**Evaluación del riesgo de dengue en Costa Rica utilizando variables climáticas y redes neuronales**

**Práctica Profesional I XS-4410**

**Estudiantes:**

**Jimena Murillo Montero B75335**

**Jose Pablo Gómez Mata B83319**

**Abril, 2022**

**Tabla de contenido**

[Aspectos de la entidad 3](#_Toc101778754)

[Antecedentes 3](#_Toc101778755)

[Justificación 5](#_Toc101778756)

[Objetivos 5](#_Toc101778757)

[Estado actual de desarrollo del proyecto 6](#_Toc101778760)

[Descripción del problema 6](#_Toc101778761)

[Aspectos metodológicos 6](#_Toc101778762)

[Tipo de asistencia estadística 7](#_Toc101778763)

[Descripción de los datos disponibles 7](#_Toc101778764)

[Posibles riesgos 7](#_Toc101778765)

[Cronograma semanal del proyecto 8](#_Toc101778766)

[Referencias Bibliográficas 8](#_Toc101778767)

# **Aspectos de la entidad**

La entidad a la que pertenece este proyecto es el Centro de Investigación de Matemática Pura y Aplicada (CIMPA) de la Universidad de Costa Rica que es una unidad especializada de investigación avanzada en matemática pura y aplicada, abarca tanto la parte teórica como sus aplicaciones y su equipo trabaja en áreas desde fundamentos de matemática hasta aplicaciones para problemas cotidianos del país.

Este proyecto tuvo comienzo el 28 de marzo del 2022 y va a ser desarrollado junto al Dr. Shu Wei Chou Chen ([shuwei.chou@ucr.ac.cr](mailto:shuwei.chou@ucr.ac.cr)), profesor de la Universidad de Costa Rica e investigador del CIMPA especializado en métodos estadísticos, series temporales y análisis espacio-temporal. De manera análoga el Dr. Luis Barboza Chinchilla ([luisalberto.barboza@ucr.ac.cr](mailto:luisalberto.barboza@ucr.ac.cr)), también profesor de la Universidad de Costa Rica e investigador del CIMPA es un colaborador importante del proyecto, quién dará un apoyo rutinario como seguimiento de las investigaciones precedentes a esta trabajadas en el CIMPA de las que ha sido partícipe.

Por otro lado, se debe destacar el papel del Ministerio de Salud de Costa Rica que aportó en gran medida a la recolección de los datos, ya que como institución su misión es proteger y mejorar el estado de salud física, mental y social de los habitantes, además de que se involucra en proyectos que incentiven y estudien temas de interés. En este caso la colaboración con este proyecto se podría promocionar la prevención de la enfermedad estudiada.

Consecuentemente el proyecto estudia dos áreas, el de la salud pública que trata sobre el análisis del vector el cual está bajo observación para prevenir en un futuro la propagación de la enfermedad y propiciar un ambiente humano de calidad. Por otro lado, la correlación asociada a esta enfermedad obliga a estudiar la parte del medio ambiente, ya que el mosquito se desenvuelve según las condiciones climáticas en donde se ven envueltos parámetros del medio ambiente y su comportamiento.

# **Antecedentes**

El dengue es una enfermedad viral, de carácter endémico-epidémico que es transmitida por los vectores de mosquitos *Aedes aegypti* y *Aedes albopictus*, que constituye la arbovirosis más importante a nivel mundial en términos de morbilidad, mortalidad y afectación económica (Guzmán et al., Kindhauser,2003).

Múltiples estudios han demostrado que la enfermedad va de la mano con el clima, según Portela et al., las temperaturas en los primeros trimestres del período 1986-2003, especialmente las temperaturas mínimas, fueron significativamente más altas en los años de inicio de las epidemias de dengue en la Ciudad de Río de Janeiro. Lo cual hace más desafiante construir un modelo que prediga con exactitud dicho riesgo relativo, ya que por los microclimas que se forman en Costa Rica no se cuenta con un comportamiento estándar para todas las regiones. Esto puede inducir a errores de predicción en distintas zonas por el comportamiento del clima, factor determinante en el ciclo reproductivo del mosquito que transmite la enfermedad. Las técnicas de predicción pueden no estar ajustándose de manera adecuada a todas las zonas por la variabilidad en el clima.

Tal cual entre los análisis de dicho tema tenemos por ejemplo a Zhu et al. (2019) que plantean un modelo de regresión probit para predicción de casos de dengue en Guangdong, China. Ellos se basan en variables meteorológicas como lo son temperatura promedio, presión atmosférica mínima/máxima y promedio y humedad relativa. Entre los resultados de este estudio se tiene que la predicción de casos de dengue en las primeras 41 semanas del 2017 tuvo una bondad de ajuste de 0.60.

Un modelo con un objetivo similar es el presentado por Salim et al. (2021) pretende predecir brotes de dengue en Selangor, Malasia utilizando IBM SPSS Modeler 18. El modelo ideal fue seleccionado entre diferentes métodos como lo son árboles de decisión, redes neuronales, SVM (máquinas de vectores de soporte) y red bayesiana. Entre las variables ambientales utilizadas están la humedad relativa, la precipitación y la velocidad del viento. En conclusión, se escoge el modelo SVM usando un núcleo lineal, la sensibilidad para la muestra de prueba fue de 63.54%. Entrando más al campo de redes neuronales se tiene el artículo de Tatavarty y Satsangi (2021). Para esta investigación se desea predecir los casos de dengue en cuatro ciudades del distrito Agra en la India. Para este modelo se utilizaron datos de condiciones ambientales como precipitación, humedad relativa y temperatura entre el 2005 y 2018. Las redes neuronales utilizadas eran de tres nodos (input, hidden y output). Utilizando el error porcentual cuadrático medio se obtuvieron resultados diferentes para las distintas ciudades, siendo el mayor error 84% y el menor 27.3%. Los modelos presentados en los artículos anteriores muestran la utilidad de ciertas variables climáticas para la predicción de dengue, además de denotar la variabilidad que ocurre entre regiones climáticas distintas y como un modelo puede dar buenos resultados para algunas regiones y no otras.

En base a los estudios realizados por los expertos del CIMPA, cuyo objetivo es predecir el riesgo relativo de la enfermedad en diferentes cantones del país, cuentan con distintos modelos y algoritmos previamente ajustados y entrenados que tienen como meta pronosticar el riesgo relativo de manera precisa en un determinado lugar y tiempo. El artículo de Vásquez, Loría, Sánchez y Barboza (2019) utiliza dos modelos diferentes para la predicción de dengue en cinco cantones del país que se ubican en distintas regiones climáticas. Los modelos utilizados son el Modelo aditivo generalizado (GAM, siglas en inglés) y bosques aleatorios (RF) tomando como variables predictoras la temperatura mínima/máxima y promedio, precipitación, humedad relativa y la temperatura superficial del mar semanal. Los modelos dieron resultados diferentes dependiendo de la localidad, sin embargo, como mínimo lograron predecir las semanas en las que ocurre un incremento o disminución de los casos.

En un segundo artículo publicado recientemente continúa esta investigación abarcando ahora 32 cantones en Costa Rica. En esta investigación por parte de Barboza, et al (2022) utiliza un modelo aditivo generalizado (GAMLS, siglas en inglés) para locación, escala y forma) y vuelve a utilizar RF. En las variables climáticas seleccionadas se tienen algunos cambios, en estos modelos se incluyen estimados de precipitación diaria, anomalías en la temperatura superficial del mar semanal, índice de diferencia en vegetación normalizado, temperatura de la superficie terrestre diurna e índice del norte atlántico tropical. Los modelos fueron entrenados con datos del 2000 a diciembre del 2020 y probados en los primeros tres meses del 2021. No hubo diferencia significativa entre ambos métodos y lograron predecir de manera relativamente efectiva los casos de dengue.

# **Justificación**

El dengue es una enfermedad que se ha establecido globalmente, la infección en seres humanos puede pasar desapercibida, pero también podría tener manifestaciones clínicas importantes. Los esfuerzos actuales que operan sobre el vector, el mosquito Aedes sp., no han logrado detener el incremento de la incidencia de epidemias de dengue (Bhatt, et al., 2013).

La carga global del dengue es grande, se estima que alrededor de 50 millones de personas se infectan al año en 100 países, y es posible que siga aumentando. Esta enfermedad también impone una gran demanda en el sistema de salud, ya que, aunque los síntomas severos ocurran en un porcentaje pequeño de los infectados, es difícil identificar a personas de riesgo por lo que se hospitalizan, aunque no tengan infecciones complicadas (Simmones, Farrar, van Vinh Chau & Wils, 2012).

En Costa Rica el dengue ha sido un reto para el sistema de salud pública desde 1993 cuando se diagnosticaron los primeros casos. Desde ese momento, se ha reportado al ministerio de salud un total de 376 158 casos sospechosos y confirmados (Sánchez, Arroyo-Esquivel & Vásquez, 2019). A pesar de las medidas preventivas, las infecciones siguen progresando; es por esto que se debe plantear una forma más estratégica para aplicar estas medidas y protocolos con la esperanza de así comenzar a reducir los casos.

En este proyecto se estará creando un modelo de redes neuronales para la predicción de casos de dengue en varios cantones del país de diferentes zonas climáticas. Utilizando este modelo sería posible determinar los patrones de aumento y disminución de casos en regiones específicas con lo que se podría advertir al Ministerio de Salud para administrar los recursos de manera más efectiva y dar apoyo a las comunidades más afectadas en el momento indicado. De esta forma se evitaría que cientos de costarricenses sufran una enfermedad incapacitante cada año y de la mano se reducirían las pérdidas económicas a lo que esto conlleva.

# **Objetivos**

## Objetivo general

Predecir el riesgo relativo de dengue para 32 cantones de Costa Rica utilizando variables climáticas y modelos de redes neuronales con el fin de apoyar la toma de decisiones en el Ministerio de Salud.

## Objetivos específicos

1. Plantear modelos de redes neuronales para la predicción de casos de dengue en base a variables climáticas
2. Medir el rendimiento de los modelos de redes neuronales planteados por medio de las métricas de precisión.
3. Comparar la habilidad predictiva de los modelos de redes neuronales entre sí y con respecto a un modelo RF.
4. Encontrar el modelo de redes neuronales con la mejor capacidad predictiva entre los planteados.

# **Estado actual de desarrollo del proyecto**

Es un proyecto nuevo el cual se diferencia de los demás estudios preliminares ya que se pondrá en práctica, una diferente metodología, se haría el intento con redes neuronales para observar el comportamiento y adaptabilidad que estas tienen con los datos; por consiguiente, se tendría un alcance más amplio respecto al tema. Al haberse realizado trabajos anteriores la recolección de datos no es necesaria y se arrancará a partir del diseño de los modelos elegidos.

# **Descripción del problema**

Para predecir el comportamiento de casos de dengue se han utilizado previamente modelos estadísticos como el GAM, GAMLSS y un algoritmo de aprendizaje automático conocido como bosque aleatorio (RF), los cuales permiten predecir el riesgo relativo del dengue para cinco cantones del país con clima variado y posteriormente para 32 municipalidades de interés. Dichos modelos poseen una gran ventaja, según Chouldechova A. y Hastie T. (2015), estos modelos permiten modelar el predictor lineal con mayor flexibilidad de un modelo lineal generalizado, sumando las funciones generales de cada variable. Del mismo modo se eligió el algoritmo de bosque aleatorio, ya que como mencionan Schonlau M. y Yuyan Zou R. este se adapta fácilmente a las no linealidades que se encuentran en los datos y por lo tanto tienden a predecir de mejor manera que una regresión lineal.

Con la intención de lograr predecir de mejor manera para todos los cantones o municipalidades, ya que estas cuentan con micro climas que en ciertas zonas entorpecen la predicción, se opta por utilizar las redes neuronales para predecir los casos de dengue de una manera más precisa.

# **Aspectos metodológicos**

El análisis de redes neuronales analiza patrones y aprendizaje de clasificación, inspirada en la estructura biológica de una neurona toma las entradas de datos (variables), se les asigna un peso, luego se calcula el sesgo el cual sería posteriormente sumado a la multiplicación del peso por las variables (entradas), todo esto con el fin de obtener un resultado (Y). La salida o el resultado de la neurona puede dar valores de menos infinito hasta más infinito, para eso se utiliza un mecanismo de mapeo de la entrada y salida de la neurona el cual se conoce como la función de activación.

Ante un análisis de este tipo, se requiere de competencias analíticas para determinar el funcionamiento correcto de dichos modelos en donde la selección de la muestra es parte fundamental a la hora de entrenar un modelo para que haya presencia baja de subestimación o sobreestimación y si se requiere posteriormente comparar resultados se deben aplicar conceptos como lo son pruebas de hipótesis.

# **Tipo de asistencia estadística**

Específicamente el tipo de asistencia estadística es de pronóstico o predictivo, ya que gran peso de esta se vería mayormente cargado en la parte de modelación, aprendizaje supervisado y aprendizaje de máquinas, con el objetivo de crear uno o varios modelos robustos que sean capaces de dar la mayor exactitud posible.

# **Descripción de los datos disponibles**

Se utilizaron dos tipos de información diferente, casos de dengue semanales y datos meteorológicos. Según explican Vásquez, P et al (2019) los casos de dengue provienen del ministerio de salud del país (Costa Rica), donde se tienen casos clínicamente sospechosos y confirmados son reportados al departamento de vigilancia de salud del ministerio, se cuenta con una base del 2000 hasta el 2021 para 32 cantones. Además, se cuenta con los datos meteorológicos desde enero del 2007 hasta diciembre del 2017 los cuales fueron proporcionados por el instituto meteorológico nacional (IMN) de Costa Rica. En donde para el primer estudio se contaban con cinco estaciones localizadas en el área de estudio, de esta base se obtuvieron tres variables (precipitación, humedad relativa y la temperatura del océano) mientras que para el segundo estudio se tomaron cinco variables del 2000 hasta el 2021. La precipitación estimada que se obtuvo mediante la precipitación infrarroja del grupo de peligros climáticos (CHRIPS, siglas en ingles), la anomalía de la temperatura de la superficie del mar (SSTA, siglas en ingles) la cual se obtuvo del centro de predicción del clima (CPC, siglas en ingles) de la administración nacional oceanográficas y atmosférica de los estados unidos (NOAA, siglas en ingles), el índice se vegetación de diferencia normalizada (NVDI, siglas en ingles) el cual se obtuvo de la imagen del satélite de resolución moderada espectrorradiómetro (MODIS, siglas en ingles), la temperatura terrestre durante el día la cual se consigue a partir de la covariable NDVI (siglas en ingles) y por último el índice del atlántico norte tropical el cual es una índice de anomalía de la temperatura en la superficie del océano.

# **Posibles riesgos**

La gran variabilidad que se tiene con respecto al clima puede ocasionar que las estimaciones en algunos casos sean muy exactas y en otros todo lo contrario, es por esto que se busca utilizar las redes neuronales para analizar si se pueden obtener patrones del clima que orienten de mejor manera las predicciones deseadas, sin embargo, un posible riesgo son las predicciones a un futuro un poco más lejano, ya que como indica el Instituto meteorológico nacional del país (Costa Rica) en las últimas décadas los ciclos naturales se han visto caracterizados por fuertes variaciones, uno de estos el efecto antropogénico, asociado a la contaminación con gases de efecto invernadero, es uno de los generadores de estas marcadas oscilaciones de la variabilidad climática. En otras palabras, el clima está asociado con el nivel de contaminación que puede producir cambios repentinos, los cuales deberían estar incluidos en el modelo y en este caso no lo están debido a que los datos ya fueron recolectados, lo cual posiblemente estaría influyendo en los patrones que se estudiarán.

Otro aspecto para tomar en consideración es que los modelos generados en este proyecto se basan en los datos reportados por el Ministerio de Salud Pública, sin embargo, para el caso del dengue muchas personas que lo padecen no llegan a reportar y estos datos no quedan registrados por lo que hay un subregistro de la enfermedad. Al basar las predicciones en datos con subregistro es importante tomarlo en cuenta los intervalos de predicción y así no confiar en la predicción puntual.

# **Cronograma semanal del proyecto**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mes | Marzo | Abril | | | | Mayo | | | | | Junio | | | | Julio | | | |
| Semana | IV | I | II | III | IV | I | II | III | IV | I | | II | III | IV | I | II | III | IV |
| Actividad: |
| Investigar y familiarizarse con las redes neuronales y paquetes de redes neuronales para R. |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Trabajar en la primera entrega del anteproyecto |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Familiarizarse con los datos y realizar análisis descriptivo de estos, buscar alternativas de paquetes en R como tidymodel. |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Crear redes neuronales con los paquetes escogidos para 4 cantones, también encontrar un modelo de RF para comparar. Trabajar en correcciones del anteproyecto. |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Buscar y aplicar mediciones de error apropiadas para estas redes neuronales |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Comparar las redes neuronales creadas entre sí, para buscar los mejores ajustes. Última revisión del Anteproyecto |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Aplicar las redes neuronales seleccionadas a todos los cantones de los que se tienen datos. |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Comparar las redes neuronales seleccionadas entre sí para cada cantón y con respecto a RF. Trabajar en el avance del proyecto |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| Trabajar en avance de proyecto. |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |

# **Referencias Bibliográficas**

Barboza, L., Calvo, J., Chou, S., García, Y., Hidalgo H., Sanchez, F., Vásquez, P. (2022). Assessing dengue fever risk in Costa Rica by using climate variables and machine learning techniques, Universidad de Costa Rica. <https://arxiv.org/abs/2204.01483>

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J., & Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504–507. https://doi-org.ezproxy.sibdi.ucr.ac.cr/10.1038/nature12060

Martinez, E. (2008). Dengue. Estudos avançados. <https://www.scielo.br/j/ea/a/j4JhLfhXsVL3RSqNHMsrH9t/?format=pdf&lang=es>

Kindhauser, M., (2003). Dengue Y Fiebre Hemorrágica Dengue. In: Defensa Global Ante La Amenaza De Enfermedades Infecciosas. Ginebra: Organización Mundial De La Salud. P.140-3.

Cardoso, D., Fagundes, A., Portela, F., Teixeira dos Santos, G. (2009). Clima e epidemias de dengue no Estado do Rio de Janeiro. Revista da Sociedade Brasileira de Medicina Tropical. <https://www.scielo.br/j/rsbmt/a/fvWZZYsM5Ms8kjzt6yx7qWN/?format=pdf&lang=pt>

Carpenter, G. (1989). Neural network models for pattern recognition and associative memory. Neural networks. <http://techlab.bu.edu/files/resources/articles_cns/carpenter1989.pdf>

Chouldechova, A., Hastie, T. (2015). Generalized Additive Model Selection. arXiv. <https://arxiv.org/pdf/1506.03850.pdf>

Salim, N.A.M., Wah, Y.B., Reeves, C. et al. Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. *Sci Rep* 11, 939 (2021).<https://doi.org/10.1038/s41598-020-79193-2>

Sánchez, F., Arroyo-Esquivel, J., & Vásquez, P. (2019). Transmission dynamics of dengue in Costa Rica: the role of hospitalizations. Revista De Matemática: Teoría Y Aplicaciones, 27(1), 241-266. doi: 10.15517/rmta.v27i1.39977

Schonlau, M., Zou, R. (2020). The random forest algorithm for statistical learning. *The Stata Journal*. <https://journals.sagepub.com/doi/pdf/10.1177/1536867X20909688>

Simmons, C., Farrar, J., van Vinh Chau, N., & Wills, B. (2012). Dengue. New England Journal Of Medicine, 366(15), 1423-1432. doi: 10.1056/nejmra1110265

Tatavarty, G. S., & Satsangi, N. (2021). Potential Use of Artificial Neural Networks in Prediction of Dengue Abundance in Agra District. IOP Conference Series.Materials Science and Engineering, 1116(1)<http://dx.doi.org/10.1088/1757-899X/1116/1/012201>

Vásquez, P., Loría, A., Sánchez, F., & Barboza, L. (2019). Climate-driven statistical models as effective predictors of local dengue incidence in costa rica: a generalized additive model and random forest approach. Revista De Matemática: Teoría Y Aplicaciones, 27(1), 1-21. doi: 10.15517/rmta.v27i1.39931

Zhu B, Wang L, Wang H, Cao Z, Zha L, Li Z, et al. (2019) Prediction model for dengue fever based on interactive effects between multiple meteorological factors in Guangdong, China (2008–2016). PLoS ONE 14(12): e0225811.<https://doi.org/10.1371/journal.pone.0225811>