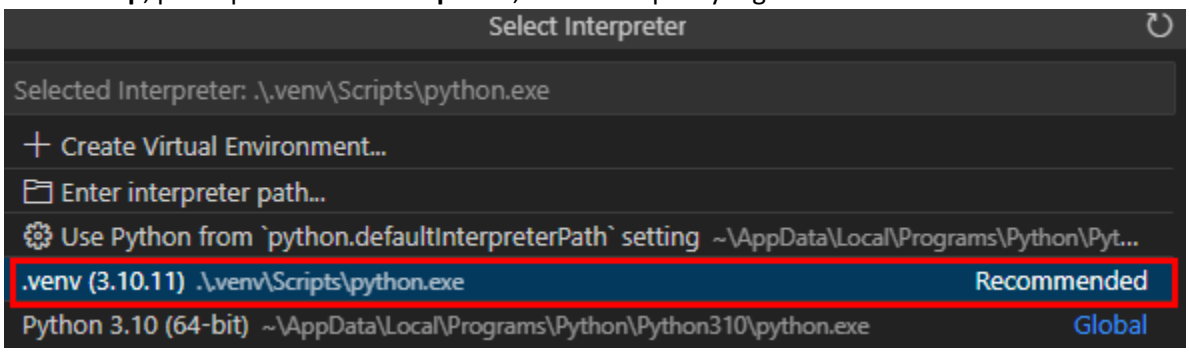
DOKUMENTASI MACHINE LEARNING PERTEMUAN 4

Muhammad Ridho Irfani|231011402243

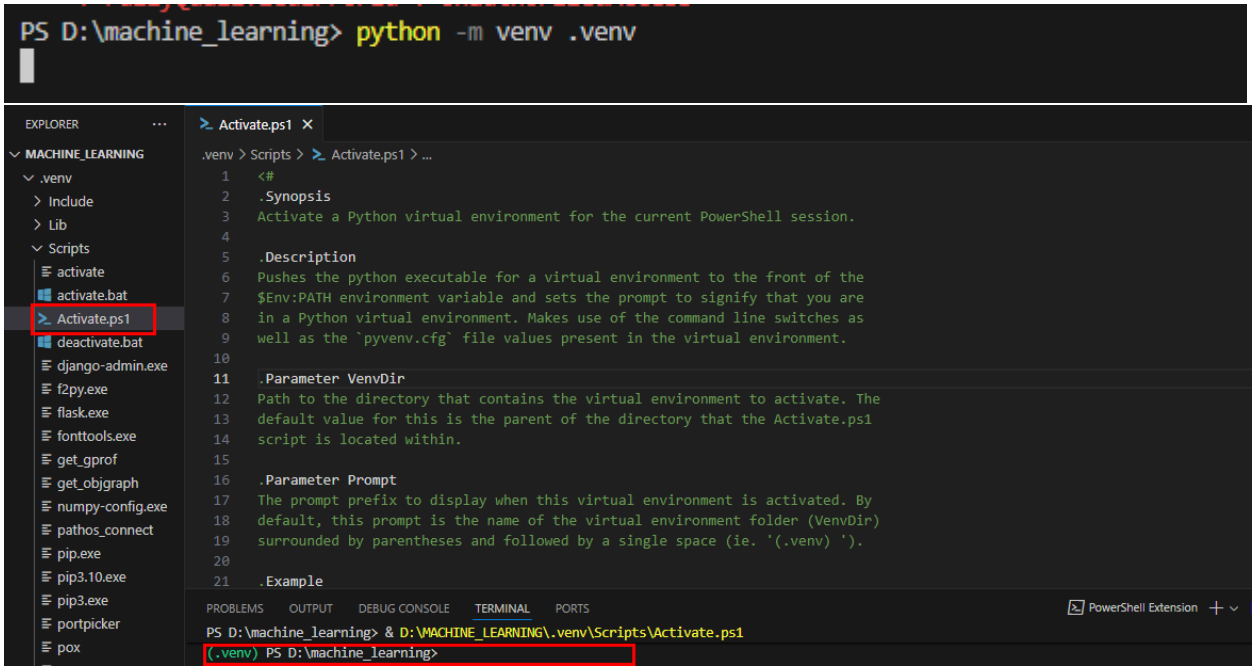
05TPLE017

Jadi di pertemuan4inidengan materi **Data Preparation.** Dengan tujuan membersihkan, memvisualisasikan, dan menyiapkan data sebelum modeling. Untuk langkah-langkahnya sebagai berikut;

1. Untuk langkah awal yang kita butuhkan yaitu aplikasi Visual Studio Code, aplikasi python (saya pakai versi 3.10.11), folder baru yang nantinya akan digunakan di dalam vs code, extention powershell, python, jupyter untuk kebutuhan coding.
2. Buka folder di vscode, di saya pakai **"MACHINE_LEARNING",** kemudian pencet tombol *shortcut* **ctrl+shift+p**, pilih opsi **selected interpreter**, kemudian pilih yang venv.



3. Kemudian masuk ke folder machine learning dan install venv dengan cara ketik **python –m venv .venv**, setelahnya proses akan jalan. Jika sudah, langsung masuk ke folder venv tersebut.



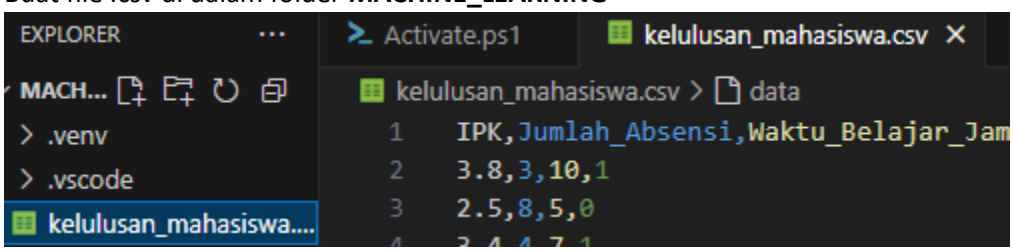
4. Lalu install kebutuhan berikut

Numpy, pandas, matplotlib, skicit-learn, seaborn, tk, pyqt5, pygame, flask, Django, tensorflow

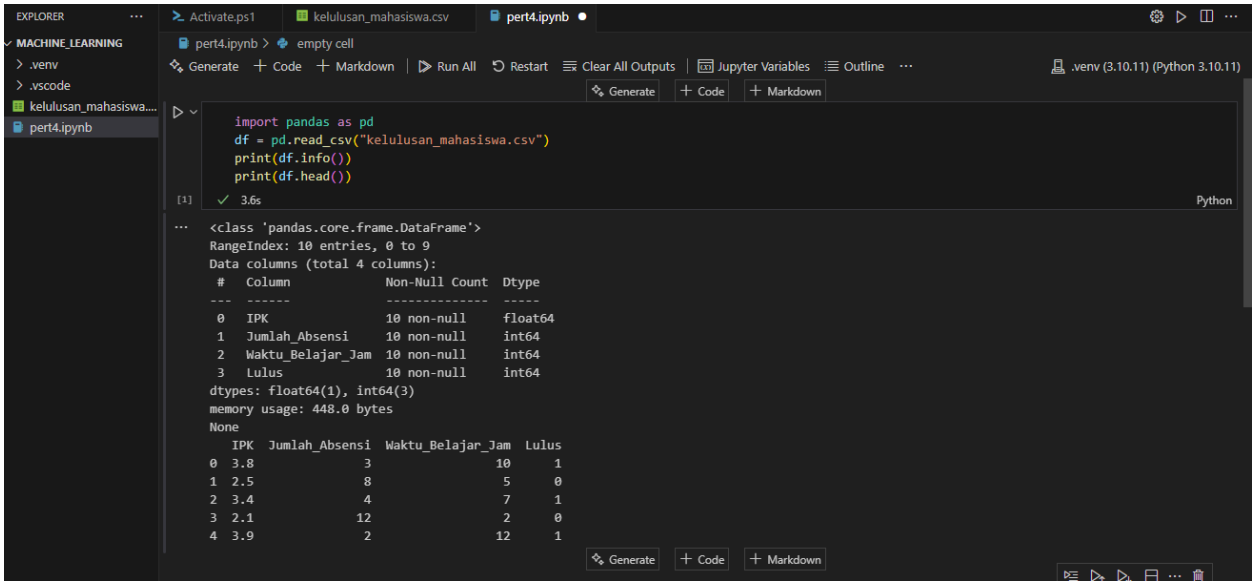
Dengan cara **pip install**

```
(.venv) PS D:\machine_learning> pip install numpy pandas matplotlib scikit-learn seaborn tk pyqt5 pygame flask Django tensorflow
Requirement already satisfied: numpy in d:\machine_learning\.venv\lib\site-packages (2.2.6)
Requirement already satisfied: pandas in d:\machine_learning\.venv\lib\site-packages (2.3.3)
Requirement already satisfied: matplotlib in d:\machine_learning\.venv\lib\site-packages (3.10.6)
Requirement already satisfied: scikit-learn in d:\machine_learning\.venv\lib\site-packages (1.7.2)
Requirement already satisfied: seaborn in d:\machine_learning\.venv\lib\site-packages (0.13.2)
Collecting tk
  Downloading tk-0.1.0-py3-none-any.whl (3.9 kB)
Collecting pyqt5
  Downloading PyQt5-5.15.11-cp38-abi3-win_amd64.whl (6.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/6.9 MB 4.8 MB/s eta 0:00:02
```

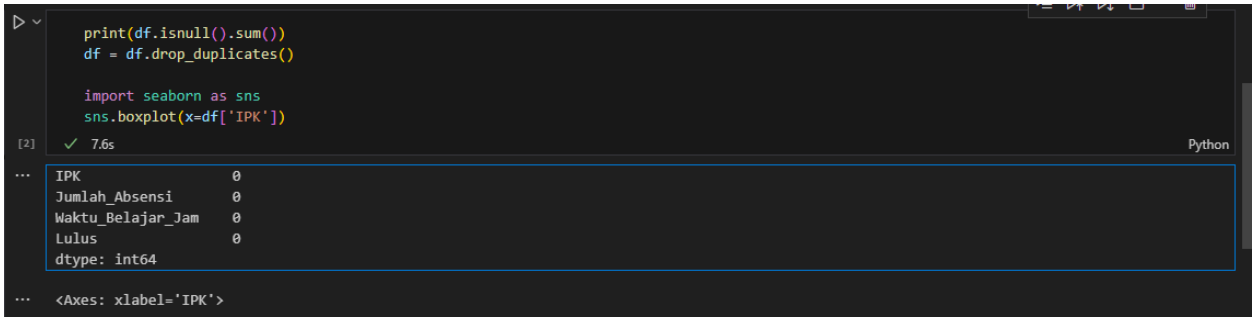
5. Setelah semua kebutuhan awal selesai, selanjutkan fokus mengerjakan sesuai modul pertemuan 4. Buat file .csv di dalam folder **MACHINE_LEARNING**

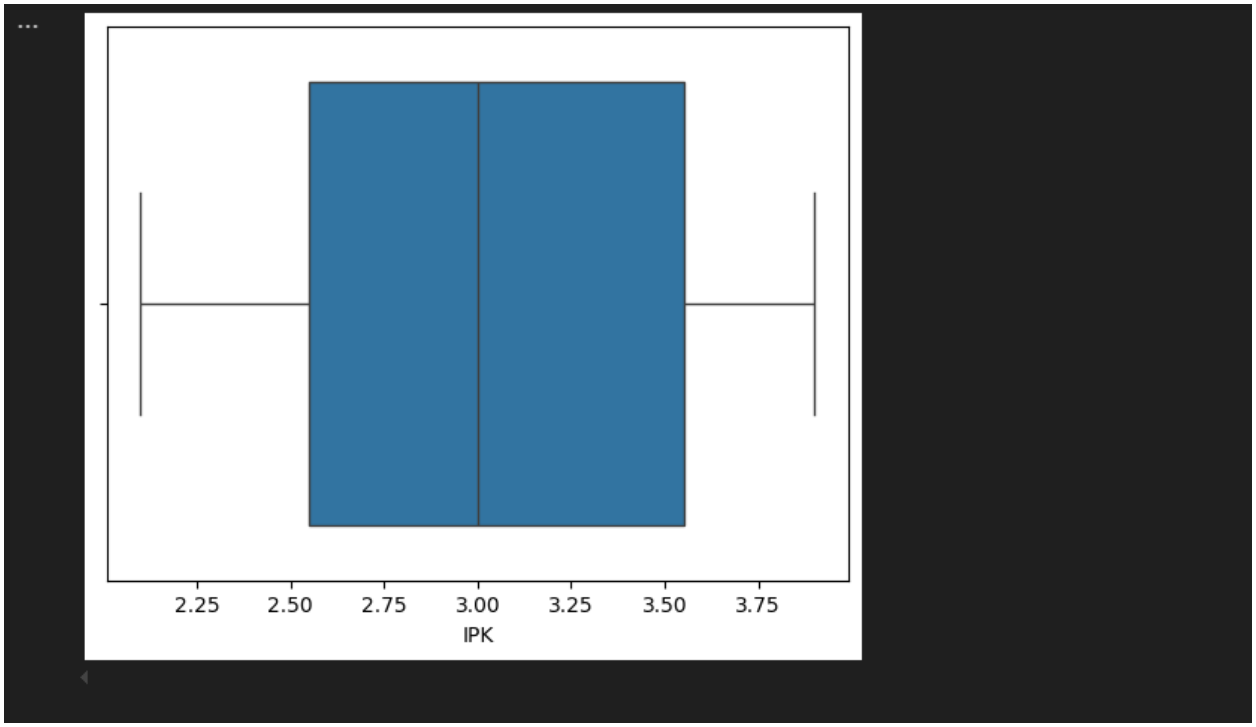


6. Collection

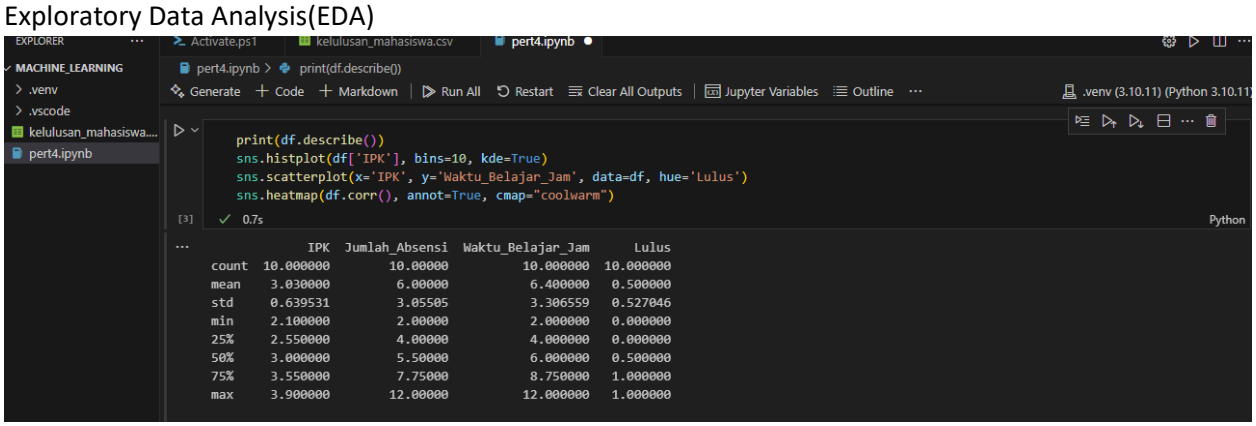


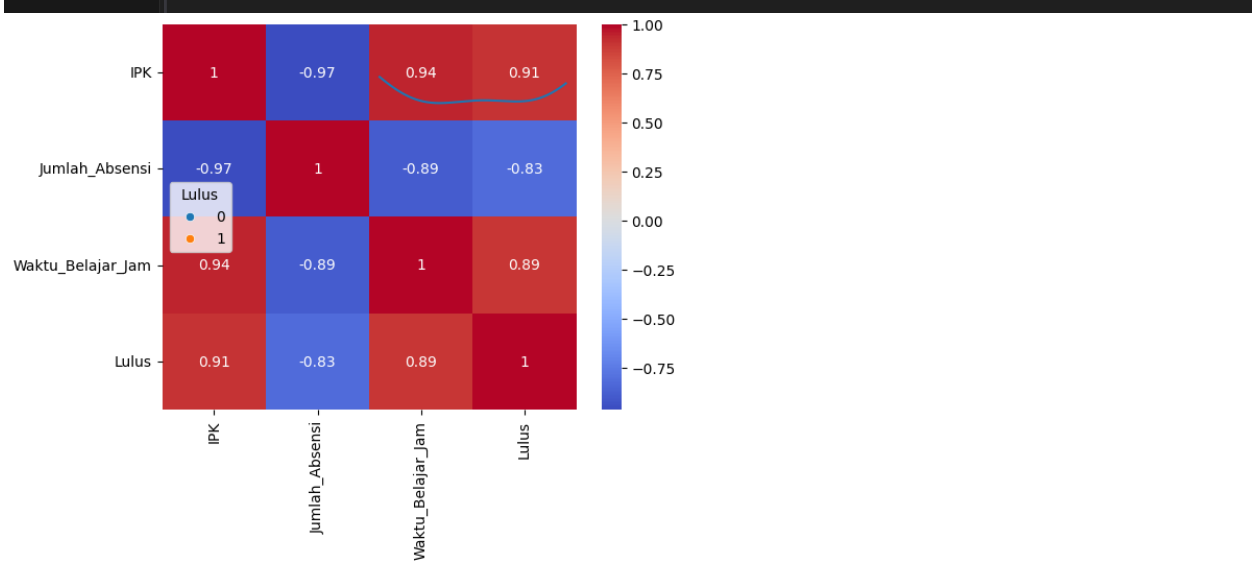
7. Cleaning





8. Exploratory Data Analysis(EDA)





9. Featuring Engineering. **Tambahkan df.head()** di bagian bawah codingan, untuk mengecek hasil

```python
df['Rasio_Absensi'] = df['Jumlah_Absensi'] / 14
df['IPK_x_Study'] = df['IPK'] * df['Waktu_Belajar_Jam']
df.to_csv("processed_kelulusan.csv", index=False)
df.head()
```

[6] ✓ 0.0s Python

	IPK	Jumlah_Absensi	Waktu_Belajar_Jam	Lulus	Rasio_Absensi	IPK_x_Study
0	3.8	3	10	1	0.214286	38.0
1	2.5	8	5	0	0.571429	12.5
2	3.4	4	7	1	0.285714	23.8
3	2.1	12	2	0	0.857143	4.2
4	3.9	2	12	1	0.142857	46.8

10. Splitting dataset. Di langkah ini saya menemukan error seperti berikut

```
    7     X, y, test_size=0.3, stratify=y, random_state=42)
----> 9 X_val, X_test, y_val, y_test = train_test_split(
     10     X_temp, y_temp, test_size=0.5, stratify=y_temp, random_state=42)
     12 print(X_train.shape, X_val.shape, X_test.shape)

File d:\MACHINE_LEARNING\.venv\lib\site-packages\sklearn\utils\_param_validation.py:218, in validate_params.<locals>.decorator.<locals>.wrap
    212 try:
    213     with config_context(
    214         skip_parameter_validation=(
    215             prefer_skip_nested_validation or global_skip_validation
    216         )
    217     ):
--> 218         return func(*args, **kwargs)
    219 except InvalidParameterError as e:
    220     # When the function is just a wrapper around an estimator, we allow
    221     # the function to delegate validation to the estimator, but we replace
    222     # the name of the estimator by the name of the function in the error
    223     # message to avoid confusion.
    224     msg = re.sub(
    225         r"parameter of \w+ must be",
...
   2351         "The train_size = %d should be greater or "
   2352         "equal to the number of classes = %d" % (n_train, n_classes)
   2353     )

ValueError: The least populated class in y has only 1 member, which is too few. The minimum number of groups for any class cannot be less th
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Dan setelah saya cari tahu, ternyata error tersebut dikarenakan dataset hanya memiliki 10 baris data, saat dilakukan pembagian 70:30 dan kemudian 50:50 untuk validation serta test, salah satu kelas hanya memiliki 1 sampel. hal ini menyebabkan error saat menggunakan stratify pada train_test_split. untuk mengatasinya, stratify dihapus agar proses splitting tetap dapat berjalan meskipun dataset menjadi tidak seimbang.

11. Hasil Akhir nya seperti berikut

```python
from sklearn.model_selection import train_test_split

X = df.drop('Lulus', axis=1)
y = df['Lulus']

# split pertama: train (70%) dan sisanya (30%)
X_train, X_temp, y_train, y_temp = train_test_split(
    X, y, test_size=0.3, random_state=42)

# split kedua: dari 30% sisa dibagi jadi validation (15%) dan test (15%)
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, random_state=42)

print(X_train.shape, X_val.shape, X_test.shape)
```

[10] ✓ 0.0s Python

```
(7, 5) (1, 5) (2, 5)
```