



Universidad de los Andes

Departamento de Ingeniería Eléctrica y Electrónica

PROYECTO FIN DE CARRERA

*Modelo de máxima entropía para la predicción de criminalidad con
técnicas de Machine Learning*

por

Juan Sebastian Murcia Ramirez - 201814401

js.murcia@uniandes.edu.co

Asesor - Fernando Enrique Lozano Martinez, Profesor asociado

Co-asesor - Juan Jose Garcia Cardenas, Profesor instructor

Jurado - Luis Felipe Giraldo Trujillo, Profesor asistente

Índice

1. Introducción	3
2. Objetivos	3
A. Objetivo General	3
B. Objetivos específicos	4
3. Justificación del trabajo	4
4. Marco teórico y conceptual	5
A. Marco teórico	5
A.1. Principio de Máxima entropía	5
A.2. Maxent - Software	6
B. Marco conceptual	8
B.1. Modelos 'Presence/Background'	8
B.2. Gain	8
B.3. AUC (Area Under the ROC Curve)	9
5. Metodología	9
A. Recolección de datos	10
A.1. Recolección de datos - Criminalidad	10
A.2. Recolección de datos - Variables	11
B. Preprocesamiento de los datos de criminalidad	12
C. Realización de los mapas	12
D. Sintonización del modelo	13
E. Obtención de modelos finales	13
E.1. Primer filtro	14
E.2. Segundo filtro	14
6. Desarrollo - Trabajo realizado	14
A. Sintonización del modelo	14
A.1. Constante de regularización	14
A.2. Generación de puntos aleatorios - Background	15
A.3. Entrenamiento del modelo con los promedios de las variables	16
B. Resultado final	17
B.1. Tipo de crimen 1	17
B.2. Tipo de crimen 2	18
B.3. Tipo de crimen 3	19
B.4. Tipo de crimen 4	20
B.5. Tipo de crimen 5	21
B.6. Tipo de crimen 6 - Robo de vehículos	22
7. Discusión	23
8. Conclusiones	25

9. Agradecimientos	26
10. Referencias	26

1. Introducción

La criminalidad es la cantidad o proporción de crímenes cometidos en un lugar o en un periodo de tiempo determinados, [1] lo cual puede tener repercusiones negativas en el estilo de vida de una comunidad o población. En la actualidad existen diferentes formas en las que se cometen crímenes, como lo son el hurto a objetos, el robo de vehículos, asaltos a casas, etc. Este tipo de acciones impulsan la creación de herramientas tecnológicas que permitan mitigar o controlar la frecuencia en la ocurrencia de los diferentes tipos de crimen. Por ende, en este proyecto se desarrolló un modelo que permite predecir y visualizar donde es más probable que ocurra un crimen.

En este caso se desarrolló el modelo para 6 tipos de crimen, el cual fue entrenado con ocurrencias entre los años 2015 al 2018 en la ciudad de Vancouver, y probado con datos del 2019 en la misma ciudad. Esta implementación se realizó utilizando el software Maxent, como también se realizaron procesos de sintonización de parámetros, con el objetivo de obtener un rendimiento óptimo en el modelo.

Tipos de crimen

- $Type_1$ = Break and enter commercial (Entrar en un establecimiento comercial con la intención de cometer una ofensa.)
- $Type_2$ = Break and enter Residential/Other. (Entrar a una casa/apartamento/garaje con intención de cometer una ofensa.)
- $Type_3$ = Mischief (Dañar propiedad publica o privada con intención maliciosa)
- $Type_4$ = Other theft (Robo de objetos personales)
- $Type_5$ = Theft of bycycle (Robo de bicicletas)
- $Type_6$ = Theft of vehicle (Robo de vehículos, motos o cualquier vehículo con motor)

2. Objetivos

A. Objetivo General

Desarrollar un modelo de Machine Learning que permita predecir y visualizar la probabilidad de ocurrencia en una grid de la ciudad de Vancouver.

B. Objetivos específicos

- Encontrar las variables más representativas para la predicción de la criminalidad.
- Implementar un método para la visualización de los resultados del algoritmo de criminalidad.
- Seleccionar el mejor modelo, según la capacidad de generalización, como también en el desempeño con los datos de prueba, basándose en la función de perdida.

3. Justificación del trabajo

El almacenamiento y análisis de datos permiten generar herramientas con la capacidad de reconocer patrones lineales o no lineales de los datos con el propósito de desarrollar un instrumento que respalde la toma de decisiones futuras o que brinde predicciones. Por lo tanto, estas herramientas se pueden implementar para la predicción de la distribución de probabilidad de ciertos tipos de crimen en una ciudad, por medio de un algoritmo entrenado con datos históricos geográficos de criminalidad, como también factores climáticos o distribuciones demográficas de la ciudad. Esto permite generar un instrumento con el cual las instituciones policiales, podrán generar un plan de distribución eficiente de su capital humano respaldado por datos históricos, aumentando la probabilidad de la disponibilidad de los policías en los lugares que se les necesiten.

Además, este tipo de implementaciones permitirá reducir el impacto de decisiones tomadas que presenten sesgos sociales, culturales y/o económicos, ya que el algoritmo le brindará la información basándose en otros factores, dificultando que la persona enfoque su atención erradamente.

Un modelo de Máxima entropía es implementado cuando solo se tienen datos de presencia, es decir que no se podrá calcular específicamente si va a ocurrir o no un crimen, pero si se lograra encontrar las condiciones que permitirán entender donde es mas probable que exista un crimen, además es difícil poder establecer la no ocurrencia de un crimen, porque no es posible afirmar que todos los crímenes fueron notificados.

Por otra parte, un modelo de Máxima entropía es altamente utilizado en implementaciones en campos como [4]: la geografía; modelos biológicos, ecológicos y médicos; planificación urbana, regional y transportes; termodinámica y mecánica estadística; reconstrucción de imágenes, etc. Sin embargo también se ha implementado para distribución de poblaciones, normalmente biológicas [6], por lo que es llamativo abarcar la criminalidad por medio de este modelo.

4. Marco teórico y conceptual

A. Marco teórico

A.1. Principio de Máxima entropía

El principio de máxima entropía se basa en encontrar una distribución de probabilidad desconocida por medio de un problema de optimización, en donde la intuición está en encontrar una distribución de probabilidad que presenta un cierto factor de ignorancia desde la perspectiva del fenómeno que se está analizando [2], sin dejar de lado la información que ofrecen los datos. En pocas palabras, se maximiza la entropía sujeta a que la distribución que se elija sea compatible con los datos observados, obteniendo como beneficios encontrar relaciones no lineales y poder predecir.

La función de entropía, definido por Shannon [3] (1), se formula como:

$$S = - \sum_{m=0}^M P_m \ln(P_m) \quad (1)$$

donde

$$\sum_{m=0}^M P_m = 1$$

En donde P_m son las probabilidades de cada variable aleatoria x_m . Por lo tanto, se puede interpretar que el valor mínimo de S , se presenta cuando el valor de una probabilidad P_j tiende a uno y el resto a cero, es decir cuando no existe casi incertidumbre. Por otra parte, es máxima cuando todas las probabilidades tienen el mismo valor (2), conformando una distribución de probabilidad uniforme.

$$p_m = \frac{1}{M}, \forall m = 1, \dots, M \quad (2)$$

Ahora desde la perspectiva del modelo de optimización, el sistema se define como:

$$\text{Max } S = - \sum_{i=1}^L P(z_i(x)) \ln(P(z_i(x))) \quad (3)$$

Sujeto a

$$\begin{aligned} g_1 &\rightarrow z_1(x) \geq C_1 \\ g_2 &\rightarrow z_2(x) \geq C_2 \end{aligned}$$

$$g_L \rightarrow < z_L(x) > = C_L$$

$$g_{L+1} \rightarrow \sum_{i=1}^L P(z_i(x)) = 1$$

donde $P(z_i(x))$ es la probabilidad del valor de la variable aleatoria x , para cada parámetro z_i del modelo (L cantidad de parámetros). Las restricciones g_1, g_2, \dots, g_L , se refieren a que se desea ajustar la distribución de probabilidad al valor esperado C_i de la variable aleatoria x de cada z_i . La restricción g_{L+1} , restringe la suma de todas las probabilidades igual a 1.

Para resolver este sistema de optimización, se opta por el método de multiplicadores de Lagrange [5], en donde la derivada de la función objetivo (4), será igual a una combinación lineal de las derivadas de las restricciones escaladas por los multiplicadores de Lagrange.

$$\frac{\partial f}{\partial x} = \sum_{i=1}^L \lambda_i \frac{\partial g_i}{\partial x} \quad (4)$$

Este método conlleva a obtener una función exponencial (5), en donde los parámetros a encontrar son los multiplicadores de Lagrange, lo cual para un modelo de múltiples variables puede ser una tarea exhaustiva, por lo que se implementan técnicas de Machine learning para encontrar el valor de estas variables.

$$P(x) = \frac{e^{-\Lambda \cdot Z(x)}}{N} \quad (5)$$

donde Λ , es igual al vector de los multiplicadores de Lagrange $[\lambda_1, \lambda_2, \dots, \lambda_L]$ con todos los $\lambda \geq 0$ y $Z(x)$ es el valor de cada parámetro $[z_1(x), z_2(x), \dots, z_L(x)]$ evaluado en x . Finalmente N es una constante de normalización para que $P(x)$ sume 1.

A.2. Maxent - Software

Maxent es un software diseñado por Steven J.Philips. et.al [6], basado en el método de Máxima entropía para el modelamiento de la distribución y el niche de especies [7]. El software tiene una interfaz de usuario en donde las personas pueden editar las variables de ingreso al modelo, las gráficas de salida o parámetros asociados al modelo.

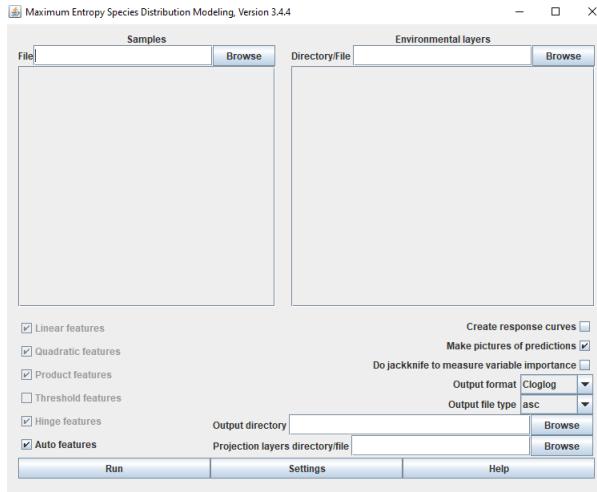


Figura 1: Interfaz de usuario - Software Maxent

Internamente lo que el software realiza es obtener una distribución de probabilidad $E(x)$ muestreando el mapa de las variables en los puntos en donde se haya dado una presencia, con lo cual se obtiene una caracterización de las variables en donde se presentan crímenes (Figura 2). Por otra parte, crea puntos aleatorios en todo el mapa ('Background') y obtiene una distribución de probabilidad $Q(x)$ con todas las variables, en donde se obtiene una caracterización de todo el mapa del estudio [8].

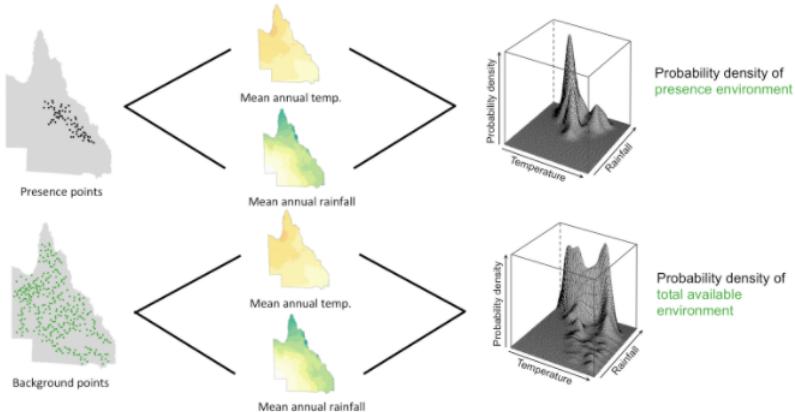


Figura 2: Ejemplo de calculo de las distribuciones de probabilidad [8]

Finalmente Maxent calcula una proporción entre estas dos densidades de probabilidad (6), obteniendo una distribución de probabilidad $E'(z(x))$ que reconoce las condiciones de las variables que mas se ajustan para que se de una

presencia. Ya que al maximizar la entropía relativa entre $E'(z(x))$ y $Q(z(x))$ (7), se obtiene una distribución de probabilidad lo mas cerca posible a la distribución $Q(z(x))$, del Background, mientras que se respeta las restricciones establecidas por la densidad de probabilidad calculada usando los datos de presencia $E(z(x))$ [8].

$$\frac{E(z(x))}{Q(z(x))} = \frac{E'(z(x))}{Q(z(x))} = \frac{e^{-\Lambda Z(x)}}{\text{constant}} \quad (6)$$

$$D(E' || Q) = - \sum_{i=1}^L E'(z_i(x)) \cdot \log \left(\frac{E'(z_i(x))}{Q(z_i(x))} \right) \quad (7)$$

B. Marco conceptual

B.1. Modelos 'Presence/Background'

Existen métodos (Maxent, ENFA, etc) [9] que modelan la distribución de especies por medio de datos de presencia y el muestreo de las variables ambientales ('Background'). Estos modelos, se enfocan en encontrar la relación que existe entre las condiciones ambientales en las que se provocaron presencias, con respecto a las condiciones ambientales de toda la ciudad, permitiendo diferenciar las condiciones o características del ambiente en las cuales es mas probable que se realice un avistamiento. En el caso del estudio de la criminalidad se quiere encontrar esas características por medio de los mapas socio-demográficos y climáticos que hacen mas probable la ocurrencia de un tipo de crimen.

También existen otros tipos de modelos como lo son de solo *Presencia* (Gower Metric) o de *Presencia/Ausencia* (Genetic algorithm, Artificial Neural Network, Regression, etc) [9].

B.2. Gain

Es una métrica implementada por el software Maxent (8) y utilizada en cada iteración con la intención de maximizarla, para encontrar el mejor modelo que tiene la capacidad de diferenciar entre locaciones en donde hay presencia o background [8]. Es decir, que el gain permitirá apreciar que tan concentrado se encuentra el modelo a los datos de presencia. Por ejemplo, si se tiene una ganancia de 1.5, significa que en promedio la distribución de probabilidad de los datos de presencia se encuentran concentrados ($e^{1.5}$) 4.48 veces más que en los puntos del background [10].

$$gain = \underbrace{\frac{1}{M} \sum_{m=0}^M P_m \cdot \lambda}_{\text{Sum of predicted values at presence locations}} - \underbrace{\log \sum_{m=0}^N Q_m e^{P_m \lambda}}_{\text{Sum of predicted values at background locations}} - \underbrace{\sum_{i=0}^I |\lambda_i| \cdot \beta \cdot \sqrt{\frac{s^2[S_j]}{M}}}_{\text{LASSO regularization}} \quad (8)$$

Se aprecia que el Gain utiliza *LASSO*, con el objetivo de prevenir un sobre ajuste a los datos, debido a que el propósito principal es encontrar una distribución de probabilidad que cumple las restricciones que imponen los datos, sin embargo no se quiere un modelo con las mismas restricciones, si no con la capacidad de generalizar las principales características para intentar realizar predicciones.

B.3. AUC (Area Under the ROC Curve)

El *Receiver Operating Characteristic* ROC, es una métrica que nos permite analizar el comportamiento del modelo de clasificación y el AUC es el area bajo la curva del ROC (Figura 3), la cual permite dimensionar de una manera mas fácil, la capacidad del modelo para diferenciar entre un positivo y un negativo, evaluándolo en todo el rango de umbral.

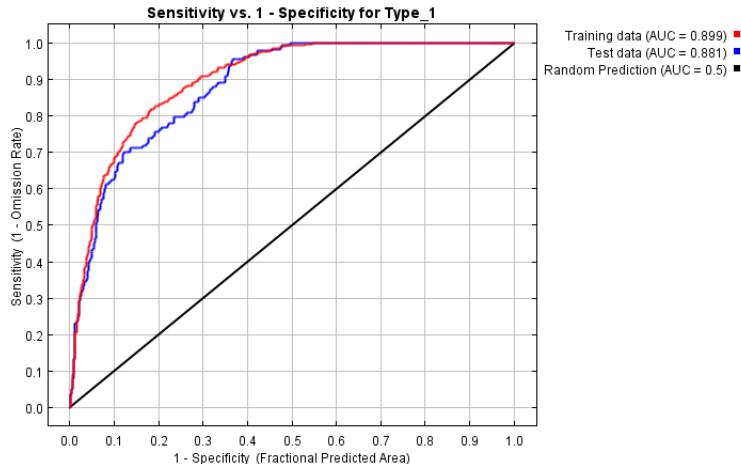


Figura 3: AUC-ROC

Para esta implementación el AUC, es una métrica que permite evaluar y comparar los diferentes modelos que se implementen, en donde lo que se busca medir, es la capacidad del modelo para diferenciar entre una ocurrencia y un punto de Background, sin llegar a hacer un análisis de predicción binaria por medio de la selección de un threshold.

5. Metodología

Para la resolución de este proyecto, el proceso se realizo en 6 etapas listadas a continuación:

- Recolección de datos.

- Preprocesamiento de los datos y variables.
- Realización de los mapas.
- Inicio de simulaciones.
- Sintonización del modelo.
- Obtención de resultados finales.

A. Recolección de datos

A.1. Recolección de datos - Criminalidad

Antes de iniciar la etapa de recolección de datos, se realizó una búsqueda de las ciudades que brindan datos acerca de la criminalidad, como también que estos fueran de libre acceso y tuviera una licencia que permitiera usarlos en este proyecto.

- **Vancouver**
- **Boston**
- **Denver**
- **Philadelphia**
- **New York**
- **Atlanta**

Finalmente, después de hacer un análisis de cada dataset y basados en el tamaño de la ciudad, tamaño del archivo, formato de los datos y distribución de la ciudad para una implementación en grids. Se escogió la ciudad de Vancouver.



Figura 4: Mapa de vancouver [11]

A.2. Recolección de datos - Variables

En el momento que se escogió la ciudad de Vancouver, se procedió a buscar las variables que se deseaban utilizar en el modelo, para lo cual se implementaron dos categorías. Mapas sociodemográficos y climatológicos. Para el primero se procedió a utilizar el censo del 2016 de la ciudad de Vancouver [12], por lo cual se decide realizar el estudio para el intervalo de tiempo entre 2015 al 2019, y además escoger las siguientes estadísticas del censo, las cuales están divididas para cada barrio que hay en la ciudad de Vancouver (Figura 4).

- Cantidad de hombres en el rango de edad - 15 a 64 años.
- Cantidad de mujeres en el rango de edad - 15 a 64 años.
- Promedio de ingresos totales en 2015.
- Proporción de desempleo en 2016.
- Proporción de desempleo en 2016 - Hombres.
- Proporción de desempleo en 2016 - Mujeres.
- Actividad laboral durante 2016 - No trabaja.

Además, se agregaron dos características; localización de las cámaras de la ciudad y la distribución de la policía en la ciudad. Por otra parte, para las variables climatológicas se utilizó los mapas mensuales desde 2015 al 2018 de temperatura máxima, temperatura mínima, temperatura promedio y precipitación de la base de datos climatológica *CHELSA* (Climatologies at high resolution for the earth's land surface areas). [13]

B. Preprocesamiento de los datos de criminalidad

Inicialmente lo que se realizó es eliminar dos columnas de los datos de criminalidad que contenían información acerca del nombre del barrio y la dirección ('Neighbourhood' y 'HundredBlock'), ya que lo realmente importante era la ubicación georeferenciada. Por lo tanto las características finales para los datos de criminalidad son el año, mes, hora, minuto, coordenada X y coordenada Y. Sin embargo, se necesitó realizar el cambio de tipo de coordenadas, las cuales se encontraban en un formato estándar UTM (Zona 10, Hemisferio N) a coordenadas en grados decimales (WGS 84), debido a que este tipo de coordenadas es como están establecidas las variables climatológicas, además de ser el formato utilizado por el software Maxent.

Además se eliminaron todos los datos que no se encuentren dentro de estas coordenadas geográficas [(49.313348 N, 49.20089 N), (-123.224020 W, -123.02346 W)], principalmente porque es la región de Vancouver en donde se va a realizar el análisis, como también para eliminar los datos que se encontraban mal digitados.

Finalmente, una vez que realizó el procedimiento anteriormente descrito, se procedió a dividir los datos en tres franjas horarias, datos tipo *Dia*; El cual contiene crímenes realizados todo el día, datos tipo *Mañana*; El cual contiene crímenes realizados desde las 00:00 hasta las 11:59, y finalmente datos tipo *Noche*; El cual contiene crímenes realizados desde las 12:00 hasta las 23:59 del mismo día. Esto con el objetivo de hacer modelos para estas franjas horarias y así realizar una comparación. Esta implementación se realizó para dividir los datos anualmente; datos conjuntos desde 2015 a 2018. Como también para dividir los datos mensualmente; datos conjuntos mensuales desde el 2015 al 2018, por ejemplo, todos los datos de enero del 2015, 2016, 2017 y 2018, para realizar una predicción de enero del 2019.

C. Realización de los mapas

En la realización de los mapas se implementó el software geográfico libre QGIS. En este se realizaron la edición de todos los mapas utilizados en este proyecto. Para los datos climatológicos, se debió de extraer de todos los mapas mes a mes desde 2015 hasta 2018 la ciudad de Vancouver (Un total de 48 mapas) y finalmente exportarlos al formato ASC, para que pueda ser leído por el software Maxent.

Por otra parte, para la realización de las variables sociodemográficas se realizó cada mapa disponiendo del valor que establecía el censo para cada barrio (Figura 5), también se realizó la exportación correspondiente al formato ASC.



Figura 5: Mapa de la variable '*avrg_total_income_2015*'

D. Sintonización del modelo

En el momento que se entendió el manejo básico del software Maxent gracias a la documentación [10]. Se procedió a realizar una sintonización de algunas variables que posee un modelo de máxima entropía, midiendo el rendimiento por medio del *Gain* y el *AUC*.

- Se vario la constante de regularización.
- Se vario la forma de generación de los puntos aleatorios pertenecientes al background.
- Se realizo entrenamiento de los modelos con todas las variables climatológicas, como también con solo los promedios de estas variables.

Es relevante mencionar que se escogieron el *Gain* y el *AUC* como métricas de rendimiento porque el *Gain*, permite comparar la concentración de la distribución de probabilidad de los datos de entrenamiento y los de prueba, para así poder comprender si el modelo esta realizando un sobre ajuste de los datos de entrenamiento. En cambio la implementación del *AUC*, permite entender si el modelo tiene la capacidad de diferenciar entre un punto de presencia y un punto del Background, lo cual es relevante si se desea realizar una predicción binaria. Sin embargo en esta implementación va a dar otra perspectiva a parte del *Gain*, de como se comporto el modelo en el entrenamiento, para diferenciar entre las presencias y los puntos aleatorios del Background.

E. Obtención de modelos finales

Finalmente una vez obtenido los mapas de todas la variables para cada año, como también los mejores parámetros del modelo, se procedió a escoger el mejor modelo comparando el rendimiento en las métricas *AUC* y *Gain*, en los datos de prueba.

E.1. Primer filtro

Para escoger el primer grupo de modelos, se comparo el rendimiento de un *Modelo₁* entrenado con datos de la franja de horario *Día*, con dos modelos (*Modelo₂*), con la franja de horario *Mañana* y *Noche*. Esto con el objetivo de intentar averiguar si las dinámicas de criminalidad son totalmente diferentes antes o después del medio día.

Para poder hacer comparable estos dos tipos de modelos, lo que se realizo es entrenar el *Modelo₁* con los datos de la franja de horario *Día*, pero probarlo con los datos de prueba del *Modelo₂*, para la *Mañana* y *Noche* por separado. Finalmente se tabulo los rendimientos de los modelos para todos los meses con los datos de prueba y se comparo el rendimiento promedio, para cada tipo de crimen.

E.2. Segundo filtro

Después del primer filtro, se inicio la comparación entre el *Modelo₃*, el cual fue entrenado con datos de todos los meses, desde el 2015 al 2018, en la franja de horario *Día*. Con un *Modelo₄*, que fue entrenado para cada mes desde el 2015 al 2018, en la franja de horario *Día*.

Para poder realizar la comparación el *Modelo₃* y *Modelo₄*, fueron puestos a prueba con los datos de cada mes en el año 2019, en la franja de horario *Día*.

6. Desarrollo - Trabajo realizado

A. Sintonización del modelo

A.1. Constante de regularización

Realizar la sintonización de la constante de regularización es un proceso obligatorio para modelos de Máxima entropía, debido a que sin este tratamiento, la distribución de probabilidad predicha tendría las mismas características que las restricciones de la distribución de probabilidad de las variables de entrenamiento, instancia que no es deseada, ya que se desea realizar una predicción en el año 2019, con datos del 2015 al 2018.

por lo cual se procede a hacer un entrenamiento del modelo para cada tipo de crimen con las siguientes constantes de regularización [0.001, 0.01, 0.1, 1, 2, 5], obteniendo el siguiente comportamiento (Figura 6).

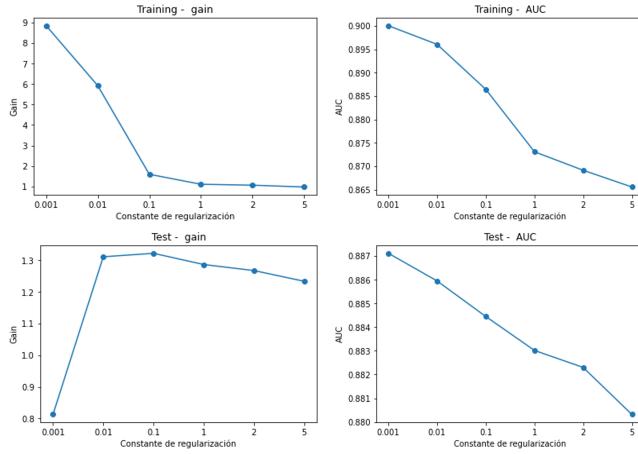


Figura 6: Comportamiento en promedio de los modelos, variando la constante de regularización.

Se logra apreciar, que en promedio para todos los tipos de crimen, cuando se disminuye el parámetro de regularización se obtiene un AUC cada vez mas alto. Sin embargo, al analizar la ganancia se entiende que el modelo se esta sobre ajustando a los datos de entrenamiento, entregando como consecuencia un rendimiento no deseado en la ganancia de los datos de prueba cuando se disminuye el parámetro de regularización. Por lo tanto, se escoge a **0.1**, como el valor por defecto para el entrenamiento de todos los modelos.

A.2. Generación de puntos aleatorios - Background

En este caso el software Maxent permite tres configuraciones para localizar los puntos del Background.

- Todo_presencia = Puede agregar todos los puntos de presencia a los puntos de background.
- Alguna_presencia = Puede agregar algunos puntos de presencia a los puntos de background.
- Sin_presencia = No agregar ningn dato de presencia a los puntos de background.

Esto con el objetivo de comprobar si era mejor, que el modelo influenciara la distribución de probabilidad de las variables a puntos en donde se haya dado una presencia, es decir que el modelo caracterizara mas las partes del terreno en donde se dio una presencia. Sin embargo al realizar esta prueba se aprecio que el modelo tenia mejor rendimiento cuando no se agregaba información especifica de los puntos de ocurrencia, por lo tanto es mejor permitir que el modelo haga un muestreo aleatorio del terreno, para esta implementación.

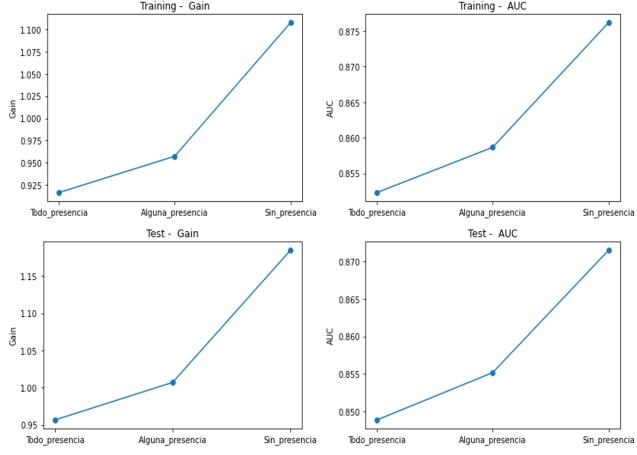


Figura 7: Comportamiento en las métricas de rendimiento al variar el método del Background

A.3. Entrenamiento del modelo con los promedios de las variables

Otro análisis que se llevo a cabo, desde la perspectiva de las variables climatológicas, fue realizar el mismo modelo para cada tipo de crimen, pero entrenado con diferentes distribuciones temporales de las variables climatológicas, en donde:

- Anual = Se realizo el entrenamiento con las variables climatológicas de los años 2015 al 2018.
- Promedio = Se realizo el entrenamiento con el promedio de las variables climatológicas de los años 2015 al 2018.
- Todos = Se realizo el entrenamiento con las variables climatológicas discretizadas por cada año, como también junto al promedio de los 4 años.

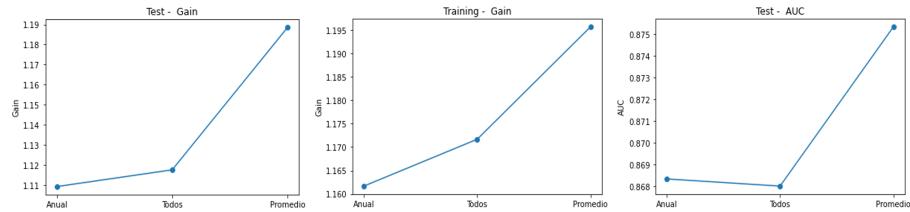


Figura 8: Comportamiento de los modelos al variar la distribución de las variables climatológicas.

Se percibe en la Figura 8, que el mejor modelo era el que solo tenia el promedio de la información del clima. Esto se debe a que un modelo de Máxima

entropía por naturaleza, va a intentar restringir la distribución de probabilidad predicha a los promedios de las variables que entran, lo cual implica que al agregar mas información de una misma variable, va a producir que el modelo se concentre mas en patrones del clima de cada año y no se enfoque en los patrones generales.

B. Resultado final

Se presentan desde la Figura 9 hasta la Figura 14, la distribución de probabilidad predicha, como también se ven los puntos, los cuales simbolizan el lugar en donde ocurrió un crimen en el año 2019. Es importante mencionar que el modelo final y el cual obtuvo mejor rendimiento en promedio en todos los tipos de crimen y para todo el año, fue el que se entrenaba con variables mensuales y con la franja de horario *Día*.

B.1. Tipo de crimen 1

Para los crímenes de ingreso a establecimientos comerciales, se aprecia una alta concentración en la zona norte de la ciudad. Lo cual es lo esperado ya que es el 'centro' de la ciudad en donde esta concentrado el comercio y el turismo de la ciudad.

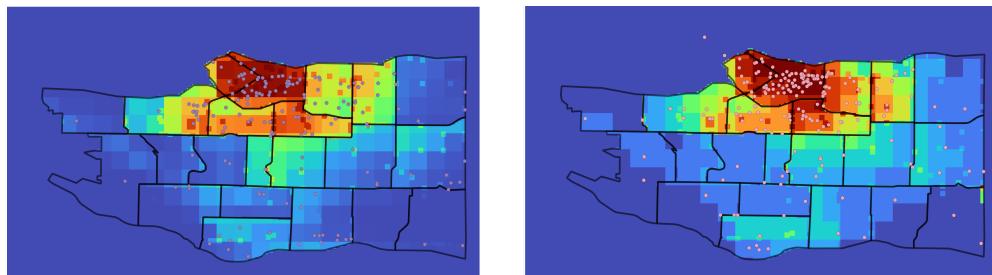


Figura 9: Predicción del modelo final para Enero y Julio - Tipo de crimen 1.

	Test_AUC	Test_Gain
Enero	0.881	1.227
Julio	0.907	1.497
Promedio	0.911	1.5057

Cuadro 1: Rendimiento del modelo - Crimen tipo 1

Variable	Porcentaje de contribución
ZonasPolicias	26.9
distriEdades15_64_male	18.5
avrg_total_income_2015	16.3
distriEdades15_64_female	18.5
Otras	19.8

Cuadro 2: Análisis de contribución de las variables - Crimen tipo 1

B.2. Tipo de crimen 2

A diferencia del crimen Tipo 1, este se encuentra distribuido a lo largo de la ciudad, ya que es el robo a establecimientos privados como casas/apartamentos/garaje, sin embargo se aprecia que hay zonas donde es mas probable que ocurra un crimen.

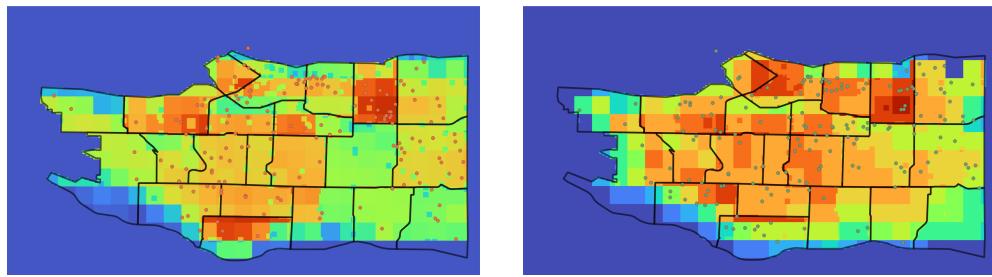


Figura 10: Predicción del modelo final para Enero y Julio - Tipo de crimen 2

	Test_AUC	Test_Gain
Enero	0.821	0.770
Julio	0.807	0.754
Promedio	0.8118	0.77117

Cuadro 3: Rendimiento del modelo - Crimen tipo 2

Variable	Porcentaje de contribución
ZonasPolicias	28.6
distriEdades15_64_male	24.3
avrg_total_income_2015	20.4
distriEdades15_64_female	8.8
Otras	10.7

Cuadro 4: Análisis de contribución de las variables - Crimen tipo 2

B.3. Tipo de crimen 3

El daño a propiedad publica o privada, también se encuentra distribuido a lo largo de la ciudad, pero con una concentración en el 'centro' de la ciudad.

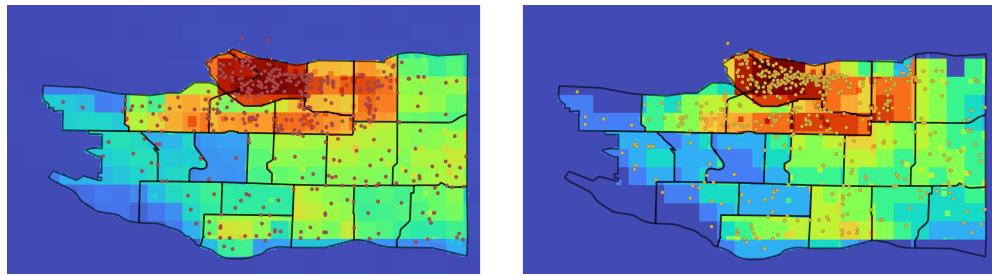


Figura 11: Predicción del modelo final para Enero y Julio - Tipo de crimen 3

	Test_AUC	Test_Gain
Enero	0.882	1.227
Julio	0.884	1.252
Promedio	0.8855	1.2614

Cuadro 5: Rendimiento del modelo - Crimen tipo 3

Variable	Porcentaje de contribución
ZonasPolicias	32.3
distriEdades15_64_male	20.3
avrg_total_income_2015	16.9
distriEdades15_64_female	9.6
Otras	20.9

Cuadro 6: Análisis de contribución de las variables - Crimen tipo 3

B.4. Tipo de crimen 4

El robo de objetos personales, parece tener dinámicas diferentes en los meses de Enero (Invierno) y Julio (Verano), lo cual en este ultimo mes se concentra mas en hacia el ‘centro’ de la ciudad, posiblemente debido a que es la zona turística de la ciudad.

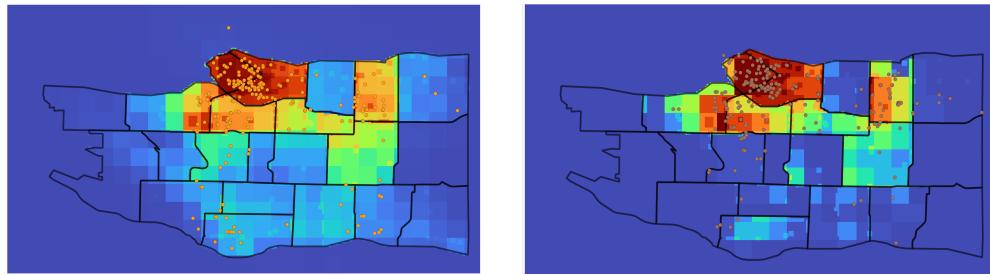


Figura 12: Predicción del modelo final para Enero y Julio - Tipo de crimen 4

	Test_AUC	Test_Gain
Enero	0.924	1.683
Julio	0.941	1.874
Promedio	0.9350	1.8145

Cuadro 7: Rendimiento del modelo - Crimen tipo 4

Variable	Porcentaje de contribución
ZonasPolicias	35
distriEdades15_64_male	26
distriEdades15_64_female	9.1
avrg_total_income_2015	6.2
Otras	23.7

Cuadro 8: Análisis de contribución de las variables - Crimen tipo 4

B.5. Tipo de crimen 5

El aumento de casos de robos de bicicletas aumenta en verano, lo cual seria lo esperado, debido a que es la temporada del año, en donde es mas ideal el uso de este transporte por las condiciones climáticas.

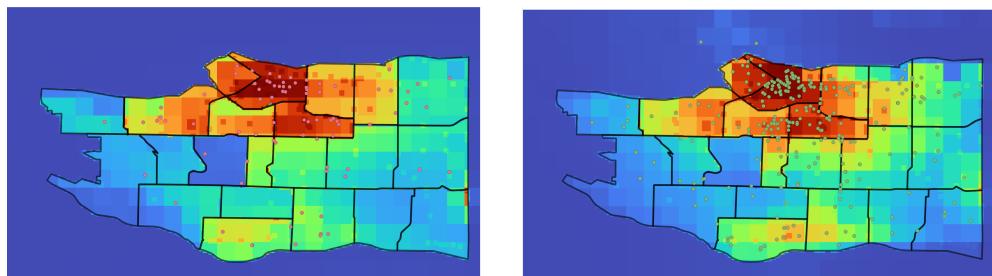


Figura 13: Predicción del modelo final para Enero y Julio - Tipo de crimen 5

	Test_AUC	Test_Gain
Enero	0.883	1.306
Julio	0.882	1.248
Promedio	0.8814	1.2458

Cuadro 9: Rendimiento del modelo - Crimen tipo 5

Variable	Porcentaje de contribución
ZonasPolicías	39.7
distriEdades15_64_male	18.9
distriEdades15_64_female	10.1
avrg_total_income_2015	8.9
Otras	22.4

Cuadro 10: Análisis de contribución de las variables - Crimen tipo 5

B.6. Tipo de crimen 6 - Robo de vehículos

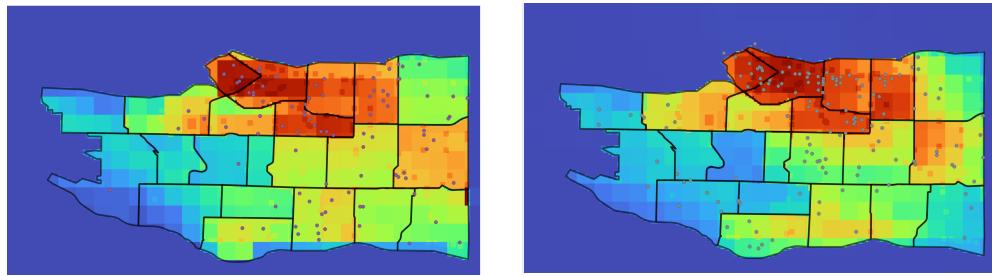


Figura 14: Predicción del modelo final para Enero y Julio - Tipo de crimen 6

	Test_AUC	Test_Gain
Enero	0.860	0.984
Julio	0.845	0.922
Promedio	0.8599	1.0239

Cuadro 11: Rendimiento del modelo - Crimen tipo 6

Variable	Porcentaje de contribución
ZonasPolicías	38
distriEdades15_64_male	22.3
avrg_total_income_2015	11.5
distriEdades15_64_female	9.2
Otras	19

Cuadro 12: Análisis de contribución de las variables - Crimen tipo 6

7. Discusión

Se aprecia que las dos variables mas representativas en este proyecto, son las ZonasPolicías y distriEdades15_64_male. La primera se podría comprender debido a que es una variable categórica, es decir toma un valor entre 1 y 4 (Figura 15) dependiendo del distrito policial [14], sin embargo se pudo apreciar que existe una gran concentración de crímenes en el distrito o zona 1, el cual se conoce como el centro de la ciudad, punto en donde hay mas atractivos turísticos como también de un alto impacto comercial, por lo tanto es esperarse que esta zona de la ciudad tome la atención del modelo, esto también esta ligado a un posible patrón de comportamiento de las diferentes zonas en donde el distrito 1 tiene mas características deseables, según el modelo, para que ocurra un crimen y va descendiendo hasta la zona 4 (Figura 16).

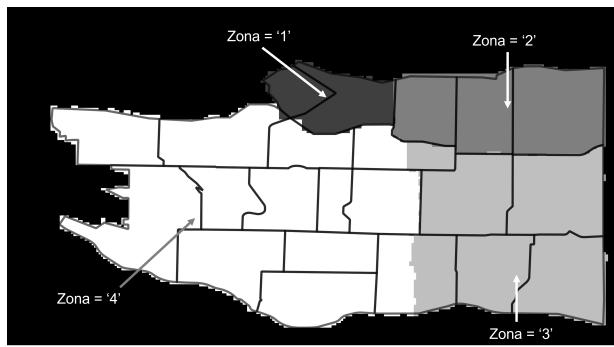


Figura 15: Distribución de Vancouver diseñada por distritos policiales

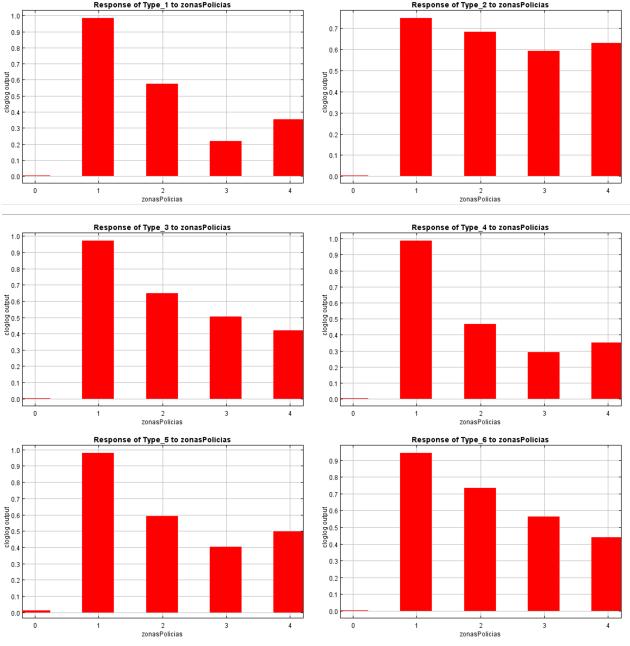


Figura 16: Valor promedio de distribución de probabilidad en Zonas Policías.

Por otra parte, analizando la variable distriEdades15_64_male, se aprecia que en barrios en donde hay más cantidad de hombres, comprendidos en la edad entre 15 y 64 años, hay una característica que puede estar asociada con la criminalidad, según el modelo (Figura 17). Concepto que puede estar asociado con lo sucedido en Canadá entre los años de 1960 y 1980, el cual menciona que el aumento entre el número de hombres entre los 15 y 25 años, está relacionado con el aumento de la criminalidad [15], sin embargo no es algo que se pueda respaldar con los resultados de este modelo, ya que se está analizando la variable desde una totalidad y no desde una proporción, además este estudio tiene un rango de edad mucho mayor. Sin embargo, se propone en futuras investigaciones analizar esta suposición por medio de este modelo, como también poder implementar un modelo más actual gracias al censo que saldrá en el año 2022.

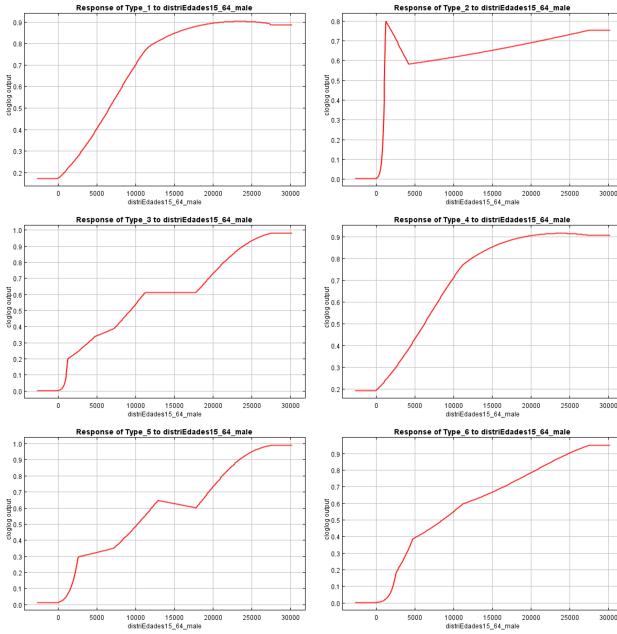


Figura 17: Valor promedio de la distribución de probabilidad con distriEdades15_64_male

8. Conclusiones

- Se obtienen las variables mas representativas para la predicción de las dinámicas de la criminalidad en la ciudad de Vancouver, esto teniendo en cuenta que el modelo no tiene la capacidad de producir una ocurrencia, ya que para esto necesitaría datos de ocurrencia/ausencia, valor que es difícil obtener ya que no se puede asegurar que todos los crímenes fueron denunciados. Sin embargo se puede predecir una distribución de probabilidad que tiene la capacidad de tener en cuenta las dinámicas o patrones que tienen los datos de entrenamiento sin sobre ajustarlos, y poder predecir donde es mas probable que ocurra un crimen.
- Gracias al uso del software Maxent, la implementación de una visualización de los resultados del algoritmo se pueden apreciar, ya sea desde los archivos de salida que entrega el mismo software o por el uso de programas libres como lo es QGIS.
- Se obtuvo un modelo final, el cual fue evaluado con varios procesos de pruebas que eran comparables entre si, con el objetivo de escoger el mejor modelo en esta implementación. Es posible seguir realizando el proceso de escoger el modelo mas exhaustivamente, sin embargo es una característica que se propone como trabajo futuro, debido a la gran carga temporal que

tiene esta labor, teniendo en cuenta que se tiene un modelo para cada mes del año y para diferentes tipos de crimen.

9. Agradecimientos

Agradezco a mis padres por apoyarme desde mi primer semestre y creer en mi. Al profesor Fernando Lozano por ayudarme a concretar la idea de este proyecto y a Juan Jose por siempre estar dispuesto a apoyar y dar su opinión.

10. Referencias

Referencias

- [1] 'Criminalidad'. [Internet] <https://www.lexico.com/es/definicion/criminalidad>
- [2] J. Rodriguez, A. Álvarez, C. Arias y E. Fernández. (2009). "La contribución a la producción: Estimación por Máxima Entropía", *Revista de economía aplicada*, [Internet]. Vol. 17, n.º 50, pp.77-96. Disponible en <https://www.redalyc.org/pdf/969/96912320004.pdf>
- [3] C. Shannon. (1948, oct). .[^] mathematical theory of communication", *The Bell System Technical Journal*, [En linea]. Vol. 27, pp. 379–423, 623–656. Disponible en <https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>
- [4] L. Llorente. "Teoria de la información Estadística", *ESTADÍSTICA ESPAÑOLA*, vol. 35, n.º 133, pp. 195 - 268, 1993
- [5] Complexity Explorer, Maximum Entropy Methods Tutorial: Review Of Max Ent. Disponible en https://www.youtube.com/watch?v=Fhr_0GgHTgc
- [6] S. Phillips, R. Anderson y R. Schapire. (2005, mzo). "Maximum entropy modeling of species geographic distributions", *ELSEVIER*, [En linea]. Disponible en <http://www.bio-nica.info/biblioteca/Phillips2006MaximumEntropy.pdf>
- [7] S. Phillips, R. Anderson y R. Schapire. (2013, mzo) Maxent software for modeling species niches and distribution. (Version 3.4.1). Disponible en https://biodiversityinformatics.amnh.org/open_source/maxent/
- [8] C. Merow, M. Smith y J. Silander. 'A practical guide to Maxent for modeling species' distributions: what it does, and why inputs and settings matter". [En linea]. Disponible en <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0587.2013.07872.x>

- [9] R. Pearson (2010), "Species' Distribution Modeling for conservation educators and practitioners", *The American Museum of Natural History*. [En linea]. Disponible en https://www.amnh.org/content/download/141368/2285424/file/LinC3_SpeciesDistModeling.pdf
- [10] *A Brief Tutorial on Maxent*, ATT Research, Disponible en https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf
- [11] https://commons.wikimedia.org/wiki/File:Stadtgliederung_Vancouver_2008.png
- [12] <https://opendata.vancouver.ca/explore/dataset/census-local-area-profiles-2016/information/>
- [13] <https://chelsa-climate.org>
- [14] Vancouver Police Department - Patrol districts. [En linea]. Disponible en <https://vpd.ca/about-the-vpd/organizations-divisions/>
- [15] Department of Justice Canada. "Predicting crime: A review of the research". p.8. [En linea]. Disponible en https://www.justice.gc.ca/eng/rp-pr/csj-sjc/jsp-sjp/rr02_7/rr02_7.pdf