MLflow 실습 3

1. MLflow 를 사용한 서빙 Example

1. MLflow 를 사용한 서빙 Example

- MLflow 를 사용하여 간단하게 서빙도 할 수 있습니다.
  - https://mlflow.org/docs/latest/tutorials-and-examples/tutorial.html

```
mlflow models serve -m $(pwd)/mlruns/0/<run-id>/artifacts/model -p <port>

mlflow models serve -m $(pwd)/mlruns/0/63d1a9cde7f84190a5634648467be195/artifacts/model -p 1234
```

  - 원하는 모델의 run id 를 확인한 다음, port 를 지정하여 `mlflow models serve` 명령을 수행합니다.
    - 모델 서빙이라는 의미는 쉽게 말하면 127.0.0.1:1234 에서 REST API 형태로 `.predict()` 함수를 사용할 수 있는 것을 의미합니다.

- 이제 해당 서버에 API 를 보내서, `predict()` 의 결과를 확인해보겠습니다.

- API 를 보내기 위해서는, request body 에 포함될 data 의 형식을 알고 있어야 합니다.
  - diabetes data 의 column 과 sample data 를 확인해보겠습니다.
    - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html

  미리 확인한 결과

```
data = load_diabetes()
print(data.feature_names)

df = pd.DataFrame(data.data)
print(df.head())
```

```
print(data.target[0])
```

```
['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']
          0         1         2         3         4         5         6  \
0  0.038076  0.050680  0.061696  0.021872 -0.044223 -0.034821 -0.043401
1 -0.001882 -0.044642 -0.051474 -0.026328 -0.008449 -0.019163  0.074412
2  0.085299  0.050680  0.044451 -0.005671 -0.045599 -0.034194 -0.032356
3 -0.089063 -0.044642 -0.011595 -0.036656  0.012191  0.024991 -0.036038
4  0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142

          7         8         9
0 -0.002592  0.019908 -0.017646
1 -0.039493 -0.068330 -0.092204
2 -0.002592  0.002864 -0.025930
3  0.034309  0.022692 -0.009362
4 -0.002592 -0.031991 -0.046641

151.0
```

- 127.0.0.1:1234 서버에서 제공하는 `POST /invocations` API 요청을 수행해보겠습니다.

```
curl -X POST -H "Content-Type:application/json" --data '{"columns":["age", "sex",
 "bmi", "bp", "s1", "s2", "s3", "s4", "s5", "s6"],"data":[[0.038076, 0.050680,  0.
061696,  0.021872, -0.044223, -0.034821, -0.043401, -0.002592,  0.019908, -0.01764
6]]}' http://127.0.0.1:1234/invocations
```

  - prediction value 가 API 의 response로 반환 되는 것을 확인할 수 있습니다.

- 정해진 Data size 와 다르게 `POST /invocations` API 요청을 수행해보겠습니다.
  (10개 컬럼이 필요한데 11개로 변경)

```
curl -X POST -H "Content-Type:application/json" --data '{"columns":["Age", "Sex", "Bod
y mass index", "Average blood pressure", "S1", "S2", "S3", "S4", "S5", "S6", "S7"],"da
ta":[[0.038076, 0.050680,  0.061696,  0.021872, -0.044223, -0.034821, -0.043401, -0.00
2592,  0.019908, -0.017646]]}' http://127.0.0.1:1234/invocations
```

- data size 가 predict 하기에는 안 맞는다는 에러가 반환되는 것을 확인할 수 있습니다.