Kode program yang akan diajarkan di sini bisa diunduh di tautan [berikut ini](https://colab.research.google.com/drive/1WLORwkaqSXBU-XlpxQZpe-wocSXWYCSI?usp=sharing). Untuk membukanya, upload berkasnya Google Colab.

**Tujuan**

Pada codelab ini kita akan belajar membagi dataset menggunakan fungsi Train Test Split dari *library*SKLearn.

**Tahapan Latihan**

Untuk latihan membagi dataset terdiri dari tahapan-tahapan sebagai berikut:

1. Persiapkan dataset ke dalam Notebook.
2. Impor library SKLearn.
3. Buat variabel untuk menampung data training dan data testing.
4. Panggil fungsi *train\_test\_split()*.

**Codelab**

Pada Google Colab, import library yang dibutuhkan.

1. import sklearn
2. from sklearn import datasets

Library sklearn menyediakan **dataset iris** yakni sebuah dataset yang umum digunakan untuk masalah klasifikasi. Dataset ini memiliki jumlah 150 sampel. Untuk mendapatkan dataset, kita bisa menulis kode berikut pada cell baru.

1. # load iris dataset
2. iris = datasets.load\_iris()

Dataset iris dari library sklearn belum dapat langsung dipakai oleh sebuah model ML. Sesuai dengan yang telah dibahas pada modul terdahulu, kita harus memisahkan antara atribut dan label pada dataset.

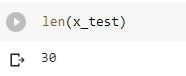
1. # pisahkan atribut dan label pada iris dataset
2. x=iris.data
3. y=iris.target

Untuk membuat train set dan test set kita tinggal memanggil fungsi train\_test\_split. Train\_test\_split memiliki parameter x yaitu atribut dari dataset, y yaitu target dari dataset, dan test\_size yaitu persentase test set dari dataset utuh. Train\_test\_split mengembalikan 4 nilai yaitu, atribut dari train set, atribut dari test set, target dari train set, dan target dari test set.

1. from sklearn.model\_selection import train\_test\_split
3. # membagi dataset menjadi training dan testing
4. x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=1)

Ketika kita print panjang dari x\_test, kita bisa melihat bahwa panjang dari atribut test set adalah **30 sampel**, sesuai dengan parameter yang kita masukkan pada fungsi train\_test\_split yaitu 0.2 atau 20% dari 150 sampel. Kode untuk print panjang dari x\_test seperti di bawah ini.

1. # menghitung panjang/jumlah data pada x\_test
2. len(x\_test)



Pada tahap ini dataset kita telah siap dipakai untuk pelatihan model machine learning.