DATOS ABIERTOS Y ALERTAS SOBRE DENGUE

Juan Pane [1], Julio Paciello [2], Yohanna Lisnichuk [3], Héctor Martínez [4], Santiago Valdez [5], Nelson Durañona [6]
jpane@ceamso.org.py[1], juliopaciello@cds.com.py[2], {yohanitalisnichuk [3], hmartinez.py [4], santiago.kenshinvaldez [5], pdmnelson [6]}@gmail.com
Centro de Estudios Ambientales y Sociales; Centro de Desarrollo Sostenible S.A., Asunción, Paraguay
PROGRAMA PROCIENCIA – CONVOCATORIA 2015 - PROYECTO PINV15-327

RESUMEN

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. 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, basada en una necesidad endógena, de herramientas de gestión de la información de todas las variables y el estudio de co-variables 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. 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.

PROBLEMA

El problema a resolver es la actual falta de herramientas *open source* que permitan agregar datos de distintas fuentes a una base de datos integrada, estándar y que permita la reutilización de los datos fomentando la investigación. Además, la falta de un framework extensible, escalable y *open source*, orientado a epidemiólogos, de *machine learning* específico para predicciones sobre enfermedades transmisibles por vectores, como el dengue, que permita usar, entrenar, comparar e implementar modelos de predicción; no pudiendo actualmente comparar fácilmente diversos algoritmos de predicción de casos ni reutilizar los datos utilizados por estos.

SOLUCIÓN IMPLEMENTADA

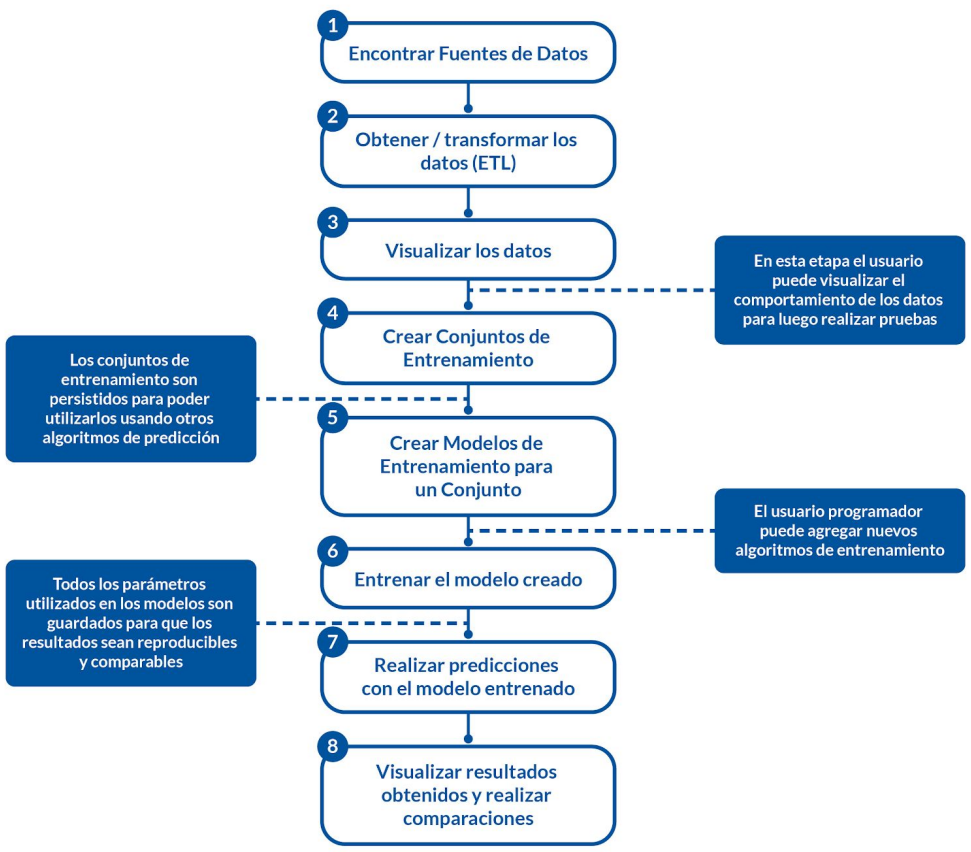

FIgura 1. Componentes de la arquitectura del proyecto

La solución desarrollada consiste en una Plataforma de Análisis de Datos Abiertos y Alertas sobre Dengue. Compuesto por el Módulo de Recolección y Publicación de Datos, Módulo de Análisis y Visualización Dinámica de datos y el Módulo de Alertas Tempranas.

El Módulo de Recolección y Publicación de Datos posibilita a un usuario, potencial experto de dominio en epidemiología, encontrar fuentes de datos de interés (paso 1 de la Figura 1) y crear y subir procesos de extracción, transformación y carga (ETL) a la plataforma, calendarizando su ejecución de manera a permitir la carga periódica de datos al repositorio de datos centralizado de esta. Adicionalmente, permite publicar estos datos de manera a que sean accesibles para otros usuarios de la plataforma, utilizando estándares en formato de datos abiertos. El módulo contempló el uso de un gestor NoSQL dada la diversidad de datos y atributos que debe permitir cargar, por este motivo la persistencia se realiza en utilizando por debajo el gestor MongoDB (Figura 2) . Los resultados de las ejecuciones de las transformaciones son almacenados para permitir al usuario consultar el estado de sus procesos en un periodo de tiempo dado. Entre las funcionalidades proveídas por una API (Interfaz de Programación de Aplicaciones) es posible obtener el listado de los conjuntos datos públicamente accesibles, la posibilidad de selección de un conjunto de datos específico y obtener sus registros, permitiendo la consulta y reutilización de estos. El Módulo de Análisis Dinámico permitirá la generación de visualizaciones (paso 2 de la Figura 1), facilitando a los investigadores realizar los análisis y correlaciones necesarios mejorando la comprensión de los datos y las condiciones que potencialmente podrían disparar epidemias. Como se observa en la Figura 2, las herramientas utilizadas en este módulo son: Elasticsearch, utilizado como repositorio auxiliar de datos y Kibana, cuya función es la generación de visualizaciones. Se contempla el desarrollo de un módulo de autenticación y privacidad de datos basados en Nginx. Finalmente, con el Módulo de Alertas Tempranas se implementó un framework extensible para permitir la inclusión de modelos de alertas tempranas sobre potenciales epidemias de dengue. Para poder crear modelos de entrenamiento, la plataforma permite utilizar los datos cargados desde el Módulo de Recolección de datos y aplicar diferentes algoritmos de entrenamiento con diferentes parámetros para poder crear modelos de predicción (paso 5). La plataforma ya incluye como prueba de concepto tres algoritmos que pueden ser utilizados, sin embargo, como está basada en estándares open source, el usuario del tipo programador puede añadir otros algoritmos de entrenamiento a la plataforma. Además, todos los parámetros utilizados en el modelo son guardados para permitir que los resultados obtenidos sean reproducibles y comparables con otros resultados que podrán ser visualizados y usados para realizar predicciones con datos nuevos (paso 7 y 8)

En la Figura 2, se pueden observar las herramientas utilizadas para desarrollar el framework, las cuales son: el lenguaje de programación Java, en su versión 8, el framework de aplicaciones Spring Boot junto con su servidor embebido Tomcat para el despliegue de la aplicación, exponiendo servicios REST para interactuar con las funcionalidades de la aplicación. Para almacenar las configuraciones de los usuarios, los resultados obtenidos por los modelos y los metadatos de los mismos se utiliza una base de datos relacional Postgresql. Para aumentar la escalabilidad y rapidez del entrenamiento de los algoritmos de predicción se utiliza Apache Spark, a modo de distribuir las tareas de entrenamiento en un clúster de computadoras en las que el proceso se ejecutará de manera paralela. Esta forma de realizar procesos distribuidos ofrecerá el entrenamiento de los algoritmos de manera más rápida y distribuida. Los datos a ser utilizados para el entrenamiento son cargados temporalmente en un clúster de almacenamiento Hadoop, para que todos los nodos de trabajo del clúster puedan acceder a ellos de forma paralela.

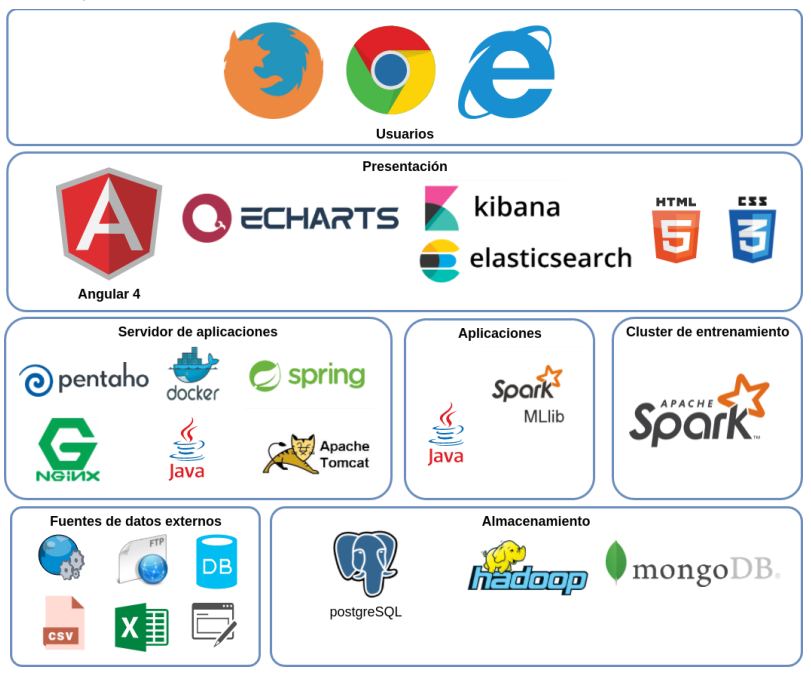

Figura 2. Tecnologías Propuestas

RESULTADOS ESPERADOS

RE#1: Herramienta de recolección y publicación de datos de variables y co-variables del dengue basada en estándares de datos abiertos que fomente la utilización de dichos datos para la investigación e innovación en la gestión de la información de datos epidemiológicos.

RE#2: Herramienta de análisis dinámico de datos relacionados al dengue que permita 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.

RE#3: Framework extensible que permita la inclusión de modelos de alertas tempranas sobre potenciales epidemias de dengue. La herramienta estará basada en estándares open source lo que permitirá a investigadores realizar extensiones y mejoras en los modelos de alertas.

RE#4: Investigación, implementación y evaluación de 3 modelos de alertas tempranas para potenciales epidemias del dengue, basados en las herramientas de recolección y publicación de datos (RE #1) e integrados a la herramienta extensible de alerta temprana (RE #3)

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. Debido a estos problemas, se plantea la implementación de una herramienta que sea capaz de solucionarlos, 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 esto, se pretende llegar a los cuatro resultados esperados mencionados en la sección anterior: 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.

REFERENCIAS

- Juan Pane, et. al. Dengue Open Data. Open Data Research Symposium, IOCD-2015.
- Boletín Epidemiológico Semanal #25-2015. DGVS/MSPBS - Paraguay.
- Maíra Aguiar, et. al. Descriptive and predictive models of dengue epidemiology: an overview. 2012.
- http://www.healthmap.org/dengue/en/
- Lowe et. al. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. 2010.

“Este proyecto es financiado por el CONACYT a través del Programa PROCIENCIA con recursos del Fondo para la Excelencia e investigación – FEEI del FONACIDE”