

# UNIVERSIDAD DEL VALLE DE GUATEMALA

Data Science

Sección 20



*Excelencia que trasciende*

**DELVALLE**  
GRUPO EDUCATIVO

## “Proyecto final”

CHRISTOPHER EMANUEL ALEXANDER GARCIA PIXOLA 20541

MARIA ISABEL SOLANO BONILLA 20504

ALEJANDRO JOSE GOMEZ HERNANDEZ 20347

ROBERTO VALLECILLOS CHINCHILLA 20441

GABRIEL ALEJANDRO VICENTE LORENZO 20498

**GUATEMALA, Octubre 2023**

# Índice

<b>Introducción.....</b>	<b>2</b>
<b>Objetivos.....</b>	<b>4</b>
Generales:.....	4
Específicos:.....	4
<b>Marco teórico.....</b>	<b>5</b>
<b>Metodología.....</b>	<b>6</b>
<b>Resultados y Análisis de Resultados.....</b>	<b>8</b>
<b>Referencias:.....</b>	<b>14</b>

# Introducción

El creciente avance de las enfermedades transmitidas por mosquitos en diversas regiones del mundo, incluyendo Guatemala, ha configurado un escenario crítico que requiere de acciones contundentes para su control y mitigación. Entre las enfermedades transmitidas por mosquitos, el Dengue, Zika y Chikungunya, propagadas principalmente por el mosquito *Aedes aegypti*, constituyen una tríada de desafíos sanitarios que han afectado a grandes segmentos de la población, especialmente a personas embarazadas, individuos sin acceso a atención médica adecuada, y personas de bajos recursos. La situación se complica aún más con la persistencia de la malaria en áreas rurales, un mal transmitido por mosquitos del género *Anopheles*. Estas condiciones, sumadas a problemas socioeconómicos y la resistencia a los insecticidas, han creado una coyuntura que demanda soluciones innovadoras y efectivas. En un país como Guatemala, resulta una de las causas principales de muerte, totalmente evitables, de ahí radica la importancia de realizar un proyecto de esta índole.

Una de las vías para abordar este problema es a través de la identificación precisa y temprana de las especies de mosquitos, lo cual puede facilitar la implementación de medidas de control vectorial más efectivas y específicas. Sin embargo, la identificación de especies de mosquitos es una tarea que tradicionalmente ha requerido de la intervención de expertos entomólogos, lo cual puede ser un proceso lento y no siempre disponible en todas las regiones afectadas. Por este motivo, al tener tecnologías que nos dejen identificar de forma automática y con alta precisión, podemos brindar una solución bastante esperanzadora.

Con el auge tremendo de las tecnologías de aprendizaje profundo y la evolución constante en el campo de la visión por computadora, ahora es posible desarrollar modelos de clasificación automatizada con una alta tasa de precisión. Las redes neuronales convolucionales (CNN), nos han demostrado ser muy buenas y eficaces en la tarea de clasificación de imágenes. De esta forma, se abre una vía prometedora para la identificación de especies de mosquitos a partir de imágenes digitales. Este enfoque no solo puede acelerar el proceso de identificación, sino también hacerlo más accesible y extensible a diferentes regiones, contribuyendo significativamente a las estrategias de control de enfermedades vectoriales.

El presente proyecto se enfoca en el diseño y desarrollo de un modelo de red neuronal que permita la identificación precisa de diferentes especies de mosquitos a partir de un conjunto

de datos compuesto por 10,700 imágenes auténticas de mosquitos. Estas imágenes fueron capturadas por participantes empleando sus dispositivos móviles y posteriormente etiquetadas por expertos entomólogos, proporcionando una base sólida para el entrenamiento y validación del modelo propuesto. Las especies de mosquitos incluidas en el conjunto de datos abarcan a *Aedes aegypti*, *Aedes albopictus*, los géneros *Anopheles* y *Culex*, así como un conjunto de especies conocido como *Aedes japonicus/Aedes koreicus*, todas de relevancia en el contexto de salud pública de Guatemala.

## Objetivos

Generales:

- Desarrollar un modelo de red neuronal que permita la identificación precisa de las diversas especies de mosquitos presentes en las imágenes del conjunto de datos proporcionado, contribuyendo así a una mejor gestión y control de los mosquitos portadores de enfermedades en Guatemala. De esta manera se cumplirá con los requisitos solicitados en el proyecto del curso de Data Science de la Universidad del Valle de Guatemala.

Específicos:

- Alcanzar una precisión mínima del 90% en la identificación de las especies de mosquitos mediante la red neuronal propuesta, al obtener una métrica alta en cuanto a la identificación de especies, se validará el modelo realizado.
- Crear un modelo capaz de procesar imágenes fuera del conjunto de datos proporcionado, manteniendo de igual forma una precisión bastante alta para poder contrastar el modelo pre entrenado con nueva data.
- Seleccionar e implementar la arquitectura de red neuronal más adecuada para la clasificación de imágenes, considerando la naturaleza de los datos, el objetivo del proyecto y la facilidad de implementación, tomando en cuenta lo necesario para realizar el proyecto.

- Evaluar el desempeño del modelo propuesto mediante métricas pertinentes y comparar los resultados obtenidos con otros enfoques y algoritmos de aprendizaje de máquinas, tomando como base la precisión como la métrica definitiva que evalúe al modelo.

## Marco teórico

### Procesamiento de Datos

El procesamiento de los datos es un paso crucial en cualquier proyecto de aprendizaje automático. En este proyecto, se utilizó un conjunto de datos compuesto por 10,700 imágenes de mosquitos, dividido en un conjunto de entrenamiento (80% de las imágenes) y un conjunto de prueba (20% de las imágenes). Las imágenes fueron etiquetadas con anotaciones precisas proporcionadas por expertos entomólogos, incluyendo las coordenadas del cuadro que rodea al mosquito y la categoría de mosquito a la que pertenecen. Además, se proporcionó un archivo CSV con dichas anotaciones.

### Aprendizaje Profundo y Redes Neuronales

El aprendizaje profundo es una subcategoría del aprendizaje automático que se centra en algoritmos que pueden aprender representaciones de los datos a varios niveles de abstracción. Las redes neuronales son la base del aprendizaje profundo y están diseñadas para imitar la manera en que el cerebro humano procesa la información. Para este proyecto, se exploraron diversas arquitecturas de redes neuronales para la clasificación de imágenes, incluyendo redes neuronales convolucionales (CNN), que son especialmente eficaces para tareas de visión por computadora.

### Redes Neuronales Convolucionales (CNN)

Las CNN son una clase de redes neuronales profundas que han demostrado ser muy eficaces en la clasificación y segmentación de imágenes. Estas redes utilizan capas convolucionales para procesar partes de la imagen, identificando características como bordes, texturas y patrones, que luego son combinadas para reconocer formas y objetos complejos en las imágenes. La eficacia de las CNN en tareas de clasificación de imágenes sugiere que podrían ser una opción viable para la identificación precisa de especies de mosquitos en este proyecto.

## Streamlit

Streamlit es una plataforma de código abierto diseñada para simplificar la creación de aplicaciones web interactivas y visualizaciones de datos. Con su enfoque minimalista, permite a los desarrolladores y científicos de datos crear aplicaciones web con facilidad, incluso sin experiencia previa en desarrollo web, gracias a su integración con bibliotecas de visualización de datos populares, la capacidad de crear aplicaciones interactivas y la actualización en tiempo real. Streamlit se caracteriza por su sintaxis clara y directa, que permite a los usuarios crear aplicaciones mediante un flujo de secuencia de comandos. Los elementos de la interfaz de usuario, como gráficos, widgets y componentes interactivos, se incorporan directamente en el código Python, lo que simplifica la construcción de interfaces de usuario dinámicas. Además, Streamlit se integra sin problemas con bibliotecas populares de visualización de datos como Matplotlib, Plotly y Altair, lo que facilita la creación de gráficos y representaciones visuales atractivas.

## Evaluación del Modelo

Para evaluar el desempeño del modelo propuesto, se emplearán diversas métricas como la precisión, la sensibilidad, la especificidad y la matriz de confusión. Además, se comparará el desempeño del modelo con otros algoritmos y enfoques de aprendizaje automático para asegurar que la red neuronal seleccionada es la más adecuada para la tarea en cuestión.

## Metodología

Para esta segunda parte se buscó crear diferentes modelos que permitieran el entrenamiento y predicción de tipos de mosquitos en base a imágenes. Para poder lograr esto se utilizaron los datos provistos por la competencia [MosquitosAlert-2023](#), los cuales habían sido utilizados previamente para el análisis exploratorio.

Se trabajaron 4 modelos diferentes, 3 de ellos fueron realizados utilizando los conocimientos vistos en clase de redes neuronales convolucionales y 1 de estos se realizó a través de un modelo precargado compartido en la competencia anteriormente mencionada para tener una idea de los resultados óptimos. El modelo precargado únicamente se utilizó para obtener una vista general de las clasificaciones y predicciones correctas del set de entrenamiento.

Para los modelos se trabajó con el mismo proceso, en cada iteración del modelo se buscó elevar las métricas de desempeño, y a continuación se explica el desarrollo de los mismos:

- Luego de haber cargado las imágenes de prueba exitosamente y el csv con la información de las mismas se procedió a dividir el conjunto de datos en 3 partes (entrenamiento, prueba y validación)
- Para cargar las imágenes se utilizó una función que obtenía la imagen en sí, evaluaba que esta cumpliera el formato RGB (para mejorar la identificación de mosquitos)
- Se normalizaron y ajustaron de tamaño las imágenes y una vez realizado este proceso con todo el conjunto se procedió a crear un array de numpy para los pasos posteriores.
- Posterior a estos ajustes se realizó la división del conjunto de imágenes y se procedió a generar, con ayuda de TensorFlow, más imágenes con diferentes características para mejorar el entrenamiento del modelo
- Luego de esto se creó el modelo convolucional que dependiendo de la iteración poseía más o menos capas y otras técnicas para regularizar, se utilizó relu, softmax, dropout, entre otras capas y técnicas.
- Con el modelo creado se procedió a realizar la evaluación del mismo y también se guardó el modelo con la finalidad de utilizarlo en la aplicación creada.
- Este mismo procedimiento se realizó en las 3 iteraciones, modificando valores en el modelo o el manejo de imágenes, obteniendo un 63%, 75% y 80% de accuracy en los modelos V1, V2 y V3 respectivamente.

Los modelos y redes mencionados en esta entrega fueron vistas en clase por lo que optamos trabajar con las mismas ya que eran implementaciones conocidas y de las cuales teníamos un conocimiento básico con el cual llegar a un MVP. Se utilizó Colab para realizar las pruebas y el desarrollo de los Notebooks ya que este cuenta con una GPU que permite acelerar el proceso de entrenamiento y brinda un entorno con herramientas útiles para trabajar con modelos de esta índole. Se utilizaron herramientas, lenguajes de programación y librerías con las que ya se había trabajado en el curso como lo son Python, Tensorflow, Pandas, Joblib, Torch, Sklearn, entre otras similares.

En cuanto a la aplicación, se trabajó con Streamlit ya que su implementación es bastante sencilla, ofrece miles de componentes y además se adapta bastante bien a la implementación y muestra de resultados de modelos. En esta primera iteración, la aplicación es bastante

simple e intuitiva, ofrece un apartado para realizar predicciones, un apartado para mostrar los performance de los modelos y un apartado para mostrar toda la información extra (información de los datos, del uso de la aplicación, etc.)

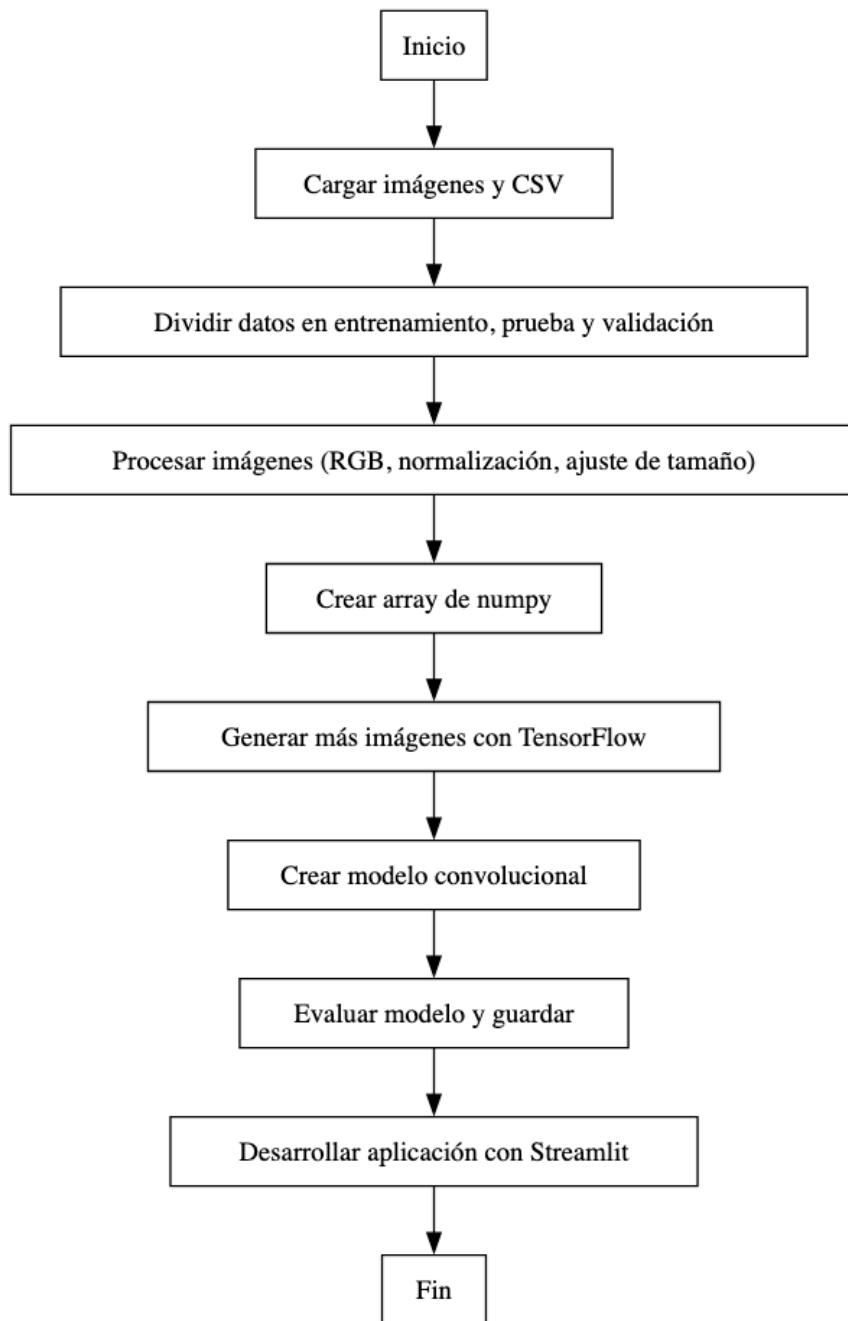


Gráfico No.1: Diagrama de flujo explicativo de metodología

# Resultados y Análisis de Resultados

En el caso de la sección de resultados y análisis de resultados podemos decir que se basó en la medición y evaluación de la eficacia y precisión del modelo de red neuronal desarrollado para identificar las distintas especies de mosquitos a partir de las imágenes. Este análisis es crucial para comprender cómo el modelo se puede usar en estrategias de control vectorial, centrado especialmente en Guatemala con las enfermedades transmitidas por mosquitos. La base de datos, la cuál se detalla en esta sección, contó con 10,700 imágenes auténticas de mosquitos, y esto proporciona una plataforma robusta para entrenar, validar y probar el modelo.

Las tecnologías emergentes han demostrado ser muy efectivas en tareas de clasificación y análisis de imágenes, lo que sugiere un camino bastante bueno para automatizar la identificación de especies de mosquitos.

- **Datos**

Los datos originales, tal como fue mencionado anteriormente, fueron obtenidos a partir de la página Alcrowd para el desafío de MosquitoAlertChallenge2023, como un desafío de detección y clasificación de estos insectos. Los datos brindados fueron una serie de 10, 700 imágenes capturadas por los participantes utilizando sus celulares. El dataset se divide en 8,025 (80%) de las imágenes utilizadas para el entrenamiento del modelo, y 2,675 (20%) para probar el modelo. Cabe resaltar que las imágenes del dataset varían en dimensionalidad. Un ejemplo de una de estas imágenes es el siguiente:



*Img 1. Ejemplar de una imagen de un mosquito albopictus (1024 x 1820)*

Junto a las imágenes, el desafío también provee un archivo de csv con no solo el nombre de la clase del mosquito para cada una de las imágenes del dataset, sino que también las coordenadas de cada una de las esquinas de frontera de dónde exactamente se encuentra el mosquito dentro de la imagen.

Las clases de mosquitos que este dataset presenta son las que se encuentran a continuación, donde 2 son especies y 3 son géneros. En el

- Aedes aegypti - Species
- Aedes albopictus - Species
- Anopheles - Genus
- Culex - Genus
- Culiseta - Genus
- Aedes japonicus/Aedes koreicus - Species complex

Entre los comentarios que las instrucciones del desafío agregan son que la clasificación de las especies de Culex es muy desafiante, por ello se clasifican como un género completo. Y en el caso de la última clase, Aedes japonicus/Aedes koreicus, esta se trabaja como una clase única porque incluso a nivel humano es difícil la clasificación de ellos.

- **Limpieza de datos**

Durante el proceso de obtención de imágenes, tal como se mencionó anteriormente, se verificó que las imágenes fueran a color, ya que es necesario para la clasificación. Al mismo tiempo también se redimensionó todas las imágenes para que estas fuesen 64 x 64. Otra medida que se tomó fue que las etiquetas de los nombres se cambiaron a etiquetas binarias. Con ese proceso realizado se procedió a realizar la separación entre datos de prueba y datos de entrenamiento.

- **Ajuste de parámetros**

El primer modelo se utilizaron 3 capas convolucionales con una función de activación reLu, por último se usó una capa fully connected con relu y un softmax. Como optimizador se utilizó Adam y para la función de pérdida una categorical cross entropy.

En el caso del segundo modelo se utilizó un modelo se utilizaron 3 capas convolucionales con función de activación relu, batch normalization, y max pooling. Luego de ello se aplanó el resultado para el final utilizar una capa fully connected densa con relu y softmax. Nuevamente se utilizó Adam, y categorical crossentropy.

Para el tercer modelo se utilizó como base el primer modelo, solo con la pequeña modificación de que los datos de entrenamiento se normalizaron. El resto de las capas eran similares.

- **Comparación de algoritmos**

Los tres modelos presentados comparten un enfoque común en el procesamiento de imágenes de mosquitos o objetos similares utilizando redes neuronales convolucionales (CNN). Sin embargo, difieren en sus detalles y enfoques específicos.

El Modelo 1 se caracteriza por su sencillez y simplicidad. Emplea capas convolucionales y Max Pooling para extraer características, pero su accuracy resulta relativamente bajo, lo que sugiere que este enfoque inicial se utiliza principalmente para familiarizarse con el dataset.

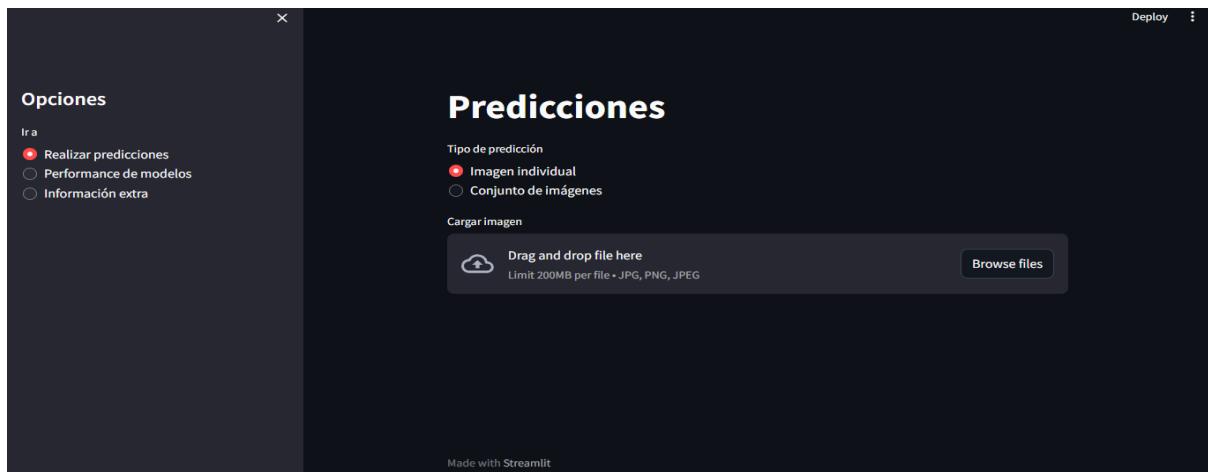
El Modelo 2, por otro lado, sigue una estructura similar a la del Modelo 1 pero con la adición de Batch Normalization para normalizar las activations intermedias de las capas convolucionales. Esto puede mejorar el rendimiento y la convergencia del modelo durante el entrenamiento. Aunque es una mejora con respecto al Modelo 1, aún no alcanza los mejores resultados.

Finalmente, el Modelo 3 destaca por su capacidad para manejar imágenes de mayor resolución (128x128 píxeles). Con tres capas de convolución para la extracción de características, seguidas de Max Pooling para reducir la dimensionalidad, y dos capas completamente conectadas para la clasificación, este modelo logra proporcionar los mejores resultados en términos de precisión. Su capacidad para procesar imágenes más detalladas y su enfoque en mantener un alto nivel de precisión lo convierten en la opción preferida cuando se buscan los mejores resultados en la clasificación de imágenes de mosquitos.

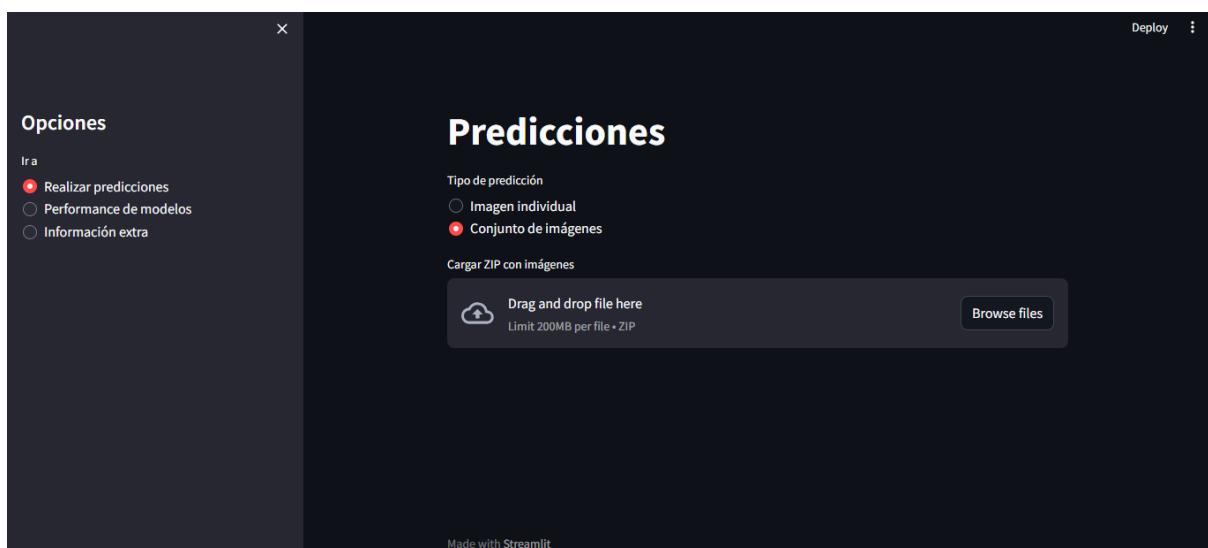
En resumen, mientras que los tres modelos son valiosos en su propio contexto, el Modelo 3 se destaca como el más efectivo en términos de precisión, especialmente cuando se trata de imágenes de mayor resolución. Su estructura y enfoque detallados lo convierten en la elección ideal cuando se buscan los mejores resultados en la tarea de clasificación de imágenes.

- **Aplicación**

La aplicación se desarrolló con Streamlit y consta con 3 apartados: Predicciones, Información de los modelos, Información Extra. En el primer apartado se realizan predicciones, valga la redundancia, tanto de imágenes individuales como de grupos de imágenes, en este apartado se muestran las imágenes cargadas, posterior a esto se muestra el tipo de mosquito y una tabla que cuenta con los valores resultantes (coincidencias con los diferentes tipos) en base a la predicción y finalmente en base al tipo de mosquito se despliega un poco de información referente al mismo para incentivar al usuario a conocer más del mismo y tomar precauciones en caso estos contagien enfermedades de alto impacto. Por otro lado el apartado de Información de los modelos muestra la información respecto a cada implementación desarrollada, esta cuenta con métricas de desempeño, gráficas y otros datos que nos permitieron determinar que modelo utilizar para la aplicación y que este presentara los mejores resultados. La última parte nos brinda un poco de detalles generales tanto de las herramientas y datos utilizados como del enfoque que le estamos dando al proyecto con la finalidad de mantener informado al usuario de los motivos por los que se tomaron las decisiones sobre tema, algoritmos, modelos, datos, herramientas y demás.



Img 2. Pantalla de predicciones para imágenes individuales



Img 3. Pantalla de predicciones para grupos de imágenes



Img 4. Pantalla de muestra de performance para diferentes modelos

The screenshot shows a mobile application's user interface. On the left, there is a sidebar with the title "Opciones" (Options) at the top. Below it are three circular buttons: "Realizar predicciones" (Perform predictions), "Performance de modelos" (Model performance), and "Información extra" (Extra information). The third button is highlighted with a red dot. The main content area is titled "Información extra" in bold. Below it is a section titled "Sabías que..." (Did you know...). The text in this section discusses the role of mosquitoes in transmitting diseases like chikungunya, dengue, and malaria. It highlights their status as vectors for these viruses and parasites, noting that they bite humans to obtain blood and then transmit the pathogen to the next victim through their saliva. The text also mentions that mosquitoes are responsible for more deaths than any other animal in the world due to the diseases they spread, with approximately 700 million people contracting mosquito-borne illnesses annually. At the bottom of the text block, there is a small note in very small font: "Estos datos subrayan la importancia de tomar medidas preventivas, como el uso de repelentes y la".

*Img 5. Pantalla de información extra sobre el proyecto*

## Referencias:

- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88.
- Rajaraman, S., Antani, S., & Poostchi, M. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6, e4568.