Aplicación de modelos de regresión y técnicas de clustering en datos de COVID -19

Trabajo práctico final - Cluster AI - 2020

Ambrogi, Tomás; Rocamora, Leandro; Suli, Solange
Universidad Tecnológica Nacional, Buenos Aires, Argentina

Resumen
La enfermedad por coronavirus 2019 (COVID-19) es una enfermedad infecciosa causada por el SARS-CoV-2 y ha provocado una enorme cantidad de muertes a nivel mundial. El objetivo de este trabajo se basa en encontrar un modelo de regresión que permita predecir la cantidad de fallecidos del día siguiente a partir de algunas variables relacionadas con la enfermedad. Además se utilizaron algoritmos de Clustering para evaluar si existía una posible segmentación de los datos.

Palabras clave: Coronavirus, modelos de regresión, clustering, fallecidos.

1. INTRODUCCIÓN

La enfermedad por coronavirus 2019 (COVID-19) es una enfermedad infecciosa causada por un síndrome respiratorio severo coronavirus 2 (SARS-CoV-2). Fue identificado por primera vez en diciembre de 2019 en la ciudad de Wuhan, capital de la provincia de Hubei, en la República Popular China.

Según el Reporte Diario Vespertino Nro 389, del 30 de septiembre del 2020 emitido por Ministerio de Salud de la República Argentina (https://www.argentina.gob.ar/coronavirus/informes-diarios/reportes/septiembre2020) fueron registrados, hasta el día de la fecha, 751.001 casos positivos en todo el territorio, de los cuales 594.645 fueron pacientes recuperados y 139.419 correspondían a casos confirmados activos. En particular, para la Ciudad Autónoma de Buenos Aires (CABA), fueron registrados 125.068 casos positivos hasta el 30 de septiembre del año correspondiente.

El objetivo de este trabajo se basa en poder entrenar, escoger y evaluar un modelo de regresión que permita predecir cantidad de fallecidos del día siguiente por coronavirus en base a distintas variables relacionadas con la enfermedad. A su vez, se estudió la relación de los datos a través de aprendizaje no supervisado usando algoritmos de clustering para ver si existía algún modo de segmentar y exponer relaciones que no son evidentes.

2. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

2.1. Preprocesamiento de datos

Para el siguiente trabajo se utilizó la base de datos del Ministerio de Salud de la República Argentina (https://datos.gob.ar/dataset/salud-covid-19-casos-registrados-republica-argentina) para el período Marzo - Septiembre 2020. Se filtraron aquellos casos registrados para la Ciudad Autónoma de Buenos Aires (CABA) y se tomó como fecha de referencia del caso la “fecha de apertura”. Si bien, no necesariamente corresponde a la fecha de toma de muestra, ésta no presentaba datos faltantes como sí era el caso de fecha de diagnóstico. Luego, se

descartaron algunas variables que no eran pertinentes para el estudio y se reestructuraron otras de manera tal que sean de utilidad para el mismo. Se obtuvo un dataset correspondiente a los casos de CABA el cual contaba con 18 features y 361133 samples.

2.1.1. Tratamiento de datos faltantes

Se encontraron datos faltantes en las features correspondientes a fechas (como "fecha_inicio_sintomas", "fecha_internación", "fecha_fallecimiento", entre otras) y en las edades. En primer lugar, se decidió descartar aquellos datos faltantes en la "edad" (37 datos) ya que era una proporción muy pequeña en relación a la cantidad de datos confirmados de CABA totales (127054 samples). En segundo lugar, las features de las fechas fueron reemplazadas por variables booleanas para que fueran de utilidad para el estudio realizado. Por lo tanto, no fueron utilizadas para la creación del dataset final.

2.1.2. Extracción de features y dataset acumulado por fecha

Con el fin de contar con un dataset completo y con gran información, se decidió crear nuevas features de interés ("argentino", "mayor_65", entre otras). Además para poder realizar un estudio temporal acerca de la cantidad de fallecidos diarios, se acumularon casos diarios (según "fecha_apertura"), confirmados de CABA a partir de una Tabla Pivot, obteniendo así el dataset final para trabajar. Las features utilizadas se encuentran detalladas en la **Tabla 1**.

Tabla 1: descripción de las features utilizadas en el dataset final. El mismo corresponde a una tabla Pivot en la que se muestran los acumulados diarios de cada una de las features.

Features	**Descripción**
argentino	Cantidad de argentinos
asistencia_respiratoria	Cantidad de personas que requirieron asistencia respiratoria mecánica
confirmados	Cantidad de casos confirmados
cuidado_intensivo	Cantidad de personas que requirieron cuidado intensivo
edad	Edad
fallecidos	Cantidad de fallecidos
mayor_65	Cantidad de personas mayor a 65 años
mayor_65_internacion	Cantidad de personas mayor a 65 años que requirieron internación
prensentan_sintomas	Cantidad de personas que presentaron síntomas
privado	Origen de financiamiento privado
publico	Origen de financiamiento público
internacion	Cantidad de personas que requirieron internación
sexo_F	Cantidad de personas de sexo femenino
sexo_M	Cantidad de personas de sexo masculino

A partir del análisis exploratorio de los datos se realizaron algunas figuras (**Figura 1 y 2**) para visualizar las distribución de los datos obtenidos.

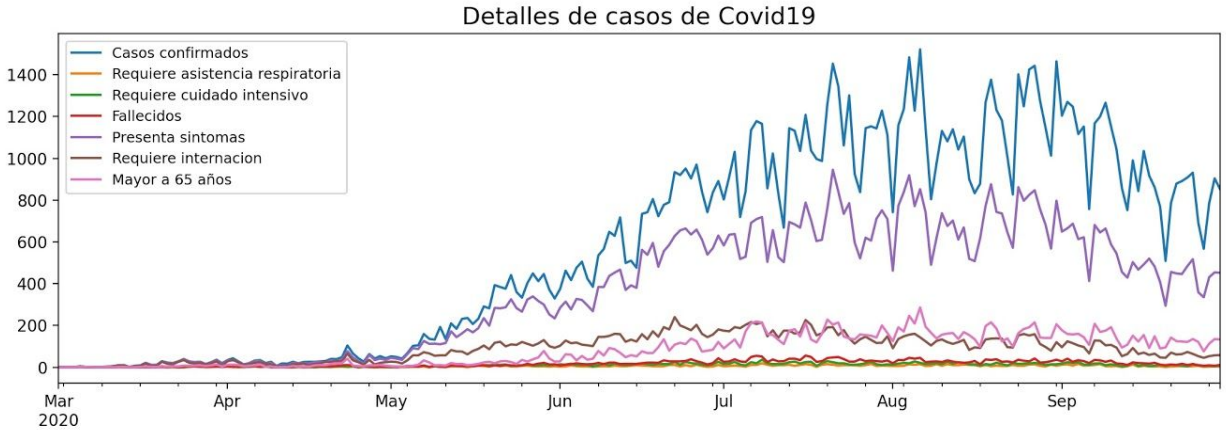

Figura 1: Series temporales (Marzo - Septiembre 2020) de algunas de las features obtenidas.

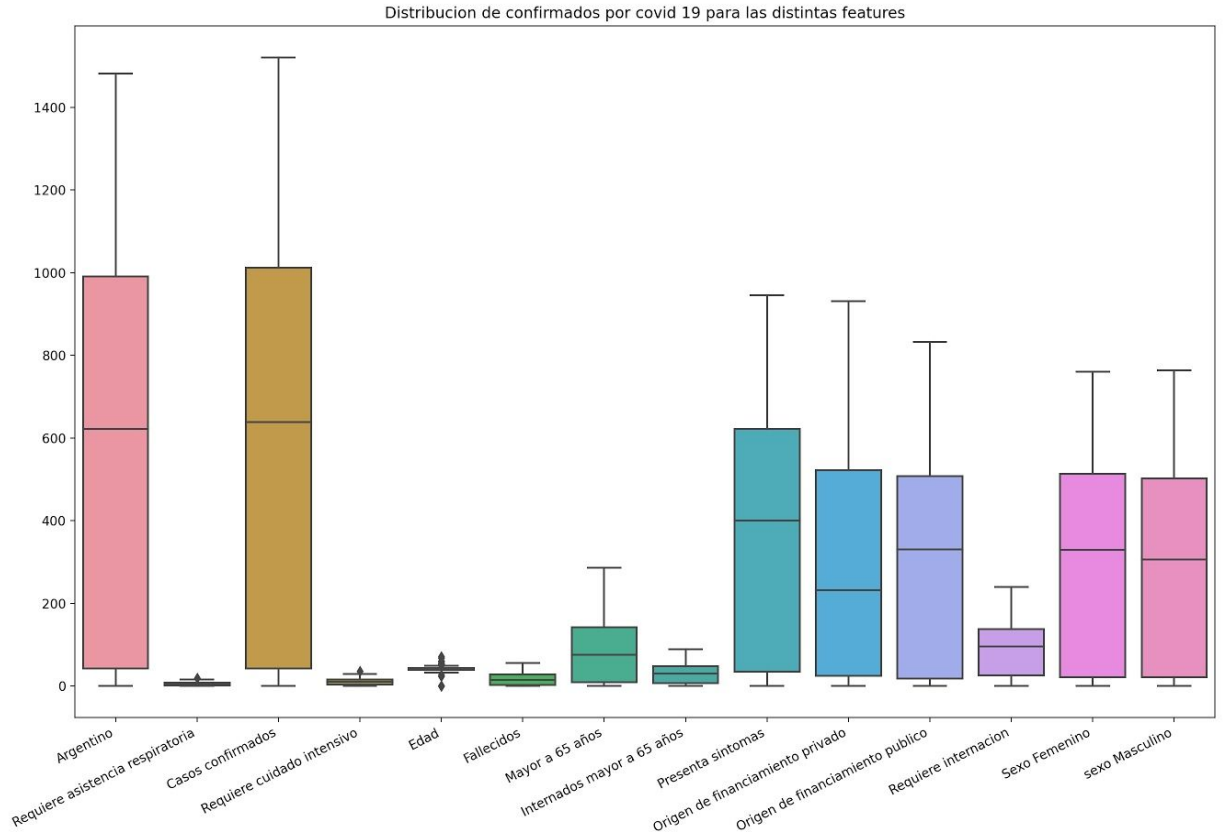

Figura 2: Distribución de casos confirmados para las distintas features obtenidas.

3. MATERIALES Y MÉTODOS

3.1. Clustering

A la hora de plantear algoritmos de clustering se propuso trabajar con el dataset original así como también con sus componentes principales. Además del dataset original, se descompone en otros dos dataset, uno con 30 y otro con 6 componentes principales. Cada uno de los datasets es sometido al mismo proceso de clustering por medio del algoritmo KMeans. Se estudian los resultados en base a 2, 4 y 6

clusters. La asignación de los clusters en el algoritmo de KMeans está dada por:

$$S_i^{(t)} = \{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \}$$

$$\forall j \ 1 < j < k$$

Donde cada x_p es asignado a solo un centroide $m_i^{(t)}$. Los centroides luego se re calculan dada la ecuación:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

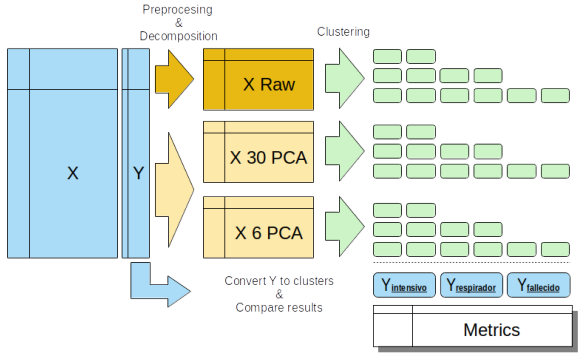

Figura 3: Esquema de las pruebas realizadas en clustering.

3.2. Modelos de regresión

Con el objetivo de predecir la cantidad de fallecidos del día siguiente planteamos distintos modelos de regresión, entre ellos: Linear Regression, Ridge Regression, Support Vector Machines y K Nearest Neighbors Regressor.

Para entrenar los modelos se utilizaron los datos agregados por día desde el inicio de la pandemia hasta septiembre, dejando los datos de los primeros quince días de septiembre para validar los modelos. Se hicieron pruebas con el dataset sin modificar, otras usando el logaritmo de algunas de las variables y por último, empleando un promedio móvil de 2 días de las variables de entrada. En todos los casos se utilizó un escalado normal de los datos para que las variables contarán con $= 0$ y $= 1$.

Como se puede ver en la **Figura 1**, dada la naturaleza exponencial del contagio del COVID-19, probamos linealizando algunas de las variables de entrada a partir del logaritmo. Debido a que había días con valores nulos (por ejemplo, días con cero contagios) se optó por la siguiente expresión de linealización para evitar valores infinitos en las entradas: $\widehat{x} = log(1 + x)$.

Por otro lado, de la **Figura 1** también se pueden apreciar fluctuaciones significativas día a día a partir de las cuales se decidió plantear una media móvil para suavizar y reducir la variabilidad de los datos de entrada. Se probaron distintos valores para el ancho de la ventana y se obtuvieron los mejores resultados con una ventana de 2 días. Matemáticamente la operación para cada variable de la matriz de datos X es la siguiente: $\widehat{x}_j = \frac{1}{2} \sum_{i=0}^{1} x_{j-i}$. Donde x_j es el sample j de la matriz X.

3.2.1. Evaluación y Métricas

Para garantizar la elección del mejor modelo utilizamos la técnica conocida como validación cruzada, que dada la temporalidad de los datos se plantea de manera distinta a la tradicional (ver TimeSeriesSplit: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html). A la hora de evaluar los modelos de regresión entrenados se plantearon distintas métricas, buscando siempre maximizar el R^2, y al mismo tiempo, minimizar tanto el Mean Absolute Error (MAE) como el Root Mean Squared Error (RMSE).

4. RESULTADOS

4.1. Clustering

De la **Tabla 2** se puede observar que el preprocesamiento de los datos con PCA mejora los clusters obtenidos. Las primeras 6 componentes principales dan como resultado clusters más agrupados y definidos dado que aumenta el Silhoutte Score.

PREPROS.	Cluster: CLUSTERS	Silhouette Score
raw	2	0.1737
	4	0.1054
	6	0.1056
30	2	0.2142
	4	0.2347
	6	0.1411
6	2	0.4000
	4	0.4319
	6	0.4748

Tabla 2: Silhouette Score para cada preprocesamiento del dataset.

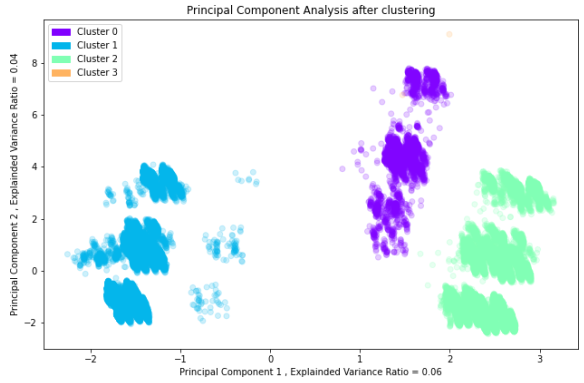

Figura 4: Clusters graficados sobre las dos componentes principales.

Además, a partir de la **Figura 4**, se puede observar que al definir 4 o más clusters estaría generando al menos un cluster vacío, lo cual indica que para segmentar los datos 3 o menos clusters serían suficientes.

4.2. Regresión

A continuación se muestra en la **Tabla 3** las métricas obtenidas para todos los modelos mencionados anteriormente. De todos ellos, el que presenta mejores resultados es la Ridge Regression *(Hoel, et al. (1970) y Suganya, et al. (2020))* usando el dataset promediado (**M**oving **A**verage).

	model	r2	rmse	mae
7	Ridge Reg MA	0.671654	4.748721	4.193563
6	Support Vector Reg MA	0.668763	4.769576	4.302948
4	Linear Reg MA	0.668659	4.770325	4.168058
1	KNeighbors Reg	0.547466	5.574887	4.619048
9	KNeighbors Reg Log	0.438724	6.208673	5.119048
2	Support Vector Reg	0.281275	7.025739	5.838129
10	Support Vector Reg Log	0.268670	7.087077	5.749135
3	Ridge Reg	0.263805	7.110613	6.035530
0	Linear Reg	0.246215	7.195060	6.179572
8	Linear Reg Log	0.186789	7.473296	5.948271
11	Ridge Reg Log	0.184786	7.482493	5.957129
5	KNeighbors Reg MA	0.096251	7.878336	6.764637

Tabla 3: Métricas para los distintos modelos entrenados.

Se puede observar que el mejor modelo tiene en promedio un error absoluto de 4,19 personas.

En la **Figura 5** se detallan las muertes para los días 1 al 14 de septiembre y las predicciones de dos de los modelos entrenados.

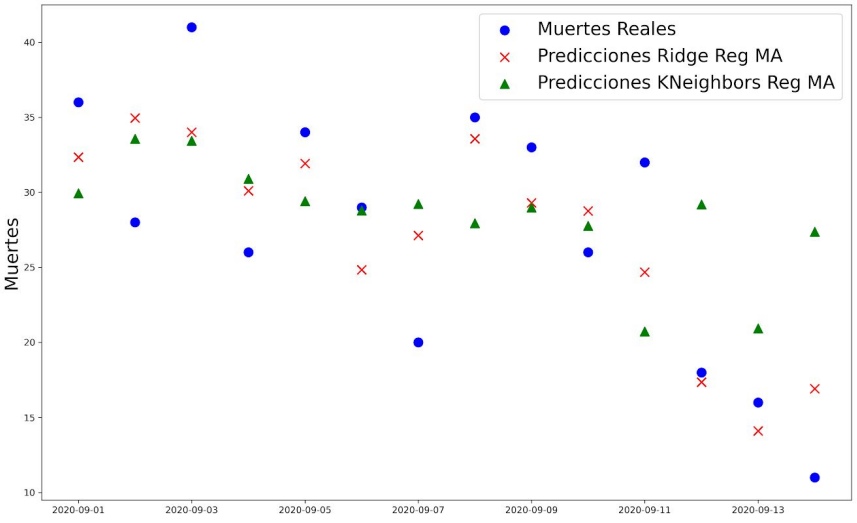

Figura 5: gráfico de dispersión en el que se detallan: las muertes reales (en azul), las predicciones con el mejor modelo (en rojo) y las predicciones con el peor modelo (en verde) para las fechas del 1 al 14 de septiembre.

A partir de la **Figura 5**, se observa que existe una relación entre las variables de entrada y los fallecidos del día siguiente. Esta relación se presenta más clara al suavizar los datos a través de una media móvil uniforme de dos días.

5. CONCLUSIONES

5.1. Clustering

En lo que respecta a clustering se observó que definir 4 o más clusters solía generar al menos un cluster vacío, indicando que para segmentar los datos 3 o menos clusters son suficientes. Además, es posible que el dataset presente outliers que terminen agrupados por su cuenta cuando la cantidad de clusters aumenta. El Silhouette Score aumenta considerablemente cuando se disminuye la dimensionalidad de los datos, encontrando su máximo en el estudio al descomponerse en únicamente 6 componentes principales. Se considera pertinente continuar con el estudio de clustering, analizando posibles métricas que permitan obtener mayores conclusiones para el estudio.

5.2. Regresión

Se encontró una relación entre las variables de entrada y los fallecidos del día siguiente. Esta relación se hace más clara si se suavizan los datos a través de una media móvil uniforme de dos días. Como estrategia para utilizar este modelo se propone re entrenarlo cada noche con los últimos datos disponibles y así predecir los fallecidos del día siguiente.
Queda para futuras pruebas evaluar distintas ventanas para el promedio móvil, como también la utilización de redes neuronales recurrentes.

REFERENCIAS

1. Ghosal, S., Sengupta, S., Majumder, M., & Sinha, B. (2020). Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.
2. Goodfellow, I. Deep Learning. (2015).
3. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.
4. Parbat, D., & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons & Fractals*, *138*, 109942.
5. Suganya, R.; Arunadevi, R.;Buhari, S. M. COVID-19 Forecasting using Multivariate Linear Regression. 1–17 (2020).