MS Azure ML Designer를 활용한 개인 실습

Lab : 와인 데이터

강명호

배경 및 모델링 목표

배경 및 목표

Wine Quality 데이터 세트를 분석하여 와인의 종류(white or red)를 분류하고,
와인의 특성을 통해 품질을 예측하는 모델 구현

- **Story :**
  - **와인 메이커는 와인 분석가를 고용하여 산도, 단맛, 알코올 수준 등 생산 중인 와인의 특성을 측정하고 개선함**
  - **최근, 와인 분석가가 다른 프로젝트에 배정되어서 와이너리의 와인 분석 작업을 진행하지 못하게 되었음**
  - **와인 분석가를 대체하기 위한 머신러닝 모델이 필요한 상황임**

배경 및 목표

Wine Quality 데이터 세트를 분석하여 와인의 종류(white or red)를 분류하고,
와인의 특성을 통해 품질을 예측하는 모델 구현

- **데이터 소스 : UCI Machine Learning Repository**
- **포르투갈의 Vinho Verde의 레드 및 화이트 와인 데이터에서 일부 발췌 (민감한 정보 제외)**

와인 데이터 세트를 이용하여 모델 구현

- **와인의 종류를 분류하는 모델, 또는**
- **와인의 품질을 측정하는 모델**

데이터 수집

데이터 수집 (참고)

Wine Quality

Donated on 10/6/2009

Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009],...

Dataset Characteristics
Multivariate

Subject Area
Business

Associated Tasks
Classification, Regression

Feature Type
Real

 # Instances
4898

 # Features
11

Dataset Information

Additional Information

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: http://www.vinhoverde.pt/en/ or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection

DOWNLOAD

IMPORT IN PYTHON

CITE

1 citations

369950 views

Keywords

Chemistry

Creators

Paulo Cortez

A. Cerdeira

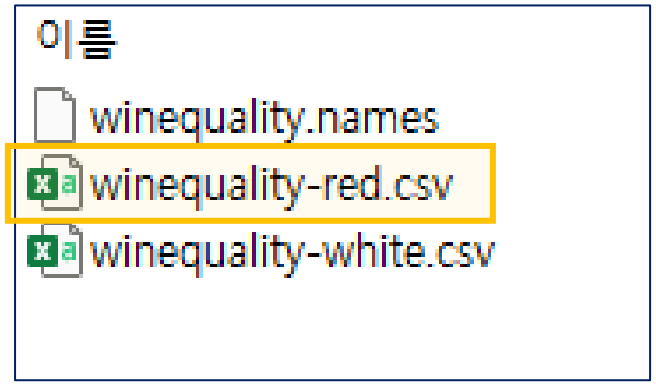
F. Almeida

T. Matos

J. Reis

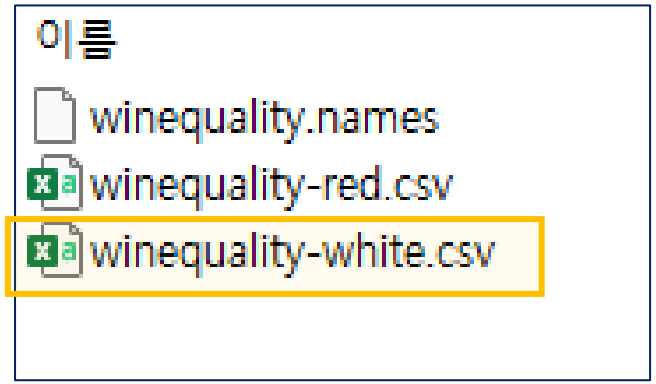
DOI

데이터 수집 (참고)



	A	B	C	D	E	F	G	H	I	J
1	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide
2	7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
3	7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
4	7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
5	11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
6	7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7	7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5
8	7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5
9	7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7
10	7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7
11	7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
12	6.7;0.58;0.08;1.8;0.097;15;65;0.9959;3.28;0.54;9.2;5
13	7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
14	5.6;0.615;0;1.6;0.089;16;59;0.9943;3.58;0.52;9.9;5
15	7.8;0.61;0.29;1.6;0.114;9;29;0.9974;3.26;1.56;9.1;5
16	8.9;0.62;0.18;3.8;0.176;52;145;0.9986;3.16;0.88;9.2;5
17	8.9;0.62;0.19;3.9;0.17;51;148;0.9986;3.17;0.93;9.2;5
18	8.5;0.28;0.56;1.8;0.092;35;103;0.9969;3.3;0.75;10.5;7
19	8.1;0.56;0.28;1.7;0.368;16;56;0.9968;3.11;1.28;9.3;5
20	7.4;0.59;0.08;4.4;0.086;6;29;0.9974;3.38;0.5;9;4
21	7.9;0.32;0.51;1.8;0.341;17;56;0.9969;3.04;1.08;9.2;6
22	8.9;0.22;0.48;1.8;0.077;29;60;0.9968;3.39;0.53;9.4;6
23	7.6;0.39;0.31;2.3;0.082;23;71;0.9982;3.52;0.65;9.7;5

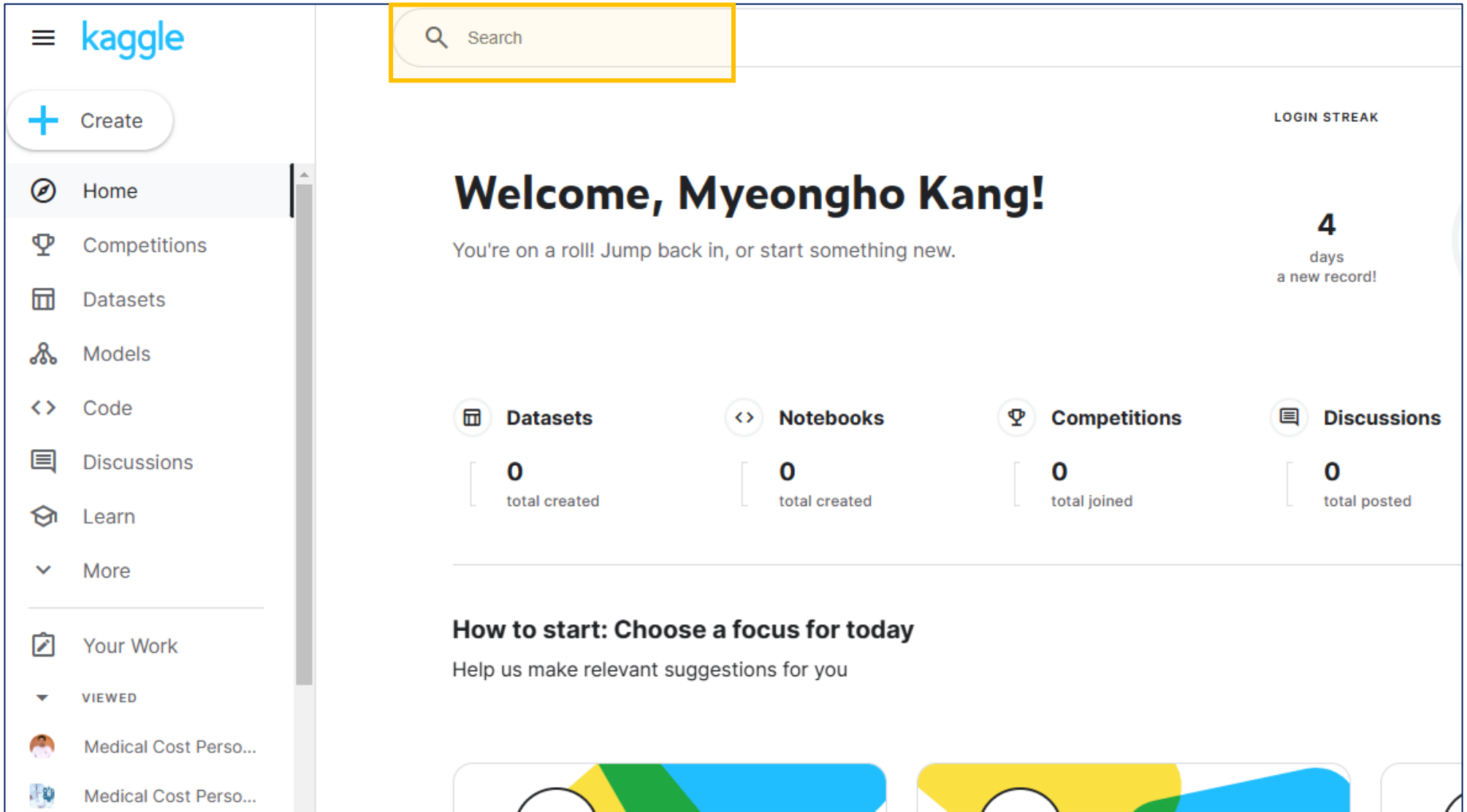
데이터 수집 (참고)



	A	B	C	D	E	F	G	H	I	J	K
1	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"										
2	7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6										
3	6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6										
4	8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6										
5	7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6										
6	7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6										
7	8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6										
8	6.2;0.32;0.16;7;0.045;30;136;0.9949;3.18;0.47;9.6;6										
9	7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6										
10	6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6										
11	8.1;0.22;0.43;1.5;0.044;28;129;0.9938;3.22;0.45;11;6										
12	8.1;0.27;0.41;1.45;0.033;11;63;0.9908;2.99;0.56;12;5										
13	8.6;0.23;0.4;4.2;0.035;17;109;0.9947;3.14;0.53;9.7;5										
14	7.9;0.18;0.37;1.2;0.04;16;75;0.992;3.18;0.63;10.8;5										
15	6.6;0.16;0.4;1.5;0.044;48;143;0.9912;3.54;0.52;12.4;7										
16	8.3;0.42;0.62;19.25;0.04;41;172;1.0002;2.98;0.67;9.7;5										
17	6.6;0.17;0.38;1.5;0.032;28;112;0.9914;3.25;0.55;11.4;7										
18	6.3;0.48;0.04;1.1;0.046;30;99;0.9928;3.24;0.36;9.6;6										
19	6.2;0.66;0.48;1.2;0.029;29;75;0.9892;3.33;0.39;12.8;8										
20	7.4;0.34;0.42;1.1;0.033;17;171;0.9917;3.12;0.53;11.3;6										
21	6.5;0.31;0.14;7.5;0.044;34;133;0.9955;3.22;0.5;9.5;5										
22	6.2;0.66;0.48;1.2;0.029;29;75;0.9892;3.33;0.39;12.8;8										
23	6.4;0.31;0.38;2.9;0.038;19;102;0.9912;3.17;0.35;11;7										
24	6.8;0.26;0.42;1.7;0.049;41;122;0.993;3.47;0.48;10.5;8										

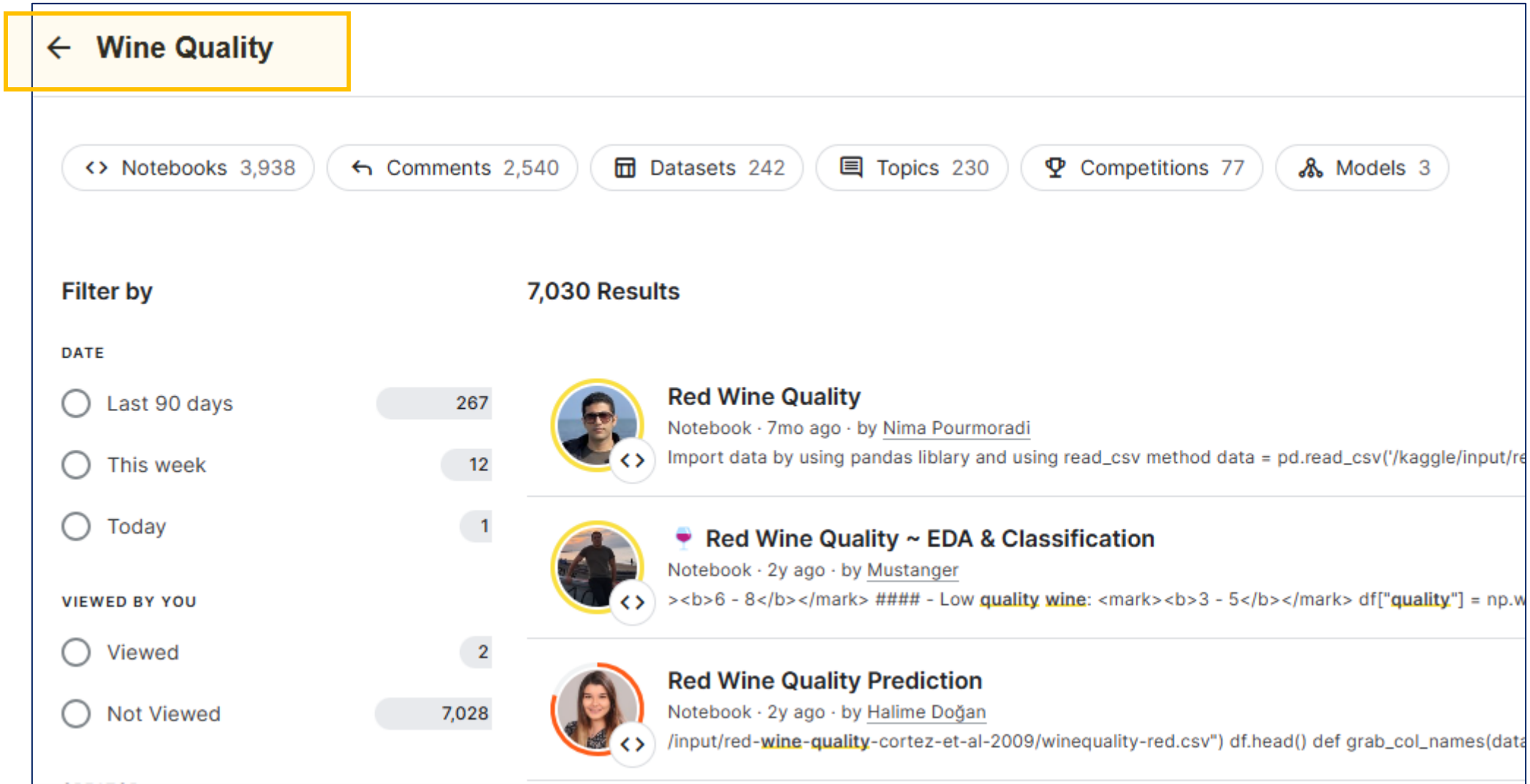
데이터 수집

Kaggle에 접속하여 Wine Quality 데이터 세트 검색



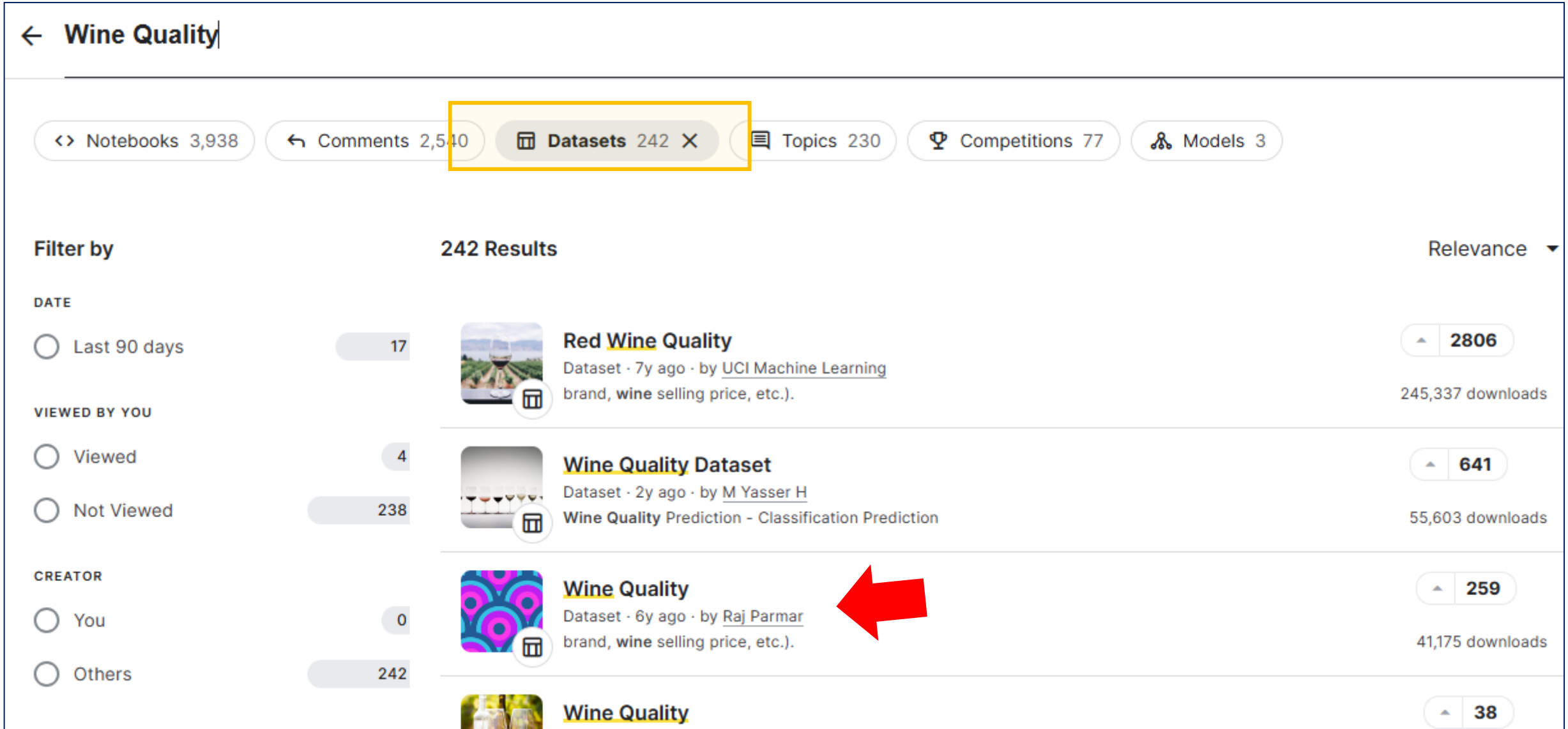
데이터 수집

Kaggle에 접속하여 Wine Quality 데이터 세트 검색



데이터 수집

Kaggle에 접속하여 Wine Quality 데이터 세트 검색



데이터 수집

RAJ PARMAR · UPDATED 6 YEARS AGO

259 | New Notebook | Download (100 kB)

Wine Quality

Data Card | Code (171) | Discussion (3) | Suggestions (0)

About Dataset

Data Set Information:

The dataset was downloaded from the UCI Machine Learning Repository.

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. The reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more

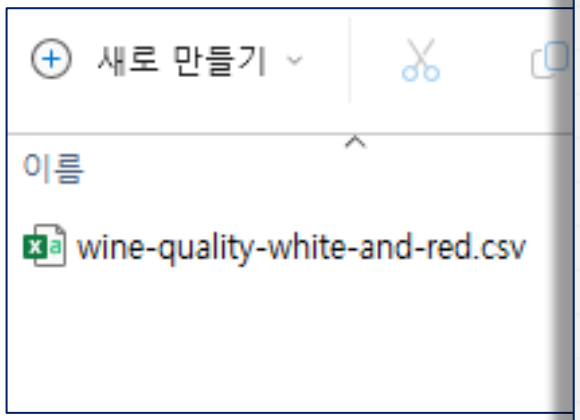
Usability ⓘ
7.06

License
Other (specified in descript

Expected update frequ
Not specified

Tags

데이터 수집



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	type	fixed acidi	volatile ac	citric acid	residual su	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	quality
2	white	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
3	white	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
4	white	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
5	white	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
6	white	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
7	white	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
8	white	6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6
9	white	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
10	white	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
11	white	8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6
12	white	8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5
13	white	8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5
14	white	7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5
15	white	6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7
16	white	8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5
17	white	6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7
18	white	6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	6
19	white	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
20	white	7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	6
21	white	6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	5
22	white	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
23	white	6.4	0.31	0.38	2.9	0.038	19	102	0.9912	3.17	0.35	11	7
24	white	6.8	0.26	0.42	1.7	0.049	41	122	0.993	3.47	0.48	10.5	8

데이터 수집

About Dataset

Data Set Information:

The dataset was downloaded from the UCI Machine Learning Repository.

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. The reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Two datasets were combined and few values were randomly removed.

Attribute Information:

For more information, read [Cortez et al., 2009].

Usability ⓘ

7.06

License

Other (specified in description)

Expected update frequency

Not specified

Tags

Earth and Nature

Classification

Alcohol

Regression

데이터 수집

- fixed acidity : 고정 산도
- volatile acidity : 휘발성 산도
- citric acid : 시트르산
- residual sugar : 잔류 당분
- chlorides : 염화물
- free sulfur dioxide : 자유 이산화황
- total sulfur dioxide : 총 이산화황
- density : 밀도
- pH
- sulphates : 황산염
- alcohol
- quality : 0 ~ 10(높을 수록 좋은 품질)

Attribute Information:

For more information, read [Cortez et al., 2009].

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)