Universidad Tecnológica Nacional - Buenos Aires, Argentina
Noviembre, 2020
Ciencia de Datos - Cluster AI

Cita bibliográfica: *Demaestri, J. I., & Gutierrez, M. (Noviembre de 2020). Modelo de Regresión para la predicción de la evolución del COVID-19 en la Ciudad de Buenos Aires a partir de factores ambientales y demográficos. Ciudad de Buenos Aires, Argentina: Cluster AI, Universidad Tecnológica Nacional.*

Modelo de Regresión para la predicción de la evolución del COVID-19 en la Ciudad de Buenos Aires a partir de factores climáticos y demográficos.

Demaestri, Juan Ignacio
Gutierrez, Manuel

Abstract—

El 11 de marzo de 2020, la Organización Mundial de la Salud (OMS) declaró pandemia al brote de la enfermedad COVID-19, producida por el nuevo coronavirus SARS CoV-2. La transmisión del SARS-CoV-2 se produce mediante pequeñas gotas expulsadas al hablar, estornudar, o toser, que al ser despedidas por un portador pasan directamente a otra persona mediante la inhalación, o quedan sobre los objetos y superficies que rodean al emisor, y luego, a través de las manos, que lo recogen del ambiente contaminado, toman contacto con las membranas mucosas orales, nasales y oculares, al tocarse la boca, la nariz o los ojos (Xian Peng, Xin Xu, Yuqing Li, Lei Cheng, Xuedong Zhou, & Biao Ren, 2020)**. Esta última es la principal vía de propagación, ya que el virus puede permanecer viable hasta por días en todo tipo de lugares.**

Debido a la alta contagiosidad de este virus y su rápida propagación en las grandes urbes del Mundo se ha establecido la hipótesis de que los factores demográficos y sociales, influidos también por los factores climatológicos, tienen alta incidencia en la aceleración de contagios.

En el presente trabajo se aborda la hipótesis antes mencionada, que identifica a los factores climáticos y demográficos como variables que puedan llegar a favorecer la transmisión del virus COVID-19, estudiando para ello el caso en la Ciudad de Buenos Aires. A través de herramientas de Machine Learning, se busca encontrar un modelo de regresión, empleando datos gubernamentales del clima, transporte urbano y densidad poblacional, para intentar predecir la evolución de contagios por COVID-19.

On March 11, 2020, the World Health Organization (WHO) declared the outbreak of the COVID-19 disease, caused by the new SARS CoV-2 coronavirus, a pandemic. The transmission of SARS-CoV-2 occurs through small droplets expelled when speaking, sneezing, or coughing, which when released by a carrier pass directly to another person through inhalation, or remain on objects and surfaces that surround the emitter , and then, through their hands, which pick it up from the contaminated environment, they come into contact with the oral, nasal and ocular mucous membranes, by touching their mouth, nose or eyes (Xian Peng, Xin Xu, Yuqing Li, Lei Cheng, Xuedong Zhou, & Biao Ren, 2020)**. The latter is the main route of spread, since the virus can remain viable for up to days in all kinds of places.**

Due to the high contagion of this virus and its rapid spread in large cities of the world, the hypothesis has been established that demographic and social factors, also influenced by weather factors, have a high incidence in the acceleration of infections.

In this paper, the aforementioned hypothesis is addressed, which identifies climatic and demographic factors as variables that may favor the transmission of the COVID-19 virus, studying the case in the City of Buenos Aires. Through Machine Learning tools, it is sought to find a regression model, using government data on the climate, urban transport and population density, to try to predict the evolution of COVID-19 infections.

I. Introducción

Los coronavirus son una extensa familia de virus que pueden causar enfermedades tanto en animales como en humanos. En los humanos, se sabe que varios coronavirus causan infecciones respiratorias que pueden ir desde el resfriado común hasta enfermedades más graves como el síndrome respiratorio de Oriente Medio (MERS) y el síndrome respiratorio agudo severo (SARS). El virus SARS COV-2 produce síntomas similares a los de la gripe, entre los que se incluyen fiebre, tos, disnea, mialgia y fatiga. También se ha observado la pérdida súbita del olfato y el gusto (sin que la mucosidad fuese la causa). En casos graves se caracteriza por producir neumonía, síndrome de dificultad respiratoria aguda, sepsis y choque séptico que conduce a alrededor del 3 % de los infectados a la muerte, aunque la tasa de mortalidad se encuentra en 4,48 % y sigue ascendiendo (Pérez Abreu, Gómez Tejeda, & Dieguez Guach, 2020). Además, el COVID-19 cambió completamente la forma de vida de las personas, se modificaron los hábitos, las relaciones personales, gran parte a causa del distanciamiento social y llevó a los científicos a una carrera por encontrar la forma de vencerlo para ponerle fin a la pandemia. El virus no distingue entre países desarrollados y no desarrollados y llevó a pelear a través del sistema de salud en algunos casos hasta el colapso.

Para poder vencerlo se debe entender la forma en que se comporta el mismo, y además se debe entener que hay dos clases de virus: envueltos y desnudos. Los virus envueltos tienen una membrana lipídica adicional llamada envoltura que rodea la cápside proteica. La envoltura contiene fosfolípidos y proteínas derivadas de las membranas de la célula huésped. Los virus envueltos adquieren este sobre durante la replicación viral y la liberación. El VIH, el VHS, el VHB y el virus de la gripe son varios ejemplos de virus envueltos. Los desnudos son más resistentes, al frío, al calor y a las radiaciones. El SARS COV-2 es un virus envuelto. Los especialistas han investigado cómo es afectado el virus en ambientes de humedad, presión atmosférica, frío y calor. El transporte de partículas influirá en la propagación del coronavirus y determinará la implementación de pautas sobre distanciamiento social, uso de máscaras, reuniones abarrotadas, así como prácticas cotidianas de comportamiento social en entornos privados, públicos y comerciales. Al estornudar o toser, las gotas más grandes se forman por la saliva y las gotas más pequeñas por la capa mucosa de los pulmones y las cuerdas vocales. Las gotas más pequeñas suelen ser invisibles a simple vista. Investigaciones anteriores han demostrado que la mayoría de las gotas respiratorias no viajan de forma independiente en sus trayectorias. (Dbouk & Drikakis, 2020).

Se sabe que los factores climáticos, como por ejemplo los días de frío, modifican las características físicas de las personas, ya que el tracto respiratorio sufre la entrada del aire frío porque la vellosidad de la mucosa se mueve menos y por consiguiente también se modifica el comportamiento de los virus y su evolución. Los países sufren picos de influenza durante la etapa otoño-invierno y en el verano por la entrada de personas del hemisferio norte que vienen con el virus en su cuerpo. El coronavirus tiene predilección por los receptores de la enzima convertidora de angiotensina (ECA) que está en el tracto respiratorio y en los pulmones. Los niños tienen menos de ECA y por eso no son tan afectados como los adultos.

Como se mencionó anteriormente, el principal objetivo de este trabajo es es entender cómo afectan a la evolución del virus, diferentes variables, tales como factores climáticos, aglomerados de personas, el movimiento urbano, entre otros. Para ello, se tratará de predecir la evolución del virus acorde a dichos factores, empleando herramientas de Inteligencia Artificial y Machine Learning.

II. Análisis exploratorio de datos

Para comenzar a procesar la información, se partió de diferentes sets de datos (ver inciso Datasets). Para dicho procesamiento, primero se debió llevar a cabo una limpieza de los mismos, extrayendo información de menor relevancia o información nula/errónea. Entre las herramientas de limpieza de datos, se han aplicado algoritmos de feature extraction, para intentar que los algoritmos de Machine Learning(ML) puedan llegar a aprender mejor de los datos que se brindan.

Una vez filtrados los sets de datos, se dispuso de los datos referidos a los casos de COVID-19 confirmados por el Gobierno de la Ciudad de Buenos Aires y mediante herramientas de visualización de datos, pudieron observarse diferentes variables demográficas y climáticas que pudieron afectar a la evolución del COVID-19.

En la figura a continuación (Figura 1) se puede obtener una primera visión del target del virus. Se puede observar a continuación, segregado por barrios, la edad de los contagiados. Se pueden observar casos como Barracas, la Boca, Nueva Pompeya Villa Soldati y Retiro, donde el promedio de edad es menor que el resto, presentando una media en torno a los 30 años. Es necesario destacar que son zonas muy cercanas entre sí, por lo que podría intuirse a priori un "contagio barrial comunitario" afectado por el transporte, lo cual podría haber provocado el traslado del virus a la zona. Por otro lado, barrios aledaños de la zona norte de la ciudad, tales como Belgrano, Colegiales, Palermo y Villa Devoto, presentan patrones similares entre si, con promedios mayores, que se encuentran por sobre los 40 años. Se podría suponer también la hipótesis antes mencionada siendo el transporte público un posible transmisor del virus. También es menester destacar una gran cantidad de casos por encima de la línea de 80 años.

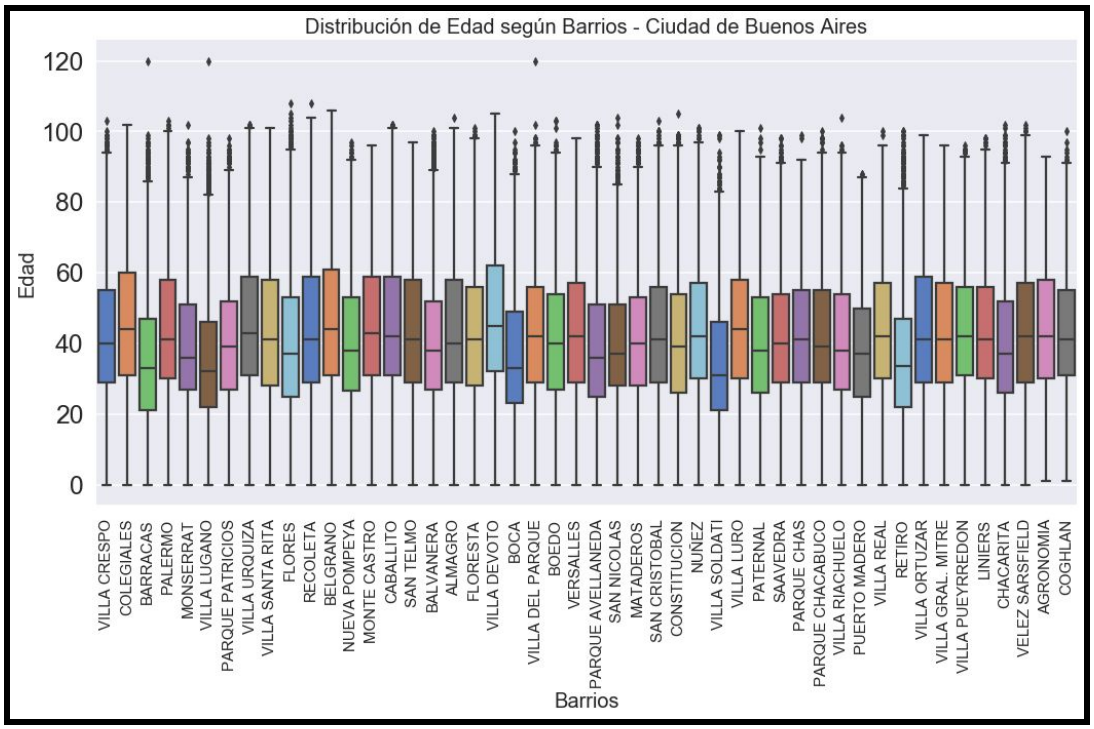


Figura 1 - Distribución de casos de COVID-19 en la Ciudad de Buenos Aires según Edad, segregado por barrios.

Elaboración Propia

Para continuar con el análisis de datos, se procedió a la búsqueda de aglomeraciones de personas, por tanto, se llevó a cabo un análisis de la densidad poblacional en los distintos barrios de la ciudad. La ciudad presenta una superficie total en torno a los 200km^2 y dentro de la misma, los barrios con mayor superficie son Palermo (15,9 km^2), Villa Lugano (9 km^2) y Villa Soldati (8,6 km^2). En las siguientes figuras se muestra la distribución de los habitantes de CABA (en habitantes por superficie), con el objeto de encontrar los barrios con mayor densidad de habitantes de la ciudad.

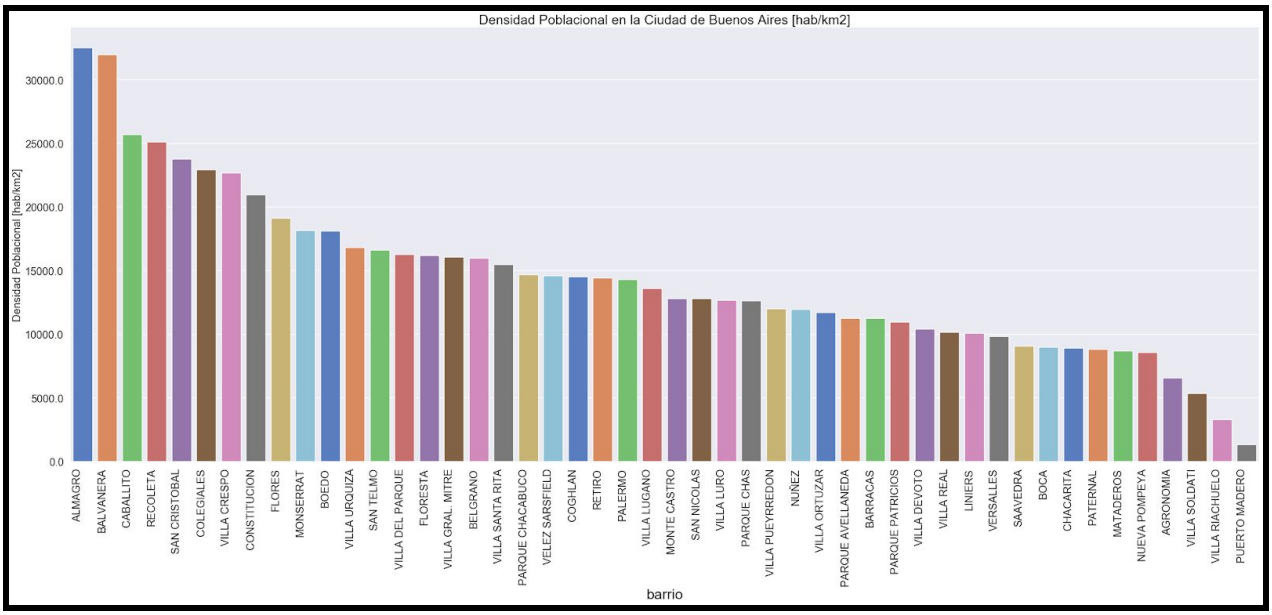


Figura 2 - Densidad poblacional en la Ciudad de Buenos Aires, discriminado por barrios [hab/km2].

Elaboración Propia

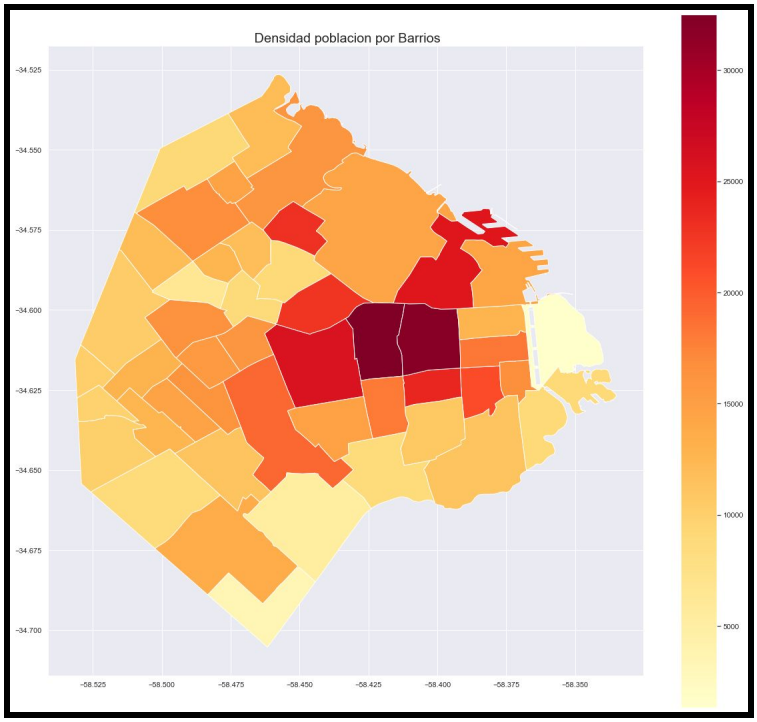


Figura 3 - Densidad poblacional en la Ciudad de Buenos Aires, discriminado por barrios [hab/km2].

Elaboración Propia

Para intentar comparar a través de una visualización geográfica, se realizó un análisis de la cantidad de casos de COVID-19 en la ciudad de Buenos Aires también segregando por barrios. En las siguientes figuras se muestra la cantidad de casos de COVID-19 según cada barrio. Se obtuvo un resultado sorprendente de casos en el barrio de Flores, siendo este el barrio con mayor cantidad de casos hasta Octubre del 2020. Luego de ello, se pueden observar Palermo (el barrio con mayor superficie de la ciudad), Balvanera (uno de los barrios con mayor cantidad de habitantes por superficie) y Villa Lugano, otro de los barrios con gran superficie.

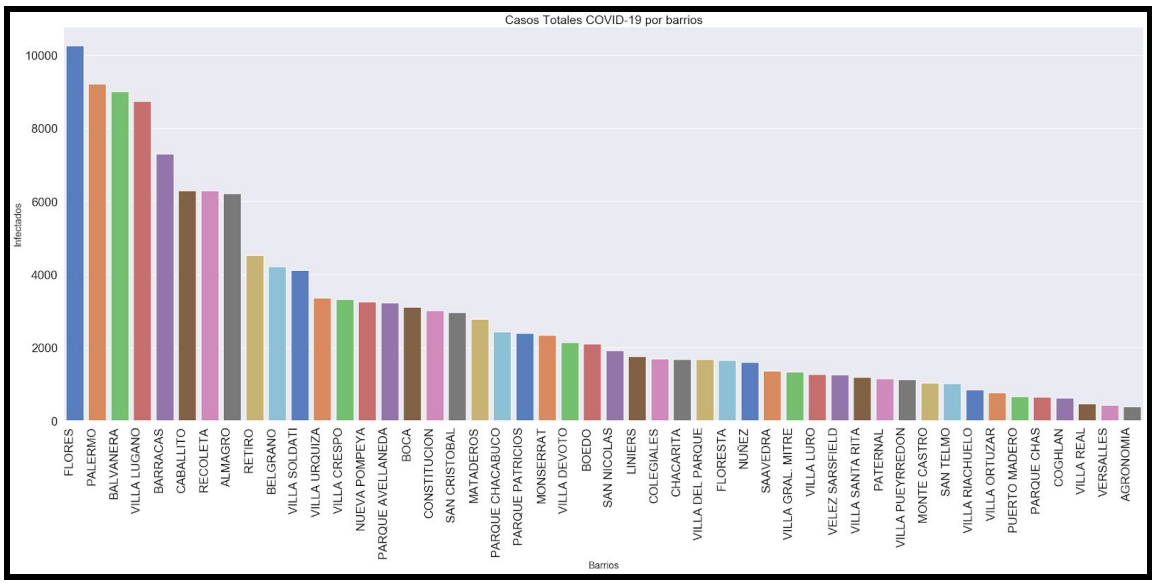


Figura 4 -Cantidad de casos de COVID-19 en la Ciudad de Buenos Aires, discriminado por barrios.

Elaboración Propia

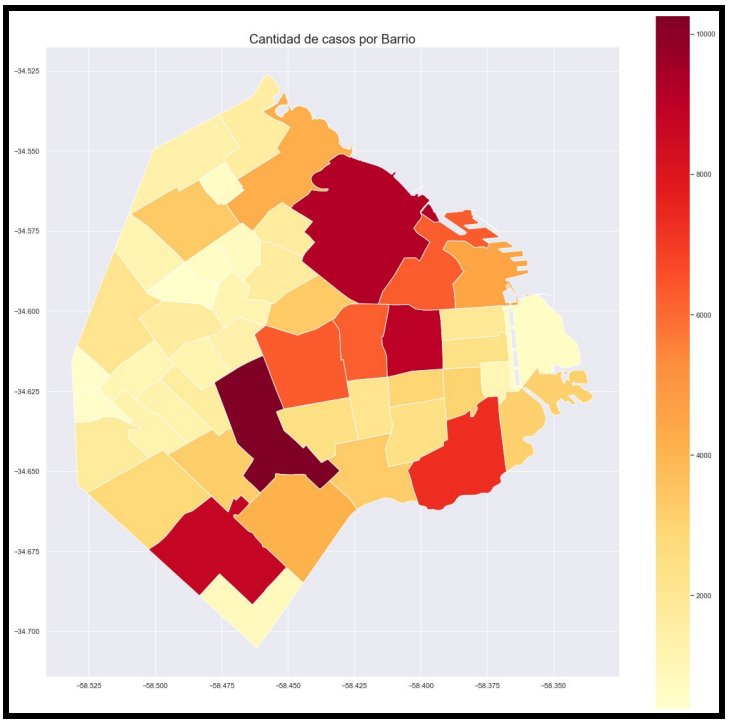


Figura 5 -Cantidad de casos de COVID-19 en la Ciudad de Buenos Aires, discriminado por barrios.

Elaboración Propia

A partir de los resultados que se fueron observando con anterioridad, se cruzaron datos demográficos con los datos de ubicación de las entradas al transporte subterráneo con el objetivo de encontrar algún patrón dado por el movimiento de las personas en el transporte público. En la siguiente figura, se puede observar cierta correlación entre los barrios afectados y la ubicación de las estaciones de subte.

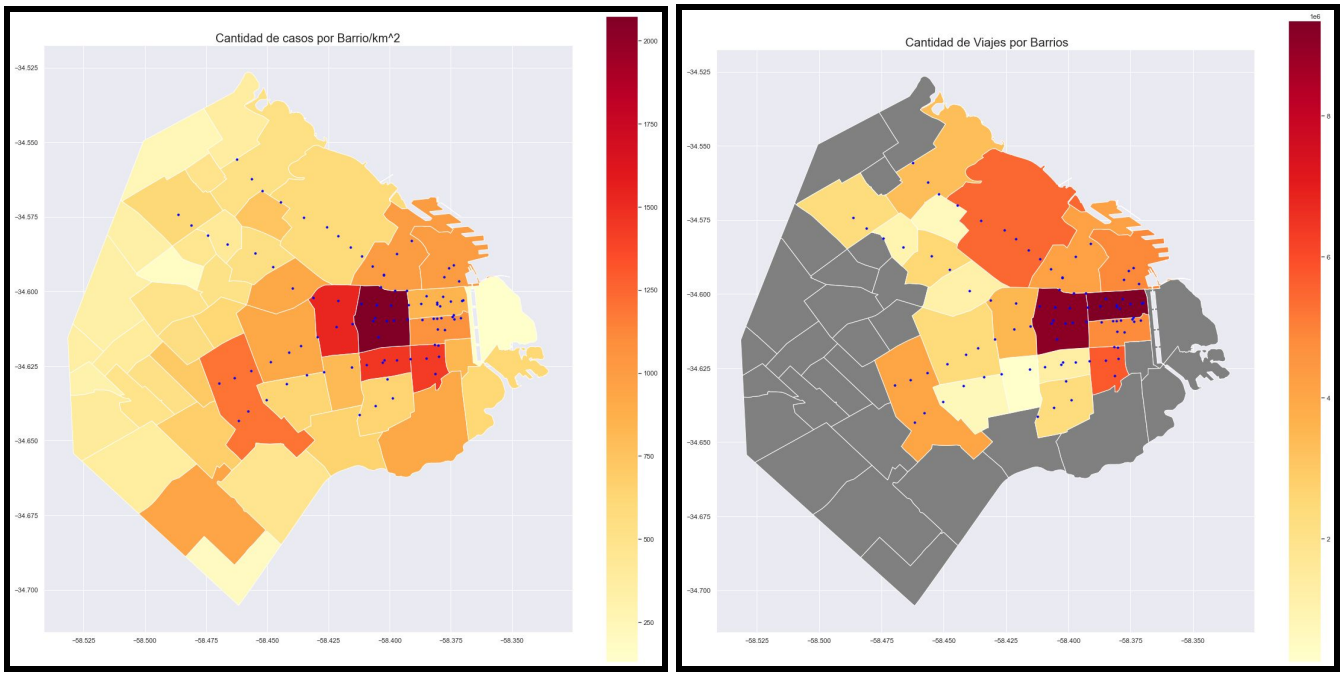


Figura 6 -Cantidad de casos de COVID-19 en la Ciudad de Buenos Aires, relación con Subterráneo.

Elaboración propia.

Con los mapas se puede llegar a intuir dónde se encuentran los lugares con alta circulación de virus y con ello se pueden llegar a tomar mejores decisiones, como se puede ver en la gráfica una de las zonas más afectadas fue Balvanera, con los datos se pudo establecer la idea de que es un lugar de alta circulación de personas y es uno de los lugares con más infectados con el virus según su superficie.

Para concluir el análisis de datos, se realizó un análisis de las variables climáticas del año, teniéndose en cuenta temperatura, humedad, presión, dirección del viento e intensidad del mismo. Mediante los factores climáticos se pudo observar que luego de picos (inversos) de bajas temperaturas se observaron picos de casos, posiblemente incrementando la transmisibilidad del virus, porque la vellosidad de la mucosa de las personas cambia sus características, se mueve menos y facilita la entrada del virus. Otro factor que puede llegar a ser clave es la presión atmosférica, ya que al aumentar la misma, se promueve la persistencia en el aire de microgotitas provocadas por la respiración, tos, estornudo, etc de las personas. La influencia de la velocidad del viento ayuda a la propagación si es combinada con baja temperatura y alta presión atmosférica.

En la siguiente figura, se puede observar la relación entre la temperatura media de la ciudad y la evolución de contagios en los barrios de Balvanera y Palermo.

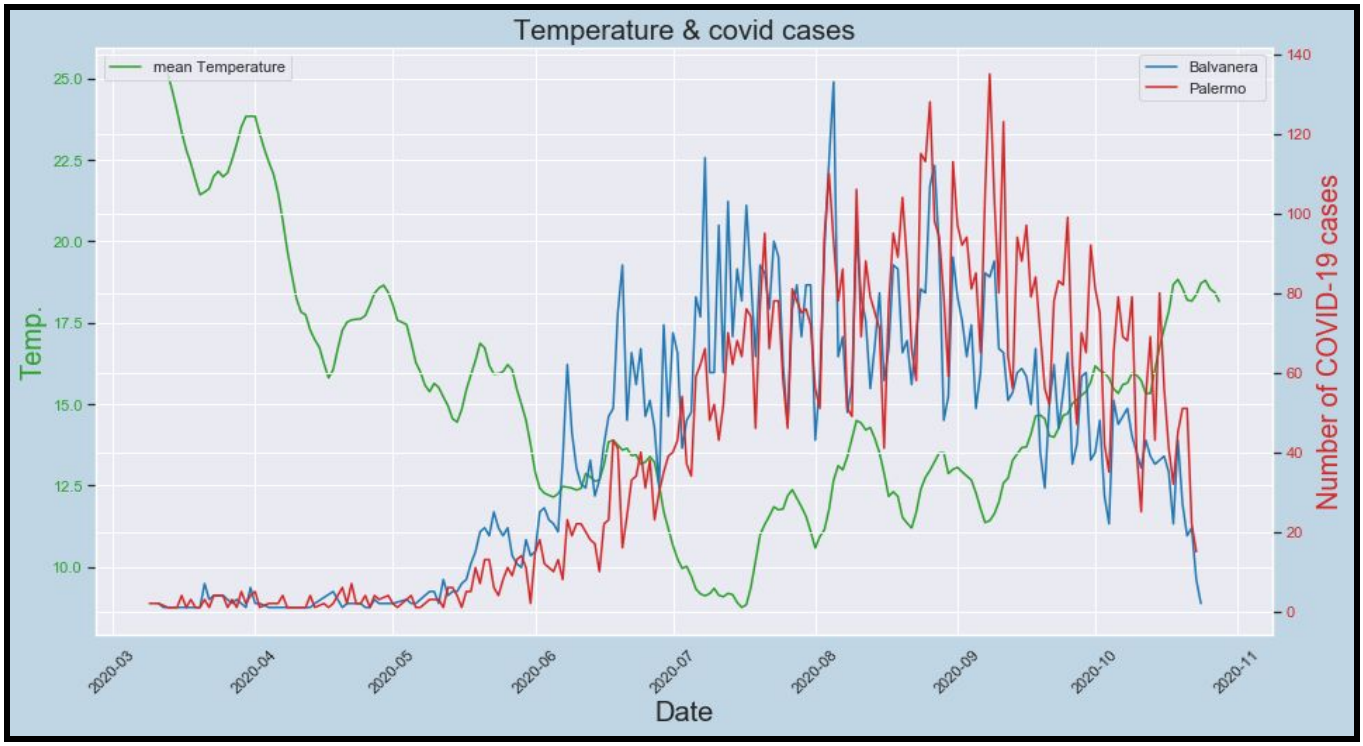


Figura 7 - Relación entre la temperatura media de la ciudad y la evolución de contagios en los barrios de Balvanera y Palermo. Elaboración Propia.

III.DATASETS

Los datasets empleados para el armado del modelo que verifique la hipótesis son los que se detallan a continuación:

- Casos COVID-19. (Gobierno de la Ciudad de Buenos Aires, 2020). Incluye la cantidad de casos positivos, altas institucionales, fallecidos y descartados por COVID-19 según lo reportado por el Sistema Integrado de Información Sanitaria Argentino (SISA).
- Datos censo 2010 CABA. (Gobierno de la Ciudad de Buenos Aires, 2020)
- Molinetes 2020. (Gobierno de la Ciudad de Buenos Aires, 2020). Cantidad de pasajeros por molinete en cada estación en rangos de a 15 minutos y discriminando según tipo de pasaje correspondiente al año 2020.
- Estaciones Subtes(geolocalización) (Gobierno de la Ciudad de Buenos Aires, 2020)
- Barrios CABA(mapa) (Gobierno de la Ciudad de Buenos Aires, 2020)
- Medición factores meteorológicos CABA. (Gobierno de la Ciudad de Buenos Aires, 2020). Información meteorológica recolectada por Torres de Monitoreo Inteligente (TMI) distribuidas en distintos puntos de la Ciudad.

IV.MÉTODOS

Con el objetivo de predecir las cantidades de contagios por días, se decidió utilizar un modelo de aprendizaje supervisado con algoritmos de regresión. Un modelo de regresión es un modelo matemático que busca determinar la relación entre una variable dependiente (Y), con respecto a otras variables de entrada (features), llamadas explicativas o independientes (X). La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores (X1, X2, X3…). Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella (esto último se debe que analizar con cautela para no malinterpretar causa-efecto).

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Yi=(\beta 0+\beta 1X1i+\beta 2X2i+\cdots+\beta nXni)+ei$$

$$f(x) = y \qquad y \in \Re \text{ Pilar de los algoritmo de regresión}$$

- β0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
- βi: es el efecto promedio que tiene el incremento en una unidad de la variable predictora Xi sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.
- ei: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

El residuo es un elemento fundamental para la medición de error del sistema, o dicho en otras palabras para la eficiencia del modelo. El objetivo es intentar minimizar en un proceso iterativo las métricas de error para poder realizar predicciones más precisas. Algunas de esas métricas son el error cuadrático medio(MSE), raíz cuadrada del error cuadrático medio(RMSE) , media del error(MAE) y el coeficiente de determinación R^2.

$$R^2 = \frac{TSS-RSS}{TSS} \qquad MSE = \frac{\Sigma(\overline{Y}-Y)^2}{n} \qquad RMSE = \sqrt{\frac{\Sigma(\overline{Y}-Y)^2}{n}} \qquad MAE = \frac{|\Sigma(\overline{Y}-Y)^2|}{n}$$

Para obtener la función de regresión se determinaron como parámetros.

y(target) = Cantidad de Casos (predicción)

En cuanto a las features de entrada (X) se emplearon las siguientes dentro del modelo de regresión.

- Fecha
- Temperatura minima
- Temperatura maxima
- Humedad
- Velocidad del Viento
- Dirección del Viento
- Presion atmosferica
- Densidad Poblacional
- Pasajes Totales por día
- Casos dia anterior

Todas las features mencionadas anteriormente son númericas con excepción de la dirección del viento, por lo cual, mediante herramientas de Feature Engineering, se transformaron las variables categóricas en variables binarias.

Los algoritmos de regresión que se utilizaron son del tipo supervisado y se mencionan a continuación:

- Support Vector Regression
- KNN Regression
- Random Forest Regression

Una vez definidos los modelos, se escalaron los datos empleando un StandardScaler, para que los datos del dataset a emplear respeten una distribución standard. Una vez definidos los modelos que se van a utilizar para entrenar y testear el dataset, se realizó un split entre datos de entrenamiento y datos de testeo determinando:

- Datos de entrenamientos = 80%
- Datos de testeo= 20%

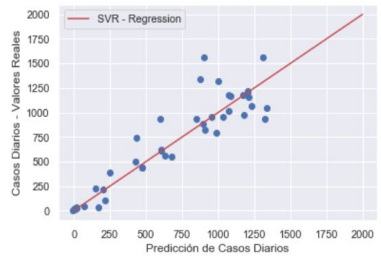
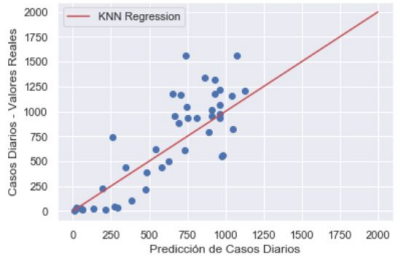
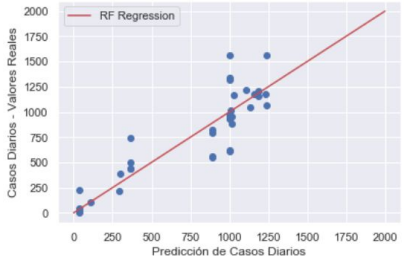
Mediante herramientas de ScikitLearn, tales como GridSearch y CrossValidation, se determinó la mejor combinación de hiperparámetros para cada modelo para así obtener un mayor rendimiento del modelo.

Los hiperparámetros obtenidos por CrossValidation para cada algoritmo fueron los siguientes:

Hiper Parámetros obtenidos por GridSearch y CrossValidation		
SVR	KNN	RFR
'C': 2000, 'gamma': 0.01, 'kernel': 'rbf'	'n_neighbors': 8	'n_estimators': 170

.

V. Resultados y conclusiones

		
Support Vector Regression	KNN Regression	Random Forest Regression
R2 score: 0.846155 MSE: 34788.083885 MAE: 118.891602	R2 score: 0.665938 MAE: 209.630814 MSE: 75539.543968	R2 score: 0.847 MAE: 124.590 MSE: 34639.112

	Model	R2	MSE	MAE
1	SVR	0.846	34788.084	118.892
2	KNN	0.666	75539.544	209.631
3	Random Forest	0.847	34639.112	124.590

Figura 8 - Resultados obtenidos por los diferentes modelos de Regresión.

Elaboración Propia

Los modelos de Support Vector Regression y Random Forest Regression presentaron un rendimiento aceptable, alcanzando una precisión del 84%. El modelo KNN no presentó resultados aceptables hasta el momento. Empleando estos modelos, podría predecirse con una considerable precisión la evolución del COVID-19 en la ciudad de Buenos Aires, a partir de los factores climáticos y demográficos antes mencionados.

References

[1] Pérez Abreu, Manuel Ramón, Gómez Tejeda, Jairo Jesús, & Dieguez Guach, Ronny Alejandro. (2020). Características clínico-epidemiológicas de la COVID-19. Revista Habanera de Ciencias Médicas, 19(2), e3254. Epub 22 de abril de 2020. Recuperado en 17 de noviembre de 2020, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1729-519X2020000200005&lng=es&tlng=es.

[2] Gutiérrez-Hernández, O., & García, L.V. (2020). ¿Influyen tiempo y clima en la distribución del nuevo coronavirus (SARS CoV-2)? Una revisión desde una perspectiva biogeográfica. Investigaciones Geográficas, (73), 31-55. https://doi.org/10.14198/INGEO2020.GHVG

[3] Peng, X., Xu, X., Li, Y. et al. Transmission routes of 2019-nCoV and controls in dental practice. Int J Oral Sci 12, 9 (2020). https://doi.org/10.1038/s41368-020-0075-9

[4] G. Dbouk, T., & Drikakis, D. (2020). On coughing and airborne droplet transmission to humans. Physics of Fluids, 32(5), 053310.

[5] Liu, H., & Motoda, H. (Eds.). (1998). Feature extraction, construction and selection: A data mining perspective (Vol. 453). Springer Science & Business Media., pp.206–207.

[6] Torres, J. (2018). DEEP LEARNING Introducci—n pr‡ ctica con Keras. Lulu. com. Pp.78

[7] Material extraído de la asignatura “Ciencia de Datos”, correspondiente al plan de carrera de Ingeniería Industrial en Universidad Tecnológica Nacional (UTN) Facultad Regional Buenos Aires, año 2020.

[8] Gobierno de la Ciudad de Buenos Aires. (2020). Buenos Aires Data / Dataset. Recuperado el Octubre de 2020, de Casos COVID-19: https://data.buenosaires.gob.ar/dataset/casos-covid-19

[9] Gobierno de la Ciudad de Buenos Aires. (2020). *Buenos Aires Data / Dataset* . Recuperado el Octubre de 2020, de Molinetes 2020: https://data.buenosaires.gob.ar/dataset/subte-viajes-molinetes/archivo/a43d8d7e-0e5e-4706-853b-303f567d82d0

[10] Gobierno de la Ciudad de Buenos Aires. (2020). *Buenos Aires Data / Datasets.* Recuperado el Octubre de 2020, de Información Meteorológica: https://data.buenosaires.gob.ar/dataset/informacion-meteorologica

[11] Gobierno de la Ciudad de Buenos Aires. (s.f.). *Estadística Ciudad.* Recuperado el Octubre de 2020, de Datos Censo BA 2010: https://www.estadisticaciudad.gob.ar/eyc/?p=28011