붓꽃 종 분류

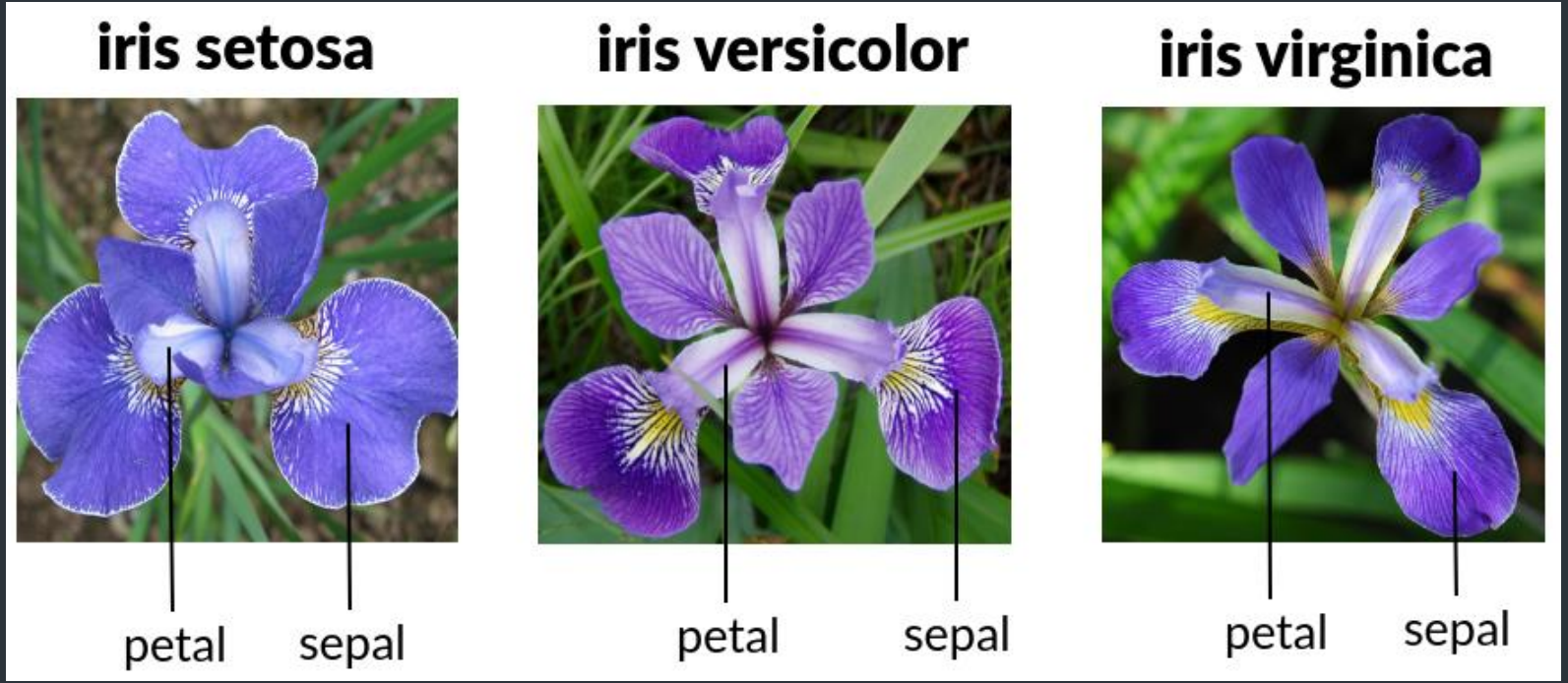
Azure Machine Learning Studio

학습 목표

- 붓꽃의 외관적 특징을 **지표**로 삼고, 종을 분류하는데 있어서 각각의 지표들의 **가중치**를 알아내며, 이를 기반으로 종을 분류할 수 있는 **학습 모델**을 만든다.
- **로지스틱 회귀 분석**에 대한 이론과 관련 모듈에 대한 지식을 습득한다.
- **예측 모델**을 작성해서 웹에 배포한 후, 이를 토대로 사용자가 직접 입력한 데이터에 대한 예측 결과를 확인한다.

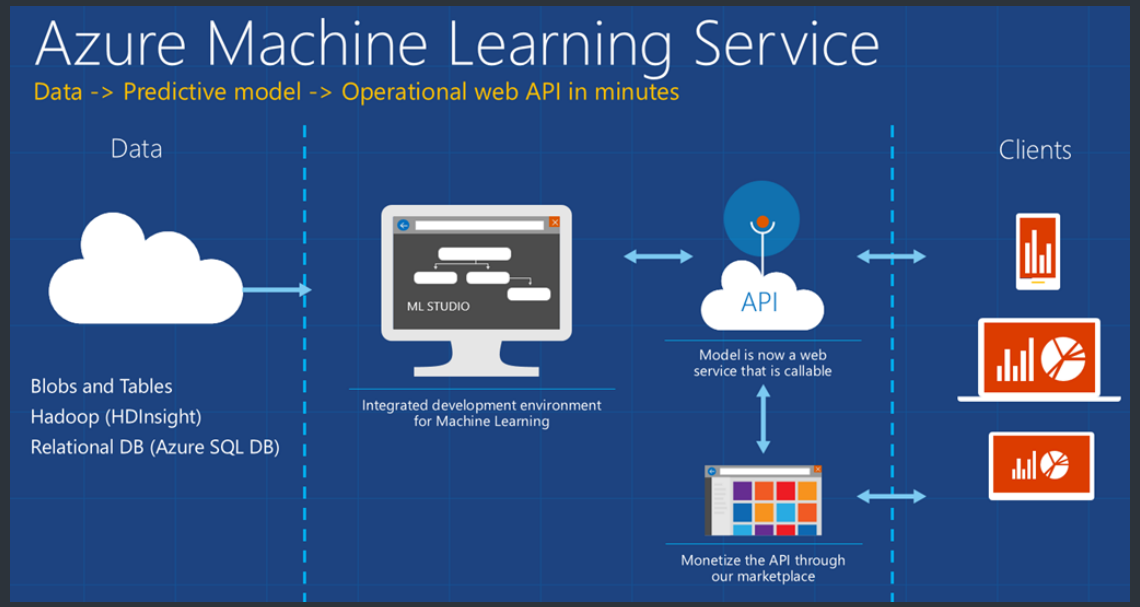
분석할 데이터

- 붓꽃 분류는 기계 학습 분야에서 가장 유명한 예제이다.
- 꽃받침의 폭과 길이, 꽃잎의 폭과 길이, 총 4개의 지표로 종을 분류한다.
- 이 학습에선 5개 features, 150개 데이터로 이루어진 데이터셋을 사용한다.



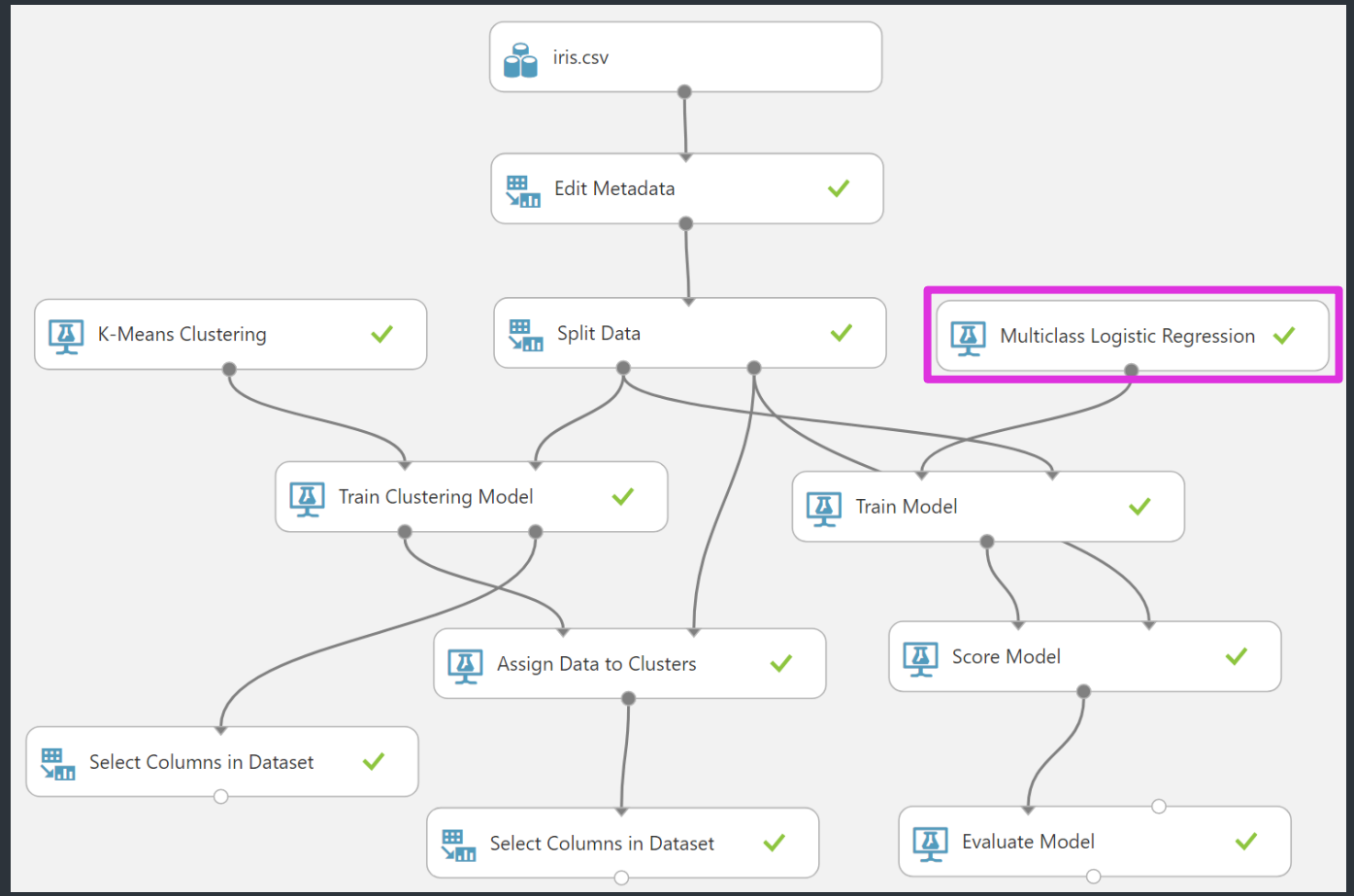
학습 절차

- 데이터 취득(Data Acquisition)
- 데이터 준비(Data Preparation)
- 모델 학습 및 평가(Model Training and Evaluation)
- 웹 배포(Web Deployment)



신규 모듈

Multiclass Logistic Regression



범주형 데이터를 대상으로 하는 회귀 분석

데이터 취득

학습에 사용할 데이터셋을 추가하는 절차

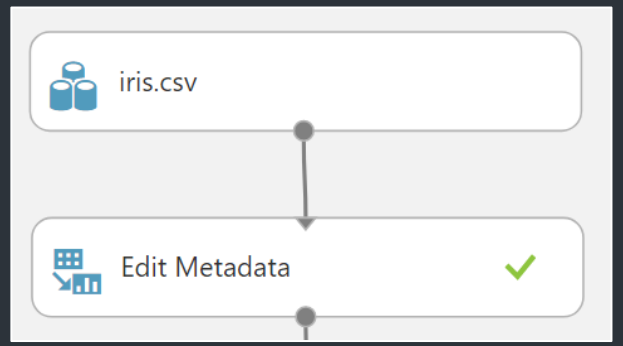
1. **Kaggle**에서 Iris Flower Dataset을 다운로드한다.
2. CSV 파일을 업로드하고 새 실험을 생성한다.
3. 업로드한 CSV 파일을 모듈로 추가한다.
4. 데이터를 확인한다.

데이터 준비

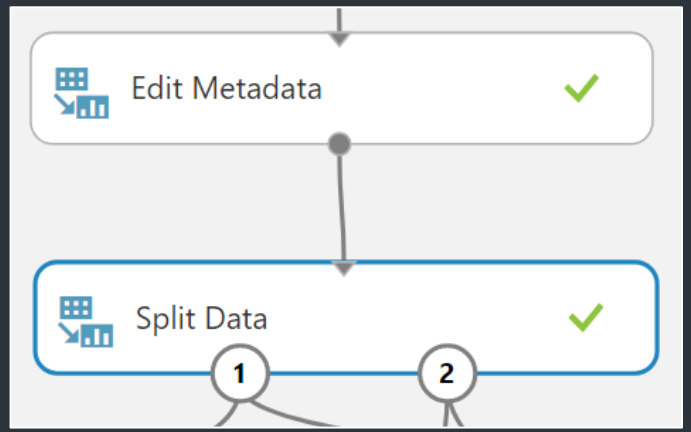
취득한 데이터를 학습에 용이하게 설정하는 절차

1. **Edit Metadata** 모듈을 추가하고 CSV 파일 모듈을 연결한다.
2. *Column selector*를 실행하고 all columns include all features 조건을 선택한다.
3. **Split Data** 모듈을 추가하고 **Edit Metadata** 모듈을 연결한다.
4. *Fraction*을 0.7로 설정해서 데이터셋을 학습 세트와 시험 세트로 나눈다.

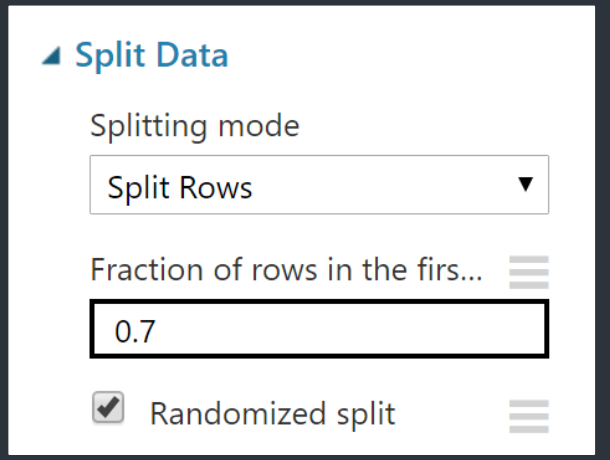
①



③



④

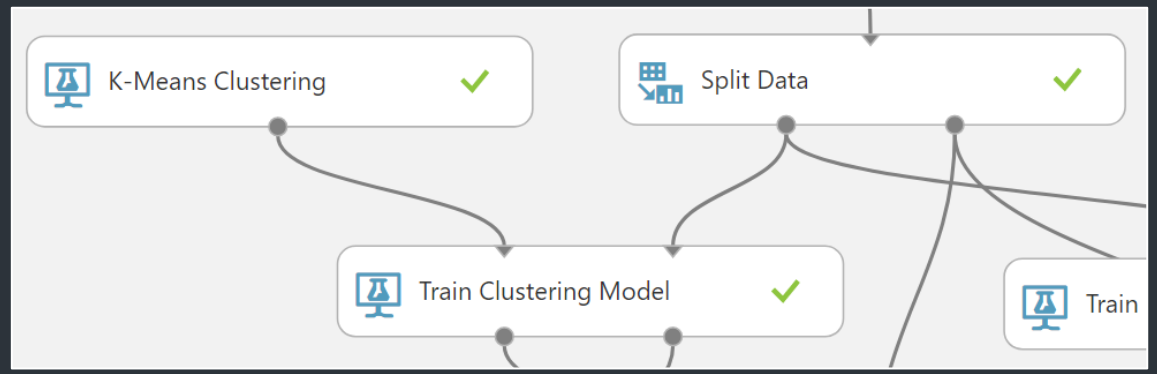


학습 및 평가: 군집화 - 1

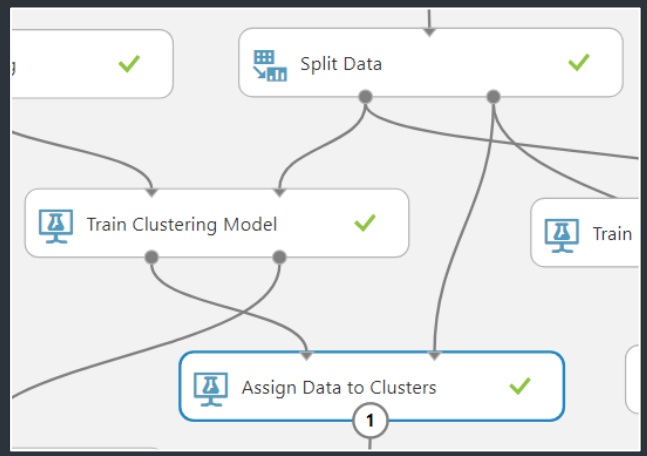
학습을 진행하여 모델을 생성하고 평가하는 절차

1. **K-Means Clustering** 모듈을 추가하고 *Number of Centroids*를 3으로 설정한다.
2. **Train Clustering Model** 모듈을 추가하고 알고리즘 모듈과 학습 세트를 연결한다.
3. *Column selector*를 실행하고 species 열을 제외한 모든 열을 선택한다.
4. *RUN*을 실행하고 학습 결과를 확인한다.
5. **Assign Data to Clusters** 모듈을 추가하고 학습된 모델과 시험 세트를 연결한다.
6. *RUN*을 실행하고 시험 결과를 확인한다.

①
②



⑤

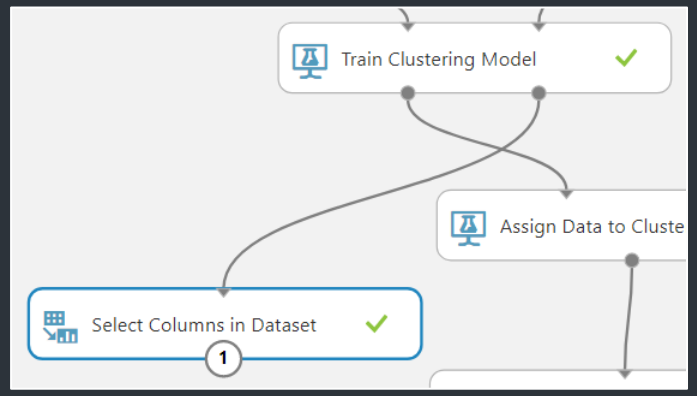


학습 및 평가: 군집화 - 2

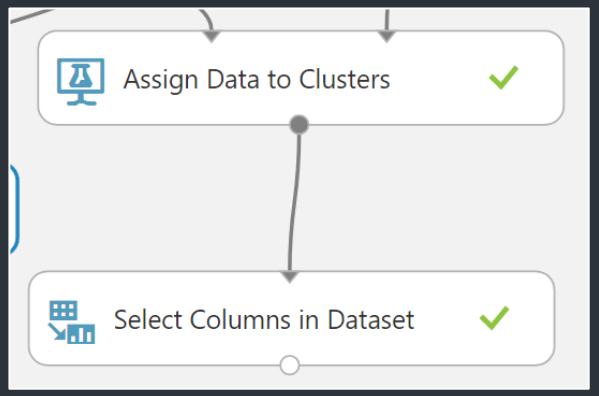
학습을 진행하여 모델을 생성하고 평가하는 절차

1. **Select Columns in Dataset** 모듈을 추가하고 학습 모듈의 결과 세트를 연결한다.
2. *Column selector*를 실행하고 species, Assignments 열을 선택한다.
3. *RUN*을 실행하고 학습 세트의 군집 할당 결과를 확인한다.
4. **Select Columns in Dataset** 모듈을 추가하고 할당 모듈의 결과 세트를 연결한다.
5. *Column selector*를 실행하고 species, Assignments 열을 선택한다.
6. *RUN*을 실행하고 시험 세트의 군집 할당 결과를 확인한다.

①



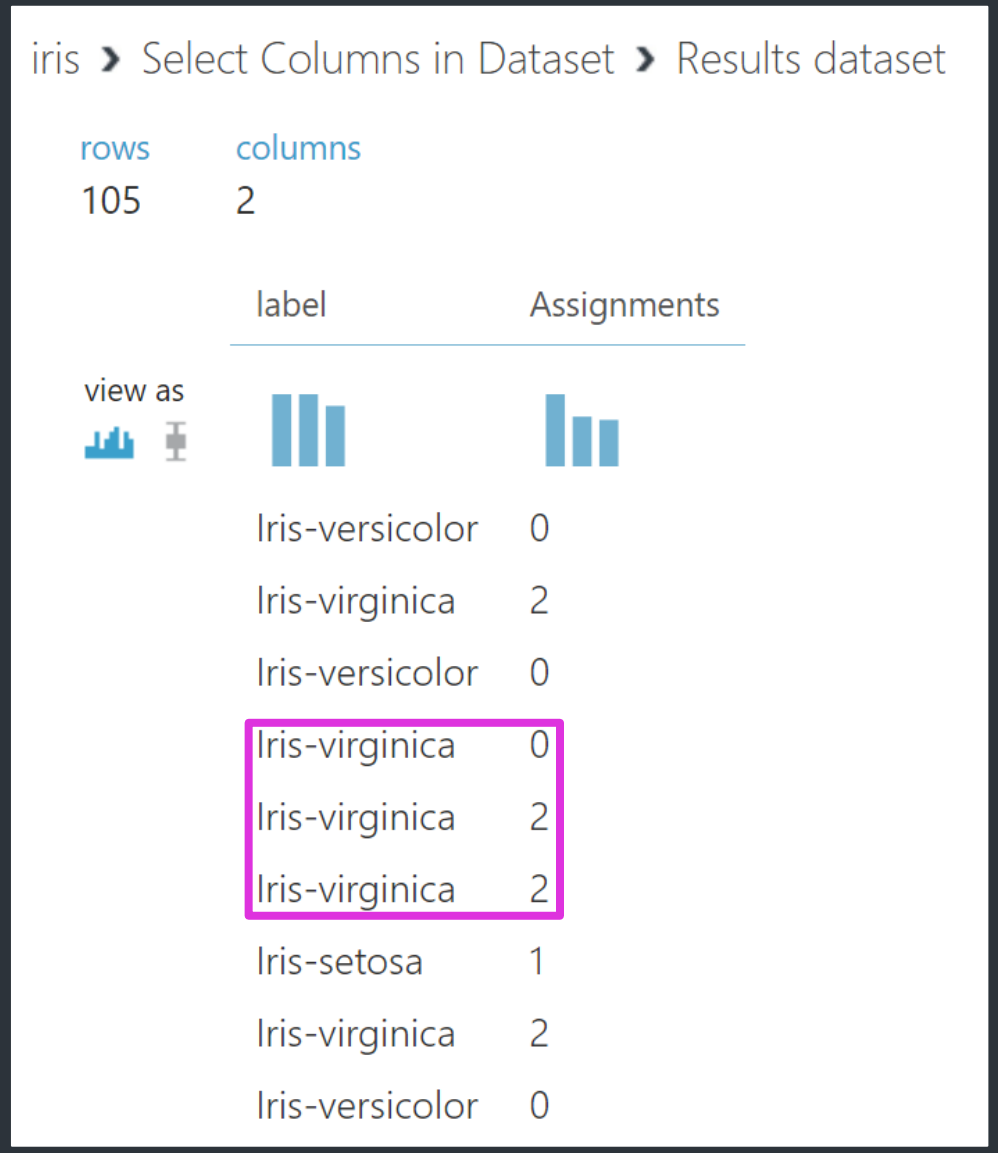
④



군집 할당은 분류가 아니다.

MS Azure ML Studio에서 군집화 학습 결과는 주성분 분석법(PCA)으로 표시된다. 여기에 Select Columns in Dataset 모듈을 통해 각각의 데이터가 어떤 군집에 할당(Assignments)되었는지 확인할 수 있는데, 이는 특정 지표(ex. species)에 따라 분류한 결과를 나타내는 것이 아니다. **군집 할당 결과는 각각의 데이터들 중 가장 연관성이 큰 데이터들끼리 묶은 결과**일 뿐이다.

PCA의 원리에 따라, 학습에 사용된 지표들을 가장 큰 영향력이 있는 두 개의 차원으로 환원하고, 이를 통해 구성된 2차원 평면에 군집 학습 결과를 표시해준다. 군집 학습 결과가 PCA의 원리를 따르기에, **결과로써 표현된 지표들은 원래 지표들과는 연관이 있었지만 결국은 완전히 다른 새로운 지표**가 되는 것이며 이는 분류와는 상관이 없다.



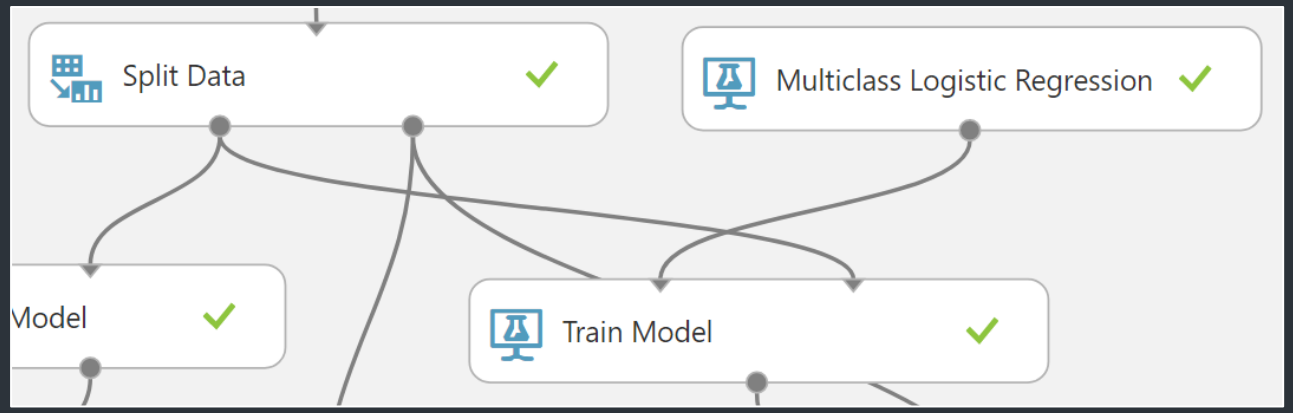
같은 종이지만 **다른 군집**으로 할당된 결과

학습 및 평가: 분류 - 1

학습을 진행하여 모델을 생성하고 평가하는 절차

1. **Multiclass Logistic Regression** 모듈을 추가한다.
2. **Train Model** 모듈을 추가하고 알고리즘 모듈과 학습 세트를 연결한다.
3. *Column selector*를 실행하고 species 열을 제외한 모든 열을 선택한다.
4. *RUN*을 실행하고 학습 결과를 확인한다.

①
②



다중 클래스 로지스틱 회귀 분석
Multiclass Logistic Regression

- 독립 변수의 선형 결합으로 종속 변수를 설명하는, 선형 회귀 분석의 일종.
- 종속 변수가 범주형 데이터를 대상으로 함.
- 입력에 대한 결과가 특정 분류로 나누어지므로 **분류 분석 알고리즘**으로 취급.
- 결과물이 두 종류로 분류되면 ***Two-Class*** *Logistic Regression*, 세 종류 이상으로 분류되면 ***Multiclass*** *Logistic Regression*으로 구분.

Feature Weights

Feature	Iris-setosa	Iris-versicolor	Iris-virginica
F4	-2.19621	0	2.3101
Bias	1.55612	0.742738	-2.29887
F3	-2.27011	0	1.5686
F2	0.894292	-0.720547	0
F1	-0.58357	0	0.679231

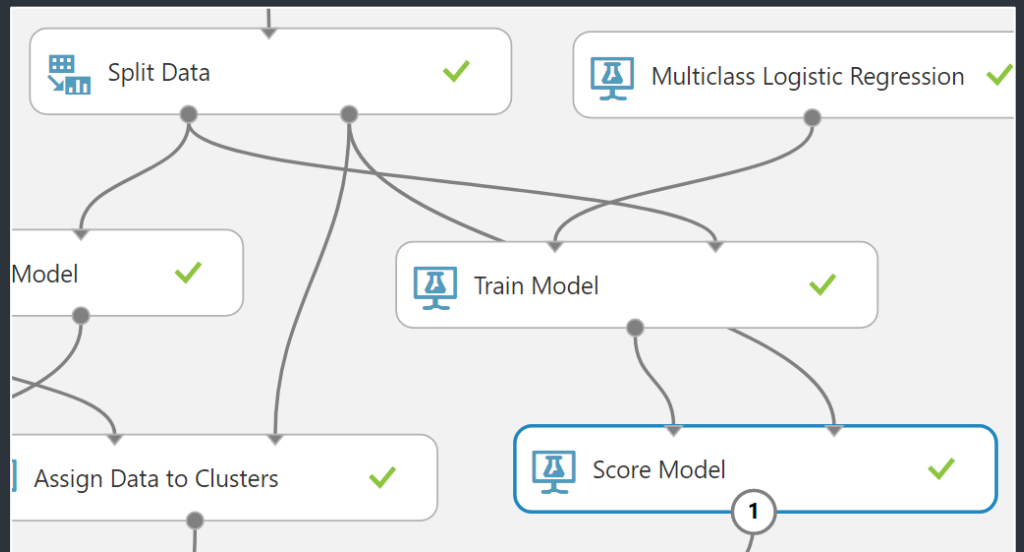
각 지표의 가중치를 구해 회귀선을 그리고, 입력 데이터를 대입하여 **분류**하는 게 목적

학습 및 평가: 분류 - 2

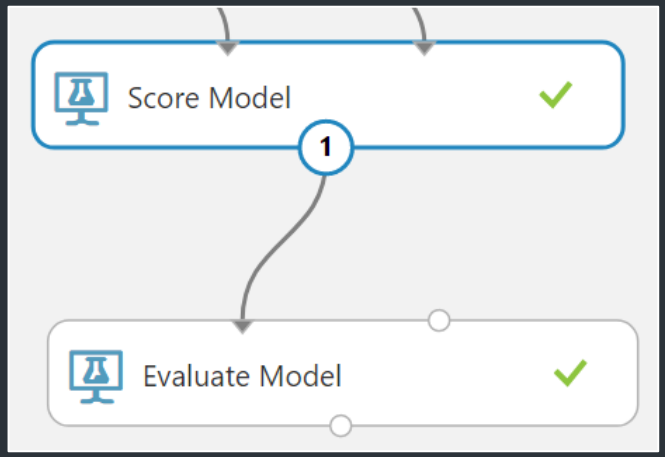
학습을 진행하여 모델을 생성하고 평가하는 절차

1. **Score Model** 모듈을 추가하고 학습된 모델과 시험 세트를 연결한다.
2. *RUN*을 실행하고 채점 결과를 확인한다.
3. **Evaluate Model** 모듈을 추가하고 채점 모듈의 결과 세트를 연결한다.
4. *RUN*을 실행하고 평가 결과를 확인한다.

①

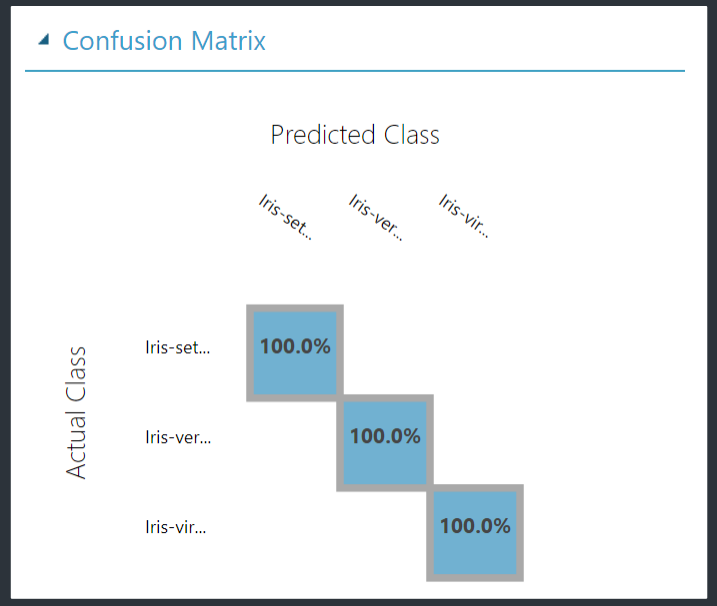


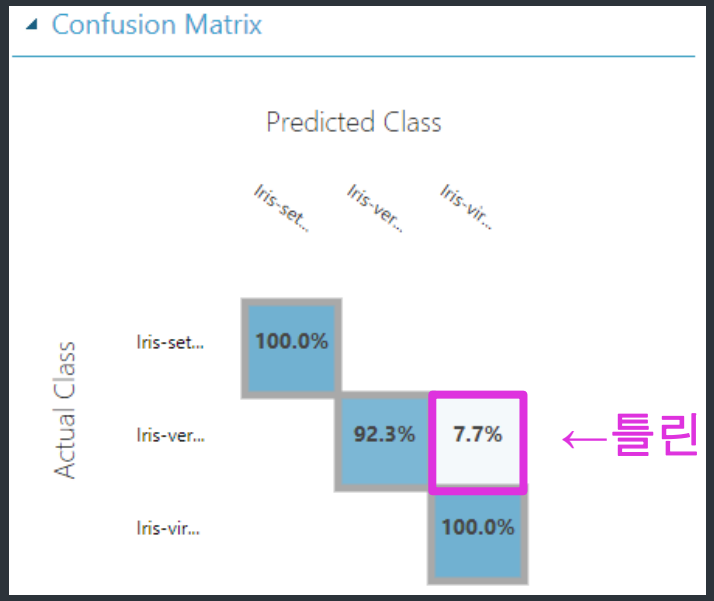
③



혼동 행렬
Confusion Matrix

- 학습 모델에 의한 예측 결과를 한 축, 실제 결과를 또 다른 한 축으로 하여 정답 확률을 2차원 평면의 형태로 표현한 것.
- 분류 분석에서 유용.





웹 배포

웹에 업로드하여 새로운 입력을 받고 예측할 수 있도록 만드는 절차

1. 학습 모델들 중에서 예측 모델 생성에 사용할 모듈을 하나 선택한다.
2. SET UP WEB SERVICE - Predictive Web Service 메뉴를 클릭한다.
3. RUN 실행 후 DEPLOY WEB SERVICE 버튼을 클릭한다.
4. Test 클릭 후 새로운 데이터를 입력하고 결과를 확인한다.