MAYO, 2025

MODELO DE PREDICCIÓN DE CASOS DE DENGUE EN EL DISTRITO DE PIURA

ANALÍTICA Y MINERÍA DE DATOS

DOCENTE: WALTER ALEJANDRO SILVA SOTILLO

INTEGRANTES:

***ALBINO HUERTAS, EDER ALBERTO***

***HUILLCA ALARCÓN, EDSON JOEL***

***NIÑO ALCÁNTARA, ULISES EMILIO***

***TOLEDO CHÁVEZ, SALLY DAMARIS***



Contenido

[RESUMEN 2](#_heading=h.estjsbk76qwg)

[INTRODUCCIÓN 3](#_heading=h.dcqyzt2aal7i)

[MARCO TEÓRICO 4](#_heading=h.6bv6hiqb9gde)

[1.](#_heading=h.jkmelvn054vc) Regresión Lineal 4

[2.](#_heading=h.yi77az9sdlrb) Random Forest Regressor 4

[3.](#_heading=h.p2sauje703yn) XGBoost Regressor 5

[4.](#_heading=h.a4pno7o7iovi) SVR (Kernel RBF) 5

[5.](#_heading=h.zhhcjz63nbic) Regresión de Poisson 6

[6.](#_heading=h.z9fdbia3ygjx) Regresión Binomial Negativa 6

[PREPROCESAMIENTO 7](#_heading=h.iw47rlrza4vg)

[MODELOS PROPUESTOS 9](#_heading=h.6we1rnv8ua9x)

[CONCLUSIONES 16](#_heading=h.3ksovtycqvt1)

Índice de figuras

[Figura 1. Valores faltantes por cada variable de la data meteorológica 2016-2023 7](#_heading=h.f72lvzkbjmws)

[Figura 1. Predicción de población del año 2016. 7](#_heading=h.ggg9seih6vgz)

[Figura 2. Matriz de correlación de variables 8](#_heading=h.rh7oolr2kane)

[Figura 2. Matriz de correlación de variables 8](#_heading=h.fcflwb2tqb1j)

[Figura 3. Regresión lineal- Valores reales vs predichos 9](#_heading=h.yewe4sig9o9)

[Figura 4. Random Forest- Valores reales vs predichos 10](#_heading=h.nfpz4i5sisxi)

[Figura 5. XGBoost- Valores reales vs predichos 11](#_heading=h.f4qh0or45f3q)

[Figura 6. SVR- Valores reales vs predichos 12](#_heading=h.y04td4vw9ctj)

[Figura 7. Regresión Poisson - Valores reales vs predichos 13](#_heading=h.yc7cvhqu4ngu)

[Figura 8. Regresión Binomial Negativa - Valores reales vs predichos 14](#_heading=h.ce1z41lv1kic)

[Figura 9. Red Neurona l- Valores reales vs predichos 15](#_heading=h.tb78n6i7cb3o)

Índice de tablas

[Tabla 1. Resultados del modelo de Regresión lineal 9](#_heading=h.akcezc7cnyme)

[Tabla 2. Resultados del modelo Random Forest Regressor 10](#_heading=h.6inod0sv3db7)

[Tabla 3. Resultados del modelo XGBoost Regressor 11](#_heading=h.ogek90jw2wen)

[Tabla 4. Resultados del modelo SVR 12](#_heading=h.wd7zwupigtdo)

[Tabla 5. Resultados del modelo Poisson 13](#_heading=h.mjx0hlbo372k)

[Tabla 6. Resultados del modelo Binomial Negativa 14](#_heading=h.lok4b5qlen3d)

[Tabla 7. Resultados del modelo Red Neuronal 15](#_heading=h.dnera4vikw09)

[Tabla 8. Resultados de los modelos predictivos 16](#_heading=h.irjraaa8f92t)

# RESUMEN

Este informe presenta el desarrollo de un modelo predictivo aplicado a datos epidemiológicos y climáticos en el distrito de Piura, utilizando fuentes como la Red Nacional de Epidemiología (RENACE) y El Servicio de Cambio Climático de Copernicus (C3S); centrado en las variables clave: semana epidemiológica, año, temperatura media, humedad, precipitación acumulada, población estimada y número de casos reportados.

El objetivo fue evaluar técnicas de procesamiento y modelado de datos para predecir la incidencia de dengue en función de variables climáticas y demográficas utilizando técnicas estadísticas para mejorar la calidad y consistencia de la información. Se realizó una limpieza de datos para manejar valores faltantes y outliers, seguida de un análisis y evaluación de múltiples algoritmos de aprendizaje automático.

Se evaluaron múltiples modelos predictivos, tales como el de Regresión Lineal, Random Forest Regressor, XGBoost Regressor, SVR con kernel RBF, Regresión de Poisson, Regresión Binomial Negativa y una Red Neuronal con MLPRegressor.

Tras el proceso de validación, destacó XGBoost como el modelo con mejor rendimiento, alcanzando un error cuadrático medio (RMSE) de 260.01 y un coeficiente de determinación (R²) de 0.48.

# INTRODUCCIÓN

El dengue, una enfermedad viral transmitida por mosquitos del género *Aedes*, representa un desafío significativo para la salud pública, especialmente en regiones tropicales y subtropicales como el distrito de Piura, Perú. Su incidencia se ve influenciada por una compleja interacción de factores climáticos, ambientales y demográficos, lo que subraya la necesidad de desarrollar modelos predictivos robustos para anticipar brotes y facilitar la implementación de medidas preventivas oportunas.

Este informe detalla el proceso de desarrollo y evaluación de diversos modelos de predicción de casos de dengue en el distrito de Piura. Para ello, se integraron datos epidemiológicos históricos proporcionados por la Red Nacional de Epidemiología (RENACE) con información climática obtenida del Servicio de Cambio Climático de Copernicus (C3S). El análisis se centró en variables clave como la semana epidemiológica, el año, la temperatura media, la humedad, la precipitación acumulada y la población estimada del distrito.

El objetivo principal de este estudio fue explorar y comparar la capacidad predictiva de diferentes técnicas de modelado estadístico y aprendizaje automático, incluyendo modelos de regresión lineal, Random Forest Regressor, XGBoost Regressor, Support Vector Regression (SVR), regresión de Poisson, regresión binomial negativa y redes neuronales (MLPRegressor). A través de un riguroso proceso de preprocesamiento de datos, que abarcó la imputación de valores faltantes y la detección de outliers, se buscó garantizar la calidad y consistencia de la información utilizada para el entrenamiento y la evaluación de los modelos. Los resultados obtenidos permitirán identificar la metodología más eficaz para la predicción de la incidencia del dengue en el contexto específico del distrito de Piura, contribuyendo así a la mejora de las estrategias de vigilancia y control de esta importante enfermedad.

# MARCO TEÓRICO

Desde una perspectiva teórica fundamental, los modelos predictivos operan bajo el marco de la teoría del aprendizaje estadístico, que establece los principios matemáticos para inferir funciones de mapeo entre variables predictoras y objetivos a partir de datos muestrales.

Los modelos predictivos se basan en principios estadísticos que permiten inferir relaciones entre variables a partir de datos observados. La teoría de la regresión, desarrollada inicialmente por Gauss y Legendre, establece las bases para modelar relaciones lineales mediante la minimización del error cuadrático. Conceptos como correlación, covarianza y coeficientes de determinación (R²) proveen medidas cuantitativas de la calidad del ajuste del modelo. Estos fundamentos son esenciales para entender cómo las variables predictoras influyen en la variable objetivo y permiten evaluar la significancia estadística de las relaciones encontradas

## Regresión Lineal

La Regresión Lineal es un modelo estadístico que busca establecer una relación lineal entre una variable dependiente y una o más variables independientes.

Ventajas

✔ Fácil interpretación.

✔ Rápido entrenamiento.

Limitaciones

✖ Asume linealidad entre variables.

✖ Sensible a outliers.

## Random Forest Regressor

El Random Forest es un método de ensamble basado en múltiples árboles de decisión, donde cada árbol se entrena con una muestra aleatoria del dataset (bootstrapping) y un subconjunto aleatorio de características. La predicción final es el promedio de las predicciones individuales.

Ventajas

✔ Reduce el sobreajuste (overfitting) vs. un solo árbol.

✔ Maneja bien datos no lineales y relaciones complejas.

Limitaciones

✖ Menos interpretable que modelos lineales.

✖ Mayor costo computacional.

## XGBoost Regressor

XGBoost es un algoritmo de boosting que combina múltiples modelos débiles (generalmente árboles de decisión) en secuencia, donde cada nuevo modelo corrige los errores del anterior. Utiliza:

* Gradient Descent para optimizar la función de pérdida.
* Regularización (L1/L2) para evitar overfitting.

Ventajas

✔ Alta precisión en problemas estructurados.  
✔ Eficiente en velocidad y memoria.

Limitaciones

✖ Requiere ajuste de hiperparámetros.  
✖ Puede sobreajustar si no se regula adecuadamente.

## SVR (Kernel RBF)

El SVR es una variante de las Máquinas de Soporte Vectorial (SVM) para regresión. Busca encontrar una función que minimice el error, manteniendo la mayor cantidad de puntos dentro de un margen.

Ventajas

✔ Efectivo en espacios de alta dimensión.  
✔ Flexible con kernels no lineales.

Limitaciones

✖ Computacionalmente costoso con datasets grandes.  
✖ Sensible a la elección de hiperparámetros.

## Regresión de Poisson

Diseñada para datos de conteo, asume que la variable dependiente sigue una distribución de Poisson.

Ventajas

✔ Ideal para variables discretas no negativas.

✔ Evita predicciones negativas.

Limitaciones

✖ Asume equidispersión, lo que no siempre se cumple

## Regresión Binomial Negativa

Similar a la regresión de Poisson, pero relaja el supuesto de equidispersión al introducir un parámetro de dispersión (α*α*). Útil cuando los datos presentan sobredispersión.

Ventajas

✔ Maneja mejor datos con alta varianza.  
✔ Más robusta que Poisson en casos reales.

Limitaciones

✖ Más compleja computacionalmente.

1. **Redes Neuronales (MLPRegressor)**

Las redes neuronales artificiales (RNA) imitan el funcionamiento del cerebro humano mediante capas de neuronas interconectadas. El MLPRegressor (Perceptrón Multicapa) utiliza función de activación (ReLU, tanh), Backpropagation para ajustar pesos y Optimizadores (Adam, SGD).

Ventajas

✔ Aproxima relaciones no lineales complejas.

✔ Escala bien con grandes volúmenes de datos.

Limitaciones

✖ Requiere gran cantidad de datos.

✖ "Caja negra" (difícil interpretación).

# PREPROCESAMIENTO

**Muestra:** Se tienen datos de 281 semanas epidemiológicas

**Variables:** Semana, año, temperatura, humedad, precipitación, población, casos

**Procedimiento de recojo de datos:** Los datos fueron conseguidos de Red Nacional de Epidemiología (RENACE) y El Servicio de Cambio Climático de Copernicus (C3S).

**Tratamiento y análisis de datos:**

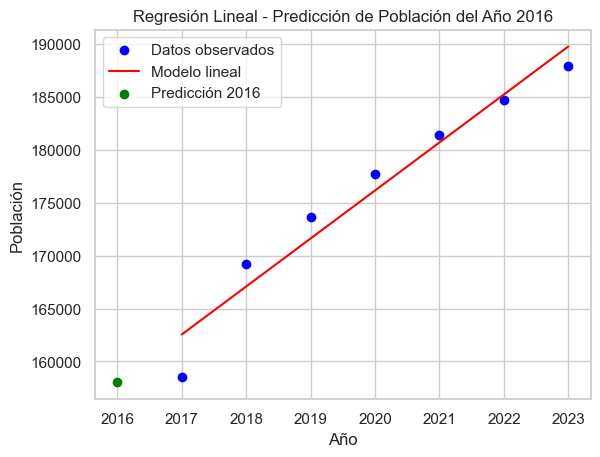
Se realizó un análisis exploratorio de los datos para identificar las variables con mayor cantidad de valores faltantes, su distribución y su relación con otras variables.

Imagen que contiene Texto

El contenido generado por IA puede ser incorrecto.

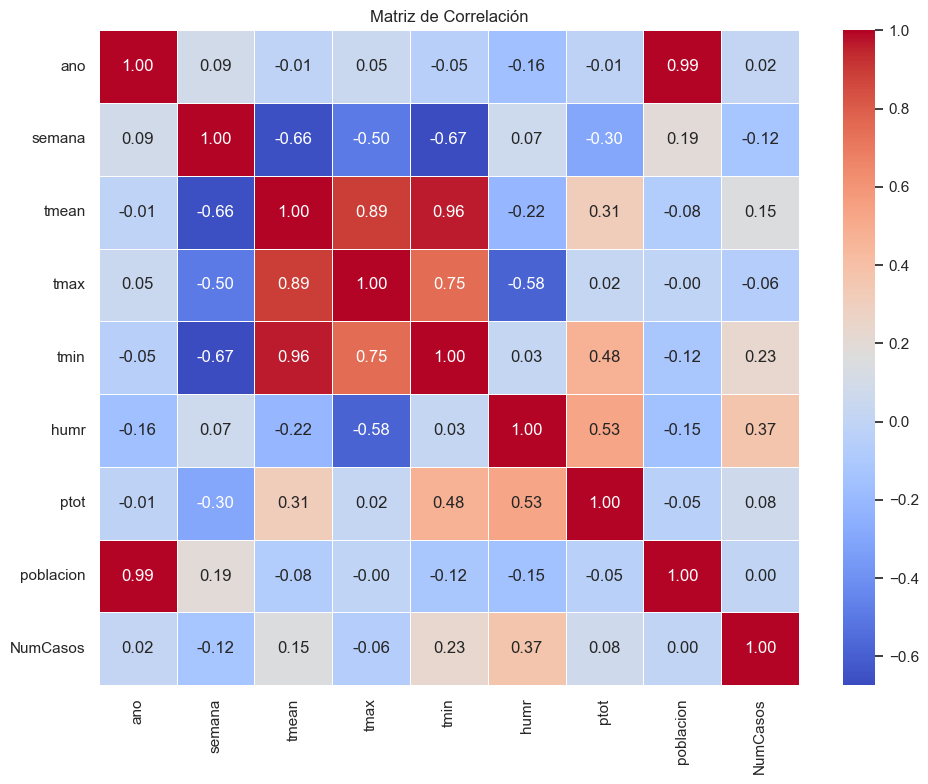
*Figura 1. Valores faltantes por cada variable de la data meteorológica 2016-2023*

Se implementó un procesamiento robusto de datos que incluyó la imputación de valores faltantes mediante técnicas jerarquizadas, junto con corrección de outliers mediante regresión iterativa. Los métodos seleccionados preservan la estructura de los datos y mostraron alta precisión, integrándose finalmente en un dataset completo y consistente.

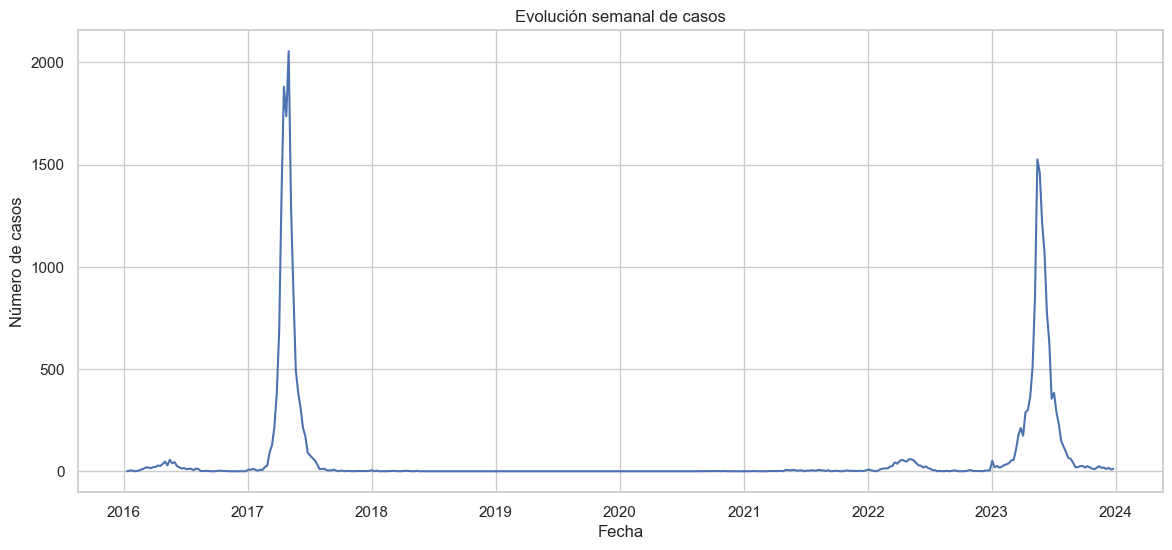


*Figura 2. Predicción de población del año 2016.*

Posteriormente se realizó la matriz de correlación.



*Figura 3. Matriz de correlación de variables*



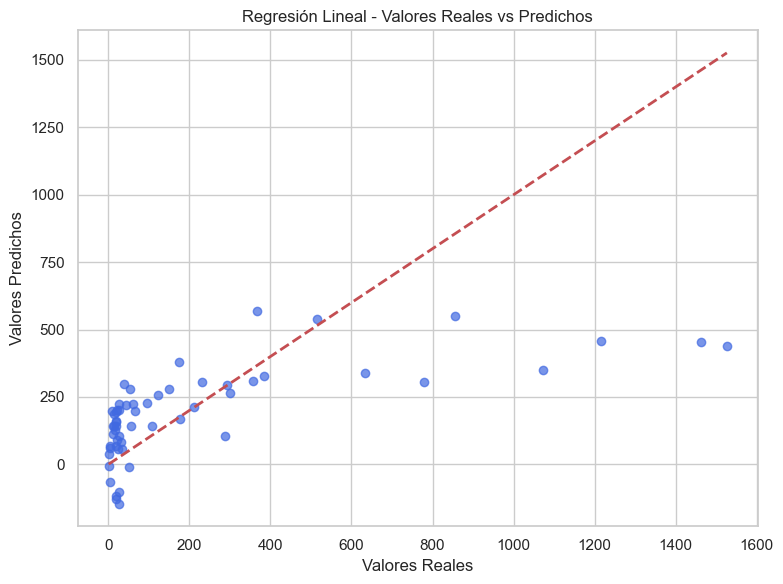
*Figura 4. Evolución semanal de casos.*

# MODELOS PROPUESTOS

Los resultados obtenidos del Modelo de Regresión Lineal fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 177.56 | 281.42 | 0.3909 |

*Tabla 1. Resultados del modelo de Regresión lineal*

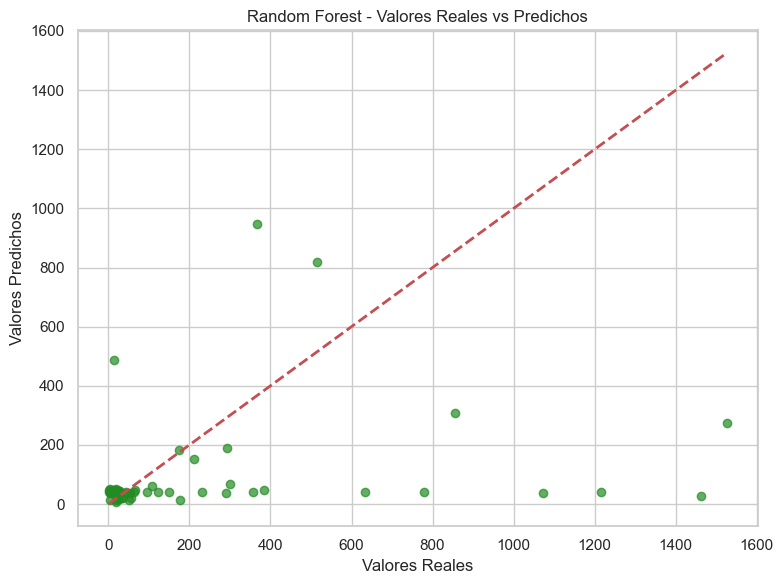


*Figura 5. Regresión lineal- Valores reales vs predichos*

Los resultados obtenidos del Modelo Random Forest Regressor fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 190.21 | 383.58 | -0.1316 |

*Tabla 2. Resultados del modelo Random Forest Regressor*

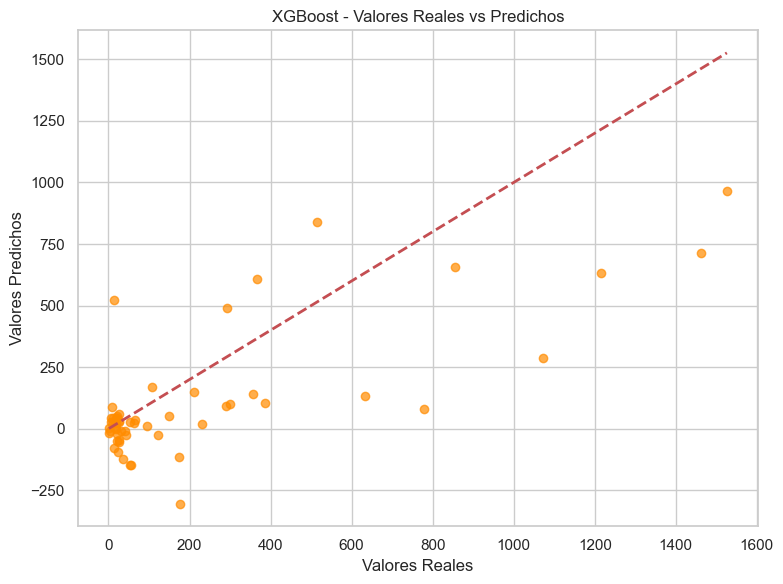


*Figura 6. Random Forest- Valores reales vs predichos*

Los resultados obtenidos del Modelo XGBoost Regressor fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 164.15 | 260.01 | 0.4800 |

*Tabla 3. Resultados del modelo XGBoost Regressor*

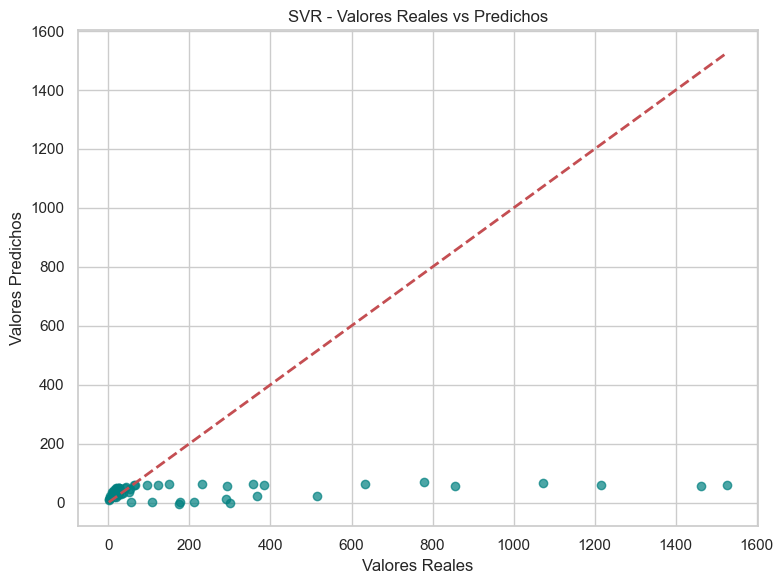


*Figura 7. XGBoost- Valores reales vs predichos*

Los resultados obtenidos del Modelo Support Vector Regressor (SVR) fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 192.19 | 394.93 | -0.1997 |

*Tabla 4. Resultados del modelo SVR*

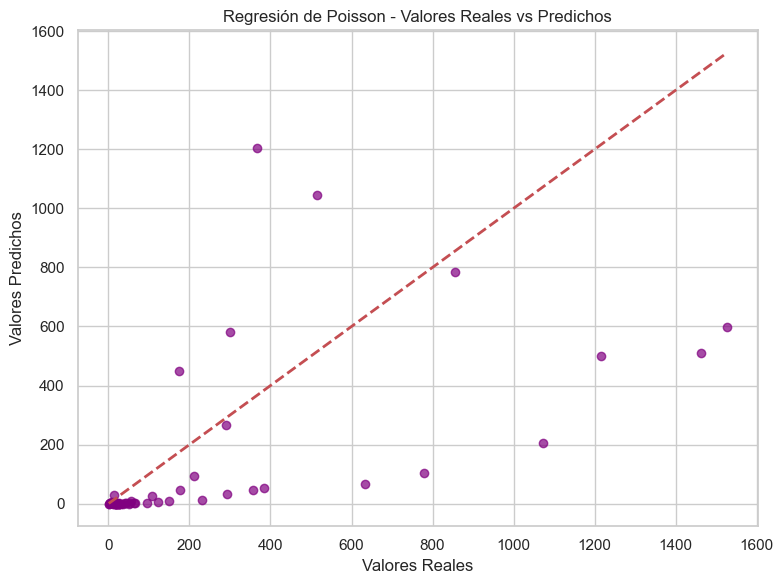


*Figura 8. SVR- Valores reales vs predichos*

Los resultados obtenidos del Modelo de Regresión de Poisson

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 163.90 | 307.19 | 0.2742 |

*Tabla 5. Resultados del modelo Poisson*

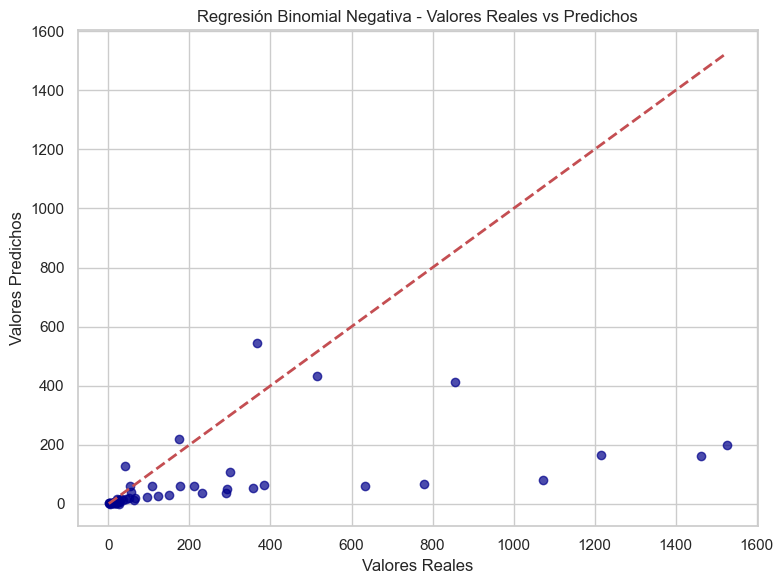


*Figura 9. Regresión Poisson - Valores reales vs predichos*

Los resultados obtenidos del Modelo Regresión Binomial Negativa con statsmodels fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 164.54 | 352.39 | 0.0449 |

*Tabla 6. Resultados del modelo Binomial Negativa*

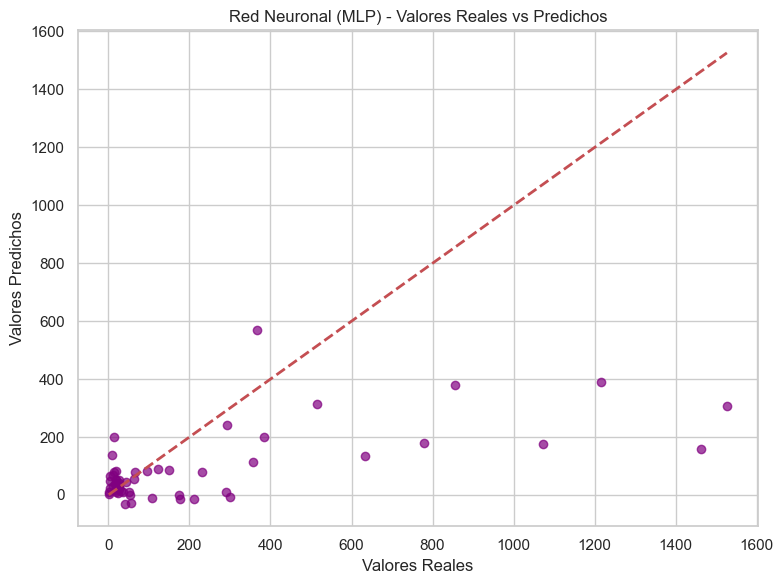


*Figura 10. Regresión Binomial Negativa - Valores reales vs predichos*

Los resultados obtenidos del Modelo para Red Neuronal con MLPRegressor fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| MAE: | RMSE: | R2: |
| 165.67 | 327.24 | 0.1764 |

*Tabla 7. Resultados del modelo Red Neuronal*



*Figura 11. Red Neurona l- Valores reales vs predichos*

En resumen, los resultados se muestran en la siguiente tabla:

Interfaz de usuario gráfica, Texto, Tabla

El contenido generado por IA puede ser incorrecto.

*Tabla 8. Resultados de los modelos predictivos*

**Discusión**

Los resultados del presente estudio, basado en la aplicación de modelos de regresión y aprendizaje automático para predecir casos de dengue en el distrito de Piura, revelan tanto el potencial como las limitaciones de estos enfoques. El modelo XGBoost destacó como el de mejor desempeño (R² = 0.48), seguido por la regresión de Poisson, lo que indica que tanto los modelos no lineales como los estadísticos clásicos ofrecen valor predictivo, aunque con restricciones en cuanto a su capacidad explicativa.

El bajo poder predictivo global observado en todos los modelos sugiere que la dinámica del dengue es altamente compleja y no puede explicarse únicamente mediante variables climáticas y demográficas. Estudios previos han advertido que factores como la densidad del vector, el comportamiento humano, la movilidad urbana y las condiciones de saneamiento son críticos en la transmisión del virus (Gubler, 2012; WHO, 2023). Esto justifica la necesidad de considerar modelos más integrados y multicausales.

Asimismo, se evidenció una fuerte sensibilidad de los modelos a la estacionalidad y a desfases temporales entre variables meteorológicas y casos reportados. La inclusión de variables derivadas (por ejemplo, medias móviles y lags) mejoró el rendimiento de los modelos, lo cual coincide con investigaciones similares (Dom et al., 2013).

## Conclusiones

1. El modelo XGBoost presentó el mejor rendimiento entre los evaluados, lo que refuerza su utilidad en tareas de predicción no lineales relacionadas con fenómenos epidemiológicos.
2. Ninguno de los modelos superó el umbral de un R² = 0.50, lo que evidencia la necesidad de integrar más variables explicativas relacionadas con factores sociales, espaciales y comportamentales.
3. La regresión de Poisson se posiciona como una alternativa robusta para datos de conteo, confirmando su aplicabilidad en contextos epidemiológicos, especialmente cuando se enfrenta a sobredispersión.
4. El preprocesamiento de datos y la ingeniería de variables tienen un impacto decisivo en el rendimiento predictivo de los modelos. La inclusión de lags y variables estacionales se demostró fundamental.

## Observaciones

* La calidad y granularidad de los datos disponibles son limitadas. Se trabajó con datos agregados semanales y proyecciones de población, lo cual puede generar pérdidas de información relevantes.
* Las variables climáticas, aunque necesarias, no fueron suficientes para explicar la variabilidad de los casos de dengue, lo cual concuerda con evidencias previas sobre el rol de determinantes sociales en enfermedades transmitidas por vectores (Morin et al., 2013).
* La elección del modelo debe considerar no solo el R², sino también la interpretabilidad y utilidad práctica del modelo para las autoridades de salud pública.

## Trabajo Futuro

1. **Ampliar el conjunto de variables**: Incorporar datos sobre movilidad humana, densidad del vector *Aedes aegypti*, cobertura de servicios básicos, niveles de educación y condiciones de vivienda.
2. **Modelado espacio-temporal**: Emplear técnicas como redes neuronales recurrentes (LSTM) o modelos bayesianos espacio-temporales que permitan capturar mejor la dinámica geográfica y temporal del dengue.
3. **Validación externa**: Aplicar los modelos desarrollados en otras provincias o distritos de características similares para evaluar su capacidad de generalización.
4. **Sistemas de alerta temprana**: Desarrollar dashboards que integren visualización de predicciones semanales para su uso por parte de autoridades locales.
5. **Integración con modelos epidemiológicos**: Complementar la predicción estadística con modelos SIR o SEIR que representen la dinámica poblacional de infección, susceptibles y recuperados.

## Mejora del Preprocesamiento

El preprocesamiento de datos fue una fase crítica para el desarrollo efectivo de los modelos predictivos. Las mejoras implementadas incluyeron:

* **Imputación avanzada**: Se utilizaron técnicas como KNN imputer y regresión múltiple para imputar valores faltantes sin distorsionar la estructura temporal.
* **Escalado y transformación**: Todas las variables numéricas fueron normalizadas utilizando Min-Max y transformadas mediante logaritmo natural y Box-Cox donde fue necesario, para estabilizar la varianza y cumplir supuestos de normalidad.
* **Detección de outliers**: Se aplicó la técnica de Isolation Forest para eliminar valores atípicos extremos que podrían afectar el entrenamiento del modelo.
* **Codificación temporal**: Las variables como semana del año y mes fueron transformadas mediante codificación cíclica (seno y coseno) para capturar estacionalidad.
* **Generación de nuevas variables**: Se agregaron lags temporales de 1 a 4 semanas para temperatura y precipitación, junto con medias móviles y diferencias semanales, para modelar efectos retardados.

## Referencias

* Dom, N. C., Hassan, A. A., Latif, Z. A., Ismail, R., & Pradhan, B. (2013). *An Artificial Neural Network model for predicting dengue outbreak*. Applied Geography, 40, 58–69. https://doi.org/10.1016/j.apgeog.2013.02.008
* Gubler, D. J. (2012). *Dengue viruses: their evolution, history and emergence as a global public health problem*. In Dengue and dengue hemorrhagic fever (pp. 1–22). CABI.
* World Health Organization. (2023). *Dengue and severe dengue*.<https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>
* DataSet de: Datos Abiertos de Vigilancia del Dengue, del año 2000 al 2023. <https://datosabiertos.gob.pe/dataset/vigilancia-epidemiol%C3%B3gica-de-dengue/resource/f854c3fb-0364-423c-a2f9-e69b67fa58d0>
* DataSet de: Poblaciones anuales del distrito de Piura, del año 2017 (Pág. 28) y del 2018 al 2023 (Pág. 71), del Compendio Estadístico Piura 2023.
* DataSet de: Indicadores climatológicos para el distrito de Piura, por semana, del 2016 al 2025. <https://app7.dge.gob.pe/maps/clima-salud/>