Predicting dengue spread

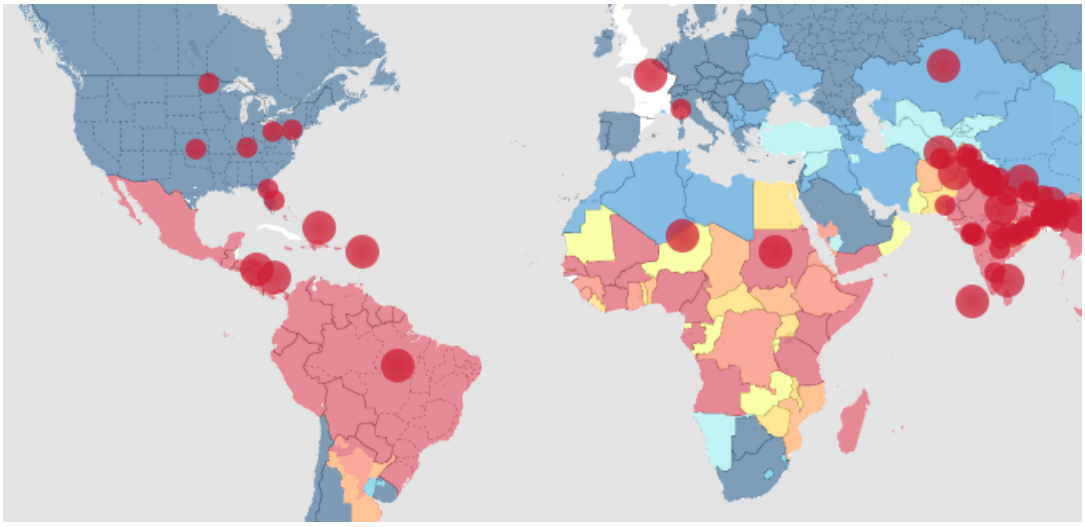
Camilo Bonilla Rafael Santofimio Nicolás Velásquez

December 12, 2022

1 Introduction

Según la OPS (2020) en América Latina cerca de 500 millones de personas están en riesgo de contraer dengue, tanto que en la última década se registraron 16.2 millones de casos. En el continente se encuentran cuatro (4) serotipos (DENV-1, DENV-2, DENV-3 y DEN-V 4) que pueden ocasionar la muerte cuando un individuo contrae más de uno. Es por esto que los gobiernos generan planes de contingencia, vigilancia y control de brotes y epidemias de esta enfermedad. En cuanto al impacto económico, se estima que, para el 2016, la región de las Américas tuvo un costo anual de USD 3 billones asociado al dengue (OPS, 2018).

Figure 1.1: Mapa mundial de dengue



Fuente: Elaborado a partir (DrivenData, 2022)

Por lo tanto, resulta indispensable desarrollar herramientas que permitan predecir el brote y esparcimiento de esta enfermedad con la mayor precisión posible, dado que no solo representa un costo en términos de tratamiento sino también la perdida de bienestar social y destrucción del aparato productivo cuando se presentan muertes. Estas deben propender por enfocar los esfuerzos de los gobiernos en mejorar la vigilancia en zonas apartadas o con altos factores riesgo, y a su vez ser de fácil acceso y uso tanto para las autoridades como cualquier ciudadano.

Es por esto, que el presente trabajo genera una herramienta que permite hacer uso de una serie de variables medioambientales para predecir la cantidad de casos de dengue en San Juan (Puerto Rico) e Iquitos (Perú). Esta comprende no solo el desarrollo modelos con diferentes especificaciones sino ir más allá, al desplegar un aplicativo web que permite capturar parámetros ingresados por un usuario (autoridades de salud y comunidad en general) y con base en estos, realizar predicciones del brote o esparcimiento de esta enfermedad. Es importante aclarar, que el proyecto se enmarca en una competencia en curso promovida por DrivenData, de manera que la base de datos ha sido provista por esta plataforma.

Así entonces, se empieza por entrenar una serie de modelos (ridege, lasso, elascticnet random forest, xgboost, red neuronal y superlearner) de aprendizaje supervisado con distintas especificaciones, teniendo como medida de desempeño el error absoluto medio - MAE, que es el criterio de evaluación establecido en la mencionada competencia. Se encuentra que xgboost es el modelo que mejor se desempeña, siendo el hyperparametro de tasa de aprendizaje el principal generador de la disminución esta metrica.

Con base en los resultados del superlearner, se desarrolla una herramienta con interfaz gráfica que predice el número de casos de dengue por cada ciudad a patir de un rago de fechas ingredas por el usuario. Para esto se emplea el paquete shiny, el cual facilita la creación de aplicaciones web interactivas directamente desde R.

El documento está estructurado así: II. Data, se presenta la fuente, estructura y tipo de información, estadísticas descriptivas y justificación de los datos de entrenamiento y testeo; III. Modelos y resultados, se reporta las variables, modelos, hiperparámetros y principales resultados; IV. Conclusiones y recomendaciones, se describe los principales hallazgos y aportes de los modelos en la predicción de los casos de dengue en San Juan Puerto Rico e Iquitos Perú, y sus respectivas limitaciones; V. Apéndice, contiene las referencias empleadas para el trabajo, y algunas tablas, gráficos y mapas que apoyan la descripción de los datos. Los archivos, tablas, bases de datos y códigos totalmente replicables se pueden obtener en el siguiente link: Final el cual redirige al repositorio en Github y la aplicación desplegada en el siguiente link: Dengue Dashboard

2 Data

Para el desarrollo del modelo, se toman como insumo dos bases de datos a nivel de localidad en San juan (Puerto Rico) e Iquitos(Perú), cada una posee data por semana entre los años 1990 y 2010, lo que representa 1456 observaciones. La base llamada dengulabelstrain.csv contiene los casos de dengue y denguefeaturestrain.csv contiene información proveniente de las siguientes agencias de Estados Unidos: el Centro de Control y Prevención de Enfermedades, la Administración Nacional Atmosférica y Oceánica y el Departamento de Comercio. En la última base se dispone de una serie de mediciones de sensores satelitales y medioambientales para las dos ciudades mencionadas. Finalmente la plataforma www.drivendata.org tiene el archivo submissionformat.csv, el cual contiene para ambas ciudades, las semanas futuras (datos fuera de muestra). Estos periodos corresponde

a 2008/2010 y 2013 para Sanjuan/Iquitos respesctivamnete, predicciopnes que se muestran en la aplicación.

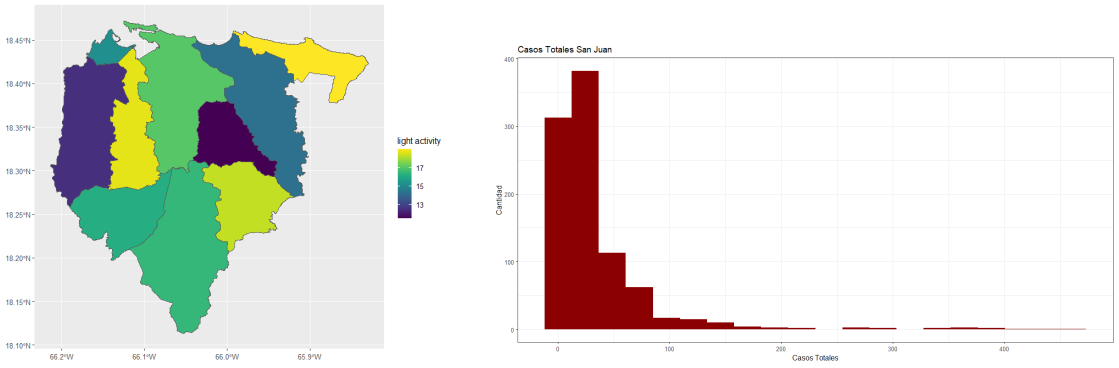
Table 1: Descriptivos

Variable	N	Mean	St. Dev.	Min	Max
year	1,872	2,003.195	6.292	1,990	2,013
weekofyear	1,872	26.489	15.006	1	53
total_cases	1,872	24.805	39.536	0	461
ndvi_ne	1,635	0.139	0.146	−0.463	0.508
ndvi_nw	1,809	0.130	0.125	−0.456	0.649
ndvi_se	1,849	0.205	0.075	−0.016	0.538
ndvi_sw	1,849	0.202	0.086	−0.063	0.546
precipitation_amt_mm	1,857	44.109	42.066	0.000	390.600
reanalysis_air_temp_k	1,860	298.728	1.387	294.554	302.200
reanalysis_avg_temp_k	1,860	299.254	1.273	294.893	303.329
reanalysis_dew_point_temp_k	1,860	295.285	1.528	289.643	298.450
reanalysis_max_air_temp_k	1,860	303.471	3.206	297.800	314.100
reanalysis_min_air_temp_k	1,860	295.725	2.609	286.200	299.900
reanalysis_precip_amt_kg_per_m2	1,860	40.601	44.705	0.000	570.500
reanalysis_relative_humidity_percent	1,860	82.237	7.204	57.787	98.610
reanalysis_sat_precip_amt_mm	1,857	44.109	42.066	0.000	390.600
reanalysis_specific_humidity_g_per_kg	1,860	16.787	1.547	11.716	20.461
reanalysis_tdtr_k	1,860	4.953	3.546	1.357	16.029
station_avg_temp_c	1,817	27.227	1.281	21.400	30.800
station_diur_temp_rng_c	1,817	8.004	2.206	4.043	15.800
station_max_temp_c	1,849	32.471	1.950	26.700	42.200
station_min_temp_c	1,849	22.161	1.613	14.200	26.700
station_precip_mm	1,845	38.202	44.961	0.000	543.300

Fuente: Elaborado a partir de calculos propios.

Como se observa en la tabla 1, la base posee 23 variables numéricas. Algunas de estas tienen datos faltantes, que no supera el 10-15% del total, es por esto que en el procesamiento de los modelos se remplazó por la media de la variable. De igual manera la variable dependiente, total_cases no presenta datos faltantes y esta presente para toda la línea de tiempo de cada una de las ciudades.

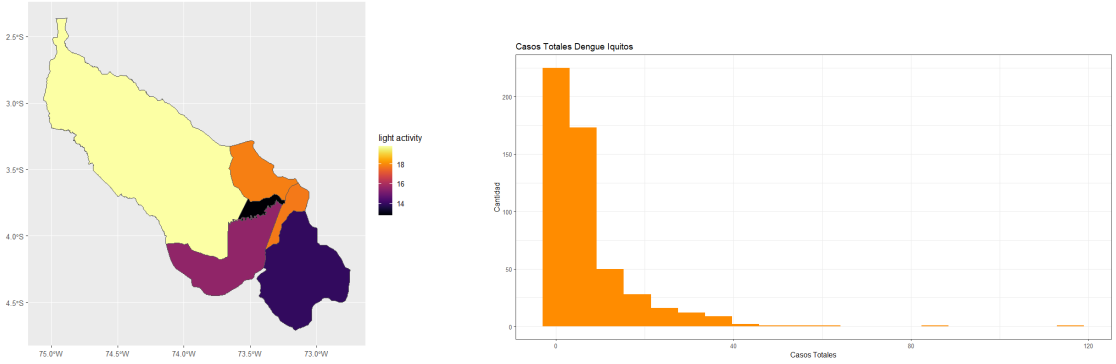
Figure 2.1: Mapa e histograma San Juan



Fuente: Elaborado a partir Maps Streets

En la figura 2.1 y 2.2 se aprecia los mapas con la densidad poblacional de las ciudades, las cuales se harán las predicciones de casos activos de dengue y el histograma respectivo de los casos, se puede apreciar que San Juan a diferencia de Iquitos posee una media de casos más alta. Esto puede deberse a una mayor densidad poblacional a pesar que Iquitos sea región amazónico, con mayor probabilidad de contagio.

Figure 2.2: Mapa e histograma Iquitos



Fuente: Elaborado a partir Maps Streets

Como método de aproximación a la data se generaron correlogramas entre las variables numéricas por ciudad para ver la relación entre estas, esto se puede observar en la figura 1 y 2 del apendice. En enlace al sitio de donde provienen los datos se encuentra a continuación: link o también se pueden conseguir en el repositorio del proyecto predicting-dengue-spread

3 Models and results

Como se ha venido documentado, la propuesta fue elaborada en dos fretes de trabajo. Uno que corresponde a los modelos (back) que realizan las predicciones de los casos de dengue por ciudad/semana y otra que proyecta los resultados en un tablero interactivo del paquete shiny de r-code, este último ideado para los tomadores de decisión política pública en salud. Así entonces esta sección dividen en estos dos componentes.

.1 Modelos

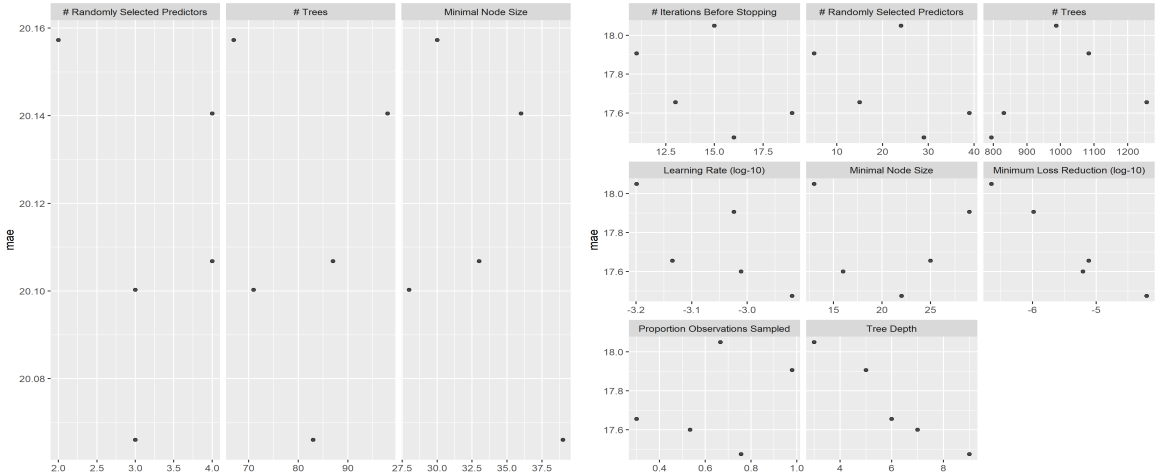
Se diseñaron seis modelos de aprendizaje de máquinas los cuales corresponden a ridge, lasso, elasctic net, random forest, xgboots y una red neuronal, con base en estos se empleó un superlearner, creando así un modelo final con el promedio ponderado óptimo de estos modelos, utilizando el rendimiento de los datos de prueba. El error absoluto medio (MAE por sus siglas en ingles) es la medida de desempeño que se va a tener en cuenta para entrenar cada uno de los modelos, la cual fue la establecida por la competencia , dado que captura la precisión de los modelos en términos de la diferencia de los casos actuales de dengue por ciudad y los predichos, por lo tanto, este debe ser un valor lo más pequeño posible sin llegar a sobreajustar sobre la muestra.

Table 2: Grillas inicial de búsqueda

Hiperparámetros	Min.	Max.	Particiones
Penalizacion	0.0001	0.9	500
Mixtura	0.01	0.99	500
Árboles	1	2000	500
Profundidad	2	16	500
Folds	5	8	3
Tamaño nodo	2	40	500
Predictores	2	80	500
Tasa aprendizaje	0.00001	0.99	500
Reducción pérdida	10^-8	10^-1	100

Para ser esto posible, los datos de entrenamiento se dividieron en dos muestras, una para ajustar la grilla de hiperparametros (15%), y la otra para entrenar los modelos (85%). Así entonces, se optimizó paso a paso el espectro de los hiperparametros finales para cada uno de los modelos, en la tabla 2 se muestra la grilla inicial. Se obtuvo que rigde se estanca en un MAE de 21.38, siendo el modelo con peor desempeño de los 7 construido. Provoca particular atención, que este modelo genera el mismo comportamiento con diferentes penalidades, cuestión que no sucede con rigde, pues con una penalización entre 0.15 y 0.34 presenta una disminución abrupta en el MAE. Por su parte, elastic net mejora su comportamiento con una penalización del 0.4 al 0.2, pero al parecer la mixtura más allá de los datos atípicos (outliers) no muestra cambios significativos.

Figure 3.1: Resultados MAE RF y XGBOOST con grilla optmizada

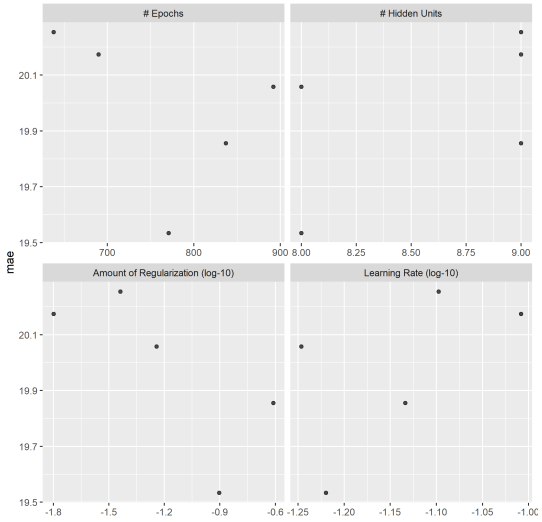


(a) RF (b) XGB

Fuente: Elaborado a partir de (**?**), (**?**), (**?**), (**?**) y (**?**)

En el caso de los modelos más sofisticados, xgboost fue el que menor MAE reportó con 17.5, seguido por random forest y la red neuronal con 19.9. Se observa que el hiperparametro relacionado con la tasa de aprendizaje es el más influyente en los modelos de xgboost y la red, los otros como profundidad del árbol o red, unidades ocultas, o número de arboles no reportan una tendencia clara a disminuir esta métrica. El numero de predictores y el tamaño del nodo son hisperparametros que al disminuir y/o aumentar respectivamente mejoran el desempeño del random forest. Con base en esto, se realizaron numerosas corridas afinando la grilla para cada modelo, centrando especial atención en la tasa aprendizaje, numero de predictores y tamaño del nodo.

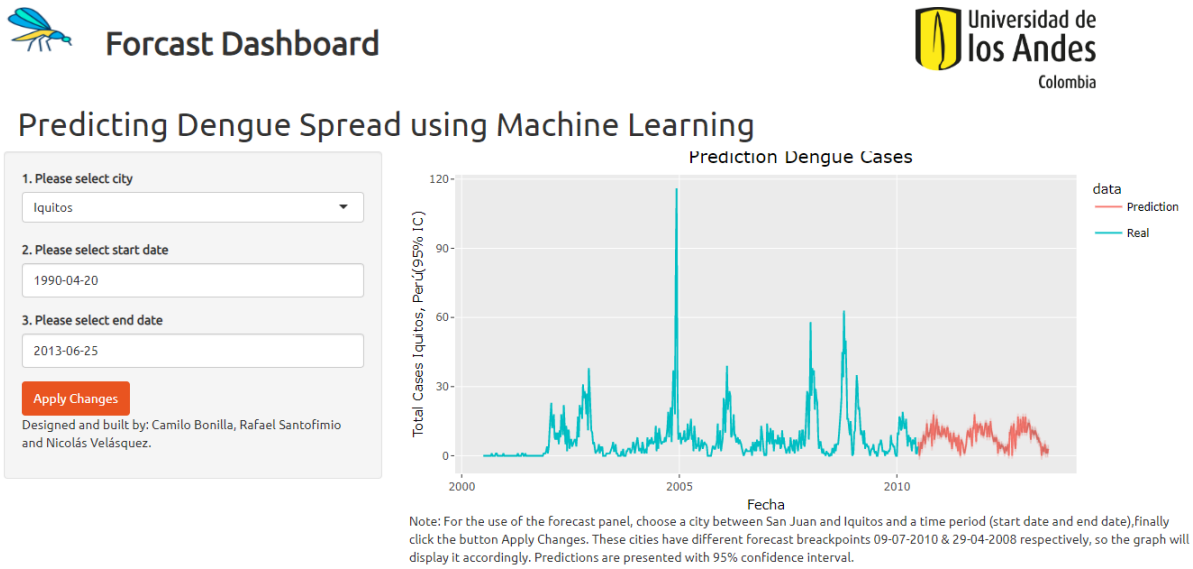
Figure 3.2: Resultados MAE red neuronal con grilla optmizada



.2 Aplicativo web

Tanto las estadísticas descriptivas y los resultados del superlearner fueron insumo principal para construir el tablero con las predicciones de los casos de dengue para cada ciudad, se utilizó el paquete shiny, el cual facilita la creación de aplicaciones web interactivas directamente desde R. Puede alojar aplicaciones independientes en una página web o incrustarlas en documentos de R markdown o crear paneles. Aplicativo está compuesto por una caja que recibe como datos de entrada o configuración, la ciudad y fechas para las cuales se requiere realizar las predicciones. Y una línea de tiempo que proyecta los valores reales y los predichos. A su vez, proyecta un intervalo de confianza del 95% sobre la proyección.

Figure 3.3: Dengue Forcast Dashboard



Aplicativo de predicción de casos de dengue link: Dengue Dashboard

Con este sencillo tablero, se busca equipar a un potencial tomador de decisiones con un elemento útil y parctico que le permita, con anticipación, saber cuántos casos o brotes de dengue puede haber en una locación y momento determinados, lo cual le brindaría información que puede ser útil a la hora de prevenir y controlar el dengue entre la población. De igual manera esta herramienta es pensada para anticipar choques en la población vulnerable por los creadores y diseñadores de políticas públicas, y optimizar la destinación de recursos ya sean jornadas de vacunación, cuarentenas y gasto en medicamentos según los casos predichos.

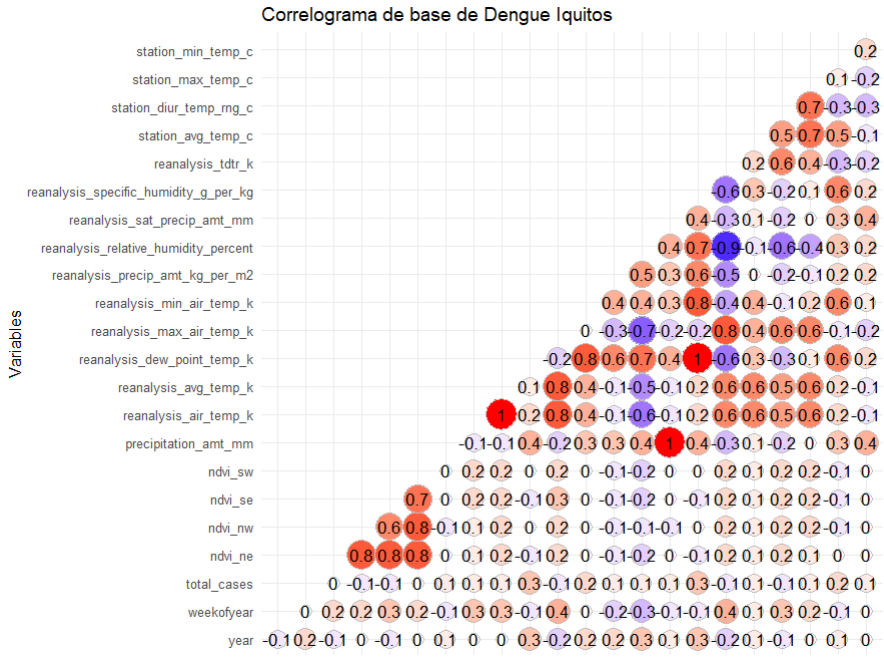
4 Conclusions and recommendations

A la luz de los resultados se encuentra que los modelos diseñados, si bien llegan un MAE del 16.4, este sigue siendo muy alto, lo que puede provocar pedidas potenciales en los tomadores de decisión de política en salud, a la hora de planear estrategias de prevención o en su defecto de mitigación de dengue. Modelos como ridge no son lo sufriente buenos a la hora de adaptarse a los cambios a través de tiempo teniendo como insumos variables medioambientales. Por otro lado, xgboost, random forest y la red neuronal disminuye abruptamente el error absoluto medio cuando se explora los hiperparametros de tasa de aprendizaje, numero de nodos y arboles según aplique. Probar nuevas configuraciones y la introducción de otras variables ambientales podrían mejorar el desempeño de estos modelos. La estacionalidad de la aparición de los casos de dengue no es capturada por superlearner, el tablero construido para proyectar ambas series de tiempo deja en evidencia esta limitación, pues el modelo no predice correctamente los picos de esta enfermedad. Una posible razón de esto es que nuestras características no miran lo suficientemente lejos en el pasado, es decir, se predice los casos de dengue al mismo tiempo que se está midiendo variables como la precipitación. Esto debido a que el ciclo de vida del mosquito que provoca enfermedad depende del ciclo del agua, por lo que es indispensable tener en cuenta tanto la vida de un

misquito como el tiempo entre la infección y los síntomas al modelar el dengue. Esta es una vía crítica para explorar al mejorar este modelo. La herramienta como medio para tomar decisiones en la gestión, monitoreo y control de la enfermedad resulta útil y práctica, no obstante el éxitos de la misma penderá la precisión con que se construya los modelos que están detrás de la misma.

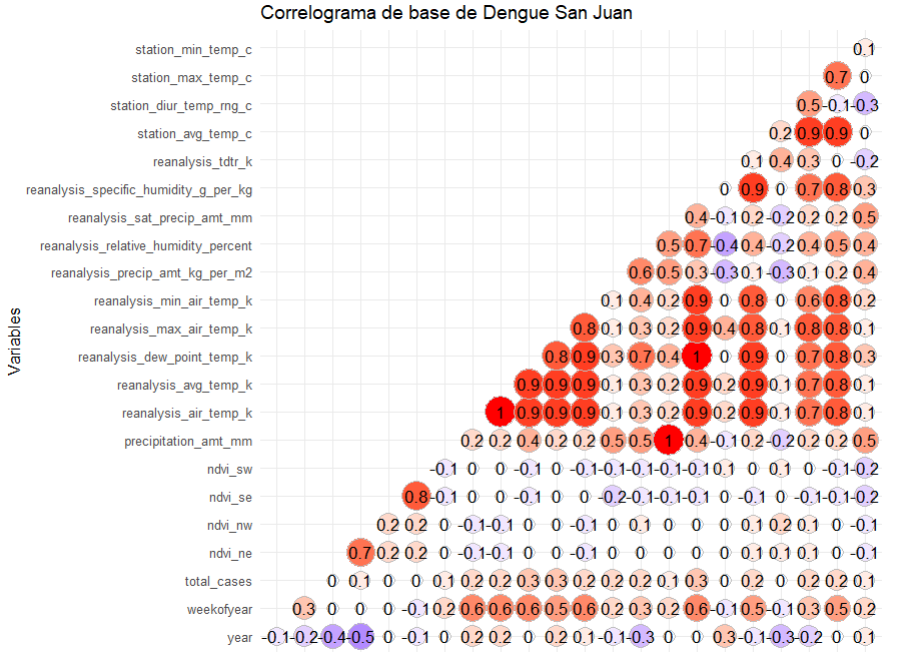
5 Appendix

Figure 5.1: Correlograma Iquitos



Fuente: Elaborado a partir de calculos propios.

Figure 5.2: Correlograma San Juan



Fuente: Elaborado a partir de calculos propios.

References

DrivenData (2022). Dengai: Predicting disease spread. *Driven Data ORG*. Disponible en: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/.

OPS (2018). Economic impact of dengue fever in latin america and the caribbean: a systematic review. *OPS*. Disponible en: https://iris.paho.org/handle/10665.2/49454.

OPS (2020). Dengue: datos y estadisticas. *OPS*. Disponible en: https://www.paho.org/es/temas/dengue.