INFB6074 - INFRAESTRUCTURA PARA CS. DE DATOS
INGENIERÍA CIVIL EN CIENCIA DE DATOS

Etapa Proyecto final° 1

"Diseño de Anteproyecto y Presentación Inicial."

Dr. Ing. Michael Miranda Sandoval

Mayo de 2025

UNIVERSIDAD TECNOLÓGICA METROPOLITANA
UTEM del Estado de Chile

ETAPA PROYECTO FINAL° 1

INTEGRANTES:

Nombre: Felipe Martínez González **Email:** fmartinezgo@utem.cl
Nombre: Christian Pérez Flores **Email:** cperezfl@utem.cl
Nombre: Benjamín Zamorano Soto **Email:** bzamoranos@utem.cl

1. Introducción

La agricultura es uno de los pilares fundamentales para el desarrollo económico y social de numerosos países, especialmente en regiones donde constituye una fuente esencial de alimentación, empleo y exportación. Sin embargo, este sector enfrenta constantemente amenazas que afectan su productividad, siendo las plagas agrícolas una de las más significativas. Estas pueden provocar pérdidas considerables en la producción, deterioro de la calidad de los cultivos y un aumento en el uso de pesticidas, lo que repercute negativamente tanto en la economía como en el medio ambiente.

En este contexto, la ciencia de datos y la inteligencia artificial se han convertido en herramientas clave para desarrollar soluciones innovadoras y eficientes. En particular, el uso de visión por computador permite automatizar procesos de monitoreo agrícola, mejorando la capacidad de respuesta ante situaciones críticas como infestaciones.

Este trabajo propone el diseño de un sistema para la detección indirecta de plagas a partir de imágenes de cultivos, identificando signos visuales de daño como manchas, decoloraciones, agujeros y otros indicadores en las hojas de las plantas. A diferencia de los enfoques tradicionales que requieren la observación directa de la plaga, este sistema se basa en la detección de sus efectos, lo que permite una respuesta temprana y con menores recursos.

El objetivo principal es construir una arquitectura técnica que permita procesar grandes volúmenes de imágenes, entrenar modelos de aprendizaje profundo para la clasificación de síntomas, y presentar resultados mediante una interfaz accesible. Se analizan diversas arquitecturas paralelas, distribuidas y de Big Data, evaluando su aplicabilidad al problema planteado y proponiendo una solución híbrida que integre componentes escalables y eficientes.

Este informe detalla la justificación del proyecto, el análisis preliminar de infraestructura, el diseño de la arquitectura propuesta, las tecnologías seleccionadas y un plan de trabajo para su desarrollo e implementación.

2. Selección y Justificación del Objetivo

El presente proyecto tiene como objetivo diseñar e implementar un sistema de detección temprana de plagas en cultivos agrícolas, basado en el análisis de imágenes de hojas y plantas que presenten signos visuales de infestación. A través de técnicas de visión por computador y modelos de aprendizaje profundo, se busca identificar daños provocados por plagas, tales como decoloraciones, manchas, perforaciones o necrosis, sin necesidad de observar directamente al agente biológico.

La elección de este objetivo responde a una problemática real y de alto impacto. Las plagas agrícolas representan una amenaza constante para la productividad del sector, generando pérdidas económicas considerables y elevando el uso de agroquímicos. Detectar la presencia de plagas de manera oportuna puede reducir los daños, mejorar la gestión de recursos y disminuir el impacto ambiental.

Desde la perspectiva de la ciencia de datos, este desafío es especialmente pertinente, ya que involucra tareas de clasificación de imágenes, entrenamiento de modelos supervisados, manejo de conjuntos de datos de gran volumen y despliegue de soluciones accesibles para usuarios no especializados. Además,

permite explorar infraestructuras escalables y modernas, combinando procesamiento paralelo, almacenamiento distribuido y visualización interactiva.

3. Análisis Preliminar: Arquitecturas y Tecnologías

Para abordar el problema de detección indirecta de plagas en imágenes de cultivos, se evaluaron distintas arquitecturas aplicables en el contexto de ciencia de datos: paralela, distribuida y Big Data. A continuación, se presentan sus características más relevantes:

- **Arquitectura Paralela**: permite el procesamiento simultáneo de operaciones, siendo especialmente útil para el entrenamiento de modelos de redes neuronales convolucionales (CNN) mediante GPUs.
- **Arquitectura Distribuida**: utiliza múltiples nodos para procesar datos de manera concurrente. Herramientas como Apache Spark son ideales para procesar grandes volúmenes de imágenes de forma eficiente.
- **Big Data**: orientada al almacenamiento y gestión de datos a gran escala. Emplea tecnologías como Hadoop HDFS, bases de datos NoSQL y servicios en la nube. Aunque su implementación puede ser compleja, es adecuada para sistemas que crecen en volumen.

Comparación de arquitecturas

Arquitectura	Ventajas	Desventajas	Aplicación al proyecto
Paralela	Alta velocidad con GPU	Requiere hardware específico	Entrenamiento de modelos CNN
Distribuida	Escalable horizontalmente	Configuración más compleja	Procesamiento masivo de imágenes
Big Data	Almacenamiento eficiente	Exceso para volúmenes pequeños	Escalabilidad futura

Table 1: Comparación entre arquitecturas aplicadas al problema

A partir de este análisis, se propone una arquitectura **híbrida** que combine el uso de procesamiento paralelo (para el entrenamiento) con herramientas distribuidas y de almacenamiento escalable para futuras etapas del sistema.

4. Propuesta de Solución y Diseño Preliminar

La solución planteada consiste en el desarrollo de un sistema capaz de detectar signos visuales de infestación en imágenes de hojas o cultivos, utilizando técnicas de aprendizaje profundo. El sistema incluye los siguientes componentes:

- **Interfaz Web**: desarrollada con Streamlit, permitirá al usuario cargar imágenes y visualizar los resultados.
- **Preprocesamiento**: ajuste de imágenes mediante scripts en Python o Apache Spark, en función del volumen de datos.
- **Modelo de detección**: una red neuronal convolucional (CNN), entrenada con datos como PlantVillage, detectará signos de plagas.
- **Almacenamiento**: MongoDB se utilizará para guardar resultados, mientras que las imágenes se almacenarán en disco o S3.
- **Contenerización**: se utilizará Docker para facilitar el despliegue del sistema completo.

5. Plan de Trabajo

El proyecto se estructura en etapas semanales para su desarrollo incremental, desde la recolección de datos hasta el despliegue funcional del sistema:

Semana	Actividad
1	Revisión bibliográfica y exploración de datos
2	Preprocesamiento de imágenes y etiquetado
	Entrenamiento inicial del modelo
3	Evaluación del modelo y ajustes
	Desarrollo de la interfaz web
4	Integración de componentes y pruebas
	Documentación y presentación final

Table 2: Cronograma del proyecto por semana

6. Conclusión

El proyecto propuesto busca aplicar herramientas de la ciencia de datos para resolver una problemática concreta y de alto impacto en el sector agrícola: la detección temprana de plagas. Al enfocarse en los signos visibles de infestación, el sistema permite una detección indirecta, más simple y escalable que otras soluciones tradicionales. La propuesta combina enfoques de aprendizaje profundo con arquitecturas modernas de infraestructura, ofreciendo una solución tecnológica flexible, eficiente y con potencial de crecimiento. Este anteproyecto establece las bases conceptuales y técnicas para su desarrollo, permitiendo avanzar hacia una implementación funcional en las etapas posteriores del curso.

7. Referencias

1. PlantVillage Dataset. (2018). Kaggle. https://www.kaggle.com/datasets/emmarex/plantdisease