Control y seguimiento de la transmisión del dengue a través del mosquito tigre

Descripción general del proyecto

El dengue es una enfermedad infecciosa transmitida a través de los mosquitos tigre. Presenta 4 niveles de gravedad, pudiendo llegar a ser potencialmente mortal en los niveles más altos. En los últimos años se ha incrementado notablemente su presencia especialmente en zonas con climas tropicales. Según la *Organización Mundial de la Salud* se cree que, en la actualidad, más de la mitad del planeta está en riesgo de contraer la enfermedad.

En el presente trabajo se intenta proveer de los datos y las herramientas necesarias para facilitar el control y seguimiento de esta enfermedad así como de los transmisores de ella. Los principales objetivos a perseguir son:

* Proveer de fuentes de datos y metodologías de recolección útiles acerca del dengue y sus principales transmisores
* Proveer procesos de curación y limpieza de datos que faciliten el análisis posterior de los propios datos
* Estandarización de los datos en repositorios
* Extracción de conclusiones útiles acerca del estado de la enfermedad
* Caracterización del comportamiento y la presencia del mosquito tigre

Con la consecución de estos objetivos esperamos obtener los siguientes resultados:

* Generar datos útiles que permitan el estudio profundizado de la problemática
* Generar datos fácilmente reutilizables y explotables
* Impulsar la colaboración de organizaciones e individuos alrededor del globo tanto en el análisis como en la recolección
* Generar medidas preventivas más eficaces
* Ayudar a frenar o erradicar, a largo plazo, la enfermedad del dengue

Lo que se pretende conseguir es facilitar y animar a las distintas organizaciones a generar mejores datos para un análisis más eficaz que permita actuar de forma óptima en lo que respecta a la enfermedad y su transmisión.

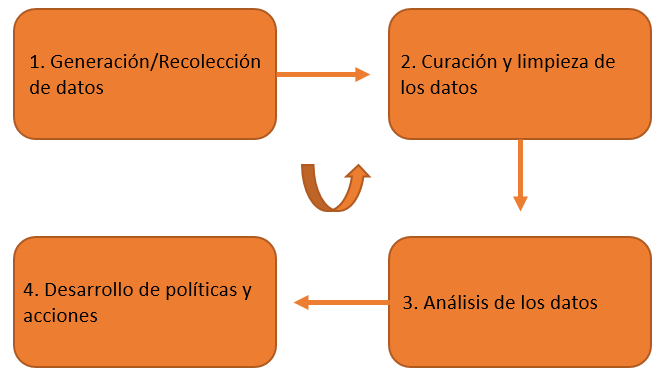
Creemos que se necesitan ciertas condiciones básicas para que este proyecto alcance sus propósitos con éxito, la principal es que se debe utilizar *software* abierto y fuentes *Open Data*, esto relajará las condiciones necesarias para que un individuo u organización pueda contribuir al proyecto.

Este proyecto recae, en un principio, únicamente sobre nuestro grupo por lo que en un futuro necesitaremos apoyo de un mayor número de organizaciones e individuos con mayor capacidad productiva y recursos técnicos. Para conseguir esto intentaremos dar comienzo al proyecto con unos conjuntos de datos limitados y unos análisis restringidos a algunos países claves. A través de actividades divulgativas pretendemos llegar a organizaciones más grandes que nos permitan una temprana consecución de nuestros objetivos, llevándonos así a los resultados esperados de forma más directa. A pesar de ser un proceso relativamente largo creemos que goza de la ventaja de la continua consecución de objetivos, haciendo que el esfuerzo se vea recompensando de forma continua durante la consecución del proyecto.

En la siguiente sección del presente informe se presentará el plan a desarrollar para hacer el proyecto viable y sostenible a medio-largo plazo.

Pipeline del proyecto

El proyecto seguirá unas etapas bien diferenciadas orientadas a cumplir los objetivos anteriormente propuestos, para ello se definirá una *pipeline* del proceso, la cual se refinará y optimizará para facilitar e impulsar su uso en organizaciones con más poder y capacidades a fin de conseguir los resultados esperados propuestos. A continuación se muestra el esqueleto de esta *pipeline*:



Veamos con más detalle cada una de estas etapas:

* **Generación/recolección de datos**: A través de la ejecución de esta etapa se espera encontrar la tipología de *dataset* útil que permita la consecución de los resultados ya propuestos. Una vez identificados se podrá desarrollar procesos de generación más eficientes de los propios datos.
* **Curación y limpieza de los datos**: En esta etapa se desarrollaran metodologías para curar los datos en crudo o casi en crudo a fin de obtener una representación de estos que facilite el posterior análisis. Se podrán crear *pipelines* que automaticen este proceso gracias a la continua adquisición de experiencia sobre los datos y la repetición del proceso.
* **Análisis de los datos**: Una vez tenemos los datos ya procesados se podrán analizar en búsqueda de resultados útiles acerca de la enfermedad y su transmisión. Con la adquisición de experiencia se podrán identificar técnicas útiles para estos tipos de datos e información contenida en ellos.
* **Desarrollo de políticas y acciones**: Una vez se ha conseguido extraer la información útil de los datos se podrá desarrollar políticas y tomar acciones más eficientes en cara a conseguir los resultados esperados de este proyecto.

Lo interesante de esta *pipeline* es que a medida que más organizaciones e individuos participen en ella se retroalimentará e irá evolucionando hacia una *pipeline* única para este tipo de problemática y que, dados unos pasos muy concretos, permita la ejecución de acciones y políticas para actuar de forma más eficiente en la lucha contra el dengue y su expansión.

A medida que se adquiera más experiencia en la recolección de datos se podrán definir metodologías estandarizadas para recopilar datos de calidad, de forma que con el paso del tiempo si algún país u organización quiere estudiar el dengue no tenga más que seguir estas pautas para la correcta recolección. Una vez se tienen ciertos datos en una estructura común se podrán aplicar técnicas de curación con una eficacia ya comprobada con estudios anteriores. Estos datos se podrán analizar con *scripts* ya utilizados y comprobados para poder obtener una información con la que se sabe ya que sirve para tomar medidas y acciones eficaces.

En el presente proyecto llevaremos a cabo una ejecución inicial de esta *pipeline* definiendo y llevado a cabo unas fases que nos permitan en cierta manera cumplir los resultados esperados del proyecto.

Descripción de las fuentes de los datos

En esta ejecución inicial del *pipeline* se han utilizado datos de tres fuentes distintas. En este apartado describiremos tanto las fuentes como el contenido de los propios *datasets*:

* **The global compendium of Aedes aegypti and Ae. albopictus occurrence**: En este *dataset* se muestran avistamientos de las especies *Aedes Aegypti* y *Aedes Albopictus*. Son dos especies de mosquitos distintas pero que viven bajos climas y condiciones similares y ambos transmiten el dengue. Los datos cubren el intervalo comprendido entre los años 1960 y 2014. Contiene datos geográficos de los distintos avistamientos por lo que podremos localizar el lugar de ocurrencia de cada avistamiento. El archivo únicamente se puede descargar en formato csv, se explorará más en detalle en secciones posteriores.

Estos datos se han extraído de [Dryad](https://datadryad.org/). Esta web es un repositorio interdisciplinar que pone a disposición de cualquier usuario los datos utilizados en diversos estudios científicos, a fin de garantizar la replicabilidad y la reutilización.

* **Global compendium of Aedes albopictus occurrence:** En este *dataset* también se muestran avistamiento de mosquitos pero en este caso únicamente de la especie *Aedes Albopictus*. Aunque parece ser el mismo *dataset* que el anterior se aprecian diferencias si se estudia más en detalle, es por ello que utilizaremos los dos a fin de utilizar la información contenida en ambos y hacer el posterior análisis más robusto. En este dataset las coordenadas geográficas parecen estar en otro formato, además se aprecian nuevas columnas no presentes en el anterior. El formato del archivo es también csv.

Este *dataset* se ha extraído de la web [GBIF](https://www.gbif.org/) (Global Biodiveristy Information Facility). Es una organización internacional cuya principal labor es unificar los datos de biodiversidad haciéndoles accesibles a una mayor cantidad de público. Estos datos son proporcionados por muchas instituciones y tratan principalmente las temáticas relativas a la distribución de las especies alrededor del globo.

* **Reported cases of dengue fever in the Americas:** En este *dataset* se muestran los casos de Dengue para el periodo comprendido entre los años 1980 y 2015. Se distingue entre dengue, dengue grave y muertes ocasionadas por la enfermedad. Aunque es posible obtener información sobre cualquier país de las Américas nosotros solo hemos utilizado la información de Brasil y Estados Unidos debido a limitaciones encontradas en los dos *datatsets* anteriores (se explicarán en futuras secciones). El archivo se ha descargado en formato csv.

Estos datos se han extraído de la web [OPS](https://www.paho.org/hq/index.php?lang=es) (Organización Panamericana de la Salud). Esta organización, afiliada a la OMS, se encarga de promover medidas que favorezcan la salud y el bienestar en los países del continente americano. Dispone de gran cantidad de datos referentes a temas de salud pública, entre los que se encuentran los de seguimiento de enfermedades graves como el dengue.

Data Management Plan (DMP)

El DMP se adjunta en un archivo aparte de este. Está basado en la plantilla para proyectos de la Comisión Europea (H2020). En algunos casos se ha supuesto la generación de datos aunque, como bien se ha explicado aquí, la generación de datos se definirá una vez se haya llevado a cabo varias iteraciones del *pipeline*, de esta forma tendremos una generación de datos óptima para lograr los objetivos y resultados del presente proyecto.

Se ha intentado plasmar en el DMP la naturaleza colaborativa y altruista del proyecto aquí definido.