

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 %.

El principal objetivo de este trabajo 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 Machine Learning.

Datasets

- Casos COVID-19.
- Datos censo 2010 CABA.
- Molinetes 2020.
- Estaciones Subtes(geolocalización)
- Barrios CABA(mapa) (Gobierno de la Ciudad de Buenos Aires, 2020)
- Medición factores meteorológicos CABA.

Métodos

Con el objetivo de predecir las cantidades de contagios por día, se decidió utilizar un modelo de aprendizaje supervisado con algoritmos de regresión. 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. Los modelos lineales múltiples siguen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}) + e_i$$

Las métricas para medir la eficiencia del modelo son: error cuadrático medio (MSE), raíz cuadrada del error cuadrático medio (RMSE), error absoluto medio (MAE) y el coeficiente de determinación R².

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

- Support Vector Regression
- KNN Regression
- Random Forest Regression

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.

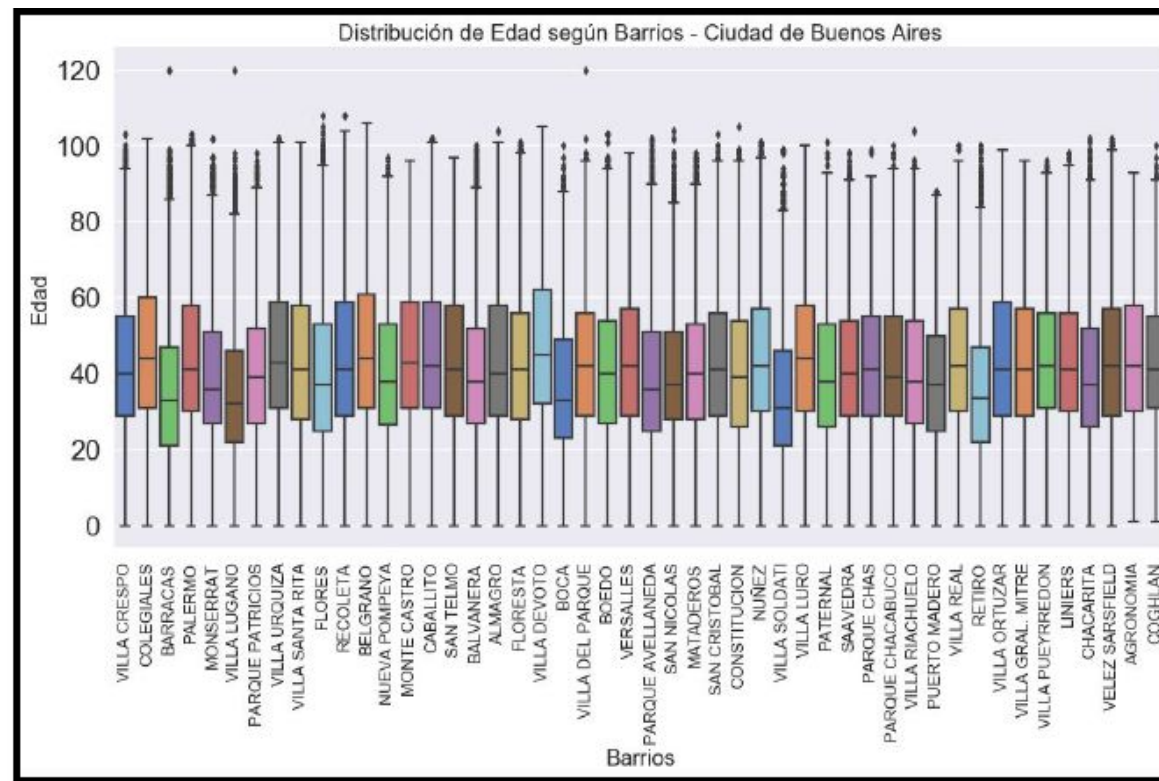
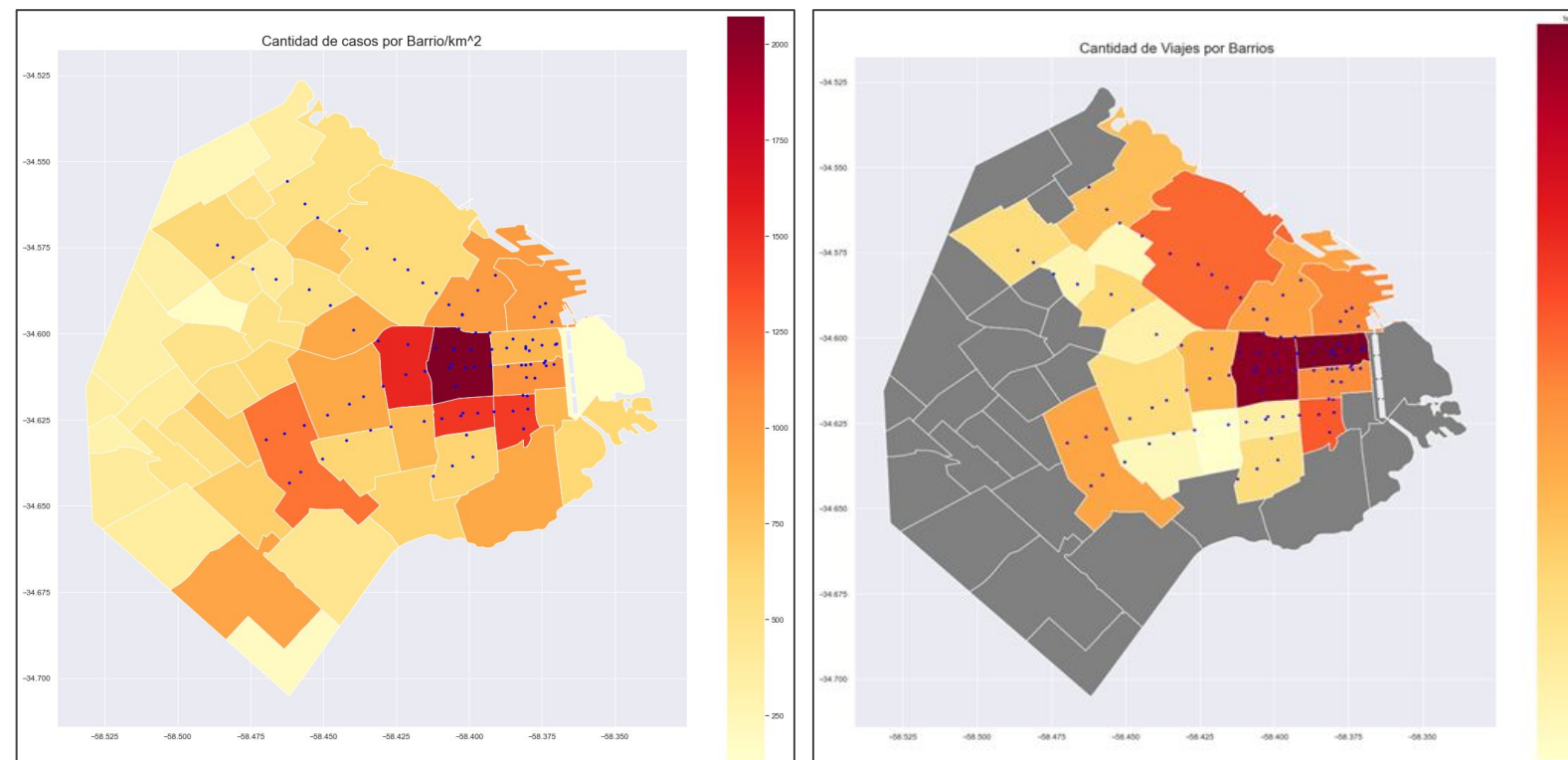


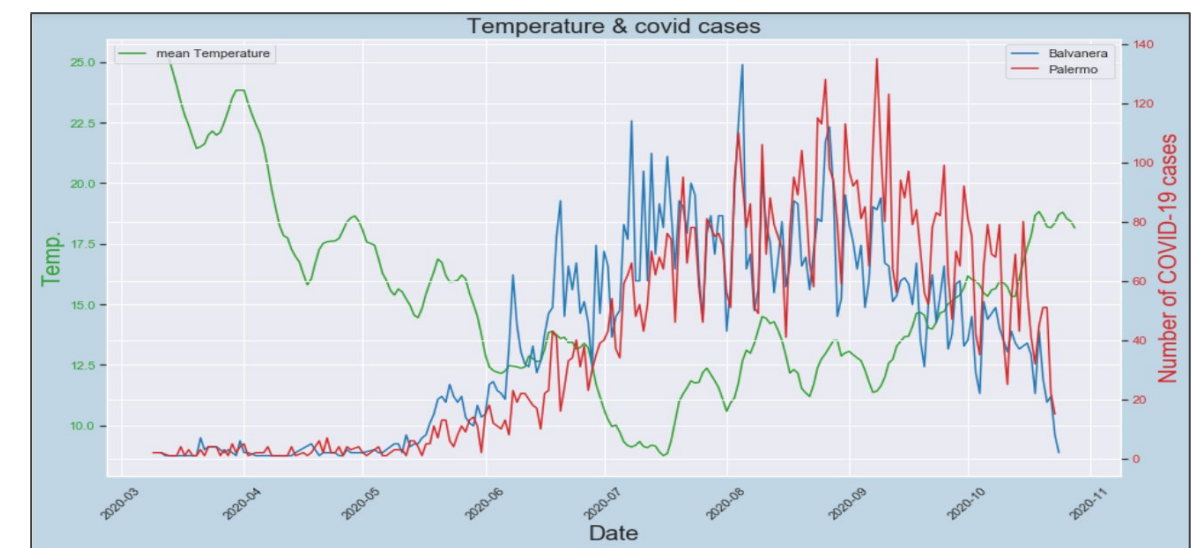
Figura 1 - Distribución de casos de COVID-19 en la Ciudad de Buenos Aires según Edad, segregado por barrios.

Elaboración Propia

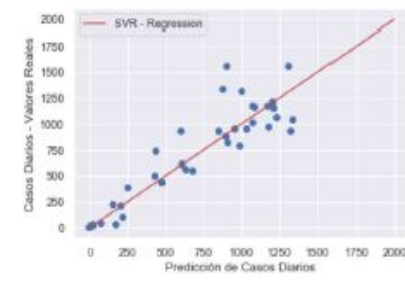
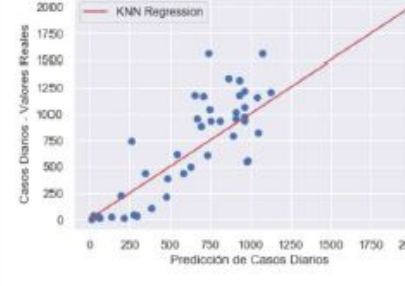
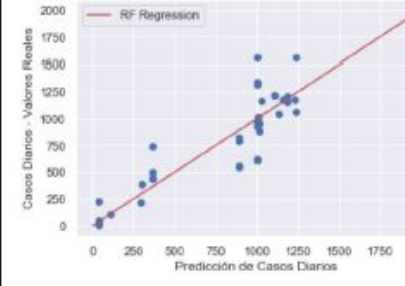
Se trabajaron con datos demográficos para encontrar relación entre la cantidad de personas por barrio, la superficie (en km²) y la cantidad de casos de COVID. En la imagen también pueden observarse las distintas entradas al subterráneo de la ciudad.



Para concluir el análisis de datos, se realizó un análisis de las variables climáticas del año, teniendo 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. 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.



Conclusiones y Resultados

		
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

Los modelos de Support Vector Regression y Random Forest Regression tuvieron un rendimiento aceptable, alcanzando una precisión del 84%.

	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	34639.1120	124.59