숭실대학교 Soongsil University

오렌지3를 이용한 알고리즘에 따른 예측 결과 비교

숭실대학교 베어드교양대학
서유화 교수
yhsuh@ssu.ac.kr

목차

학습 목표

- 오렌지3를 이용해 머신러닝을 이용해 예측할 수 있다.
- 다양한 알고리즘에 따른 예측 결과를 비교할 수 있다.
- 머신러닝 알고리즘의 성능지표를 이해할 수 있다.

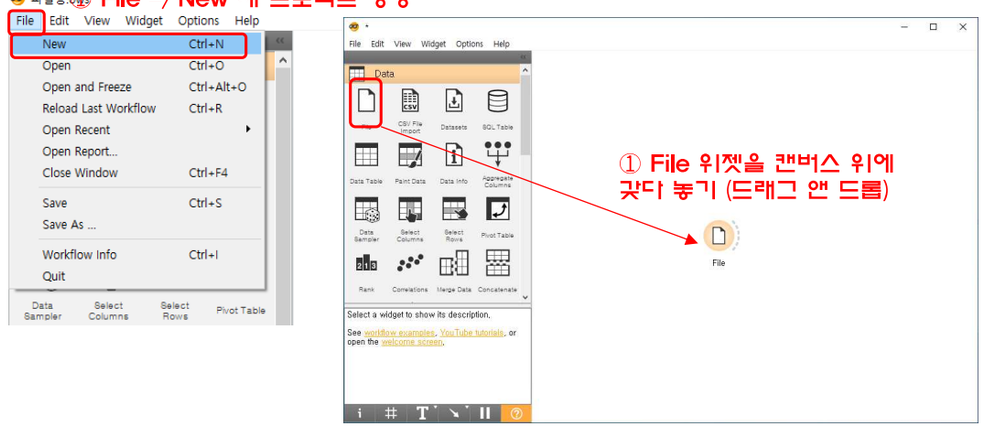
지난시간 배운 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티쳐블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘과 모델링의 개요 2. 알고리즘 기반 의사결정 시스템의 한계 3. 윤리가 필요한 인공지능 4. 오렌지3 설치 및 사용법	1. 오렌지3를 이용한 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	

머신러닝(지도학습)을 이용한 보스톤 집값 예측

학습할 데이터 불러오기

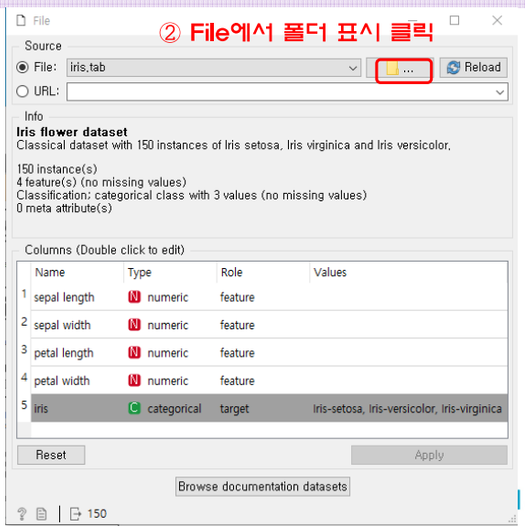
① File -> New 새 프로젝트 생성



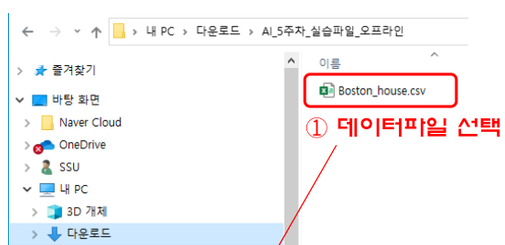
학습할 데이터 불러오기

① 캔버스 위의 File 위젯 더블클릭

② File에서 폴더 표시 클릭

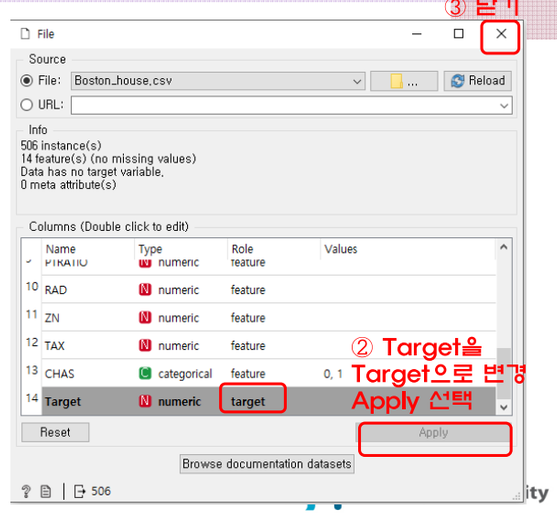


학습할 데이터 불러오기

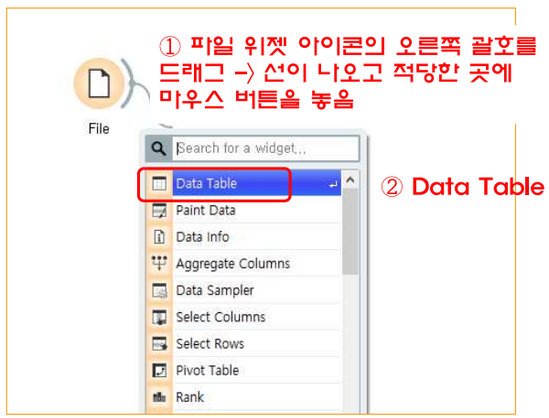


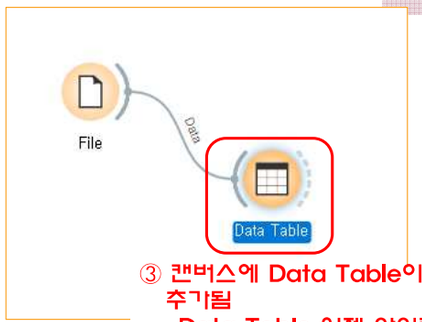
Boston_house.csv 데이터

③ 닫기

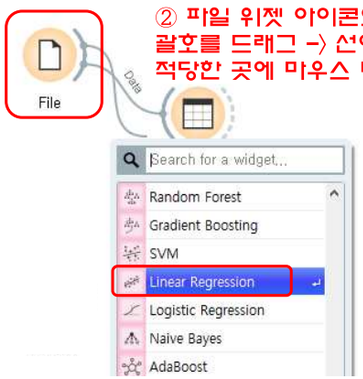


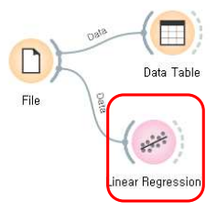
학습할 데이터 불러오기



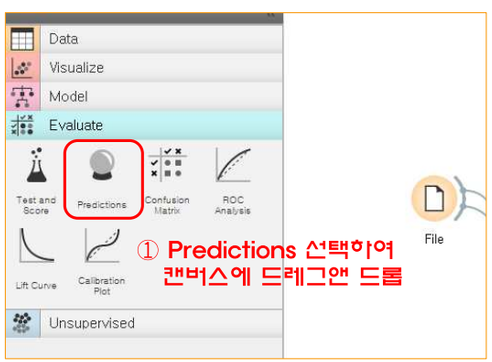


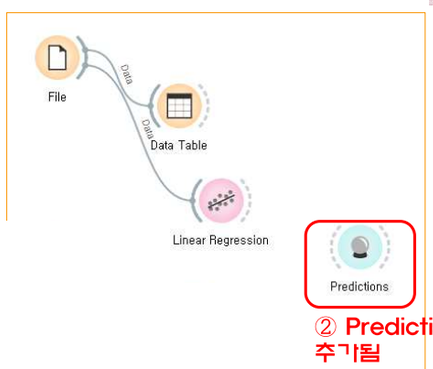
선형회귀 모델 만들기



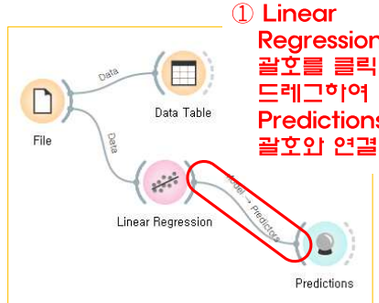


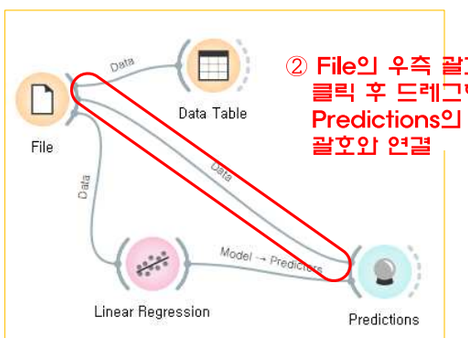
선형회귀로 집값 예측하기





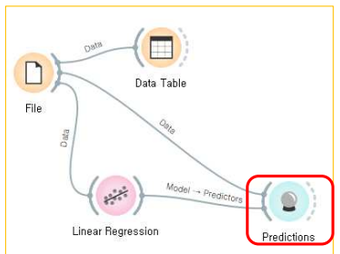
선형회귀로 집값 예측하기





- Prediction에는 두 가지 데이터가 필요함
  - 첫번째는 모델, 두번째는 예측하고 싶은 원인데이터
  - 별도의 파일에 저장된 데이터를 가져올 수도 있으나, 테스트 목적임으로 File 위젯 데이터 그대로 연결

선형회귀로 집값 예측하기



① Predictions 더블클릭

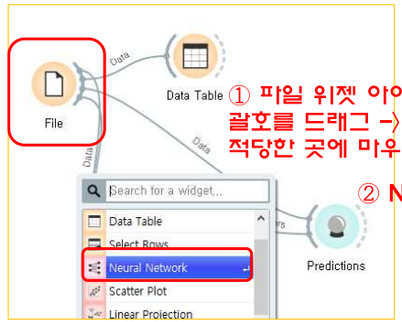
② 정답

	Linear Regression	Target	AGE	B	RM
1	30.0	24.0	65.2	396.90	6.575
2	25.0	21.6	78.9	396.90	6.421
3	30.6	34.7	61.1	392.83	7.185
4	28.6	33.4	45.8	394.63	6.998
5	27.9	36.2	54.2	396.90	7.147
6	25.3	28.7	58.7	394.12	6.430
7	23.0	22.9	66.6	395.60	6.012
8	19.5	27.1	96.1	396.90	6.172
9	11.5	16.5	100.0	386.63	5.631
10	18.9	18.9	85.9	386.71	6.004
11	19.0	15.0	94.3	392.52	6.377
12	21.6	18.9	82.9	396.90	6.009
13	20.9	21.7	39.0	390.50	5.889
14	19.6	20.4	61.8	396.90	5.949
15	19.3	18.2	84.5	380.02	6.096
16	19.3	19.9	56.5	395.62	5.834

Model	MSE	RMSE	MAE	R2
Linear Regression	21.895	4.679	3.271	0.741

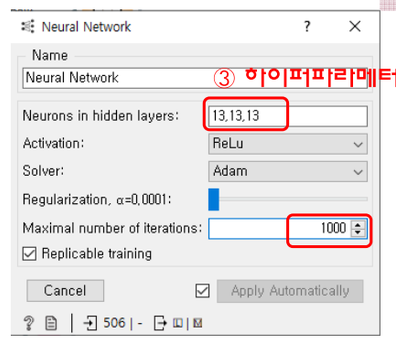
② Linear Regression으로 예측한 결과

다른 모델과 비교하기



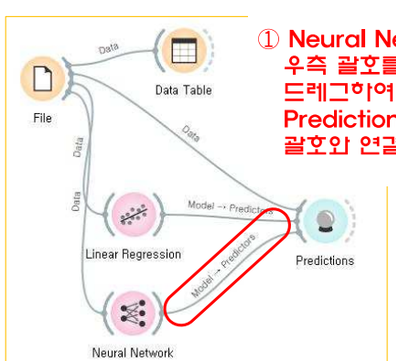
① 파일 위젯 아이콘의 오른쪽 괄호를 드래그 -> 선이 나오고 적당한 곳에 마우스 버튼을 놓음

② Neural Network선택

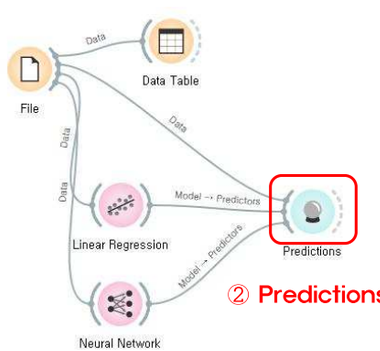


③ 하이퍼파라메터 설정

다른 모델과 비교하기



① Neural Network의 우측 괄호를 클릭 후 드레그하여 Predictions의 좌측 괄호와 연결



② Predictions 더블클릭

다른 모델과 비교하기

	Linear Regression	Neural Network	Target
1	30.0	24.7	30.1
2	25.0	23.9	20.3
3	30.6	31.6	22.0
4	28.6	31.5	21.4
5	27.9	32.6	18.8
6	25.3	25.8	20.5
7	23.0	19.3	17.3
8	19.5	19.4	15.7
9	11.5	14.5	29.6
10	18.9	18.4	26.4
11	19.0	20.5	50.0
12	21.6	18.9	37.9
13	20.9	20.1	37.2
14	19.6	17.7	32.5

Model	MSE	RMSE	MAE	R2
Linear Regression	21.895	4.679	3.271	0.741
Neural Network	5.519	2.349	1.784	0.935

주로 RMSE를 확인
Neural Network의 성능이 더 우수함

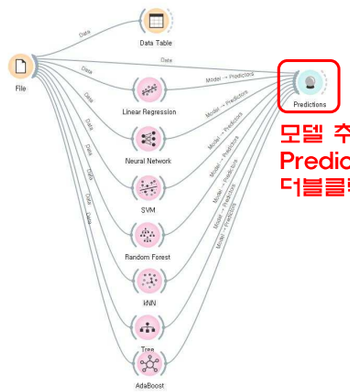
회귀모델의 대표 성능지표

- MSE(Mean Squared Error) : 오차(실제값과 예측값의 차이를 제곱하여 더한 후 평균을 낸 값 (작을 수록 좋음, 과도하게 줄이면 과적합의 오류 가능성)
- RMSE(Root Mean Squared Error) : MSE에 루트를 씌운값, 오류의 제곱임으로 실제 오류 평균보다 커지는 특성이 있어 루트를 씌움 (작을 수록 좋음)
- MAE(Mean Absolue Error) : 실제값과 예측값이 차이를 절대값으로 변환해 평균한 것 (작을 수록 좋음)
- R2(R Square) : 예측값 분산/실제값 분산 (1에 가까울 수록 좋음)

다른 모델과 비교하기

▶ 다른 모델들도 추가하여 실행해보기

▶ SVM, Random Forest, KNN, Tree, AdaBoost



모델 추가 후 Predictions 더블클릭

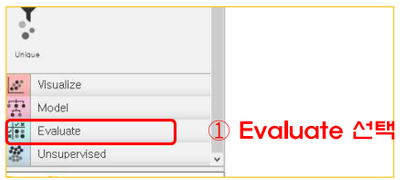
	Linear Regression	Neural Network	SVM	Random Forest	kNN	Tree	AdaBoost	Target
1	30.0	24.7	26.0	26.4	21.8	26.4	24.0	24.0
2	25.0	23.9	21.5	21.7	22.9	21.9	21.6	21.6
3	30.6	31.6	28.2	35.0	25.4	34.8	34.7	34.7
4	28.6	31.5	26.7	34.4	26.1	33.2	33.4	33.4
5	27.9	32.6	26.9	36.0	27.1	37.2	36.2	36.2
6	25.3	25.8	22.6	28.7	27.1	28.9	28.7	28.7
7	23.0	19.3	19.5	21.1	20.9	22.3	22.9	22.9
8	19.5	19.4	18.2	21.8	19.1	22.1	27.1	27.1

Model	MSE	RMSE	MAE	R2
AdaBoost	0.042	0.205	0.058	1.000
Tree	1.858	1.363	0.715	0.978
Random Forest	2.432	1.559	1.029	0.971
Neural Network	5.519	2.349	1.784	0.935
Linear Regression	21.895	4.679	3.271	0.741
kNN	23.967	4.896	3.375	0.716
SVM	38.050	6.168	3.949	0.549

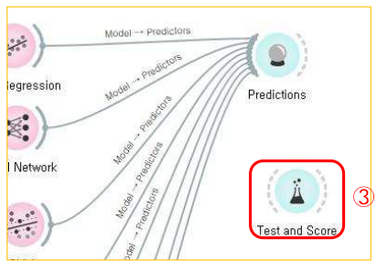
AdaBoost 성능이 가장 좋음

공정하게 평가하기

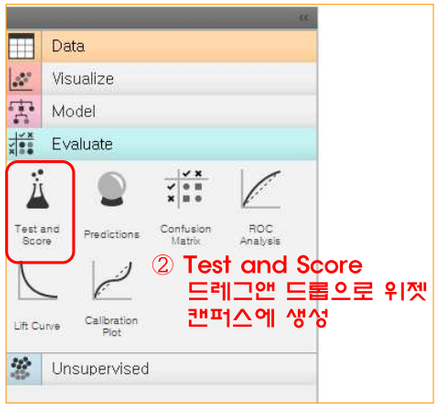
▶ 학습되지 않은 데이터(테스트 데이터)로 평가하기



① Evaluate 선택



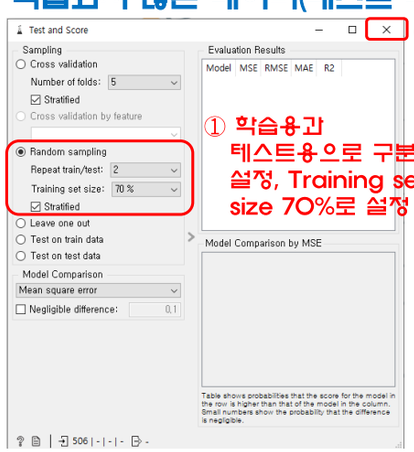
③ Test and Score 위젯 더블클릭



② Test and Score 드래그앤 드롭으로 위젯 캔버스에 생성

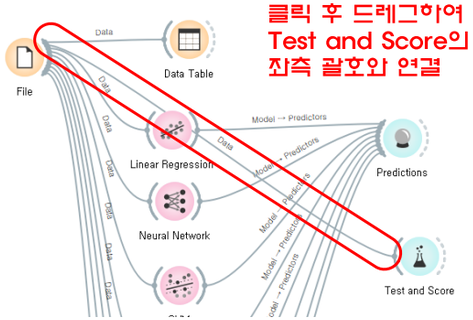
공정하게 평가하기

▶ 학습되지 않은 데이터(테스트 데이터)로 평가하기



② 닫기

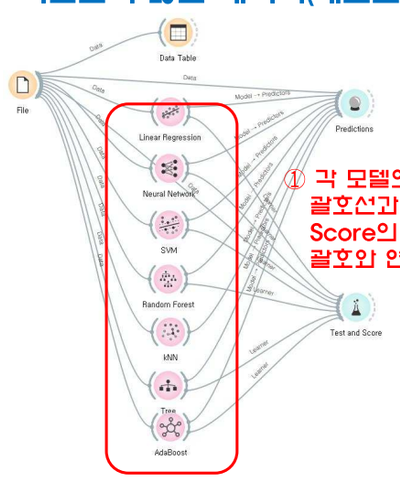
① 학습용과 테스트용으로 구분 설정, Training set size 70%로 설정



③ File의 우측 괄호를 클릭 후 드래그하여 Test and Score의 좌측 괄호와 연결

공정하게 평가하기

▶ 학습되지 않은 데이터(테스트 데이터)로 평가하기



① 각 모델의 우측 괄호선과 Test and Score의 좌측 괄호와 연결

② 테스트 데이터의 성능
- 전반적으로 RMSE가 다 올라갔음
- AdaBoost 성능이 가장 좋음

Evaluation Results

Model	MSE	RMSE	MAE	R2
AdaBoost	8.256	2.873	1.960	0.885
Random Forest	8.448	2.906	2.111	0.882
Neural Network	11.089	3.330	2.393	0.845
Tree	14.875	3.857	2.927	0.793
Linear Regression	19.310	4.394	3.103	0.731
SVM	29.614	5.442	3.285	0.587
kNN	35.186	5.932	4.252	0.510

앞서 수행한 학습데이터의 성능

Model	MSE	RMSE	MAE	R2
AdaBoost	0.042	0.205	0.058	1.000
Tree	1.858	1.363	0.715	0.978
Random Forest	2.432	1.559	1.029	0.971
Neural Network	5.519	2.349	1.784	0.935
Linear Regression	21.895	4.679	3.271	0.741
kNN	23.967	4.896	3.375	0.716
SVM	38.050	6.168	3.949	0.549

학습활동 – 폐암환자 생존여부 예측

- 오렌지를 이용해 thoracic_surgery.csv 데이터로 지도학습 분류모델을 이용해 예측하고 모델 비교하기
  - SVM, kNN, Tree, Random Forest, AdaBoost, Neural Network (Neurons in hidden layers를 17,17,17), Logistic Regression
  - 학습데이터로 학습 및 예측하기
  - 테스트데이터로 예측하기

  - 파일 저장하기 파일명 5주차_학습활동_학번_이름.ows 파일 업로드

학습활동

- 분류모델의 성능지표

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.426	0.809	0.766	0.732	0.809
Tree	0.443	0.741	0.734	0.728	0.741
SVM	0.570	0.851	0.783	0.724	0.851
Random Forest	0.619	0.823	0.779	0.750	0.823
Neural Network	0.513	0.777	0.778	0.779	0.777
Logistic Regression	0.668	0.844	0.785	0.763	0.844
AdaBoost	0.520	0.734	0.744	0.756	0.734

- CA((Classification Accuracy, 분류정확도) : 전체데이터 중 몇 건의 데이터가 맞았는지 알려주는 값 (클 수록 좋음, 1에 가까울 수록)
- Precision(정밀도) : 모델이 True라고 분류한 것 중 실제 True인 것의 비율
- Recall(재현율) : 실제 True인 것 중에서 모델이 True라고 예측한 비율
- F1 : Precision과 Recall의 조화평균

다음시간에 배울 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성,) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티처블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘과 모델링의 개요 2. 알고리즘 기반 의사결정 시스템의 한계 3. 윤리가 필요한 인공지능 4. 오렌지3 설치 및 사용법	1. 오렌지3를 이용한 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	