Gestión de riesgos en la Ciudad Autónoma de Buenos Aires: modelos de clasificación para predecir delitos violentos entre 2016 y 2019

María Florencia Ain
Universidad de Buenos Aires

Resumen

La seguridad pública se encuentra, históricamente, entre las principales inquietudes de los ciudadanos y la comisión de delitos se relaciona directamente con ella. La predicción de delitos está destinada a la anticipación de posibles hechos futuros, tomando como punto de partida el procesamiento y análisis de la información disponible en la actualidad. Incorporar la predicción de delitos al ámbito de la planificación de la política de seguridad constituye un hecho innovador que puede redundar en programas basados en la evidencia y más adecuados a las particularidades de los contextos en que se insertan, con un desarrollo de sistemas de vigilancia y prevención eficaces y adecuados a los recursos disponibles de la jurisdicción.

El presente trabajo tiene por objeto predecir la ocurrencia de delitos violentos en la Ciudad Autónoma de Buenos Aires (CABA) a partir de la aplicación de métodos de clasificación a las bases de datos (*datasets*) de delitos publicadas por el Gobierno de la Ciudad de Buenos Aires (GCBA) para encontrar el método de cuantificación de riesgo que mayor exactitud arroje para dicha predicción.

Recién a partir del año 2017, el GCBA publica el conjunto de los datos sobre la criminalidad registrada en el portal BA Data https://data.buenosaires.gob.ar/. Por consiguiente, los datos que se emplean son a partir de enero de 2016 y hasta diciembre de 2019. El criterio de selección de las dimensiones espacio temporales consiste en utilizar los datos disponibles. Cabe mencionar que los delitos cometidos a partir del año 2020 no se incorporan al análisis por motivo de la pandemia de COVID 19, la cual ha provocado algunas distorsiones en los valores consignados que afectan cualquier tipo de predicción.

El enfoque del estudio es cuantitativo exploratorio y el tipo de diseño longitudinal. Se realiza un análisis exploratorio de los *datasets* originales de delitos violentos y se procede a la transformación de los atributos del conjunto de datos para mejorar la calidad de las predicciones. A partir de la confección de una única base y su posterior procesamiento en *Rapidminer Studio*, se predice la ocurrencia de delitos violentos utilizando los métodos de aprendizaje supervisado Regresión Logística (RL) y *k* vecinos más cercanos (k-*Nearest Neighbour*, k-NN) para conocer cuál es el que mayor exactitud arroja. Estas herramientas de cuantificación de riesgo, etapa fundamental dentro de lo que se denomina gestión de riesgo, permiten optimizar la identificación de oportunidades y amenazas, crear un marco para la toma de decisiones informadas, perfeccionar los métodos de seguimiento y monitoreo y mejorar la prevención de hechos delictivos.

Por último, y en base a los resultados obtenidos, se elaboran sugerencias para prevenir el riesgo inherente a la comisión de delitos violentos y/o mitigar su impacto.

Palabras claves: delito violento, gestión de riesgo, *Rapidminer Studio*, predicción.
Clasificación JEL: D81**,** C88, E17

1. Introducción

La gestión de la seguridad pública se ha convertido en un eje de política prioritario para los gobiernos en tanto la inseguridad constituye una de las principales preocupaciones de la opinión pública. El delito, como expresión material de la inseguridad, es un fenómeno complejo y multidimensional. En este sentido, las encuestas de victimización, realizadas conjuntamente por el Instituto Nacional de Estadística y Censos (INDEC) y el Ministerio de Seguridad de la Nación, se erigen como un insumo fundamental para comprenderlo y elaborar estrategias de abordaje.

De la última Encuesta Nacional de Victimización (ENV) publicada, se desprende que el 85,1% de la población del país considera la inseguridad en su ciudad de residencia como un problema "bastante o muy grave" (Instituto Nacional de Estadística y Censos [INDEC], 2018, pp. 8-9). Asimismo, resulta destacable que la tasa de no denuncia o "cifra negra" asciende al 66,3% de los delitos contra las personas, porcentaje que en la CABA alcanza al 75,9%, la más elevada del país (Observatorio de Seguridad Ciudadana, s.f.). La desmotivación para no realizar la denuncia puede encontrar justificación en los bajos niveles de confianza que tiene la ciudadanía en las instituciones encargadas de administrar la justicia. Adicionalmente, las políticas de seguridad parecen no contentar a una ciudadanía que al tiempo que descree de la justicia, reclama mayores niveles de punición. Resulta probable que un abordaje más integral del problema del delito reduzca significativamente esta demanda.

El análisis de datos de cualquier índole –en este caso, de seguridad- puede resultar un factor clave para mejorar la administración gubernamental a través de la implementación de políticas públicas basadas en evidencia que satisfagan las necesidades ciudadanas, pero que también resuelvan los problemas que abordan. En este sentido, la predicción de delitos está destinada a la anticipación de posibles hechos futuros, tomando como punto de partida el procesamiento y análisis de la información disponible. Incorporar la predicción de delitos a la planificación de la política de seguridad constituye un hecho innovador que puede redundar en programas más adecuados al contexto en el que se insertan, con datos fiables que pueden ser permanentemente actualizados.

En las últimas décadas, el acceso a datos abiertos y el surgimiento de nuevas tecnologías basadas en internet han posibilitado la difusión de los modelos predictivos como técnicas idóneas para cuantificar los riesgos delictivos. Estos modelos utilizan técnicas propias del aprendizaje automático (*machine learning)* y desarrollan algoritmos que permiten aprender comportamientos de experiencias pasadas (Stamp, 2017). Con la utilización de técnicas analíticas adecuadas se pueden identificar, cuantificar y predecir problemáticas de interés para cualquier gestión de gobierno: a esto se refiere con la denominada "inteligencia de valor público" (Rodríguez et al., 2017, p. 9).

El objetivo general del presente trabajo consiste en proponer una gestión de riesgos de los delitos violentos en la CABA mediante la predicción de su ocurrencia aplicando los métodos de clasificación RL y k-NN. Para ello, se utilizan las bases de datos de delitos publicadas en el portal BA Data para el periodo comprendido entre los años 2016 y 2019, con el objeto de contribuir al desarrollo de sistemas de vigilancia y prevención de delitos eficaces y adecuados a los recursos disponibles de la jurisdicción.

Los resultados de la aplicación del modelo predictivo más preciso implica una mejora de calidad en la gestión integral de riesgos, contribuyendo al desarrollo de sistemas de vigilancia y prevención de la inseguridad y al diseño de dispositivos de mitigación del impacto de estos hechos en las víctimas; constituyéndose así en una herramienta valiosa también para el uso y la distribución racional de los recursos financieros disponibles, siempre limitados.

2. Gestión del riesgo delictivo con minería de datos

Existen múltiples definiciones y tipos de riesgo. Según la Real Academia Española, la palabra riesgo tiene dos acepciones: por un lado, se lo define como una contingencia o proximidad de un daño y, por otro, como cada una de las contingencias que son pasibles de un contrato de seguro. En otras palabras, el riesgo puede definirse como la probabilidad de ocurrencia de un evento desfavorable y sus consecuencias inherentes. Esta conceptualización permite distinguir no sólo la probabilidad de ocurrencia en sí misma, relacionada indefectiblemente al concepto de incertidumbre, sino también la pérdida asociada a dicha ocurrencia (Dorofee et al., 1996, p. 20). En este sentido, la gestión o gerenciamiento de riesgos comprende aquellas actividades coordinadas para identificar, analizar, evaluar y clasificar los riesgos con el objeto de mitigar sus consecuencias (Organización Internacional de Normalización, 2010).

Según Garland (2005), el delito constituye un riesgo que debe ser calculado o un accidente que debe evitarse. Este enfoque se encuadra dentro de las denominadas "nuevas criminologías de la vida cotidiana", que el autor define como un "conjunto de marcos teóricos afines que incluyen la teoría de las actividades rutinarias, del delito como oportunidad, del análisis de los estilos de vida, de la prevención situacional del delito y ciertas versiones de la teoría de la elección racional." (p. 217). A partir de este marco teórico, el delito se percibe de manera prospectiva y agregada con el objetivo de calcular los riesgos y formular políticas preventivas.

En este marco, la gestión de riesgos abarca las iniciativas coordinadas para identificar, analizar, calcular y clasificar los delitos con el objeto de predecir hechos delictivos que permitan implementar políticas preventivas en materia de seguridad. Este proceso también incluye el análisis de datos como parte de una gestión integral de riesgos por parte del estado. Si bien éste puede transferir una parte del riesgo delictivo a otras entidades (sean empresas de seguridad privada u organizaciones vecinales), es su obligación gestionarlo, siendo los modelos predictivos los instrumentos más eficaces para intentar disminuir la probabilidad de ocurrencia de delitos violentos.

Las técnicas propias del aprendizaje automático brindan sistemas de medición de riesgos más precisos y personalizados en función de las demandas de cada organización que les permiten crear una estrategia de gestión de riesgos integral. Al evitar utilizar hipótesis previas para la creación de modelos, como lo hace la estadística clásica, permite descubrir patrones y tendencias ocultos en los conjuntos de datos que redundan en modelos con mayor poder predictivo.

Los beneficios inherentes a la utilización y análisis de datos por parte de las organizaciones resultan innegables: contribuyen en la predicción, anticipación y minimización de los riesgos vinculados a su operatoria específica. La existencia de grandes volúmenes de datos

requiere de modelos propios del aprendizaje automático que permitan procesarlos, lo que genera una mejora significativa de las capacidades analíticas en gestión de riesgos. Al basarse en un aprendizaje continuo y automático, disminuye el margen de error y evalúa permanentemente patrones y desvíos de manera más eficiente que las herramientas estadísticas tradicionales.

En el presente trabajo, se aborda la gestión de riesgo delictivo desde su cuantificación. A partir del cálculo de la probabilidad de ocurrencia de delitos violentos, la aplicación de los modelos de clasificación RL y k-NN permite optimizar la identificación de oportunidades y amenazas, crear un marco para la toma de decisiones informadas, perfeccionar los métodos de seguimiento y monitoreo y mejorar la prevención de hechos delictivos.

2.1 Predicción de delitos: RL y k-NN

Los métodos de aprendizaje supervisado permiten resolver problemas de regresión o clasificación, dependiendo de si la variable objetivo es continua o discreta (Musumeci et al., 2018). La clasificación permite extraer información significativa a partir de grandes conjuntos de datos y puede utilizarse para predecir clases desconocidas.

Los métodos de regresión son un tipo de técnica de análisis predictivo en la que la variable objetivo se relaciona funcionalmente con las variables de entrada. La RL es técnicamente un método de clasificación pero estructuralmente es similar a la Regresión Lineal (Kotu y Deshpande, 2015, p. 14).

La idea subyacente en la Regresión Lineal es la construcción de una función que explique y prediga el valor de la variable objetivo, dados los valores de las variables predictoras. De este modo, el problema se reduce a encontrar la línea recta que mejor explique esta tendencia (Kotu y Deshpande, 2015, pp. 167-168). Con más de dos predictores, la variable dependiente puede expresarse como una combinación lineal de las variables independientes:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \quad (1)$$

El objetivo de la RL consiste en encontrar el modelo que mejor se ajusta para describir la relación entre la variable dependiente y las variables independientes (Prabakaran y Mitra, 2018, p. 6). Es decir, resuelve el problema de predicción de una variable objetivo, la cual puede ser binomial o binaria, usando atributos numéricos.

Si la variable dependiente *y* es binomial (delito violento y delito no violento), el desafío es encontrar una ecuación que conecte funcionalmente los predictores *x* con la variable dependiente *y*, que sólo puede tomar dos valores: 0 ó 1. Sin embargo, los predictores no tienen restricciones ya que pueden ser continuos o categóricos, siendo su rango funcional también irrestricto entre $-\infty$ y $+\infty$. Para superar este problema, se debe transformar la función continua en una discreta. El concepto de *logit* es lo que permite conseguir este cometido.

Ahora bien, cabe preguntarse cómo la RL encuentra la curva sigmoidea. Como se desprende de la Ecuación 1, una línea recta puede representarse con dos parámetros: la

pendiente b_1 y el intercepto b_0. La forma en que las *x* y la *y* están relacionadas puede especificarse a través de b_1 y b_0. No obstante, la curva sigmoidea es más compleja y representarla paramétricamente no resulta tan sencillo, por ello, la clave radica en encontrar los parámetros matemáticos que relacionan ambas variables.

Si se transforma la variable objetivo *y* al logaritmo de las *odds* de *y*, entonces dicha variable transformada está linealmente relacionada a los predictores *x*. En la mayoría de los casos en los que se necesita usar RL, la *y* es usualmente un tipo de respuesta del estilo sí/no. Esto suele interpretarse como la probabilidad de ocurrencia de un evento (y = 1) o no (y = 0). Esto puede explicitarse de la siguiente manera:

- Si *y* es un evento (sí/no), y
- *p* es la probabilidad de que el evento ocurra (y = 1),
- entonces (1 – p) es la probabilidad de que el evento no ocurra (y = 0), y
- p/(1 – p) es el *odds* ratio, es decir, las *odds* de que el evento suceda.

El logaritmo de las *odds* de *y*, log (p/1 – p), se denomina función *logit* de *y*. Puede expresarse como una función lineal de los predictores *x*, similar a la planteada en la Ecuación 1:

$$\text{logit} = \log p/(1-p) = b_0 x + b_1 \quad (2)$$

En lugar de predecir *y*, se predice el logaritmo de las *odds* de obtener un 1 en la variable dependiente. El *odds ratio* brinda una estimación para una única unidad de incremento en la variable independiente. Como no es una función lineal de los coeficientes, no se puede afirmar que para cada unidad de incremento en la variable dependiente, el *odds ratio* aumenta en la misma proporción.

En el caso que se involucren múltiples variables independientes, la ecuación se estructura de la siguiente manera:

$$\text{logit} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \quad (3)$$

El *logit* puede tomar cualquier valor entre –∞ y +∞. Para cada fila de predictores de un *dataset*, se puede computar el *logit*. Esto es, los valores predichos de *y* obtenidos de la ecuación de regresión varían entre 0 y 1, a pesar de que los resultados de la regresión pueden alcanzar cualquier valor entre –∞ y +∞. Con el *logit*, es sencillo computar la probabilidad de ocurrencia o no de *y*. Esta ecuación se conoce con el nombre de función de probabilidad acumulada:

$$p = e^{\text{logit}}/(1 + e^{\text{logit}}) \quad (4)$$

El modelo de RL de la Ecuación 3 proporciona, en última instancia, la probabilidad de ocurrencia de *y* (y = 1) dado valores específicos de *x* mediante la Ecuación 4.

Utilizando las Ecuaciones 3 y 4, y conociendo previamente los valores de *x*, se puede calcular el valor de *p*. Para ello, se necesita determinar los coeficientes *b* de la Ecuación 3.

Suponiendo un valor inicial de prueba para *b*, y dada una muestra de datos de entrenamiento, se puede calcular:

$$p^y * (1 - p)^{(1 - y)} \tag{5}$$

donde *y* es la variable objetivo original y *p* es la probabilidad estimada usando la Ecuación 4.

Como afirman Fix y Hodges (1951, como citó García Jiménez, 2010, p. 24), el método k-NN es una técnica de aprendizaje conocida como aprendizaje basado en instancias o ejemplos (*instance-based learning*), que se circunscribe a "memorizar" los datos de entrenamiento. La premisa subyacente a este razonamiento radica en que los miembros de una población comparten características y propiedades similares con los individuos que la rodea (García Jiménez, 2010, p. 24).

En el aprendizaje supervisado, los problemas de agrupamiento consisten en organizar un conjunto de elementos en grupos según una lógica de similitud o proximidad, para lo cual se utiliza una función o métrica de distancia (Pérez Verona y Arco García, 2016, p. 44). La cercanía entre las instancias determina la pertenencia a determinado grupo: se presume que un elemento es más similar o afín con los elementos de su grupo que con relación a los elementos de un grupo diferente. Si por ejemplo, se elige un valor k =1, una muestra cualquiera x se clasifica en la clase asociada a su instancia más cercana. Para valores $k \geq 2$, dicha muestra se asigna a la clase más representada entre las k instancias más próximas a la muestra (García Jiménez, 2010, pp. 24-25).

De este modo, y considerando la estructura de los datos del conjunto de entrenamiento, el cálculo de distancias para estimar la proximidad entre dos instancias resulta fundamental (Cover y Hart, 1967; Cong et al., 2015). La exactitud que arroja el método de clasificación depende del modo en que se calculan las distancias entre los diferentes ejemplos (Cong et al., 2015). Como sostienen Deza y Deza (2009, como citó Pérez Verona y Arco García, 2016, p. 44), la medida de similitud más popular para datos numéricos por su simplicidad y características de generalización es la distancia euclidiana. La misma se calcula con la fórmula detallada a continuación y el valor escalar resulta de los valores de los atributos de las dos instancias (García Jiménez, 2010, p. 25):

$$d(\mathbf{y}, \mathbf{x}) = \sqrt[2]{\sum_{j=1}^{n} (y_j - x_j)^2} \tag{6}$$

Como se mencionó precedentemente, los modelos explicados son aquellos que arrojan los resultados más exactos con relación a la predicción de ocurrencia de delitos violentos. La RL y el algoritmo k-NN, al emplearse como metodologías de cuantificación del riesgo para predecir los delitos violentos, constituyen herramientas flexibles capaces de adaptarse a un contexto dinámico y cambiante producto de la utilización de conjuntos de datos abiertos, y se convierten en aliadas estratégicas en la gestión de riesgos delictivos.

3. Metodología

3.1 Tipo de estudio

El enfoque del estudio es cuantitativo exploratorio y el tipo de diseño longitudinal. Se realiza un análisis exploratorio de los *datasets* de delitos publicados en el portal de datos abiertos del GCBA, denominado BA Data. A pesar de haber sido creado en 2012, recién a partir del año 2017, se publica el conjunto de los datos sobre la criminalidad registrada. Por consiguiente, los datos utilizados en el presente trabajo corresponden al periodo 2016-2019. El criterio de selección de las dimensiones espacio temporales es utilizar los datos disponibles. Cabe mencionar que los delitos cometidos a partir del año 2020 no se incorporan al análisis por motivo de la pandemia, la cual ha provocado algunas distorsiones en los valores consignados que afectan cualquier tipo de predicción.

Con relación a las bases de datos disponibles para realizar el presente trabajo, existen cuatro tipos de delitos publicados por el GCBA, a saber: homicidios, lesiones, hurtos y robos. No obstante, esta última tipología representa entre el 55% y el 60% de los delitos denunciados durante todo el periodo de estudio. Por su parte, el porcentaje de homicidios sobre el total de delitos denunciados representa menos del 1% y, en el caso de las lesiones, sólo se encuentran publicadas aquellas correspondientes al año 2019. Por dichos motivos, ambos tipos delictivos no se incluyen en la predicción. De esta manera, el análisis se centra en los hurtos, definidos como delitos sin violencia, y en los robos, que constituyen delitos violentos.

El *dataset* original utilizado se denomina “Delitos 2019” y suministra información mensual a través de los siguientes atributos: número de identificador, fecha, franja horaria, tipo de delito, subtipo de delito, cantidad registrada, comuna, barrio, latitud y longitud. Los *datasets* “Delitos 2016”, “Delitos 2017” y “Delitos 2018”, cuyos datos se incorporan en una instancia posterior, poseen idéntica estructura. A continuación, se detalla cada atributo con su descripción así como el tipo de dato:

- Id: identificador de cada delito. Tipo de dato entero.
- Fecha: fecha de comisión del delito. Tipo de dato fecha AAAA-MM-DD.
- Franja_horaria: franja horaria de comisión del delito. Los valores oscilan entre 0 y 23, es decir, cada franja horaria se corresponde a una hora reloj. Tipo de dato entero.
- Tipo_delito: tipo de delito cometido. Puede tomar cuatro valores: homicidio, hurto, lesiones y robo. Tipo de dato polinomial.
- Subtipo_delito: subtipo de delito cometido. Puede tomar cuatro valores: doloso (homicidio), automotor (hurto o robo), no automotor (hurto o robo) y siniestro vial (lesiones). Tipo de dato polinomial.
- Cantidad_registrada: cantidad de delitos registrada en cada franja horaria. Puede tomar distintos valores pero se supone que en una franja horaria con una latitud y longitud específica, sólo puede ocurrir un delito. Tipo de dato entero.
- Comuna: número de comuna en que se cometió el delito. Los valores oscilan entre 1 y 15. Tipo de dato entero.
- Barrio: nombre del barrio en que se cometió el delito. Puede tomar 48 valores. Tipo de dato polinomial.
- Lat: latitud del lugar donde ocurrió el delito. Tipo de dato real.
- Long: longitud del lugar donde ocurrió el delito. Tipo de dato real.

A partir de los *datasets* mencionados, la Figura 1 exhibe la cantidad de delitos cometidos en la CABA durante el período de estudio con el objetivo de conocer la tendencia de los tipos de hechos delictuales.

Figura 1

Cantidad de tipos de delitos por año en valores absolutos

	Cantidad			
Tipo de delito	2016	2017	2018	2019
Lesiones	8890	9851	10061	10106
Homicidio	289	266	277	198
Hurto	46178	42150	42274	49351
Robo	71226	68297	71121	62829
Total	126583	120564	123733	122484

3.2 Tratamiento de los datos

Los *datasets* originales necesitan ser pre procesados para completar y/o eliminar celdas vacías o valores atípicos, suprimir columnas innecesarias y agregar elementos relevantes. Para ello, se utilizan métodos de extracción y selección de atributos que permiten identificar y suprimir aquellos datos que puedan ser redundantes o irrelevantes. En este sentido, se procede a eliminar aquellas celdas que no contienen información sobre el barrio en cuestión o sobre la franja horaria en que se comete el delito. Por otro lado, no se encuentran valores atípicos. La Figura 2 exhibe la cantidad de delitos cometidos en la CABA durante el período de estudio después del tratamiento de los datos.

Figura 2

Cantidad de tipos de delitos después del tratamiento de los datos

	Cantidad			
Tipo de delito	2016	2017	2018	2019
Lesiones	8656	9809	7947	9591
Homicidio	265	244	252	194
Hurto	46070	42050	40941	49269
Robo	70963	66601	69279	62734
Total	125954	118704	118419	121788

A partir del *dataset* “Delitos 2019” obtenido luego del tratamiento de los datos, se calcula el primer modelo denominado línea de base (*baseline*) utilizando *Microsoft Excel*. En este sentido, cualquier modelo seleccionado con posterioridad para realizar las predicciones debe superar el umbral de la línea de base. Para ello, se computa el cociente entre los delitos violentos y el total de los delitos con el objetivo de obtener una primera medida de exactitud.

Para un problema de dos clases, se dice que un conjunto de datos está desbalanceado cuando el número de muestras de la clase mayoritaria es significativamente superior al de la clase minoritaria (García Jiménez, 2010, p. 27). Existen algoritmos de clasificación, como la RL, sensibles a las proporciones de las diferentes clases. Si el conjunto de entrenamiento está desbalanceado, dichos algoritmos suelen favorecer a la clase mayoritaria lo que puede generar métricas de exactitud sesgadas.

Es dable destacar que esta etapa, así como la transformación del conjunto de datos al formato requerido por el algoritmo, resulta de capital importancia para que los patrones descubiertos al finalizar el estudio sean de calidad (Valenga et al., 2007). La calidad de las predicciones depende de la transformación de los atributos del conjunto de datos que se realice previamente. Aquellas que se utilizan frecuentemente consisten en el aumento o disminución de la dimensionalidad, la discretización de atributos numéricos, la numerización de atributos nominales y la normalización.

Con relación al aumento de la dimensionalidad, este se logra mediante la creación de nuevos atributos a partir de los existentes. En el caso de la fecha, el atributo original con formato AAAA-MM-DD brinda escasa información si se lo emplea directamente; no obstante, al crearse los atributos día, mes y día de la semana se logra capturar la misma información de manera más eficiente. Otro ejemplo similar es la creación de los atributos delitos por barrio, construidos a partir de los *datasets* de los años 2016, 2017 y 2018, que reflejan la cantidad de tipos de delitos cometidos por barrio expresada en porcentaje.

Cabe mencionar que la transformación de atributos mencionada en el párrafo precedente se realiza en las hojas de cálculo de *Microsoft Excel*. Las columnas representan los atributos para cada serie de datos y las filas son un ejemplo de los mismos. Para exportar el archivo al *Rapidminer Studio,* se utiliza el operador *Read Excel*.

Por otra parte, se utiliza la numerización en los casos en que los atributos nominales u ordinales deban convertirse en números. Para los nominales suele utilizarse una representación binaria y para los ordinales suele utilizarse una representación entera. Tal es el caso de la variable objetivo, en la que se reemplazan los valores “si” y “no” por 1 y 0, y del atributo día de la semana, cuyos valores oscilan entre 1 y 7, respectivamente.

Para realizar esta operación en *Rapidminer Studio*, se utiliza el operador *Nominal to Numerical* que tiene por objeto no sólo cambiar los valores de los atributos no numéricos a numéricos, sino que también asigna todos los valores de estos atributos a un tipo numérico. Los atributos numéricos del *dataset* original no sufren modificaciones. Los atributos binarios los convierte en 0 y 1.

De esta manera, se agregan los siguientes atributos a la enumeración realizada previamente:

- Dia: día en que ocurrió el delito. Puede tomar valores entre 1 y 30 o 31, dependiendo del mes. Tipo de dato entero.
- Mes: mes en que ocurrió el delito. Puede tomar valores entre 1 y 12, donde 1 representa el mes de enero, 2 el mes de febrero, 3 el mes de marzo, 4 el mes de abril, 5 el mes de mayo, 6 el mes de junio, 7 el mes de julio, 8 el mes de agosto, 9 el mes de septiembre, 10 el mes de octubre, 11 el mes de noviembre y 12 el mes de diciembre. Tipo de dato entero.

- Dia_ semana: día de la semana en que ocurrió el delito. Puede tomar valores entre 1 y 7, donde 1 representa el día lunes, 2 el día martes, 3 el día miércoles, 4 el día jueves, 5 el día viernes, 6 el día sábado y 7 el día domingo. Tipo de dato entero.
- Delito_violento: indica si el delito cometido es violento (robo) o no violento (delitos restantes). Puede tomar dos valores: 1 en el primer caso y 0 en el segundo. Es la variable objetivo que se quiere predecir (*label*). Tipo de dato binomial.
- Delitos2019_barrio: porcentaje de tipo de delito cometido por barrio en el año de referencia. Tipo de dato real.
- Delitos2018_barrio: porcentaje de tipo de delito cometido por barrio en el año de referencia. Tipo de dato real.
- Delitos2017_barrio: porcentaje de tipo de delito cometido por barrio en el año de referencia. Tipo de dato real.
- Delitos2016_barrio: porcentaje de tipo de delito cometido por barrio en el año de referencia. Tipo de dato real.

3.3 Aplicación de los modelos de RL y k-NN en *Rapidminer Studio*

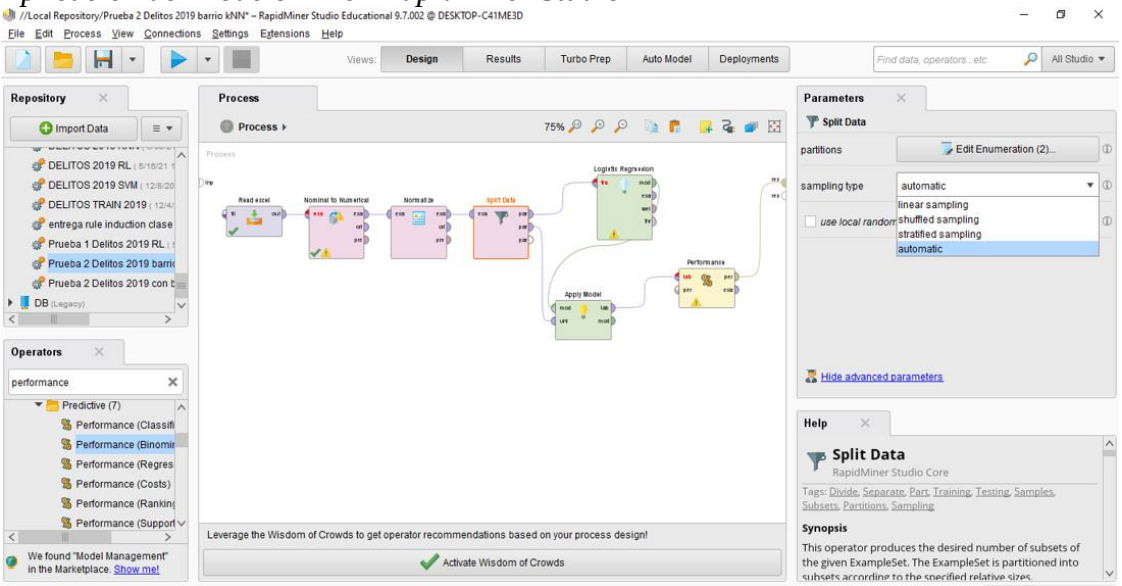
Además del aumento de la dimensionalidad y la numerización de atributos nominales, resulta necesario aplicar la normalización de atributos para los modelos seleccionados. El operador *Normalize* normaliza los valores de los atributos seleccionados para que se ajusten a un determinado rango. Ajustar el valor del rango resulta una tarea prioritaria cuando los atributos están en diferentes unidades y escalas. En el presente trabajo, se utiliza la transformación *z* también denominada normalización estadística: se resta de todos los valores la media de los datos y luego se los divide por la desviación estándar.

Previamente a la aplicación de los algoritmos seleccionados, resulta necesario dividir el conjunto de datos original para que los modelos puedan "aprender" los datos del conjunto de entrenamiento. Para ello, se utiliza el operador *Split Data* que produce la cantidad de subconjuntos (particiones) deseados a partir del *dataset* original, es decir, este último se divide de acuerdo con las proporciones especificadas en los parámetros. En este caso, se eligió la proporción 0,8 y 0,2, es decir que el puerto de salida con el 80% de los datos se conecta al operador del modelo elegido (RL o k-NN) y el puerto de salida con el 20% se conecta al operador *Apply Model*.

Como se explicitó precedentemente, el objetivo de la aplicación de las técnicas de aprendizaje supervisado consiste en obtener una clasificación binaria de los delitos, según sean o no violentos. En este contexto, la variable objetivo o dependiente toma valor 1 si el delito es violento y 0 sino lo es, y se emplean las variables independientes o predictoras descriptas oportunamente. Como se exhibe en la Figura 3, y mediante la aplicación del operador *Logistic Regression*, se genera un modelo de RL que permite clasificar clases binarias.

Figura 3

Aplicación del modelo RL en Rapidminer Studio

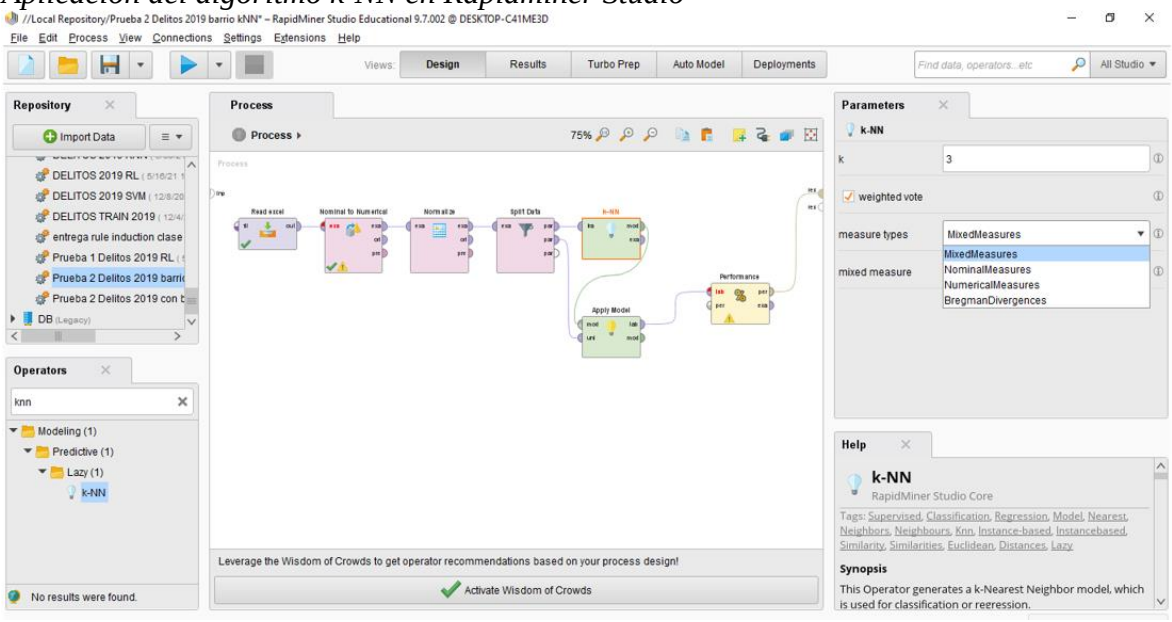


Nota. Salida de *Rapidminer Studio*

A su vez, el operador *k-NN* genera un modelo de k-vecinos más cercanos y se basa en comparar un ejemplo desconocido con los *k* ejemplos de entrenamiento. Dentro de los parámetros disponibles, se encuentran distintos tipos de medida (measure_types) para calcular la distancia entre dichos ejemplos. En este caso, se selecciona el parámetro medidas numéricas con el objetivo de aplicar la distancia euclidiana, tal como se desprende de la Figura 4.

Figura 4

Aplicación del algoritmo k-NN en Rapidminer Studio



Nota. Salida de *Rapidminer Studio*

Asimismo, el algoritmo k-NN clasifica el ejemplo desconocido por un voto mayoritario de los vecinos encontrados. Si se habilita el parámetro *weighted vote*, los vecinos que tengan una distancia menor al ejemplo que se quiere predecir tienen mayor importancia que aquellos que se encuentran más alejados. Si se desestima el uso de dicho parámetro, todos los vecinos más próximos tienen el mismo peso relativo en la predicción.

Finalmente, el operador *Apply Model* se encarga de accionar el modelo con base en el conjunto de datos de entrada y el operador *Performance Binomial* se utiliza para evaluar el desempeño del modelo predictivo. En esta instancia, se elige el criterio de desempeño que se quiere utilizar, en este caso, la exactitud a través de la matriz de confusión y el área bajo la curva ROC (acrónimo de *Receiver Operating Characteristic* o Característica Operativa del Receptor), llamada comúnmente AUC (acrónimo de *Area Under the Cuve* o Área Bajo la Curva).

3.4. Métricas para la evaluación de los modelos

Existen distintas herramientas disponibles para evaluar la capacidad predictiva de un modelo de clasificación binario, a saber: la matriz de confusión, la curva ROC y el AUC (Kotu y Deshpande, 2015, pp. 257-259). En el presente trabajo, se utiliza la matriz de confusión para evaluar la *performance* de los modelos. Como se desprende de la Figura 5, las columnas representan el valor real de la variable objetivo mientras que las filas representan el valor predicho de la misma variable.

Figura 5

Matriz de confusión

		Valor Real	
		Positivo	Negativo
Valor predicho	Positivo	Verdadero positivo (VP)	Falso positivo (FP)
	Negativo	Falso negativo (FN)	Verdadero negativo (VN)

En los casos de los VP y los VN, los valores predichos coinciden con el valor real de la variable objetivo. En el primer caso, el valor real es positivo así como la predicción que arroja el modelo. En el segundo, el valor real es negativo y también la predicción. Para los FP y FN, el valor predicho es falso. En el primer caso, el valor real es negativo pero el modelo predice un valor positivo (comúnmente conocido como error de tipo I); en el segundo caso, el valor real es positivo pero el modelo predice un valor negativo (denominado error de tipo II).

Asimismo, a partir de esta matriz se pueden calcular las siguientes métricas: la exactitud (*accuracy*), la especificidad (*precision*), la sensibilidad (*recall*) y el error. Mientras que la exactitud explica cuántas, de todas las clases, se predicen correctamente, la sensibilidad indica cuántas se predicen acertadamente de todas las positivas. Por su parte, la especificidad refleja qué proporción de positivos reales se condicen con la

predicción y el error constituye el complemento de la exactitud. A continuación, se enumeran las respectivas fórmulas:

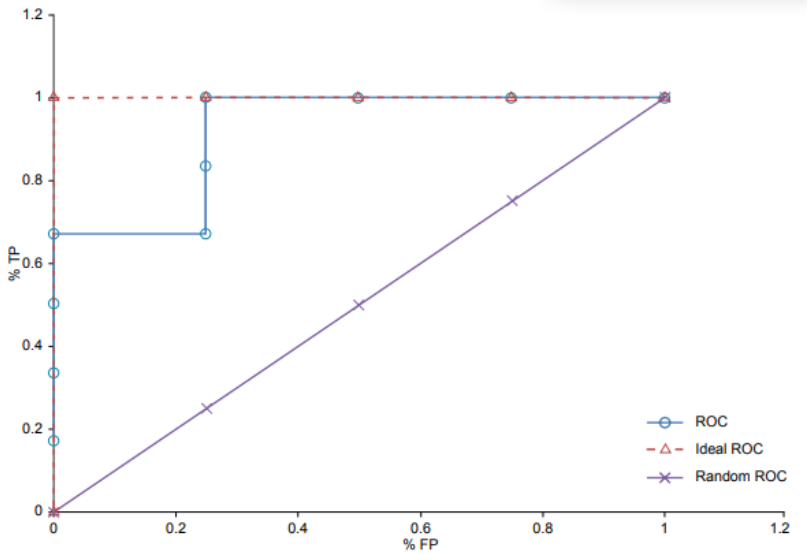
Exactitud = (VP + VN) / (VP + VN + FP + FN) (7)
Especificidad = VP / (VP + FP) (8)
Sensibilidad = VP / (VP + FN) (9)
Error = 1 – exactitud (10)

En una clasificación binaria, se puede establecer un umbral arbitrario para distinguir entre ejemplos falsos y verdaderos. Al aumentar el valor de dicho umbral, se reduce la cantidad de instancias clasificadas como positivas y se incrementa el número de aquellas clasificadas como negativas. Esto genera una disminución de los VP con el consiguiente aumento de los FN y un aumento de los VN con el correspondiente descenso de los FP. Por lo tanto, las tasas de VP (TVP) y FP (TFP) disminuyen (Musumeci et al, 2018, p. 1401).

Para distintos valores de umbrales, la curva ROC traza la TVP o la sensibilidad en el eje vertical y la TFP o (1- especificidad) en el eje horizontal. Esta curva constituye una representación gráfica de la sensibilidad frente a la especificidad según varía el umbral de discriminación, es decir, explicita los intercambios entre los VP y los FP. Un modelo predictivo con buen desempeño genera una curva ROC por encima de la diagonal del plano (TFP, TVP), por ello, los puntos ubicados por debajo de la diagonal representan resultados de clasificación pobres. El modelo de predicción perfecta se sitúa en la esquina superior izquierda o punto (0,1); con una sensibilidad del 100% (no existe ningún FN) y una especificidad del mismo valor (ningún FP). Esta coordenada en el espacio ROC recibe la denominación de clasificación perfecta (Figura 6).

Figura 6

Curva ROC



Nota. Tomado de *Predictive Analytics and Data Mining. Concepts and Practice with Rapidminer Studio* (p. 262), por V. Kotu, y B. Deshpande, 2015, Elsevier.

Cabe mencionar que, si bien la curva ROC es una herramienta gráfica eficiente para evaluar el desempeño de un clasificador, el AUC es una medida numérica que sintetiza la

performance del algoritmo independientemente del umbral que se haya elegido (Musumeci et al., 2018, p. 1401).

3.5 Selección del modelo con validación cruzada

Los métodos de remuestreo (*resampling methods*) o de validación permiten estimar la capacidad predictiva de uno o más modelos cuando se aplican a nuevas observaciones, utilizando solamente el conjunto de datos de entrenamiento. Como el error medido en la muestra de entrenamiento es un indicador deficiente para realizar una generalización del resultado del modelo, este último se ajusta empleando un subconjunto de observaciones del conjunto de entrenamiento y se evalúa con las observaciones restantes. Este proceso es iterativo y los resultados obtenidos se agregan para luego promediarse, permitiendo compensar las desviaciones que puedan surgir por la manera en que se distribuyeron las observaciones.

Entre los métodos de remuestreo más utilizados, cabe destacar el *bootstrap* y la validación cruzada (*cross-validation*). La técnica de *bootstrap* consiste en obtener, al menos de forma aproximada, la distribución de un estadístico utilizando la información que se deriva de una sola muestra (y sus réplicas). Por su parte, la validación permite evaluar el desempeño o rendimiento de un clasificador o, en su defecto, decidir a partir de una serie de clasificadores cuál es el mejor. A partir de la división aleatoria de la muestra en dos submuestras, una de entrenamiento y otra de prueba, permite obtener un error de predicción relativamente realista. El mejor clasificador es aquel cuya tasa de error proporciona el menor error de generalización (Sayeh y Bellier, 2014). Existen diversos procedimientos para aplicar este proceso, entre los que cabe mencionar la validación cruzada aleatoria, los métodos *Leave-One-Out Cross-Validation* (LOOCV), *k-fold cross-validation* y *repeated k-fold cross-validation* (Refaeilzadeh et al., 2009).

En el caso de *k-fold cross validation* (validación cruzada con *k* iteraciones), el conjunto de datos original se divide aleatoriamente en *k* subconjuntos mutuamente excluyentes, cada uno aproximadamente del mismo tamaño (Han et al., 2012, pp. 370-371). El modelo se entrena *k* veces utilizando cada uno de los *k* subconjuntos para la validación y los (k -1) restantes para el entrenamiento (Hastie et al., 2009, pp. 241-243). Este proceso se repite durante *k* iteraciones, alternando con cada uno de los subconjuntos de prueba disponibles y, por último, se calcula la media aritmética de los resultados de cada iteración. En este sentido, la validación cruzada sólo estima de manera efectiva el error promedio.

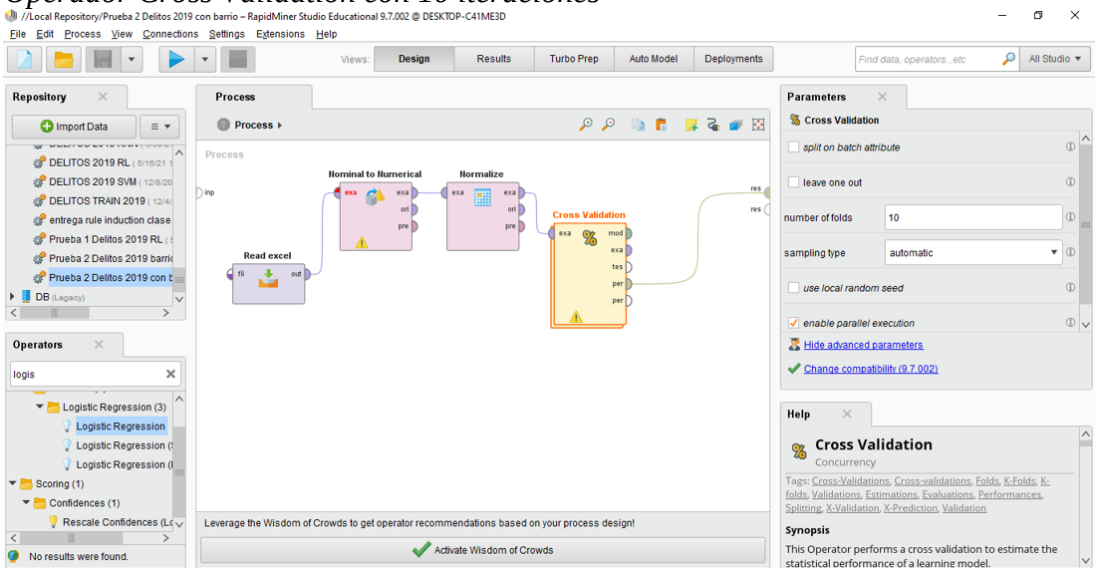
Cuando el objetivo de la clasificación es predecir un *output* y conocer la precisión del modelo, el método con *k* iteraciones garantiza que la evaluación de los resultados sea independiente de la partición entre el conjunto de datos de entrenamiento y el de prueba. Ahora bien, cabe preguntarse qué valor debe elegirse para el parámetro configurable *k*.

Existe consenso en la literatura sobre el valor que este parámetro debe adoptar en la práctica: suele utilizarse $k = 10$, es decir, para calcular la predicción de un modelo se utiliza repetidamente el 90% de los datos y se prueba su precisión en el 10% restante. Si el conjunto de entrenamiento es lo suficientemente grande y ese 10% de los datos tiene una distribución similar a las instancias etiquetadas, se obtiene una estimación confiable (Refaeilzadeh et al., 2009).

Tal como se muestra en la Figura 7, el operador *Cross Validation* realiza una validación cruzada dividiendo aleatoriamente el conjunto de entrenamiento y prueba y realizando una evaluación del modelo. Para realizar esta operación, resulta preciso eliminar el operador *Split Data.*

Figura 7

Operador Cross Validation con 10 iteraciones



Nota. Salida de *Rapidminer Studio*

4. Resultados obtenidos

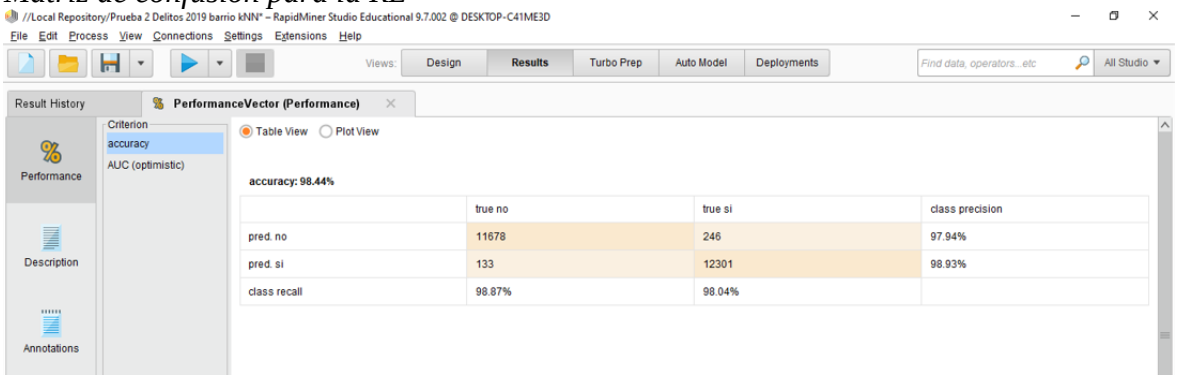
A partir del *dataset* "Delitos 2019" obtenido luego del tratamiento de los datos, se calcula el primer modelo denominado línea de base (*baseline*) utilizando *Microsoft Excel.* Para ello, se computa el cociente entre los delitos violentos y el total de los delitos con el objetivo de obtener una primera medida de exactitud. Para el año 2019, esta métrica asciende a 51,51% (62734/121788), lo que también demuestra que las clases están balanceadas. Asimismo, cualquier modelo predictivo seleccionado con posterioridad debe superar el valor de la línea de base.

Por su parte, el modelo de RL para todos los atributos, excluyendo latitud y longitud, barrio[1] y delitos por barrio en 2019, arroja una exactitud del 98,44%, con la matriz de confusión correspondiente a la Figura 8.

[1] La información del barrio está contenida en los atributos delitos por barrio, por ello excluirlo no representa una pérdida de información.

Figura 8

Matriz de confusión para la RL

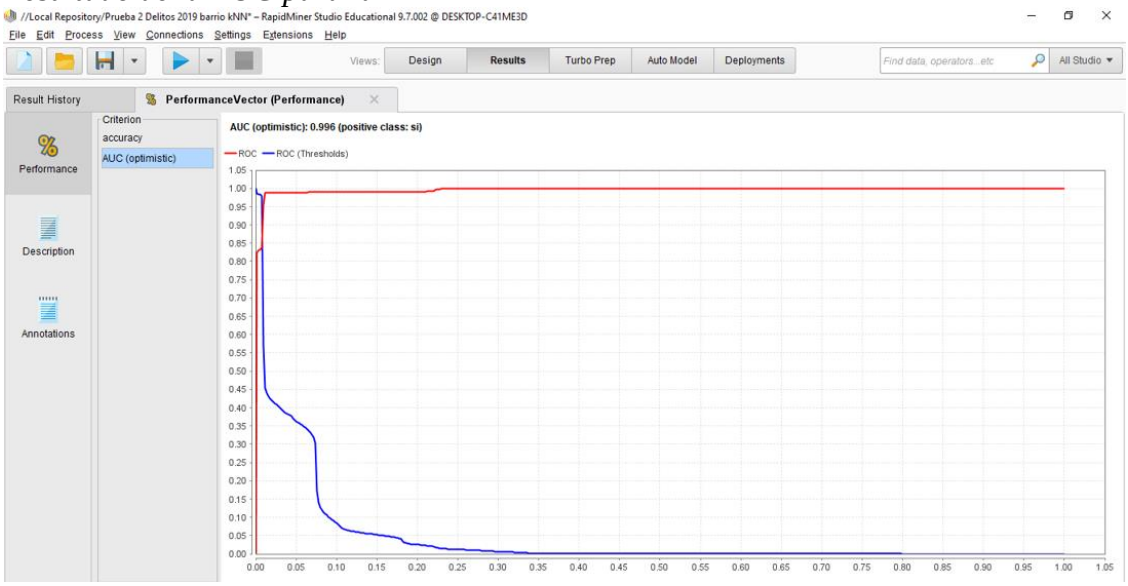


Nota. Salida de *Rapidminer Studio*

Con relación al AUC, la Figura 9 exhibe un resultado de 0,996.

Figura 9

Resultado de la AUC para la RL

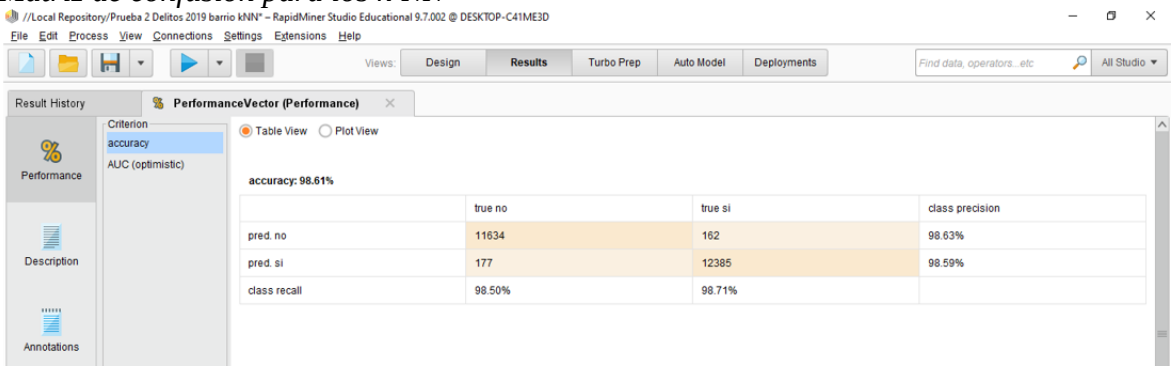


Nota. Salida de *Rapidminer Studio*

Utilizando los mismos atributos que para la RL, el algoritmo k-NN para $k = 3$ y *weighted vote*, arroja una exactitud de 98,61% con la matriz de confusión que se exhibe en la Figura 10.

Figura 10

Matriz de confusión para los k-NN

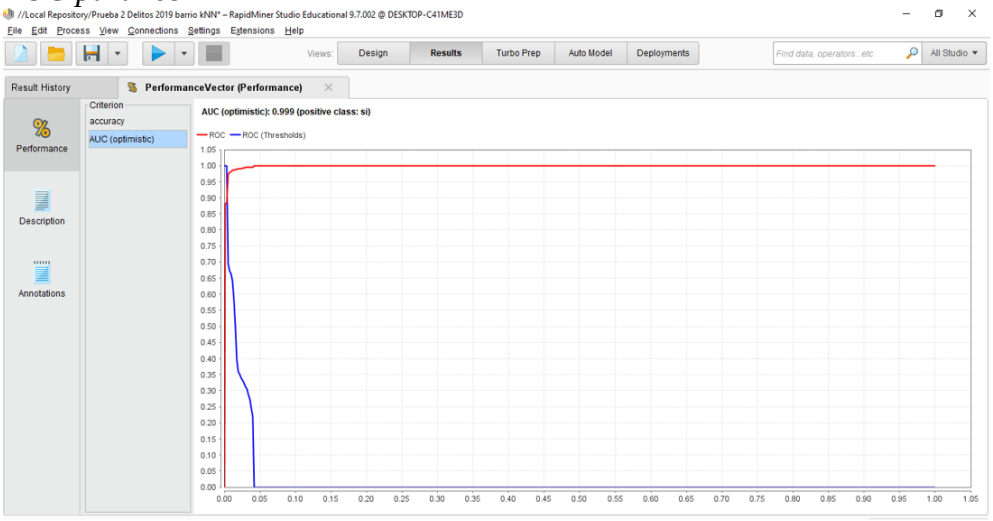


Nota. Salida de *Rapidminer Studio*

Con relación al AUC, la misma alcanza un valor de 0,995 (Figura 11).

Figura 11

AUC para los k-NN



Nota. Salida de *Rapidminer Studio*

Para comparar las métricas de ambos modelos, se confecciona la Figura 12 con el objetivo de simplificar la exposición de los resultados obtenidos.

Figura 12

Comparación de las métricas de la RL y los k-NN

Modelo/Métrica	Exactitud	Error	Especificidad	Sensibilidad	AUC
RL	0.9844	0.0156	0.9893	0.9804	0.996
k-NN	0.9861	0.0139	0.9859	0.9871	0.995

Por su parte, al realizar la validación cruzada en *Rapidminer Studio* para ambos modelos y con 10 iteraciones (k = 10), se obtienen los resultados exhibidos en la Figura 13.

Figura 13

Validación cruzada con RL y los k-NN

Modelo/Métrica	Exactitud	Desvío +/-	AUC
RL	0.9849	0.0010	0.996
k-NN	0.9864	0.0008	0.990

De los resultados obtenidos, se desprende que el algoritmo k-NN presenta una *performance* superior a la RL para predecir la ocurrencia de delitos violentos en CABA: la exactitud de los modelos alcanzan, con los parámetros especificados precedentemente, un 98,61% (AUC = 0,995) y un 98,44% (AUC=0,996) respectivamente. El resultado de la validación cruzada con 10 iteraciones utilizando los k-NN tiene una exactitud del 98,64% con una desviación de +/- 0,08% y el AUC alcanza un 0,999. Manteniendo la cantidad de iteraciones pero aplicando la RL, se obtiene una exactitud del 98,49% con una desviación de +/- 0,1% y el AUC asciende a 0,996.

5. Conclusiones

En las últimas décadas, el acceso a datos abiertos y el surgimiento de nuevas tecnologías basadas en internet han posibilitado la difusión de los modelos predictivos como técnicas idóneas para cuantificar los riesgos delictivos. A partir del cálculo de la probabilidad de ocurrencia de delitos violentos, la aplicación de los modelos de clasificación RL y k-NN permite optimizar la identificación de oportunidades y amenazas, crear un marco para la toma de decisiones informadas, perfeccionar los métodos de seguimiento y monitoreo y mejorar la prevención de hechos delictivos.

La gran cantidad de información circulante y las múltiples variables que intervienen para predecir la ocurrencia de los delitos, justifica la utilización de herramientas más potentes que los métodos estadísticos convencionales. En este sentido, las técnicas propias del aprendizaje automático brindan sistemas de medición de riesgos delictuales más precisos y personalizados. Al basarse en un aprendizaje continuo y automático, disminuye el margen de error y evalúa permanentemente patrones y desvíos de manera más eficiente que las herramientas estadísticas tradicionales.

Resulta fundamental, para llevar a cabo esta tarea, contar con bases de datos confiables y actualizadas que permitan diseñar políticas públicas adecuadas al territorio en el que se aplican, que sean flexibles y puedan contener el contexto cambiante y las nuevas formas de delincuencia. Al utilizar *datasets* dinámicos, todo nuevo conjunto de datos que se agregue permite no sólo una mejora con relación a la evaluación de riesgos sino también la identificación de riesgos emergentes o cambios en el contexto. Por eso, los datos constituyen un pilar prioritario en la gestión de riesgos. Asimismo, el GCBA debería fomentar no sólo una cultura de intercambio de información, tanto al interior de su sistema

como con sistemas externos, sino también procesos de colaboración y transparencia que posibiliten integrar todo el reservorio de datos de la jurisdicción para tener un diagnóstico más preciso que permita describir el fenómeno del delito en su totalidad.

En el GCBA, los procesos de recopilación, registro y almacenamiento de datos continúan siendo obsoletos y esto no se debe únicamente a la falta de sistemas tecnológicos. Las múltiples fuentes de información disponibles para evaluar la actividad delictiva (encuestas de victimización, registros administrativos policiales y judiciales; datos del registro civil, estadísticas confiables sobre el nivel de ingreso, el nivel educativo formal, entre otras) no pueden ser unificadas y estandarizadas con un criterio homogéneo. Sumado a lo antedicho, las bases de datos de delitos publicadas contienen pocos atributos y no representan la totalidad de los tipos delictivos. La inexistencia de ciertos datos o su falta de disponibilidad se puede deber a diversos factores, que incluyen desde el desconocimiento sobre el uso de los datos hasta la falta de voluntad política para que estén disponibles.

En síntesis, este trabajo permite vislumbrar la importancia del uso de datos para el diseño de políticas públicas, tradición que no parece estar extendida en el ámbito de gobierno, no solo en el área de seguridad. La subestimación o la subutilización de los datos para el armado de programas públicos suele ser un factor silenciado en la explicación de los fracasos de las políticas. Y también una excusa su falta de actualización o de interoperabilidad entre sistemas. Para utilizar datos y actualizarlos, es necesario que las políticas se diseñen con herramientas que permitan absorber el dinamismo del contexto e incorporar los cambios que se suceden en el ámbito donde son aplicadas. El uso de datos y la aplicación de modelos de predicción no es una garantía de éxito en el desarrollo de las políticas, pero sí la forma más adecuada de gestionar riesgos.

6. Bibliografía

Cong, B.N., Pérez, J.L.R. y Morell, C. (2015). Aprendizaje supervisado de funciones de distancia: estado del arte. *Revista Cubana de Ciencias Informáticas*, *9*(2), 14-28. https://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1014

Cover, T. y Hart, P (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. http://dx.doi.org/10.1109/TIT.1967.1053964

Dorofee, A., Walker, J., Alberts, C., Higuera, R., Murphy, R. y Williams, R. (1996). *Continuous Risk Management Guidebook***.** Software Engineering Institute, http://jodypaul.com/SWE/ContinuousRiskManagement.pdf

García Jiménez, V. (2010). *Distribución de clases no balanceadas: métricas, análisis de complejidad y algoritmos de aprendizaje*. [Tesis de Doctorado, Universitat Jaume I]. Repositorio UJI.

Garland, D. (2005). *La cultura del control*. *Crimen y orden social en la sociedad contemporánea* (Trad. M. Sozzo). Gedisa. (Trabajo original publicado en 2001).

Gobierno de la Ciudad de Buenos Aires (s.f.). *Delitos*. https://data.buenosaires.gob.ar/dataset/delitos

Han, J., Kamber, M. y Pei, J. (2012). Classification: Basic Concepts. En J. Han, M. Kamber y J. Pei (Eds.), *Data Mining: Concepts and Techniques* (3ª ed., pp. 327-392). Morgan Kaufmann. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Hastie, T., Tibshirani, R. y Friedman, J. (2009). Model Assessment and Selection. En T. Hastie, R. Tibshirani y J. Friedman (Eds.), *The elements of statistical learning* (2ª ed., pp. 219-260). Springer. https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Instituto Nacional de Estadística y Censos (2018). *Encuesta Nacional de Victimización 2017*. Ministerio de Economía. https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-27-137

Kotu, V. y Deshpande, B. (2015*). Predictive Analytics and Data Mining. Concepts and Practice with Rapidminer Studio*. Elsevier.

Musumeci F., Rottondi C., Nag A., Macaluso I., Zibar D., Ruffini M. y Tornatore, M. (2018). An overview on application of machine learning techniques in optical networks. *IEEE Communications Surveys & Tutorials, 21*(2), 1383-1408. https://doi.org/10.1109/COMST.2018.2880039.

Observatorio de Seguridad Ciudadana. (s.f.). *Encuesta Nacional de Victimización*. http://www.seguridadciudadana.org.ar/estadisticas/datos-a-nivel-subnacional/victimizacion-y-percepcion

Organización Internacional de Normalización. (2010). *Gestión de riesgos* (31000). https://www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:es

Pérez Verona, I. y Arco García, L. (2016*). Una revisión sobre aprendizaje no supervisado de métricas de distancia. *Revista Cubana de Ciencias Informáticas, 10*(4), 43-67. https://www.researchgate.net/publication/317514053_Una_revision_sobre_aprendizaje_no_supervisado_de_metricas_de_distancia

Prabakaran, S. y Mitra, S. (5-6 de enero de 2018). *Survey of analysis of crime detection techniques using data mining and machine learning* [Presentación en papel]. Congreso Nacional de Técnicas Matemáticas y sus Aplicaciones, Kattankulathur, India.

Refaeilzadeh, P., Tang, L. y Liu, H. (2009). Cross-Validation. En L. Liu y M.T. Özsu, (Eds.), *Encyclopedia of Database Systems*. Springer. https://doi.org/10.1007/978-0-387-39940-9_565

Rodríguez, P., Palomino, N. y Mondaca, J. (2017). *El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe*. Banco Interamericano de Desarrollo. http://dx.doi.org/10.18235/0000694

Sayeh, W. y Bellier, A. (12-13 de diciembre de 2014). Neural Networks versus Logistic Regression: The Best Accuracy in Predicting Credit Rationing Decision [Presentación en

papel]. Simposio Mundial de Banca Financiera 2014. Escuela de Negocios de Nanyang, Singapur.

Stamp, M. (2017). *Introduction to machine learning with applications in information security*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315213262

Valenga, F., Perversi, I., Fernández, E., Merlino, H., Rodríguez, D. Britos, P. y García-Martínez, R. (1-5 de octubre de 2007). *Aplicación de la minería de datos para la exploración y detección de patrones delictivos en Argentina* [Presentación en papel]. XIII Congreso Argentino de Ciencias de la Computación, Resistencia, Chaco, Argentina.