número de habitantes del municipio, para evitar que el indicador únicamente señale a los municipios más grandes. Denotaremos a este indicador como Δ/hab.

El valor máximo de Δ/hab es del 29,99 %, y se alcanza en un municipio de 230 habitantes. Hay otro pequeño municipio de 20 habitantes que alcanza un 29,75 %, otro de 105 con un 24 %, y a partir de ahí una lista de 35 pequeños municipios, el mayor de ellos con 303 habitantes, hasta llegar a *Sant Josep de sa Talaia* con 25.849 habitantes y un Δ/hab =7,32 %.

En estas localidades tan pequeñas, cuando ocurren unos pocos hechos delictivos por encima de los previstos, el indicador Δ/hab se dispara.

Es llamativo que en esta lista ordenada por Δ/hab hay una serie de municipios de más de 10.000 habitantes intercalados con estos pequeños municipios. El Cuadro II muestra los que tienen un Δ/hab por encima del 2 %.

Municipio	Nº hab	Δ/hab	Hechos	Predicción	pos
S. Josep de sa Talaia	25.849	7,32 %	5.578	3.685	38
Calvià	49.580	4,79 %	6.281	3.907	80
Torrevieja	84.213	3,18 %	7.312	4.635	149
S. Antony de Portmany	24.478	2,96 %	3.260	2.536	177
Borriana	34.643	2,29 %	2.347	1.556	254
Las Rozas de Madrid	94.471	2,16 %	4.160	2.126	276
Benicasim	17.957	2,03 %	1.356	991	301
Guardamar del Segura	15.386	2,02 %	1.280	969	304

Cuadro II
MUNICIPIOS CON MÁS DE 10.000 HABITANTES Y $\Delta/hab > 2\,\%$.
POS=POSICIÓN EN EL RANKING.

Estas localidades podrían estar siendo predichas mal debido a que en su mayoría son conocidas plazas turísticas, y su tamaño real, teniendo en cuenta los turistas, seguramente difiera mucho del tamaño por habitantes empadronados. De hecho, todas las predicciones son siempre a la baja. No obstante, los valores de Δ/hab son bastantes moderados.

Por tanto, se aprecia que una línea de mejora a futuro podría venir por incorporar al conjunto de datos características nuevas que cuantifiquen el fenómeno turístico en los municipios.

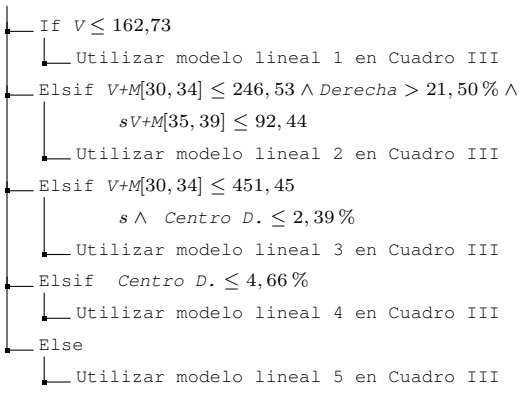
IV-B. Interpretación con M5Rules

Se ha utilizado para interpretar el conjunto de datos la generación de reglas mediante el algoritmo *M5Rules* [13]. *M5Rules* genera un árbol M5P, selecciona la mejor hoja, y haciendo AND con los nodos en el recorrido entre la raíz y dicha hoja, genera una primera regla. Después, descarta del árbol los datos que han caído en esa hoja, y vuelve a construir un nuevo M5P repitiendo todo el proceso para generar una segunda regla, y así sucesivamente va generando nuevos árboles por cada regla con los datos de entrenamiento que no son cubiertos por las reglas anteriores, hasta que finalmente a toda instancia la corresponda una regla.

La implementación WEKA de *M5Rules*, en la configuración por defecto del algoritmo, incluye un proceso de poda, de manera que elimina las reglas finales, que tienden a concentrarse en casos particulares, sustituyéndolas por un modelo lineal.

En el cuadro I de la sección I podía verse que no hay diferencias significativas en términos de *RRSE* entre el mejor método y *M5Rules*, por lo que parece una aproximación aceptable para tener una cierta interpretación de los datos desde la óptica de los árboles de decisión.

Para ello, se utilizó el conjunto de entrenamiento al completo y se generaron las siguientes cinco reglas:



V representa el número total de varones del municipio, V[x,y] representa el número de varones en el rango de edades [x,y], V+M[x,y] representa la suma de varones y mujeres en el rango de edades [x,y], *Derecha* y *Centro D* es el porcentaje de votos a formaciones de derecha y de centro derecha respectivamente. La última regla (i.e., *ELSE*) no contiene una condición lógica por que es la que proviene de la poda de las reglas menos importantes, ya comentada anteriormente.

Nótese que las reglas de la ilustración no son las originales generadas por el algoritmo, el cual, como ya se ha indicado trabaja con datos normalizados en el intervalo [0,1]. En su lugar, y para facilitar su comprensión, esas reglas se han traducido con los valores correspondientes a los datos sin normalizar.

Se aprecia que las expresiones lógicas en las cinco reglas toman como variables los rangos de edad, prestando mucha atención a los rangos en torno a 30-39 años para caracterizar los cinco grupos de municipios. Es probable que el modelo esté tomando esas edades para caracterizar el tamaño de los municipios. Como novedad importante frente a otros trabajos relacionados, la orientación del voto también ha sido tenida en cuenta. Por el contrario, el origen de la población (e.g., extranjera, nacida en el mismo municipio, etc...) no ha sido utilizada.

Los cinco grupos generados por las reglas, se describen en el Cuadro IV, mientras que sus respectivos modelos lineales están en el Cuadro III. En dicho cuadro se ha incluido una columna *Peso* que se calcula a partir de los valores absolutos de los coeficientes, de manera que representa el cociente entre el valor absoluto del coeficiente de ese atributo, dividido por la suma de los valores absolutos de todos los coeficientes. Una vez calculado el peso, los atributos se ordenan por el mismo descendentemente. El cuadro solo muestra los de mayor absoluto, concretamente los necesarios para que su suma supere el umbral del 33,34 % del peso total.

Los modelos lineales de este cuadro son confusos y no arrojan conclusiones sobre la influencia de un determinado colectivo en la aparición de hechos delictivos. Esto se debe principalmente a las relaciones de inclusión que existen entre gran parte de los atributos. Es decir, algunos atributos representan colectivos que están incluidos dentro de colectivos representados por otros atributos. Por ejemplo, en el modelo

lineal número 1, aparentemente la variable más importante es el número de extranjeros americanos, que contribuye a la aparición de hechos delictivos con signo positivo y un peso del $8,01\,\%$. Sin embargo, el coeficiente con tercer mayor peso son las mujeres de ese colectivo, que contribuyen negativamente con un peso del $4{,}70\,\%$, mientras que los varones de ese colectivo están en undécimo lugar con un peso, también negativo del $2,93\,\%4$, de manera que unos coeficientes están contrarrestando el peso de otros, y dado que el modelo está representando los coeficientes para predecir la raíz cuadrada de los hechos, no es directo establecer la contribución neta de los tres coeficientes. Hay más relaciones de inclusión, por ejemplo, en el modelo 2, los varones y mujeres de 30 a 34 años son un subconjunto de los varones y mujeres, y a la vez es superconjunto de los varones en ese rango de edad, etc ...

Modelo Lineal 1		
Coef.	Variable	Peso
+86.053,06	Americanos	8,01 %
-66.181,54	Varones+mujeres de 0 a 4 años	6,16 %
-50.466,39	Mujeres de América	4,70 %
-48.402,00	Varones+mujeres de 30 a 34 años	4,50 %
+40.735,51	Varones+mujeres de 40 a 44 años	3,79 %
+34.253,68	Varones+mujeres de 60 a 64 años	3,19 %
-33.294,81	Varones	3,10 %
+0,61	Término independiente	

Modelo Lineal 2		
Coef.	Variable	Peso
+31.480,21	Varones+mujeres de 30 a 34 años	16,76 %
-29.679,75	Varones+mujeres de 0 a 4 años	15,80 %
-15.372,52	Varones de 30 a 34 años	8,18 %
+2,07	Término independiente	

Modelo Lineal 3		
Coef.	Variable	Peso
-14.576,13	Extranjeros	21,68 %
+9.533,03	Mujeres extranjeras	14,18 %
+3.10	Término independiente	

Modelo Lineal 4		
Coef.	Variable	Peso
-295,28	Varones de 70 a 74 años	11,48 %
-250.56	Varones de 90 a 94 años	9,74 %
+247.15	Varones nacidos en el extranjero	9,61 %
-246.47	Mujeres asiáticas	9,58 %
+10,87	Término independiente	

Modelo Lineal 5		
Coef.	Variable	Peso
-120.03	Varones de 75 a 79 años	9,62 %
+118.35	Varones de 85 a 89 años	9,49 %
+110.85	Varones y mujeres de 80 a 84 años	8,88 %
-106.74	Mujeres nacidas en el extranjero	8,56 %
+0.126	Término independiente	

Cuadro III

MODELOS LINEALES OBTENIDOS PARA EL CONJUNTO DE REGLAS *M5Rules*. LA COLUMNA PESO REPRESENTA EL PESO DEL VALOR ABSOLUTO DE ESE COEFICIENTE EN EL MODELO LINEAL. SE MUESTRAN SOLO LOS DE MAYOR PESO.

Aunque los modelos lineales no parecen arrojar ninguna conclusión plausible, el análisis de los grupos que generan las reglas si que es algo más revelador. En el Cuadro IV se han incluido el mínimo, máximo, promedio y desviación típica de la población, número de hechos delictivos y Δ/hab para cada uno de los grupos de municipios definidos por las cinco reglas. Asimismo, la Figura 1 muestra también para los cinco grupos cómo se distribuyen los hechos delictivos frente al tamaño de los municipios. Los colores en la figura representan valores Δ/hab altos a medida que toman valores más claros.

Parece que los grupos correspondientes a las cuatro primeras reglas mantienen una cierta similitud en que simplemente constatan que a medida que aumenta el tamaño del municipio aumenta el número de hechos delictivos. En el grupo de la regla 1 llama la atención el máximo número de hechos (146 hechos en Escorca, Mallorca), que se corresponde con el municipio con el Δ/hab máximo que ya se comentó en la sección IV-A. Todos los demás municipios de R1 excepto éste están, sin embargo, por debajo de los 45 hechos.

El ratio Δ/hab es algo mas grande para la regla R1 (promedio de $0,74\,\%$), debido a que en los municipios pequeños se penaliza mucho una leve variación en unos pocos delitos, pero en los grupos R2 a R4 se estabiliza en torno a $0,30\,\%-0,38\,\%$.

El grupo correspondiente a la regla R5, sin embargo, es diferente a los demás. A pesar de englobar municipios grandes y muy grandes, mantiene un promedio de hechos delictivos muy bajo (3,22). Un análisis identificativo de dichos municipios nos revela que son todos municipios de Cataluña y País Vasco, donde hay transferidas muchas competencias a las policías autonómicas, y donde por tanto, el número de denuncias que tramita la Guardia Civil tiende a ser marginal.

Por tanto, podemos apuntar como debilidad del modelo predictivo que no predice los hechos delictivos en si, sino únicamente los denunciados a la Guardia Civil, como por otro lado es lógico, ya que son las denuncias que se han utilizado en el estudio. Por tanto, otra línea de mejora es la incorporación de datos de las policias autonómicas.

	R1	R2	R3	R4	R5
Municipios	3.118	1.846	1.437	830	894
Min Habs	5	258	299	4.027	291
Max Habs	341	2.036	8.483	3.165.541	1.608.746
Prom Habs	133,5	706,0	3.014,2	38.194,6	9.848,5
Dev Habs	79,2	342,0	1.574,4	127.095,5	58.499,7
Min Hechos	0	0	0	0	0
Max Hechos	146	85	459	10.025	792
Prom Hechos	2,55	14,32	70,06	542,83	3,22
Dev Hechos	4,31	11,22	55,25	775,15	27,86
Min Δ/hab	0,00 %	0,00 %	0,00 %	0,00 %	0,00 %
Max Δ/hab	29,99 %	10,50 %	5,90 %	7,30 %	0,70 %
Max Δ/hab	33,70 %	37,10 %	33,70 %	7,90 %	1,30 %
Prom Δ/hab	0,74 %	0,38 %	0,30 %	0,37 %	0,03 %
Dev Δ/hab	1,68 %	0,52 %	0,39 %	0,50 %	0,06 %

Cuadro IV

DESCRIPCIÓN ÁREAS CUBIERTAS POR CADA REGLA. R_i =Nº DE REGLA EN EL ORDEN DE INTERPRETACIÓN *M5rules*, MIN-MAX-PROM Y DEV = MÍNIMO, MÁXIMO, PROMEDIO Y DESVIACIÓN ESTÁNDAR RESPECTIVAMENTE. HABS=Nº DE HABITANTES, HECHOS=Nº HECHOS DELICTIVOS POR MUNICIPIO, Δ/HAB =PORCENTAJE DE ERROR POR HABITANTE.

V. Conclusiones y líneas futuras

En el presente trabajo se ha obtenido un modelo predictivo basado en *Random Forests* que permite predecir el número

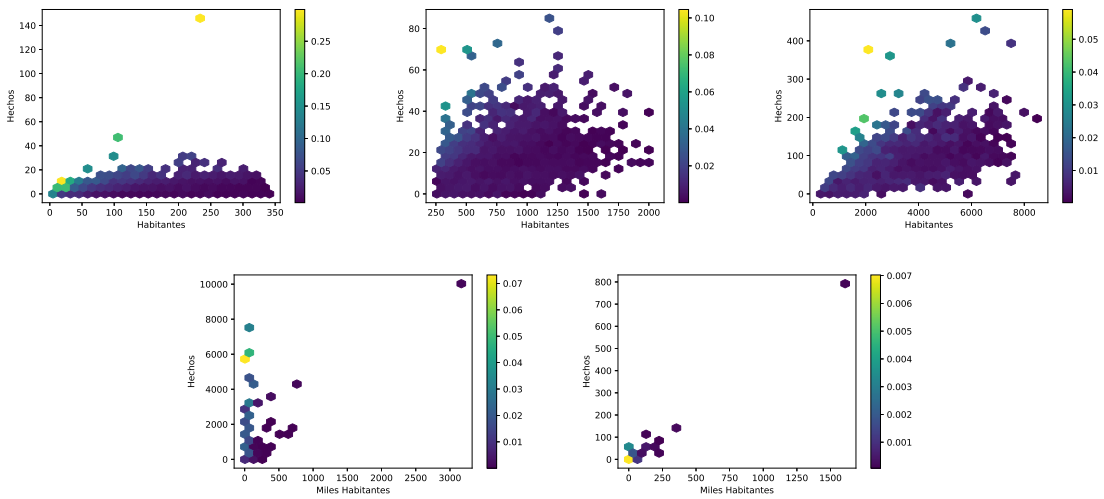


Figura 1. Un diagrama de *binning* hexagonal por cada uno de los subconjuntos resultantes de aplicar las reglas. En el eje y se muestran el número de habitantes, en el eje x el número de hechos y el color representa Δ/hab en la predicción en los municipios situados en esas coordenadas

de hechos delictivos que se denuncian a la Guardia Civil anualmente en cada municipio. El valor de RRSE alcanzado es de 41.23, pero en general todos los *ensembles* dan resultados que no son significativamente diferentes.

Los datos utilizados combinan la localización de las denuncias de la Guardia Civil con fuentes de datos públicas (i.e., INE y datos electorales de 2016). Una novedad que incorpora el presente trabajo es precisamente la incorporación de datos electorales.

Del análisis de los municipios en los que peor se comporta el modelo se deriva una posible mejora incorporando datos relativos a la ocupación turística, probablemente enfocados al turismo de playa.

Por otro lado, las reglas *M5Rules* han discriminado cinco grupos de municipios. Los cuatro primeros grupos parecen segmentar los municipios por su tamaño. El quinto grupo aglutina las denuncias en las comunidades autónomas en las que la policía autonómica ha sustituido a la Guardia Civil en muchas de sus competencias; por lo que otra línea de mejora cara identificar puntos calientes, es incorporar datos de denuncias de esos cuerpos policiales.

Referencias

[1] Luiz G.A. Alves, Haroldo V. Ribeiro, and Francisco A. Rodrigues. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505:435 – 443, 2018.

[2] Iniciativa Aporta. Acerca de la iniciativa aporta. http://datos.gob.es/es/acerca-de-la-iniciativa-aporta. [Internet; descargado 16-mayo-2018].

[3] Maria Jeseca C. Baculo, Charlie S. Marzan, Remedios de Dios Bulos, and Conrado Ruiz. Geospatial-temporal analysis andclassification of criminal data in manila. In *Procs. of 2nd IEEE International Conference on Computational Intelligence and Applications*, pages 6–11. IEEE, 2017.

[4] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[6] Leo Breiman. Using iterated bagging to debias regressions. *Machine Learning*, 45(3):261–277, Dec 2001.

[7] Elise Clougherty, John Clougherty, Xiaoqian Liu, and Donald Brown. Spatial and temporal analysis of sex crimes in charlottesville, virginia. In *Procs. of IEEE Systems and Infromation Engeneering Design Symposium*, pages 69–74. IEEE, 2015.

[8] Harris Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 107–115, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[9] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[10] Jerome H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, February 2002.

[11] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[13] Geoffrey Holmes, Mark Hall, and Eibe Frank. Generating rule sets from model trees. In *Proceedings of the 12th Australian Joint Conference on Artificial Intelligence: Advanced Topics in Artificial Intelligence*, AI '99, pages 1–12, London, UK, UK, 1999. Springer-Verlag.

[14] Hyeon-Woo Kang and Hang-Bong Kang. Prediction of crime occurrence from multimodal data using deep learning. *PLoS One*, 12(4):e0176244, 2017.

[15] Keivan Kianmehr and Reda Alhajj. Effectiveness of support vector machine for crime hot-spots prediction. *Applied Artificial Intelligence*, 22(5):433–458, 2008.

[16] J. C. Leitão, J. M. Miotto, M. Gerlach, and E. G. Altmann. Is this scaling nonlinear? *Royal Society Open Science*, 3(7), 2016.

[17] G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

[18] C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(239–281), 2003.

[19] Comisión Económica para América Latina y el Caribe. ¿qué son los datos abiertos? https://biblioguias.cepal.org/EstadoAbierto/datospublicos. [Internet; descargado 16-mayo-2018].

[20] Ross J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.

[21] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.