Enfermedad Dengue en la República Argentina

Aplicación de Machine Learning para su estudio y predicción de casos

Fresca, Lorenzo
Ingeniería Industrial - UTN.FRBA

Mojico, Ailín
Ingeniería Industrial - UTN.FRBA

Rosselló, Matías
Ingeniería Industrial - UTN.FRBA

***Abstract* — El interés por descubrir patrones y poder desarrollar un modelo de predicción de casos de Dengue en la República Argentina, nos incentivó a desarrollar el presente estudio analitico de datos.**

El objetivo es poder predecir la cantidad de casos de dengue que puedan contabilizarse en cada provincia argentina, según determinadas variables meteorológicas y basándonos en los casos históricos.

Se aspira a poder generar una herramienta útil para el sistema de salud y que se pueda aplicar para programar campañas de prevención y fumigación en las zonas del país que posean el mayor índice de casos.

***Keywords* —** *Dengue, Humedad, Precipitaciones, Regresión dengue, Predicción casos dengue, Machine Learning dengue, SVR, KNN*

I. INTRODUCCIÓN

El virus del dengue se transmite por mosquitos hembra principalmente de la especie Aedes aegypti y, en menor grado, por medio de la A. albopictus. Dichos mosquitos transmiten también la fiebre amarilla y la infección por el virus de Zika.

El dengue se presenta en los climas tropicales y subtropicales de todo el planeta, sobre todo en las zonas urbanas y semiurbanas. En las últimas décadas ha aumentado enormemente la incidencia de esta enfermedad a nivel mundial. Alrededor de la mitad de la población del mundo corre el riesgo de contraer esta afección [1]. De hecho, en la República Argentina, en el último año fue considerada como epidemia, por la cantidad de casos registrados en el país, afectando principalmente al centro-norte del mismo.

En el presente mapa de calor se pueden observar las principales ciudades del país en función de los casos autóctonos [2].

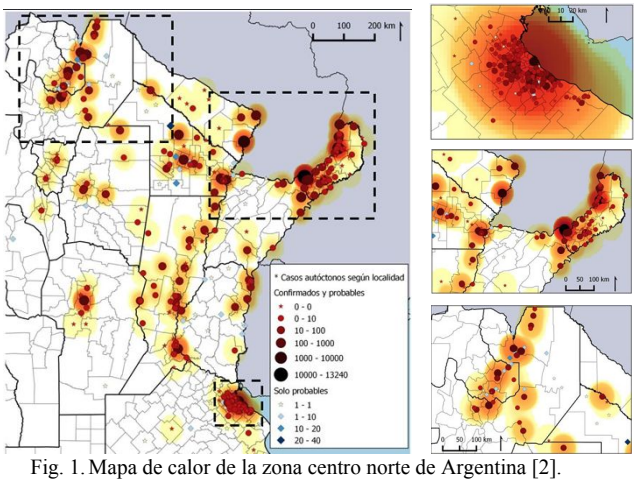

Fig. 1. Mapa de calor de la zona centro norte de Argentina [2].

Sabiendo esto, nos surgió el interés de poder dimensionar los volúmenes de los casos y las variables climáticas que los condicionan.

Realizamos el análisis con registros de dengue de 2018 y 2019 y con un dataset con el promedio de datos meteorológicos de los últimos 30 años, ambos aportados por el Gobierno de la Nación. No conformes con los resultados obtenidos, ya que eran escasos los registros de la enfermedad en cuestión, y el dataset meteorológico poco preciso, debido a que no variaba año a año, recurrimos a la plataforma de Trámites a Distancia y solicitamos la información pública que deseábamos.

Así descubrimos que el Ministerio de Salud de la Nación lleva a cabo un registro de los mismos por cada año, semana epidemiológica, rango de edad de la persona contagiada, departamento geográfico de donde reside la misma y provincia, con el fin de reportar de forma semanal los casos en un boletín epidemiológico nacional.

Por otro lado, el Servicio Meteorológico Nacional cuenta con registros de todos los centros de medición de las condiciones climáticas del país.

De ambos entes públicos mencionados, se recibió el acceso a los datasets que recopilan los datos iniciales necesarios para llevar a cabo el análisis de los mismos.

II. DESCRIPCIÓN DE LOS DATASETS

A. Vigilancia_dengue

Casos confirmados de Dengue y Zika correspondientes al Registro del Sistema Nacional de Vigilancia de la Salud del periodo 2018 y 2019 [3]. Aquellos referentes al 2020 fueron solicitados mediante la plataforma Trámites a Distancia [4].

Cada fila, también llamada sample, representa un reporte realizado en una semana determinada del año en una ciudad determinada, indicando la cantidad de casos autóctonos confirmados en ese periodo de tiempo. Ej: Avellaneda - Provincia de BSAS - Sem_23 - De 45 a 64 años - 33 casos

B. Exp_horarios

Por cada estación meteorológica se dispone una medición con una frecuencia de 6 horas, de las siguientes variables: temperatura [ºC], humedad relativa [%] e intensidad de los vientos [Km/hs].

C. Exp_precipitaciones

Se dispone de los valores mensuales de precipitaciones por cada estación[mm] , desde el primero de enero de 2018 hasta los primeros días de octubre del 2020.

D. Exp_observatorios

Se cuenta con la información de la provincia en donde se encuentra la estación con su respectiva altura, longitud y latitud.

La cantidad de samples a disposición en cada dataset fueron resumidos en la Tabla I.

Dataset	Significado samples	Año		
		2018	2019	2020
Vigilancia_dengue	Cantidad de reportes semanales de casos	**679**	**807**	**9.002**
Exp_horarios	Cantidad Mediciones	**97.690**	**98.179**	**48.128**
Exp_precipitaciones	Cantidad de mediciones	**840**	**840**	**840**
Exp_observatorios	Cantidad de centros	**70**		

TABLA I. Dimensiones dataset input

III. ANÁLISIS EXPLORATORIO DE DATOS

Se trabajó con el lenguaje de programación Python, desde la plataforma Google Collaboratory para llevar a cabo el análisis de datos y desarrollar el archivo de trabajo.

El objetivo de dicho estudio fue el de encontrar la naturaleza de los datos a disposición, entender qué influencia tienen entre ellos y si existía algún patrón que los definiera.

A partir de los datasets antes mencionados, se realizó un pre-processing en el cual se eliminaron los registros que contenían valores inválidos o vacíos, y sucesivamente se unieron los tres dataset de contenido meteorológico con un merge, mediante el ID number de cada estación.

En una segunda instancia se eliminaron las features meteorológicas que se consideraron como información no útil para el análisis, tales como: nombre estación, latitud y longitud de las estaciones. El resultado fue una base de datos agrupada por provincia, mes y año con la información de temperatura, precipitación, humedad relativa e intensidad de vientos de cada una de ellas (de aquí en más llamada Clima).

Con respecto al dataset de *Vigilancia_dengue*, se concatenaron las samples, enlazando los años 2018, 2019 y 2020; se agruparon las semanas del año por cada mes (en formato float) y se eliminaron las features de *grupo_edad_id*, *grupo_edad_desc*, los cuales se consideró que no aportaban información al análisis. El resultado fue la tabla denominada Dengue, la cual contenía la información de cantidad de casos, por cada provincia, mes y año.

Para unir las tablas Dengue y Clima se modificó el formato de minúsculas a mayúsculas, para que todos los registros tuvieran igual formato. A raíz de esto, se combinaron las tablas en una nueva, llamada Casos, mediante el nombre de la provincia, mes y año.

En los meses que no había registros de casos, apareció nuevamente las siglas *NaN*, pero ahora es considerado un dato, por lo que se reemplazó por el valor 0.

TABLA II. Recorte tabla output pre-processing

Luego de todo el *pre-processing*, se obtuvo una tabla con 750 filas (samples) y 6 columnas (features). A partir de esta última tabla se realizó el análisis exploratorio de datos (EDA), donde se obtuvo la siguiente información relevante:

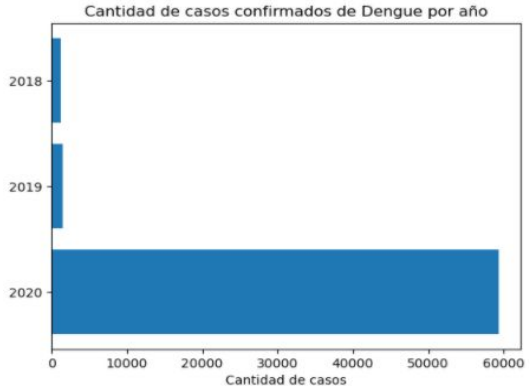

Fig. 2. Gráfico a barras de cantidad de casos según año.

Se contabilizaron la cantidad de casos confirmados de Dengue por año, por lo que se determinó llevar a cabo el análisis con el año 2020, con las provincias con cantidad de casos relevantes. Así, de 24 provincias, se empezó a trabajar con 9. Las mismas se encuentran ordenadas según ranking en la Tabla III.

	Provincia	cantidad_casos
0	CABA	9172
1	TUCUMÁN	7756
2	SALTA	7571
3	MISIONES	5411
4	JUJUY	5235
5	SANTA FE	4680
6	BUENOS AIRES	4574
7	CÓRDOBA	3679
8	CHACO	3370

TABLA III. Top 9 Provincias según la cantidad de casos en 2020.

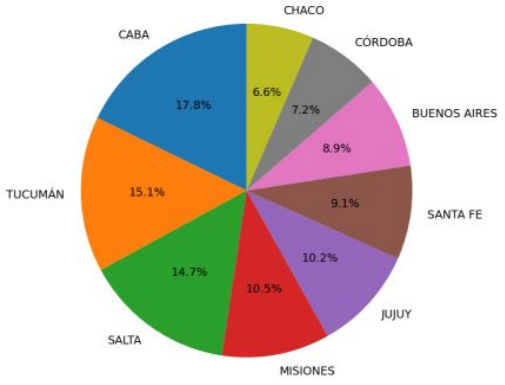

Fig. 3. Gráfico torta de los casos confirmados de Dengue en el año 2020.

Tal como se había anticipado en la introducción, las provincias con mayor cantidad de casos son aquellas ubicadas en la zona norte del país y las provincias cercanas al Río de la plata y Río Paraná.

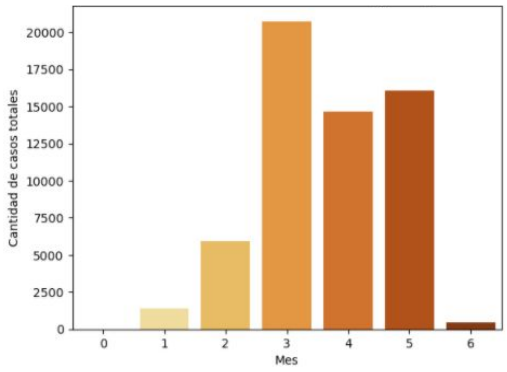

Fig. 4. Histograma de evolución de casos acumulados de las 9 provincias top entre enero (mes 0) y julio (mes 6) del año 2020.

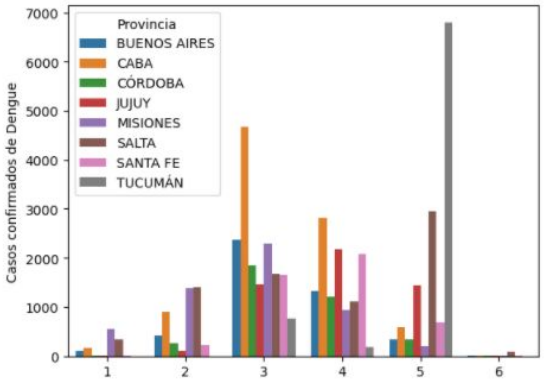

Fig. 5. Histograma de evolución de casos distinguiendo las 9 provincias top entre enero (mes 0) y julio (mes 6) del año 2020.

Con el gráfico de barras agrupado se buscó visualizar la evolución de la cantidad de casos en los primeros 6 meses del año por cada provincia. Podemos observar la subida abrupta en la cantidad de casos en la provincia de Tucumán.

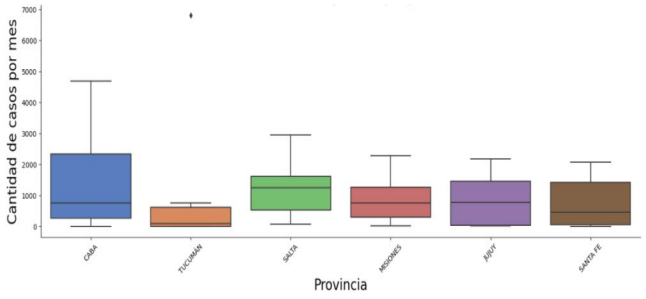

Fig. 6. Box plot de las primeras 6 provincias con más casos en el año 2020.

Para observar la distribución de la cantidad de casos por provincias y meses epidemiológicos, decidimos visualizarlo a través de un diagrama de cajas (box plot).

Entre Marzo y Mayo se concentran más del 85% de los casos reportados. Haciendo foco en las provincias, la conducta de contagio de CABA fue más distribuida respecto a Tucumán, donde a pesar de tener la menor media y variabilidad entre las primeras 6 provincias, en Mayo se presentaron más de 6.500 casos, generando un outlier (punto periférico) en la box de la provincia.

Siguiendo con el análisis, se realizó una visualización de correlación lineal (Pearson) para identificar qué features tuviesen mayor relación directa entre ellas.

	Mes	Temperatura	Precipitacion	Humedad	Vel_vientos	cantidad_casos
Mes	1.000000	-0.700837	-0.464744	0.322507	-0.178209	0.043850
Temperatura	-0.700837	1.000000	0.472183	-0.278081	-0.374547	0.158690
Precipitacion	-0.464744	0.472183	1.000000	0.398157	-0.124899	0.119148
Humedad	0.322507	-0.278081	0.398157	1.000000	-0.018600	0.205906
Vel_vientos	-0.178209	-0.374547	-0.124899	-0.018600	1.000000	-0.231722
cantidad_casos	0.043850	0.158690	0.119148	0.205906	-0.231722	1.000000

TABLA IV. Correlación lineal entre features del dataset procesado

Pudimos observar que no hay una fuerte correlación lineal entre la cantidad de casos con las demás variables, siendo las que mayor relación lineal tienen con la misma es la humedad y velocidad de vientos.

Además, escalaron los datos y se realizó un Pair plot para visualizar la relación entre variables.

Se tomaron las features de temperatura, humedad y cantidad de casos y se observó la relación entre ellas a través de un gráfico de dispersión. Cuanto más grande el círculo, mayor la cantidad de casos.

Entre los 22 y 26ºC se presentaron la mayor cantidad de casos con una humedad promedio del 74%; como también a menores temperaturas y con un aire más seco (humedad del 70%).

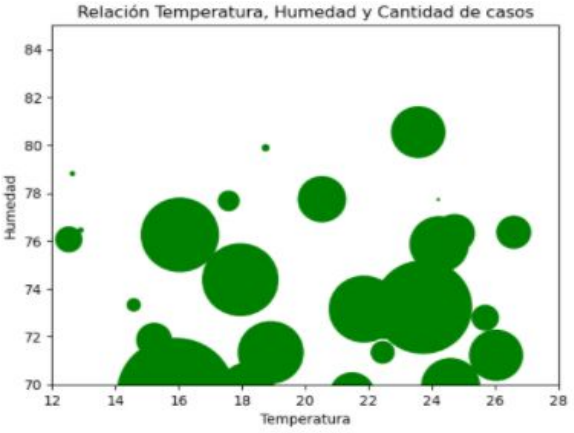

Fig. 7. Histograma de evolución de casos acumulados de las 9 provincias top entre enero (mes 0) y julio (mes 6) del año 2020.

Se partió con la hipótesis de que las variables meteorológicas de cada provincia tuviesen una relación lineal con la cantidad de casos de dengue. Con el presente EDA se dedujo que no necesariamente sea así. Por otro lado, se confirmó que en las primeras 5 provincias se aglomeran casi el 80% de los casos detectados hasta la fecha, en el 2020.

IV. MODELOS DE MACHINE LEARNING

Luego de haber ejecutado el EDA, se procedió al entrenamiento de varios modelos de aprendizaje supervisado. Los mismos se podrían definir como el "(...)subconjunto de Machine Learning donde se generan modelos para predecir el resultado de salida en base a ejemplos históricos de esa variable de salida. Los modelos se construyen a partir de los algoritmos de Machine Learning y características o atributos de los datos de entrenamiento para que podamos predecir el valor utilizando otros valores obtenidos a partir de datos de entrada".

Para desarrollar los modelos de aprendizaje supervisado, procedimos a generar dummies para la features categóricas de provincia. Una vez acondicionado el dataset, se divide el mismo en una porción para entrenamiento y otra para evaluar si el modelo aplicado es eficiente (partición en train y test).

Los métodos que se aplicarán son el de Regresión lineal, Ridge Regression (L1), Lasso (L2), Super Vector Regression (SVR), KNN Regression, Regresión Polinómica y Random Forest Regression.

En todos los casos se calcularán las siguientes métricas: R2 , MSE y MAE. Los mismos servirán para comprender que tan eficiente y representativa es la función resultado de cada modelo.

A. Modelo Regresión Lineal

En este caso, el objetivo es predecir la recta que mejor se ajusta a las variables del dataset. Para hacerlo se mide el error con respecto a los puntos de entrada y el valor "Y" de salida real. El algoritmo deberá minimizar el coste de una función de error cuadrático. Para emplear dicho modelo, precisamos definir el hiper-parámetro y la cantidad de K folds con las cuales dividiremos el training set y luego iteraremos K veces.

Empleamos K=5 y, luego del proceso de Grid Search y del entrenamiento, calculamos las métricas para analizar la eficiencia del modelo

Modelo	R2	MSE	MAE
Linear	0.055411	138142.776964	234.900743

TABLA V. Errores calculados a raíz de testear el modelo de regresión lineal

El valor de MSE en este caso, nos indica que el modelo de Regresión lineal tiene baja precisión a la hora de predecir los próximos valores. El MAE nos indica que, en promedio, la diferencia lineal entre los valores observados del x_train y de aquellos predecidos con el *y_test*, fueron 234 veces más lejanos. Cuantos más lejanos menos útil es el modelo.

El resultado del R-cuadrado nos indica la aptitud del modelo. El mismo puede valer entre menos infinito 1. Por consiguiente, se puede afirmar que el modelo propuesto no mejora la predicción sobre el modelo medio.

B. Modelo de Ridge regression y Lasso

Estas alternativas de aprendizaje consisten en ajustar el modelo incluyendo todos los predictores pero empleando un método que fuerce a que las estimaciones de los coeficientes de regresión tiendan a cero, es decir, que tienda a minimizar la influencia de los predictores menos importantes. En el caso de Ridge regression, se aproxima a cero los coeficientes de los predictores pero sin excluir ninguno. Para el caso de Lasso, se llegan a excluir predictores.

Tal como realizamos en el modelo anterior, se define el objeto para el Grid Search y entrenamos el train escalado. El mejor estimador es alpha=1 . Se utiliza la porción del dataset apartada para validar el modelo y verificar el resultado del mismo.

Modelo	R2	MSE	MAE
Ridge Regression	0.060763	137360.102330	232.822357
Lasso	0.062751	137069.401930	231.725174

TABLA VI. Errores calculados a raíz de testear el modelo de Ridge regression y Lasso.

Los valores obtenidos para R-cuadrado, MSE y MAE no difieren mucho del modelo de Regresión lineal.

Frente a esto, se explorarán otros modelos que se adapten mejor a relaciones no lineales con el fin de poder compararlos con los ya obtenidos.

C. Modelo Super Vector Regression (SVR)

Este algoritmo se basa en buscar la curva o hiperplano que modele la tendencia de los datos de entrenamiento y según ella predecir cualquier dato en el futuro. Esta curva siempre viene acompañada con un rango (máximo margen), tanto del lado positivo como en el negativo, el cual tiene el mismo comportamiento o forma de la curva. Todos los datos que se encuentren fuera del rango son considerados errores por lo que es necesario calcular la distancia entre el mismo y los rangos. Esta distancia lleva por nombre epsilon y afecta la ecuación final del modelo.

Realizamos Grid Search, luego entrenamos con el train escalado y corroboramos la eficacia con la porción definida por el test.

Modelo	R2	MSE	MAE
SVR	0.317942	99748.575227	164.337727

TABLA VII. Errores calculados a raíz de testear el modelo de SVR

Los valores obtenidos para los indicadores que empleamos para evaluar qué tan apto es el modelo para predecir, en este caso, fueron 6 veces mejores en el caso del R-cuadrado; 1,3 veces mejores respecto al MSE. Aun frente a este mejora, en términos absolutos, el modelo de SVR tiene un error de predicción elevado.

D. Modelo de K Nearest Neighbours (KNN) Regression

Este algoritmo identifica los k vecinos de cada punto mediante la distancia euclidiana y considerará el valor que toma la etiqueta para cada uno de ellos y devolverá como predicción el valor medio de dichos valores.

En nuestro caso particular, para definir el valor de K se aplicó un proceso de iteración. Desde K=0 hasta K=15. El valor que mejor resultado trajo fue K=2.

Con el parámetro K definido, se entrenó el modelo y luego se testeó calculando, a raíz de esto el MSE, el MAE y el R2 .

Modelo	R2	MSE	MAE
KNN (K=2)	0.320381	99391.862500	95.441667

TABLA VIII. Errores calculados a raíz de testear el modelo de KNN Regression.

Con el presente modelo, se obtuvieron errores relativamente más pequeños y un R-cuadrado, levemente mejor.

E. Modelo Random Forest Regression

El Random Forest Regressor o bosque aleatorio es un algoritmo de aprendizaje supervisado que utiliza el método de aprendizaje por conjuntos para la regresión. Un bosque aleatorio funciona construyendo varios árboles de decisión durante el tiempo de entrenamiento y generando la media de las clases como la predicción de todos los árboles.

Dicho modelo fue empleado para observar si este algoritmo podría devolver mejores resultados de eficiencia y errores más bajos dado que hasta el momento, el que mejor predijo arrojaba un valor de R2 de 0,32.

Utilizando la misma metodología explicada en los casos anteriores, se entrena y luego se evalúa. Previo a este proceso fue necesario definir la cantidad de estimadores (definimos 1000) y un valor de random_state=42.

Aquí de seguido se pueden observar los resultados obtenidos:

Modelo	R2	MSE	MAE
Random Forest Regressor	0.241314	110955.225157	124.674517

TABLA IX. Errores calculados a raíz de testear el modelo Random Forest Regression

Comparativamente, mejoró el error MAE respecto a la media de los otros modelos (ya que devolvió un valor por debajo de 125) pero el error MSE y el R-cuadrado siguen siendo lejanos a un valor aceptable para considerarse un algoritmo eficiente para nuestro objetivo.

V. CONCLUSIONES

Partiendo de la hipótesis de que existía una relación directa entre las variables meteorológicas y la cantidad de casos reportados de dengue, incursionamos en el presente análisis y desarrollo de modelos de aprendizaje supervisado de machine learning.

En la etapa de pre-processing, vimos que no disponíamos de suficientes valores para empezar con el análisis exploratorio de datos, motivo por el cual iniciamos un expediente y solicitamos al ente pertinente un volumen mayor, actualizado y más preciso.

Sucesivamente a eso, en la etapa análisis de los datos (EDA), pudimos observar que los primeros seis meses del año son los que engloban más del 98% de los casos reportados en un año, con un pico entre los meses de marzo y mayo.

A su vez, de los tres años en estudio, el 2020 tuvo un volumen de casos tal que eran despreciables los reportados en 2018 y 2019.

Se tomó la decisión de acotar el análisis al corriente año y adicionalmente a las 9 provincias con más casos. Las mismas, en orden descendente son: CABA, Tucumán, Salta, Misiones, Jujuy, Santa Fe, Buenos Aires, Córdoba y Chaco. Se confirmó que en las primeras 5 se aglomeran casi el 80% de los casos detectados hasta la fecha, en el 2020.

Entre las variables meteorológicas analizadas (precipitaciones, humedad, temperatura y velocidad de viento) no se encontraron fuertes correlaciones lineales respecto a la cantidad de casos. Aún así, no quiere decir que no haya una relación definida por alguna función específica.

Por tal razón, desarrollamos siete algoritmos distintos de aprendizaje supervisado.

En el siguiente cuadro resumen, se tabularon los resultados obtenidos por cada modelo.

Modelo	R2	RMSE	MAE
KNN (K=2)	0.320381	99391.862500	95.441667
SVR	0.317942	99748.575227	164.337727
Random Forest Regressor	0.241314	110955.225157	124.674517
Lasso	0.062751	137069.401930	231.725174
Ridge Regression	0.060763	137360.102330	232.822357
Linear	0.056678	137957.510123	234.022931
Polynomial Regression	-3.246194	788.029767	620990.913268

TABLA X. Resumen de los resultados de cada modelo empleado.

Se puede concluir a primera vista que, el modelo KNN con K=2, resultó ser el algoritmo de aprendizaje más preciso y con menores errores de predicción.

Además, el modelo SVR presentó resultados similares de precisión, pero con mayores errores respecto al KNN. El algoritmo de regresión Random Forest fue el tercer modelo con mejores resultados, diferenciándose del SVR en tener un menor error medio absoluto, pero un mayor error cuadrático medio.

El resto de los modelos mostraron, en base a los datos empleados como input, no ser útiles para la predicción de los casos dado que el error cuadrático (R2) fue muy bajo (menos de 0,07).

En el caso particular del Polynomial Regression, el hecho que nos haya dado un valor negativo el R2 , significa que el modelo es peor que predecir la media.

Respecto a nuestra hipótesis inicial, podemos concluir que las variables tomadas en cuenta no son suficientes para relacionar con una aptitud elevada y un error reducido.

Para seguir desarrollando el estudio de esta problemática que afecta múltiples países, sería conveniente incorporar otros factores tales como la densidad poblacional, flora y fauna de cada región geográfica y mayores cantidades de registros.

REFERENCIAS

[1] (2020, junio 24). Dengue y dengue grave - World Health Organization. Se recuperó el 05 de noviembre de 2020. https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue

[2] (2016, mayo 26). Solución colaborativa contra el dengue – Agencia TSS. Se recuperó el 13 de noviembre de 2020. https://www.unsam.edu.ar/tss/solucion-colaborativa-contra-el-dengue/

[3] (2018, septiembre 13). Vigilancia de las enfermedades por virus ... - Datos Argentina. Se recuperó el 12 de noviembre de 2020 https://datos.gob.ar/dataset/salud-vigilancia-enfermedades-por-virus-dengue-zika

[4] (n.d.). Trámites a Distancia. Se recuperó el 13 noviembre de 2020. https://tramitesadistancia.gob.ar/

[5] Jake VanderPlas, “Python Data Science Handbook, essential tools for working with data”. pp. 428-429.