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.

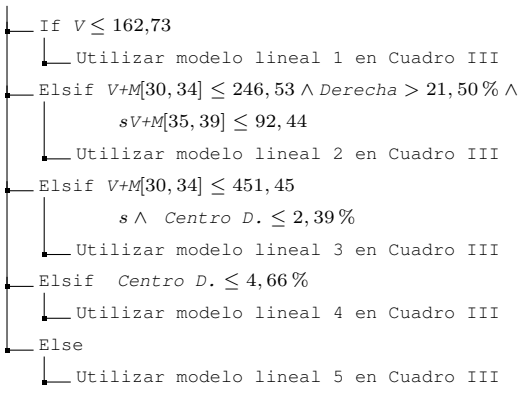
IV-B. Interpretación con M5Rules

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

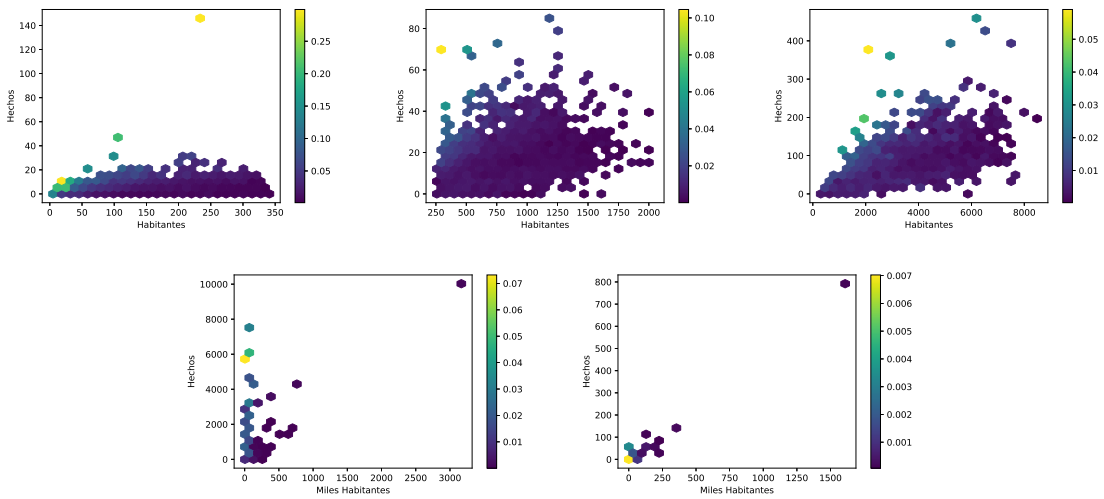


Figura 1. Un diagrama de *binning* hexagonal por cada uno de los subconjuntos resultantes de aplicar las reglas. En el eje y se muestran el número de habitantes, en el eje x el número de hechos y el color representa Δ/hab en la predicción en los municipios situados en esas coordenadas

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.