Predicciones de Covid 19 usando Algoritmos de Machine Learning

Anabel Dalila Paredes Flores
anabel.paredes@ucsp.edu.pe
Universidad Católica San Pablo

Abstract

Este documento contiene la explicación de los pasos para construir un predictor del número de casos y el número muertes producidas por el Covid-19 en el Perú y específicamente en la región de Arequipa. Para esto se realiza una implementación de Regresión Lineal Multivariada usando gradiente descendiente; además de algunos modelos proporcionados por la librería de $Sci\text{-}kit\ learn$.

1 Introduction

El Covid 19 ha producido daños económicos y pérdidas humanas ocasionando estrés a los profesionales de la salud y administrativos. Los modelos de predicción son requeridos justamente para estimar la frecuencia de número de casos y muertes que pueden ocurrir en un futuro para evaluar y emplear otras estratégias. Para realizar la predicción en un rango de fechas del número de casos y muertes del Covid-19, primero se analizaó el conjunto de datos resolviendo que el problema debía trabajar con series temporales por ser una secuencia de datos medidos en momentos a lo largo del tiempo y ordenados de forma cronológica. También se consideraron modelos de aprendizaje supervisado como Regresión Lineal Múltiple y Regresión Polinomial.

2 Metodología

La predicción de casos y muertes del Covid-19 presentados en este documento utiliza dos modelos de aprendizaje supervisado:

2.1 La Regresión Lineal Múltiple(RLM)

La RLM explica la relación que existe entre la variable respuesta y y una variable explicativa X.
En el aprendizaje supervisado se calcula la hipótesis h, esto se hace para aproximar y como una función lineal x.

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n = \theta^T x \tag{1}$$

También se debe calcular la Función de Costo a minimizar:

$$J(\theta_0, \theta_1, ..., \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta x^{(i)} - y^{(i)})^2 \tag{2}$$

donde m es la cantidad de datos.

Para encontrar los parámetros se halla la Gradiente Descendente:

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x)^{(i)} - y^{(i)}) x_j^{(i)} \tag{3}$$

2.2 Regresión Polinomial(RP)

La RP es una modificación de RLM donde la hipótesis h esta definida por la Ecuacion 4.

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2^2 + ... + \theta_n x_n^m = \theta^T x \tag{4}$$

El modelo al ser un polinomio tiene la posibilidad de ajustarse mejor a los gráficos de curva. El problema se da cuando en el conjunto de datos encuentra vacios.

3 Estado del Arte

La investigación de los modelos de Aprendizaje de máquina utilizados para la Predicción en casos y muertes del Covid-19 fue basandose de algunas ideas y definiciones encontradas en *Kaggle* [1]

4 Experimentos y Resultados

La sección explica como fue realizada la predicción de casos y muertes por Covid-19 en el Perú y la región de Arequipa. En el caso de la RP se utilizó la librería *sci-kit learn* y el caso de la RLM se utilizó una implementación propia.

4.1 Pre-procesamiento de Datos

Los datos fueron extraídos de la recolección hecha por José M. Castagnetto [2]. La cantidad total de datos que registra al Perú como único país es de 5261 y a nivel de región en este caso la región Arequipa con 208 registros. En la Fig.1 se visualiza que el conjunto de datos tiene vacios por momentos en el tiempo. El conjunto de datos fue dividido en dos grupos 40% para pruebas y 60% para entrenamiento.

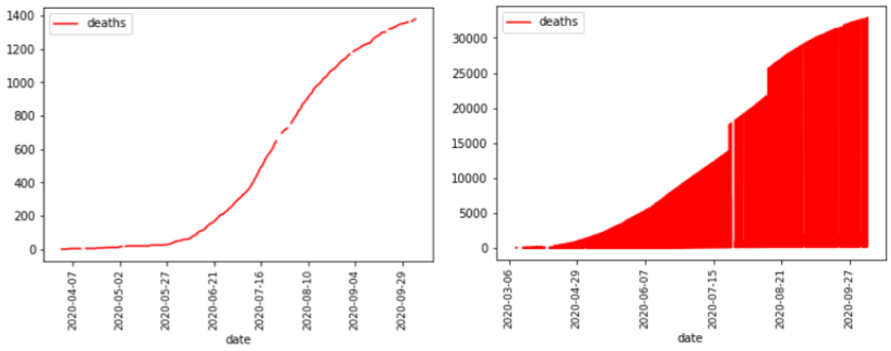

Figure 1: Series temporales con vacios

4.2 Modelos de Aprendizaje de Máquina

La tasa de aprendizaje y número de iteraciones fue de 0.0001 y 6000 respectivamente para el conjunto de datos de Perú y Arequipa. En la Tabla 1 se visualiza la predicción de los casos confirmados y muertes por Covid-10 dado el rango de fecha del 19 de octubre al 29 de octubre del 2020.

Las Figuras 2 y 3 muestran el comportamiento que sigue el Costo en la predicción de casos y muertes de Arequipa y Perú respectivamente.

[1] https://www.kaggle.com/therealcyberlord/coronavirus-covid-19-visualization-prediction
[2] https://github.com/jmcastagnetto/covid-19-peru-data/blob/master/datos/covid-19-peru-data.csv

	RLM				RP			
	Arequipa		Perú		Arequipa		Perú	
Fecha	Casos	Muertes	Casos	Muertes	Casos	Muertes	Casos	Muertes
19-10-2020	578.0	5.0	3561.0	97.0	120.0	49.0	-1535.0	-9.0
20-10-2020	612.0	6.0	3566.0	97.0	93.0	45.0	-1531.0	-9.0
21-10-2020	647.0	8.0	3570.0	97.0	68.0	41.0	-1528.0	-9.0
22-10-2020	681.0	9.0	3575.0	97.0	44.0	38.0	-1525.0	-9.0
23-10-2020	715.0	10.0	3579.0	97.0	23.0	34.0	-1522.0	-9.0
24-10-2020	750.0	11.0	3583.0	98.0	3.0	31.0	-1519.0	-9.0
25-10-2020	784.0	12.0	3588.0	98.0	-16.0	27.0	-1515.0	-9.0
26-10-2020	818.0	13.0	3592.0	98.0	-33.0	24.0	-1512.0	-9.0
27-10-2020	853.0	15.0	3597.0	98.0	-48.0	21.0	-1509.0	-9.0
28-10-2020	887.0	16.0	3601.0	98.0	-61.0	18.0	-1506.0	-9.0
29-10-2020	921.0	17.0	3606.0	98.0	-73.0	15.0	-1502.0	-9.0

Table 1: Predicción de casos y muertes por Covid-19

La Figura 4 muestra las gráficas que comparan el modelo de la RLM y el modelo de la RP de acuerdo al tiempo transcurrido.

Los errores para el modelo de RP estan definidos por el Error Cuadrático Medio(MSE) y el Error Absoluto Medio (MAE) tal cual se muestran en la Tabla 2 a nivel de Arequipa y Perú.

	Arequipa		Perú	
	MSE	MAE	MSE	MAE
Casos	11182.869	9406.539	123441.655	62605.517
Muertes	311.465	221.531	5255.410	2473.100

Table 2: MSE y MAE para modelos de Aprendizaje

La implementación está disponible en github [3].

5 Conclusiones

La predicción realizada por RP presenta valores no claros puesto que algunos son negativos; además que el margen de error es elevado es probable que ocurra esto por los vacios mostrados en el conjunto de datos. En cuanto al modelo de RLM los valores que se muestran indican un nivel de incremento de casos y muertes a nivel nacional y regional. También se visualiza en el ploteo del historial de costo que presentan un comportamiento ideal porque se muestra una convergencia después de la cantidad de iteraciones.

References

Coronavirus (COVID-19) Visualization Prediction from Kaggle, https://www.kaggle.com/therealcyberlord/

Andrew Ng. Lecture Notes.

[3] https://github.com/anabeL19/TopicosT-ai-/blob/main/parcial/Proyect-covid19.ipynb

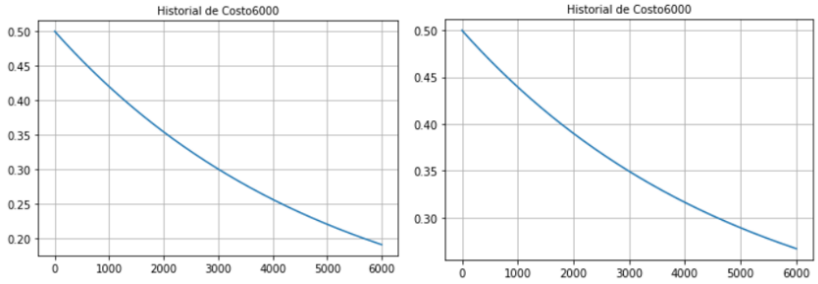

Figure 2: Historial de Costo de Arequipa

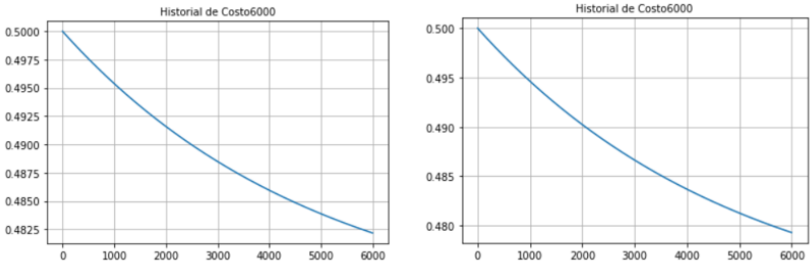

Figure 3: Historial de Costo del Perú

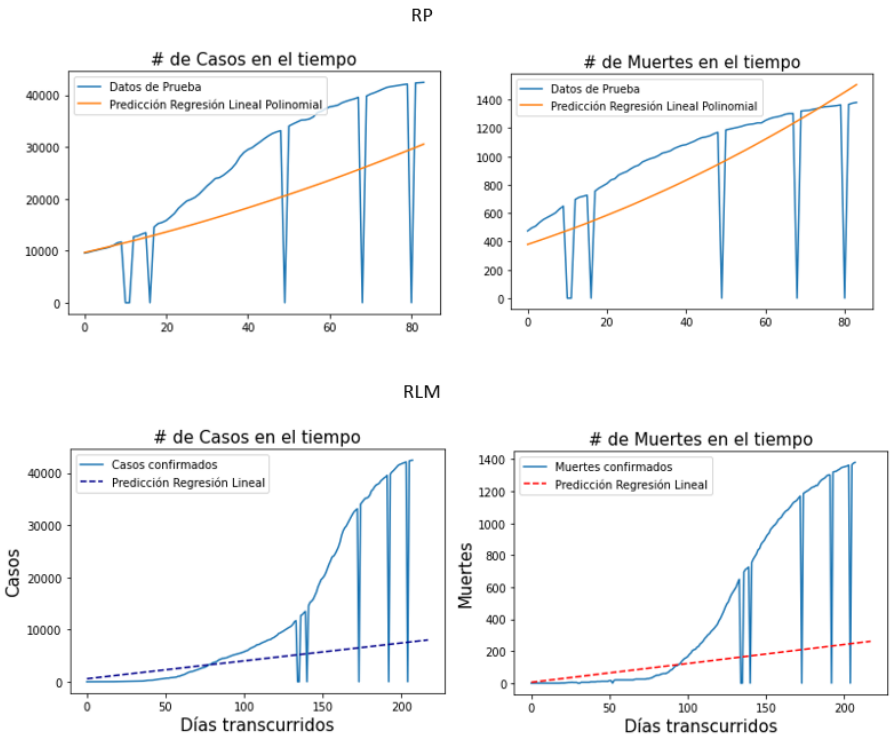

Figure 4: Historial de Costo de Perú