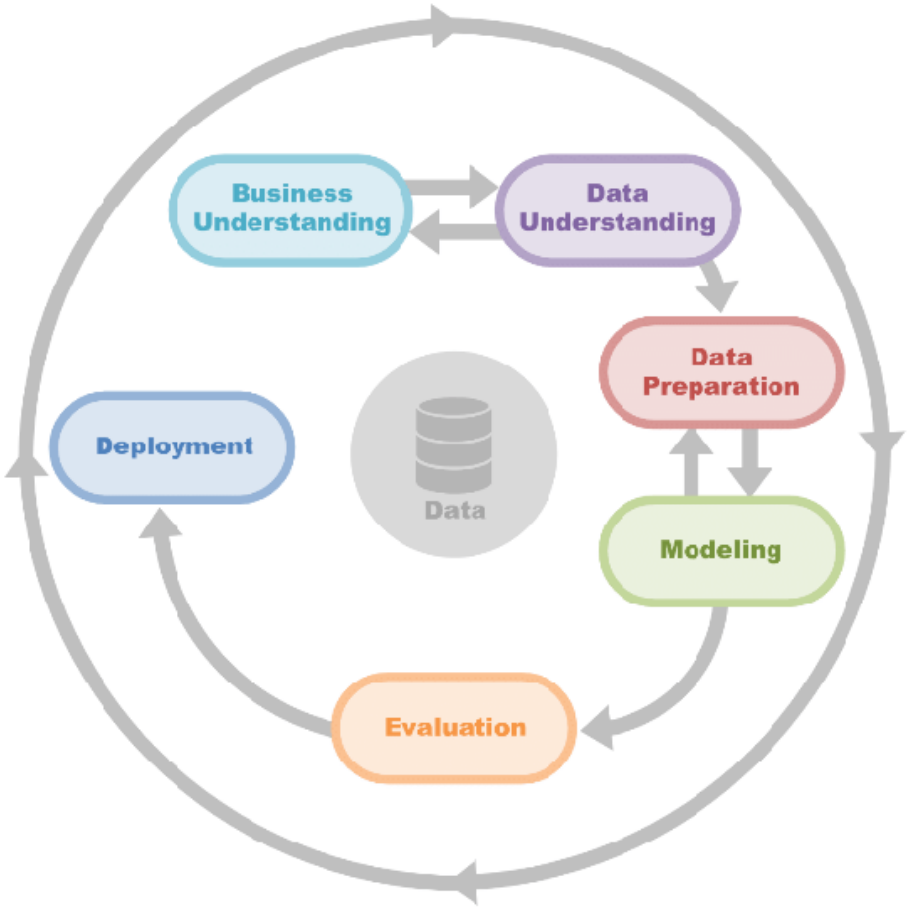
Modul Praktikum 1
Data Mining

PROGRAM STUDI SAINS DATA
JURUSAN SAINS
INSTITUT TEKNOLOGI SUMATERA
2024

Pengantar Data Mining

Data mining adalah proses menemukan pola, tren, dan informasi berguna dari kumpulan data besar dengan menggunakan teknik statistik, algoritma, dan metode analisis. Tujuannya untuk mengidentifikasi informasi yang mungkin tidak terlihat secara langsung atau yang tersembunyi dalam data tersebut. Proses ini sering digunakan untuk membantu membuat keputusan bisnis, memahami perilaku pelanggan, atau mengidentifikasi peluang baru. Proses data mining umumnya melibatkan beberapa langkah yang dirancang untuk mengambil data mentah dan mengubahnya menjadi informasi yang berguna.



Gambar 1. Proses Data Mining

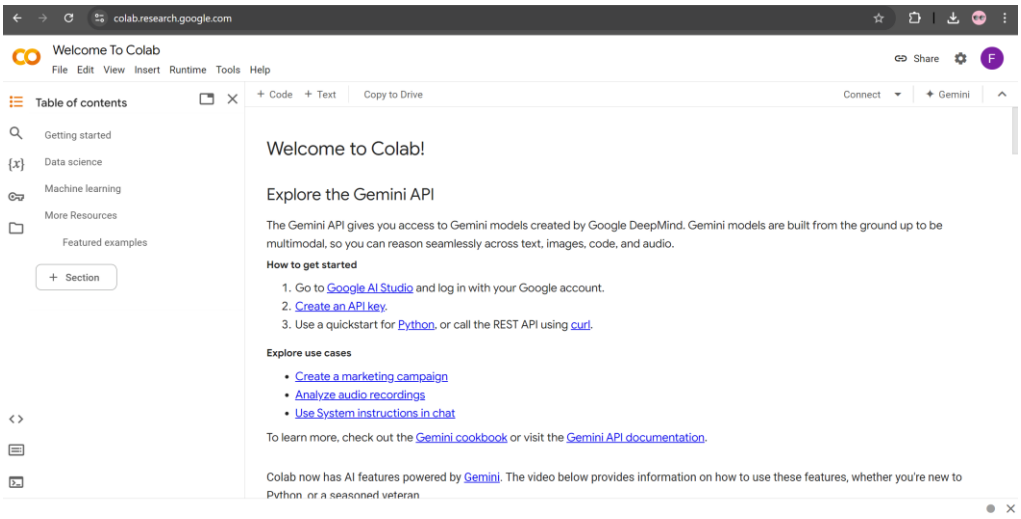
Dalam proses data mining, tahapan pertama yang harus dilakukan adalah Data Understanding dan Data Preparation. Untuk itu Praktikum pertama mata kuliah Data Mining akan membahas tentang Data Understanding dan konsep dasar Data Preparation (Data Preprocessing).

Tujuan Praktikum

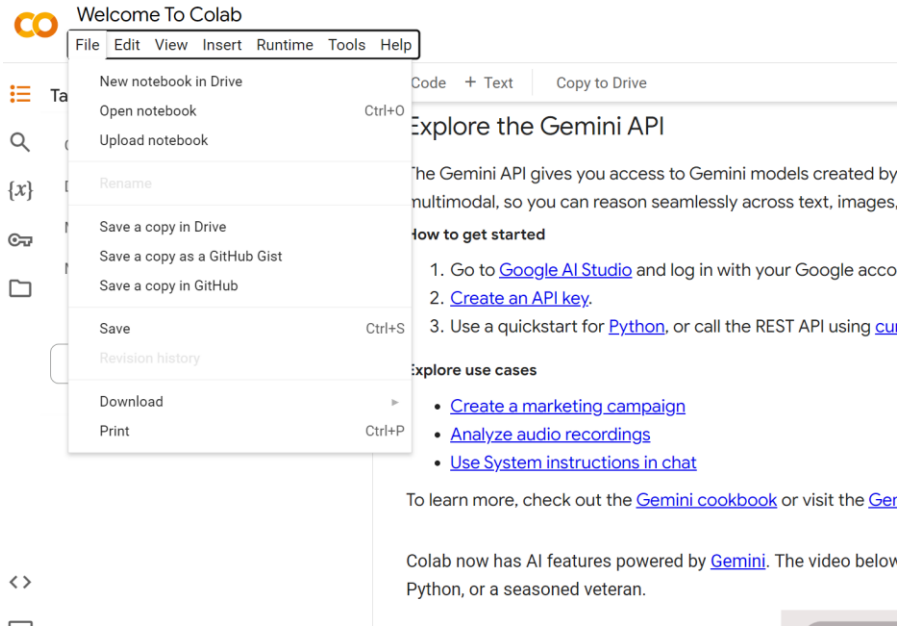
1. Memahami konsep dasar tentang data mining
2. Mempelajari konsep Data Understanding yang mencakup mendeskripsikan data (objek data, type data, atribute data, baris dan kolom) dan eksplorasi data
3. Mempelajari konsep dasar Data Preparation yang mencakup verifikasi kualitas data (Data Preprocessing).

Pada praktikum Data Mining, akan digunakan IDE Python Service dari Google Colab. Praktikan dipersilahkan menyiapkan akun google agar dapat mengakses Google Colab. Langkah-langkah membuka Google Colab dapat dilakukan dengan cara berikut:

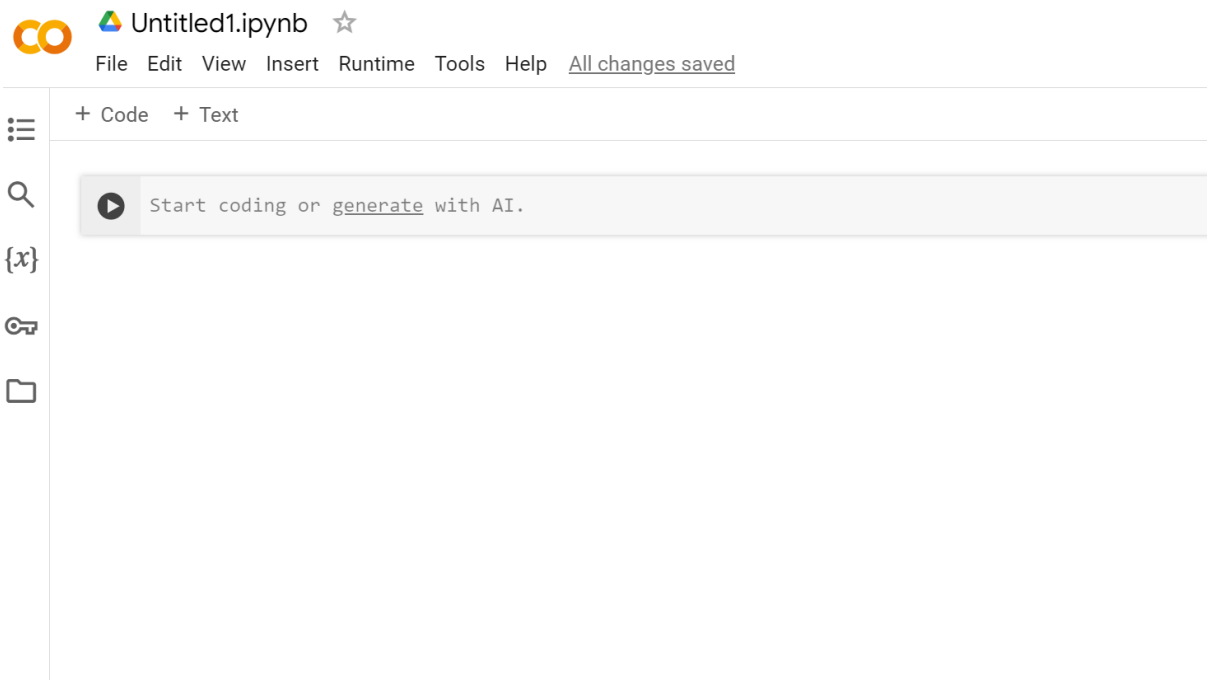
- Pergi ke laman platfrom google colabs yaitu https://colab.research.google.com/
- Akan ditampilkan laman google colabs seperti berikut



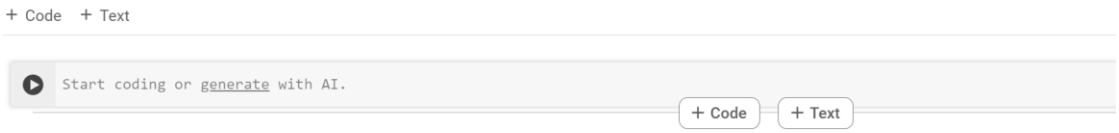
- Klik File lalu pilih New notebook in Drive



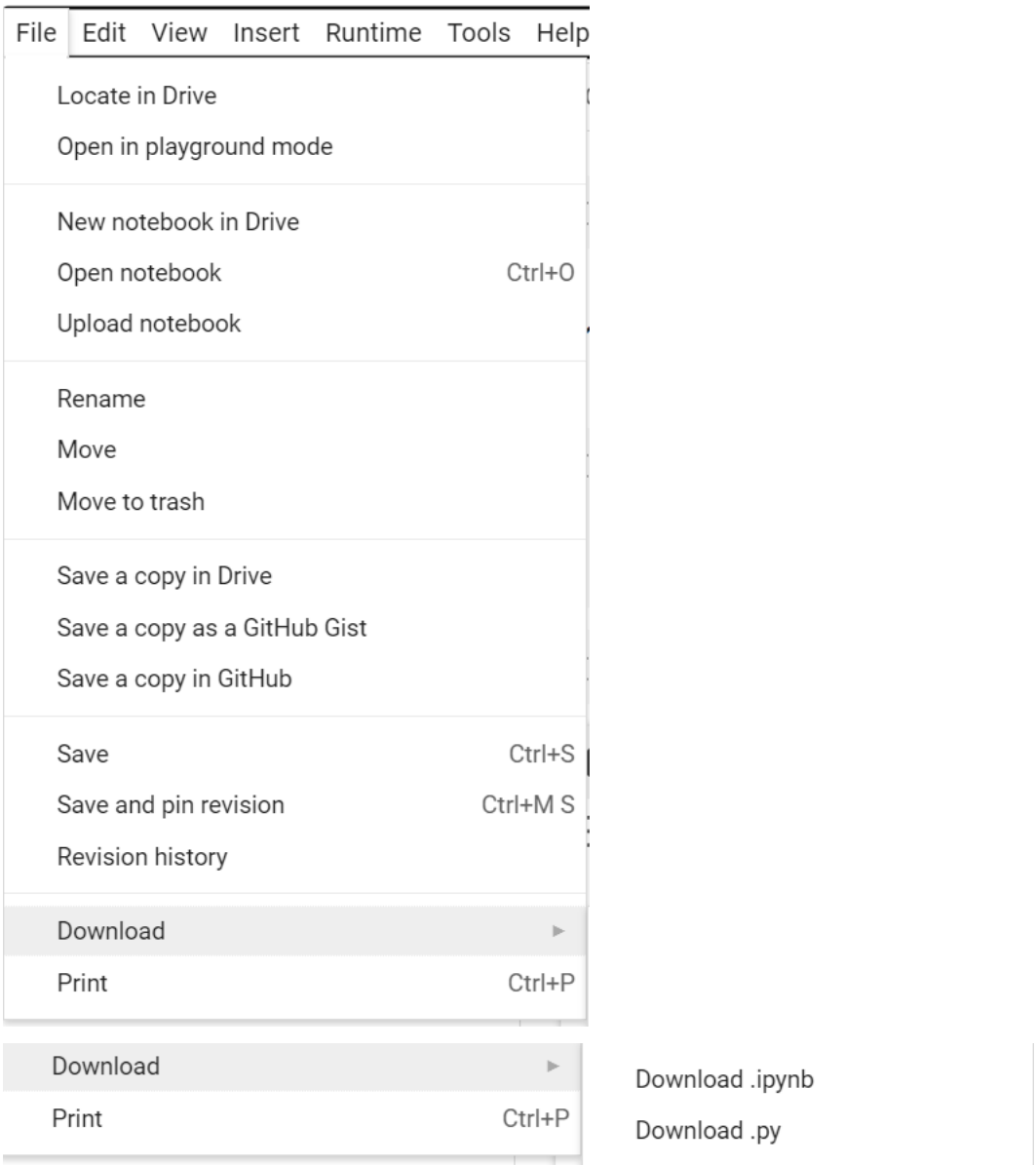
- Ubahlah nama notebook pada “Untitled1.ipynb” menjadi “Praktikum1_DataMining_NIM.ipynb”



- Klik + Code untuk menambahkan baris coding



- Silahkan mulai Praktikumnya, lakukan perubahan nama untuk setiap kali praktikan melakukan praktikum dengan mengubah angka pada nama sesuai pertemuan praktikum.
- Setelah praktikum pastikan praktikan mendownload codingan beserta hasilnya dengan cara ke menu File – Lalu Downlod – pilih download “.ipynb”

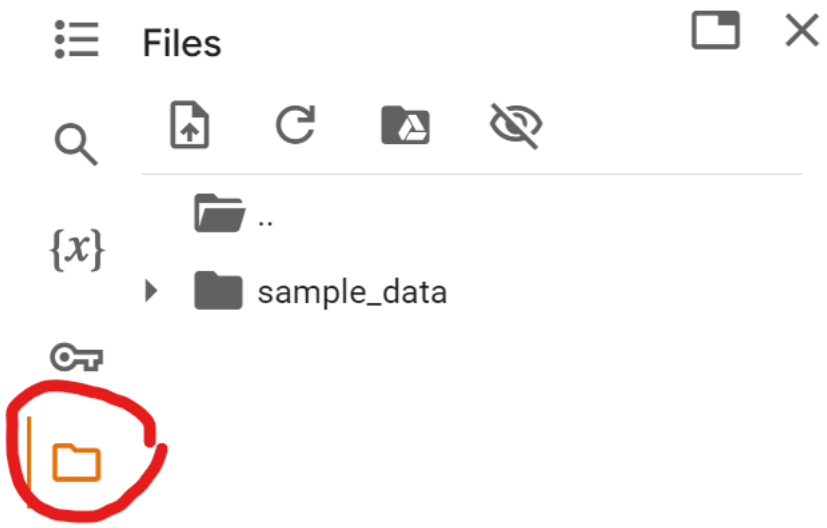


- Lalu, silahkan kumpulkan hasil kodingan dan tugas individu ke link berikut https://forms.gle/wkoKSZZphf9z4Ar28

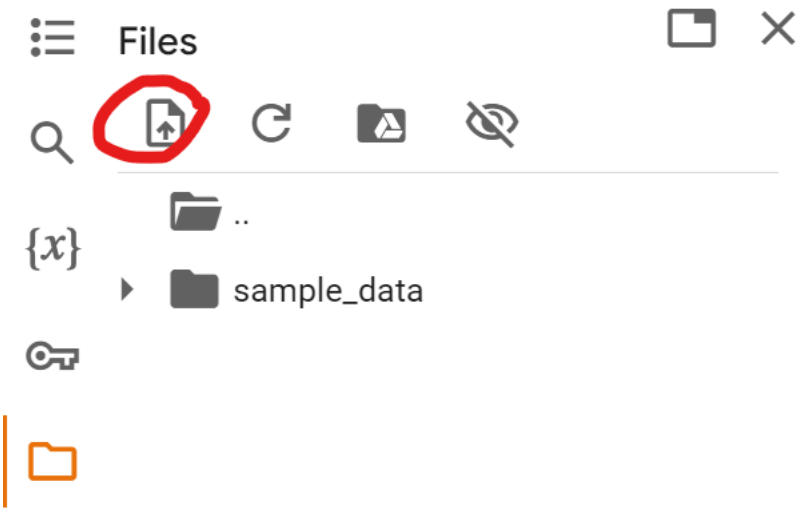
Data Preprocessing

Proses Data Mining seperti Data Understanding dan Data preparation menggunakan Python seringkali menggunakan beberapa pustaka seperti Pandas, NumPy, Matplotlib, dan Seaborn. Berikut tahapan-tahapan dalam Data Understanding dan Data preparation.

1. Upload datasets ke Google Colab



Klik gambar file seperti gambar diatas lalu klik Upload to session storage



Lalu pilih file yang diunduh pada link https://shorturl.at/NosAU

2. Mengimport Pustaka

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

3. Load datasets yang sudah diupload

```
df=pd.read_csv("diabetes.csv") #membuka file csv
```

4. Menampilkan 5 baris pertama dari Data Frame

```
df.head() #Menampilkan 5 baris pertama dari Data Frame
```

5. Menampilkan 5 baris terbawah

```
df.tail() #menampilkan 5 baris terbawah
```

6. Menampilkan baris “n” secara random

```
df.sample(5) #menampilkan baris "n" secara random
```

7. Memeriksa jumlah baris dan kolom

```
df.shape #memeriksa jumlah baris dan kolom
```

8. Melihat informasi dasar tentang dataset

```
df.info() # Melihat informasi dasar tentang dataset
```

9. Menampilkan jenis data tiap kolom

```
df.dtypes #menampilkan jenis data tiap kolom
```

10. Melihat statistik deskriptif dari data numerik

```
df.describe() # Melihat statistik deskriptif dari data numerik
```

11. Menampilkan deskripsi data berupa nilai perhitungan statistik pada atribut tertentu

df[attribute].describe()

```
df['Pregnancies'].describe() #menampilkan deskripsi data berupa nilai perhitungan statistik pada atribut tertentu
```

12. Menampilkan komposisi semua data

```
df.value_counts() #menampilkan komposisi semua data
```

13. Menampilkan komposisi data suatu atribut

df[attribute].value_counts()

```
df['Insulin'].value_counts() #menampilkan komposisi data suatu atribut
```

Selanjutnya akan dilakukan data preprocessing yaitu verifikasi data seperti identifikasi nilai yang hilang, identifikasi duplikat data, identifikasi outlier, dan identifikasi data imbalance. Berikut cara untuk mendeteksi semua masalah data yang biasanya terdapat pada data mentah/*raw data*.

1. Identifikasi nilai yang hilang (Missing Values)

```
df.isnull().sum() #menampilkan jumlah nilai yang hilang di setiap kolom
```

2. Identifikasi Duplikat Data

```
df.groupby('BMI')['BMI'].agg("count") #identifikasi duplikat data pada attribut tertentu
```

```
duplikat = df[df.duplicated(keep=False)] #identifikasi baris duplikat
print("Baris Duplikat:")
print(duplikat)
```

3. Identifikasi Outlier

```
correlation_matrix = df.corr() #menghitung korelasi antar kolom
sns.heatmap(correlation_matrix) #menampilkan matriks korelasi sebagai heatmap
print(correlation_matrix) #menampilkan nilai korelasi
plt.show() #menampilkan plot
sns.boxplot(x='BMI',data=df) #menampilkan data sebagai boxplot
```

4. Identifikasi data tidak berimbang (Data Imbalanced)

```
df['Outcome'].unique() #menampilkan variabel pada atribut Outcome
```

```
df['Outcome'].value_counts() #menampilkan komposisi pada atribut Outcome
```

Setelah mengidentifikasi missing values, duplikat data, outlier data dan data imbalance, selanjutnya akan dilakukan penanganan seperti data cleaning, data integrasi, data transformation, dan data reduksi untuk memperbaiki data error tersebut. Data cleaning, data integrasi, data transformation dan data reduksi akan dijelaskan pada praktikum selanjutnya.