

Iniciamos

6:10 am

:::Gracias::

Mentor: Jonnatan Arias

Garcia



Machine Learning Operations

MLops es un enfoque para construir, desplegar y mantener sistemas de aprendizaje automático en producción de forma eficaz y escalable.



by Jonnatan Arias Garcia

Contenido I

Modulo 1: Introducción MLOps

MLOps

- MLOps y Devs
- Importancia MLOps
- Ciclo de vida MLOps

Entorno de desarrollo

- Google Colab
- Configuración de Python para MLOps
- Control de Versiones Git y Github

Modulo 2: Fundamentos de Machine Learning

ML básico

- Aprendizaje Supervisado – No Supervisado
- Evaluación de Modelos: Accuracy, precisión recall, F1 score

Gestión de Datasets

- Preprocesamiento
- Librerías Pandas y Numpy

Módulo 3: Automatización de procesos

Pipeline de ML

- Definición y usos
- Pipelines en Scikit-learn
- Implementación en Colab

Control de Versiones de datos

- DVC colab
- Versionado de Datasets.

Módulo 4. Modelos en Producción

Modelos Reproducibles

- Reproducibilidad y Semilla
- Tracking con MLflow
- Mlflow en colab

Modelos y entrenamiento

- Modelos con tensorflow y keras
- Guardado y Exportación de modelo entrenado
- Deploy con FastAPI, otras herramientas API

Contenido II

Módulo 5. Pruebas y Monitoreo de Modelos

Pruebas

- Validación Cruzada
- Pruebas unitarias y de integración de pipelines en ML
- Pytest

Monitoreo

- Introducción al monitoreo de modelos
- Herramientas (Prometheus, Grafana)
- Implementación de Alertas Básicas

Módulo 6. Gestión de Recursos

Optimización de Recursos

- CPU, GPU, TPU
- Uso eficiente y paralelaje

Optimización de Parámetros

- GridSearch y RandomSearch

Módulo 7. Despliegue en la Nube

Google Cloud AI

- Integración de Colab con Google Cloud
- Despliegue en Google Cloud AI

Integración CI/CD

- CI/CD: Integración continua + entrega e implementación continua
- Github Actions

Módulo 8. Documentación y Buenas Practicas

Sphinx

- Documentación de pipelines, modelos y librerías

Buenas Prácticas

- Recomendaciones, Seguridad y privacidad



Modulo I

MLOps Basics y Entorno (Colab)



¿Qué es MLops?

Combina prácticas de ingeniería de software con DevOps para automatizar y optimizar el ciclo de vida del desarrollo de modelos de aprendizaje automático.

Se encarga del forma general del Desarrollo, implementación, despliegue y supervisión de proyectos de Machine learning

Integración Continua

Automatización de la construcción, prueba y despliegue de modelos.

Entrega Continua

Despliegue continuo de modelos actualizados a entornos de producción.

Monitoreo Continuo

Seguimiento del rendimiento del modelo y modelo y detección de problemas.



Importancia de MLops en el desarrollo de modelos de aprendizaje automático

MLops ayuda a garantizar la calidad, la fiabilidad y la escalabilidad de los modelos de aprendizaje automático, lo que permite un desarrollo y una implementación más rápidos y eficientes.

1

Mejorar la eficiencia

Automatización de tareas repetitivas, como el entrenamiento de modelos y la implementación.

2

Mayor colaboración

Fomenta la colaboración entre científicos de datos, ingenieros y equipos de operaciones.

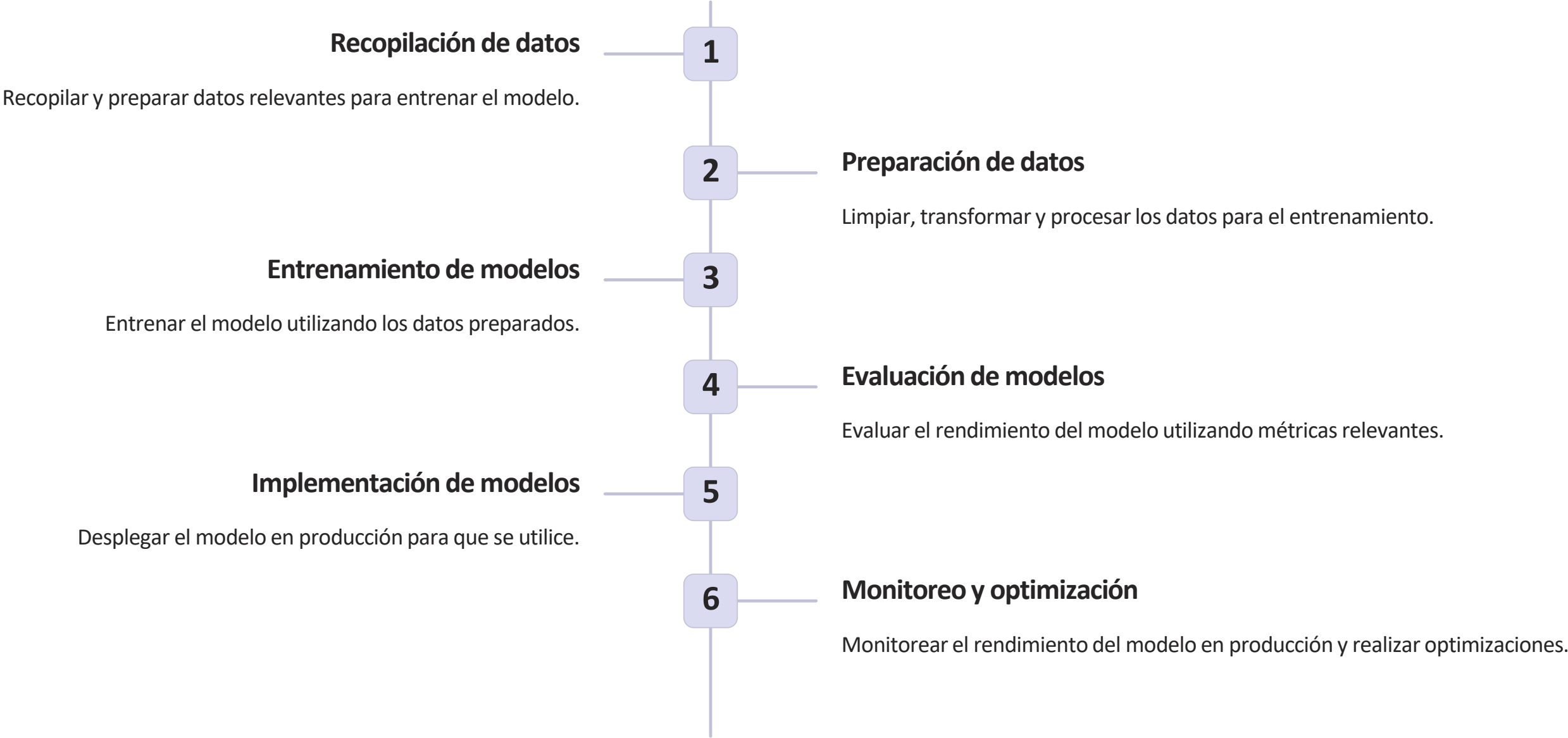
3

Reducir riesgos

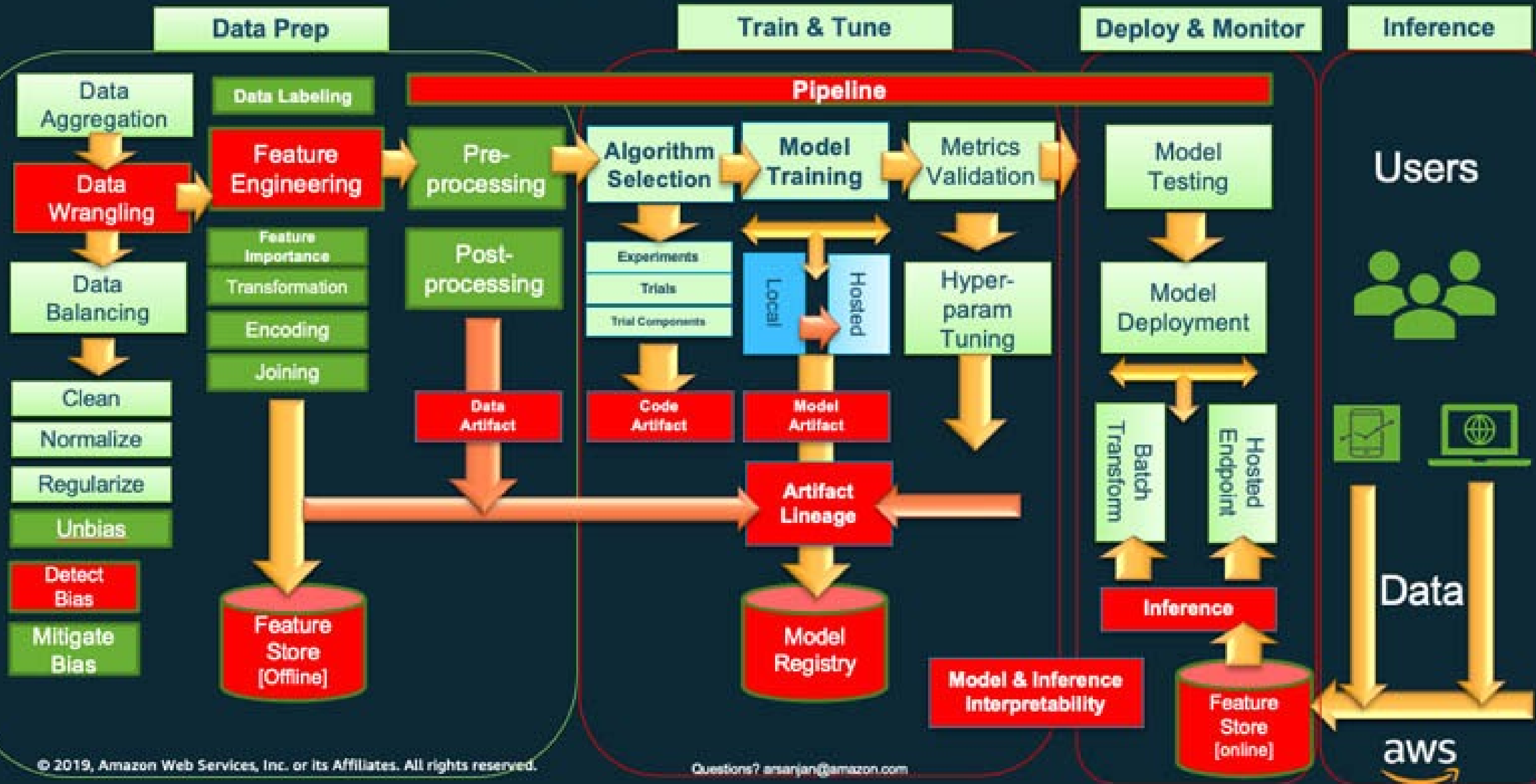
Previene errores y problemas potenciales durante el desarrollo y la implementación.

Ciclo de vida del modelo de aprendizaje automático

El ciclo de vida del modelo de aprendizaje automático abarca desde la recopilación de datos hasta la implementación y el seguimiento del modelo en producción.



The ML-Lifecycle: Detailed View



Google
colab



 + colab
python



1. It is a software

2. It is installed locally on the system

3. It is a command line tool

4. It is a tool to manage different versions of edits, made to files in a git repository

5. It provides functionalities like Version Control System Source Code Management

1. It is a service

2. It is hosted on Web

3. It provides a graphical interface

4. It is a space to upload a copy of the **Git** repository

5. It provides functionalities of Git like VCS, Source Code Management as well as adding few of its own features



Modulo II

Fundamentos de ML (Supervisado – No
Supervisado)

Gestión Datasets (Data-lake & Data-
WareHouse)

Aprendizaje Supervisado vs No Supervisado

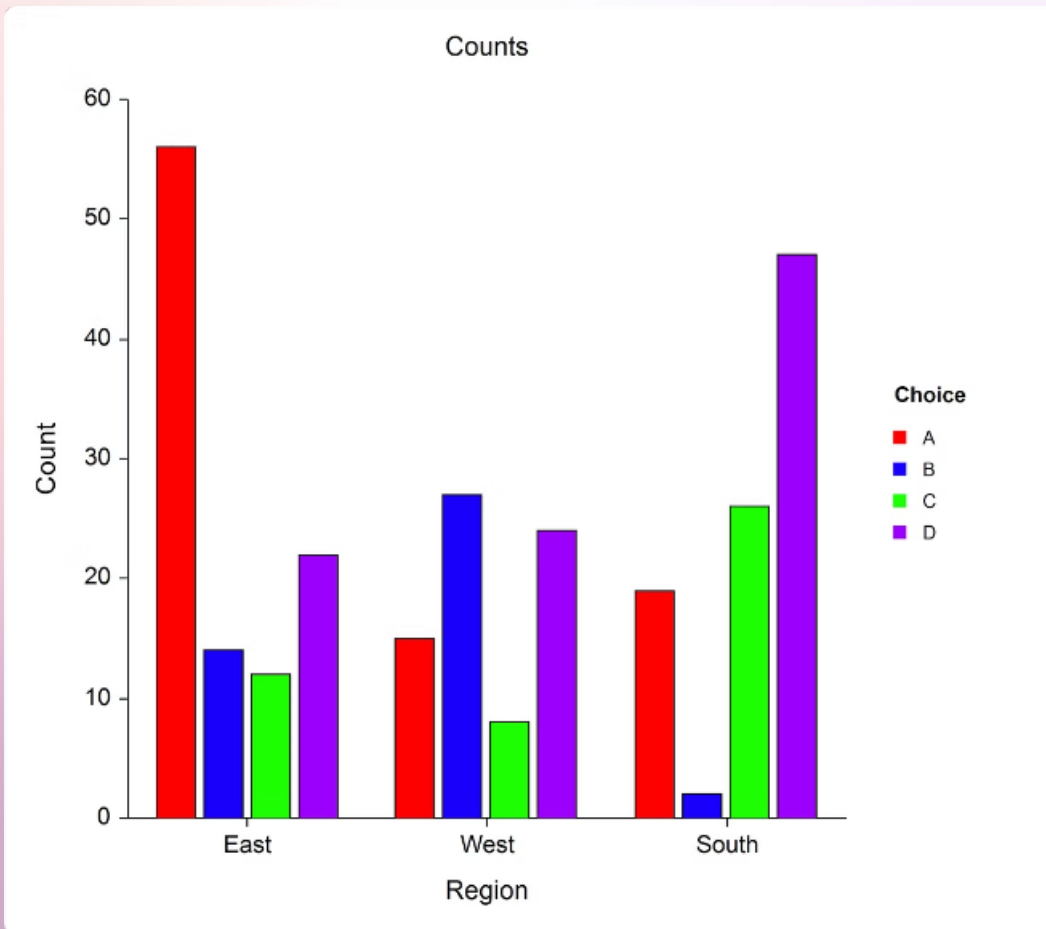
Aprendizaje Supervisado

En el aprendizaje supervisado, los modelos se entrenan con datos etiquetados, lo que significa que cada ejemplo de entrenamiento tiene una salida conocida. Se utiliza para tareas como clasificación y regresión.

Aprendizaje No Supervisado

El aprendizaje no supervisado trata con datos no etiquetados. Los modelos deben descubrir patrones y estructuras ocultas en los datos. Este tipo de aprendizaje se utiliza para tareas como la agrupación en clústeres y la reducción de dimensionalidad.

Métricas de Evaluación de Modelos



1

Accuracy

La exactitud es la proporción de predicciones correctas sobre el total de predicciones. Es una métrica útil, pero pero no siempre es la mejor, especialmente cuando se trata de conjuntos de datos desequilibrados.

2

Precisión

La precisión es la proporción de predicciones positivas correctas sobre el sobre el total de predicciones positivas. Es positivas. Es una medida de la precisión de precisión de un modelo.

3

Recall

La sensibilidad, también llamada recuerdo, es la proporción de predicciones predicciones positivas correctas sobre el sobre el total de casos positivos reales. reales. Mide la capacidad del modelo para para identificar casos positivos.

4

F1 Score

La puntuación F1 es la media armónica de armónica de precisión y sensibilidad. Es Es una medida útil para evaluar modelos modelos cuando se requiere un equilibrio equilibrio entre precisión y sensibilidad. sensibilidad.

Gestión de Datasets



Recopilación de Datos

La recopilación de datos es el primer primer paso en la gestión de datasets. Es crucial recopilar datos datos relevantes y de alta calidad para para el modelo de aprendizaje automático.

Limpieza de Datos

La limpieza de datos implica la eliminación de datos faltantes, duplicados o inconsistentes. Es un paso crucial para garantizar la precisión del modelo.

Almacenamiento de Datos

El almacenamiento de datos es esencial para la gestión eficiente de de datasets. Es importante elegir un un sistema de almacenamiento que que sea escalable y seguro.

Acceso a Datos

El acceso a datos debe ser rápido y y eficiente. Los sistemas de gestión de gestión de bases de datos permiten un permiten un acceso eficiente a los los datos para el entrenamiento del del modelo.

Preprocesamiento Estándar de Datos

1

Escalado de Datos

El escalado de datos es esencial para mejorar el rendimiento del modelo. Los algoritmos de aprendizaje automático funcionan mejor cuando los datos están en una escala similar.

2

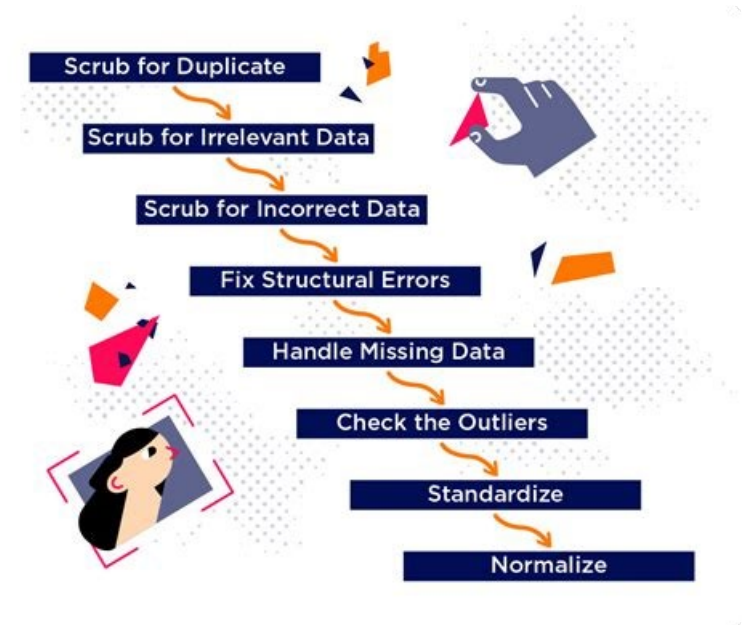
Codificación de Datos Categóricos

Las variables categóricas deben transformarse en formato numérico antes de antes de poder usarse en los modelos de aprendizaje automático.

3

Gestión de Datos Faltantes

Los datos faltantes pueden afectar el rendimiento del modelo. Es importante importante manejar los datos faltantes de forma adecuada, ya sea eliminándolos o eliminándolos o imputándolos.



Conceptos de DataLake

1

Almacenamiento Centralizado

Un data lake es un repositorio centralizado para almacenar todos los datos, independientemente de su formato o estructura.

2

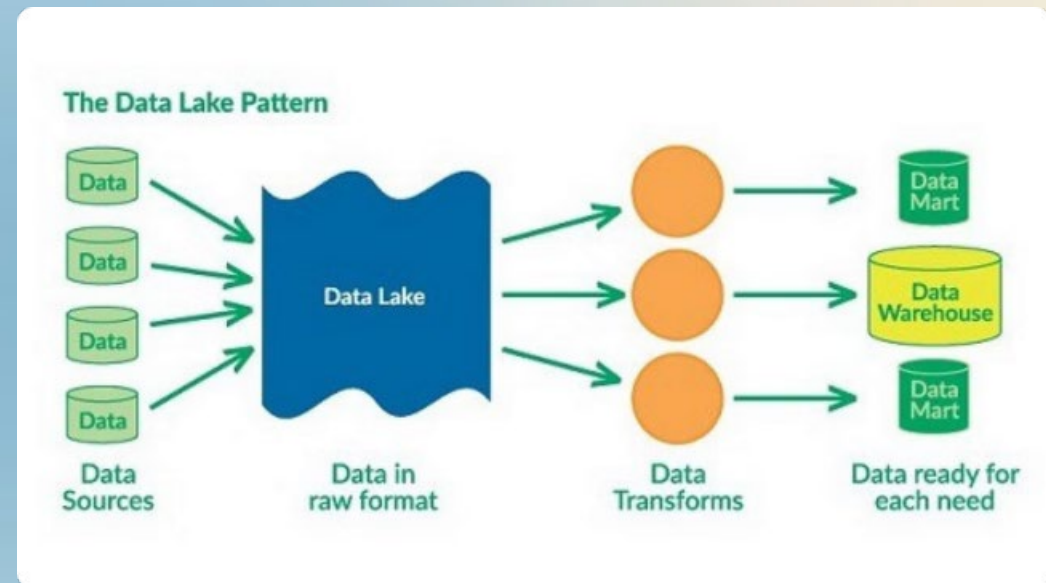
Datos en Bruto

Los datos se almacenan en su formato original, lo que permite un análisis flexible y la posibilidad de aplicar diferentes esquemas a los datos.

3

Escalabilidad

Los data lakes son altamente escalables, lo que permite el almacenamiento de grandes volúmenes de datos.



Conceptos de DataWarehouse

DataWarehouse

Data Lake

Almacenamiento de datos sin procesar

Esquema flexible

Orientado a la ingestión

Data Warehouse

Almacenamiento de datos estructurados

Esquema predefinido

Orientado a la consulta



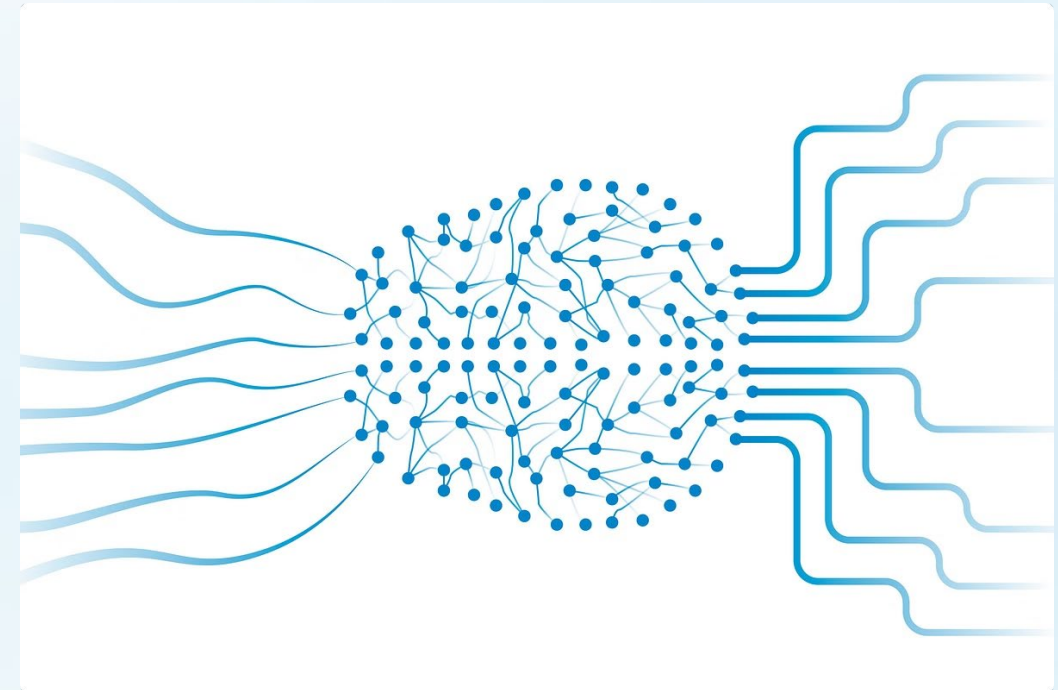


Modulo III

Pipelines & Control de Versiones

Pipelines de ML

En el ámbito de la ciencia de datos, un pipeline de aprendizaje automático (ML) se refiere a un proceso automatizado que organiza y gestiona las diferentes etapas de un proyecto de ML, desde la preparación de datos hasta la evaluación del modelo.



Definición de un pipeline de ML

Formalmente, un pipeline es una secuencia de pasos que automatizan el flujo de trabajo de un proyecto de ML, desde la preparación de datos hasta el entrenamiento y la evaluación del modelo.

Preprocesamiento de datos

Transformaciones como la limpieza, la codificación y la estandarización de datos para que sean compatibles con los algoritmos de ML.

Entrenamiento del modelo

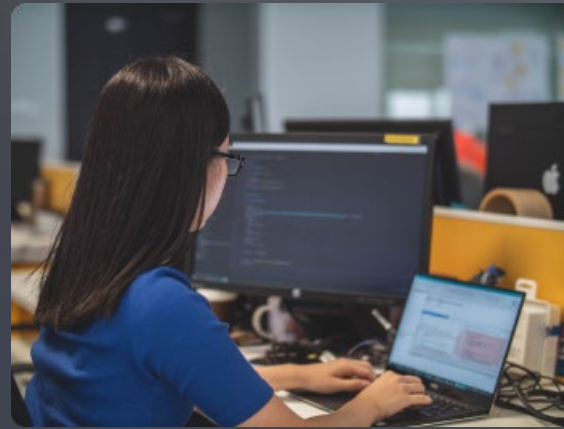
El entrenamiento del modelo de ML utilizando los datos preprocesados para optimizar su desempeño en la tarea específica.

Evaluación del modelo

Evaluar el rendimiento del modelo entrenado utilizando métricas adecuadas para la tarea en cuestión.

Predicción

Utilizar el modelo entrenado para hacer predicciones sobre nuevos datos no vistos.



Usos y beneficios de los pipelines de ML

Los pipelines de ML son utilizados en una amplia gama de aplicaciones de aprendizaje automático, como la clasificación de imágenes, la detección de fraudes y el análisis predictivo.

1 Automatización

Automatizar las tareas repetitivas de un proyecto de ML, lo que reduce el tiempo y esfuerzo manual.

2 Reproducibilidad

Garantizar la coherencia y reproducibilidad de los resultados del proyecto de ML.

3 Escalabilidad

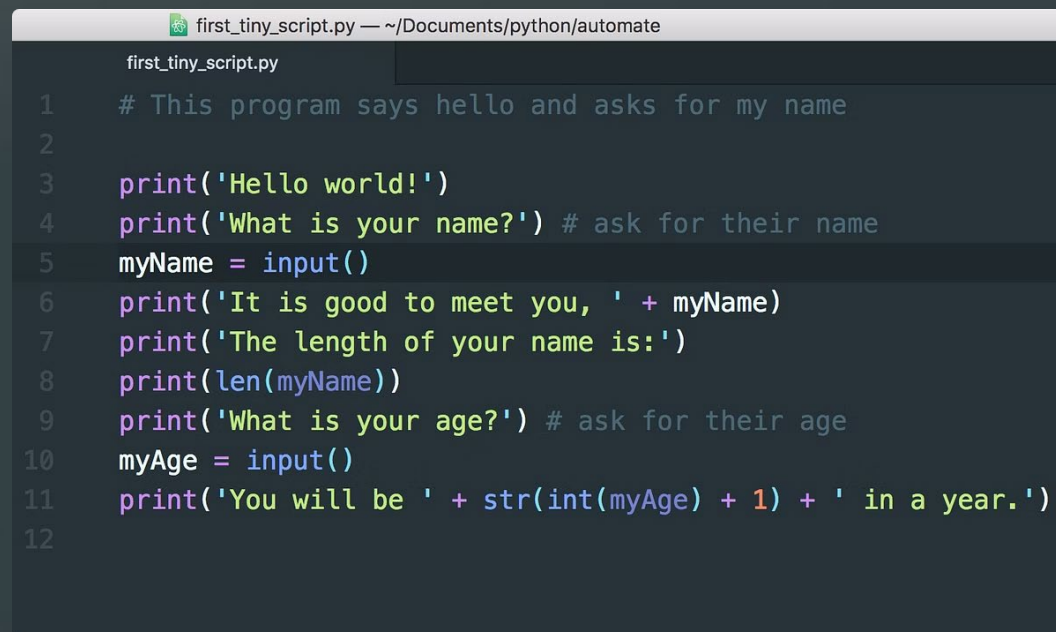
Facilitar la escalabilidad del proyecto de ML para manejar grandes conjuntos de datos.

4 Codificación

Mejorar la eficiencia del código y simplificar la gestión de dependencias.

Pipelines en Scikit-learn (python)

Scikit-learn es una biblioteca de aprendizaje automático de Python que ofrece herramientas para la creación de pipelines de ML.



```
first_tiny_script.py — ~/Documents/python/automate
first_tiny_script.py
1 # This program says hello and asks for my name
2
3 print('Hello world!')
4 print('What is your name?') # ask for their name
5 myName = input()
6 print('It is good to meet you, ' + myName)
7 print('The length of your name is:')
8 print(len(myName))
9 print('What is your age?') # ask for their age
10 myAge = input()
11 print('You will be ' + str(int(myAge) + 1) + ' in a year.')
12
```

1

Crear un pipeline

Definir los pasos del pipeline utilizando la clase Pipeline de Scikit-learn.

2

Añadir etapas

Agregar las etapas de preprocesamiento, entrenamiento del modelo y evaluación del modelo al pipeline.

3

Ajustar el pipeline

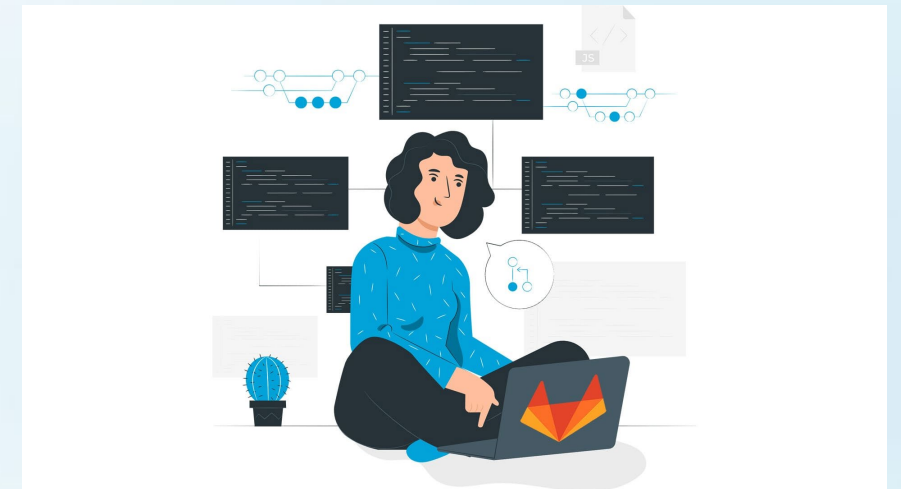
Ajustar el pipeline a los datos de entrenamiento para entrenar el modelo.

4

Realizar predicciones

Utilizar el pipeline entrenado para realizar predicciones sobre nuevos datos.

Versionado (control de versiones)



Importancia del control de versiones de versiones de datos

El control de versiones de datos es fundamental para la gestión de proyectos de ML, ya que permite que permite rastrear los cambios realizados en los datos y volver a versiones anteriores si es necesario.

1

Reproducibilidad

Garantizar la reproducibilidad de los resultados del proyecto de ML, al poder acceder a versiones específicas de los datos.

2

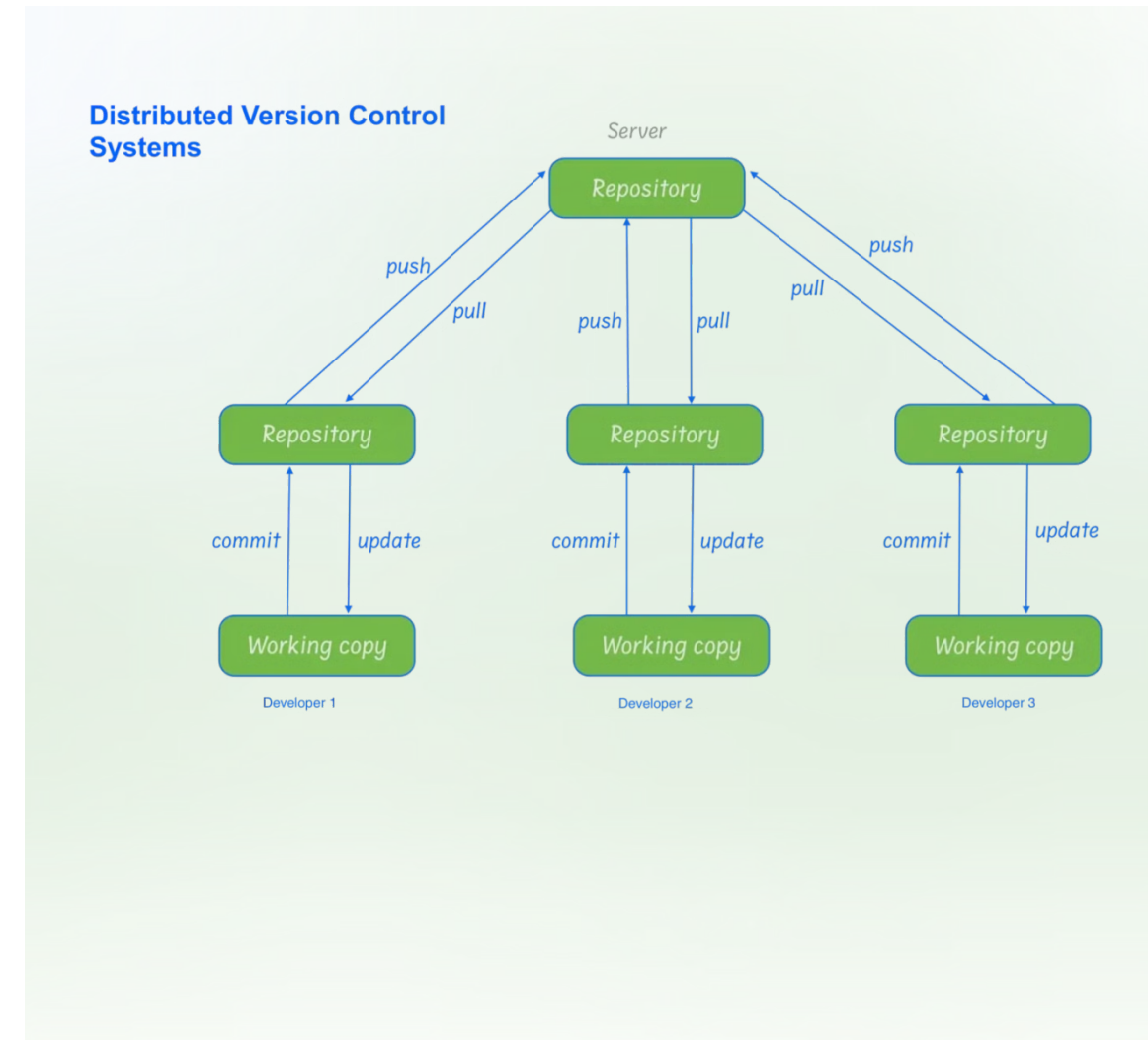
Colaboración

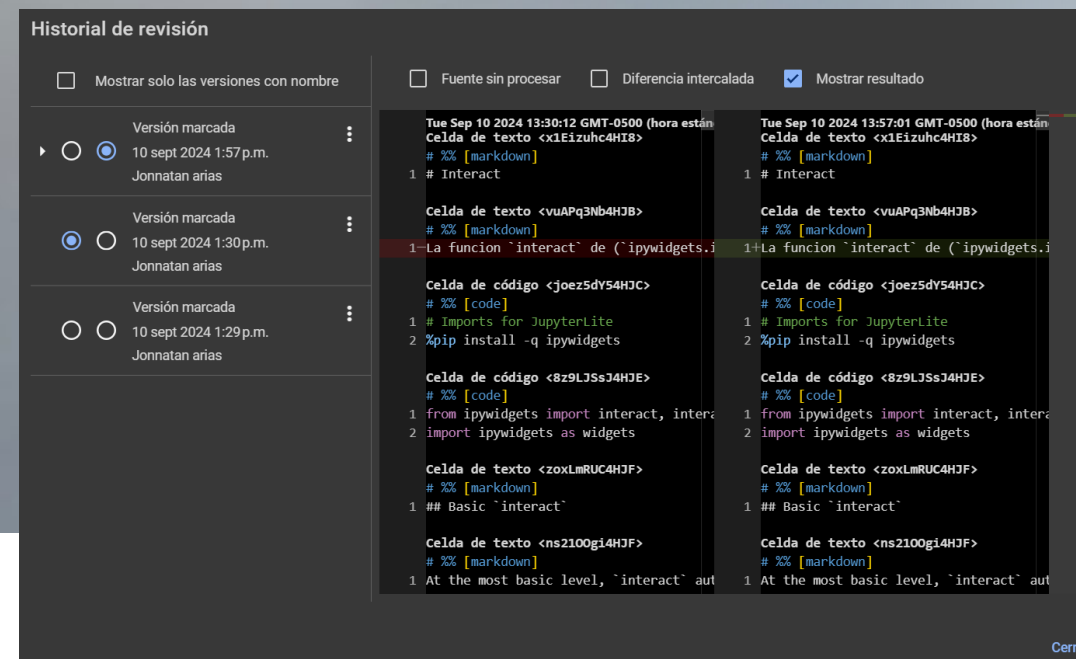
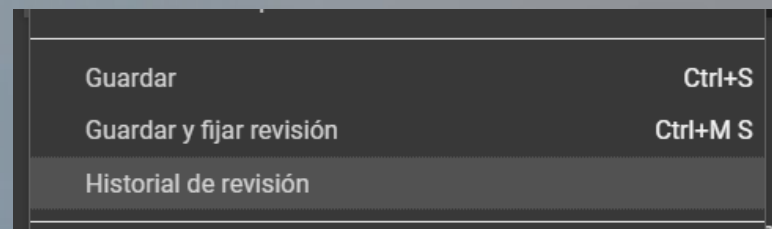
Facilitar la colaboración entre diferentes miembros del equipo de ML.

3

Control de errores

Permitir la detección y corrección de errores en los datos, al poder revertir a versiones anteriores.





Introducción a DVC (Data Version Control) en Colab

DVC (Data Version Control) es una herramienta de código abierto que extiende Git para el control de versiones de datos, archivos y código de ML.



Control de versiones de datos

Gestionar versiones de los datos utilizados en un proyecto de ML.



Control de versiones de código

Gestionar versiones del código del proyecto de ML.



Seguimiento de experimentos

Rastrear las diferentes configuraciones y resultados de los experimentos de ML.



Versionado de datasets en proyectos de ML

DVC facilita el versionado de datasets al crear una rama independiente para cada versión del dataset, lo que permite acceder a versiones específicas del dataset cuando sea necesario.

Dataset	Versión	Fecha	Descripción
data.csv	v1	2023-08-01	Versión inicial del dataset
data.csv	v2	2023-08-05	Actualización de datos con nuevos registros
data.csv	v3	2023-08-10	Corrección de errores en los datos



Siguientes modulos:

4-5 Jueves

6-8 Viernes

Asistencia:

<https://tally.so/r/mD5PqR>