PROPUESTA DE ENTREGABLES DATATON MINSA 2023

INTEGRANTES:

Ballon Peralta Christopher Irvin, 45375156, Físico

Hernández Rodriguez Hernán Eder, 70810968, Matemático

Neyra Salas Ruben Cornelio, 09456722, Ing. Informático

Ballon Peralta Christian Joseph, 70391306, Psicólogo

MODELO PREDICTIVO

DESCRIPCIÓN DEL PROBLEMA: No existe un modelo de predicción actualizado que determine con eficacia la propagación del dengue en el país.

OBJETIVO: Estimar el número de casos de dengue por zonas, considerando el clima, grupo etareo, la frecuencia diaria de casos de dengue.

CONTEXTO: En el año 2023 ha habido un aumento considerable de casos de dengue en comparación con años anteriores.

PREGUNTA: ¿Cuántos casos de dengue habrá en un lugar determinado y en un momento determinado?

POR QUÉ ES IMPORTANTE ESTA PREGUNTA: Porque nos permitirá tomar medidas preventivas y así reducir el impacto causado por el dengue en nuestra sociedad.

DATASETS UTILIZADOS:

NOMBRE	TIPO DE ARCHIVO	NÚMERO DE FILAS	NÚMERO DE COLUMNAS	PESO	ORIGEN
Dengue_2013_2023	csv	603187	10	37.2 MB	Proporcionado por la Dathaton
Lima_Sur	csv	4318	7	160 KB	Fuentes externas
Lima_Norte	csv	4318	7	160 KB	Fuentes externas
Lima_Este	csv	4318	7	160 KB	Fuentes externas
Lima_Oeste	csv	4318	7	160 KB	Fuentes externas

TRANSFORMACIONES REALIZADAS

1. Archivo “dengue_2013_2023.csv”

2. Archivo “Lima_Sur.csv”
   2.1. Creación de una nueva columna:
   Se usó los datos de la columna ‘YEAR’ y de la columna ‘DOY’ para crear una nueva columna llamada ‘Fecha_Meteor’ con el objetivo de que contenga fechas en el formato aaaammdd.
   Por ejemplo, se usó el dato ‘2012’ de la columna ‘YEAR’ junto al dato ‘4’ de la columna ‘DOY’ para almacenar el dato 20120104 (representando la fecha 04 de marzo del 2012) en la nueva columna ‘Fecha_Meteor’

3. Archivo “Lima_Norte.csv”
   3.1. Creación de una nueva columna:

Se usó los datos de la columna 'YEAR' y de la columna 'DOY' para crear una nueva columna llamada 'Fecha_Meteor' con el objetivo de que contenga fechas en el formato aaammdd.
Por ejemplo, se usó el dato '2012' de la columna 'YEAR' junto al dato '4' de la columna 'DOY' para almacenar el dato 20120104 (representando la fecha 04 de marzo del 2012) en la nueva columna 'Fecha_Meteor'

4. Archivo "Lima_Este.csv"
   4.1. Creación de una nueva columna:
   Se usó los datos de la columna 'YEAR' y de la columna 'DOY' para crear una nueva columna llamada 'Fecha_Meteor' con el objetivo de que contenga fechas en el formato aaammdd.
   Por ejemplo, se usó el dato '2012' de la columna 'YEAR' junto al dato '4' de la columna 'DOY' para almacenar el dato 20120104 (representando la fecha 04 de marzo del 2012) en la nueva columna 'Fecha_Meteor'

5. Archivo "Lima_Oeste.csv"
   5.1. Creación de una nueva columna:
   Se usó los datos de la columna 'YEAR' y de la columna 'DOY' para crear una nueva columna llamada 'Fecha_Meteor' con el objetivo de que contenga fechas en el formato aaammdd.
   Por ejemplo, se usó el dato '2012' de la columna 'YEAR' junto al dato '4' de la columna 'DOY' para almacenar el dato 20120104 (representando la fecha 04 de marzo del 2012) en la nueva columna 'Fecha_Meteor'

Análisis de datos

El modelo elegido fue de tipo supervisado, ya que se poseen datos para la variable objetivo, además de las variables independientes.

Se utilizó un árbol de clasificación multinomial, para determinar la caracterización de los participantes en el evento correspondiente al diagnóstico (PREGUNTA: DADAS LAS CARACTERÍSTICAS DEL PACIENTE QUÉ DIAGNÓSTICO ES EL MÁS PROBABLE?)

Se procesó los datos en R, empleando las librerías siguientes:

```
library(rpart)
library(tibble)
library(bitops)
library(rattle)
library(rpart.plot)
install.packages("modeldata")
library(modeldata)

library(caret)
library(dplyr)
library(readr)
library(ggplot2)
library(reshape)
library(knitr)
```

En el caso de las tablas que presentaron información solamente del año 2019, se generó una columna categórica de trimestres. Para el caso de las tablas con datos entre 2019 y 20123 se creó una columna para indicar períodos anuales.

Una vez organizados los datos, se realizó la limpieza y exploración. Al tratarse de tablas de propósito específico, su procesamiento se desarrolló por separado.

Metodología y técnicas utilizadas para desarrollar el modelo.

El modelo elegido fue de tipo supervisado, ya que se poseen datos históricos de los años 2013 a 2023, divididos en conjuntos de entrenamiento y prueba para la validación haciendo énfasis en predecir y validar el número de casos de la última ola de dengue en 2023.

Concretamente utilizamos un modelo predictivo basado en análisis de series temporales, realizamos el entrenamiento con datos de número de casos confirmados por fecha en cada provincia del país, debido a su elevada densidad poblacional, tomamos a Lima en regiones Norte, Sur, Este y Oeste y como conjunto de validación el número de casos de los últimos 30 días, además consideramos como variables regresoras para el modelo, la temperatura máxima de la región en cuestión y precipitación.

Dada la naturaleza del dataset proporcionado, puramente número de casos confirmados y presuntivos por centro de salud y la elevada data histórica, (10 años aproximadamente), es que utilizamos la librería Prophet implementada por Meta para poder escoger variables regresoras que implementen la calidad del modelo a proponer.

En este caso establecimos una periodicidad anual y 30 días que la ventana de datos predichos que planteamos alcanzar, además realizamos la validación calculando el MAE (error absoluto promedio) el cual establece la variación promedio de los datos predichos con los datos verdaderos (número de casos), además del MAPE (error porcentual absoluto promedio) que nos especifica el porcentaje de discrepancia de los valores predichos y verdaderos. Además nos valimos de una matriz de correlación de las variables para poder seleccionar los mejores regresores para el modelo.

Realizamos limpieza y preprocesado de datos cómo eliminar caracteres inválidos, esto nos permitió realizar la integración con las tablas maestras que nos proporcionaron correspondiente a los datos de centros de salud, también con datos meteorológicos obtenidos para cada región.

Describir cómo se evaluó y validó el modelo, incluyendo conjuntos de datos de entrenamiento/ prueba/ validación, técnicas como validación cruzada, etc.

Se conformaron dos conjuntos uno de entrenamiento con datos del número de casos de las fechas hasta el 2023-09-20 y los últimos 30 días como conjunto de validación.

Se calcularon los parámetros MAE y MAPE como se definieron anteriormente luego de comprobar las predicciones con el conjunto de pruebas.

Se obtuvieron los parámetros para dos casos:

Modelo considerando únicamente número de casos:

MAE 30.62 (diferencia promedio de número de casos predichos y verdaderos)

MAPE 8.96 (diferencia porcentual del promedio de número de casos predichos y verdaderos)

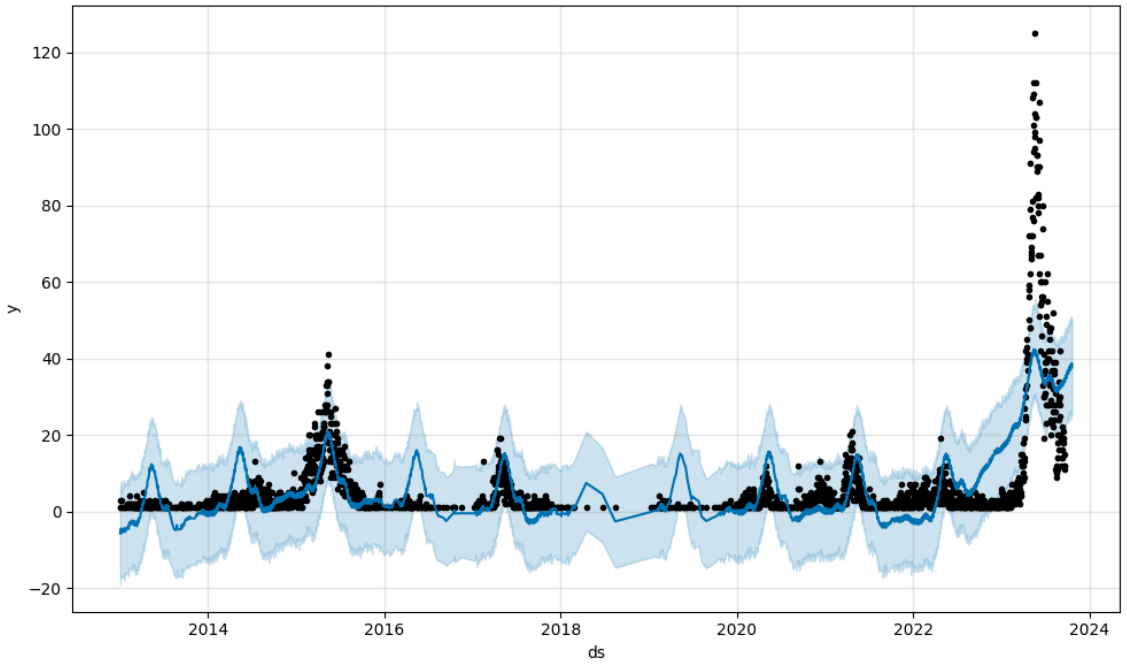
Modelo considerando número de casos, más las variables regresoras: Temperatura Máxima y Precipitación:

MAE 27.56 (diferencia promedio de número de casos predichos y verdaderos)

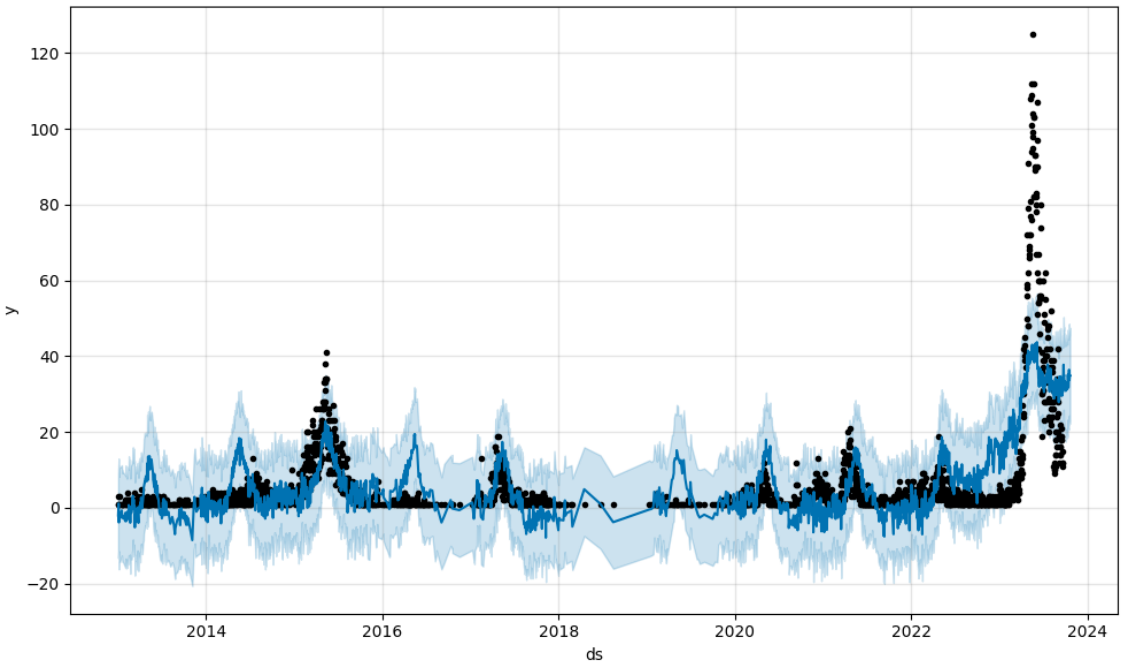
MAPE 8.04 (diferencia porcentual del promedio de número de casos predichos y verdaderos)

Presentar los resultados y predicciones del modelo de una manera visual y fácil de entender

Datos historicos con el numero de casos (puntos en negro) y predicción (curva en azul) considerando solo el numero de casos



Datos historicos con el numero de casos (puntos en negro) y predicción (curva en azul) considerando el numero de casos y las variables temperatura maxima y precipitación



MAE 27.56 (diferencia promedio de número de casos predichos y verdaderos)

MAPE 8.04 (diferencia porcentual del promedio de número de casos predichos y verdaderos)

Incluir posibles mejoras o iteraciones futuras para el modelo.

Como se conversó con el mentor para este reto, en este caso pudieron agregarse otras variables como si la localización contaba con servicio de agua potable y otros datos relacionados a pobreza que permitieran mejorar la estimación del número de casos.

Por otro lado, con un mayor tiempo hubiéramos podido realizar el procesado de datos para realizar estimaciones del número de casos por distrito, en este caso solo se realizó por departamento y considerando Lima en 4 regiones.

Existen otras técnicas que también permiten evaluación de series temporales como redes neuronales y otros algoritmos como xgboost que de evaluarlos se podría considerar como mejora a este trabajo.

Proporcionar recomendaciones claras sobre cómo desplegar/implementar el modelo en producción y monitorear su rendimiento con el tiempo.

Una vez probado el modelo e implementado los puntos de mejora, el código puede fácilmente integrarse a sistemas de evaluación (dashboard) o alerta para poder evaluar el impacto del niño costero al número de casos de dengue en el país.

Entre más datos se vayan considerando para el modelo, su exactitud mejorará, por lo que consideramos que el mantenimiento del mismo requiere poca concentración.

ANEXO:

Repositorio con el código y datasets.

https://github.com/chrisballon/Dathaton-MINSA