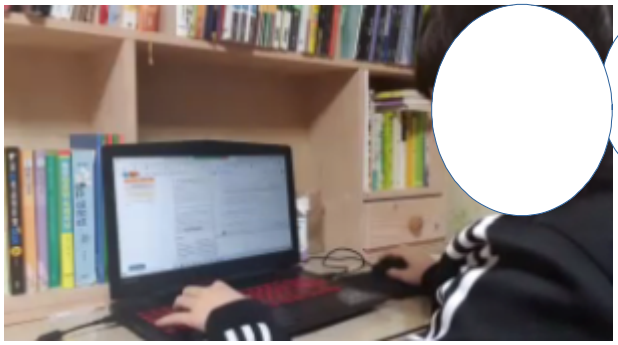
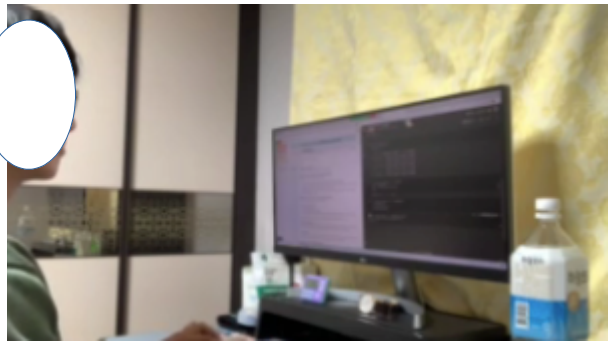
2022.기계학습.중간고사.시험지

[온라인 시험 바른 자세 예시]

아래와 같이 바른 자세 예시를 따르지 않는 학생은 퇴실 조치 합니다.

자세 검사를 마친 학생에 한하여 시험지의 비밀번호가 제공됩니다.

온라인 시험 주의 사항

- 시험 중 화장실에 다녀올 수 없음
- 화면 녹화 본 제출 필수 (대용량 이메일), 화면 녹화 중단시 녹화된 곳 까지만 점수 인정
- **화면 녹화는 반드시 7시부터 시작, 녹화 시작시 컴퓨터 하단에 그 어떠한 창도 실행되어 있으면 안됨**
- 코드 제출 필수(구글드라이브), 반드시 제출 코드로 **리더보드 재현**이 가능해야 인정
- OJ를 통한 IP 필터링을 통해 모여서 시험을 보는 행위 방지함, 이에 테더링 사용 불가
- **고사 중 부적절한 자세로 인한 경고 3회 이상 누적시 시험 0점 처리**
- 듀얼모니터 사용 불가, **시험 시작 후 듀얼 모니터 사용 중이 아님을 입증하는 화면 캡쳐 제출 필수**
- 시험은 오픈 메뉴얼 북 (메뉴얼 안에서 검색 가능)만 허용, 강의안과 책 불가, 인터넷 검색 불가
- 고시 시작 두 시간 이후 퇴실 가능하나 (퇴실 시 반드시 퇴실 하겠다고 알릴 것) 재 입실 불가
- 시험 중 캐글 리더보드 여전히 20회 제출 제한 있음
- 실시간 중계되는 감독 화면은 녹화되어 보관됨
- 고사 중 공지 및 질문을 위해서는 슬랙을 사용할 예정, 반드시 본인 PC에 슬랙 설치 필수
- 고사 중 카카오톡과 같은 메신저는 반드시 종료할 것. 적발시 F 처리
- 기존 코드를 참고하는 행위 절대 불가, 적발시 F
- 대리 시험 절대 불가, 적발시 학사경고

제출

- 화면 녹화 파일 제출: admin@rcv.sejong.ac.kr
  : 화면 녹화 본 중 듀얼모니터 부분 화면 캡쳐 제출: 동영상과 함께 이메일로 제출

이론 문제는 총 9문제로 중간고사 30점 만점 중 15점이 할당되어 있습니다.
아래의 링크에 접근하여 주관식과 객관식 답안을 온라인으로 작성 및 제출하면 됩니다.
https://forms.gle/hkcYBKEbggRrWT527

[문제1] 데이터 분석이란 무엇이고, 데이터 분석에 머신러닝이 도입된 이유를 설명하시오. (1점)

[문제2] 데이터와 정보의 정의를 각각 설명하시오. 단, 두 단어의 차별점이 들어나도록 명시적으로 작성하시오. (2점)

[문제3] MLOps의 6단계를 작성하고, 각 단계별로 가지는 과정을 설명하시오. (5점)

[문제4] 분류 문제와 회귀 문제가 각각 어떻게 다른지 설명하고, 수업시간 (시험포함) 에 배우지 않은 사례를 각각 제시하시오. (2점)

[문제5] KNN에 대한 설명 중 틀린 것을 고르시오. (1점)

[문제6] LDA와 QDA의 설명 중 잘못 된 것을 모두 고르시오. (1점)

[문제7] 의사결정 나무의 설명 중 잘못된 것을 고르시오. (1점)

[문제8] 앙상블의 개념 중 잘못된 것을 고르시오. (1점)

[문제9] 다음의 설명은 경사하강법 중 어떤 방법론에 해당하는가? (1점)

실습 문제는 총 5문제로 중간고사 30점 만점 중 15점이 할당되어 있습니다.
더불어 각 문제당 추가점수가 배당되어 있으니 참고하시기 바랍니다.

[문제1][2점] 심장 질환 환자 예측 https://www.kaggle.com/t/9f45feed2a6d4ba783baeb60fe2e62b4

심장 질환은 전 세계적으로 가장 큰 사망 원인 중 하나 이다. 심장병을 예방하는 것은 무척 중요하며, 심장병을 예측하기 위한 데이터 기반 시스템은 연구 및 예방 프로세스를 향상시켜 더 많은 사람들이 건강한 삶을 누를 수 있게 한다. 본 문제에서 제공되는 데이터 셋은 수년간 대중에게 공개된 심장병에 대한 연구에서 나온 것이며, 이 연구는 환자의 건강 및 심혈 관계 통계에 대한 다양한 측정을 수집하여 제공한다. 원본 데이터는 UCI Machine Learning 저장소를 통해 제공되고 있으며, 여러 데이터 셋 중 Cleveland Heart Disease 데이터를 가공하여 제공하였다. 원본 데이터는 총 76개의 속성 정보를 제공하고 있으나, 본 문제에서는 76개 중 14개의 속성과 해당 속성에 따른 심장 질환 여부를 학습용 데이터로 제공한다. 여러분은 지금부터 수업시간에 배운 〈머신러닝 기술〉을 활용하여, test.csv 파일로 제공된 환자들의 심장 질환 여부를 판단하는 인공지능 SW를 작성해주길 바란다. 주의사항으로, 의료 데이터는 환자 개인정보와 관련되어 있어 대용량 데이터 확보에 어려움이 많다. 이에 제공되는 학습 및 테스트 데이터의 샘플이 매우 적다는 점을 유념하기 바란다.

학습 데이터는 (환자의 건강 정보)와 (환자의 심장 질환 정보)를 제공한다.

index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
5	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

테스트 데이터는 (환자의 건강 정보)만 제공하며, 인공지능 SW를 통해 예측된 (환자의 심장 질환 유무)는 submit.csv 파일로 저장하여 캐글 리더보드에 제출해야 한다.

index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3
6	62	0	4	140	268	0	2	160	0	3.6	3	2	3
21	58	0	1	150	283	1	2	162	0	1.0	1	0	3
24	60	1	4	130	206	0	2	132	1	2.4	2	2	7
31	60	1	4	117	230	1	0	160	1	1.4	1	2	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...

(환자의 건강 정보) 중 환자고유ID를 제외한 13가지 속성이 가지는 내용은 다음과 같다. age 는 나이, sex 는 성별(0-여성, 1-남성), cp 는 흉통 유형(1~4), threstbps는 휴식 혈압, chol은 혈청콜레스테롤, fbs는 공복혈당 〉 120mg/dl 여부, restecg 는 휴식 심전도 결과 (0~2), thalach 는 최대 심박수, exang 은 운동으로 인한 흉통 여부(0-없음, 1-있음), oldpeak 은 심전도의 비정상 측정, slope 는 최대 운동 ST 세그먼트의 기울기 (1~2), ca 는 형광 투시로 채색된 주요 혈관의 수 (0~3), thal 은 심장에 혈류를 측정하는 Thallium stress test 결과(3-normal, 6-fixed defect, 7-reversible_defect)를 의미 한다.

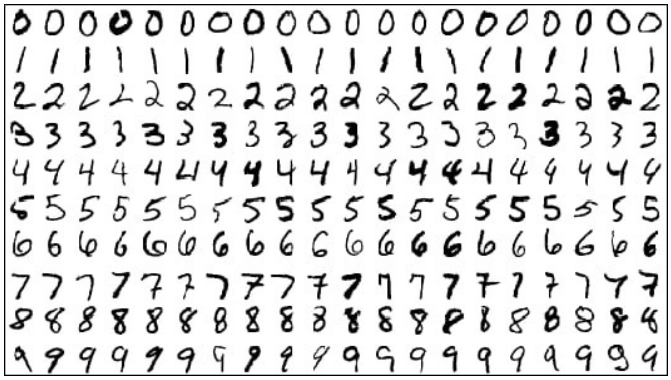
(환자의 심장 질환 여부)는 0 - 심장질환 없음, 1,2,3,4 - 심장질환의 카테고리을 의미 한다.
자, 그럼 테스트 데이터로 주어진 (환자의 건강 정보)에 맞는 (환자의 심장 질환 유무)를 예측하여 보자.

>> 심장 질환 예측 문제에 사용되는 데이터의 수가 적다는 점을 유의하기 바란다.
>> 베이스라인은 1개이며, 주어진 score 이상으로 제출해야 점수를 부여 받는다.

※ **주의사항** ※
① 모듈별 코드에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다.
(미 제출시 0.5점 감점)
② 배운 범위 내의 기계학습 방법론 2가지 이상을 사용하여 비교 실험 후 베스트 모델을 선정해주세요.
(**배운 범위 내가 아니면 0점, 2가지 이상의 비교가 없을 경우 0.5점 감점**)
③ 각 문제 제출 코드 마지막에 텍스트 셀을 추가하여 실험 결과 분석 내용을 서술형으로 작성해주세요.
(**미 제출시 0.5점 감점, 설명이 부족하고 논리적이지 않으면 부분 감점**)

[문제2][2점] 손글씨 숫자 인식 https://www.kaggle.com/t/227ddd9a7f1a401996449175b2df19d9

철수는 신용카드회사 IT팀에서 근무 중이다. 최근 고객들이 신용카드 신청시 작성하는 문서를 자동으로 전산화 하는 업무가 내려왔다. 신청서에는 고객들이 수기로 작성한 개인정보가 들어있는데, 크게 이름, 연락처, 주소, 이메일 등이 존재한다. 우선 철수는 프로토 타이핑용으로 고객의 연락처(핸드폰 번호)를 자동화 해보려고 한다. 고객들이 작성한 문서를 스캔하고, 스캔 문서 중 연락처가 작성된 표 박스를 크랍하여 데이터를 다음과 같이 모았다. 여러분은 지금부터 수업시간에 배운 <머신러닝 기술>을 활용하여, test.csv 파일로 제공된 고객의 손글씨를 인식하는 인공지능 SW를 작성해주길 바란다.



학습 데이터로는 고객의 (손글씨 영상 정보)와 고객의 (손글씨 정답) 함께 제공됩니다.

index	0	1	2	3	4	5	6	7	8	9	...	55	56	57	58	59	60	61	62	63	label
0	0	0	5	13	9	1	0	0	0	0	...	0	0	0	6	13	10	0	0	0	0
1	0	0	0	12	13	5	0	0	0	0	...	0	0	0	0	11	16	10	0	0	1
2	0	0	0	4	15	12	0	0	0	0	...	0	0	0	0	3	11	16	9	0	2
3	0	0	7	15	13	1	0	0	0	8	...	0	0	0	7	13	13	9	0	0	3
4	0	0	0	1	11	0	0	0	0	0	...	0	0	0	0	2	16	4	0	0	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

테스트 데이터로는 고객의 (손글씨 영상 정보)만 제공되며, 예측된 고객의 (손글씨 정답)은 submit.csv 파일로 저장하여 캐글 리더보드에 제출하셔야 합니다.

index	0	1	2	3	4	5	6	7	8	9	...	54	55	56	57	58	59	60	61	62	63
23	0	1	8	12	15	14	4	0	0	3	...	0	0	0	0	14	15	11	2	0	0
29	0	0	9	13	7	0	0	0	0	0	...	16	2	0	0	7	12	12	12	11	0
30	0	0	10	14	11	3	0	0	0	4	...	0	0	0	0	11	16	12	3	0	0
32	0	2	13	16	16	16	11	0	0	5	...	0	0	0	2	16	15	8	0	0	0
44	0	0	9	16	16	16	5	0	0	1	...	0	0	0	0	13	10	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

제공되는 고객의 (손글씨 정보)는 8x8 영상으로 크기를 통일하였고, [0, 255] 사이의 밝기 값을 가지는 흑백 영상의 값을 [0, 16]으로 정규화 하였다.

자, 그럼 테스트 데이터로 주어진 (손글씨 영상 정보)를 가지고 고객의 (손 글씨 정답)을 예측하여 봅시다.
>> 베이스라인은 1개이며, 주어진 score 이상으로 제출해야 점수를 부여 받는다.

※ **주의사항** ※

① 모듈별 코드에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다.
(미 제출시 0.5점 감점)
② 배운 범위 내의 기계학습 방법론 2가지 이상을 사용하여 비교 실험 후 베스트 모델을 선정해주세요.
(**배운 범위 내가 아니면 0점, 2가지 이상의 비교가 없을 경우 0.5점 감점**)
③ 각 문제 제출 코드 마지막에 텍스트 셀을 추가하여 실험 결과 분석 내용을 서술형으로 작성해주세요.
(**미 제출시 0.5점 감점, 설명이 부족하고 논리적이지 않으면 부분 감점**)

[문제3][3점] 보스턴 땅값 예측문제 https://www.kaggle.com/t/d1b847a820024d27b81e3d685bd6f345

영희는 보스턴에서 부동산중개업자로 일하기 시작한 초보 중개업자이다. 종종 집을 팔기 희망하는 고객이 적절한 판매가를 질문하곤 하지만 초보 중개업자인 관계로 이를 대답하기 쉽지 않을 때가 많다. 마침 영희 남편 철수는 IT회사 데이터 분석가로 일하고 있는 터라, 영희는 남편 철수에게 적절한 주택가격을 알려주는 프로그램을 만들어 달라고 부탁을 하려 한다. 여러분 역시 수업시간에 배운 <머신러닝 기술>을 활용하여, test.csv 파일로 제공된 보스턴 주택가격을 예측해주는 인공지능 SW를 작성해 주기 바란다.

학습 데이터는

index	a	b	c	d	e	f	g	h	i	j	k	l	m	price
0	0.14150	0.0	6.91	0	0.448	6.169	6.6	5.7209	3	233	17.9	383.37	5.81	25.3
1	0.15445	25.0	5.13	0	0.453	6.145	29.2	7.8148	8	284	19.7	390.68	6.86	23.3
2	16.81180	0.0	18.10	0	0.700	5.277	98.1	1.4261	24	666	20.2	396.90	30.81	7.2
3	0.05646	0.0	12.83	0	0.437	6.232	53.7	5.0141	5	398	18.7	386.40	12.34	21.2
4	8.79212	0.0	18.10	0	0.584	5.565	70.6	2.0635	24	666	20.2	3.65	17.16	11.7

테스트 데이터는

index	a	b	c	d	e	f	g	h	i	j	k	l	m
0	0.04932	33.0	2.18	0	0.472	6.849	70.3	3.1827	7	222	18.4	396.90	7.53
1	0.02543	55.0	3.78	0	0.484	6.696	56.4	5.7321	5	370	17.6	396.90	7.18
2	0.22927	0.0	6.91	0	0.448	6.030	85.5	5.6894	3	233	17.9	392.74	18.80
3	0.05789	12.5	6.07	0	0.409	5.878	21.4	6.4980	4	345	18.9	396.21	8.10
4	3.67822	0.0	18.10	0	0.770	5.362	96.2	2.1036	24	666	20.2	380.79	10.19

데이터로 제공되는 (주택 정보)는 순차적(a=>m)으로 자치시별 1인당 범죄율, 비소매상업지역이 점유하고 있는 토지의 비율, 10ppm당 농출 일산화질소 농도, 주택 1가구당 평균 방의 개수, 자치시별 흑인의 비율, 자치시별 학생/교사 비율, 25000 평방 피트를 초과하는 거주지역의 비율, 찰수강 근접 여부(1-경계위치, 0-아닌 경우), 1940년 이전에 건축된 주택 비율, 방사형 고속도로까지의 접근성 지수, 5개의 보스턴 직업센터까지의 접근성 지수, 1000달러 당 재산 세율 이다.

자, 그럼 테스트 데이터로 주어진 (주택 정보)에 맞는 (주택 가격)를 예측하여 보자.
>> 베이스라인은 1개이며, 주어진 score 이상으로 제출해야 점수를 부여 받는다.

※ **주의사항** ※
① 모듈별 코드에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다.
(미 제출시 0.5점 감점)
② 배운 범위 내의 기계학습 방법론 2가지 이상을 사용하여 비교 실험 후 베스트 모델을 선정해주세요.
(**배운 범위 내가 아니면 0점, 2가지 이상의 비교가 없을 경우 0.5점 감점**)
③ 각 문제 제출 코드 마지막에 텍스트 셀을 추가하여 실험 결과 분석 내용을 서술형으로 작성해주세요.
(**미 제출시 0.5점 감점, 설명이 부족하고 논리적이지 않으면 부분 감점**)

[문제4][4점] 자동차가격 예측문제 https://www.kaggle.com/t/a3461fa660d54d8a80afb5396545cdee

철수는 초보 중고차매매업자이다. 종종 자동차를 팔기 희망하는 고객이 적절한 판매가를 질문하곤 하지만 초보 매매업자인 관계로 이를 대답하기 쉽지 않을 때가 많다. 마침 철수 친구 미영이는 IT회사 데이터 분석가로 일하고 있는 터라, 적절한 중고차가격을 알려주는 프로그램을 만들어 달라고 부탁을 하려 한다. 여러분 역시 수업시간에 배운 <머신러닝 기술>을 활용하여, test.csv 파일로 제공된 중고자 자동차가격을 예측해주는 인공지능 SW를 작성해 주기 바란다.

학습 데이터는

	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

	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

데이터로 제공되는 (중고자 정보)는 순차적으로 company 자동차 제조회사, model 차량의 제품명, year 해당 차량의 제조년도, transmission 해당 차량의 변속기, mileage 해당 차량의 마일리지, fueltype 해당 차량의 연료 유형, tax 해당 차량의 세금, mpg 해당 차량의 마일당 갤런 사용량(연비), enginesize해당 차량의 엔진 크기, price 마지막으로 해당 차량의 가격을 의미합니다.

자, 그럼 테스트 데이터로 주어진 (중고자 정보)에 맞는 (중고차 가격)를 예측하여 보자.
>> 베이스라인은 1개이며, 주어진 score 이상으로 제출해야 점수를 부여 받는다.

※ **주의사항** ※
① 모듈별 코드에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다. **(미 제출시 0.5점 감점)**
② 배운 범위 내의 기계학습 방법론 2가지 이상을 사용하여 비교 실험 후 베스트 모델을 선정해주세요. (**배운 범위 내가 아니면 0점, 2가지 이상의 비교가 없을 경우 0.5점 감점**)
③ 각 문제 제출 코드 마지막에 텍스트 셀을 추가하여 실험 결과 분석 내용을 서술형으로 작성해주세요. (**미 제출시 0.5점 감점, 설명이 부족하고 논리적이지 않으면 부분 감점**)

[문제5][5점] 행동 분류 문제 https://www.kaggle.com/t/ca6984762db841fe8944f4eec2d48ac1

본 문제는 스마트폰에 장착된 자이로센서와 가속도센서를 통해 사람의 행동을 예측하는 문제입니다. 자세한 설명과 스켈레톤 코드는 캐글 리더보드 내 Overview를 참고하세요.

>> 베이스라인은 1개이며, 주어진 score 이상으로 제출해야 점수를 부여 받는다.

※ **주의사항** ※

① 모듈별 코드에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다. **(미 제출시 0.5점 감점)**
② 배운 범위 내의 기계학습 방법론 2가지 이상을 사용하여 비교 실험 후 베스트 모델을 선정해주세요. (**배운 범위 내가 아니면 0점, 2가지 이상의 비교가 없을 경우 0.5점 감점**)
③ 각 문제 제출 코드 마지막에 텍스트 셀을 추가하여 실험 결과 분석 내용을 서술형으로 작성해주세요. (**미 제출시 0.5점 감점, 설명이 부족하고 논리적이지 않으면 부분 감점**)