Series Temporales

Tercer bimestre 2020

Daniel Caicedo
UAH
djcc710@gmail.com

Ignacio Chiapella
FCEyN
ignacio.chiapella@gmail.com

Miguel Guerrero
CUFM
miguelgh72@gmail.com

Juan Knebel
FCEyN
juanknebel@gmail.com

Contents

Introducción

En el presente trabajo se analizará una serie de tiempo que contiene los delitos cometidos en la Ciudad de Buenos Aires durante los años 2016 a 2019 inclusive. Los datos fueron obtenidos de la plataforma de datos abiertos de Buenos Aires 2020 la cual colecta los datos informados por el Ministerio de Justicia y Seguridad y la Policía de la Ciudad de Buenos Aires.

Utilizando los datos previamente mencionados y las herramientas estudiadas durante el curso, se propusieron las siguientes lineas de trabajo para el estudio:

- **Análisis exploratorio**: Sobre la serie completa a modo de poder entender tendencia, estacionalidad y distribución de la mismas.
- **Análisis comparativo** de los distintos métodos de forecast para todos los tipos de delitos. Utilizaremos el periodo $2016 - 2018$ para entrenar y se realizará una predicción sobre el año 2019 para con este ultimo poder evaluar la eficacia del modelo.

Composición de los datos

A continuación describimos brevemente como está compuesto el conjunto de datos con el cuál trabajaremos a lo largo del proyecto.

Columna	Descripción
id	Identificador único de registro
fecha	Fecha cuando se produjo el delito
franja horaria	Sección del día cuando tuvo lugar
tipo delito	Tipificación del delito
subtipo delito	Sub tipificación si la hubiere
cantidad registrada	Contador
comuna	Comuna donde tuvo lugar
barrio	Barrio donde tuvo lugar
lat	Latitud de la coordenada
long	Longitud de la coordenada

El conjunto de datos cuenta con los registros acumulados por día y hora de los crímenes cometidos y denunciado en los años 2016, 2017, 2018 y 2019. Se cuenta con un total de aproximadamente 500000 registros. Además de los datos provistos, se enriqueció con los feriados nacionales de cada año y el inicio y fin del período vacacional invernal.

Herramientas utilizadas

El presente trabajo se realizó utilizando como plataforma de desarrollo **Python** en su versión 3.7 y las siguientes librerias: pandas, numpy, math, scipy.fft, scipy, shapiro, kstest, interpolate, statsmodels, fbprophet, torch, sklearn, matplotlib, seaborn.

Desarrollo

Análisis Exploratorio

En un primer momento y motivados por la curiosidad se analizó la distribución de los crímenes reportados por cada día de la semana y comprobar si efectivamente existe alguna diferencia sustancial.

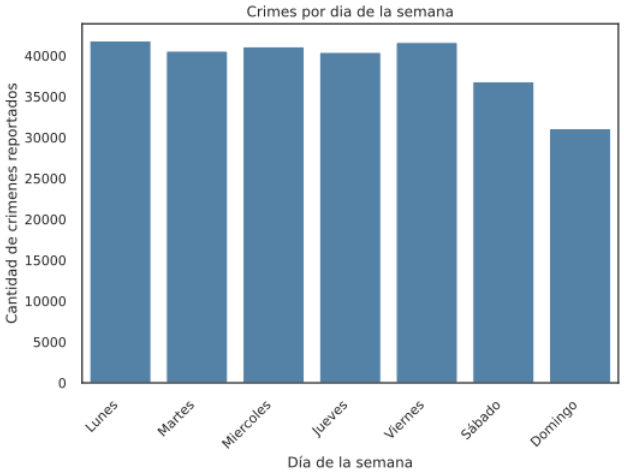


Figure 1. Crímenes por día

Se evidencia claramente que, los fines de semana hay menor cantidad de casos reportados y en particular el Domingo es el día con menos crímenes reportados. El resto de los días de la semana presentan un comportamiento similar.

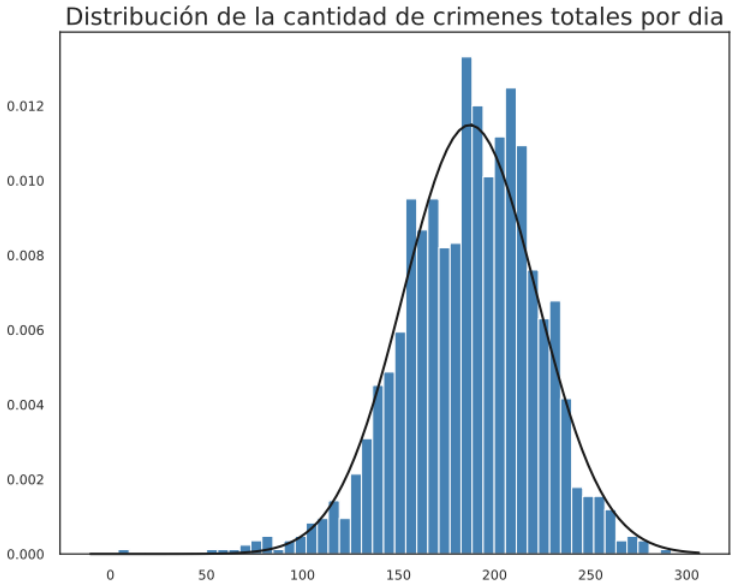


Figure 2. Distribución de crímenes por día

En un segundo estudio se buscó entender si los datos seguían alguna distribución: En el segundo análisis, el gráfico de barras de la distribución de los crímenes, contrastado con una distribución normal parece

elocuente. Para poder afirmar que estos datos siguen o no una distribución normal es que se propuso realizar dos test de normalidad:

Metodo	p-value
Shapiro test	≈ 0
Kolmogorov-Smirnov	≈ 0

Los supuestos de normalidad quedan cubiertos, debido al bajo p-value arrojado por ambos métodos.

Tendencia

Para iniciar el estudio y mejorar la comprensión de la tendencia se separó el periodo por año, y en cada estudio buscaremos identificar feriados y vacaciones de invierno en cada periodo.

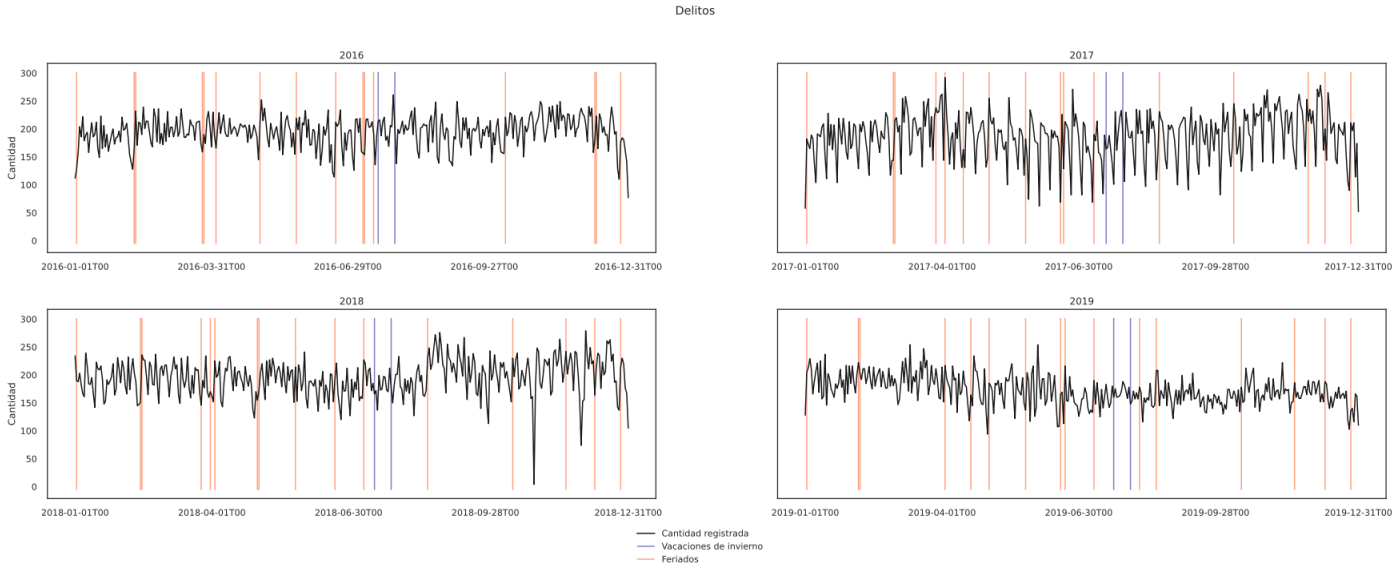


Figure 3. Crímenes por año

Del gráfico anterior se observa una similitud en el comportamiento de la curva en todos los finales de año, en especial en los del años 2016, 2017, 2018 y no tanto en 2019. Por otro lado se identificó siempre una suba de casos para el periodo comprendido en las vacaciones de invierno.

A continuación se realizó una descomposición clásica que compre solo a los años 2016 y 2018, que son los que se utilizarán para generar las predicciones futuras.

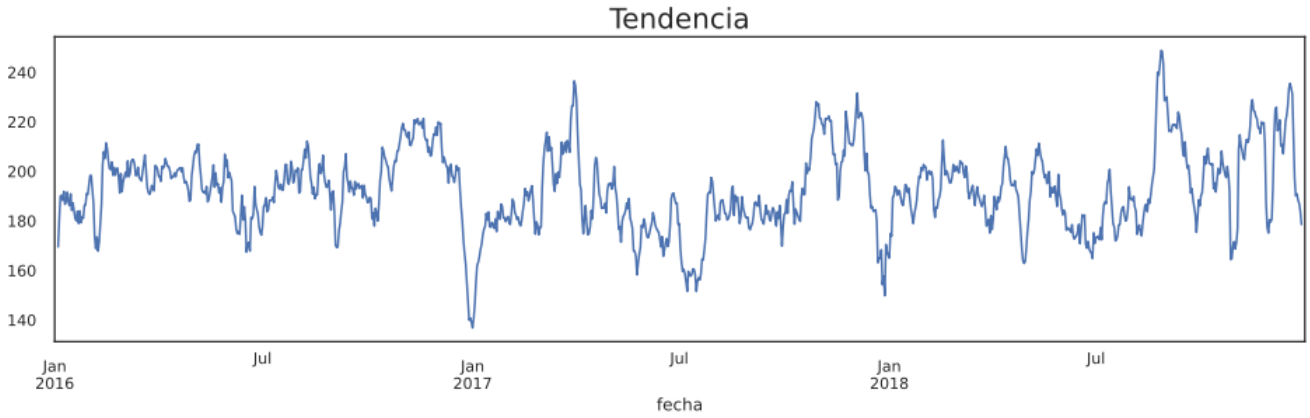


Figure 4. Tendencia

De este estudio se puede mencionar que hay un claro aumento de casos hacia fin de año, coincidiendo con las fechas de las fiestas y una caída significativa hacia comienzos de año, probablemente debido a que gran parte de los delincuentes que realizan sus actividades en la Ciudad de Buenos Aires mueven sus operaciones a destinos turísticos de la Argentina. Por último podemos destacar que no se observa ninguna tendencia y si una clara estacionalidad en la serie elegida.

A continuación se muestra un pequeño estudio que se realizó sobre los residuos y su distribución así como también los test de **Dickey-Fuller** y **KPSS** para evaluar la estacionalidad antes mencionada.

En los gráficos se puede apreciar de manera gráfica que los residuos se distribuyen de manera normal, también se observa que se distribuyen bastante cerca del 0 a lo largo de toda la serie, salvo unos pocos puntos con un comportamiento aleatorio al final de la misma. Los **p-value** de ambos tests refuerzan la hipótesis de la estacionalidad de la serie.

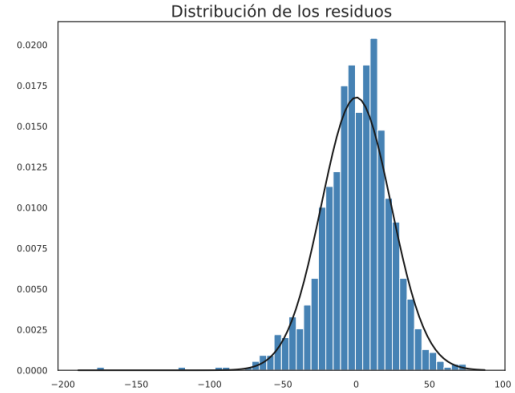


Figure 5. Distribución

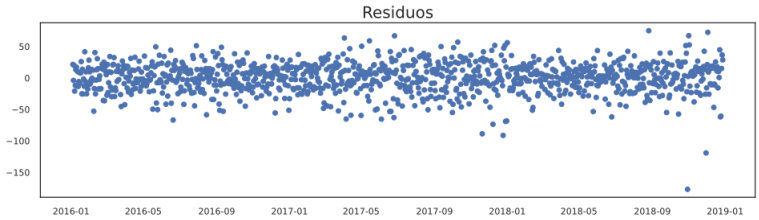


Figure 6. Residuos

	Dickey-Fuller	KPSS
p-value	≈ 0	= 0.1
lags	28	28

Para completar el análisis se decidió realizar una descomposición de Fourier la cual indica una estacionalidad en el dominio de la frecuencia 0.15 que en su representación en días representa correspondería a 7 días.

En cuanto a la autocorrelación, se puede ver que la serie esta autocorrelacionada y tambien se observa un primer pico pico en el día 1 contra si mismo, y luego picos semanales o cada 7 días, coincidiendo con los días 7, 14, 21 y 28 del mes.

En base a todo lo anterior nos lleva a concluir que los crímenes informados día por día tiene una estacionalidad de 7 días.

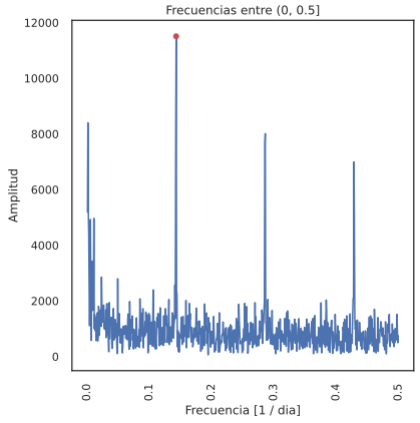


Figure 7. Descomposición de Fourier

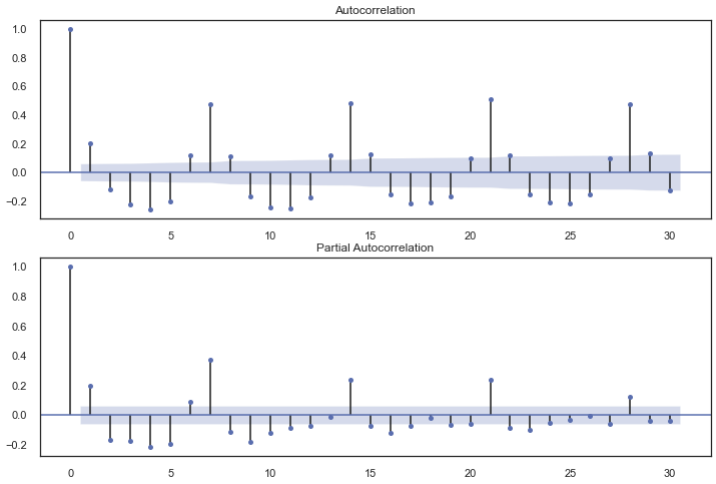


Figure 8. Autocorrelación

Prediciendo crímenes

En la siguiente sección se analizarán distintos métodos de predicción (o forecast) y luego realizaremos una comparación entre los mismos.

ARIMA (7,0,1) vs (0,0,0)

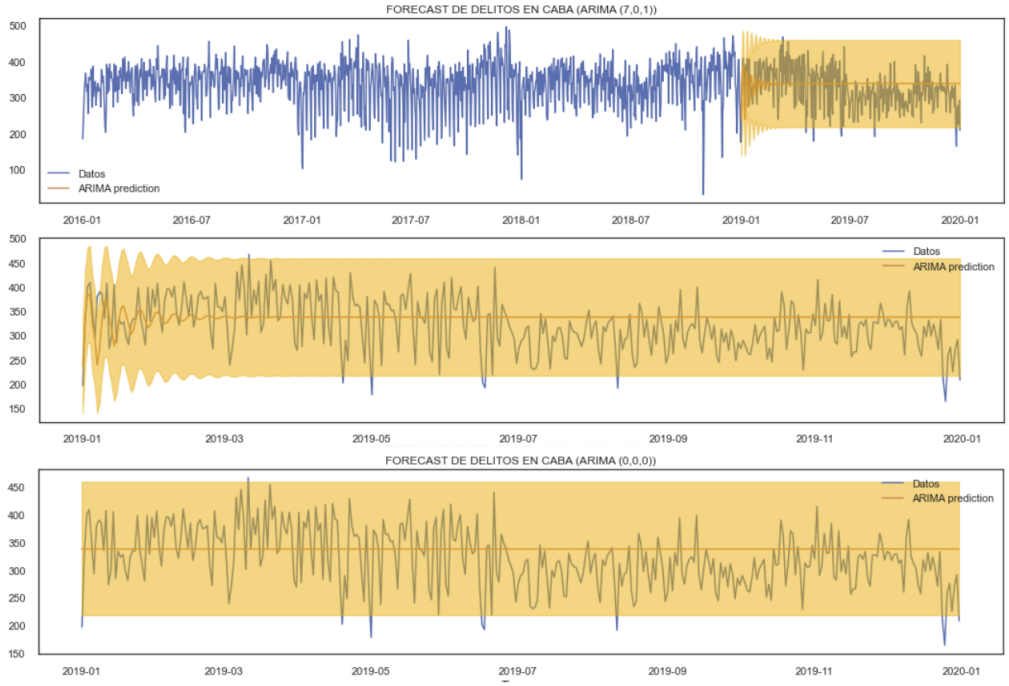
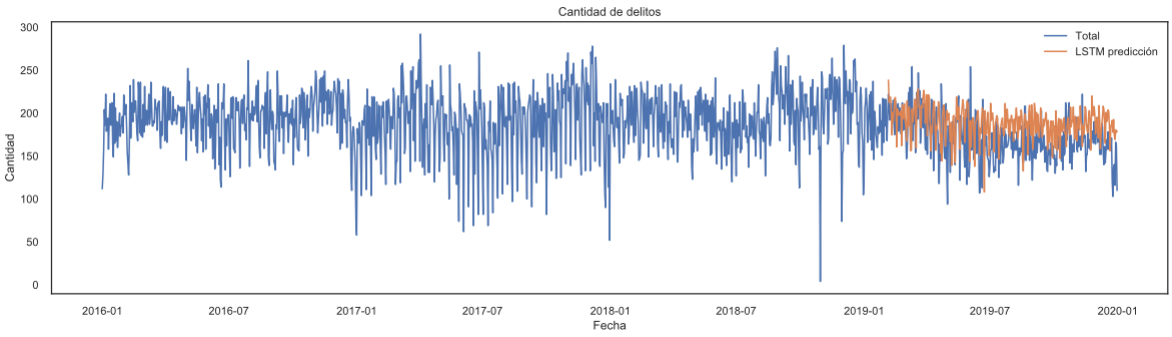


Figure 9. Arima

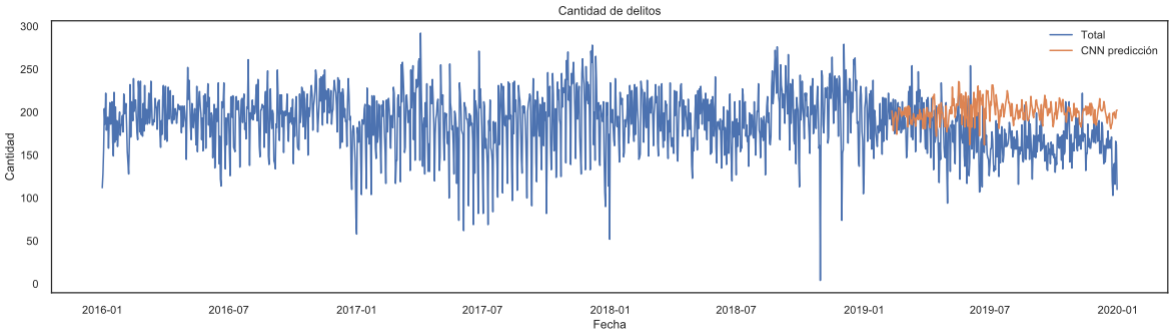
Se realizaron 2 experimentos con ARIMA con hiperparámetros distintos: 7,0,1 vs 0,0,0. El experimento con los hiperparámetros ajustados (7,0,1, primeros dos gráficos de la figura 9) predice mejor el principio de la serie, pero luego predice la media. Mientras que el experimento donde no sea ajustaron los hiperparámetros parece predecir siempre la media. Con estos experimentos se podría decir que, en primer lugar, los hiperparámetros seleccionados mejoran los resultados del análisis y en segundo lugar que ARIMA no funcionó del todo bien para predecir un año completo de nuestra serie, para trabajos a futuro se podría probar haciendo cortes mas chicos de entrenamiento y predicción.

Redes Neuronales

Se realizaron pruebas utilizando técnicas de deep learning con una configuración muy básica sin entrar en mucho detalle, solo para tener una priemera aproximación a dichas técnicas. Se usaron dos arquitecturas distintas **CNN** (Convolutional Neural Networks) y **LSTM** (Long Short Term Memory). La decisión del uso de una configuración simple se realizó teniendo en cuenta que los modelos se ejecutaron en máquinas de escritorio de uso personal.



(a) LSTM

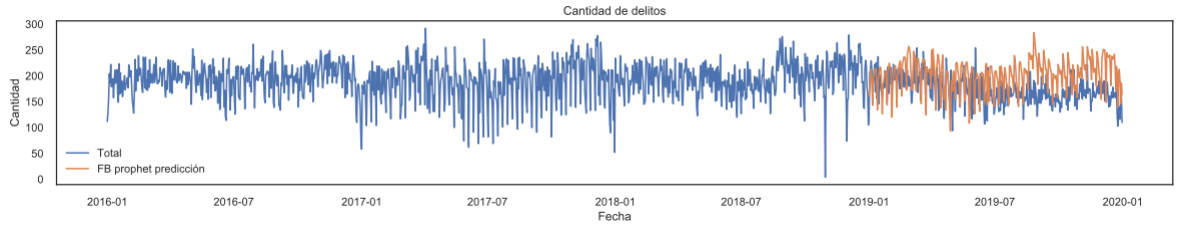


(b) CNN

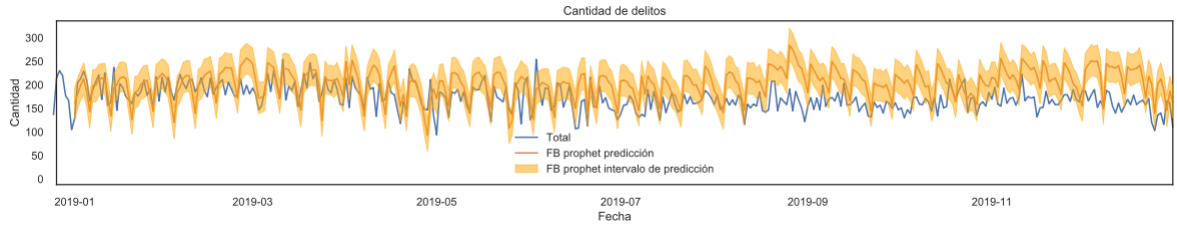
Figure 10. Deep Learning

Prophet

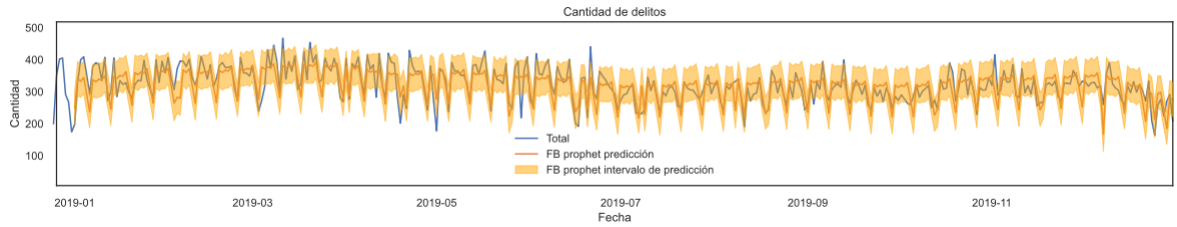
A continuación de mostraran las predicciones utilizando Prophet:



(a) Período 2016 - 2019, con estacionalidad



(b) Período 2019, con estacionalidad



(c) Período 2019, sin estacionalidad

Figure 11. Prophet

En el primer gráfico se puede ver la predicción día por día en comparación a los datos reales. En el segundo y tercer gráfico se muestran para el año de la predicción los intervalos de confianza que se generaron para cada día. El gráfico **b** se generó teniendo en cuenta una estacionalidad de 7 días, mientras que el **c** no se tuvo en cuenta ninguna.

Se puede apreciar que para la primer parte del año ambas predicciones parecieran aproximar bien, pero hacia la segunda parte del año no tener en cuenta la estacionalidad parece funcionar mejor. A pesar de esta última observación, los intervalos de confianza respetan más las altas y subidas de los crímenes en la predicción en la cual se tuvo en cuenta la estacionalidad, capturando de esta manera el comportamiento.

Comparación

Para realizar las comparaciones entre los métodos utilizados, se utilizó el error cuadrático medio, el cual realiza una comparación entre el valor real y el estimado, con el cuál obtuvimos los siguientes resultados:

Método	ECM
LSTM	2479
CNN	3283
ARIMA	3119
PROPHET	**2035**

Se puede apreciar es que si bien, **Prophet** obtiene el mejor score basado en **ECM** llama mucho la atención la buen performance obtenida en la utilización de **LSTM**, sobre todo teniendo en cuenta que se trata de una implementación muy sencilla y en dónde hay mucho camino para mejorar.

Conclusiones

- En el estudio de una serie de tiempo, queda evidenciado la importancia de conocer el dominio del problema, en este trabajo fue fundamental conocer distintas fechas importantes así como periodos particulares a modo de poder enriquecer el análisis.
- La mejor predicción mediante el la medición de ECM la dio Prophet, consideramos que la manera de medir ese error no seria la mas adecuado y se podria pensar en algo mas ajustado a los datos. Por ejemplo utilizar slots de delitos diarios diferenciados en baja cantidad de delitos, media y alta e ver si el método predijo adecuadamente en base a esa separaciones.
- También se puede pensar que si bien estudiar todos los delitos juntos da un paneo general, como un próximo paso seria casi inmediato pensar en una separación del estudio por tipo de delito para no mezclar.
- Resulto sumamente interesante el estudio con redes neuronales que si bien fue incipiente, permitió rápidamente explorar estos métodos en dos variantes y hacer una aplicación de los temas vistos en la materia.

References

Buenos Aires, C. de (2020). “Portal de Datos Abiertos,” (available online at https://data.buenosaires.gob.ar/dataset/delitos;).