Predicción de homicidios con arma de fuego utilizando redes neuronales

Omar Esteban Corona Mindiola[1] and Sergio Ivvan Valdez Peña[2]

[1]Universidad de Guadalajara, esteban.corona@alumnos.udg.mx
[2]Centro de Investigación en Geografía y Geomática Ing. Jorge L. Tamayo, A.C. (CONACYT), svaldez@centrogeo.edu.mx

13 de agosto de 2020

Resumen

En este trabajo se utiliza una red neuronal para predecir los homicidios por arma de fuego en los 50 municipios más violentos de México, usando 17 delitos de la base de datos del Secretariado Ejecutivo como predictores, encontrados mediante la correlación que tienen con el delito principal (homicidios). Se logró estimar el número de homicidios en los últimos 4 meses disponibles con un error máximo de 0.25 .

Introducción

México es el país número 37 en el índice de crímenes de 2020 [4], esto afecta seriamente la calidad general de vida de la población, como lo indica el *Better Life Index* de la OCDE donde México se ubica en la posición 38 de 40 tanto en el índice global como en la tasa de homicidios, también esta muy cerca de la última posición en sensación de seguridad al caminar por la noche, en ambos índices solo es superado [5]. La comprensión del fenomeno incluye el determinar como grupos de homicidios están ligados en la dinámica del crimen.

Por ello, una forma de hacer esto es intentar predecir los delitos de alto impacto, tomando como predictores o variables de entrada otros delitos. Utilizando algunos delitos reportados un mes antes se intenta predecir el número de homicidios por arma de fuego en los 50 municipios más violentos de acuerdo a los datos del 2019, esto es, sumando la cantidad de homicidios dolosos ocurridos en cada municipio en todo este año. Se usó una red neuronal del tipo retropropagación la cual toma varios valores de entrada en este caso los delitos de un mes anterior y regresa un único valor, correspondiente al número de homicidios que se esperarían el mes siguiente.

Método

Se utilizaron los datos brindados por el gobierno, de parte del Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública correspondientes a los datos abiertos de incidencia delictiva en el periodo de 2015-actualidad (https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva?state=published).

El primer proceso que se realiza es seleccionar columnas de los datos para quedarnos solo con las que podrían ser relevantes, es decir, el número de entidad, municipio y los delitos ocurridos en cada mes, en específico la modalidad de cada delito.
Para la primera etapa, se utilizaron los años 2018 y 2019, uno como predictor y el otro para observar la relación que podría existir entre ambos. La forma en que se compararon fue: usando todo el año 2019 y a partir de él, usar un periodo de 12 meses diferido por 1 y 2 meses. En primera instancia solo se realizó para el municipio de Guadalajara, ya que al ser uno de los más poblados del país podría sugerir como identificar cuales de los delitos están más correlacionados con los homicidios dolosos.
El primer filtro para escoger los posibles predictores fue calcular la correlación entre los homicidios cometidos con arma de fuego y los 98 delitos reportados, con un mes de desfase, es decir calculamos la correlación entre los homicidios del mes con los delitos que sucedieron un mes anterior. Seleccionamos solo los que cumplen una correlación mayor o igual a $\pm 0{,}5$, para esto se usó la función cor.test mediante los métodos "pearson y kendall"[3], todo esto aplicado a datos normalizados dividiendo entre el valor máximo de cada delito en el rango de tiempo considerado, así todos los valores estarán en el rango de [0,1].
Luego de observar que varios delitos cumplían con la condición impuesta se procedió a comparar todos los municipios y agruparlos por estados para ver las semejanzas entre estos. Para esto se hizo un histograma de frecuencias que tomaba el número de veces que un delito se correlacionaba con los homicidios en los diferentes municipios del estado. De esta manera se obtuvieron 32 histogramas de frecuencias correspondientes a cada estado, como un ejemplo de estos, se muestra en la figura (1) el estado de Jalisco.

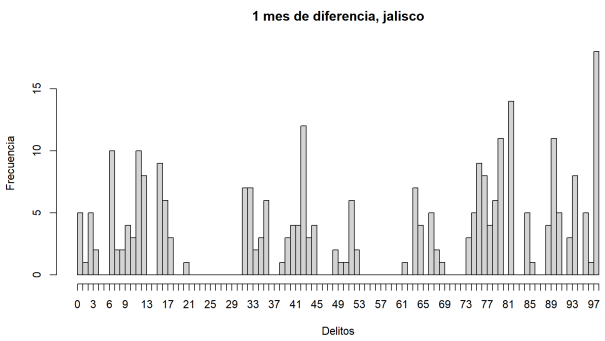

Figura 1: Histograma de frecuencias del estado de Jalisco.

Las frecuencias corresponden al número de veces en que un delito superaba la correlación mencionada anteriormente. De esta manera encontramos los delitos que consistentemente están mas correlacionados con los homicidios con arma de fuego en los diferentes estados.

Se procedió por dos métodos similares para separar los datos. El primero fue usando clustering del tipo K-means como se indica en [6] y de esta manera organizar los delitos en 5 grupos de posibles predictores, como se muestra en la Figura (2). Los valores de entrada a la función k-means corresponden al data.frame que contiene las densidades de delitos en cada estado, de esta manera el algoritmo trata de buscar las semejanzas que existen entre los 98 delitos y regresarnos de esta manera 5 subconjuntos de delitos.

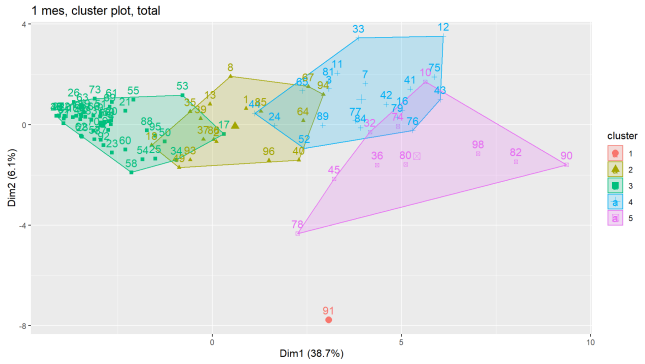

Figura 2: Grupos de posibles predictores

Podemos notar en primera instancia que en

realidad se trata de 4 grupos, sin embargo, se realizaron 5 para separar el dato anómalo que corresponde al estado de Yucatán. De estos 4 grupos, se escogió el clúster que contuviera el delito número 1, es decir, los homicidios dolosos con arma de fuego, se puede ver en la figura anterior que se trata del clúster número 2. El cual contiene 17 delitos que se enlistarán más adelante (Tabla (2)) .

Teniendo seleccionados el conjunto de delitos que se usaran como predictores,lo siguiente fue crear un data.frame que nos serviría para entrenar una red neuronal y a la vez para ponerla a prueba. El contenido de ese data.frame fueron los 54 meses que tenemos disponibles hasta la actualidad, diferidos respecto a un mes los delitos y los homicidios cometidos con arma de fuego. Se usaron los primeros 50 meses como datos de entrenamiento y los 4 restantes como prueba de nuestra red. La red neuronal fue creada con la función neuralnet, que utiliza el método de retropropagación [2] para calcular las salidas de la red.
Dado este punto, es importante mencionar el número de neuronas que se incluyeron en esta red. Para esto, se hicieron 4 pruebas, una red con 4 capas de neuronas y otra con 5, que a su vez se aplicaron a una red con una columna temporal y otro sin esta (Figuras (3) y (4)).
El agregado de esta columna es porque como la red realiza predicciones solo con valores de un mes anterior, se pierde información relacionada a tendencias a través del tiempo que podría ser relevante. Dicha columna corresponde a la enumeración de los meses usados, es decir, es una columna donde sus datos van siempre incrementando desde el 1 al 54 en este caso.
Utilizando gráficas de tipo violín para los errores de estas redes, se determinó la configuración de la red que tuviera un menor error.
En todas las gráficas se trabajó con el error absoluto medio, aplicado a los resultados normalizados.

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

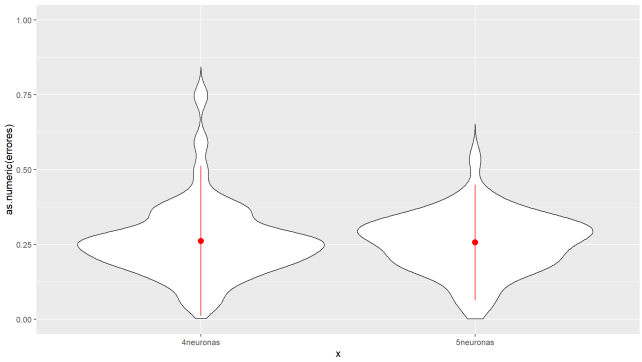

Figura 3: Errores sin columna temporal

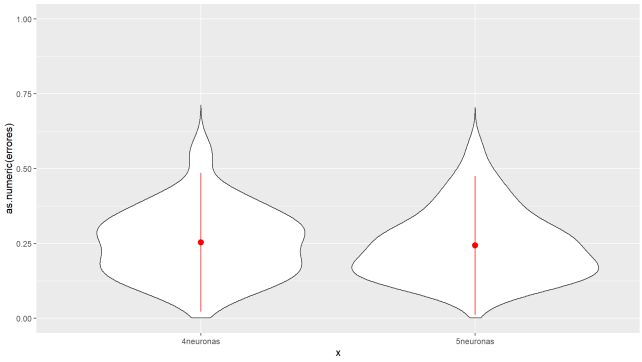

Figura 4: Errores con columna temporal

Como se puede interpretar de las gráficas anteriores, la configuración con una columna temporal y 5 capas de neuronas da mejores resultados, pues la masa de densidad en los errores es más común en valores cercanos a 0.
La manera que devolvió mejores resultados al escoger los números de neuronas en cada capa escapa de momento a este proyecto por lo que se decidió hacerlo de tal forma que fuera como el vector siguiente (34,17,9,5,1),que corresponde al doble del número de delitos utilizados y luego disminuyendo por la mitad hasta llegar a 1.

En la primera propuesta se tomo en consideración los delitos y su correlación temporal, sin

embargo estos fenómenos de crimen y violencia tienen una dimensión territorial importante, para considerarla integramos como predictores datos de municipios vecinos que pudieran compartir alguna relación. Para escoger dichos vecinos lo que se hizo fue comparar entre los 50 municipios principales y sus municipios adyacentes, la correlación entre los 17 delitos que ocurren en el municipio principal y los adyacentes, seleccionando aquellos en los que se encontraba una correlación más alta, siempre mayor a 0.7. Obteniendo así la tabla de adyacencias del apéndice. Teniendo las adyacencias, lo único que se hizo fue agregar los delitos de los municipios vecinos a nuestros datos de entrenamiento en una nueva red y comparar mediante gráficas de tipo violín, si las predicciones eran mejores o no. Por ello algunos resultados de varios municipios fueron tomados de sus datos individuales.
Se selecciona la red que entrega el mejor resultado después de haberla entrenado 10 veces, esto porque decidimos trabajar con pesos iniciales generados aleatoriamente.

Figura 5: Homicidios reales por arma de fuego,marzo 2020

Figura 6: Predicciones de homicidios por arma de fuego,marzo 2020

Resultados

Se realizó una representación visual en forma de mapa de los resultados obtenidos,esto para buscar relaciones espaciales que nos indicaran posibles indicadores en el fallo de nuestro modelo. Los datos con los que se generaron dichos mapas se encuentran anexados en el apéndice B.

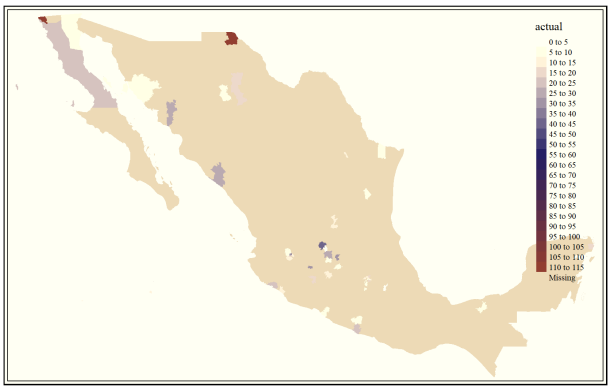

Figura 7: Homicidios reales por arma de fuego,abril 2020

Figura 8: Predicciones de homicidios por arma de fuego,abril 2020

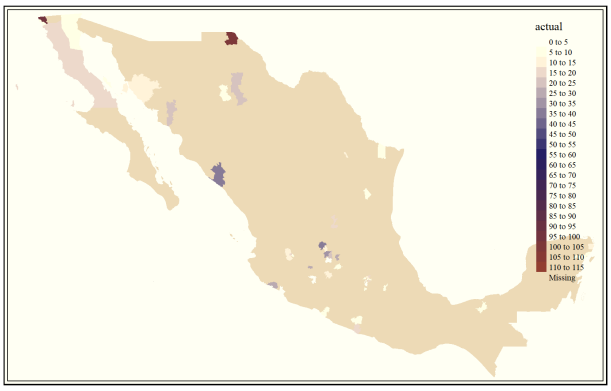

Figura 11: Homicidios reales por arma de fuego,junio 2020

Figura 9: Homicidios reales por arma de fuego,mayo 2020

Figura 12: Predicciones de homicidios por arma de fuego,junio 2020

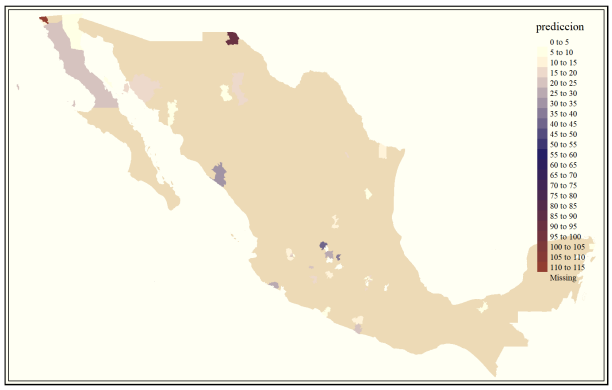

Figura 10: Predicciones de homicidios por arma de fuego,mayo 2020

Discusiones

Según los datos que se muestran en el apéndice C, podemos ver que se ha logrado una buena predicción,pues en la mayoría, los errores no superan el valor del 0.25, y en los que si lo hacen, existen varios factores que causan esos valores atípicos, los cuales discutiremos a continuación.
Debemos mencionar que no todos los resultados presentados se obtuvieron usando datos de municipios adyacentes, pues algunos de ellos no tuvieron la suficiente correlación para tomarlos en cuenta, y en otros casos, sus resultados no mejoraban al utilizarlos.
Municipios que no se incluyeron vecinos por falta de correlación: Juárez, Acapulco de Juárez,

León, Cajeme, Chihuahua, Uruapan, Hermosillo, Reynosa, Centro, Celaya y Cuauhtémoc.
Municipios en los que se obtuvieron mejores resultados usando solo datos propios: Tijuana, Manzanillo, Álvaro Obregón,Salamanca,Guadalajara,Chimalhuacán y Solidaridad.
Uno de los principales factores que hay que tomar en cuenta, es que los meses que estamos prediciendo se encuentran dentro del periodo de tiempo de la pandemia ocasionada por el virus COVID-19, por lo que en algunos municipios esto causó un decaimiento en el número de homicidios, sin embargo, los delitos que usamos para predecir quizá se mantuvieron constantes, arrojando así valores de predicción más elevados de los que realmente sucedieron, un ejemplo de esto es el municipio de Uruapan, que pasó de tener 18-15 homicidios a solo 4 en el último mes.
Podemos realizar una comparación entre el municipio que tiene un error más bajo y el que tiene como resultado un mayor error, dichos municipios son León con un error de 0.040 y Ensenada con un error de 0.253, sin embargo, el error tan grande por parte de Ensenada se debe a que la predicción del primer mes es bastante mala, en comparación con los meses siguientes, por lo que en su lugar podríamos poner al municipio de Benito Juárez, en el cual los resultados si son muy distintos a los reales. Una primera diferencia entre estos dos municipios es la cantidad de población entre ambos, por lo que se podría inferir que entre más elevado sea el número de habitantes se obtienen mejores resultados, no obstante, Manzanillo tiene una población muy inferior a la de León y las predicciones que se hicieron en dicho estado son bastante aceptables, por lo que el factor de población queda descartado.
Analizando los datos de cada municipio nos damos cuenta que en los municipios donde existe un error más grande suelen haber valores o tendencias muy raras en sus delitos, un ejemplo de esto es Benito Juárez, que pasa de tener menos de 100 robos a negocios con violencia a casi 400, este cambio se produce de un mes a otro alterando de esta manera la tendencia. De este tipo de variaciones podemos encontrar también en otros municipios, dejando esto como una variable a considerar a la hora de validar nuestro modelo. Se desconoce si esto se debe a errores en la captura de datos o si en realidad así es la varianza de delitos en nuestro país.
Los códigos y archivos utilizados se encuentran en el repositorio [1].

Conclusiones

La red neuronal que contiene una columna temporal funciona mejor al momento de realizar predicciones. De igual forma, agregar datos de municipios vecinos, disminuyó (en su mayoría) el error absoluto medio de los resultados.
Se encontraron 17 delitos que sirven muy bien en algunos de estos municipios para predecir los homicidios con arma de fuego. Dichos delitos pueden servir como una alarma para las autoridades y reducir de esta forma la violencia en estos municipios.
Siendo esta la primera aproximación que se realiza por nuestra parte para intentar predecir homicidios en el país, encontramos resultados bastante satisfactorios. Sin embargo, estamos conscientes de que la metodología utilizada se puede mejorar, por ello, se tiene pensado en un futuro trabajo, donde se cambie la manera en que se seleccionen los posibles predictores, usando métodos más especializados y consistentes que las pruebas de correlación.

Apéndice A. Nombres de municipios y delitos

*	Clave del municipio	Nombre del municipio	*	Clave del municipio	Nombre del municipio
19	02001	Ensenada	6	14039	Guadalajara
39	02002	Mexicali	16	14097	Tlajomulco de Zúñiga
1	02004	Tijuana	12	14098	San Pedro Tlaquepaque
35	02005	Playas de Rosarito	29	14101	Tonalá
17	06007	Manzanillo	18	14120	Zapopan
38	06009	Tecomán	50	15031	Chimalhuacán
40	08017	Cuauhtémoc	13	15033	Ecatepec de Morelos
11	08019	Chihuahua	34	15057	Naucalpan de Juárez
2	08037	Juárez	30	15058	Nezahualcóyotl
23	09005	Gustavo A. Madero	37	15104	Tlalnepantla de Baz
9	09007	Iztapalapa	15	16053	Morelia
42	09010	Álvaro Obregón	20	16102	Uruapan
48	09012	Tlalpan	33	16108	Zamora
46	09015	Cuauhtémoc	32	17007	Cuernavaca
49	09017	Venustiano Carranza	22	19039	Monterrey
47	11002	Acámbaro	26	21114	Puebla
27	11007	Celaya	7	23005	Benito Juárez
8	11017	Irapuato	28	23008	Solidaridad
4	11020	León	31	24028	San Luis Potosí
14	11027	Salamanca	5	25006	Culiacán
41	11037	Silao de la Victoria	10	26018	Cajeme
43	11042	Valle de Santiago	21	26030	Hermosillo
3	12001	Acapulco de Juárez	25	27004	Centro
36	12029	Chilpancingo de los Bravo	24	28032	Reynosa
44	12038	Zihuatanejo de Azueta	45	28041	Victoria

Tabla 1: Nombre de los 50 municipios utilizados en la red. La columna (*) corresponde al orden en la lista de municipios más violentos.

Número de delito	**Nombre de delito**	**Media**	**Máximo**	**Mínimo**	**Varianza**
1	Homicidio doloso con arma de fuego	0.2673	0.7567	0.0188	0.0324
2	Homicidio doloso con arma blanca	0.2396	0.6694	0.0098	0.0262
8	Homicidio culposo con otro elemento	0.2623	0.5762	0.0058	0.0261
13	Lesiones dolosas no especificado	0.309	0.7827	0.0356	0.0505
18	Lesiones culposas no especificado	0.2859	0.6862	0.0036	0.0376
35	Hostigamiento sexual	0.2617	0.7266	0.0026	0.0373
37	Violación equiparada	0.2215	0.6552	0.0049	0.0286
39	Otros delitos que atentan contra la libertad y la seguridad sexual	0.2379	0.7007	0.0247	0.0309
40	Robo a casa habitación con violencia	0.2682	0.6623	0	0.0283
49	Robo de autopartes sin violencia	0.2523	0.6794	0.0121	0.0318
64	Robo a negocio con violencia	0.2742	0.7567	0	0.0361
67	Robo de ganado sin violencia	0.2094	0.7038	0.0021	0.0321
85	Otros delitos contra la familia	0.2821	0.6846	0.0099	0.0278
86	Corrupción de menores	0.2319	0.6068	0	0.0301
93	Falsedad	0.2612	0.7915	0.0143	0.0384
94	Falsificación	0.2861	0.8303	0	0.0396
96	Delitos cometidos por servidores públicos	0.2675	0.6756	0.0072	0.0386

Tabla 2: Número y nombre de delitos utilizados para predecir junto con sus valores de correlación en los 50 municipios.

Apéndice B. Tablas de resultados

Clave del municipio	Predicción	Real	Clave del municipio	Predicción	Real
2001	10	33	14039	18	23
2002	9	10	14097	10	8
2004	101	114	14098	22	21
2005	8	10	14101	4	13
6007	24	26	14120	12	11
6009	6	3	15031	4	9
8017	8	7	15033	21	23
8019	26	18	15057	11	13
8037	85	108	15058	10	8
9005	18	10	15104	8	11
9007	12	16	16053	23	20
9010	5	3	16102	15	15
9012	8	8	16108	26	36
9015	7	6	17007	14	15
9017	6	10	19039	19	20
11002	6	4	21114	4	5
11007	40	18	23005	7	31
11017	30	28	23008	1	6
11020	35	32	24028	11	10
11027	26	22	25006	31	34
11037	5	5	26018	27	26
11042	4	2	26030	21	18
12001	20	25	27004	14	15
12029	12	13	28032	11	4
12038	7	7	28041	12	10

Tabla 3: Comparaciones del mes de marzo, 2020.

Clave del municipio	**Predicción**	**Real**	**Clave del municipio**	**Predicción**	**Real**
2001	23	23	14039	21	33
2002	8	5	14097	11	10
2004	109	129	14098	23	19
2005	9	20	14101	11	12
6007	20	20	14120	9	7
6009	8	10	15031	9	11
8017	5	8	15033	19	17
8019	11	15	15057	7	10
8037	102	114	15058	14	8
9005	14	12	15104	6	14
9007	13	16	16053	6	12
9010	5	5	16102	18	18
9012	11	7	16108	22	32
9015	8	8	17007	6	9
9017	4	4	19039	15	14
11002	6	6	21114	5	4
11007	34	34	23005	8	15
11017	31	27	23008	6	4
11020	41	42	24028	10	11
11027	21	26	25006	25	25
11037	1	0	26018	25	29
11042	4	7	26030	10	9
12001	20	24	27004	13	9
12029	7	9	28032	13	5
12038	14	7	28041	8	5

Tabla 4: Comparaciones del mes de abril, 2020.

Clave del municipio	**Predicción**	**Real**	**Clave del municipio**	**Predicción**	**Real**
2001	21	26	14039	17	18
2002	7	7	14097	10	9
2004	116	121	14098	23	17
2005	10	15	14101	9	9
6007	26	28	14120	10	10
6009	4	4	15031	8	7
8017	6	9	15033	19	20
8019	16	21	15057	8	6
8037	90	76	15058	8	6
9005	10	12	15104	7	8
9007	16	15	16053	10	16
9010	4	4	16102	15	9
9012	4	6	16108	23	27
9015	7	7	17007	15	7
9017	6	5	19039	15	13
11002	2	7	21114	3	6
11007	38	36	23005	7	14
11017	29	37	23008	1	2
11020	41	43	24028	12	21
11027	26	27	25006	32	27
11037	2	2	26018	9	34
11042	4	4	26030	16	9
12001	20	26	27004	7	6
12029	10	6	28032	13	14
12038	7	8	28041	7	3

Tabla 5: Comparaciones del mes de mayo, 2020.

Clave del municipio	**Predicción**	**Real**	**Clave del municipio**	**Predicción**	**Real**
2001	20	15	14039	19	17
2002	8	8	14097	11	14
2004	102	96	14098	14	13
2005	9	8	14101	9	8
6007	27	27	14120	10	14
6009	5	2	15031	6	5
8017	7	9	15033	15	15
8019	18	22	15057	4	6
8037	102	104	15058	7	7
9005	9	13	15104	11	11
9007	21	16	16053	14	14
9010	4	4	16102	19	4
9012	1	3	16108	22	25
9015	1	4	17007	8	5
9017	4	7	19039	9	12
11002	10	8	21114	6	8
11007	37	29	23005	13	10
11017	41	33	23008	7	6
11020	39	38	24028	18	18
11027	26	21	25006	30	37
11037	4	10	26018	13	22
11042	4	1	26030	12	11
12001	22	15	27004	6	5
12029	8	5	28032	13	9
12038	10	6	28041	6	9

Tabla 6: Comparación del mes junio, 2020.

Apéndice C. Tabla de errores

Clave del municipio	Error Absoluto Medio	Clave del municipio	Error Absoluto Medio
2001	0.2527917	14039	0.1238955
2002	0.1054227	14097	0.08861209
2004	0.06782881	14098	0.09465734
2005	0.23866	14101	0.1458915
6007	0.04265345	14120	0.08400515
6009	0.08588397	15031	0.1872655
8017	0.1798971	15033	0.04610335
8019	0.1384704	15057	0.1545306
8037	0.1124643	15058	0.1290723
9005	0.151468	15104	0.1615324
9007	0.1097735	16053	0.1055735
9010	0.05839217	16102	0.1789004
9012	0.1480095	16108	0.188397
9015	0.07049553	17007	0.2229985
9017	0.1666996	19039	0.07284502
11002	0.1333995	21114	0.1096199
11007	0.1968515	23005	0.2123625
11017	0.1317144	23008	0.1513633
11020	0.04033695	24028	0.1383799
11027	0.09625547	25006	0.05635187
11037	0.09261821	26018	0.2113732
11042	0.1676985	26030	0.1074874
12001	0.07960405	27004	0.08634068
12029	0.1140232	28032	0.2192814
12038	0.1896772	28041	0.1046551

Tabla 7: Error absoluto medio correspondiente a los cuatro meses de cada municipio

Referencias

[1] O. Corona, *Verano_delfin*. https://github.com/iOmaR-Corona/Verano_Delfin, 2020.

[2] S. Fritsch, F. Guenther, and M. N. Wright, *neuralnet: Training of neural networks*, 2019.
R package version 1.44.2.

[3] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*, vol. 751, John Wiley & Sons, 2013.

[4] NUMBEO, *Crime index by country 2020 mid-year*. https://www.numbeo.com/crime/rankings_by_country.jsp, June 2020.

Online; recuperado el 10-Agosto-2020.

[5] OCDE, *Oecd better life index*.
http://www.oecdbetterlifeindex.org/es/topics/safety-es/, 2014.

[6] U. of Cincinati, *K-means cluster analysis*.
https://uc-r.github.io/kmeans_clustering#elbow, 2017.
Online; recuperado el 28-Julio-2020.