

DATA COLLECTION



Eka Praja Wiyata Mandala, S.Kom, M.Kom

Email : ekaprajawm@upiyptk.ac.id

Dosen Teknik Informatika Universitas Putra Indonesia YPTK Padang



About Me



Eka Praja Wiyata Mandala, S.Kom, M.Kom

NIDN : 1014088502



ekaprajawm@upiypk.ac.id



ekapraja199@gmail.com



40Jr_ZQAAAAJ



085213873216



S1 – Teknik Informatika
Universitas Esa Unggul
Jakarta



S2 – Teknik Informatika
Universitas Putra Indonesia
YPTK Padang



Eka Praja Wiyata Mandala, S.Kom, M.Kom

Data Collection



Eka Praja Wiyata Mandala, S.Kom, M.Kom



Data Collection

Definisi Data Coleection

Data collection merupakan **langkah pertama** dalam **pipeline Data Science dan Machine Learning**.

Proses ini sangat penting karena **kualitas** dan **relevansi** data yang dikumpulkan akan berdampak langsung pada hasil model machine learning yang akan dibangun.

Data yang baik akan membantu **model belajar** secara **efektif**, sementara data yang **buruk** dapat menghasilkan **model yang tidak akurat** atau **bias**.



Data Collection

Tujuan Data Collection

Data collection bertujuan untuk **mengumpulkan data yang relevan, akurat, dan cukup dalam jumlah** untuk **mendukung proses analisis dan pelatihan model**

Tujuan utama meliputi :

- ❖ **Mengidentifikasi** pola atau tren.
- ❖ **Memprediksi** hasil berdasarkan data historis.
- ❖ **Mendukung** pengambilan keputusan berbasis data



Data Collection

Jenis Data

Data yang dikumpulkan untuk analisis dalam machine learning berupa:

- ❖ **Structured Data:** Data terstruktur dalam bentuk tabel (seperti di database), memiliki kolom dan baris yang jelas. Contoh: data pelanggan, data transaksi.
- ❖ **Unstructured Data:** Data yang tidak memiliki format terstruktur, seperti teks, gambar, video. Contoh: posting media sosial, review produk.
- ❖ **Semi-structured Data:** Data yang tidak sepenuhnya terstruktur namun memiliki elemen atau metadata yang bisa dianalisis, seperti XML atau JSON.



Data Collection

Sumber Data

Data dapat diperoleh dari berbagai sumber, antara lain:

- ❖ **Internal Data:** Data yang berasal dari sistem internal perusahaan atau organisasi, seperti database penjualan, sistem ERP, CRM.
- ❖ **External Data:** Data yang diperoleh dari sumber eksternal seperti survei, data dari pihak ketiga, atau data dari internet (scraping web, open data).
- ❖ **Public Data:** Data publik yang tersedia secara online, seperti data pemerintah (open data), dataset dari platform seperti Kaggle, UCI Machine Learning Repository.



Data Collection

Metode Pengumpulan Data

Ada beberapa metode utama untuk mengumpulkan data, tergantung pada sumber dan jenis datanya:

- ❖ **Manual Data Collection:** Pengumpulan data secara manual, misalnya melalui survei, kuesioner, atau wawancara.
- ❖ **Automated Data Collection:** Menggunakan skrip atau software untuk otomatisasi pengambilan data. Contoh: web scraping, API.
- ❖ **Sensors/IoT:** Pengumpulan data real-time dari sensor atau perangkat IoT, misalnya untuk data cuaca, data mesin di industri manufaktur.
- ❖ **Logs dan Tracking:** Data yang dikumpulkan dari log server, alat analitik web seperti Google Analytics, atau aplikasi mobile.



Data Collection

Tantangan dalam Pengumpulan Data

- ❖ **Volume dan Ukuran Data:** Mengumpulkan data dalam jumlah besar membutuhkan penyimpanan dan pengolahan yang efisien.
- ❖ **Kualitas Data:** Data yang dikumpulkan mungkin memiliki masalah kualitas, seperti missing data, duplikasi, atau noise.
- ❖ **Integrasi Data:** Menggabungkan data dari berbagai sumber sering kali menimbulkan tantangan dalam hal kompatibilitas format atau struktur.
- ❖ **Etika dan Privasi:** Pengumpulan data harus mematuhi regulasi terkait privasi seperti **GDPR (General Data Protection Regulation)** dan menjaga etika dalam penggunaannya.



Data Collection

Data Collection dalam Machine Learning

Untuk membangun model machine learning yang kuat, data harus:

- ❖ **Representatif:** Data harus mencerminkan populasi sebenarnya yang ingin dianalisis atau diprediksi.
- ❖ **Bersih (Clean):** Data yang tidak lengkap, bias, atau berisik harus diperbaiki melalui proses preprocessing.
- ❖ **Terkait dengan Masalah yang Ingin Diselesaikan:** Pastikan data relevan dengan tujuan analisis atau masalah yang sedang dihadapi.



Data Collection

Tools untuk Data Collection

- ❖ **Web Scraping Tools:** BeautifulSoup, Scrapy, Selenium – digunakan untuk mengambil data dari web.
- ❖ **API Access Tools:** Requests (Python), Postman – digunakan untuk mengambil data dari API publik.
- ❖ **Database Query Tools:** SQL, Pandas (untuk manipulasi data di Python) – digunakan untuk mengambil data dari database.
- ❖ **Data Collection Tools dari Platform:** Google Analytics, AWS Kinesis, Azure Data Factory.



Data Collection

Best Practices dalam Data Collection

- ❖ **Data Sampling:** Dalam kasus data yang sangat besar, sampling dapat digunakan untuk mengambil sebagian kecil data yang tetap representatif.
- ❖ **Data Quality Checks:** Setelah pengumpulan, selalu lakukan quality check untuk memastikan data yang diambil memenuhi standar kualitas.
- ❖ **Documenting Data Collection Process:** Dokumentasi yang baik mencakup sumber data, metode pengumpulan, dan potensi bias data.



Data Collection

Data Collection pada E-Commerce

Misalkan kita membangun sistem rekomendasi produk untuk platform ecommerce. Data yang perlu dikumpulkan meliputi:

- ❖ **Data Transaksi:** Pembelian produk oleh pengguna (**structured**).
- ❖ **Review Produk:** Komentar dan rating dari pengguna (**semi-structured/ unstructured**).
- ❖ **Data Perilaku Pengguna:** Log kunjungan halaman, klik pada produk, data interaksi di website (**semi-structured**).



Google Colaboratory



Eka Praja Wiyata Mandala, S.Kom, M.Kom



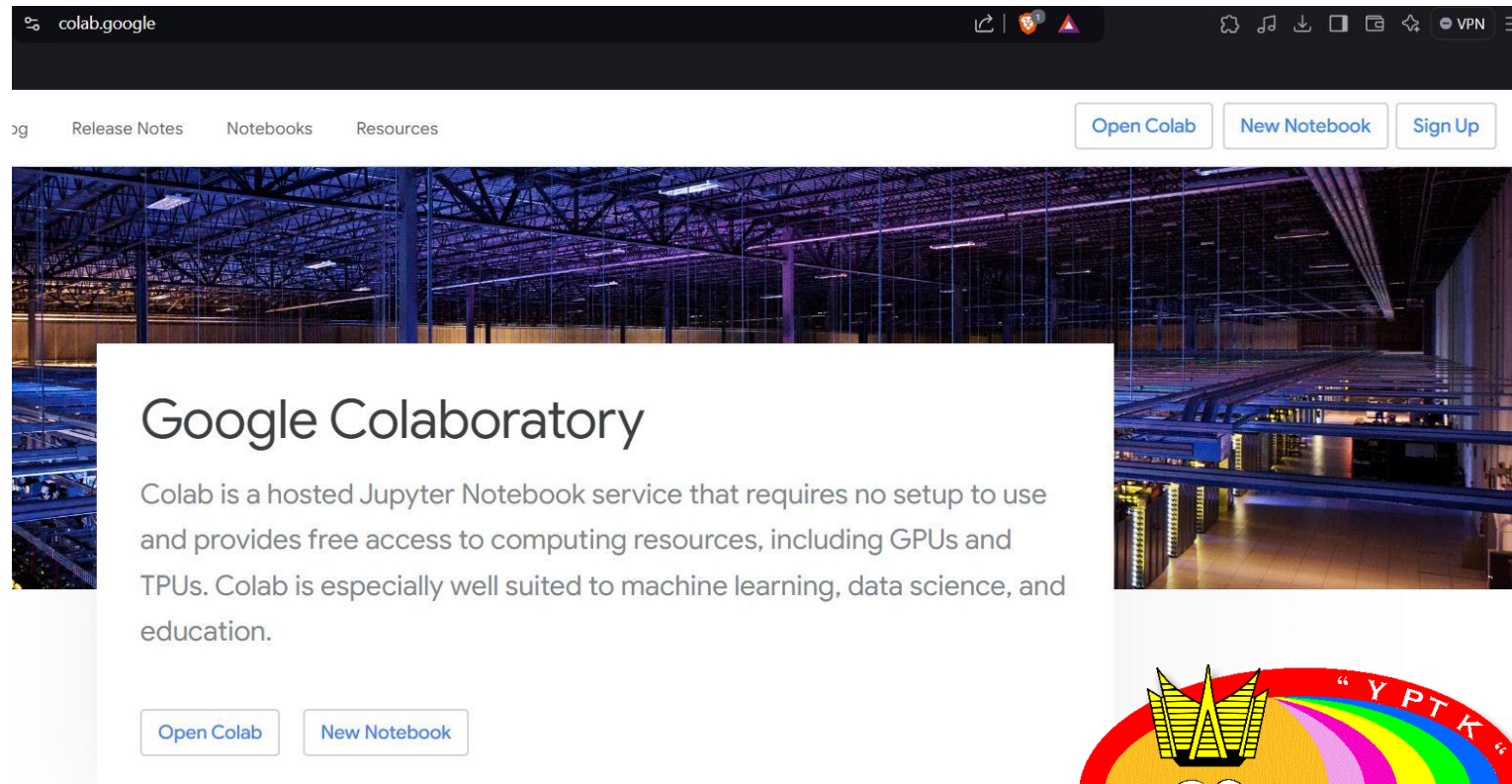
Google Colaboratory

Data Collection dengan Google Colaboratory

Akses Google Colaboratory melalui : <https://colab.research.google.com/>

Syarat :

1. Memiliki akses Internet.
2. Memiliki Account Google
3. Data diakses dengan laptop ataupun Smartphone

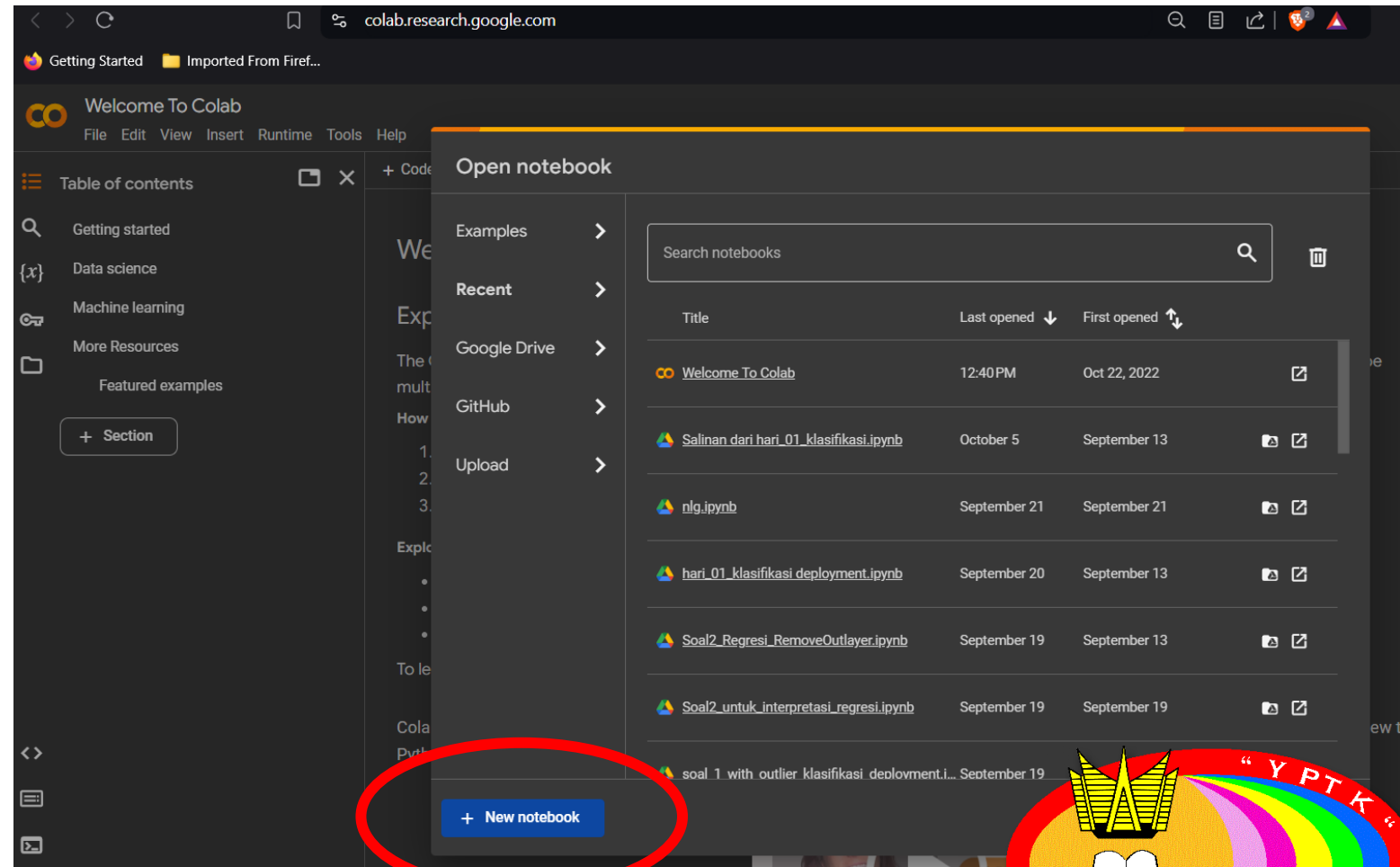


Google Colaboratory

Data Collection dengan Google Colaboratory

Untuk Pengoperasian Perdana colab tidak memiliki project apapun sehingga pada bagian project akan kosong.

Lakukan pembuatan notebook baru dengan memilih **New Notebook**



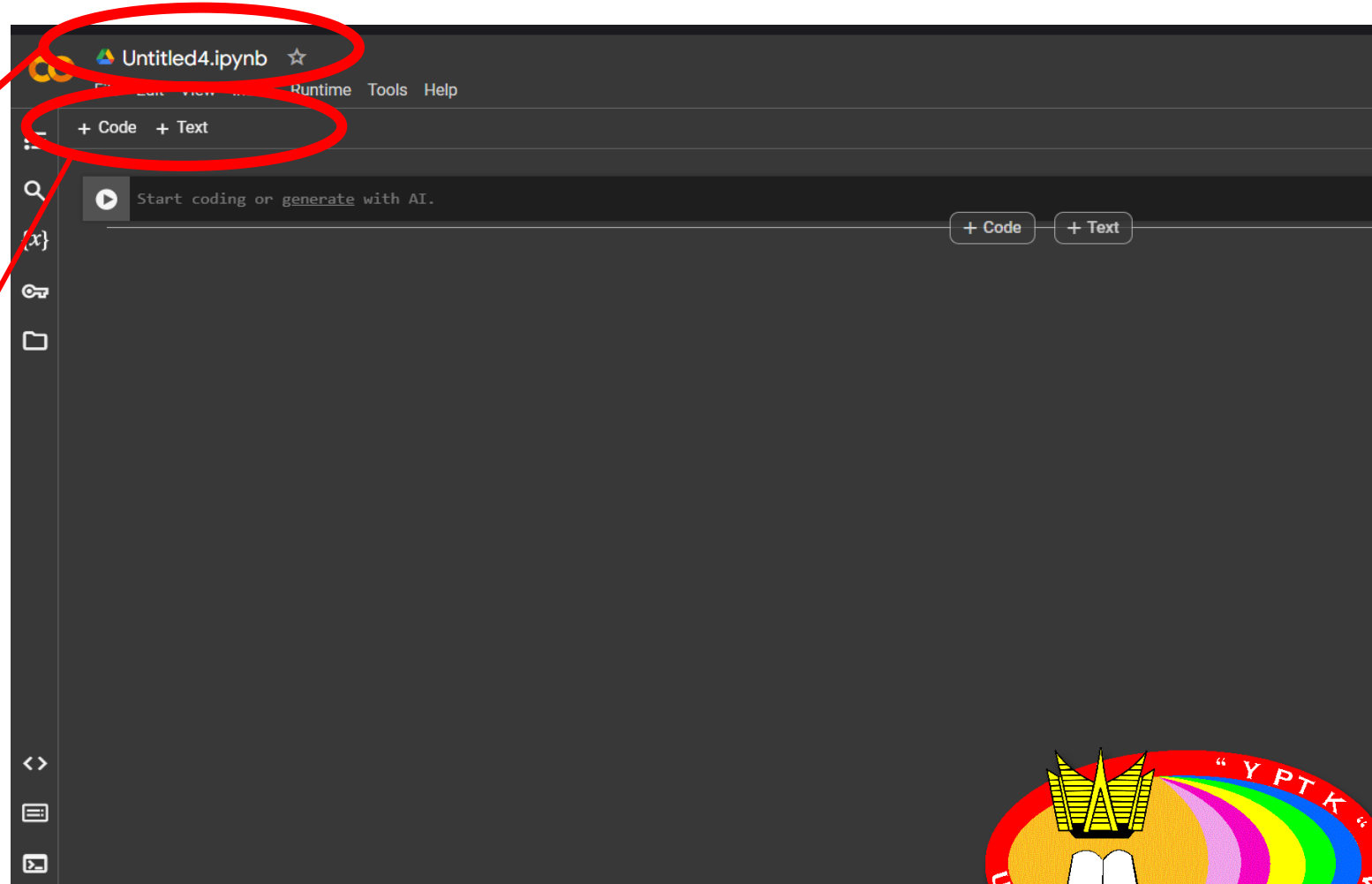
Google Colaboratory

Data Collection dengan Google Colaboratory

Untuk menamai project double klik pada nama project

Bagian ini digunakan untuk menambah baris kode dan Teks.

Catt:
Rename project dengan nama **Unit1** dan simpan (pilih **file** dan **save**).



Google Colaboratory

Data Collection dengan Google Colaboratory

Untuk melihat file tersimpan, silahkan buka **google drive** dan pastikan folder **Colab Notebooks** ada di dalam drive.

My Drive ▾

Type ▾

People ▾

Modified ▾

Pelaporan BKD 20202

PAK Lektor 300

PAK

Modul Database

Materi Ajar

Laporan PKL 20221

Jurnal Revisi Kode_...

hasil koreksi ujian

Form Perbaikan Abs...

Form Pengumpulan ...

Form Pengumpulan ...

Data Warehouse

Contoh Presentasi R...

Colab Notebooks

Classroom

Bukti Perkuliahan

Bimbingan Skripsi

Bahan Ujian Augme...

Bahan Ajar Rekayas...

AR dan VR

Absensi Perkuliahan...

Files



Google Colaboratory

Data Collection dengan Google Colaboratory

Jika penyimpanan sudah benar maka file **Unit1.ipynb** sudah ada di dalam **folder Colab Notebooks**

co Untitled0.ipynb

co Untitled

co Unit1.ipynb



Data Source



Eka Praja Wiyata Mandala, S.Kom, M.Kom



Data Source

Data Source

Terdapat beberapa cara dalam pengumpulan data untuk kegiatan analisis dengan menggunakan pembelajaran mesin.

Sumber data umum dapat diakses melalui :

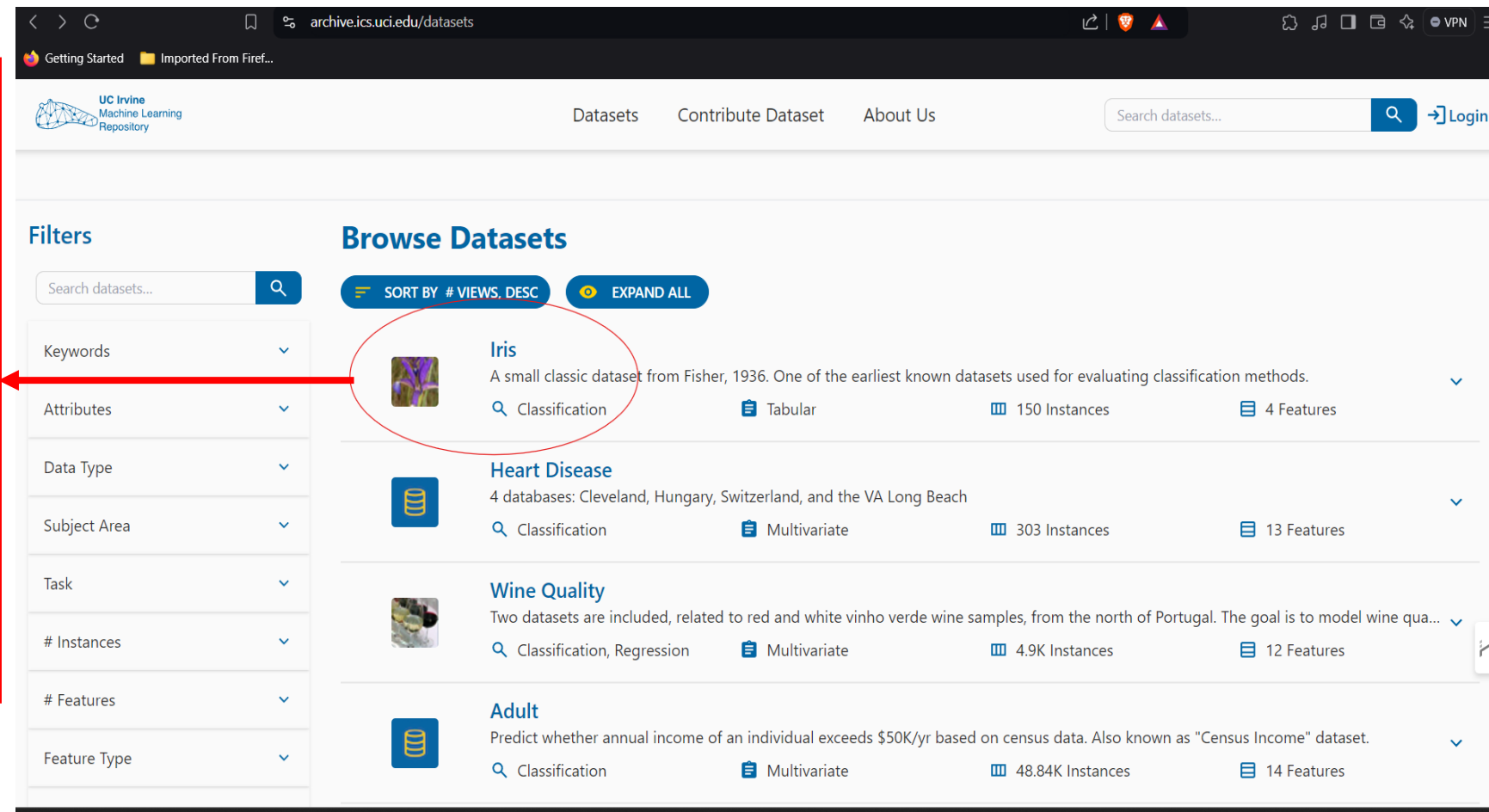
- ❖ <https://archive.ics.uci.edu/>
- ❖ <https://www.kaggle.com/datasets>
- ❖ <https://github.com/datasets/>
- ❖ <https://paperswithcode.com/datasets>



Data Source

<https://archive.ics.uci.edu/>

Contoh :
Ambil dataset dari UCI
Dataset.
Klik pada dataset Iris
untuk diambil dan
diproses dengan
Menggunakan Google
Colab.

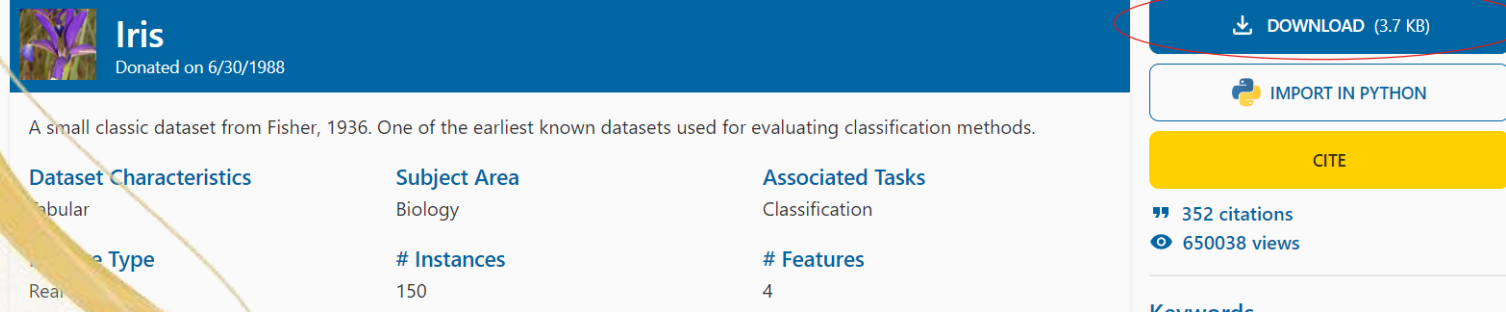


The screenshot shows the UC Irvine Machine Learning Repository website. The 'Browse Datasets' section is active, displaying a list of datasets. The 'Iris' dataset is highlighted with a red circle and a red arrow pointing to it from the text on the left. The 'Iris' dataset is described as a small classic dataset from Fisher, 1936, used for evaluating classification methods. It is a tabular dataset with 150 instances and 4 features. Other datasets listed include 'Heart Disease', 'Wine Quality', and 'Adult'.

Dataset Name	Description	Task	Instances	Features
Iris	A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.	Classification	150 Instances	4 Features
Heart Disease	4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach	Classification	303 Instances	13 Features
Wine Quality	Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine qua...	Classification, Regression	4.9K Instances	12 Features
Adult	Predict whether annual income of an individual exceeds \$50K/yr based on census data. Also known as "Census Income" dataset.	Classification	48.84K Instances	14 Features

Data Source

<https://archive.ics.uci.edu/>



The screenshot shows the Iris dataset page on the UCI Machine Learning Repository. The page has a blue header with the dataset name 'Iris' and a small image of an iris flower. Below the header, there is a description: 'A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.' The page is divided into three columns: 'Dataset Characteristics', 'Subject Area', and 'Associated Tasks'. The 'Dataset Characteristics' column lists 'Tabular' and 'Real'. The 'Subject Area' column lists 'Biology'. The 'Associated Tasks' column lists 'Classification'. Below these columns, there are statistics: '# Instances' is 150 and '# Features' is 4. On the right side, there are three buttons: 'DOWNLOAD (3.7 KB)' (highlighted with a red circle and an arrow pointing to a text box), 'IMPORT IN PYTHON', and 'CITE'. Below the buttons, there are statistics: '352 citations' and '650038 views'. At the bottom, there is a 'Keywords' section.

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Biology	Classification
Real		

Instances: 150
Features: 4

352 citations
650038 views

Keywords

Download Dataset
Iris



Data Source

<https://archive.ics.uci.edu/>

Name	Date modified	Type	Size
▼ Today			
bezdeklris.data	09/10/2024 14:23	DATA File	5 KB
Index	09/10/2024 14:23	File	1 KB
iris.data	09/10/2024 14:23	DATA File	5 KB
iris.names	09/10/2024 14:23	NAMES File	3 KB

Lakukan
pengekstrakan pada
dataset, sehingga
akan terlihat
4 file seperti pada
gambar.



Data Source

<https://archive.ics.uci.edu/>

nlq.ipynb

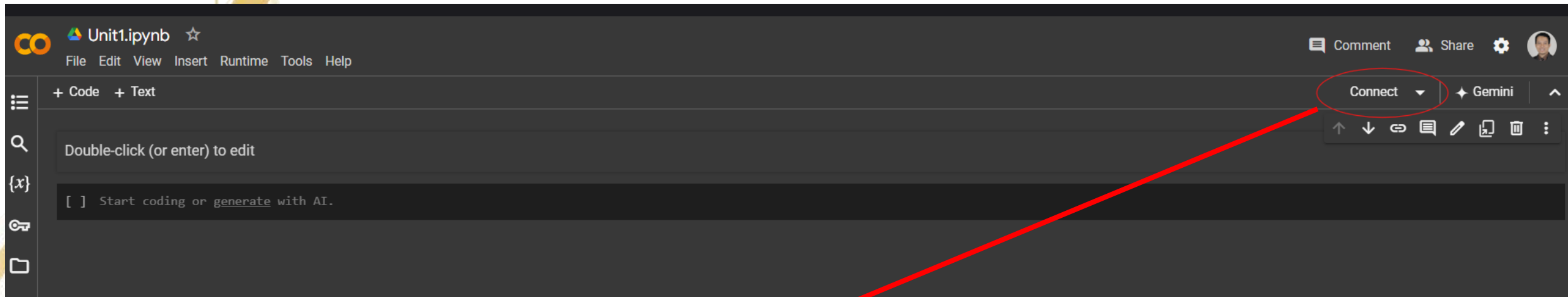


iris.data

Upload file **iris.data**
ke dalam folder
Colab Notebooks di
Google Drive



Data Source



Aktifkan Google Colab dengan memberikan perintah **Connect** dengan engine **Python google**.



Data Source

```
✓ [2] # Define the variable  
0s folder_name = "/content/drive/My Drive/Colab Notebooks/"
```

Directory dari file project dan pendukung ditentukan secara **default** akan berada pada

`"/content/drive/My Drive/Colab Notebooks/"`



Data Source

```
from google.colab import drive
drive.mount('/content/drive', force_remount=True)

import sys
sys.path.append(f'{folder_name}')
```

Permit this notebook to access your Google Drive files?

This notebook is requesting access to your Google Drive files. Granting access to Google Drive will permit code executed in the notebook to modify files in your Google Drive. Make sure to review notebook code prior to allowing this access.

No thanks

Connect to Google Drive

Kode ini digunakan untuk
Menghubungkan google colab
dengan google drive.



Data Source

```
✓ 10s ▶ from google.colab import drive  
drive.mount('/content/drive', force_remount=True)  
  
import sys  
sys.path.append(f'{folder_name}')
```

Mounted at /content/drive

Koneksi sukses maka informasi **drive mounted** akan muncul.



Data Source

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
pd.set_option('display.max_columns', None)
```

Secara umum library ini adalah :

1. Pandas = mengatur data kedalam bentuk dataframe
2. Numpy = Manipulasi Data dengan menggunakan operasi matematika
3. Matplotlib = visualisasi statis, animasi dan interaktif
4. Seaborn = Pustaka matplotlib untuk membuat Grafik



Data Source

```
▶ column_names = ['sepal_length', 'sepal_width', 'petal_lenght', 'petal_width', 'class']  
  
df = pd.read_csv(f'{folder_name}/iris.data', sep=',', names=column_names, skipinitialspace=True, na_values="?")  
  
df.head(5)
```

```
▶ url = "https://archive.ics.uci.edu/dataset/53/iris"  
column_names = ['sepal_length', 'sepal_width', 'petal_lenght', 'class']  
df.head(5)
```



Thank's

Thanks for your attention



Eka Praja Wiyata Mandala, S.Kom, M.Kom

