유전체 정보 품종 분류

1. Data set

- 출처 : https://dacon.io/competitions/official/236035/data
- 구성 : 해당 페이지의 Dataset Info. 참고

Dataset Info.

- **train.csv [파일]** id : 개체 고유 ID

  개체정보

  - father : 개체의 가계 고유 번호 (0 : Unknown)
  - mother : 개체의 모계 고유 번호 (0 : Unknown)
  - gender : 개체 성별 (0 : Unknown, 1 : female, 2 : male)
  - trait : 개체 표현형 정보
  - 15개의 SNP 정보 : SNP_01 ~ SNP_15
  - class : 개체의 품종 (A,B,C)

- **test.csv [파일]** id : 개체 샘플 별 고유 ID

  개체정보

  - father : 개체의 가계 고유 번호 (0 : Unknown)
  - mother : 개체의 모계 고유 번호 (0 : Unknown)
  - gender : 개체 성별 (0 : Unknown, 1 : female, 2 : male)
  - trait : 개체 표현형 정보
  - 15개의 SNP 정보 : SNP_01 ~ SNP_15

- **snp_info.csv [파일]** 15개의 SNP 세부 정보
  - name : SNP 명
  - chrom : 염색체 정보
  - cm : Genetic distance
  - pos : 각 마커의 유전체상 위치 정보

2. Development

- 목표

  1) DACON 10%안에 들기(이름에 색깔 칠해지기)

  2) 데이터 처리 여러 방법으로 해보기(데이터 처리 연습)

     - SNP의 특징 반영하기

- 프로젝트 소스 : 프로젝트2.유전체정부품종분류.ipynb

- 개발 환경

  Jupyter notebook

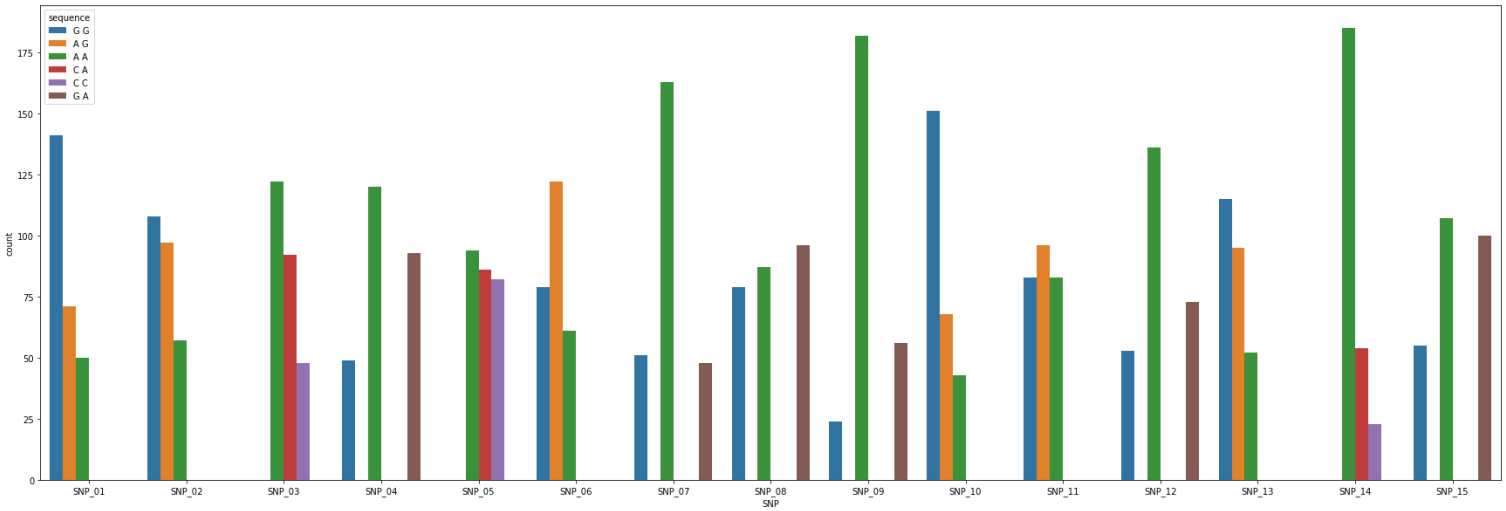
  모듈 Version

Package	Version
jupyter	1.0.0
notebook	6.4.8
keras	2.11.0
matplotlib	3.5.1
numpy	1.21.5
pandas	1.4.2
Python	3.9.12
pip	21.2.4
tensorflow	2.11.0

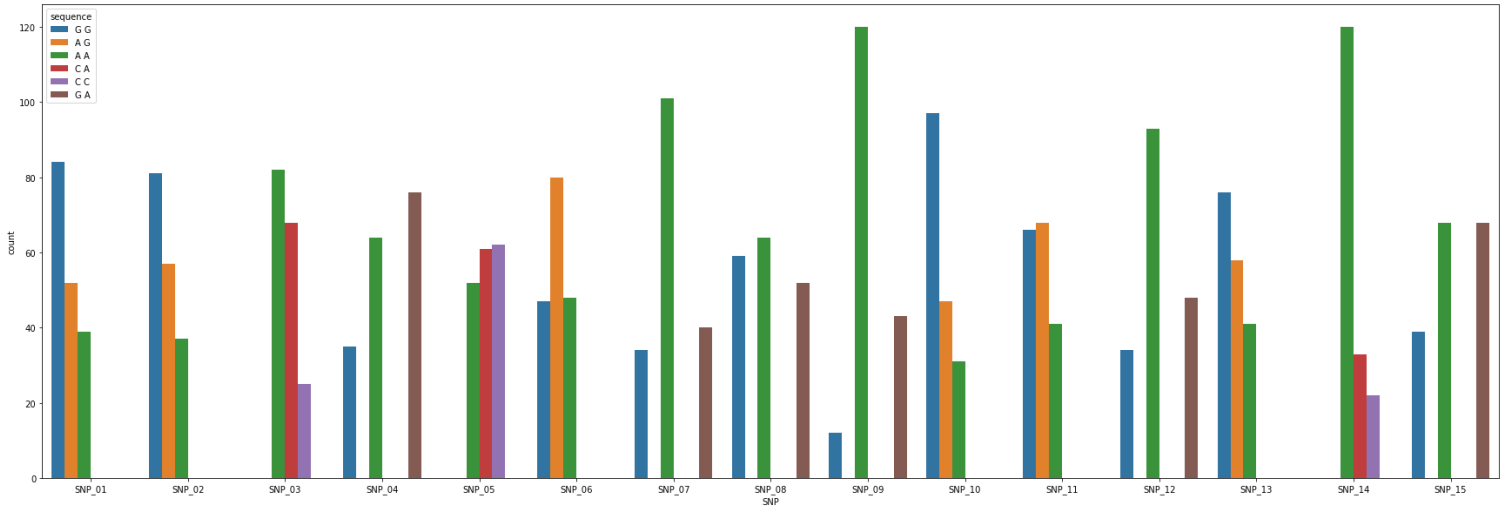
<데이터 분석 시각화>

각 SNP에서 서열이 차지하는 정도 시각화

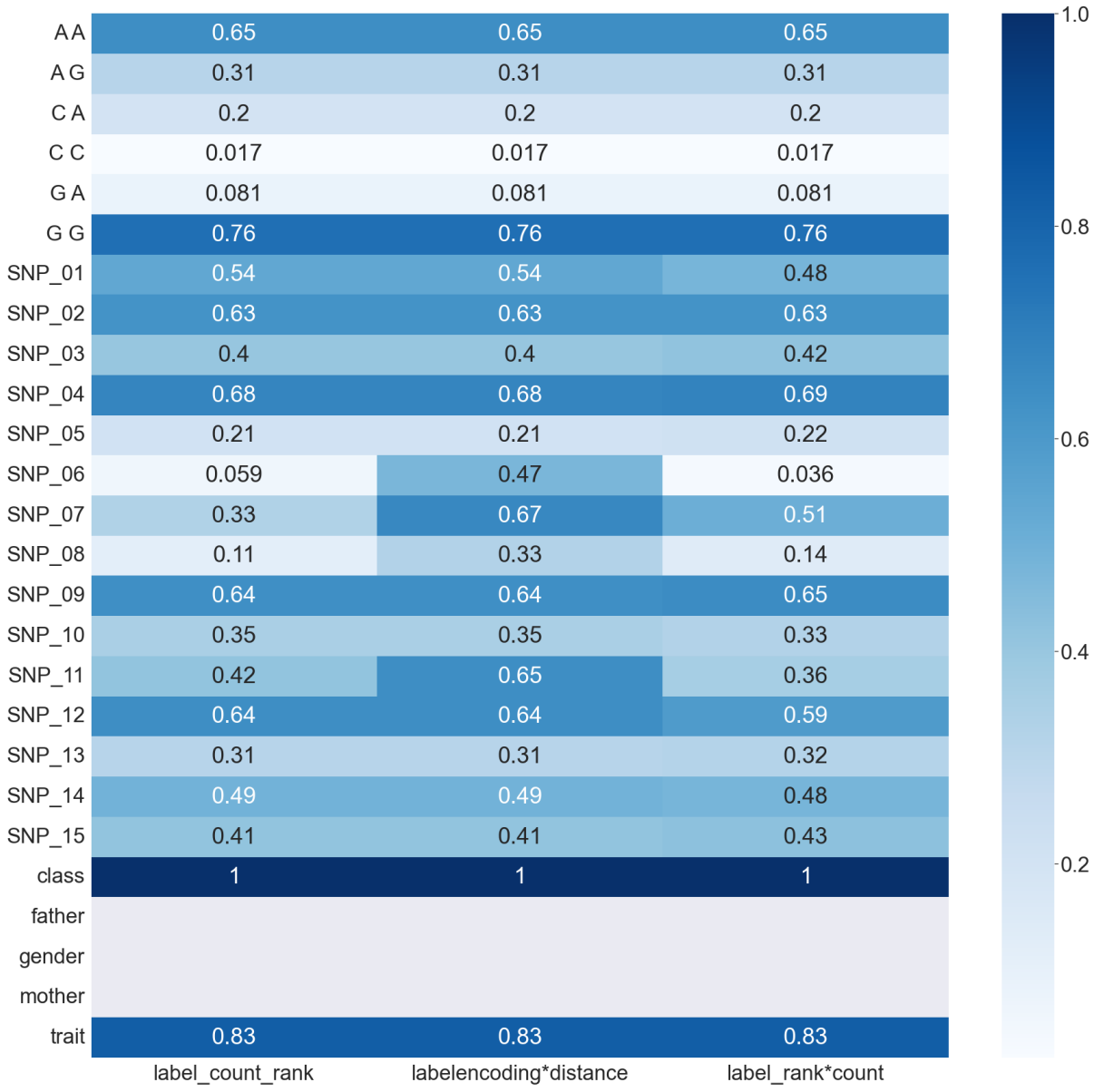
train_data



test_data



상관관계 시각화



Label_count_rank : 각 SNP의 서열 labeling을 각 SNP에서 차지하는 count순위로 대체 후 correlation 확인

Labelencoding*distance : 각 SNP의 서열 lableing을 sklearn의 labelencoder로 대체 후 snp_info의 genetic distance 값을 곱해서 genetic distance크기를 데이터에 반영 후 correlation확인

Label_rank*count : 각 SNP의 서열 labeling을 각 SNP에서 차지하는 count순위로 대체 후 각 SNP에서 서열의 실제 count값을 곱한 후 correlation 확인