Entrega semana 4 – Propuesta inicial

Grupo 13: Freddy Rodrigo Mendoza Ticona, William Alexander Romero Bolívar, Maria Paula Salamanca Delgado, Jorge Oswaldo Suárez Rodríguez.

Detección e identificación de "puntos calientes" de dengue en Colombia: una aproximación desde el Aprendizaje no Supervisado

Resumen

El dengue es una enfermedad viral transmitida por la picadura de mosquitos infectados que se caracteriza por producir fiebre, dolor corporal, pérdida del apetito y, en casos graves, sangrado de mucosas. La población más vulnerable a esta enfermedad son niños y adultos mayores, sin embargo, puede afectar a cualquier grupo demográfico. Según el Instituto Nacional de Salud (INS), a julio de 2022, en Colombia se han presentado 34.017 casos, dentro de los cuales el 52,4% han presentado signos de alarma o graves (1,2).

Es posible prevenir y controlar su propagación a través de la concientización de la población para evitar la proliferación de los mosquitos que transmiten esta enfermedad. Así, este proyecto tiene como objetivo encontrar en qué zonas de Colombia se deberían concentrar las campañas de concientización de prevención y control contra el dengue haciendo uso de métodos de aprendizaje no supervisado. De esta manera, se aplicará un análisis de clústeres basado en densidad con el cual se podrán encontrar clústeres de forma arbitraria para lograr estratificar municipios según el riesgo de tasa de incidencia de dengue. Se espera encontrar que las zonas de mayor humedad, menor altitud sobre el nivel del mar y mayor temperatura sean las zonas críticas de incidencia.

Los resultados de este proyecto pretenden ser de utilidad para entidades públicas y gubernamentales, al permitirles dirigir recursos hacia la capacitación y concientización de la población para la prevención y control del dengue eficientemente en aquellas zonas de mayor riesgo en la proliferación del dengue.

Palabras clave: Dengue, Epidemia, Aprendizaje no Supervisado, Clústeres.

Introducción

Según la Organización Mundial de la Salud (OMS), se estima que anualmente el número de personas infectadas por dengue llega a 390 millones en el mundo y el número de personas en riesgo de infección, a 3,9 billones (3). Adicionalmente, según la Organización Panamericana de la Salud (OPS), la circulación del dengue en Suramérica es hiperendémica, siendo Brasil y Colombia los que reportan el mayor número de casos; tan solo para el 2019 Colombia reportó 127.000 casos (4). Hay que considerar que, en los años 2020 y 2021 por la pandemia de COVID-19, la notificación de las enfermedades endémicas disminuyó, por lo cual hubo subregistro de casos por dengue.

Los mosquitos *aegypti* son el principal vector de transmisión del dengue y están presentes en casi todos los países de Suramérica. El ciclo de transmisión depende de la interacción humano-mosquito-humano y se da por el contagio de la sangre de un humano infectado a un humano susceptible a través de la picadura del mosquito *Aedes* hembra. El tiempo de incubación de este mosquito es de 8 a 12 días, después de este período es capaz de transmitir el virus a varios humanos durante su ciclo de vida, además, se reproduce cerca de

las casas, poniendo huevos en recipientes de agua estancada y su distancia típica de vuelo es relativamente corta (5,6).

Se ha revisado factores asociados al aumento de la incidencia de infección por dengue, considerando el efecto potencial del cambio climático, como el aumento global de la temperatura y en Suramérica el fenómeno de El Niño. En muchos países tropicales, el aumento estacional de las precipitaciones contribuye a una mayor densidad de mosquitos, dada la presencia de recipientes de almacenamiento de agua. Las temperaturas más cálidas también aumentan el tiempo que el mosquito permanece infeccioso. Adicionalmente, condiciones de hacinamiento incrementan la fracción de la población susceptible (5,6).

La presentación clínica clásica del dengue se caracteriza por fiebre de cinco a siete días, acompañada de cefalea, dolor retro orbitario, dolores musculares y óseos. El dengue se clasifica como grave cuando presenta extravasación severa del plasma, hemorragias severas y daño grave en órganos (5).

Es posible prevenir la infección por dengue mediante el control de su transmisión por el vector, para ello es importante detectar donde hay mayor circulación de la infección. De esta manera, este trabajo tiene como propósito estratificar el riesgo de infección por localización mediante una unidad de medida, como la tasa de incidencia de dengue, a través del aprendizaje no supervisado como el análisis de clústeres basado en densidad. Lo anterior sería de utilidad para las entidades gubernamentales y funcionarios de salud pública, al poder diferenciar sitios con mayor riesgo que requieran toma de decisiones sobre una intervención de control de propagación.

En la primera parte de este documento se describen los antecedentes en la literatura nacional e internacional que tratan sobre el dengue y la identificación de factores de riesgo, encontrando que no existe en la literatura el uso de algoritmos de aprendizaje no supervisado para este fin. En la segunda parte se describen los datos objeto de análisis y la metodología a seguir para el desarrollo de este proyecto. Finalmente, de exponen los resultados obtenidos del análisis y las conclusiones encontradas.

Revisión de literatura

En estudios anteriores sobre el dengue se han utilizado algoritmos de aprendizaje no supervisado. Por ejemplo, los métodos de agrupamiento son adecuados para la visualización de enfermedades, en especial los basados en la densidad para separar las regiones que tienen alta densidad de las regiones que tienen baja densidad, lo que permite comprender a detalle los datos. En el estudio de Shaukat, et al, el análisis del conjunto de datos de Pakistán mediante un algoritmo de DBSCAN encontró que el dengue atacó principalmente la ciudad de Jhelum y Tehsil Jhelum, lo que permite enfocar estrategias de prevención de la enfermedad (7). Otro estudio realizado en la India encontró conglomerados de casos de dengue en Delhi usando un algoritmo de agrupamiento DBSCAN; estos puntos críticos se caracterizaron por ser grupos socioeconómicos bajos, estar cerca al río Yamuna, lagos y lugares con agua estancada (8).

Una revisión de la literatura realizada en Colombia encontró que, a nivel mundial en general, se ha usado modelos de aprendizaje supervisado con el objetivo de predicción de dengue, en especial modelos de regresión logística con datos demográficos, clínicos y de laboratorio. Así como modelos de regresión lineal, Random Forest y Support Vector Machine, con datos socioeconómicos, demográficos, climáticos y ambientales (9). También los estudios de análisis espacio-temporal, existiendo uno realizado en Cali, Colombia, en el cual los autores encontraron que el nivel socioeconómico, la densidad poblacional, proximidad a talleres con neumáticos, viveros de plantas y sistemas de alcantarillado, están relacionados con la

enfermedad (10). No se encontró alguna publicación donde se utilice aprendizaje no supervisado para la estratificación del riesgo de dengue en Colombia, por lo cual, un análisis de clústeres basado en densidad por municipios de Colombia propuesto en este proyecto sería diferente a los presentados en la literatura.

Descripción de los datos

Para el desarrollo de este proyecto se cuenta con una base de datos con información sobre 1.121 municipios y 1.017 características, como datos demográficos y socioeconómicos, casos acumulados de dengue del 2007 al 2019 y por semana epidemiológica e información climática como mediciones de temperatura y precipitaciones mensuales. Esta información se obtuvo de las bases de datos del Sistema de Vigilancia en Salud Pública (SIVIGILA) del Instituto Nacional de Salud (INS) y del Departamento Administrativo Nacional de Estadística (DANE).

Variable	**Descripción**	**Tipo dato**
Municipality code	Código del municipio	numérico
Municipality	Descripción de municipio	texto
Population (2007-2019)	Población del municipio por año	numérico
Cases (2007-2019)	Casos dengue reportados en el municipio x año	numérico
Age0-4(%)	Porcentaje de población menor de 4 años	numérico
Age5-14(%)	Porcentaje de población entre 5 y 14 años	numérico
Age15-29(%)	Porcentaje de población entre 15 y 29 años	numérico
Age>30(%)	Porcentaje de población mayor de 30 años	numérico
AfrocolombianPopulation(%)	Porcentaje de población afrocolombiana	numérico
IndianPopulation(%)	Porcentaje de población indígena	numérico
PeoplewithDisabilities(%)	Porcentaje de población con discapacidades (física, psicológica o mental)	numérico
Peoplewhocannotreadorwrite(%)	Porcentaje de población con que no puede leer/escribir	numérico
Secondary/HigherEducation(%)	Porcentaje de población que tiene educación secundaria	numérico
Employedpopulation(%)	Porcentaje de población empleada	numérico
Unemployedpopulation(%)	Porcentaje de población desempleada	numérico
Peopledoinghousework(%)	Porcentaje de población que realizan trabajo doméstico	numérico
Retiredpeople(%)	Porcentaje de población jubilada	numérico
Men(%)	Porcentaje de población masculina	numérico
Women(%)	Porcentaje de población femenina	numérico
Householdswithoutwateraccess(%)	Porcentaje de viviendas sin acceso a agua	numérico
Householdswithoutinternetaccess(%)	Porcentaje de viviendas sin acceso a internet	numérico
Buildingstratification1(%)	Porcentaje de viviendas estrato 1	numérico
Buildingstratification2(%)	Porcentaje de viviendas estrato 2	numérico
Buildingstratification3(%)	Porcentaje de viviendas estrato 3	numérico
Buildingstratification4(%)	Porcentaje de viviendas estrato 4	numérico
Buildingstratification5(%)	Porcentaje de viviendas estrato 5	numérico
Buildingstratification6(%)	Porcentaje de viviendas estrato 6	numérico
NumberofhospitalsperKm2	Número de hospitales por km2 en el municipio	numérico

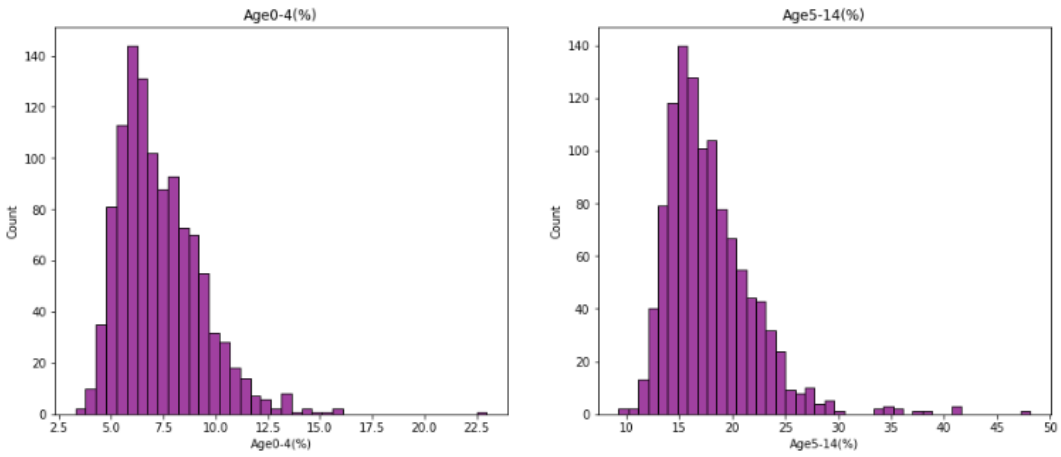
NumberofhousesperKm2	Número de casas por km2 en el municipio	numérico
TEMPERATURE_mm_AA	Temperatura promedio en el municipio por mes mm del año AA (desde 2007 hasta 201812)	numérico
PRECIPITATION_mm_AA	Precipitaciones en el municipio por mes por mes mm del año AA (desde 2007 hasta 201812)	numérico
AAAA/wi	Casos de dengue reportados x semana w(i) (1-52) del año AAAA (desde el año 2007 hasta 201912)	numérico

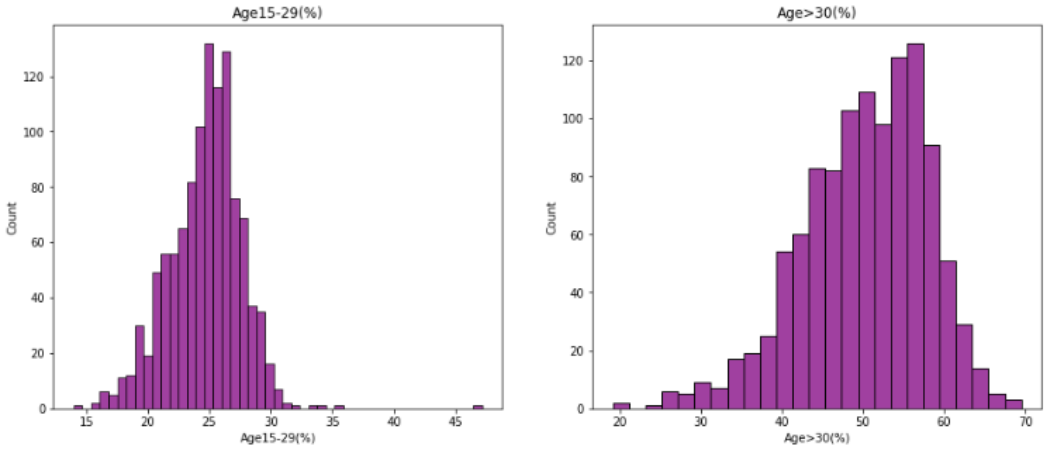
A continuación, se encuentran algunas estadísticas descriptivas sobre población y casos de dengue:

	Population2007	Population2008	Population2009	Population2010	Population2011	Population2012	Population2013	Population2014	Population2015
count	1.121000e+03	1.121000e+03	1.121000e+03	1.121000e+03	1.121000e+03	1.121000e+03	1.121000e+03	1.121000e+03	1.121000e+03
mean	3.830083e+04	3.872660e+04	3.914528e+04	3.956012e+04	3.995886e+04	4.033605e+04	4.069779e+04	4.105313e+04	4.141868e+04
std	2.349465e+05	2.372341e+05	2.394122e+05	2.415079e+05	2.434845e+05	2.452775e+05	2.466488e+05	2.476931e+05	2.485725e+05
min	0.000000e+00	1.270000e+02	1.270000e+02	1.290000e+02	1.440000e+02	1.590000e+02	1.720000e+02	1.920000e+02	2.100000e+02
25%	6.373000e+03	6.505000e+03	6.583000e+03	6.632000e+03	6.689000e+03	6.586000e+03	6.496000e+03	6.520000e+03	6.433000e+03
50%	1.200400e+04	1.220600e+04	1.232500e+04	1.252800e+04	1.260300e+04	1.256800e+04	1.267200e+04	1.267900e+04	1.268000e+04
75%	2.422200e+04	2.464200e+04	2.483100e+04	2.516200e+04	2.548000e+04	2.570500e+04	2.595700e+04	2.627200e+04	2.633500e+04
max	6.866363e+06	6.936977e+06	7.003434e+06	7.065669e+06	7.119281e+06	7.162261e+06	7.197326e+06	7.226652e+06	7.253823e+06

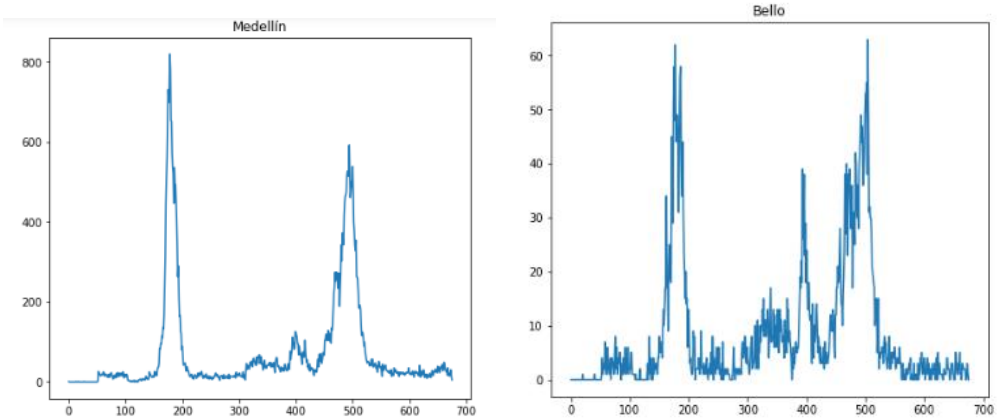
	Cases2007	Cases2008	Cases2009	Cases2010	Cases2011	Cases2012	Cases2013	Cases2014	Cases2015
count	1121.000000	1121.000000	1121.000000	1121.000000	1121.000000	1121.000000	1121.000000	1121.000000	1121.000000
mean	36.551293	32.183764	46.338983	138.495986	26.652988	47.715433	111.233720	93.465656	85.495986
std	229.110339	162.958825	286.415512	784.616567	111.911293	211.452528	660.811098	399.638060	510.500042
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000
50%	1.000000	2.000000	2.000000	10.000000	2.000000	3.000000	9.000000	8.000000	8.000000
75%	9.000000	10.000000	12.000000	54.000000	12.000000	23.000000	51.000000	51.000000	44.000000
max	5223.000000	2816.000000	6335.000000	15570.000000	1724.000000	3128.000000	17539.000000	5855.000000	14523.000000

Adicionalmente, gracias a histogramas sobre la distribución de población de cada municipio por edades se observa que, aproximadamente, el 30% de la población es vulnerable al dengue.





Finalmente, se encontraron comportamientos estacionales en los casos de dengue en Bello y dos picos importantes en Medellín.



Propuesta metodológica

Estratificar municipios según el riesgo de tasa de incidencia de dengue, pertenece al área de aprendizaje no supervisado. El enfoque de agrupación en clústeres basado en la densidad es una metodología capaz de encontrar clústeres de forma arbitraria, donde los clústeres se definen como regiones densas separadas por regiones de baja densidad.

Dado que el algoritmo DBSCAN (Density-based spatial clustering of applications with noise) agrupa los datos en función de las densidades de las observaciones, permite formar grupos más densos y homogéneos en sus características e identificar aquellos ruidos o datos atípicos, cosa que también pueda ser un output interesante en el estudio que se va a realizar, por esta razón, es un algoritmo que se ajusta muy bien para abordar este proyecto al permitir la estratificación de municipios según el riesgo de tasa de incidencia de dengue.

De otra parte, al intentar estratificar las zonas de influencia de dengue para evaluar donde se deben enfocar esfuerzos en campañas de prevención, también podría ser muy útil utilizar Clustering Jerárquico, lo que puede permitir tener una visualización de las distancias entre las características que representan a los municipios.

Bibliografía

1 . Instituto Nacional de Salud. Protocolo de Vigilancia de Dengue [Internet]. INS 2022. [citado el 20 de agosto de 2022. Disponible en: https://www.ins.gov.co/buscadoreventos/Lineamientos/Pro_Dengue.pdf

2. Instituto Nacional de Salud. Informe de Evento Dengue [Internet]. INS 2022. Citado el 20 de agosto de 2022. Disponible en: https://www.ins.gov.co/buscadoreventos/Informesdeevento/DENGUE%20PE%20VII%202022.pdf

3. World Health Organization. Dengue and severe dengue [Internet]. WHO 2022. Citado el 04 de septimebre de 2022. Disponible en: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=The%20number%20of%20dengue%20cases,affecting%20mostly%20younger%20age%20group.

4. Organización Panamericana de la Salud. Dengue [Internet]. OPS. 2020. Citado el 20 de agosto de 2022. Disponible en: https://www.paho.org/es/temas/dengue

5. Thomas SJ, Rothman AL. Dengue virus infection: Epidemiology. In: UpToDate, Shefner JM (Ed), UpToDate, Waltham, MA. (Accessed on September 04, 2022.)

6. Bhatt S, Gething PW, Brady OJ, et al. The global distribution and burden of dengue. *Nature*. 2013;496(7446):504-507. doi:10.1038/nature12060

7. Shaukat K, Masood N, Shafaat AB, Jabbar K, Shabbir H, Shabbir S. Dengue fever in perspective of clustering algorithms. arXiv preprint arXiv:1511.07353. 2015 Nov 23.

8. G. M. Nandana, S. Mala and A. Rawat, "Hotspot Detection of Dengue Fever Outbreaks Using DBSCAN Algorithm," *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, pp. 158-161, doi: 10.1109/CONFLUENCE.2019.8776916.

9. Hoyos W, Aguilar J, Toro M. Dengue models based on machine learning techniques: A systematic literature review. *Artif Intell Med*. 2021;119:102157. doi:10.1016/j.artmed.2021.102157

10. Delmelle E, Hagenlocher M, Kienberger S, Casas I. A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. *Acta Trop*. 2016;164:169-176. doi:10.1016/j.actatropica.2016.08.028