

시험지 - 퀴즈1

아래의 내용을 반드시 숙지하세요

- 현재 스마트폰 테더링 중입니까? 감독 조교에게 연락하세요
- 시험에 응시할 경우, 부정 행위에 대한 모든 책임이 학생 본인에게 있음에 동의 한 것으로 간주합니다.
- 문제 중 붉은색 글씨로 작성된 내용을 반드시 확인하세요.

- 구글 드라이브 공유 폴더: 코드(ipynb), 리더보드 제출 파일(csv)
- 대용량 이메일: admin@rcv.sejong.ac.kr 로 대용량 이메일로 전송 요망 (고사 이후 2시간 내)

[문제1][20점] <https://www.kaggle.com/t/83524cac7c2d4c4ca8554e669972436e>

철수와 영희는 도심 근교에 작은 텃밭을 하나 구입했습니다. 작은 텃밭을 주말농장으로 운영하여 작물도 재배하고 소소한 즐거움도 느껴보고자 합니다. 그런데 철수와 영희는 농사에 대한 지식이 부족하여 구입한 텃밭에 어떤 작물을 재배하는 것이 좋은지 알지 못합니다. 어떻게 해야 이 난관을 해결할 수 있을까요? 여러분이 수업 시간에 배운 <머신러닝 기술>을 통해 철수와 영희를 도와주세요. 여러분의 예측 모델이 철수와 영희에게 주말농장의 소소한 재미를 제공하길 희망합니다.

학습 데이터는 (재배 환경)과 재배 환경에 적합한 (작물의 종류)를 제공합니다.

	N	P	K	temperature	humidity	ph	rainfall	label
0	22	36	16	30.581395	50.771481	8.184229	64.585596	mothbeans
1	18	27	41	22.365094	92.308824	7.175344	104.821633	pomegranate
2	61	68	50	35.214628	91.497251	6.793245	243.074507	papaya
3	26	72	22	28.767949	37.577921	4.674942	91.720849	pigeonpeas
4	2	140	197	22.697801	92.822234	5.534567	105.050823	apple
...

테스트 데이터는 (재배 환경)만 제공하며, 재배 환경에 적합한 (작물의 종류)는 submit.csv 파일로 저장하여 캐글 리더보드에 제출하셔야 합니다.

	N	P	K	temperature	humidity	ph	rainfall
0	101	17	47	29.494014	94.729813	6.185053	26.308209
1	98	8	51	26.179346	86.522581	6.259336	49.430510
2	59	62	49	43.360515	93.351916	6.941497	114.778071
3	44	60	55	34.280461	90.555616	6.825371	98.540477
4	30	137	200	22.914300	90.704756	5.603413	118.604465
...

(재배 환경)에 대한 정보는 다음과 같습니다. N은 토양 내 질소 함량 비율, P는 토양 내 인산 함량 비율, K는 토양 내 칼륨 함량 비율을 나타내며, temperature는 섭씨 온도를, humidity는 상대 습도를 ph는 토양의 ph 값을 rainfall은 강우량을 의미합니다.

재배 환경에 맞는 (작물의 종류)에는 총 22개가 있으며 구체적인 정보는 다음과 같습니다. mothbeans, lentil, pigeonpeas, cotton, coconut, mungbean, jute, chickpea, mango, apple, blackgram, muskmelon, papaya, coffee, orange, pomegranate, kidneybeans, rice, banana, watermelon, maize, grapes

평가 지표는 Categorization Accuracy를 사용하세요.

자, 그럼 테스트 데이터로 주어진 (재배 환경)에 맞는 (작물의 종류)를 예측하여 봅시다. Let's start!!

※ 주의사항 ※

- ① 배운 범위 내의 기계학습 방법론만 사용합니다. (100% 감점)
- ② 반드시 학습 과정에서 범주형 데이터를 수치형 데이터로 변환하여 사용합니다. (50% 감점)
- ③ 상위 랭커 6인에게 추가 점수를 부여합니다. (10% 가점)
단, 공동 순위가 6인 이상일 경우 추가 점수를 부여하지 않음

[문제2][20점] <https://www.kaggle.com/t/b5f050461afe4605b48fac4764c09763>

철수네 가족은 오랫동안 사용해온 패밀리카를 팔고 새로운 차를 구입하려고 한다. 엔카 혹은 중고나라에 판매 글을 올리려니 사용해온 차량의 각종 정보와 판매 희망가를 작성하게 되어있었다. 그런데 갑자기 철수는 자신의 패밀리카에 얼마의 중고가를 책정해야 합리적인지 고민이 들었다. 이에 여러분이 수업시간에 배운 <머신러닝 기술>을 통해 철수가 합리적인 중고가를 책정할 수 있도록 도와주길 바란다.

학습 데이터는 (차량 상태 정보)과 차량의 (중고 가격)를 제공합니다.

	ID	company	model	year	transmission	mileage	fueltype	tax	mpg	engineSize	price
0	0	4	Auris	2015	1	47541	4	145	46.3	1.6	8095
1	1	0	X2	2018	3	5000	0	145	50.4	2.0	21726
2	2	2	Focus	2017	3	47018	4	145	51.4	1.0	10490
3	3	2	Focus	2020	1	1550	4	145	49.6	1.0	17490
4	4	0	3 Series	2017	3	23505	0	145	64.2	2.0	18995
...

테스트 데이터는 (차량 상태 정보)만 제공하며, 예측된 차량의 (중고 가격)은 submit.csv 파일로 저장하여 캐글 리더보드에 제출하셔야 합니다.

	ID	company	model	year	transmission	mileage	fueltype	tax	mpg	engineSize
0	0	2	Ka+	2017	1	10150	4	145	57.7	1.2
1	1	2	Kuga	2013	1	50216	0	145	53.3	2.0
2	2	3	Q5	2016	1	34189	0	200	47.9	2.0
3	3	1	I10	2020	1	1900	4	145	56.5	1.0
4	4	0	3 Series	2014	0	79809	0	20	68.9	2.0
...

제공되는 (차량 상태 정보)는 순서대로 제조회사, 차량 모델명, 변속기 종류, 운영 거리, 연료 유형, 도로세, 갤런당 마일리지, 엔진 크기입니다. 참고로 차량 가격의 단위는 천원입니다.

평가 지표는 Mean Absolute Error (MAE)를 사용하세요.

자, 그럼 테스트 데이터로 주어진 (차량 상태 정보)에 맞는 차량의 (중고 가격)을 예측하여 봅시다. Let's start!!

※ 주의사항 ※

- ① 배운 범위 내의 기계학습 방법론만 사용합니다. (100% 감점)
- ② 반드시 학습 과정에서 범주형 데이터를 수치형 데이터로 변환하여 사용합니다. (50% 감점)
- ③ 상위 랭커 6인에게 추가 점수를 부여합니다. (25% 가점)
단, 공동 순위가 6인 이상일 경우 추가 점수를 부여하지 않음

[문제3][20점] <https://www.kaggle.com/t/ebc45c1a77554c90837997311ff538bf>

영희는 평소 별자리에 관심이 많아 입학과 동시에 별자리동아리에 가입했다. 작년 한해 망원경 다루는법도 배우고 별자리의 역사에서부터 계절별 별자리 위치까지 다양한 상식을 공부했다. 요즘은 동기들과 함께 은하계에 대한 스터디를 통해 은하는 크게 Spiral, Edge, Smooth로 나뉘지는 것을 알게 되었다. 그런데 영희는 문득 천체망원경으로 촬영된 별자리 사진을 자동으로 분석해주는 프로그램을 만들면 어떨까 라는 생각을 하게 되었고, 어떻게 하면 이 문제를 해결할 수 있는지 다양하게 고민 중이나 아직 특별한 방향을 찾지 못했다. 이번학기 기계학습을 이수하고 있는 여러분이 영희를 도와주었으면 한다.

학습 데이터는 (은하 영상을 기술한 정보)과 (은하 종류)를 제공합니다.

	0	1	2	3	4	5	6	7	8	9	...	55	56	57	58	59	60	61	62	63	Category
0	13.934082	-3.077109	-13.515020	-0.844135	3.097764	0.154635	5.615488	-8.032149	2.776431	1.469208	...	0.474123	-1.481585	-2.168011	-0.472213	1.542271	0.356707	0.530720	0.171594	-0.322960	edge
1	18.757893	-0.304454	1.757282	10.702372	5.530047	-9.571358	9.296905	-2.858612	0.027188	3.768508	...	1.776886	0.905478	-1.641744	0.734237	1.231368	0.179600	1.700099	-0.001509	1.077432	edge
2	-8.063476	-3.259291	-16.577340	-5.497738	-6.616893	4.916349	-10.537728	5.398572	-1.091936	-2.561823	...	-0.962800	-0.409117	0.497765	1.440391	-0.513098	-0.477530	2.724299	-0.663966	-1.755266	edge
3	3.128894	16.911400	-10.434631	1.177685	3.228336	-1.875124	8.374058	-0.885263	5.068379	-6.400661	...	0.775439	3.694294	0.438467	-0.148669	0.227370	3.004657	0.440074	-0.087064	1.925870	edge
4	-15.744248	-1.022630	0.004898	6.656867	-2.534440	-8.309071	-1.379099	4.352854	8.783724	2.650707	...	-1.151704	0.354604	0.752026	-1.192524	-0.910384	1.549595	-2.284629	0.672590	-0.606422	edge
...

테스트 데이터는 (은하 영상을 기술한 정보)만 제공하며, 예측된 (은하 종류)는 submit.csv 파일로 저장하여 캐글 리더보드에 제출하셔야 합니다.

	0	1	2	3	4	5	6	7	8	9	...	54	55	56	57	58	59	60	61	62	63
0	-14.238203	-15.670372	-12.266742	8.772731	-1.062115	12.313971	5.447355	-10.486055	-1.319069	1.412028	...	-1.526661	-0.031426	-0.351120	0.926840	-1.687954	-1.666352	-1.295853	-0.757767	-1.793229	-1.617771
1	-11.216002	15.657775	-1.080561	10.588281	1.698980	0.221580	0.651915	2.617677	-7.410492	2.398025	...	-2.074036	0.779153	0.752769	-1.249239	-0.982266	0.347240	-1.616450	0.859141	-0.279321	0.024584
2	19.227375	-13.398582	6.344983	0.673839	2.338009	-2.120843	5.539773	7.779192	1.380348	-1.728367	...	0.113740	-0.441069	1.053713	-1.845533	-1.311030	-1.594242	-0.743166	-0.533958	0.417801	-1.707941
3	-12.524920	9.557715	15.283616	1.440119	1.804742	-3.472384	2.337877	-2.322144	1.910832	-4.620938	...	-0.671643	0.078542	0.172880	-1.075137	0.725919	-1.553811	-0.787675	0.873334	-1.319784	-0.432190
4	15.702551	9.524783	-16.462688	-5.936708	-11.013749	2.756393	-7.259588	-3.768883	-3.980812	3.195737	...	-0.195311	1.171689	0.563732	-0.960047	-0.908781	0.250107	-0.578281	0.579105	-0.327169	0.806069
...

데이터로 제공되는 (은하 영상을 기술한 정보)는 인공지능망을 이용하여 2D 영상을 4096차원 벡터로 추출한 후, PCA(차원 축소 기술, 중간 이후 배움)를 이용하여 64차원 벡터로 차원 축소하였다. 이와 같이 차원을 축소하면 모델 학습에 걸리는 시간을 단축 할 수 있는 장점이 있다.

평가 지표는 Categorization Accuracy를 사용하세요.

자, 그럼 테스트 데이터로 주어진 (은하 영상을 기술한 정보)에 맞는 (은하 종류)를 예측하여 봅시다.

※ 주의사항 ※

- ① 배운 범위 내의 기계학습 방법론만 사용합니다. (100% 감점)
- ② 필요시 학습 과정에서 범주형 데이터를 수치형 데이터로 변환하여 사용합니다. (50% 감점)
- ③ 상위 랭커 6인에게 추가 점수를 부여합니다. (25% 가점)
단, 공동 순위가 6인 이상일 경우 추가 점수를 부여하지 않음

[문제4][20점] <https://www.kaggle.com/t/a56ffacf15d44c6b0de395a2b872c5b>

영철이는 MD-Ph.D를 취득한 의사로 최근 수면시간과 우울증의 상관관계를 연구 중이다. 일반적으로 우울증에 대한 진단은 정신건강의학과를 내원하여 PHQ-9과 같은 우울증 선별 검사지를 기반으로 이루어진다. 최근 이러한 선별 검사지를 통한 우울증 진단과정을 간소화 및 자동화하려는 노력들이 이루어지고 있으며, 스마트폰/웨어러블 디바이스 등을 이용하여 우울증을 진단하고 자살 등의 사회적 문제를 예방하려는 노력들이 이루어지고 있다. 영철이는 스마트폰을 사용하여 사용자들의 수면시간을 자동으로 측정하고, 수면 패턴을 기반으로 사용자의 우울증 여부를 진단하는 프로그램을 만들려고 한다.

학습 데이터는 스마트폰 사용자의(수면 패턴 정보)와 사용자의 (우울증 여부)를 제공합니다.

	sleep_time_mean	sleep_time_std	sleep_time_min	sleep_time_25	sleep_time_50	sleep_time_75	sleep_time_max	sleep_quality_mean	sleep_quality_std	sleep_quality_min	sleep_quality_25	sleep_quality_50	sleep_quality_75	sleep_quality_max	label
0	6.290000	0.689202	5.0	6.125	6.50	6.50	7.0	3.500000	0.836660	2.0	3.25	4.0	4.0	4.0	0
1	6.418182	1.169693	3.0	6.000	6.50	6.75	9.5	3.345455	0.672700	1.0	3.00	3.0	4.0	4.0	0
2	7.034483	1.127262	5.5	6.500	6.50	8.00	11.0	3.862069	0.347839	3.0	4.00	4.0	4.0	4.0	0
3	6.453846	1.070642	3.0	6.000	6.50	7.00	9.0	3.061538	0.768052	1.0	3.00	3.0	4.0	4.0	1
4	5.902439	1.817345	3.5	4.500	5.50	7.00	10.0	2.975610	0.961452	1.0	2.00	3.0	4.0	4.0	0
5	6.179487	1.205727	3.0	5.500	6.00	6.50	10.0	3.097561	0.916648	1.0	2.00	3.0	4.0	4.0	0
6	5.630495	0.741886	4.0	5.500	5.50	6.00	7.0	2.652174	0.884652	1.0	2.00	3.0	3.0	4.0	1

테스트 데이터는 스마트폰 사용자의(수면 패턴 정보)만 제공하며, 예측된 사용자의 (우울증 여부)는 submit.csv 파일로 저장하여 캐글 리더보드에 제출하여야 합니다.

	sleep_time_mean	sleep_time_std	sleep_time_min	sleep_time_25	sleep_time_50	sleep_time_75	sleep_time_max	sleep_quality_mean	sleep_quality_std	sleep_quality_min	sleep_quality_25	sleep_quality_50	sleep_quality_75	sleep_quality_max
0	6.574468	1.202257	3.0	6.000	6.50	7.000	10.0	2.659574	0.599877	1.0	2.0	3.0	3.0	4.0
1	6.233333	1.293758	3.5	4.250	5.00	5.750	8.5	2.375000	1.147461	1.0	1.0	2.5	3.0	4.0
2	5.980000	0.820484	4.5	5.500	6.00	6.500	7.0	3.560000	0.711805	2.0	3.0	4.0	4.0	4.0
3	5.735294	0.430562	5.0	5.500	5.50	6.000	6.5	2.941176	0.547234	2.0	3.0	3.0	3.0	4.0
4	6.423077	0.832974	5.0	6.000	6.50	6.500	8.0	3.384615	0.803837	2.0	3.0	4.0	4.0	4.0
5	5.857143	1.610406	3.0	5.125	5.75	6.375	10.0	3.428571	0.755929	2.0	3.0	4.0	4.0	4.0
6	6.281250	0.659148	4.5	6.000	6.50	6.500	7.5	3.031250	0.782237	1.0	3.0	3.0	3.0	4.0

제공되는 스마트폰 사용자의(수면 패턴 정보)는 sleep_time_mean, sleep_time_std, sleep_time_min, sleep_time_25, sleep_time_50, sleep_time_75, sleep_time_max, sleep_quality_mean, sleep_quality_std, sleep_quality_min, sleep_quality_25, sleep_quality_50, sleep_quality_75, sleep_quality_max 로 총 14개 제공된다.

평가 지표는 Categorization Accuracy를 사용하세요.

자, 그럼 테스트 데이터로 주어진 스마트폰 사용자의(수면 패턴 정보)에 맞는 사용자의 (우울증 여부)를 예측하여 봅시다.

※ 주의사항 ※

- ① 배운 범위 내의 기계학습 방법론만 사용합니다. (100% 감점)
- ② 필요시 학습 과정에서 범주형 데이터를 수치형 데이터로 변환하여 사용합니다. (50% 감점)
- ③ 상위 랭커 6인에게 추가 점수를 부여합니다. (25% 가점)
단, 공동 순위가 6인 이상일 경우 추가 점수를 부여하지 않음

대찬이와 정민이는 최근 따름이를 통해 정기적인 운동을 시작했다. 다양한 운동 중 따름이는 재미적인 요소가 많아 지속적으로 운동이 가능하다고 판단했기 때문이다. 그런데 최근 따름이 사용자가 부쩍 늘어 원하는 시간에 따름이를 빌리지 못하는 사태가 발생하곤 했다. 이에 둘은 운동에 대한 열정을 불태우고자 따름이 사용자 예측 프로그램을 제작하기로 결심하게 된다. 여러분이 수업시간에 배운 <머신러닝 기술>을 통해 대찬이와 정민이가 오늘의 따름이 사용자를 예측하고 이를 기반으로 따름이를 대여하여 운동하는 삶을 지속할 수 있도록 도와주길 바란다.

[illegible][illegible]

- ① 배운 범위 내의 기계학습 방법론만 사용합니다. (100% 감점)
- ② 필요시 학습 과정에서 범주형 데이터를 수치형 데이터로 변환하여 사용합니다. (50% 감점)
- ③ 상위 랭커 6인에게 추가 점수를 부여합니다. (25% 가점)

단, 공동 순위가 6인 이상일 경우 추가 점수를 부여하지 않음