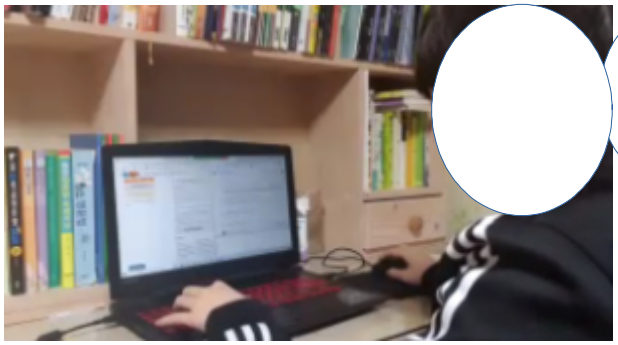
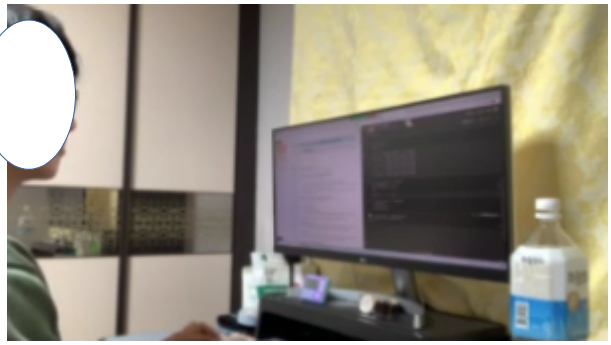
2021.기계학습.중간고사.시험지 - 2021년 4월 19일

[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
- 코드 및 기타 파일
  : 공유된 구글 드라이브

이론 문제는 총 11문제로 중간고사 30점 만점 중 15점이 할당되어 있습니다.
아래의 링크에 접근하여 주관식과 객관식 답안을 온라인으로 작성 및 제출하면 됩니다.
https://forms.gle/2LyoyHijkPYiQyar6

[문제1][2점] 최근 SW 패러다임이 변화하고 있다. 소프트웨어 1.0과 소프트웨어 2.0 시대의 철학을 설명하라.

[문제2][2점] 빅데이터 시대, 왜 데이터가 중요한지 설명하라.

[문제3][2점] 데이터 시각화란 무엇이고 데이터 시각화의 목적은 무엇인지 설명하라.

[문제4][2점] 공공데이터를 이용하여 기계학습 기반 문제해결을 하고자 한다. 본인이 이해하고 있는 ML생애주기를 기반으로 직접 풀고 싶은 문제를 정의하고, 해당 문제를 해결하기 위해 필요한 데이터를 정의하고, 데이터를 어떻게 분석하여 문제를 해결 할 것인지 시뮬레이션 하라.

[문제5][1점] 다음 중 분류 문제에 해당하는 것을 모두 고르시오.

[문제6][1점] 다음 중 데이터 전처리 과정이 아닌 것을 고르세요.

[문제7][1점] 데이터 불균형에 대한 설명 중 잘못된 것은?

[문제8][1점] 선형판별분석 설명중 옳지 않은 것은?

[문제9][1점] KNN에 대한 설명 중 틀린 것을 고르시오

[문제10][1점] 다음의 설명은 경사하강법 중 어떤 방법론에 해당하는가? "파라미터를 업데이터 할 때, 무작위로 샘플링된 학습 데이터를 하나씩만 이용하여 cost function의 gradient를 계산하고, 모델을 자주 업데이트 하며 성능 개선 정도를 빠르게 확인 가능하다. 최소 cost에 수렴했는지의 판단이 상대적으로 어렵다"

[문제11][1점] LDA와 QDA의 설명 중 잘못 된 것을 모두 고르시오.

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

[문제1][3점] 심장 질환 환자 예측 https://www.kaggle.com/t/ec8aaccd1112404fa22797db4f8a11e0

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 2개이며, 각 score 이상으로 받아야 점수를 부여 받는다.

※ **주의사항** ※
① 모듈별 코드(데이터전처리, 모델학습, 모델검증 등) 에 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**)
④ 상위 랭커 6인에게 추가 점수를 부여합니다. 단, 공동 순위가 6인 이상일 경우 점수를 부여하지 않으며, 또한 풀이방식이 문제가 있다면 상위 랭커 6인 안에 들어가더라도 가점이 부여되지 않습니다. (**1점 가점**)
⑤ 데이터 전처리는 scikit-learn에 있는 다양한 함수를 모두 사용하셔도 좋습니다.

[문제2][3점] 통신사 고객 이탈 예측 https://www.kaggle.com/t/92714abea4f64f77aef62eafff9b280a

통신 회사의 경우 신규 고객을 유치하는 동시에 수익 창출 기반을 늘리기 위해 계약 해지(=이탈)을 피하는 것이 중요합니다. 신규 고객이 이탈하는 이유를 살펴보면 더 나은 가격, 더 흥미로운 패키지, 불편한 서비스 경험 또는 고객의 개인적인 상황 변화와 같이 다양한 이유로 고객이 계약을 종료하게 됩니다. 고객 이탈 분석은 고객 이탈을 예측하고 이탈을 유발하는 근본적인 이유를 정의하는 기능을 제공합니다. 통신사는 기계 학습 모델을 적용하여 개별 고객을 기준으로 이탈을 예측하고 할인, 특별제안 또는 기타 만족을 주기 위한 대응 조치를 취하여 고객을 유지할 수 있습니다. 여러분은 지금부터 수업시간에 배운 <머신러닝 기술>을 활용하여, test.csv 파일로 제공된 통신사 고객들의 이탈 여부를 예측하는 인공지능 SW를 작성해주길 바란다.

학습 데이터로는 (통신사 고객 정보)과 해당 고객들의 (최종 이탈 여부)를 함께 제공합니다.

index	Unnamed: 0	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	1869	7010-BRBUU	Male	0	Yes	Yes	72	Yes	Yes	No	...	No internet service	No internet service	No internet service	No internet service	Two year	No	Credit card (automatic)	24.10	1734.65	No
1	4528	9688-YGXVR	Female	0	No	No	44	Yes	No	Fiber optic	...	Yes	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	88.15	3973.2	No
2	6344	9286-DOJGF	Female	1	Yes	No	38	Yes	Yes	Fiber optic	...	No	No	No	No	Month-to-month	Yes	Bank transfer (automatic)	74.95	2869.85	Yes
3	6739	6994-KERXL	Male	0	No	No	4	Yes	No	DSL	...	No	No	No	Yes	Month-to-month	Yes	Electronic check	55.90	238.5	No
4	432	2181-UAESM	Male	0	No	No	2	Yes	No	DSL	...	Yes	No	No	No	Month-to-month	No	Electronic check	53.45	119.5	No
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

테스트 데이터로는 (통신사 고객 정보)만 제공하며, 예측된 고객의 (이탈 여부)은 submit.csv 파일로 저장하여 캐글 리더보드에 제출하셔야 합니다.

index	Unnamed: 0	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	...	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
5	2215	4312-GVYNH	Female	0	Yes	No	70	No	No phone service	DSL	...	No	Yes	Yes	No	Yes	Two year	Yes	Bank transfer (automatic)	49.85	3370.20
10	3169	4578-PHJYZ	Male	0	Yes	Yes	52	Yes	No	DSL	...	Yes	Yes	Yes	Yes	No	One year	Yes	Electronic check	68.75	3482.85
13	1760	2511-MORQY	Male	0	Yes	Yes	50	Yes	Yes	DSL	...	No	Yes	No	No	No	One year	No	Bank transfer (automatic)	54.90	2614.10
18	6508	8708-XPXHZ	Female	0	Yes	Yes	42	Yes	Yes	Fiber optic	...	No	No	No	Yes	Yes	Month-to-month	Yes	Electronic check	94.20	4186.30
20	4693	0463-TXOAK	Male	0	No	Yes	52	Yes	Yes	No	...	No internet service	No internet service	No internet service	No internet service	No internet service	Two year	No	Credit card (automatic)	25.60	1334.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

제공되는 (통신사 고객 정보)는 순서대로 custumerID, gender, SeniorCitizen[고령자], Partner, Dependents[부양가족], tenure[계약유지기간], PhoneService, MultipleLines[여러가인가입], InternetService[인터넷공급자], OnlineSecurity[온라인보안], OnlineBackup, DeviceProtection, TechSupport[기술지원, StreamingTV[TV스트리밍], StreamingMovies[영화스트리밍], Contract[계약형태 -Month to Month, One year, Two year], PaperlessBilling, PaymentMethod, MonthlyCharges[월청구액], TotalCharges[전체기간청구액], Churn[이탈여부]입니다.

자, 그럼 테스트 데이터로 주어진 (통신사 고객 정보)를 가지고 고객의 (이탈 여부)를 예측하여 봅시다.
>> 베이스라인은 2개이며, 각 score 이상으로 받아야 점수를 부여 받는다.

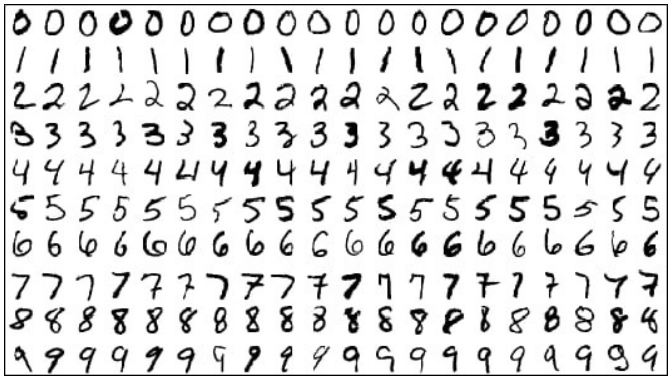
※ **주의사항** ※
① 모듈별 코드(데이터전처리, 모델학습, 모델검증 등) 에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다. **(미 제출시 0.5점 감점)**
② 배운 범위 내의 기계학습 방법론 2가지 이상을 사용하여 비교 실험 후 베스트 모델을 선정해주세요. (**배운 범위 내가 아니면 0점, 2가지 이상의 비교가 없을 경우 0.5점 감점**)
③ 각 문제 제출 코드 마지막에 텍스트 셀을 추가하여 실험 결과 분석 내용을 서술형으로 작성해주세요. (**미 제출시 0.5점 감점, 설명이 부족하고 논리적이지 않으면 부분 감점**)
④ 상위 랭커 6인에게 추가 점수를 부여합니다. 단, 공동 순위가 6인 이상일 경우 점수를 부여하지 않으며, 또한 풀이방식이 문제가 있다면 상위 랭커 6인 안에 들어가더라도 가점이 부여되지 않습니다.

(**1점 가점**)

⑤ 데이터 전처리는 scikit-learn에 있는 다양한 함수를 모두 사용하셔도 좋습니다.

[문제3][3점] 손글씨 숫자 인식 https://www.kaggle.com/t/512ae811743946da94ea115d1cc4b446

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 2개이며, 각 score 이상으로 받아야 점수를 부여 받는다.

※ **주의사항** ※

① 모듈별 코드(데이터전처리, 모델학습, 모델검증 등) 에 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**)
④ 상위 랭커 6인에게 추가 점수를 부여합니다. 단, 공동 순위가 6인 이상일 경우 점수를 부여하지 않으며, 또한 풀이방식이 문제가 있다면 상위 랭커 6인 안에 들어가더라도 가점이 부여되지 않습니다. (**1점 가점**)
⑤ 데이터 전처리는 scikit-learn에 있는 다양한 함수를 모두 사용하셔도 좋습니다.

[문제4][3점] https://www.kaggle.com/t/e9572ae7284d41ed8e3e65fe700de0da

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

	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	...	1.776686	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

	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 (주택 정보)는 순차적으로 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 2개이며, 각 score 이상으로 받아야 점수를 부여 받는다.

※ **주의사항** ※
① 모듈별 코드(데이터전처리, 모델학습, 모델검증 등) 에 주석을 반드시 달아주세요. 특히, 어떤 목적으로 API를 호출했는지 작성되어야 합니다. **(미 제출시 0.5점 감점)**
② 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**)
③ 상위 랭커 6인에게 추가 점수를 부여합니다. 단, 공동 순위가 6인 이상일 경우 점수를 부여하지 않으며, 또한 풀이방식이 문제가 있다면 상위 랭커 6인 안에 들어가더라도 가점이 부여되지 않습니다. (**1점 가점**)
④ 데이터 전처리는 scikit-learn에 있는 다양한 함수를 모두 사용하셔도 좋습니다.

[문제5][3점] https://www.kaggle.com/t/483b5b1f9dfb4daaba57ee04e32a95eb

최근 N사는 빅데이터분야와 데이터사이언스분야를 활용한 서비스가 많아지면서 해당 분야 개발자가 많이 필요해졌다. 이에 외부에서 직원을 추가 채용하는 부분을 고려하기 전 사내 부서 이동 제도를 통해 빅데이터분야와 데이터사이언스분야로 옮기고 싶어하는 직원들의 신청을 받기로 했다. 그러나 HR(인사팀)에서 근무하는 솔잎 양은 해당 직원들의 정보가 인공지능관련 부서로의 이동뿐만 아니라, 현 부서의 만족도가 낮아(분야, 임금, 동료, 기타 등등) 이직을 고려중인 직원으로 분류 가능하다는 분석결과를 도출하고 사내 직원들 중 현재 직장을 그만두고 새로운 일자리를 알아보는 직원을 예측하는 소프트웨어를 만들어 보려한다. 여러분 역시 수업시간에 배운 <머신러닝 기술>을 활용하여, test.csv 파일로 제공된 이직하기 희망하는 직원을 예측해주는 인공지능 SW를 작성해 주기 바란다.

학습 데이터는 (직원의 개인 정보)와 직원의 (이직 희망 여부)를 제공합니다.

index	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new_job	training_hours	target
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	NaN	NaN	1	36	1.0
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>4	47	0.0
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree	<1	NaN	Pvt Ltd	never	52	1.0
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99	Funded Startup	4	8	0.0
5	21651	city_176	0.764	NaN	Has relevent experience	Part time course	Graduate	STEM	11	NaN	NaN	1	24	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

테스트 데이터는 (직원의 개인 정보)만 제공하며, 예측된 직원의(이직 희망 여부)는 submit.csv 파일로 저장하여 캐글 리더보드에 제출하셔야 합니다.

index	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new_job	training_hours
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN	NaN	never	83
9	699	city_103	0.920	NaN	Has relevent experience	no_enrollment	Graduate	STEM	17	10000+	Pvt Ltd	>4	123
10	29452	city_21	0.624	NaN	No relevent experience	Full time course	High School	NaN	2	NaN	NaN	never	32
11	23853	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	5	5000-9999	Pvt Ltd	1	108
15	6588	city_114	0.926	Male	Has relevent experience	no_enrollment	Graduate	STEM	16	10/49	Pvt Ltd	>4	18
...	...	...	...	...	...	...	...	...	...	...	...	...	...

제공되는 (직원의 개인 정보)는 순서대로 직원의 고유 ID, 도시 코드, 도시 개발 지수, 직원의 성별, 직원의 데이터사이언스분야 관련 경험, 현 대학등록여부(풀타임, 파트타임, 없음), 학위(고졸, 대졸 등), 직원의 경력, 현 회사의 직원수, 고용 유형, 이전 회사 입사 년도, 이수한 교육 시간, 이직 희망 여부 이다.

자, 그럼 테스트 데이터로 주어진 (직원의 개인 정보)에 맞는 직원의 (이직 희망 여부)를 예측하여 보자.

※ **주의사항** ※

① 모듈별 코드(데이터전처리, 모델학습, 모델검증 등) 에 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**)
④ 상위 랭커 6인에게 추가 점수를 부여합니다. 단, 공동 순위가 6인 이상일 경우 점수를 부여하지 않으며, 또한 풀이방식이 문제가 있다면 상위 랭커 6인 안에 들어가더라도 가점이 부여되지 않습니다. (**1점 가점**)
⑤ 데이터 전처리는 scikit-learn에 있는 다양한 함수를 모두 사용하셔도 좋습니다.