Datos Abiertos y Alertas sobre el Dengue
Protocolo de investigación

Indice de contenidos

1. Introducción

El Dengue es una enfermedad actualmente endémica en todo el Paraguay, según la Dirección General de Vigilancia de Salud del Ministerio de Salud Pública y Bienestar Social (MSPBS). La enfermedad tiene sus brotes en ciclos epidemiológicos que tienen relación con otras co-variables como ser clima y factores sociales que permiten el desarrollo del vector de transmisión.

Este proyecto propone la creación de herramientas de gestión de la información de todas las variables y el estudio de covariables relacionadas al dengue que permitan la normalización de los datos relacionados y favorezcan el análisis, la correlación y sienten las bases para un sistema de alertas tempranas para potenciales epidemias del dengue. Se pretende construir una herramienta que permita actuar como repositorio centralizado de datos vinculados al Dengue y otros, sin limitaciones en general, junto con un módulo que permita generar visualizaciones de los datos presentes en el repositorio de manera a realizar análisis descriptivos y exploratorios sobre los datos. Además se propone incluir un módulo que permita de forma experimental entrenar y validar modelos predictivos, basados en técnicas de aprendizaje automatizado de máquinas, para incluso realizar análisis predictivo sobre los datos. Un objetivo primordial de la herramienta es garantizar la reproducibilidad de los experimentos, almacenando los datos utilizados para entrenamiento, validaciones y predicciones, atendiendo así una metodología formal de investigación que facilite a trabajos futuros incorporar nuevos métodos, estrategias y datos, y a su vez puedan ser comparados con trabajos anteriores validando así de forma más objetiva los resultados experimentales, y aportando nuevamente al enriquecimiento de la herramienta mediante nuevos modelos entrenados, y nuevos datos.

Finalmente, la creación de estas herramientas basadas en estándares *open source*, permitirá posicionar al Paraguay a la vanguardia de la investigación e innovación para herramientas de análisis de datos epidemiológicos a nivel internacional.

2. Marco Teórico

Según datos de la Organización Panamericana de la Salud [2], entre el 2003 y 2013, el número de casos de dengue registrados en la región se quintuplicaron. En Paraguay, el 2013 registró un número elevado de casos, alcanzando 128.000 entre enero y junio [3] según los datos reportados por la Dirección General de Vigilancia de Salud (DGVS) del MSPBS. Entre los motivos principales del crecimiento de esta enfermedad se encuentran la movilidad local e internacional de las personas y el crecimiento urbano no planificado [1]. El

primer motivo distribuye el virus geográficamente, y el segundo crea las condiciones óptimas para el desarrollo y propagación del vector de transmisión, el mosquito Aedes aegypti.

La literatura indica que la dinámica de transmisión del dengue depende de la interacción entre varios factores o variables que se pueden dividir en dos grandes grupos: i) variables epidemiológicas y ii) co-variables [1]. El grupo de variables epidemiológicas es el que permite identificar y comprender el fenómeno epidemiológico e incluye: i) la población afectada, ii) el lugar y iii) el tiempo en el que se desarrolla el fenómeno y iv) las características de la enfermedad o casos. Las co-variables son aquellas que afectan directa o indirectamente al aumento o disminución de casos de dengue, por ejemplo: variables socioeconómicas (densidad poblacional, pobreza relativa, etc), variables urbanas (acceso a agua corriente y a servicios sanitarios, etc), variables climatológicas (temperatura, precipitación). El problema radica en que si bien existe una propuesta de modelo de cómo se podrían publicar de manera estándar estos datos, basados en principios de datos abiertos [1], actualmente no existen herramientas open source que permitan agregar los datos de distintas fuentes a una base de datos integrada, estándar y que permita la reutilización de los datos fomentando la investigación.

La DGVS-MSPBS publica gráficos como ser curvas históricas de cantidad de casos por región, o mapas de calor en sus reportes epidemiológicos semanales en formato .pdf. Otros tipos de mapas encontrados en la literatura son: de presencia, de ocurrencia, de riesgo, de incidencia, entre otros [1]. Estos tipos de gráficos y mapas son necesarios para investigar y comprender el comportamiento de la enfermedad y los brotes epidémicos. Sin embargo, no existe en la actualidad una herramienta que permita a los investigadores generar los gráficos dinámicamente según distintos niveles de precisiones geográficos y temporales. Como resultado de esta propuesta se pretende aplicar las TICs para crear una herramienta.

El campo de análisis predictivo, mediante técnicas de machine learning, data mining y otras, busca realizar un pronóstico de futuro rápido y preciso permitiendo tomar decisiones preventivas más efectivamente. Si bien existen diversos modelos predictivos descritos en la literatura, podemos observar tres limitaciones importantes en los mismos: i) las implementaciones no se encuentran disponible para su reutilización, ii) los datos utilizados tampoco se encuentran fácilmente disponibles en formatos procesables por máquinas, y iii) los distintos modelos no comparten una estructura modular que permita adaptarlos y re-utilizarlos, facilitando la investigación y la evaluación de la precisión de los mismos. Se propone la creación de una herramienta open source genérica y modular, que permita la investigación, implementación y publicación de modelos predictivos abiertos y reutilizables, que beneficien a todos los investigadores en el área del dengue. Se pretende utilizar esta

herramienta para impulsar la investigación de alumnos de tesis de grado y postgrado creando nuevos modelos de alertas tempranas para el dengue.

3. Planteamiento del Problema

El problema principal que este trabajo pretende resolver es la falta de una herramienta que facilite la formulación de modelos de predictivos de número de casos de dengue, mediante el acceso centralizado a los datos, a las técnicas de predicción, y a la comparación de los resultados obtenidos con otros modelos ya realizados, para poder prever con cierta anticipación y con mayor precisión la cantidad de casos de dengue que se desarrollarán en una región y tiempo determinados, para así permitir a las autoridades tomar mejores decisiones en el tiempo oportuno y realizar una mejor estimación de recursos y control más efectivo de la propagación de la enfermedad disminuyendo así el número de víctimas potenciales.

De este problema principal se derivan los siguientes problemas específicos:

- Los datos utilizados en diferentes investigaciones sobre modelos de alertas tempranas no están públicos y se encuentran descentralizados en muchos casos
- Los resultados de los diferentes modelos obtenidos no se pueden comparar entre sí porque los datos utilizados no están disponibles
- Los expertos de dominio no siempre tienen las herramientas informáticas necesarias para realizar análisis más profundos con sus datos

4. Hipótesis

Las hipótesis que presenta el proyecto son las siguientes:

- La disponibilización de las notificaciones de casos de dengue en formatos que pueden ser procesados automáticamente por máquinas y bajo un modelo de datos estándar predefinido permitirá la utilización de dichos datos para la investigación e innovación en la gestión de la información de datos epidemiológicos.
- La creación de una herramienta para automatizar el proceso de análisis estadístico de datos para obtener resultados en el menor tiempo posible permitirá a los investigadores el análisis y mejorar el entendimiento de las variables y covariables que inciden en el aumento del número de casos de Dengue.
- El diseño de un modelo predictivo extensible que integre notificaciones de casos de dengue, variables climáticas y otras variables relacionadas que inciden sobre el comportamiento del vector de transmisión permitirá la creación de modelos de

predicción que correlacionen estas variables y encuentren cómo estas afectan a la enfermedad.

- El desarrollo de una herramienta *open source*, genérica y modular, reutilizable y extensible que utilice el modelo predictivo diseñado para procesar los datos y generar alertas tempranas permitirá la colaboración entre investigadores realizando extensiones y mejoras en los modelos de alertas.

5. Justificación

La justificación del proyecto se da tomando tres dimensiones diferentes: la científica, social y contemporánea, la cuales se describen a continuación:

- **Científica**: Actualmente las investigaciones sobre dengue realizan un esfuerzo considerable para obtener los datos necesarios. Los datos del dengue son tratados como una mercancía de lujo a la que solamente grupos de investigación con muchos recursos o contactos pueden acceder. Con este proyecto se pretende cambiar este *status quo*, permitiendo a todos los grupos de investigación acceder a los datos de manera rápida, gratuita y eficiente. Utilizando los datos históricos del dengue en Paraguay, se atraerá la atención de muchos grupos de investigación internacionales que pueden basar sus investigaciones e innovaciones en datos de la realidad paraguaya, lo que aportará de gran manera al entendimiento y propuesta de soluciones para nuestro contexto. Además, la creación de un framework integrado a una capa de acceso a datos públicamente disponible e integrado a una herramienta de visualización de resultados será un aporte científico en sí mismo, que servirá de base y fomentará la investigación en el campo de alerta temprana de dengue. Cada uno de los modelos creados serán un aporte a la ciencia en Paraguay, fomentando la formación de nuevos investigadores. Todos los resultados esperados permitirán afianzar un área innovativa de investigación en el Paraguay, la de datos abiertos aplicados al dengue y el uso de los mismos en sistemas de alertas tempranas. Esto colocará al Paraguay a la vanguardia de esta área de investigación.
- **Social**: Actualmente el MSPBS no utiliza los datos históricos recabados ni los correlacionan a otras variables para predecir el comportamiento de la enfermedad y en consecuencia poder tomar acciones preventivas, sino que las acciones son más bien reactivas (existe un reporte de caso y se trata de minimizar la ocurrencia de una epidemia por medio de la cuarentena y tratamiento adecuado del caso, y de la fumigación y eliminación de criaderos de mosquitos, atacando al vector de transmisión). La herramienta de análisis de datos y los modelos predictivos ayudarán a comprender y correlacionar de manera más eficiente los diferentes elementos que fomentan el inicio de una epidemia de dengue y permiten su persistencia en el tiempo. Con esta información, se podrán tomar acciones preventivas bien específicas, dados los recursos limitados, que minimicen los brotes del dengue a

nivel local. Todas estas acciones tendrán un impacto directo en la sociedad paraguaya ya que se disminuirá la carga social de las epidemias de dengue y los costos de la gestión reactiva ante las mismas.

- **Contemporánea**: Existe un auge de proyectos de investigación internacionales, especialmente en Europa (ej. Denfree, IDAMS; DengueTools, VMerge, EdeNext) en temas relacionados al dengue. Sin embargo, uno de los problemas principales de todos estos proyectos es la falta de datos públicamente accesibles basados en modelos estándares que fomenten la reusabilidad e integración de los datos. Esta propuesta utilizará las herramientas e investigaciones de un área nueva y en auge de investigación, los datos abiertos, para ayudar a solucionar este inconveniente, sirviendo de base para los grupos de investigación locales e internacionales con el fin de mejorar el entendimiento, mitigación y solución del impacto del dengue.

6. Objetivos

A continuación se detallan tanto el objetivo principal como los objetivos específicos de este proyecto.

Objetivo Principal:

- Reducir el impacto del dengue en el Paraguay mediante la investigación y la mejorar la capacidad de gestión de la información de los datos epidemiológicos del dengue mediante una herramienta que permita analizar dinámicamente las variables y co-variables relacionadas al dengue en el Paraguay, con la capacidad de integrar modelos automatizados de alertas tempranas que permita a los organismos competentes tomar las acciones necesarias ante potenciales epidemias de dengue.

Objetivos Específicos:

- Disponibilizar las notificaciones de casos de dengue en formatos que puedan ser procesados automáticamente por máquinas y bajo un modelo de datos estándar predefinido.
- Crear una herramienta para automatizar el proceso de análisis estadístico de datos para obtener resultados en el menor tiempo posible.
- Diseñar un modelo predictivo extensible que integre notificaciones de casos de dengue, variables climáticas y otras variables relacionadas que inciden sobre el comportamiento del vector de transmisión.
- Desarrollar una herramienta open source, genérica y modular, reutilizable y extensible que utilice el modelo predictivo diseñado para procesar los datos y generar alertas tempranas.

7. Metodología

La metodología de investigación abarca dos grandes áreas: datos abiertos y sistemas de alertas tempranas. Considerando los datos abiertos aplicados al dengue, el trabajo de Pane et. al [1], ha dado los primeros pasos en definir el marco conceptual en donde se definen los elementos del Ecosistema de Datos Abiertos (EDA) para el Dengue. Este ecosistema está dado por el conjunto de actores relevantes y los factores que se encuentran involucrados en la recolección, publicación y uso de los datos abiertos. Tomando como base esta caracterización, se plantea el desarrollo de una herramienta que facilite la recolección, integración y publicación de los diferentes datos necesarios para el estudio y análisis del dengue basada en estándares de datos abiertos con licencias que permitan el uso libre de los mismos. Para esta implementación, se tomará como base el modelo propuesto en [1] y se extenderá para incluir las co-variables necesarias para el desarrollo de los modelos de alerta temprana.

En base a los datos recabados y disponibilizados, se facilitará el análisis mediante el uso de herramientas dinámicas de análisis de datos que permita crear diversos tipos de gráficos (barras, líneas, treemaps) y mapas, lo que facilitará a los investigadores realizar los análisis y correlaciones necesarios mejorando la comprensión del dengue y las condiciones necesarias para las epidemias. Esta herramienta de código fuente abierto constituirá un aporte novedoso para la investigación, ya que actualmente los investigadores deben conseguir con gran esfuerzo los datos, integrarlos manualmente, y utilizar diversas herramientas para analizar los datos. Este aporte al estudio de la epidemiología del dengue, permitirá que las conclusiones tomadas en las investigaciones científicas sean replicables desde el punto de vista de los datos, y del proceso de análisis utilizado para llegar a las conclusiones, mejorando de gran manera el proceso científico nacional e internacional en materia de dengue.

Se realizará una evaluación del estado del arte de las diferentes técnicas existentes para la creación de sistemas de alertas tempranas. Dada esta evaluación, y considerando los tipos de variables y co-variables a ser disponibilizadas, se creará un framework que permita acceder a los datos mediante su interfaz programática (API) y que permita su extensión para la implementación efectiva de diferentes modelos de alertas tempranas. El framework incluirá todas las funcionalidades necesarias para la evaluación de la precisión y calidad de las alertas, en comparación con otros modelos ya implementados en el framework. La creación de este framework y la definición de las métricas a ser utilizadas para la evaluación de las implementaciones serán una investigación en sí misma, ya que en la actualidad no existe un framework similar para el dengue integrado a i) una capa de acceso a datos comunes y ii) visualización de los resultados. Esta investigación será desarrollada por

tesistas de grado y/o de maestría. Dado que se plantea que el framework extensible y de código fuente y licencia abiertos, esto facilitará a los investigadores y estudiantes de tesis de grados y maestrías crear nuevos modelos de alertas fácilmente comparables entre sí.

Finalmente, dado el framework base, y la evaluación del estado del arte previamente realizada, se realizarán al menos 3 investigaciones para la implementación de modelos de alertas tempranas basados en diferentes paradigmas, con el fin de comparar las prestaciones de cada uno de ellos, no solamente desde el punto de vista de calidad de las alertas, sino también en términos de tiempo de ejecución y uso de recursos y tipos de datos. De esta forma se pretende validar experimentalmente la herramienta propuesta, y cómo ésta puede ser utilizada para el análisis descriptivo, exploratorio y predictivo del número de casos de dengue, y de esta forma poder contar con alertas tempranas que ayuden a mitigar o encarar potenciales riesgos de epidemias.

8. Cronograma de Actividades

En la siguiente tabla se detalla el cronograma de actividades seguido

	6/17	7/17	8/17	9/17	10/17	11/17	12/17	1/18	2/18	3/18	4/18	5/18	6/18	7/18	8/18	9/18	10/18	11/18
1 - Gestión de Proyecto	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
1.1 - Gestión administrativa del proyecto	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2 - Herramienta de recolección y publicación de datos de dengue basada en estándares de datos abiertos.	■	■	■	■	■	■	■	■	■	■	■	■	■					
2.1 - Construcción del Marco conceptual	■	■	■	■	■													
2.2 - Desarrollo y validación de la herramienta						■	■	■	■	■	■	■	■					
3 - Herramienta de análisis dinámico de datos relacionados al dengue						■	■	■	■	■	■	■	■	■	■	■	■	■
3.1 - Construcción del Marco conceptual						■	■	■	■									

3.2 - Desarrollo y validación de la herramienta										■	■	■	■	■	■	■	■
4 - Framework extensible de modelos de alertas tempranas sobre potenciales epidemias de dengue	■	■	■	■	■	■	■	■	■	■	■	■	■				
4.1 - Construcción del Marco conceptual	■	■	■	■	■												
4.2 - Desarrollo y validación de la herramienta						■	■	■	■	■	■	■	■				
5 - Modelos de alertas tempranas para potenciales epidemias del dengue						■	■	■	■	■	■	■	■	■	■	■	■
5.1 - Construcción del Marco conceptual						■	■	■	■								
5.2 - Desarrollo y validación de la herramienta										■	■	■	■	■	■	■	■
5.3 - Presentación de avances del proyecto en congresos, seminarios nacionales y/o internacionales																■	■

9. Resultados Obtenidos

Los resultados obtenidos mediante este proyecto son los siguientes:

1) Una herramienta de recolección y publicación de datos de variables y co-variables del dengue basada en estándares de datos abiertos que fomenta la utilización de dichos datos para la investigación e innovación en la gestión de la información de datos epidemiológicos.
2) Una herramienta de análisis dinámico de datos relacionados al dengue que permite a los investigadores aplicar distintas metodologías de análisis de datos para visualizar patrones y correlaciones entre las diferentes variables y co-variables relacionadas al dengue.
3) Un framework extensible que permite la inclusión de modelos de alertas tempranas sobre potenciales epidemias de dengue. La herramienta está basada en estándares

open source lo que permite a investigadores realizar extensiones y mejoras en los modelos de alertas.

4) Implementación de tres modelos de alertas tempranas para potenciales epidemias del dengue, basados en las herramientas de recolección y publicación de datos e integrados a la herramienta extensible de alerta temprana, las cuales son:
    a) Predicción de casos de Dengue utilizando Redes Neuronales Artificiales: para la predicción de casos de dengue utilizando redes neuronales, se utilizó el estudio previo realizado por Ughelli et al. [12], replicando el mismo utilizando, para la realización de un modelo de predicción de casos en la ciudad de Asunción con una semana de anticipación las co-variables de: cantidad de casos de esta semana, temperatura máxima media registrada con 11 semanas de anticipación, la humedad mínima media alcanzada con 12 semanas de anticipación y la lluvia registrada con tres semanas de anticipación, alcanzando un error, representado por el *Root Mean Square Error* (RMSE) de **0,01,** el cual se considera bajo.
    b) Predicción de Brotes de Dengue utilizando Árboles de Decisión: para la predicción de brotes de dengue utilizando árboles de decisión, se utilizó el estudio previo utilizado por Ojeda et al. [13], replicando el mismo, utilizando, para la realización de un modelo de predicción de brotes en el Paraguay con una semana de antelación las co-variables relacionadas a la cantidad de casos y a la población, no así las climáticas, obteniendo un *accuracy* de **89.9%**, donde, con un máximo de 100% este valor se considera un buen resultado.
    c) Predicción de casos de Dengue utilizando Regresiones Lineales Simples: el caso de las regresiones lineales es el modelo más básico implementado, usando como co-variable solamente la cantidad de casos de la semana pasada, teniendo un RMSE de **0.2**, que es el error más elevado comparado con los resultados obtenidos con las demás técnicas

    Estos tres entrenamientos, validaciones y modelos de predicción se pudieron realizar enteramente usando la herramienta desarrollada, resultando así efectiva la prueba de concepto y también dejando tres modelos ya entrenados para poder realizar predicciones de casos o brotes de dengue.

La plataforma se encuentra públicamente accesible mediante la siguiente dirección en Internet y credenciales de prueba, además del repositorio de código fuente.

URL de acceso al repositorio de código fuente:
https://github.com/cdsparaguay/daae-framework

URL de acceso a la herramienta:

http://datos.cds.com.py/

Credenciales para pruebas
Email: contacto@cds.com.py
Clave: dengue.2018

10. Conclusiones

Este trabajo presentó el problema actual existente en cuanto a la recolección, utilización y reutilización de datos relacionados al dengue, así como la falta de una herramienta open source que permita realizar análisis y predicciones con estos datos.

A consecuencia de estos problemas, se planteó como solución la implementación de una herramienta que sea capaz de resolver estos problemas, mediante la utilización de una arquitectura que sea capaz de manejar gran cantidad de datos, visualizarlos y realizar análisis sobre ellos de una manera eficiente. Con la solución implementada se pudo llegar a los cuatro resultados esperados del proyecto: una herramienta de recolección, una de análisis, un framework extensible para realizar modelos de predicción y finalmente la implementación de tres modelos utilizando las tres herramientas desarrolladas.

Pudo constatarse la importancia de la herramienta, y la posibilidad de obtener modelos entrenados con buen desempeño para el análisis predictivo del número de casos de dengue, así como la clasificación e incidencia de las covariables en la predicción de potenciales brotes de la enfermedad.

Como trabajos futuros, se iniciará un contacto con las autoridades del MSPBS, encargadas de la Vigilancia de la Salud de manera a proveer la herramienta desarrollada y que puedan apropiarse de la misma. Además los investigadores involucrados en el proyecto continuarán utilizando la herramienta, y divulgarán la misma para que pueda albergar más datos, de potencialmente otros proyectos de investigación y temas, y permitir así enriquecer la herramienta mediante su uso.

11. Bibliografía

[1] Juan Pane, et. al. Dengue Open Data. Open Data Research Symposium, IOCD-2015.
[2] http://www.paho.org/hq/index.php?view=article&id=9657&lang=es
[3] Boletín Epidemiológico Semanal #25-2015. DGVS/MSPBS - Paraguay.

[4] Maíra Aguiar, et. al. Descriptive and predictive models of dengue epidemiology: an overview. 2012.
[5] Denguenet Web Site. WHO. http://apps.who.int/globalatlas/default.asp
[6] http://tc1.sms.fortaleza.ce.gov.br/simda/dengue/monitoramento
[7] http://www.healthmap.org/dengue/en/
[8] Lowe et. al. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. 2010.
[9] Lowe et. al. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. 2012.
[10] Buczak et. al. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. 2012.
[11] State of the art in the Prevention and Control of Dengue in the Americas. 2014.
[12] Ughelli, V., et al. Predicción de casos de dengue en el Paraguay utilizando Redes Neuronales Artificiales. Tesis de Grado. Universidad Nacional de Asunción.2017
[13] Ojeda, V. et al. Estandarización de Reporte de Casos y Predicción de Brotes de Dengue en Paraguay en Base a Datos Abiertos. Tesis de Grado. Universidad Nacional de Asunción. 2016.