데이터 분석 실습

오렌지3를 통한 AI 챌린지 도전기

제1회 AI 챌린지 문제 2번 (난이도 상)

- **아마존 리뷰 기반 긍정 부정 리뷰 예측 문제**
  - 문제: https://www.kaggle.com/t/a0f4372db7554a64b62fc8c5e2f4fe37

- **문제**
  - 본 문제는 아마존 사용자 리뷰 데이터(1~5 점 평점)를 이용하여 사용자 리뷰의 긍정 리뷰와 부정 리뷰를 분류하는 자연어 처리 문제이다. train.csv 파일에서 라벨은 1, 0 으로 아마존 사용자 리뷰 데이터의 4, 5점은 라벨1로, 1, 2점은 라벨 0으로 3점 데이터는 사용하지 않는 것으로 전처리 하여 제공하였다. 또한 unbalance 데이터 셋에 대해서는 balance 하게 데이터를 구성하였는데, 원하는 학생들은 제공되는 raw data 를 다양한 방법으로 이용하여 성능을 향상시킬 수 있다. 최종적으로는 test.csv 파일 내의 리뷰 내용을 기반으로 리뷰 내용의 긍정/부정 여부를 예측하는 것을 목표로 한다.

제1회 AI 챌린지 문제 2번

- **평가 지표**
  - Category Accuracy 를 활용하여 예측 모델의 정확도를 측정한다.

- **데이터 파일 설명**
  - train.csv : 학습 데이터 (*주의: 데이터가 많음)
  - test.csv : 테스트 데이터 (*주의: 데이터가 많음)
  - submission_sample.csv : 결과 제출 템플릿

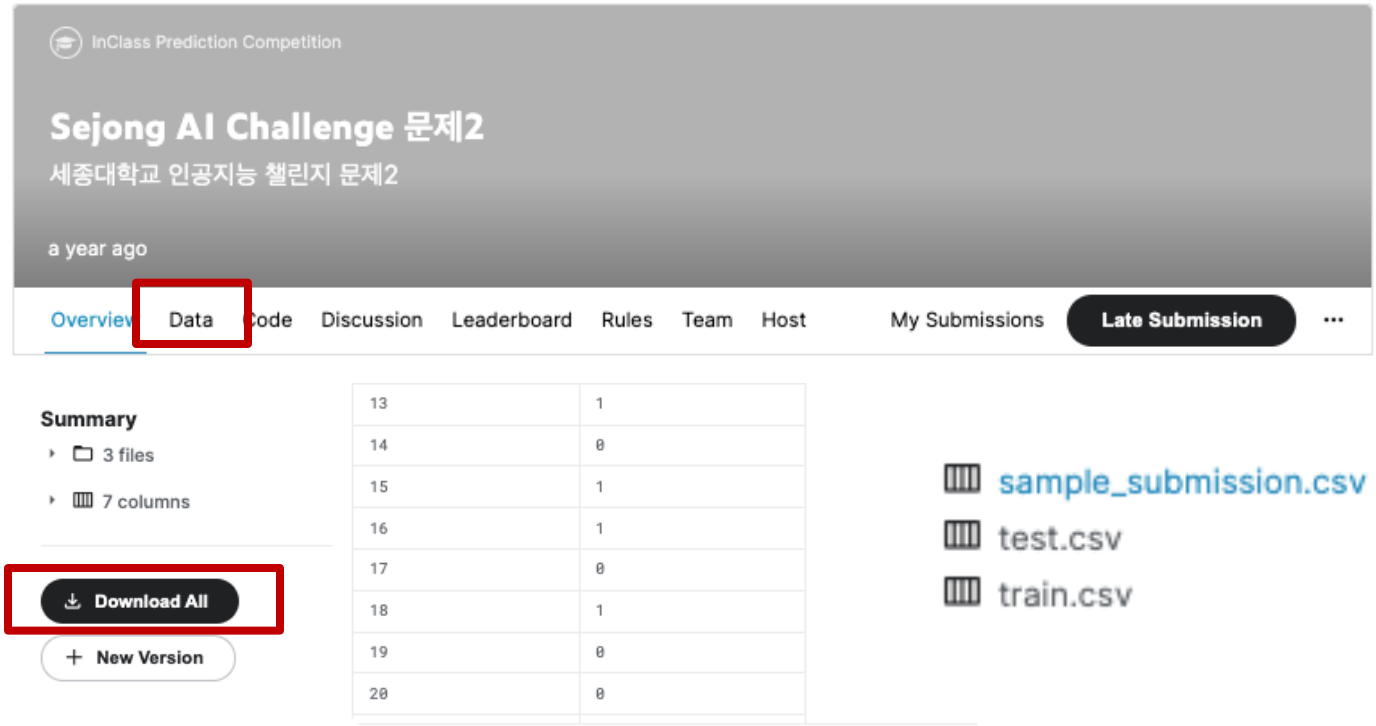
- **데이터 설명**
  - 각 데이터 샘플은 리뷰 텍스트 정보와 리뷰 점수가 아래와 같이 라벨링되어 있다.
    - id-데이터순번
    - Text – 사용자 리뷰
    - Label – 사용자 평점 (*대문자 임을 주의)

제1회 AI 챌린지 문제 2번

- **왜 난이도가 상인가?**
  - 기존 1, 3번 문제는 정형 데이터
    - 제공되는 데이터 자체가 1d 벡터로 볼 수 있음
  - 2번 문제는 비정형 데이터
    - 음성, 영상, 텍스트, 비디오 등
    - 〈Feature Extraction〉 과정을 통해서 1d 벡터를 얻을 수 있음
    - 〈Feature Extraction〉 방법론이 매우 다양하고 성능에 영향을 줌

제1회 AI 챌린지 문제 2번: 분류 문제

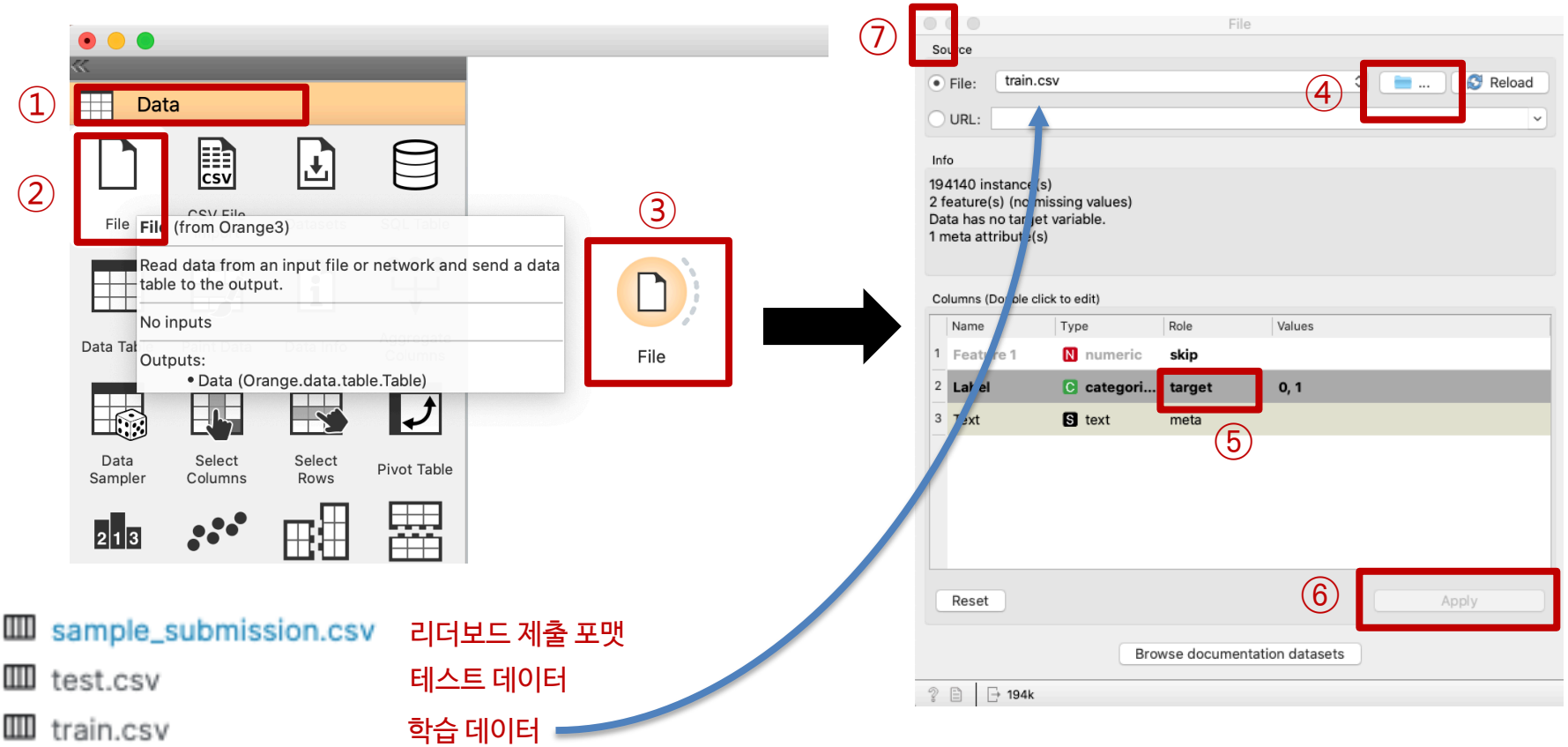
- 데이터 준비
  - 캐글 챌린지 〉〉 Data 탭 이동 〉〉 “Download All” 버튼 클릭 〉〉 압축 풀기



제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

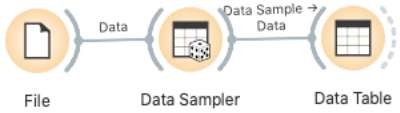
- 학습 데이터 열기
  - ① Data 탭 클릭 ›› ② File 드래그&드랍 ›› ③ File 위젯 더블클릭
  - ④ File 경로 선택 ›› ⑤ 데이터 Role 변경 ›› ⑥ Apply 버튼 클릭 ›› ⑦ Dialog 닫기

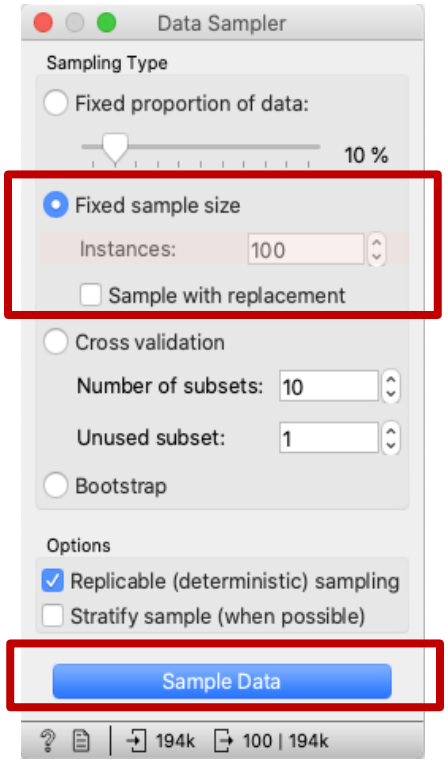


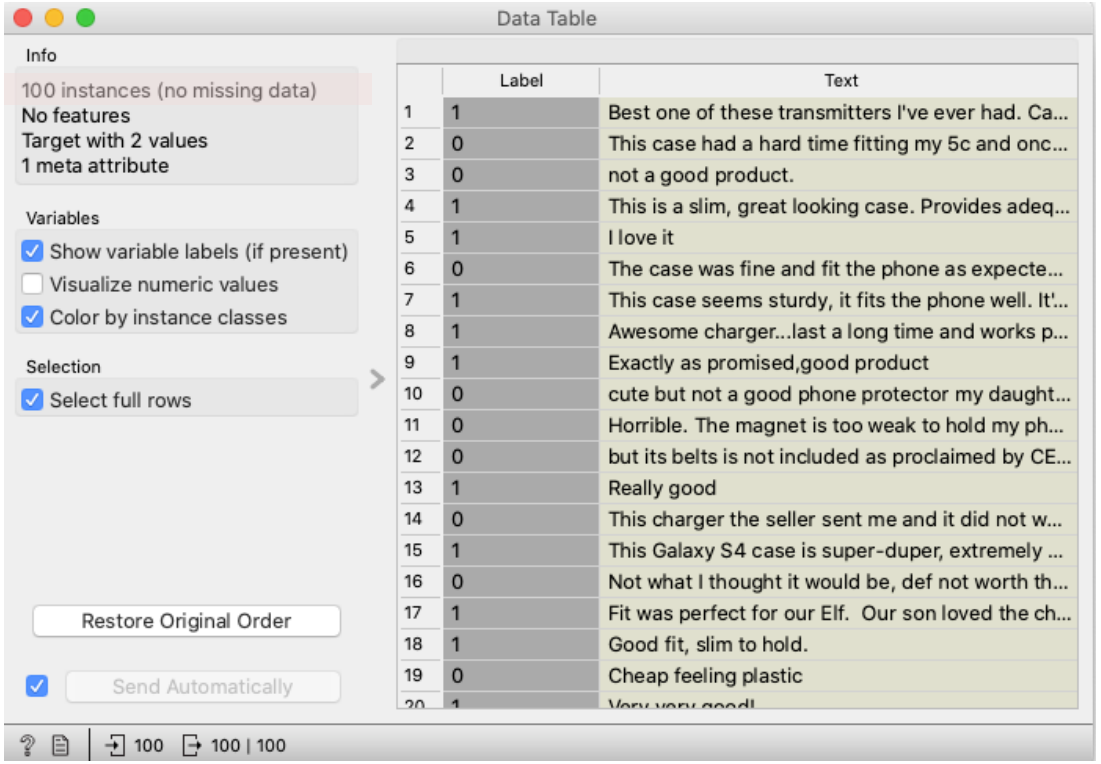
제1회 AI 챌린지 문제 2번: 분류 문제

① **학습 데이터 열기 (필수)**
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 학습 데이터 열기
  - Data Sampler 위젯 드래그&드랍 〉〉 Data Sampler 더블클릭 〉〉 위젯 연결
    - 빠른 모델 검증을 위해서 데이터 샘플링 사용



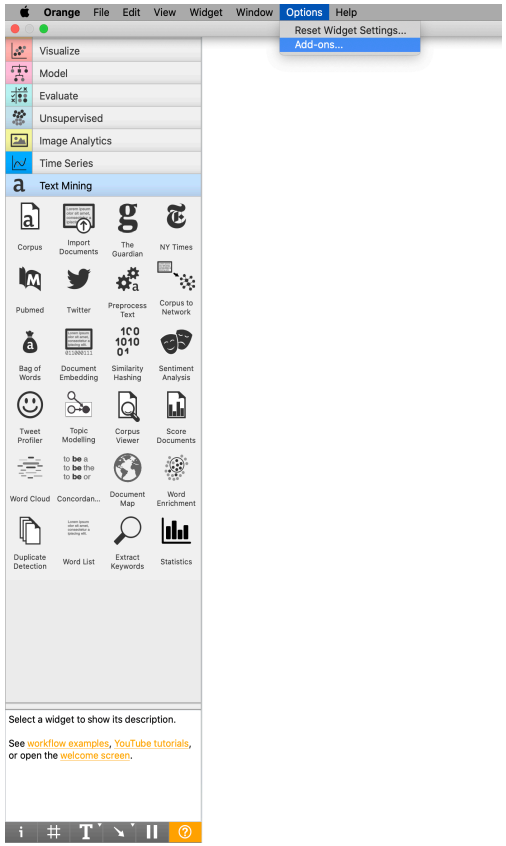


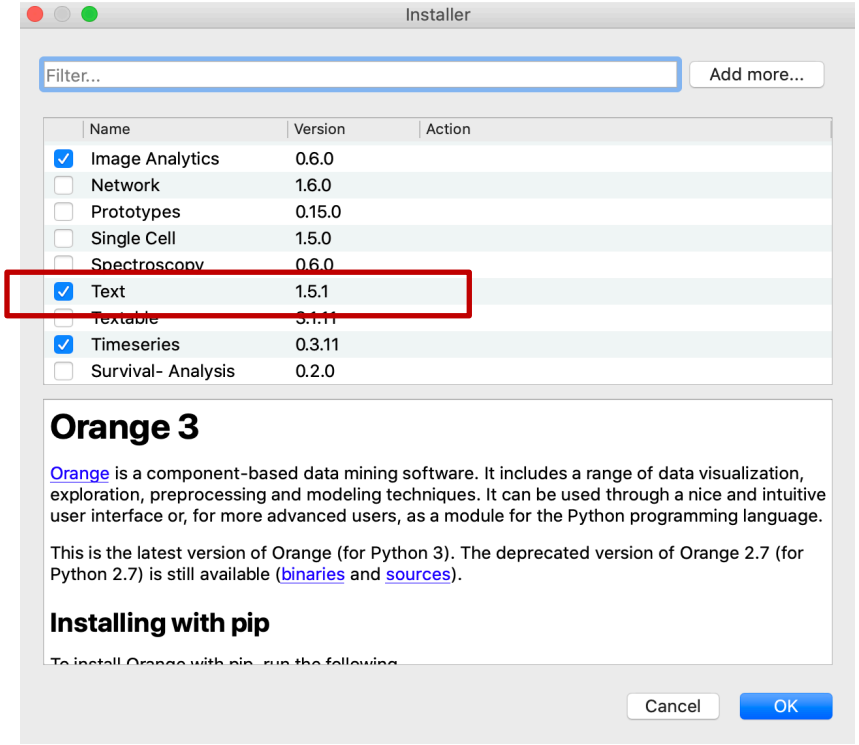


제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 학습 데이터 열기
  - 텍스트 데이터 처리를 위한 “Text Mining” 기능 추가
  - 패키지 설치하는 법: Add-on 〉〉 Text 패키지 체크

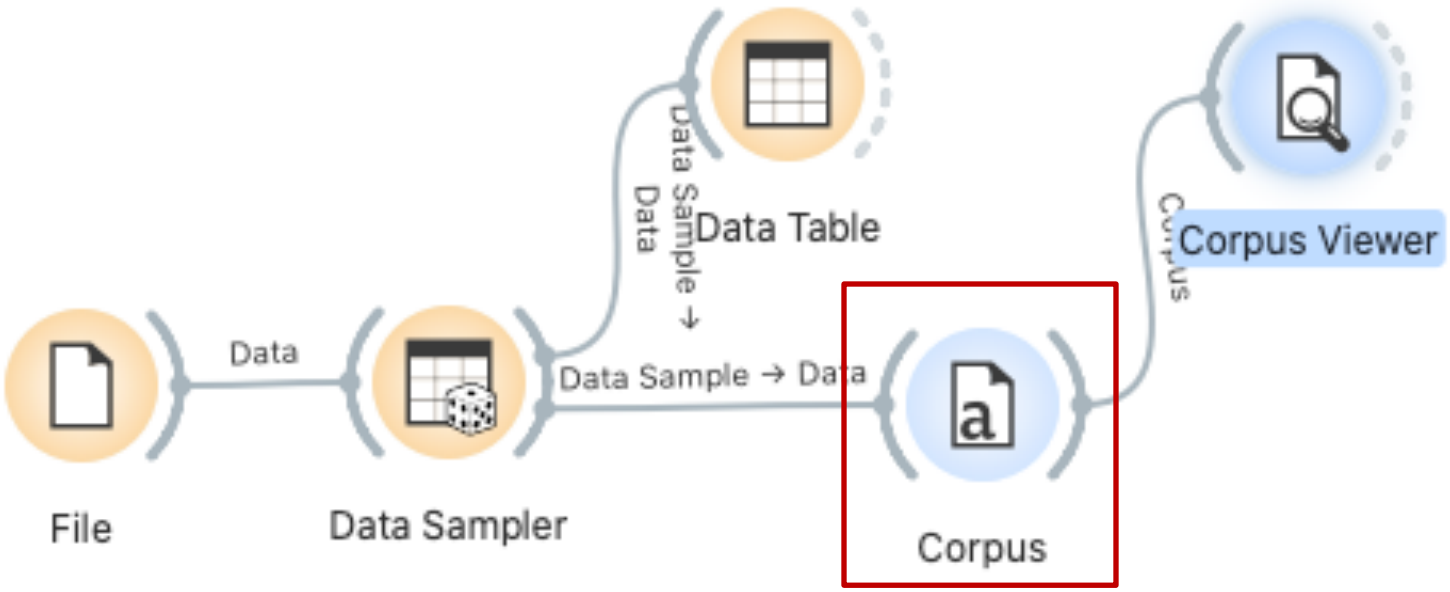




제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

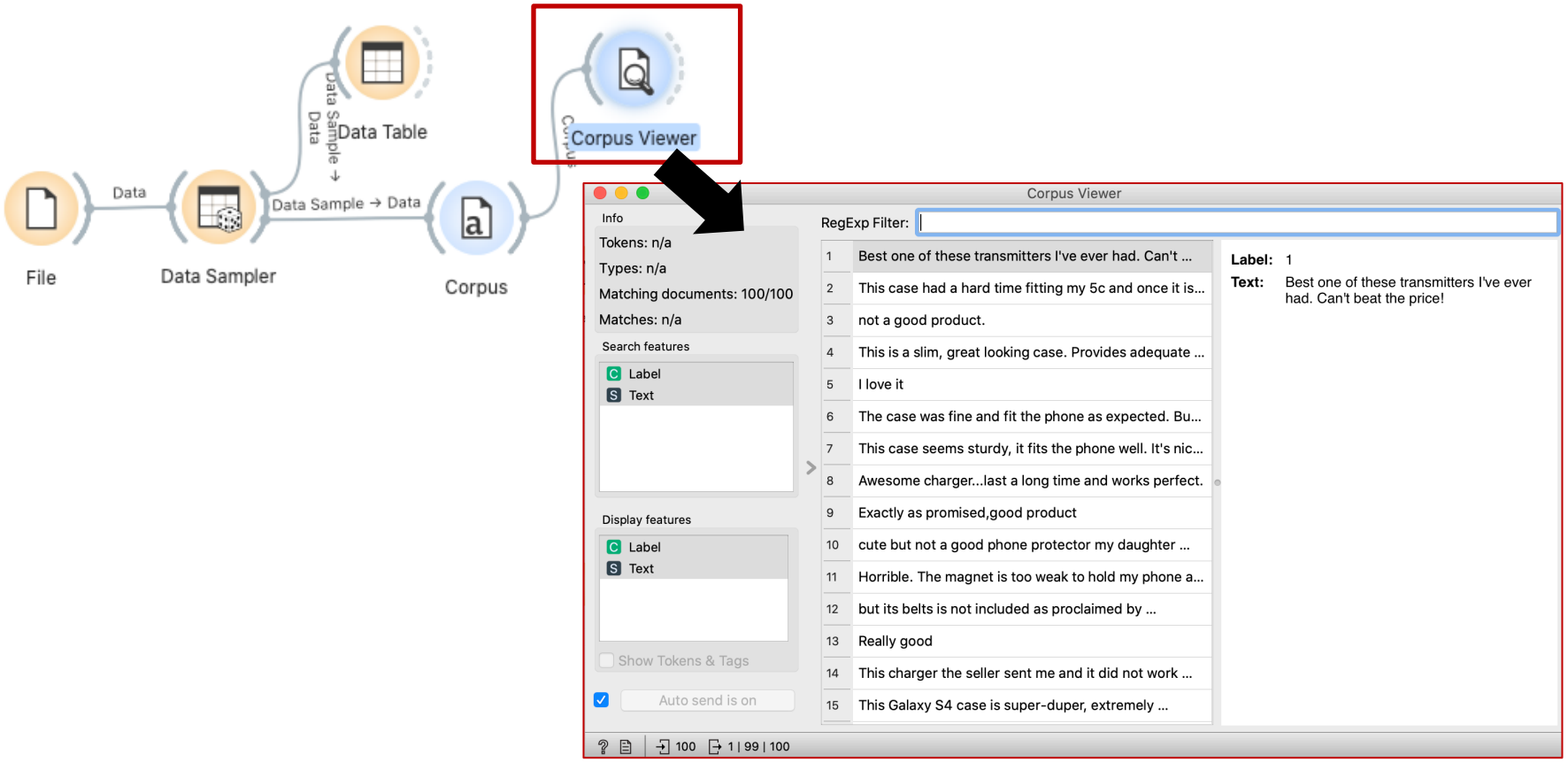
- **학습 데이터 열기**
  - **말뭉치(Corpus) 데이터로 열기: 텍스트를 말뭉치 데이터로 변경**
  - ① Corpus 위젯 드래크&드랍 〉〉 ② 위젯 연결 〉〉 ③ Corpus Viewer 로 확인



제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

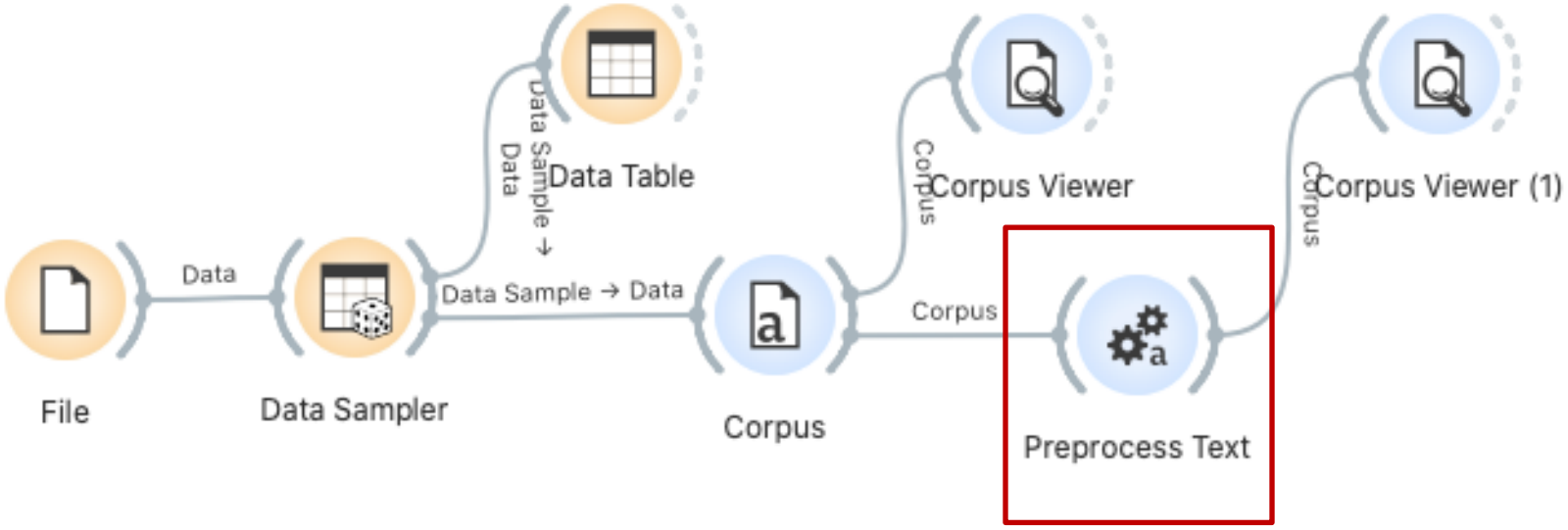
- 학습 데이터 열기
  - 말뭉치(Corpus) 데이터로 열기: 텍스트를 말뭉치 데이터로 변경
  - ① Corpus 위젯 드래크&드랍 〉〉 ② 위젯 연결 〉〉 ③ Corpus Viewer 로 확인



제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② **학습 데이터 전처리 (필수)**
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

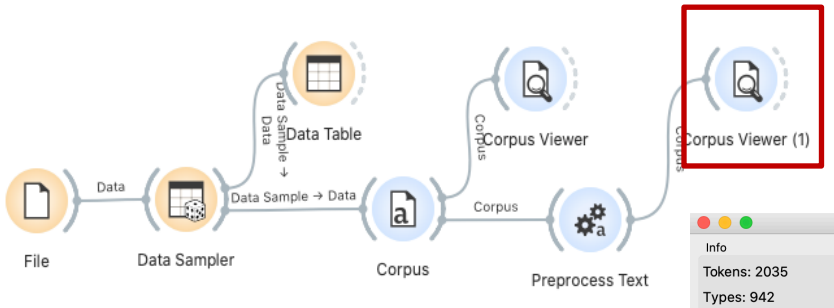
- **학습 데이터 전처리**
  - **텍스트 토큰화 (Tokenization)**
    - **예시)** "There is an apple" ➔ ”There”, “is”, “an”, “apple”

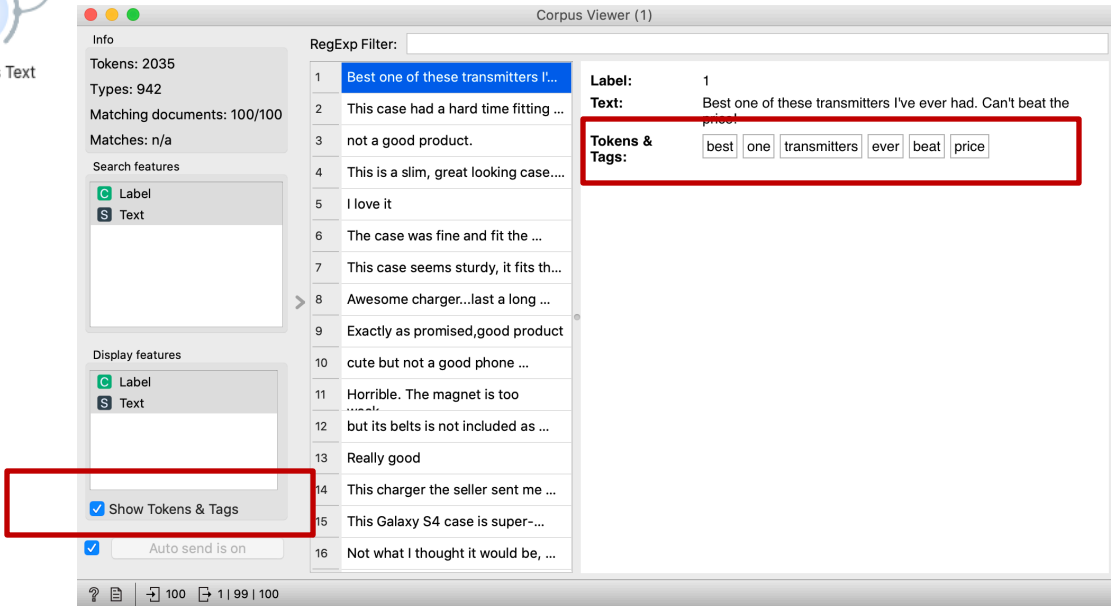


제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② **학습 데이터 전처리 (필수)**
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 학습 데이터 전처리
  - 텍스트 토큰화 (Tokenization) 시각화
    - 예시) "There is an apple" ➔ ”There”, “is”, “an”, “apple”

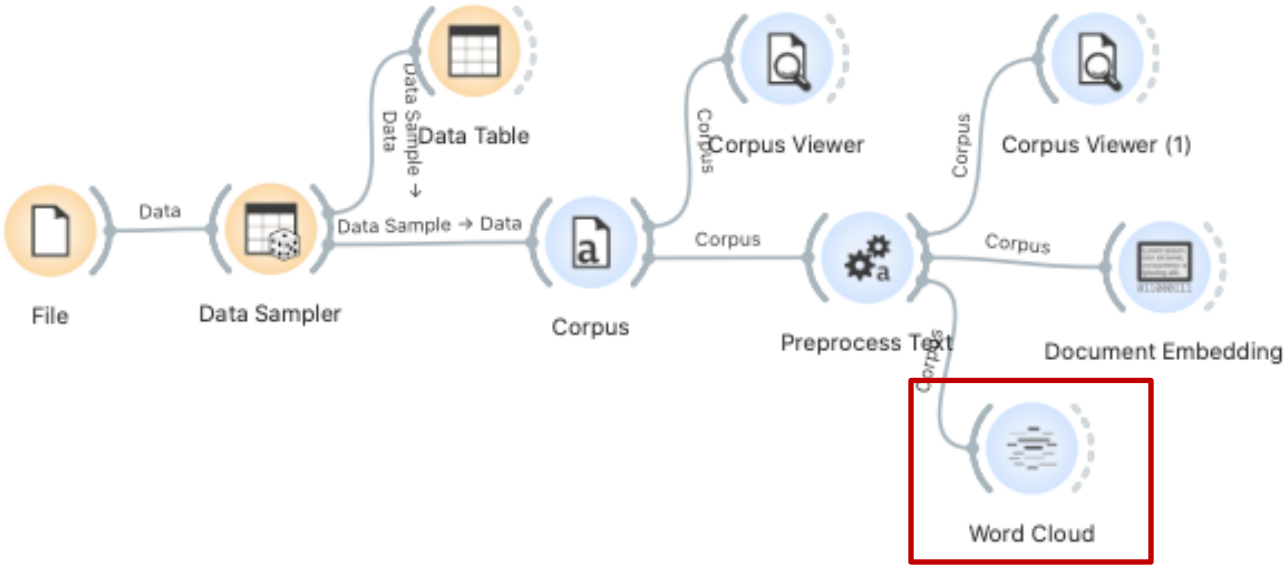




제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② **학습 데이터 전처리 (필수)**
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

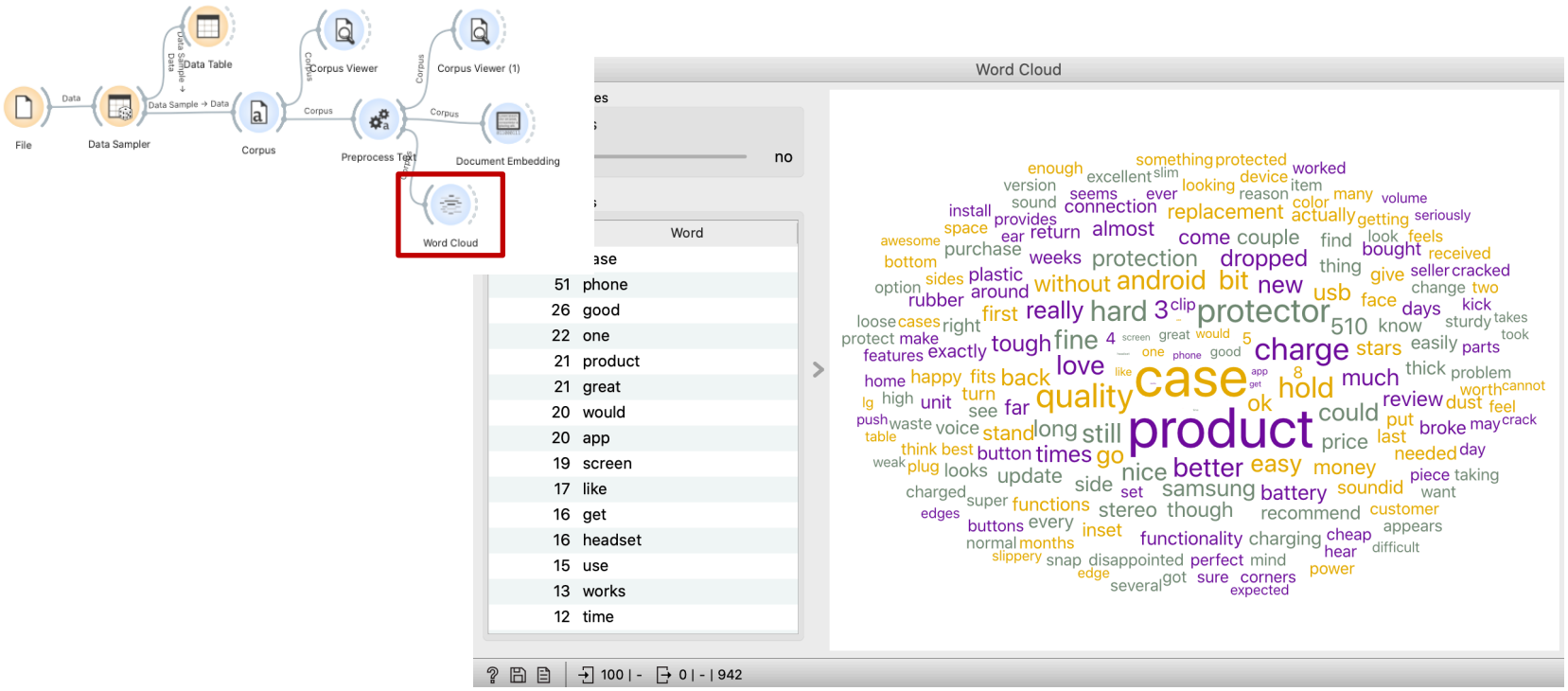
- **학습 데이터 전처리**
  - **텍스트 토큰화 (Tokenization) 시각화**
    - **예시)** "There is an apple" ➔ ”There”, “is”, “an”, “apple”



제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② **학습 데이터 전처리 (필수)**
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 학습 데이터 전처리
  - 텍스트 토큰화 (Tokenization) 시각화
    - 예시) "There is an apple" ➔ ”There”, “is”, “an”, “apple”



제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② **학습 데이터 전처리 (필수)**
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

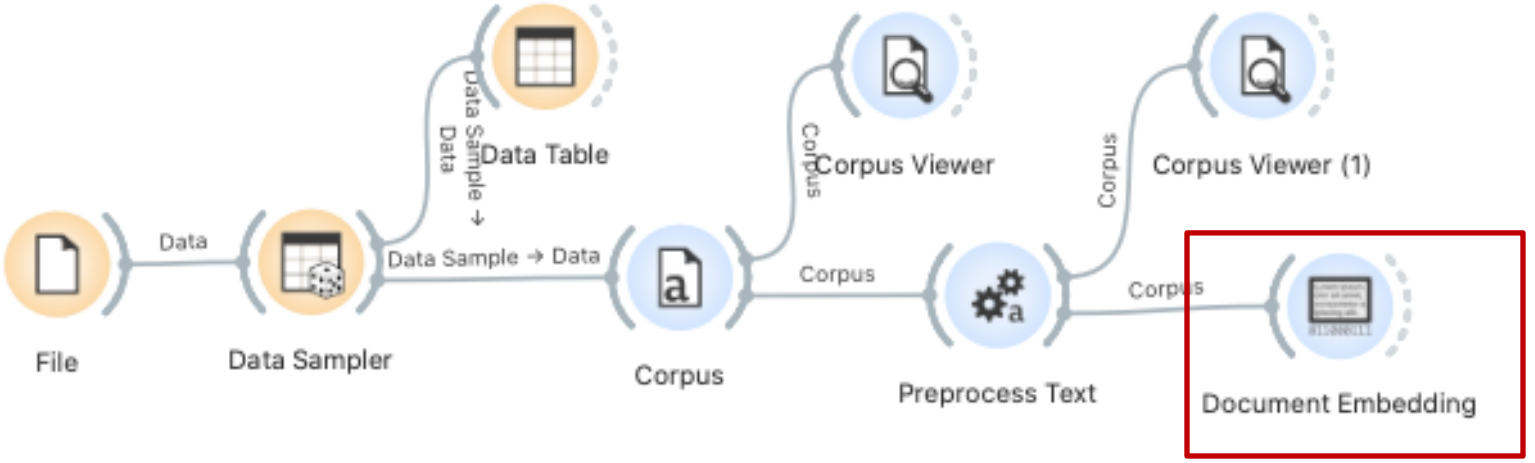
- 학습 데이터 전처리
  - 텍스트 임베딩과 단어 가방(Bag of Words)
    - 예시)

D.1: John likes to watch movies. Mary likes movies too.

D.2: Mary also likes to watch football games.

BoW.1: John:1, likes:2, to:1, watch:1, movies:2, Mary:1, too:1

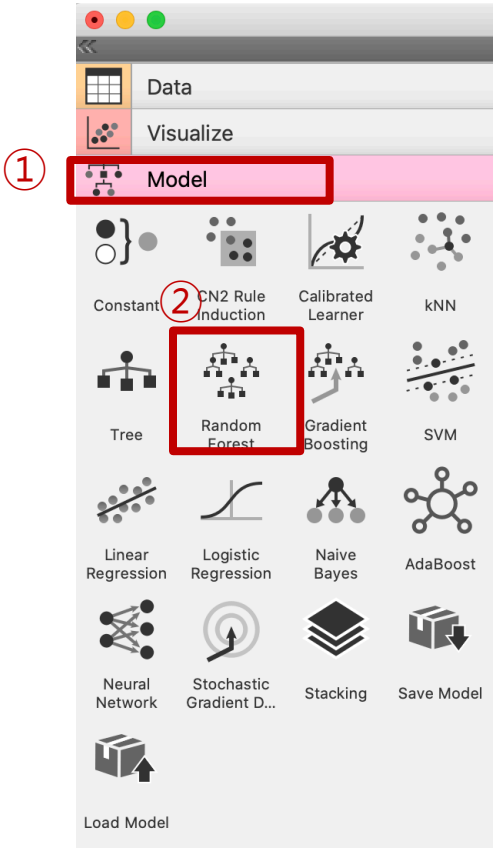
BoW.2: Mary:1, also:1, likes:1, to:1, watch:1, football:1, games:1



제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

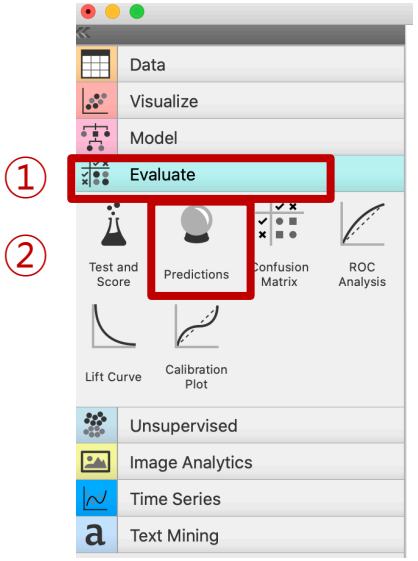
- 학습 데이터로 모델 학습
  - ① Model 탭 클릭 ⟩⟩ ② RF 드래그 & 드랍 ⟩⟩ ③ 위젯 사이 연결

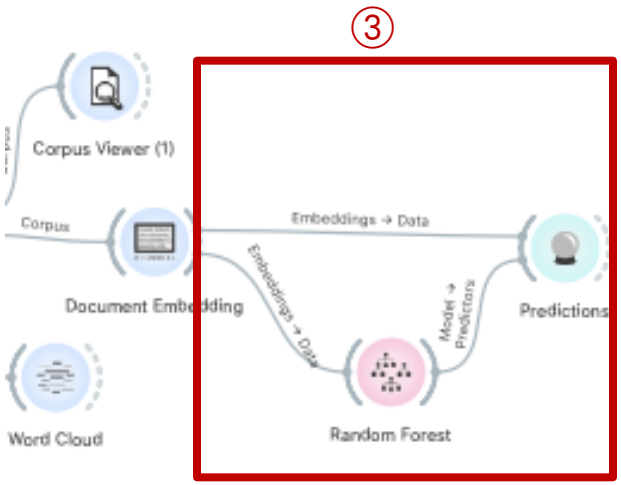


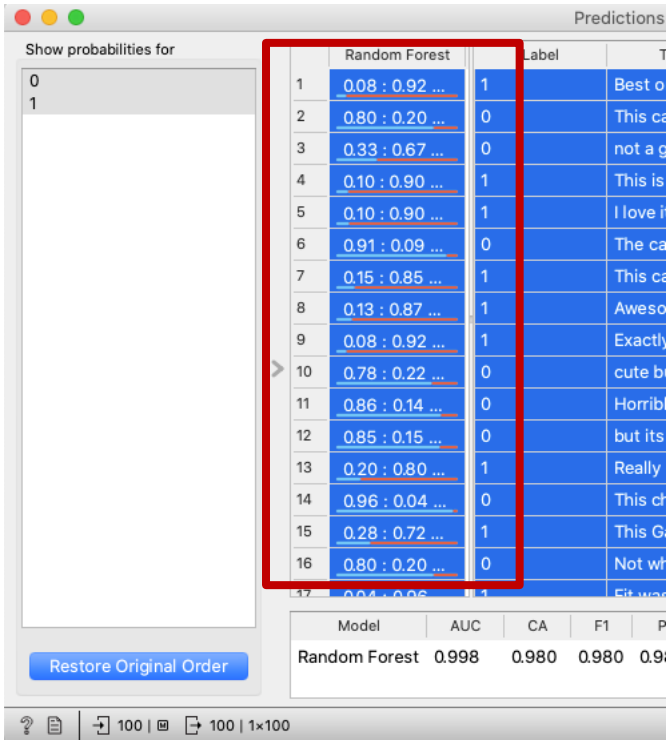
제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 학습 데이터로 모델 평가
  - ① Evaluate 탭 클릭 ⟩⟩ ② Predictions 드래그 & 드랍 ⟩⟩ ③ 위젯 사이 연결



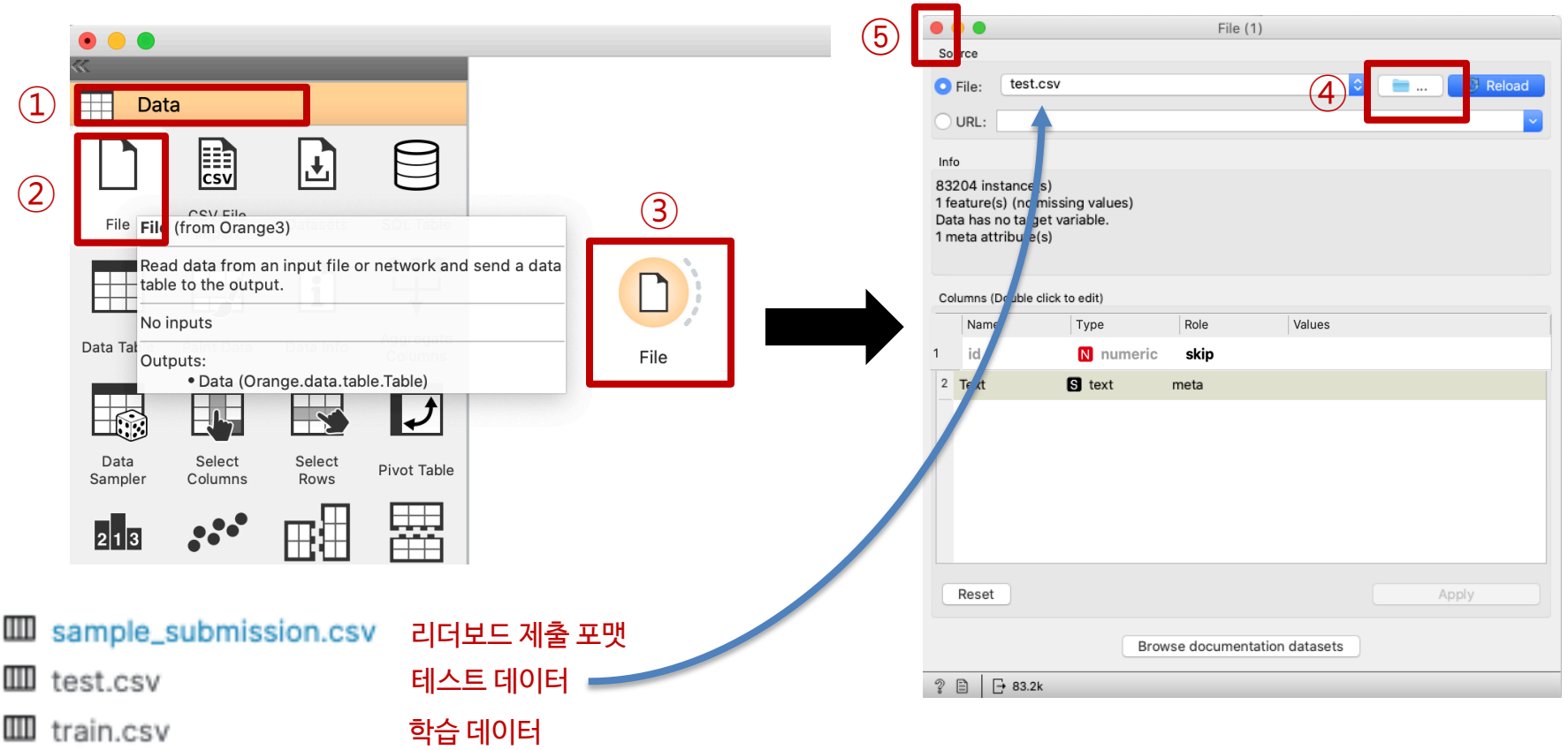




제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ **테스트 데이터 열기 (필수)**
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

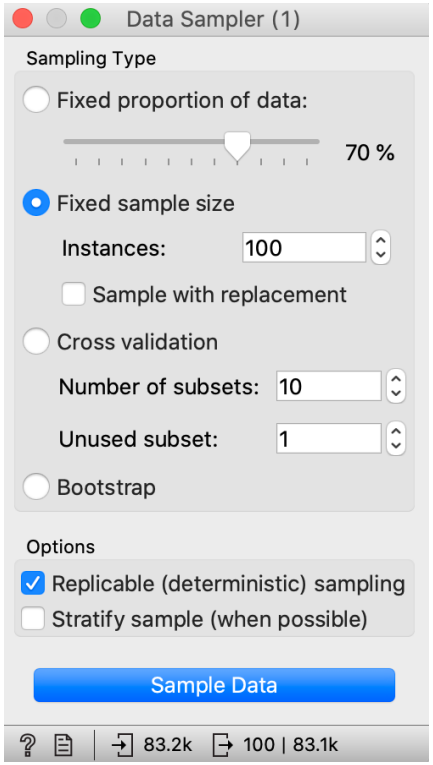
- 테스트 데이터 열기
  - ① Data 탭 클릭 ›› ② File 드래그&드랍 ›› ③ File 위젯 더블클릭
  - ④ File 경로 선택 ›› ⑤ Dialog 닫기

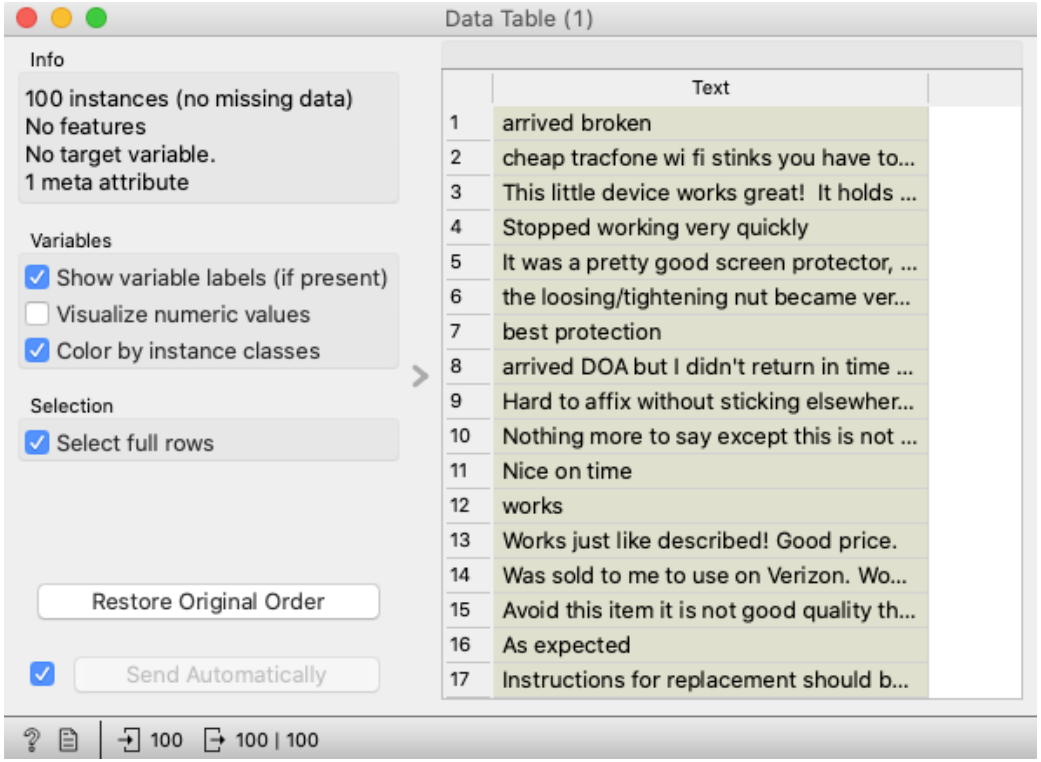


제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ **테스트 데이터 열기 (필수)**
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 테스트 데이터 열기 〉〉 테스트 데이터 샘플링

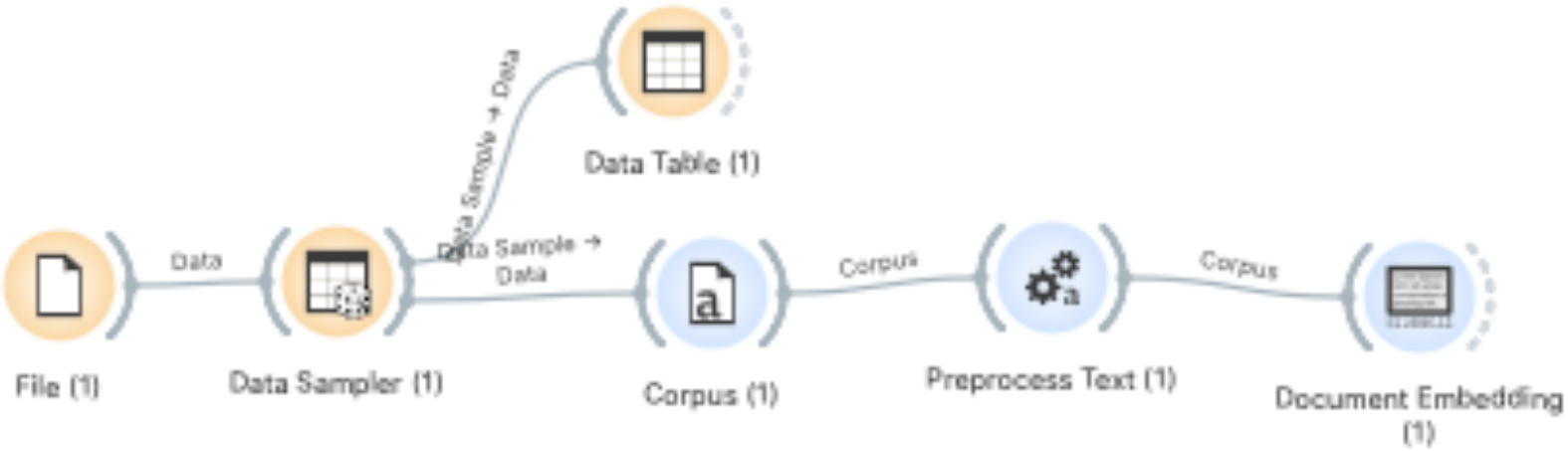




제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ **테스트 데이터 전처리 (필수)**
⑥ 테스트 데이터로 모델 평가 (필수)

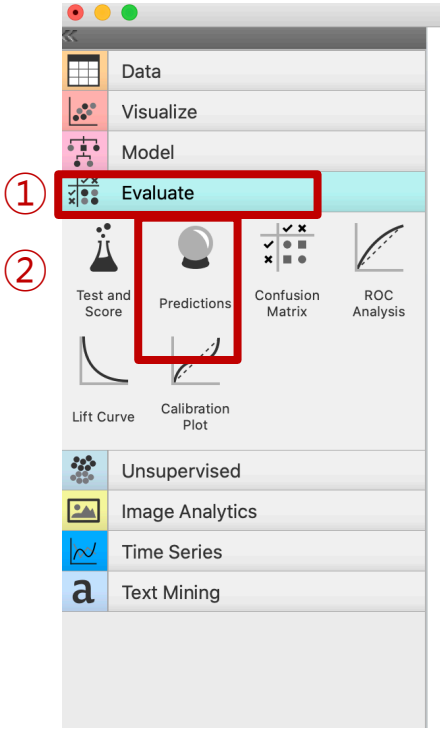
- 테스트 데이터 전처리 (학습 데이터 전처리와 동일)
  - 샘플링 데이터 〉〉 말뭉치(Corpus) 로 변환 〉〉
  - 텍스트 토큰화 〉〉 텍스트 임베딩

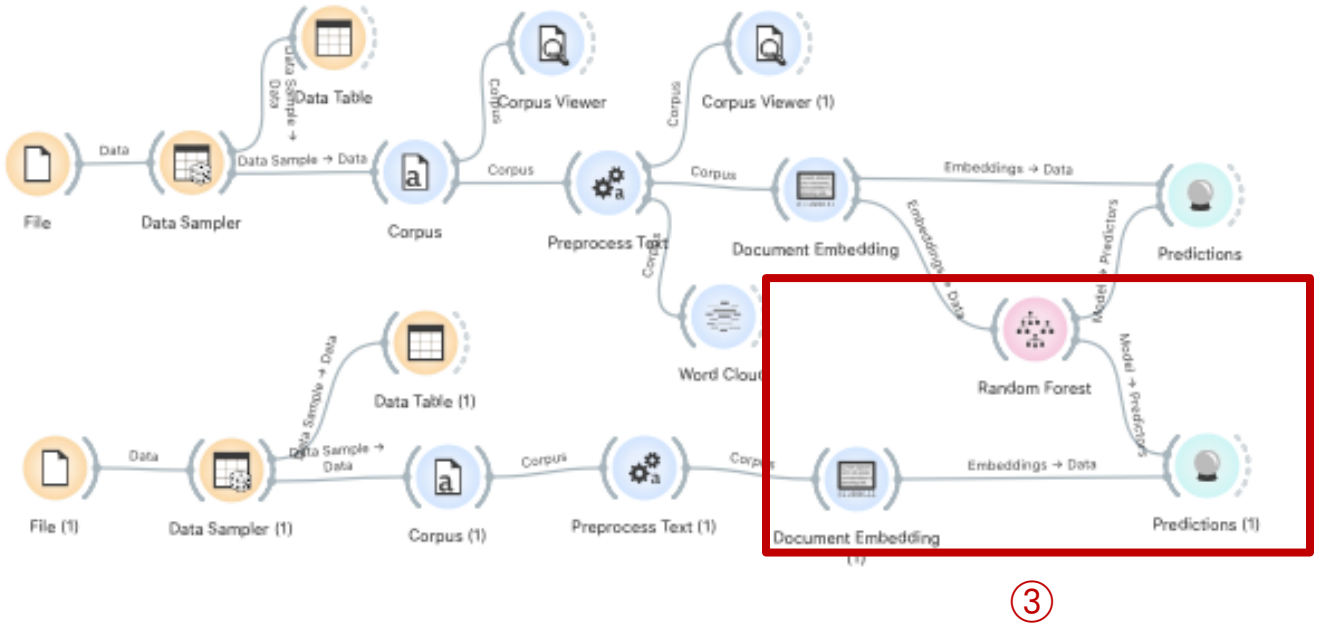


제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ **테스트 데이터로 모델 평가 (필수)**

- 테스트 데이터로 모델 평가
  - ① Evaluate 탭 클릭 ⟩⟩ ② Predictions 드래그 & 드랍 ⟩⟩ ③ 위젯 사이 연결

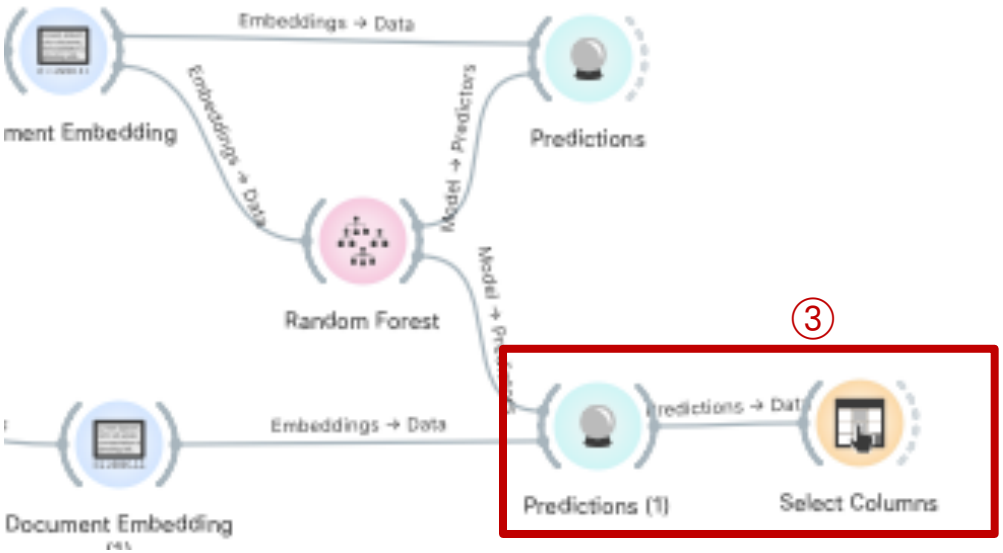


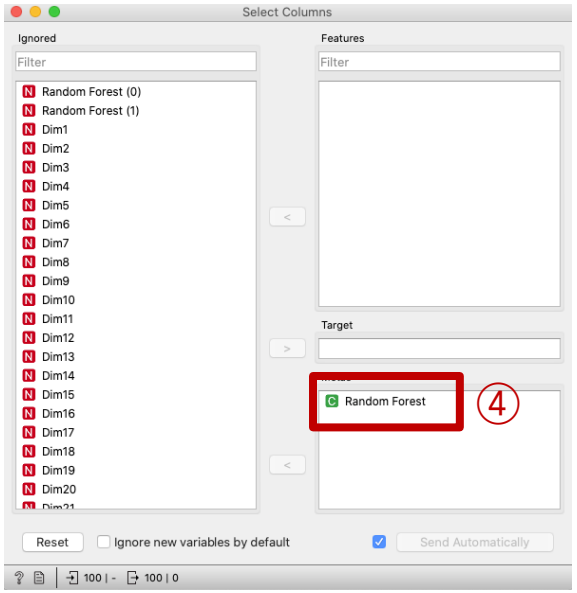


제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 테스트 데이터 예측 결과 저장
  - ① Data 탭 클릭 〉〉 ② Select Columns 드래그 & 드랍 〉〉 ③ 위젯 사이 연결
  - ④ 모델 예측(Random Forest) 데이터만 선택

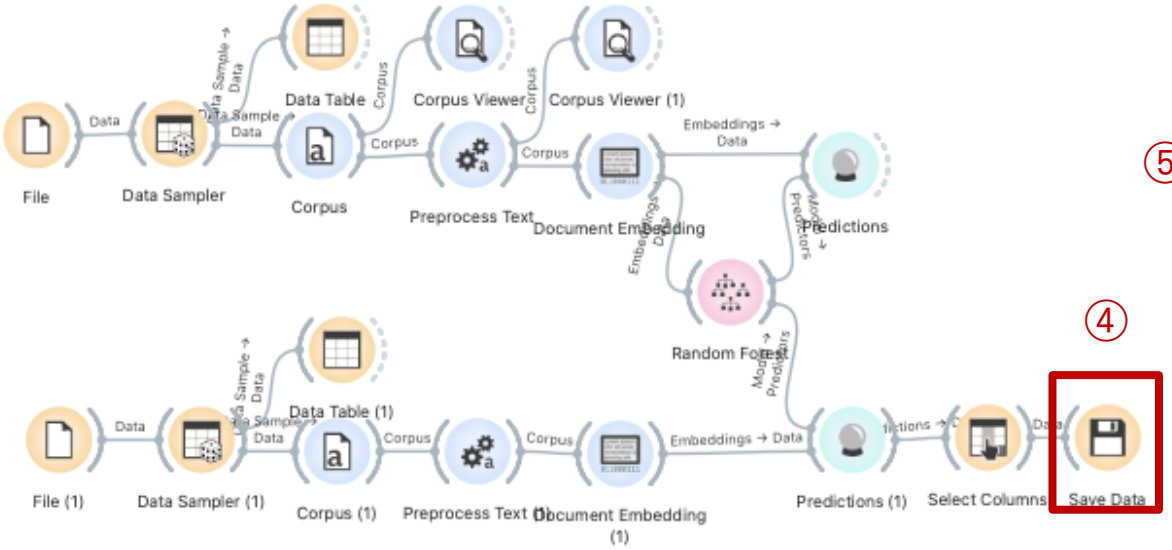


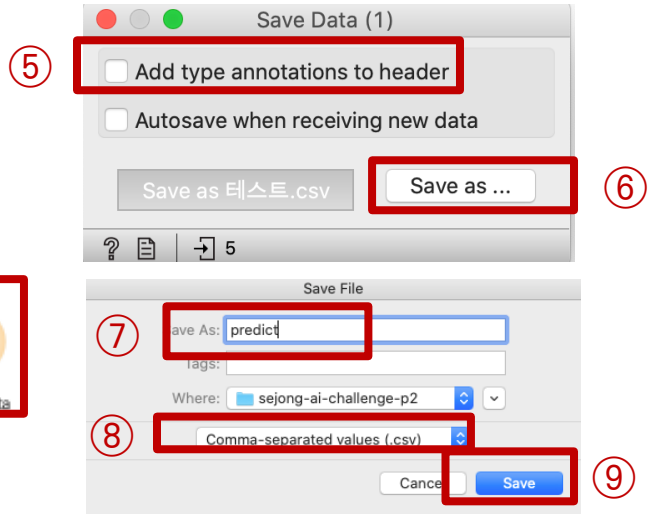


제1회 AI 챌린지 문제 2번: 분류 문제

① 학습 데이터 열기 (필수)
② 학습 데이터 전처리 (필수)
③ 학습 데이터로 모델 학습 (필수)
④ 테스트 데이터 열기 (필수)
⑤ 테스트 데이터 전처리 (필수)
⑥ 테스트 데이터로 모델 평가 (필수)

- 테스트 데이터 예측 결과 저장
  - ④ Save Data 더블 클릭 〉〉 ⑤ Add type.. 체크 해제 〉〉 ⑥ Save as 클릭
  - ⑦ 파일명 설정 〉〉 ⑧ 확장자 csv 변경 〉〉 ⑨ Save 클릭





제1회 AI 챌린지 문제 2번: 분류 문제

- **예측 결과 제출**
  - **캐글 리더보드 제출 파일 만들기**
    - 〈predict.csv〉 파일의 예측 결과를 〈sample_submission.csv〉 파일에 복붙
  - **캐글 리더보드에 답안 제출**
    - 캐글 리더보드에 제출
  - **캐글 리더보드에 답안 제출 확인**
    - 캐글 리더보드에서 등수 확인

주의사항

1) 샘플링 데이터의 학습 결과를 전체 데이터의 학습으로 변경 후
2) 전체 테스트 데이터 결과 예측 후
3) 결과 파일 제출