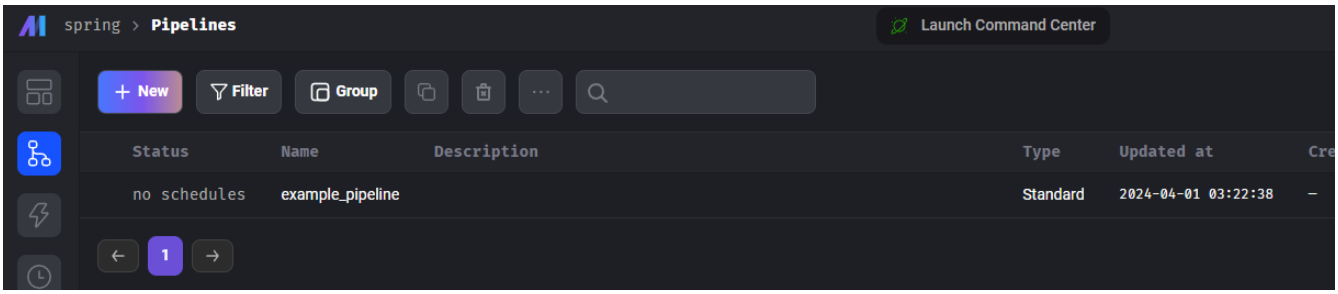
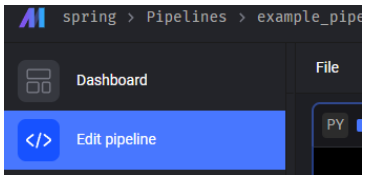
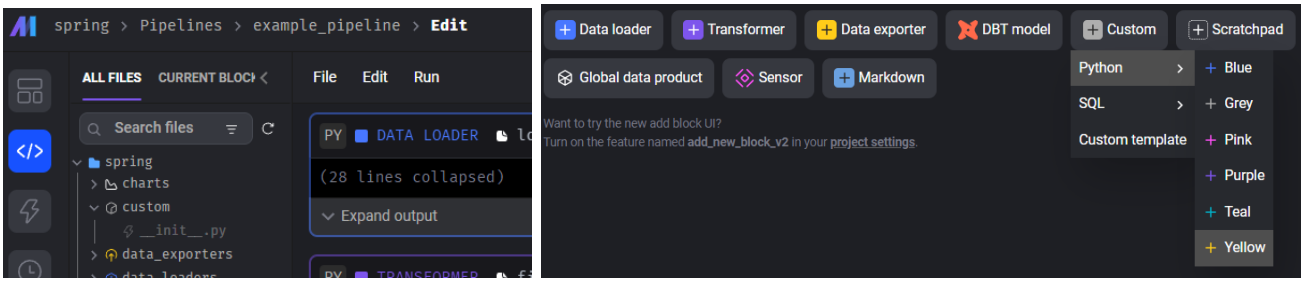
Machine Learning with mage.ai

(파이썬 파일 편집은 vscode 에서 직접 수정하는 것을 권장)

1. (mage) mage start spring
2. 

3. Example_pipline 클릭 >> 

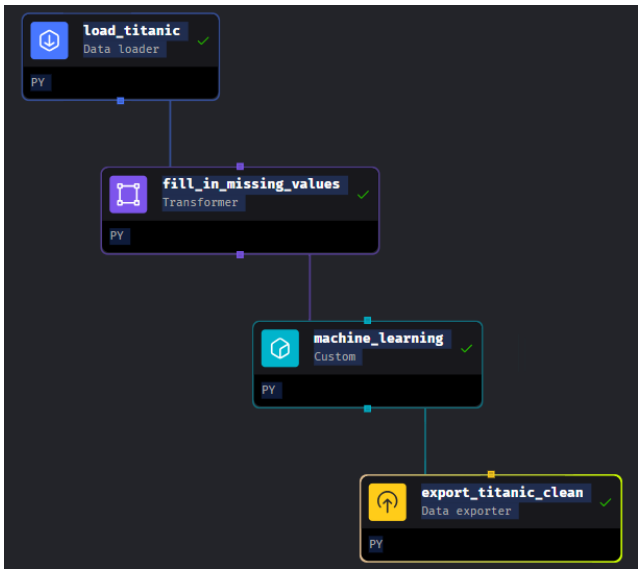
4. 블록을 확인하고 단계적으로 실행해본다.
5. Custom 으로 machine learning block 을 만들고 up/down stream 을 구성(선 연결하기)
6. 

7. 파이썬코드는 아래와 같이 만들고 저장한다. (vs code 에서 직접 작성해도 됨)

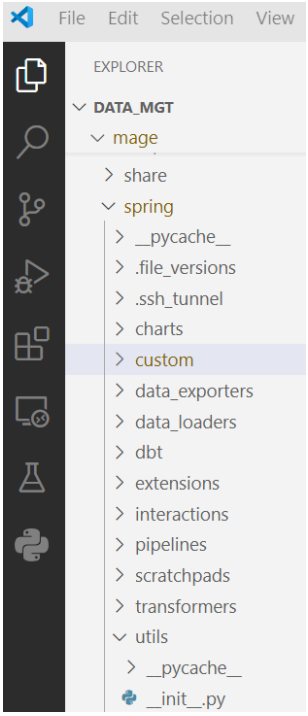
```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from spring.utils.variables import (
    X_COLS,
    Y_COLS,
)

if "custom" not in globals():
    from mage_ai.data_preparation.decorators import custom
if "test" not in globals():
    from mage_ai.data_preparation.decorators import test

def _model_save(rf_model):
    """
    Save RandomForest model to a file.
    """
    # Here you would implement the logic to save your trained model to a file
    # Example:
    # with open('random_forest_model.pkl', 'wb') as file:
    #     pickle.dump(rf_model, file)
```

```
    pass

@custom
def random_forest_train(df: pd.DataFrame, *args, **kwargs):
    """
    Train a Random Forest Classifier and predict the 'Survived' column.

    Args:
        df: Data frame containing the training data.

    Returns:
        Data frame with a new column 'Survived_predict' with predictions.
    """
    # Prepare the data
    x_train = df[X_COLS]
    y_train = df[Y_COLS].values.ravel() # RandomForest expects a 1D array for y

    # Initialize the Random Forest Classifier
    rf_model = RandomForestClassifier()

    # Train the model
    rf_model.fit(x_train, y_train)

    # Predict using the trained model
    _pred = rf_model.predict(x_train)

    # Optionally save the model
    _model_save(rf_model)

    # Assign predictions to a new column in the dataframe
    df = df.assign(Survived_predict=_pred)
    return df

@test
def test_output(output, *args) -> None:
    """
    Template code for testing the output of the block.

    Args:
        output: The output from the random_forest_train function.
    """
    assert output is not None, "The output is undefined"
    assert 'Survived_predict' in output.columns, "Prediction column is missing in the output dataframe"
    # You can add more tests to check the quality of your predictions,
    # such as accuracy score, confusion matrix, etc.
```

8. Pipeline 구성



9. C:\data_mgt\mage\spring\utils\variables.py 파일을 만든다.
10.



11.
mage > ml_project > utils > variables.py > ...

```python
TITANIC_DATA_URL = (
    "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"
)
X_COLS = [
    "Age",
    "Fare",
    "Parch",
    "Pclass",
    "SibSp",
]
Y_COLS = [
    "Survived",
]
COLS = X_COLS + Y_COLS
```

12. (mage) C:\data_mgt\mage>mage start spring 재실행해서 block 을 모두 실행한다.
13. ‘spring’ 를 다시 실행하려고 하며 반드시 폴더 위치를 지켜라 아래와 같이 다른 폴더에서 시작하면 새로운 프로젝트가 생성된다.(주의)

    (mage) C:\data_mgt\mage>mage start spring
    # mage 폴더 밑에 spring 폴더가 존재

참고자료

1. 모델을 저장하고 싶으면 model_persistence
2. 저장할 파일의 폴더는 만들고 경로를 만드는 방법 파이썬 폴더/경로 생성
   - Import os
   - os.mkdirs(“a/b/c”. exit_ok = True)
3. 오픈소스 데이터 파이프라인 툴 mage.ai (https://github.com/RektPunk/mage-ai-example)