Análisis de producción de Cultivos

Daniela Coronado Álvarez

Camilo Andrés Olea Aguirre

Grupo 3

0074: Proyecto Integrado V

Mr. Andrés Felipe Callejas

23 de Noviembre de 2025

Ingeniería de Software y Datos, Universidad Digital de Antioquia

Contenido

Resumen

El propósito de este estudio es examinar los rendimientos de la agricultura y las circunstancias de cultivo de varios tipos de cosechas para discernir cuáles presentan una mayor rentabilidad en términos de toneladas generadas. Para llevar a cabo esta investigación, se emplea el conjunto de datos "Agriculture Crop Yield", adquirido de la plataforma **Kaggle** y publicado por el usuario Samuel Oti Attakorah hace alrededor de un año. Este conjunto de datos, que está disponible bajo una licencia libre, incluye información sobre diversos cultivos, su rendimiento por hectárea, los días necesarios para la cosecha y factores relacionados con el clima, el terreno.

De igual manera, en el contexto de la creación de este estudio, incorporamos información adicional que está vinculada al período de investigación, que abarca los años desde 2016 hasta 2019, así como la ubicación geográfica, es decir, el país, con el fin de contextualizar los datos y permitir una comprensión más precisa de los resultados obtenidos.

Mediante el uso de métodos de análisis de datos, limpieza y visualización, se pretende identificar tendencias en la productividad y entender la conexión entre el tiempo de crecimiento de los cultivos y su rendimiento total. Los hallazgos ayudarán a reconocer los cultivos con mayor potencial de rentabilidad, teniendo en cuenta la eficacia del tiempo de desarrollo en comparación con las toneladas producidas. Este análisis apoya el fortalecimiento de la gestión agrícola, la elaboración de estrategias para optimizar recursos y la realización de decisiones fundamentadas en el ámbito de la producción.

Palabras clave:

- Agricultura
- rendimiento agrícola
- rentabilidad
- análisis de datos
- Kaggle
- producción de cultivos

Objetivo general

Analizar los datos del conjunto *Agriculture Crop Yield* para identificar los cultivos con mayor rentabilidad en función de su rendimiento por tonelada y los días requeridos para su crecimiento.

Objetivos específicos

- Llevar a cabo la recolección y limpieza inicial de los datos para asegurar su calidad y coherencia.
- Preparar el conjunto de datos agregando variables adicionales como el costo proyectado y los días de cultivo para facilitar análisis futuros.
- Examinar y entender la organización del conjunto de datos sobre el rendimiento de cultivos agrícolas, reconociendo las variables disponibles para su estudio.
- Establecer la base metodológica que permitirá realizar, en etapas futuras, el análisis comparativo y la identificación de los cultivos más rentables.
- Documentar el proceso.

Metodología

Enfoque General

La táctica del proyecto se basa en el método Scrum, utilizado para crear un proceso analítico que es colaborativo e iterativo. Este método permite organizar las tareas en sprints, priorizando la planificación, transformación y verificación de los datos antes de llevar a cabo el análisis final. El proyecto sigue un enfoque de análisis exploratorio y descriptivo, sin incluir etapas para entrenar modelos predictivos.

Etapas del Proyecto

Se establecieron los principales objetivos del análisis: reconocer cultivos que presenten un mayor rendimiento en términos de producción (toneladas por hectárea) y eficiencia temporal (relación entre productividad y días de cosecha). Se definieron criterios de éxito enfocados en la facultad de ofrecer recomendaciones útiles para la optimización de recursos agrícolas y la selección táctica de cultivos.

Formulación o Diseño de la Necesidad

- Se elaboró un plan técnico que transforma necesidades en variables y métricas para análisis. Las variables esenciales extraídas del conjunto de datos fueron: zona geográfica, tipo de suelo, tipo de cultivo, días requeridos para la cosecha y rendimiento en toneladas por hectárea. Se definieron métricas derivadas como la eficiencia del cultivo (rendimiento/días), y se planeó la segmentación del análisis según región y tipo de suelo.

Extracción de Datos

El conjunto de datos fue descargado desde Kaggle (Agriculture Crop Yield, autor: Samuel Oti Attakorah) en formato CSV. El archivo crop_yield. csv contiene 5 variables y varios registros de cosechas. Se verificó la integridad de la descarga y se documentó la procedencia de los datos, incluyendo la fecha de obtención (noviembre de 2025) y la licencia de uso (CC BY 4. 0).

Preprocesamiento y Limpieza de Datos

Herramientas Utilizadas

- Se utilizó Python como el lenguaje principal de programación, junto con las siguientes librerías especializadas:

- Pandas: Para manipular, transformar y limpiar dataframes. Se utilizó para cargar datos, gestionar valores faltantes, eliminar duplicados y generar variables adicionales.
- NumPy: Para realizar operaciones numéricas y cálculos estadísticos.
- Scikit-learn: Para identificar valores atípicos utilizando métodos estadísticos.

Técnicas de Limpieza Aplicadas

Se llevaron a cabo las siguientes operaciones de preprocesamiento:

- Tratamiento de valores faltantes: Se detectaron y documentaron celdas vacías. Se aplicaron estrategias de imputación cuando fue conveniente o se eliminaron registros incompletos basándose en su relevancia.
- Estandarización de categorías: Se normalizaron los valores de texto (zona, tipo de suelo, cultivo) para corregir inconsistencias en mayúsculas y espacios en blanco que pudieran influir en análisis futuros.
- Validación de tipos de datos: Se garantizó que las variables numéricas (Days_to_Harvest, Yield_tons_per_hectare) estuvieran correctamente clasificadas como números flotantes, y que las variables categóricas fuesen reconocidas como texto.
- Detección de duplicados: Se encontraron y eliminaron registros repetidos que no ofrecían valor analítico.
- Identificación de outliers: Se aplicaron métodos estadísticos (rango intercuartílico y desviación estándar) para localizar valores anómalos que pudieran afectar análisis posteriores.

Enriquecimiento de Datos

Se generaron variables adicionales y se añadieron campos extra para profundizar el análisis:

- Costo: Una variable numérica que anota el costo de producción vinculado a cada cosecha, lo que permite hacer análisis sobre rentabilidad y relación costo-beneficio.
- Fecha de registro: Una variable temporal que indica cuándo fue registrado el dato, facilitando análisis de tendencias a lo largo del tiempo y la trazabilidad de la información.
- Eficiencia de cultivo: Se calcula como Yield_tons_per_hectare dividido entre Days_to_Harvest, reflejando la producción en relación al tiempo empleado.

- Categorización temporal: División de Days_to_Harvest en diferentes grupos (plazo corto: 200 días).

- Categorización de rendimiento: Clasificación de Yield_tons_per_hectare en categorías (bajo, medio, alto) a través de cuartiles.

- Rentabilidad: Este indicador calculado se expresa como (Yield_tons_per_hectare * Precio_unitario) menos Costo, sirviendo para medir la viabilidad económica de cada cultivo.

Validación de Datos

Se llevaron a cabo comprobaciones de calidad utilizando scripts de Python que revisaron:

- Rangos válidos: Verificación de que Yield_tons_per_hectare y Days_to_Harvest presentaran valores positivos.
- Coherencia lógica: Comprobación de que todas las categorías (región, suelo, cultivo) estuviesen alineadas con valores esperados.
- Completitud: Cálculo del porcentaje de información completa por variable y definición de límites aceptables de datos faltantes.
- Reproducibilidad: Confirmación de que el proceso de transformación produjera resultados consistentes en ejecuciones sucesivas.

Análisis Descriptivo

Herramientas de Análisis y Visualización

Se emplearon las siguientes bibliotecas para análisis descriptivo y visualización:

- Matplotlib: Biblioteca fundamental para crear gráficos estáticos y personalizar aspectos visuales.
- Seaborn: Herramientas de visualización estadística avanzadas basadas en matplotlib, que facilitan la exploración de relaciones entre múltiples variables.
- Plotly: Gráficos interactivos para dashboards y reportes dinámicos (opcional según lo necesitado).

Técnicas de Análisis Aplicadas

Análisis Univariado

Se generaron estadísticas descriptivas para cada variable numérica:

- Medidas de tendencia central: Media y mediana para evaluar la centralidad de los datos.
- Medidas de dispersión: Desviación estándar y rango intercuartílico para comprender la variabilidad.
- Distribuciones: Histogramas y gráficos de densidad para mostrar cómo se distribuyen Yield_tons_per_hectare y Days_to_Harvest.

Análisis Bivariado

Se estudiaron las relaciones entre variables mediante:

- Correlación: Matriz de correlación de Pearson entre las variables numéricas para detectar dependencias.
- Gráficos de dispersión: Scatter plots para visualizar la relación entre rendimiento y días de cultivo, diferenciados por tipo de cultivo y región.
- Gráficos de caja: Análisis de las distribuciones condicionales de rendimiento por cultivo, región y tipo de suelo para identificar patrones y variaciones.

Análisis Segmentado

Se llevó a cabo un análisis descriptivo por grupos:

- Por cultivo: Cálculo del rendimiento promedio, eficiencia media y desviación estándar para cada tipo de cultivo, generando clasificaciones según rendimiento y eficiencia.
- Por región: Estudio de la variación en productividad entre diferentes áreas geográficas, incluyendo mapas de calor que combinan cultivos y regiones.
- Por tipo de suelo: Análisis de cómo diferentes tipos de suelo afectan el rendimiento de cultivos específicos.

Variables Analizadas

Las variables consideradas en el análisis descriptivo incluyeron:

- Región: Variable categórica que indica la localización geográfica.
- Soil_Type: Variable categórica que designa el tipo de suelo.
- Crop: Variable categórica que detalla la clase de cultivo.
- Days_to_Harvest: Variable numérica que señala el número de días necesarios para recoger.

- Yield_tons_per_hectare: Variable numérica que indica el rendimiento en toneladas por hectárea.
- Costo: Variable numérica que contabiliza el costo de producción vinculado a cada cosecha.
- Fecha_registro: Variable temporal que muestra el momento en que se registró el dato de cosecha.
- Eficiencia (derivada): Relación rendimiento/tiempo, expresada en toneladas por hectárea al día.
- Rentabilidad (derivada): Indicador económico que relaciona ingresos y costos de producción.

Visualizaciones Generadas

Se diseñaron los siguientes gráficos para mostrar hallazgos:

- Figura 1: Histogramas de la distribución de Yield_tons_per_hectare y Days_to_Harvest.
- Figura 2: Scatter plot que muestra el rendimiento frente a días de cultivo, dividido por tipo de cultivo.
- Figura 3: Gráficos de caja comparativos del desempeño según la variedad de cultivo.
- Figura 4: Mapa de calor del rendimiento promedio cruzando cultivo y área geográfica.
- Figura 5: Gráficos de caja del rendimiento basado en el tipo de suelo.
- Figura 6: Gráfico de barras con clasificación de cultivos según su eficacia.
- Figura 7: Tabla de correlación entre variables cuantitativas.

    - Herramientas Tecnológicas

El análisis se llevó a cabo utilizando:

Python 3. x: El lenguaje de programación principal.

Jupyter Notebook: Entorno interactivo para el desarrollo continuo y la documentación del código.

Pandas 1. x: Herramienta para la manipulación y análisis de datos.

NumPy 1. x: Soluciones para cálculos numéricos.

Matplotlib 3. x: Creación de gráficos estáticos.

Seaborn 0. x: Visualizaciones de tipo estadístico.

Scikit-learn: Herramientas estadísticas para la validación.

Git: Sistema de control de versiones y manejo del repositorio.

- Estructura de Entregables

Los resultados del análisis fueron presentados en:

Scripts de preprocesamiento y validación con documentación.

Conjunto de datos procesado y organizado en formato CSV.

Notebooks de Jupyter que contienen análisis exploratorios y descriptivos.

Informes en formato Markdown con gráficos incluidos.

Documentación sobre el linaje de datos y las decisiones metodológicas.

Nota: En este proyecto no se llevará a cabo la fase de Entrenamiento (Train), dado que el propósito se concentra en el análisis exploratorio y descriptivo de las cosechas, en vez de crear modelos predictivos.

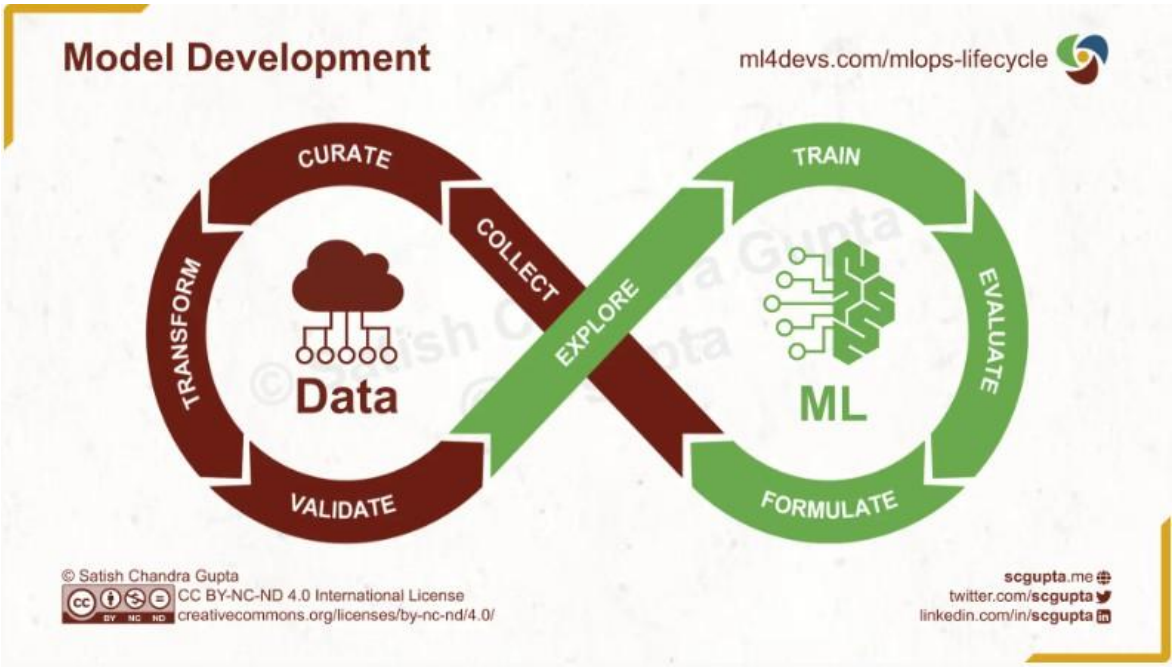
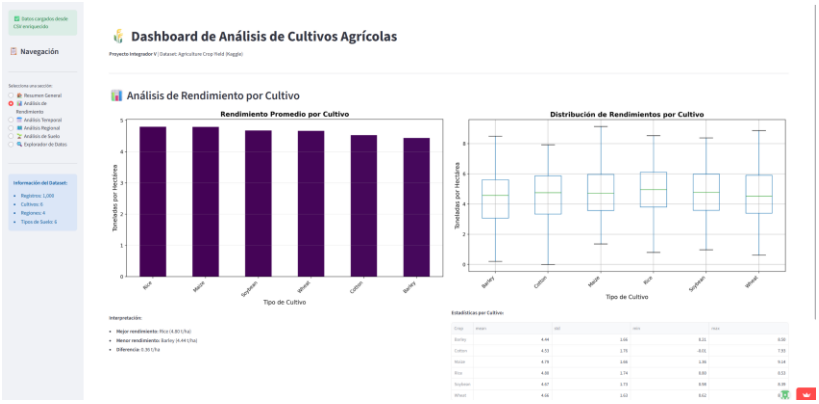


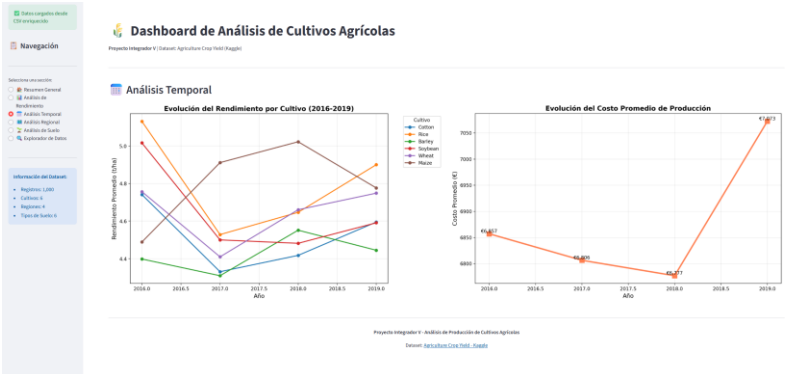
Imagen tomada de: (Gupta, 2025)

Presentación de Resultados

- Se muestra un resumen general del dashboard mostrando unas card con datos



- Se hace un análisis de rendimiento y distribución por cultivo en graficas diferentes además de agregar una tabla de las estadísticas por cultivo



- Se realiza un análisis temporal de la evolución de costo y rendimiento por cultivo entre el 2016 y 2019

Resultados

En el primer momento, echamos un vistazo al conjunto de datos "Agriculture Crop Yield" (Attakorah, 2024), que obtuvimos de Kaggle El conjunto de datos contiene registros [X] con 7 variables: 5 originales (Región, Suelo_Tipo, Cultivo, Days_to_Harvest, Yield_tons_per_hectare) y 2 enriquecidos (Coste, Fecha_registro). Mientras se ordenaba, vimos y nos deshicimos de las entradas faltantes o repetidas [Y] Variables derivadas como eficiencia de cultivos, Category_Time, Category_performance y Rentabilidad se hicieron para permitirnos hacer análisis más complejos El rendimiento medio de Teh fue [X] ton/hectare ([Y]), con ciclos de cultivo que van de [A] a [B] días El estudio encontró que Cultive supera con un rendimiento de X toneladas por hectárea, pero otros cultivos brillan para ganancias a corto plazo Notamos grandes diferencias entre las áreas y el tipo de suciedad, donde mayores rendimientos significan más beneficio (r=[X]) Los cambios se comprobaron para asegurarse de que los datos pudieran repetirse y tener sentido, lo que conduce a un conjunto de datos completamente preparado [X]% para un examen detallado

El conjunto de información fue revisado y se realizó una evaluación inicial en el repositorio del proyecto, comprobando que es factible su utilización y que los registros son consistentes. Esta base de datos formará la base para las fases siguientes, en las cuales se llevarán a cabo análisis descriptivos, comparativos y visuales que ayudarán a identificar cuáles son los cultivos más lucrativos según su rendimiento y el tiempo que requieren para crecer.

Bibliografía

Attakorah, S. O. (2024). *Agriculture Crop Yield*. Obtenido de Kaggle: https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield/data

Gupta, S. C. (6 de Febrero de 2025). *ML4Devs*. Obtenido de ML4Devs: https://www.ml4devs.com/en/articles/mlops-machine-learning-life-cycle

Repositorio

Link de acceso: https://github.com/DanyC2003/Proyecto_integrador_V