Modelos de Aprendizaje Supervisado para Predecir Dengue y Malaria a partir de Datos Climatológicos

*Análisis de la Influencia del Clima en la Propagación del Dengue y la Malaria Mediante Machine Learning

1st Paul Andre Auccacusi Huanca
Escuela Profesional de Ingeniería Informática y de Sistemas
Universidad Nacional de San Antonio Abad del Cusco
Cusco, Perú
paulauccacusihuanca@gmail.com

***Resumen*—Las enfermedades transmitidas por mosquitos, como el dengue y la malaria, representan una amenaza significativa para la salud pública, especialmente en regiones tropicales y subtropicales. Factores climáticos, como la temperatura, la humedad y la precipitación, influyen directamente en la propagación de estas enfermedades. En este trabajo, se propone el uso de modelos basados en ensambles y modelos basados en máquinas de soporte además de técnicas como Cross-Correlation para la predicción de casos de dengue y malaria a partir de variables climatológicas desde el 2000 hasta el 2022. Se recopilaron datos históricos de casos epidemiológicos y condiciones climáticas, que fueron procesados y utilizados para entrenar y comparar el desempeño de los modelos. Se utilizaron los datos de 2021-2022 para la validación, donde el modelo SVR obtuvo el mejor desempeño, obteniendo un MAE de 19.17 y un R2 de 0.91 en Piura-Dengue como uno de sus mejores resultados.**

***Index Terms*—dengue, malaria, predicción, machine learning, salud pública.**

I. Introducción

La malaria y el dengue son enfermedades que presentan un reto significativo para la salud pública en el Perú, afectando a millones de personas en regiones tropicales y subtropicales cada año. Debido a su estrecha relación con factores climáticos, se han desarrollado estudios que emplean técnicas de machine learning para clasificar y predecir su incidencia. En esta investigación, se propone aplicar modelos de aprendizaje supervisado en diversas regiones para predecir los casos de dengue, complementado con un análisis de correlación cruzada (Cross-Correlation) para mejorar el desempeño de los modelos utilizados.

II. Marco teórico

II-A. Dengue

El dengue es una infección vírica que se transmite de los mosquitos a las personas. Es más frecuente en las regiones de climas tropicales y subtropicales [1].

II-B. Malaria

El paludismo (o malaria) es una enfermedad febril aguda provocada por parásitos del género Plasmodium, que se transmiten a las personas a través de la picadura de mosquitos hembra del género Anopheles infectados. Se trata de una enfermedad prevenible y curable [2].

II-C. Métricas de Error

Utilizamos estas métricas para definir el desempeño del modelo.

- MAE: Error absoluto medio, mean absolute error; es una medida de la diferencia entre dos valores, es decir, nos permite saber que tan diferente es el valor predicho y el valor real u observado. Para que un error con valor positivo no cancele a un error con error negativo usamos el valor absoluto de la diferencia. Como nos interesa conocer el comportamiento del error de todas las observaciones y no solamente de una, entonces obtenemos el promedio de los valores absolutos de la diferencia [3].
- MSE: Error medio cuadrado, mean square error; es un caso parecido al anterior, pero en lugar de usar el valor absoluto, se eleva al cuadrado la diferencia [3].
- RMSE: Raiz del error medio cuadrado, root mean square error; es el resultado de sacar la raiz cuadrada al MSE ya que el resultado de este último se da en unidades cuadradas [3].
- R2: R cuadrada, coeficiente de determinación; nos indica que tanta variación tiene la variable dependiente que se puede predecir desde la variable independiente. En otras palabras que tan bien se ajusta el modelo a las observaciones reales que tenemos. Cuando usamos R2 todas las variables independientes que estén en nuestro modelo contribuyen a su valor. El mejor valor posible que tenemos con R2 es 1 y el peor es 0, pudiendo incluso tomar valores negativos lo que evidencia mas la mala calidad del modelo [3].

II-D. Modelos de Aprendizaje Supervisado

El aprendizaje supervisado requiere datos de entrada y salida etiquetados durante la fase de entrenamiento del ciclo de

vida del machine learning; se llama aprendizaje supervisado porque al menos parte de este modelo requiere supervisión humana. La gran mayoría de los datos disponibles son datos brutos sin etiquetar. Por lo general, se requiere la interacción humana para etiquetar con precisión los datos. Naturalmente, puede ser un proceso intensivo en recursos, ya que se necesitan grandes conjuntos de datos etiquetados [4].

II-E. Regresión en Aprendizaje Automático

La regresión en el aprendizaje automático consiste en métodos matemáticos que permiten a los científicos de datos predecir un resultado continuo y basado en el valor de una o más variables predictoras x [5].

II-F. XGBRegressor

Extreme Gradient Boosting (XGBoost) es una biblioteca de código abierto que proporciona una implementación eficiente y eficaz del algoritmo de aumento de gradiente. Poco después de su desarrollo y lanzamiento inicial, XGBoost se convirtió en el método de referencia y, a menudo, en el componente clave para obtener soluciones ganadoras para una variedad de problemas en competencias de aprendizaje automático. Los problemas de modelado predictivo de regresión implican predecir un valor numérico, como una cantidad en dólares o una altura. XGBoost se puede utilizar directamente para el modelado predictivo de regresión [6].

II-G. Support Vector Regression

Una Support Vector Machine (SVM) es un algoritmo de aprendizaje automático supervisado que clasifica los datos al encontrar una línea o hiperplano óptimo que maximice la distancia entre cada clase en un espacio N-dimensional. Las SVM fueron desarrolladas en la década de 1990 por Vladimir N. Vapnik y sus colegas, y publicaron este trabajo en un artículo titulado "Método de vectores de soporte para aproximación de funciones, estimación de regresión y procesamiento de señales en 1995. Las SVM se utilizan comúnmente en problemas de clasificación. Distinguen entre dos clases al encontrar el hiperplano óptimo que maximiza el margen entre los puntos de datos más cercanos de clases opuestas. La cantidad de características en los datos de entrada determina si el hiperplano es una línea en un espacio 2-D o un plano en un espacio n-dimensional. Dado que se pueden encontrar múltiples hiperplanos para diferenciar clases, maximizar el margen entre puntos permite que el algoritmo encuentre el mejor límite de decisión entre clases. Esto, a su vez, le permite generalizar bien a nuevos datos y hacer predicciones de clasificación precisas. Las líneas adyacentes al hiperplano óptimo se conocen como vectores de soporte, ya que estos vectores pasan por los puntos de datos que determinan el margen máximo [7].

Support Vector Regression (SVR) es una técnica de aprendizaje automático para tareas de regresión. Es una variante de Support Vector Machines (SVM) y esta diseñado para predecir números continuos, lo que lo hace adecuado para tareas como pronóstico de series de tiempo, predicción de precios de acciones y más [8].

II-H. Random Forest Regression

Random Forest Regression es una herramienta poderosa en la ciencia de datos, que permite realizar predicciones precisas y analizar conjuntos de datos complejos mediante un algoritmo avanzado de aprendizaje automático. Un modelo de regresión de bosque aleatorio combina múltiples árboles de decisión en un único conjunto. Cada árbol se construye a partir de un subconjunto diferente de los datos y realiza una predicción independiente. El resultado final se determina promediando o tomando un promedio ponderado de las predicciones de todos los árboles [9].

II-I. Grid Search

La búsqueda en cuadrícula es una técnica de optimización que se utiliza para aplicar la fuerza bruta a todas las combinaciones posibles de un conjunto de variables. Básicamente, es uno de los algoritmos de optimización más básicos y simples, pero sigue siendo bastante potente y garantiza encontrar la solución más óptima para su problema. Funciona creando una cuadrícula de todas las combinaciones posibles de valores de parámetros y probando cada combinación para encontrar la mejor. Esta cuadrícula de parámetros se define antes del paso de optimización/búsqueda, de ahí el nombre de búsqueda en cuadrícula [10].

II-J. Cross-Correlation y lead-lag relationships

Nos permite saber como se relacionan entre sí dos señales de series temporales lo que puede ayudarnos a extraer información significativa.

Cross-Correlation mide la similitud entre dos señales de series temporales en función de un desfase temporal aplicado a una de ellas. Nos indica si una señal está adelantada(lead) o rezagada(lag) respecto de la otra y cómo se alinean a lo largo del tiempo.

Matemáticamente:

$$R_{xy}(\Gamma) = \sum_{t} x(t) \cdot (y + \Gamma) \tag{1}$$

Donde:

- $R_{xy}(\Gamma)$ es la función cross-correlation a un desfase(lag) temporal Γ.
- $x(t)$ y $y(t)$ son las dos series temporales.
- Γ es el desfase(lag) temporal.

Los valores que nos devuelve esta función nos da 3 posibles valores, los cuales son:

- Desfases(lags) positivos: Indican que la variable y lidera(leads) a la variable x y predice futuros cambios en esta después de un tiempo Γ.
- Desfases(lags) negativos: Indican que la variable x lidera(leads) a la variable y y predice futuros cambios en esta después de un tiempo Γ.
- Cero: Indica que la relacion entre las series temporales es simultanea.

Si, por ejemplo, vemos una correlación positiva alta a los +5 días, esto sugiere que los movimientos en x tienden a preceder a movimientos similares en y por 5 días o las unidades temporales que se esten usando [11].

III. Trabajos relacionados

III-A. Forecast of Dengue Incidence Using Temperature and Rainfall

Este trabajo utiliza un modelo de regresión multivariable de Poisson de series de tiempo usando el promedio de temperatura y precipitaciones acumuladas durante el periodo de 2000-2010 en Singapore.

Se validó el modelo, pronosticando casos de dengue desde la semana 1 de 2011 hasta la semana 16 de 2012, utilizando únicamente datos meteorológicos.

El modelo predijo con precisión el brote de 2011 con menos de un 3 % de posibilidad de falsa alarma [12].

III-B. Traditional Machine Learning based on Atmospheric Conditions for Prediction of Dengue Presence

Este trabajo utiliza técnicas tradicionales de machine learning sobre los casos semanales registrados en Loreto-Perú desde el 1 de Enero de 2016 hasta el 31 de Enero del 2022, se utilizaron modelos como SVM, Arboles de decisión, Random Forest y AdaBoost y los parámetros definidos para evaluar los modelos son: Exactitud, Precisión, Recall y F-1. Se trata de una clasificación binaria para detectar la presencia de Dengue. Como resultado se obtuvieron valores óptimos de AUC en un rango de 0,818 a 0,996 para el SVM, Random Forest y AdaBoost [13].

III-C. A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico

Este estudio utilizó un modelo de aprendizaje automático multietapa denominado AutoTiC-NN para predecir casos de dengue en México, basándose en factores climáticos. Los resultados demostraron que AutoTiC-NN superó a otros modelos en precisión [14].

III-D. Support Vector Regression for Predicting the Number of Dengue Incidents in DKI Jakarta

En este trabajo se uso Cross-Correlation y Augmented Dickey-Fuller test en conjunto con el modelo de Support Vector Regression(SVR) para predecir el número de casos semanales, en el mismo se mostraton los diferentes valores de cross-correlation con diferentes desfases(lags) y su correlación con la variable a predecir, para el modelo se usaron diferentes hiperparametros de SVR cuyo mejor resultado para RMSE y MAE en test fueron 3,40 y 2,48 respectivamente usando el 95 % de los datos [15].

IV. Proceso Metodológico

IV-A. Dataset

El dataset se compone principalmente de variables climatológicas obtenidos de: https://power.larc.nasa.gov/ y de casos de Dengue y Malaria obtenidos de: https://www.datosabiertos.gob.pe/, para el estudio se consideraron principalmente las tres provincias con más incidencias de casos por enfermedad; los cuales fueron, Maynas(194413), Datem del Marañon(119050), Loreto(100123) y Maynas(67063), Piura(58975), Coronel Portillo(38240) para Malaria y Dengue respectivamente, los casos recopilados son desde el 2000 hasta el 2022, por lo que se extraen datos climatológicos del mismo rango de tiempo.

Para la preparación de los datos climatológicos, se transformaron los registros diarios en valores semanales calculando el promedio de las variables, las cuales son:

- T2M: Temperatura a dos metros.
- PRECTOTCORR: Precipitación.
- RH2M: Humedad relativa a dos metros.
- WS2M: Velocidad del viento a dos metros.

Para la preparación de datos epidemiológicos se sumaron los casos de una misma semana por último unir ambos dataset usando el día como se especifica en los tres primeros pasos del flujo de trabajo mostrado en la figura 1.

Al finalizar este proceso se tienen seis mini-dataset pertenecientes a cada provincia.

IV-B. Cross-Correlation Analysis

Se aplica Cross-Correlation a las variables climatológicas y versiones anteriores de la variable a predecir, de todos los mini-dataset, los mejores desfases(lags) se muestran en el cuadro I:

Enfermedad	Provincia	Variable				
		T2M	PRECTOTCORR	RH2M	WS2M	Número de Casos
Malaria	Maynas	-6	-4	-1	-1	-1
	Datem del Marañon	-9	-6	-9	-9	-1
	Loreto	-9	-3	-9	-1	-1
Dengue	Maynas	-6	-8	-1	-4	-1
	Piura	-9	-7	-6	-7	-1
	Coronel Portillo	-9	-9	-9	-9	-1

Cuadro I
Lags según Cross-Correlation por enfermedad y Provincia

Analizaremos el caso específico de Coronel Portillo-Dengue, sin embargo en el cuadro II, puede ver los resultados de Cross-Correlation de todas las provincias en el código.

Lag	T2M	PRECTOTCORR	RH2M	WS2M	Número de casos
-1	0.11	-0.03	-0.05	-0.14	0.92
-2	0.11	-0.02	-0.05	-0.13	0.81
-3	0.11	-0.00	-0.05	-0.13	0.67
-4	0.12	-0.03	-0.07	-0.13	0.53
-5	0.14	-0.01	-0.09	-0.14	0.42
-6	0.16	-0.00	-0.12	-0.16	0.33
-7	0.19	-0.01	-0.16	-0.17	0.25
-8	0.21	-0.04	-0.19	-0.17	0.20
-9	0.23	-0.06	-0.21	-0.17	0.16

Cuadro II
Correlación según lag para cada variable en Coronel Portillo-Dengue

IV-C. Modelo

Se utilizaron los datos del 2000-2020 como entrenamiento y 2021-2022 como validación para cada mini-dataset, posteriormente se realizo grid-search en cada conjunto analizando diferentes hiperpámetros propios de cada modelo usado,

entre estos, XGBoostRegressor, Support Vector Regression, RandomForestRegressor, los hiperparámetros usados se especifícan en el cuadro III.

Modelo	Hiperparámetro	Valores
XGBoost Regressor	n_estimators learning_rate max_depth	{50, 100, 150} {0.1, 0.08, 0.05} {5, 7, 9}
Support Vector Regression	n_estimators max_depth	{50, 100, 150} {5, 7, 9}
Random Forest Regressor	kernel C gamma	{linear, rbf} {1, 10, 100} {0.1, 0.01}

Cuadro III
VALORES USADOS PARA REALIZAR GRID-SEARCH POR MODELO E HIPERPARÁMETRO

Después de hacer Grid-Seach y escoger el mejor modelo para cada mini-dataset se realizan las predicciones, se puede apreciar el flujo de trabajo en la figura 1.

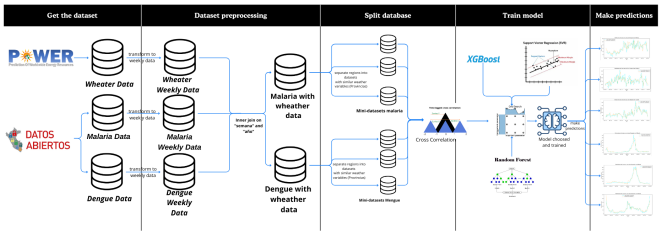

Figura 1. Flujo de trabajo del Proyecto

V. RESULTADOS

Los resultados de nuestras métricas de error se muestran en el cuadro IV.

Enfermedad	Provincia	Métrica			
		MSE	RMSE	MAE	R2
Malaria	Maynas	297.02	17.23	13.06	0.58
	Datem del Marañon	2367.10	48.65	37.80	0.10
	Loreto	930.85	30.50	24.60	0.60
Dengue	Maynas	658.58	25.66	13.89	0.83
	Piura	681.92	26.11	19.17	0.91
	Coronel Portillo	464.60	21.55	15.01	0.75

Cuadro IV
RESULTADOS USANDO EL MEJOR MODELO POR PROVINCIA USANDO LOS DATOS DE 2021-2022 COMO VALIDACIÓN

Los modelo usados se muestran en el cuadro V.

Las figuras 2-6 corresponden a las predicciones para cada mini-dataset.

Enfermedad	Provincia	Métrica	
		Modelo	Hiperparámetros
Malaria	Maynas	Random Forest Regressor	{n_estimators:50, max_depth:10}
	Datem del Marañon	SVR	{kernel:rbf, C:1, gamma:0.01}
	Loreto	SVR	{kernel:rbf, C:100, gamma:0.01}
Dengue	Maynas	XGBRegressor	{n_estimators:50, learning_rate:0.08, max_depth:9}
	Piura	SVR	{kernel:linear, C:100}
	Coronel Portillo	SVR	{kernel: rbf, C: 1, gamma: 0.01}

Cuadro V
MEJOR MODELO CON SUS RESPECTIVOS HIPERPARÁMETROS POR PROVINCIA

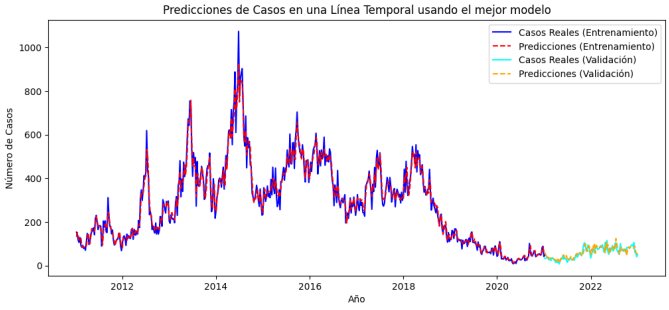

Figura 2. Predicciones para Maynas-Malaria

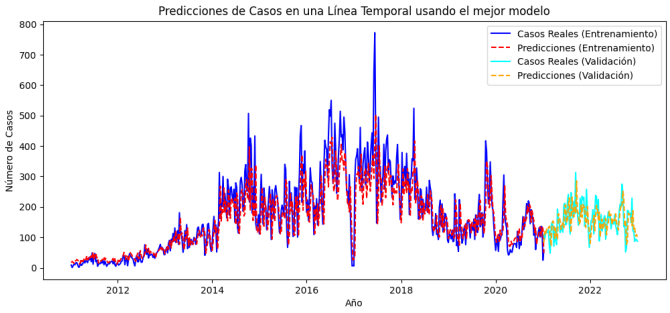

Figura 3. Predicciones para Datem del Marañon-Malaria

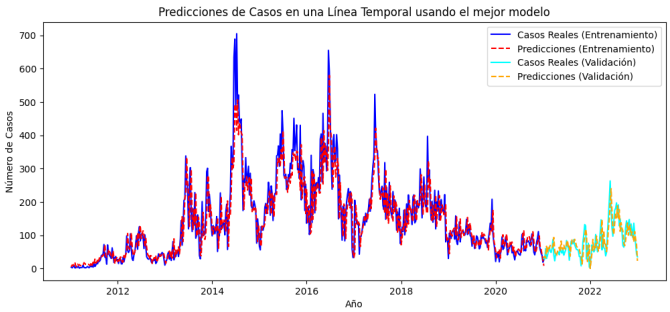

Figura 4. Predicciones para Loreto-Malaria

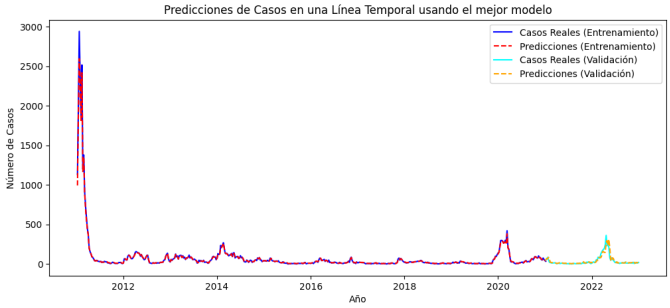

Figura 5. Predicciones para Maynas-Dengue

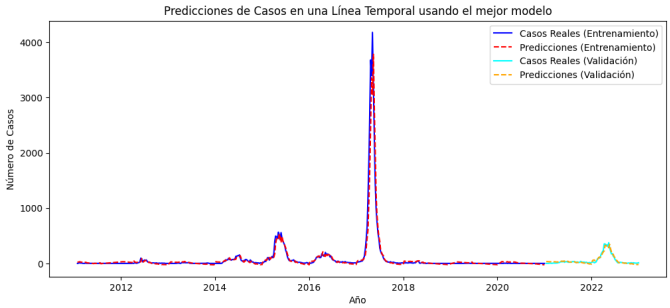

Figura 6. Predicciones para Piura-Dengue

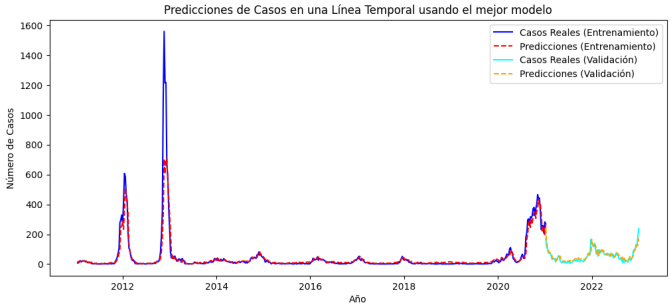

Figura 7. Predicciones para Coronel Portillo-Dengue

VI. Conclusiones

- SVR ha demostrado ser mejor al predecir los casos del 2021-2022 basados en MSE, RMSE, MAE y R2, sin embargo, la diferencia no es abismal comparado con los otros modelos con diferencias mínimas en MAE y RMSE.
- Cross-Correlation ha demostrado ser determinante en la preparación de datos para el entrenamiento de los modelos lo que demuestra que los valores pasados de las variables climatológicas son más significativos para la predicción del número de casos que los valores actuales o simultaneos.
- A partir del análisis de Cross-Correlation, se observa correlación positiva y negativa en las variables conforme se itera entre lags negativos(pasados) en las variables independientes.
- En estre trabajo se uso solo datasets pertenecientes a provincias ya que sus variables climatológicas son parecidas en toda su extensión, para posteriores investigaciones y como se suele hacer en algunas de las fuente bibliográficas, valdría la pena investigar un conjunto de estas provincias o regiones agrupadas por su similitud respecto a sus variables climatológicas.

Referencias

[1] O. M. de la Salud, "Dengue y dengue grave," Disponible en: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue, 2024.

[2] ——, "Paludismo," Disponible en: https://www.who.int/es/news-room/questions-and-answers/item/malaria?gad_source=1, 2024.

[3] N. A. L. Cosio, "Métricas en regresión," Disponible en: https://medium.com/@nicolasarrioja/m%C3%A9tricas-en-regresi%C3%B3n-5e5d4259430b, 2021.

[4] U. Europea, "Aprendizaje supervisado y no supervisado," Disponible en: https://universidadeuropea.com/blog/aprendizaje-supervisado-no-supervisado/, 2022.

[5] V. Kurama, "Regression in machine learning: Definition and examples of different models," Disponible en: https://builtin.com/data-science/regression-machine-learning#:~:text=Regression%20in%20machine%20learning%20consists,use%20in%20predicting%20and%20forecasting., 2024.

[6] J. Brownlee, "Xgboost for regression," Disponible en: https://machinelearningmastery.com/xgboost-for-regression/, 2021.

[7] IBM, "What are svms?" Disponible en: https://www.ibm.com/think/topics/support-vector-machine, 2023.

[8] N. Verma, "An introduction to support vector regression (svr) in machine learning," Disponible en: https://medium.com/@nandiniverma78988/an-introduction-to-support-vector-regression-svr-in-machine-learning-681d541a829a, 2023.

[9] AnalytixLabs, "Random forest regression — how it helps in predictive analytics?" Disponible en: https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4, 2023.

[10] K. Amani, "Understanding grid search as an optimization algorithm in machine learning," Disponible en: https://neurosnap.ai/blog/post/understanding-grid-search-as-an-optimization-algorithm-in-machine-learning/643748be49872f3862f39aed, 2023.

[11] T. Konstantinovsky, "Cross-correlation and coherence in time series analysis: How to uncover relationships between signals," Disponible en: https://medium.com/pythoneers/cross-correlation-and-coherence-in-time-series-analysis-how-to-uncover-relationships-2024.

[12] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, "Forecast of dengue incidence using temperature and rainfall," *Neglected Tropical Diseases*, 2012.

[13] B. S. S. López, D. C. Nolberto, J. A. T. Gutiérrez, and Y. G. López, "Traditional machine learning based on atmospheric conditions for prediction of dengue presence," *Universidad de Lima*, 2023.

[14] A. Appice, Y. R. Gel, I. Iliev, V. Lyubchich, and D. Malerba, "A multi-stage machine learning approach to predict dengue incidence: A case study in mexico," *IEEEXplore*, 2020.

[15] I. N. Tanawia, V. Vitoa, D. Sarwindaa, H. Tasmana, and G. F. Hertono, "Support vector regression for predicting the number of dengue incidents in dki jakarta," *ELSEVIER*, 2020.