

파이썬 머신러닝 Starter 강의

with 코랩 & 오렌지3

온라인 강의 구성

강의 1, 2 → 머신러닝 이론 설명 + 오렌지3 맛보기

강의 3 ~ 11 → 머신러닝 Lab (오렌지 3 + 구글 코랩 with 파이썬)

부록 1 → 딥러닝 이론 보강 및 트랜스포머 자동번역

간편 추천 코스

강의 1, 2 → 머신러닝 이론 설명 + 오렌지 3 맛보기

부록 1 → 딥러닝 이론 보강 및 트랜스포머 자동번역

더 배우고 싶은 분은 강의 3~11까지 추천

강의 3 ~ 11 → 머신러닝 Lab (오렌지 3 + 구글 코랩 with 파이썬)



Why 파이썬+머신러닝

초 간단 이유

1. 문과생이 배우기 최적
2. 개인 커리어, 회사 업무에 도움
3. 고급형 엑셀?!

example

1. 머신러닝 맛보기
2. 데이터 처리
3. 탐색적 데이터 분석
4. 머신러닝 revisited



CAMPUS DESTINATIONS

Admissions and Enrollment –

Nigh University Center (NUC) (#43)

Barnes & Noble Bookstore – NUC (#43)

Career Development Center – NUC (#43)

Constitution Hall – NUC (#43)

Center for Counseling and Well-Being –

NUC (#43)

CAMPUS LOCATIONS



UCO Boathouse at Oklahoma River
732 Riversport Dr.,
Oklahoma River Boathouse District,
Downtown Oklahoma City



UCO at Carnegie Centre
131 Dean A. McGee Ave.,
Downtown Oklahoma City



100 North University Drive, Edmond, OK 73034

(405) 974-2000

uco.edu

uco.bronchos

UCOBronchos

ucobronchos

강사: 임선집

jasonyimg@gmail.com

미국 주립대 UCO
MSBA, MBA 졸업

- 47. Power Plant
- 48. President's Annex
- 49. Public Safety
- 50. Administration Park
- 51. Design (future)
- 52. Central Plant
- 53. Performance
- 54. Center Annex
- 55. Annex
- 56. Performance
- 57. Center Annex
- 58. Performance
- 59. Center Annex
- 60. Performance
- 61. Center Annex
- 62. Performance
- 63. Center Annex
- 64. Performance
- 65. Center Annex
- 66. Performance
- 67. Center Annex
- 68. Performance
- 69. Center Annex
- 70. Performance
- 71. Center Annex
- 72. Performance
- 73. Center Annex
- 74. Performance
- 75. Center Annex
- 76. Performance
- 77. Center Annex
- 78. Performance
- 79. Center Annex
- 80. Performance
- 81. Center Annex
- 82. Performance
- 83. Center Annex
- 84. Performance
- 85. Center Annex
- 86. Performance
- 87. Center Annex
- 88. Performance
- 89. Center Annex
- 90. Performance
- 91. Center Annex
- 92. Performance
- 93. Center Annex
- 94. Performance
- 95. Center Annex
- 96. Performance
- 97. Center Annex
- 98. Performance
- 99. Center Annex
- 100. Performance

폴리텍 정수 캠퍼스
파이썬 머신러닝 강사

UCO at Santa Fe Plaza
101 North E.K. Gaylord, Ste. 1
Downtown Oklahoma City



100 North University Drive, Edmond, OK 73034

(405) 974-2000 uco.edu

[uco.bronchos](https://facebook.com/ucobronchos) [@UCOBronchos](https://twitter.com/UCOBronchos) [@ucobronchos](https://instagram.com/ucobronchos)

Lecturer: Sunjip Yim (JJ)
jasonyim2022ws@gmail.com

UCO
MSBA, MBA

- CAMPUS DESTINATIONS –
- Admissions and Enrollment –
 - Nigh University Center (NUC) (#43)
 - Barnes & Noble Bookstore – NUC (#43)
 - Career Development Center – NUC (#43)
 - Constitution Hall – NUC (#43)
 - Center for Counseling and Well-Being –
 - NUC (#43)
- CAMPUS LOCATIONS
- UCO Boathouse at Oklahoma River
732 Riversport Dr.,
Oklahoma River Boathouse District,
Downtown Oklahoma City
 - UCO at Carnegie Centre
131 Dean A. McGee Ave.,
Downtown Oklahoma City
 - UCO at Santa Fe Plaza
101 North E.K. Gaylord, Ste. 1
Downtown Oklahoma City
23. Evans Hall
 24. Facilities M.
 25. Forensic S.
 26. Gerry Pink
 27. Hamilton Field House
 28. Health and Physical Education
 29. Howell Hall
 30. Hurst Hall
 31. Inter
 32. J. C. Nichols Hall
 33. Kell Hall
 34. Ladd Hall
 35. Lee Hall
 36. Lewis Hall
 37. Lovelace Hall
 38. Main Hall
 39. McFarlin Hall
 40. McMurtry Hall
 41. Moore Hall
 42. Mullins Hall
 43. Nigh University Center
 44. Old Main
 45. O'Neil Hall
 46. Owen Hall
 47. Pugh Hall
 48. President's Annex
 49. Public Safety Administration
 50. Reflection Park
 51. School of Design (future)
 52. South Central Plant
 53. Sports Performance Center Annex
 54. Sports Performance Center
 55. Stamps Hall
 56. Student Union
 57. Student Union Annex
 58. Student Union Annex
 59. Student Union Annex
 60. Student Union Annex
 61. Student Union Annex
 62. Student Union Annex
 63. Student Union Annex
 64. Student Union Annex
 65. Student Union Annex
 66. Student Union Annex
 67. Student Union Annex
 68. Student Union Annex
 69. Student Union Annex
 70. Student Union Annex
 71. Student Union Annex
 72. Student Union Annex
 73. Student Union Annex
 74. Student Union Annex
 75. Student Union Annex
 76. Student Union Annex
 77. Student Union Annex
 78. Student Union Annex
 79. Student Union Annex
 80. Student Union Annex
 81. Student Union Annex
 82. Student Union Annex
 83. Student Union Annex
 84. Student Union Annex
 85. Student Union Annex
 86. Student Union Annex
 87. Student Union Annex
 88. Student Union Annex
 89. Student Union Annex
 90. Student Union Annex
 91. Student Union Annex
 92. Student Union Annex
 93. Student Union Annex
 94. Student Union Annex
 95. Student Union Annex
 96. Student Union Annex
 97. Student Union Annex
 98. Student Union Annex
 99. Student Union Annex
 100. Student Union Annex

PolyTech Campus
Python ML lecturer

문과 vs
이공대

문과용 파이썬

데이터 처리, 통계, 숫자, 텍스트 머신러닝

이공대

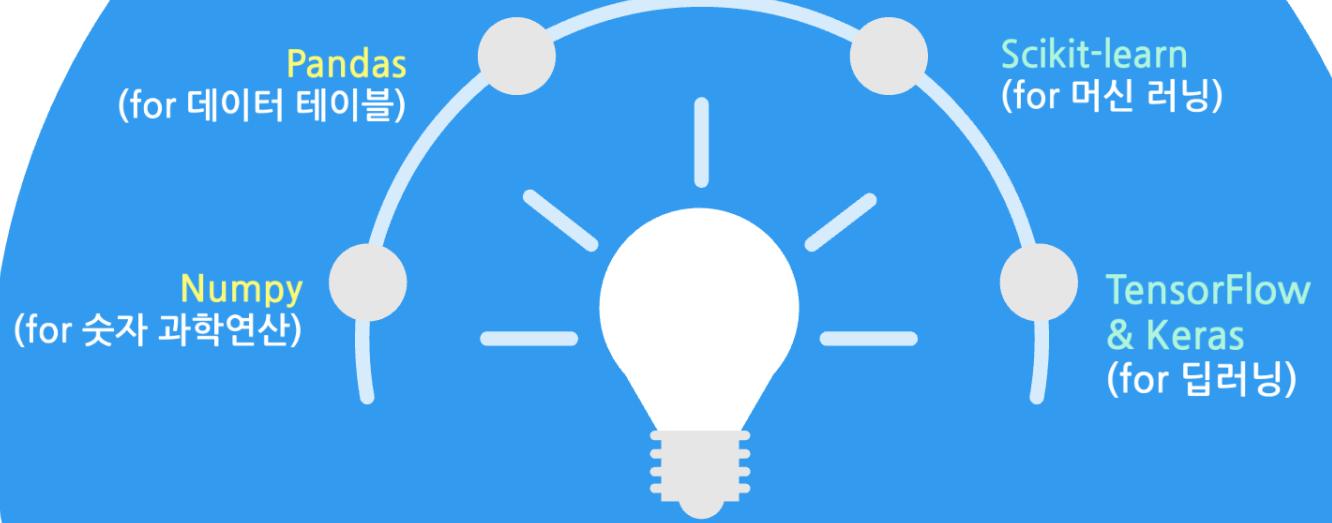
- 
1. 숫자 (**문과**)
 2. 텍스트 (**문과**)
 3. 이미지 (**이과**)

1. 지도 학습
(Supervised)
2. 비지도 학습
3. 강화 학습
(알파고, 로봇청소기)

Predictive
Modeling

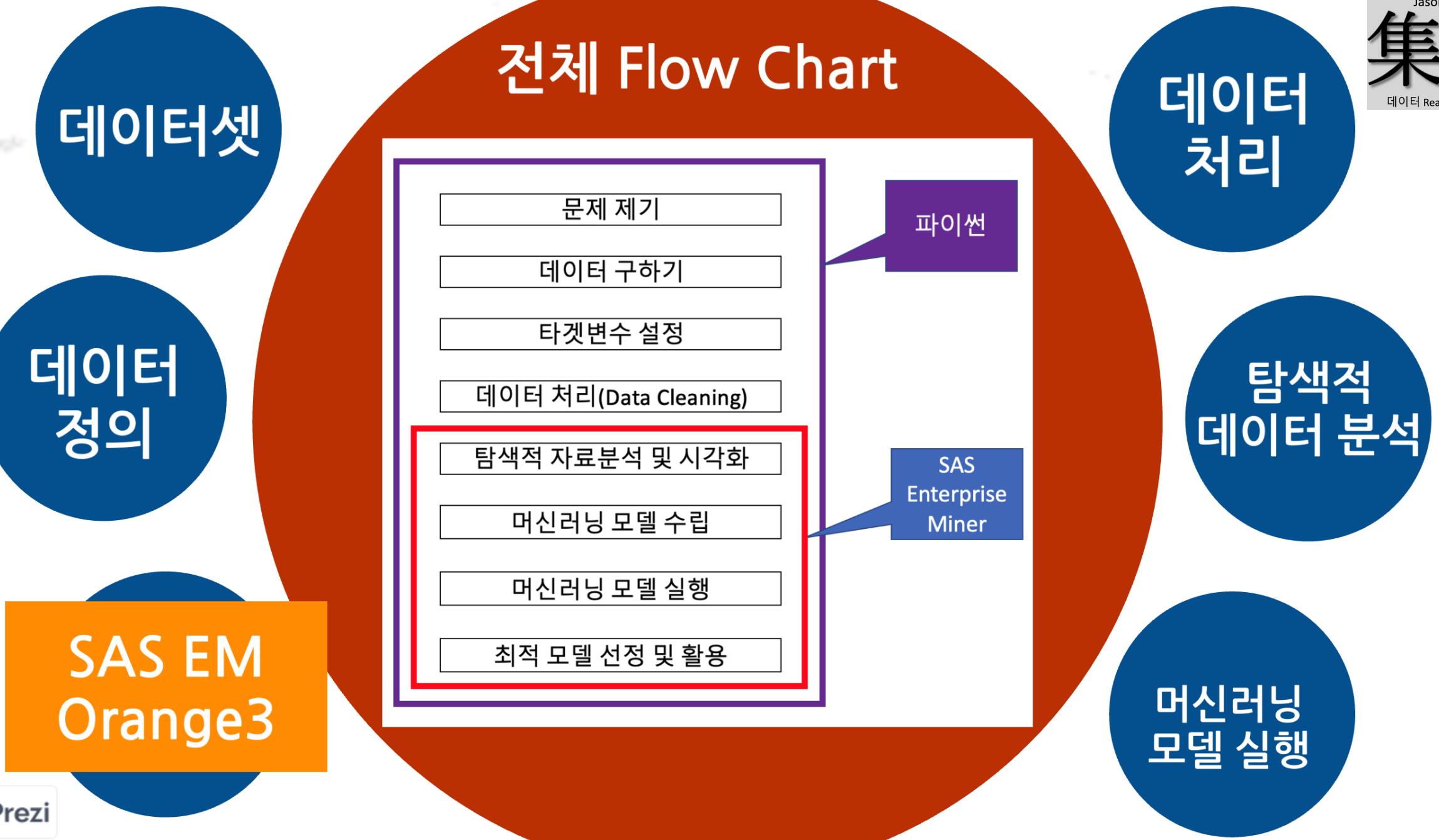
1. Decision
(Yes or No)
2. Estimates
(for 연속변수)
3. Ranking
(순위 매기기)

파이썬 라이브러리



예: 소니 플레이스테이션

전체 Flow Chart





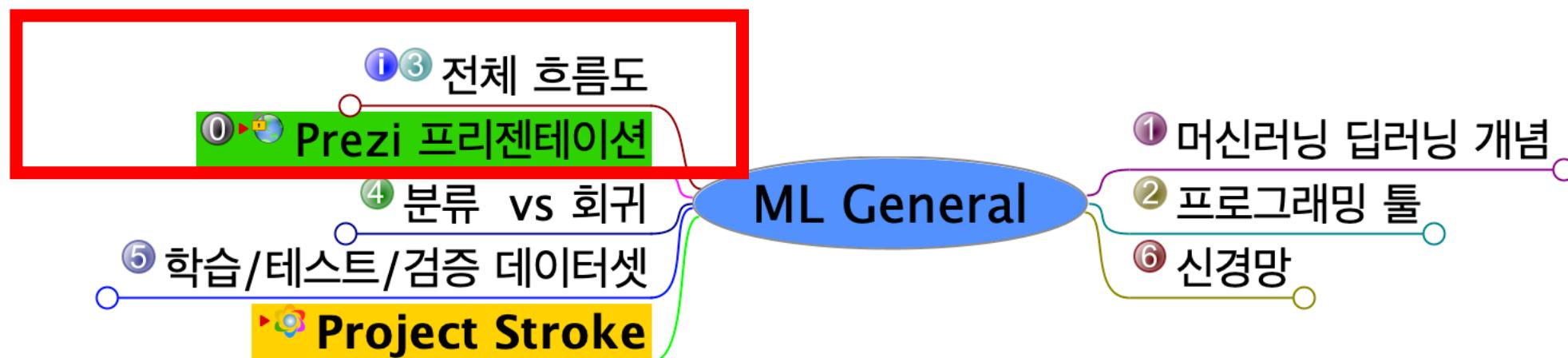
Freeplane

Application for Mind Mapping, Knowledge and Project Management

온라인 강의 Lab에서 프리젠테이션 용으로 사용하는 오픈소스
마인드맵 프로그램!

How to install → <https://bit.ly/3tYEEZw>

강의 듣는 분들은 설치하지 않으셔도 됩니다.
프리젠테이션 용이며 파이썬 머신러닝 코딩과 무관한 프로그램입니다.



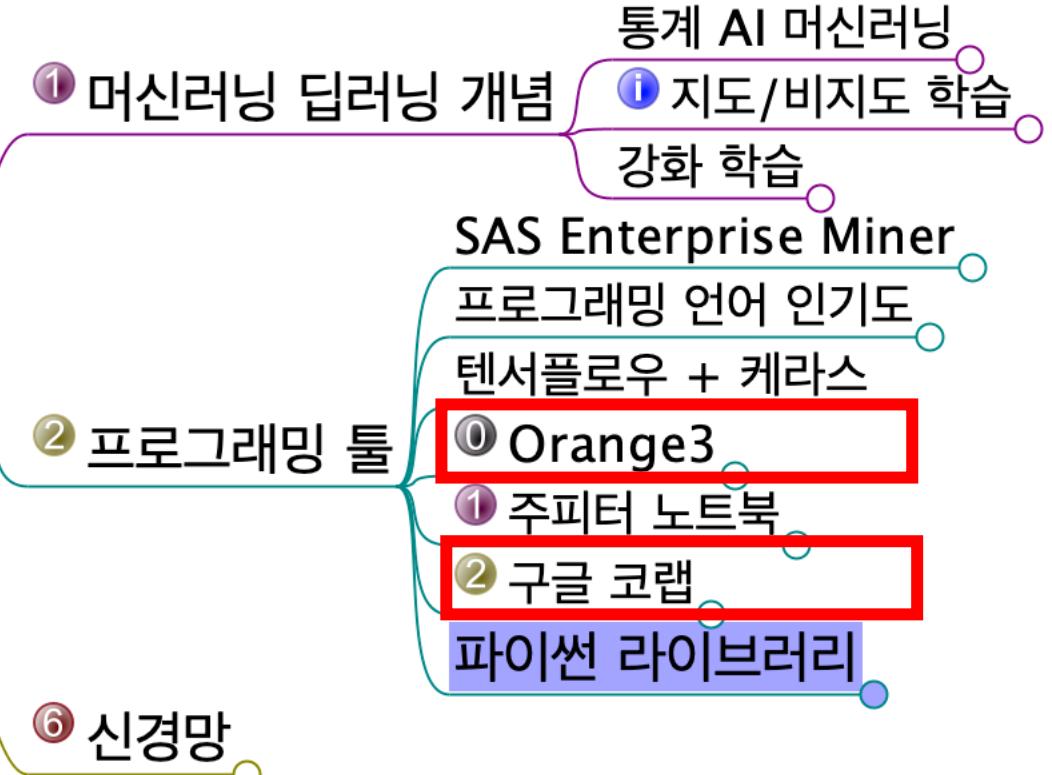


Why Orange3?

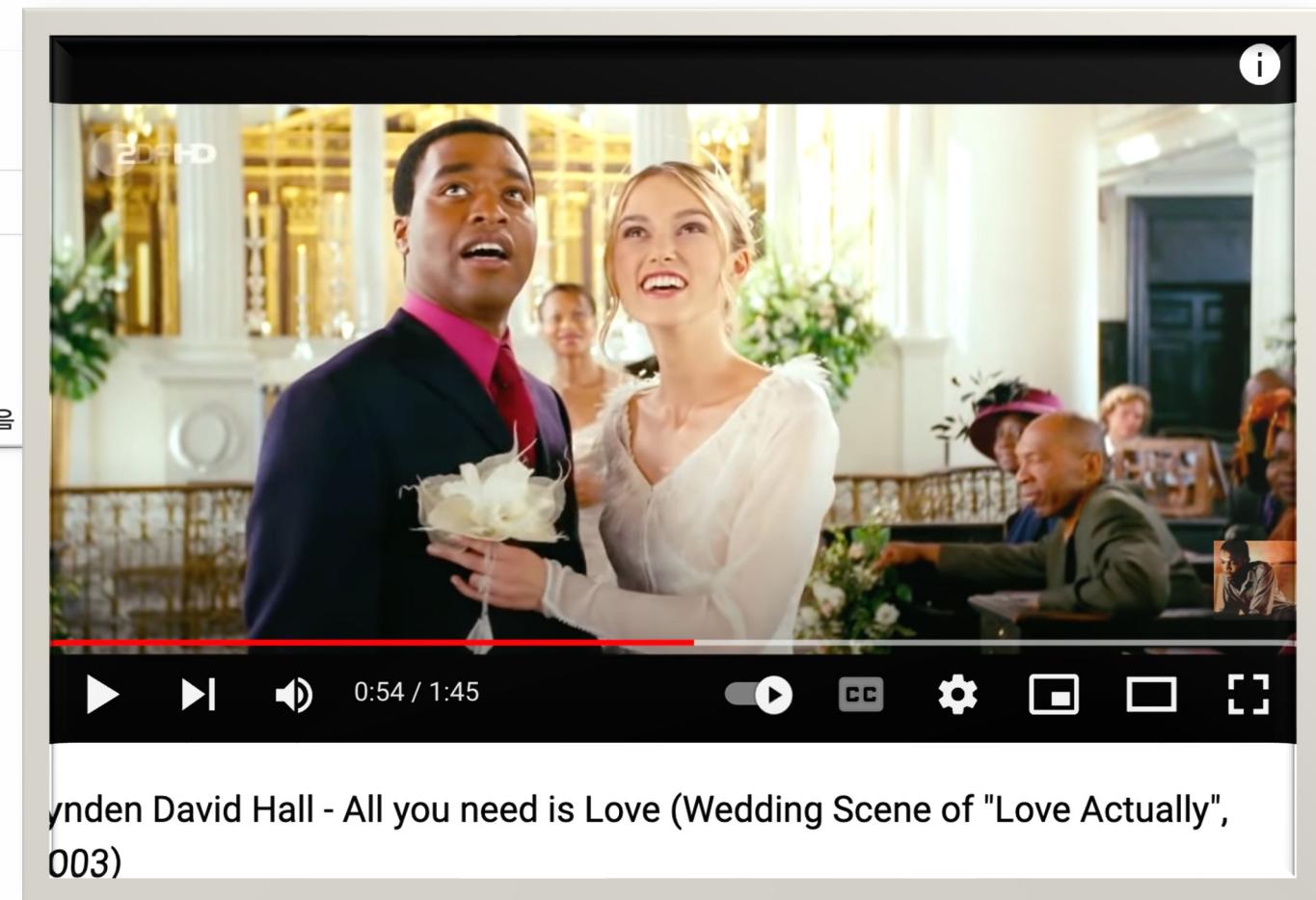
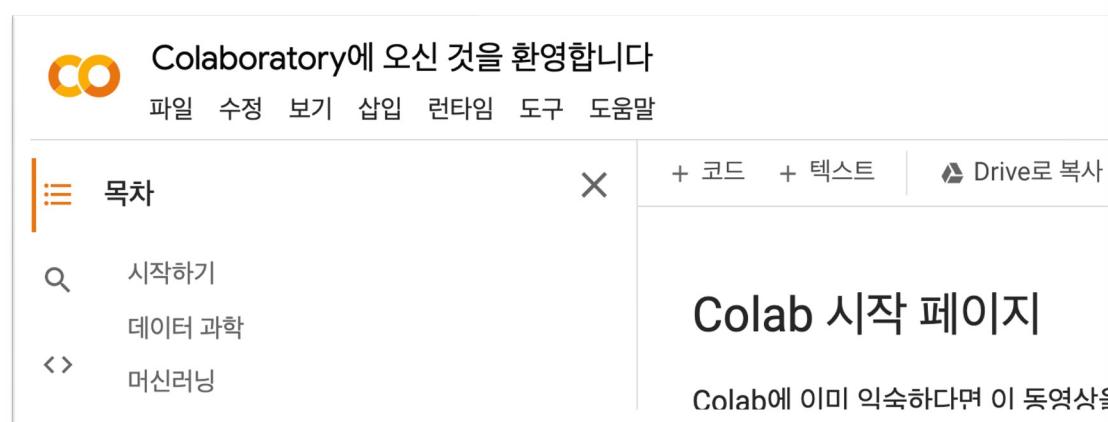
Why 구글 코랩?



ML General



All you need is Colab, but...



Orange3 can help.

The screenshot shows the Orange3 interface with the 'Data' tab selected. The top bar has tabs for 'Data', 'Visualize', 'Model', and 'Script'. Below the tabs is a toolbar with icons for 'File', 'CSV File Import', 'Datasets', 'SQL Table', 'Data Table', 'Paint Data', 'Data Info', 'Aggregate Columns', 'Data Sampler', 'Select Columns', 'Select Rows', and 'Pivot Table'. The main area displays a flow diagram:

```

graph LR
    File((File)) -- Data --> FS1((Feature Statistics (3)))
    FS1 --- Note1[File (health csv) bmi meta -- type role change needed  
bmi not showing]
    FS1 -- Data --> FS2((Feature Statistics (2)))
    FS2 --- Note2[CSV File Import (health csv,bmi OK)]
    FS2 -- Data --> Distributions((Distributions (1)))
    Distributions --- Note3[Distributions (1)]
  
```

The 'Feature Statistics (3)' node has a note: "File (health csv) bmi meta -- type role change needed
bmi not showing". The 'Feature Statistics (2)' node has a note: "CSV File Import (health csv,bmi OK)". The 'Distributions (1)' node has a note: "Distributions (1)".

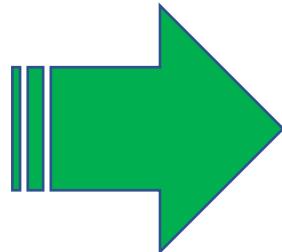




Orange3는 빠른 review용 or ML 학습용
단, 실전은 꼭 Python으로!!!

The screenshot shows the Orange3 interface with the following components visible:

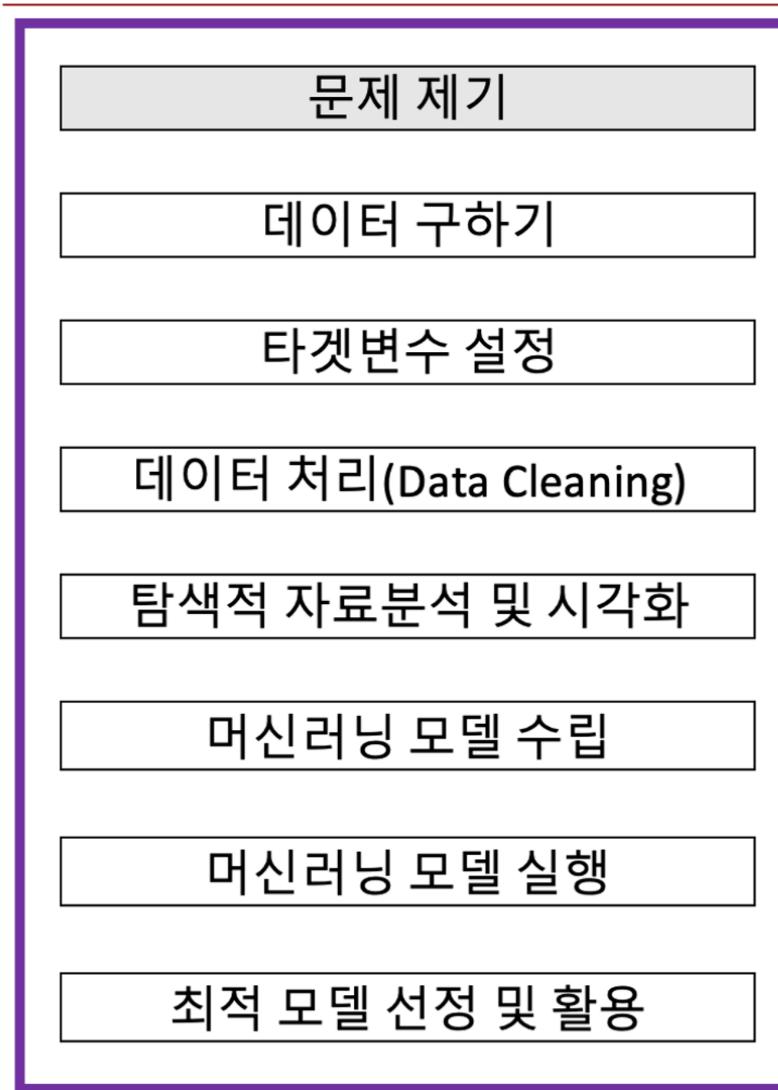
- Left Sidebar:** Contains tabs for Data, Visualize, and Model. Under Model, there are icons for Constant, CN2 Rule Induction, Calibrated Learner, kNN, Tree, Random Forest, Gradient Boosting, and SVM.
- Workflow Area:** Displays a data flow diagram with nodes and connections. Nodes include:
 - Data node connected to Feature Statistics (3) node.
 - Feature Statistics (3) node with a warning message: "File (health csv) bmi meta -- type role change needed" and "bmi not showing".
 - CSV File Import (health csv,bmi OK) node connected to Feature Statistics (2) node.
 - Feature Statistics (2) node connected to Distributions (1) node.
 - Distributions (1) node with a warning message: "Data" and "CSV File Import (health csv,bmi OK)".

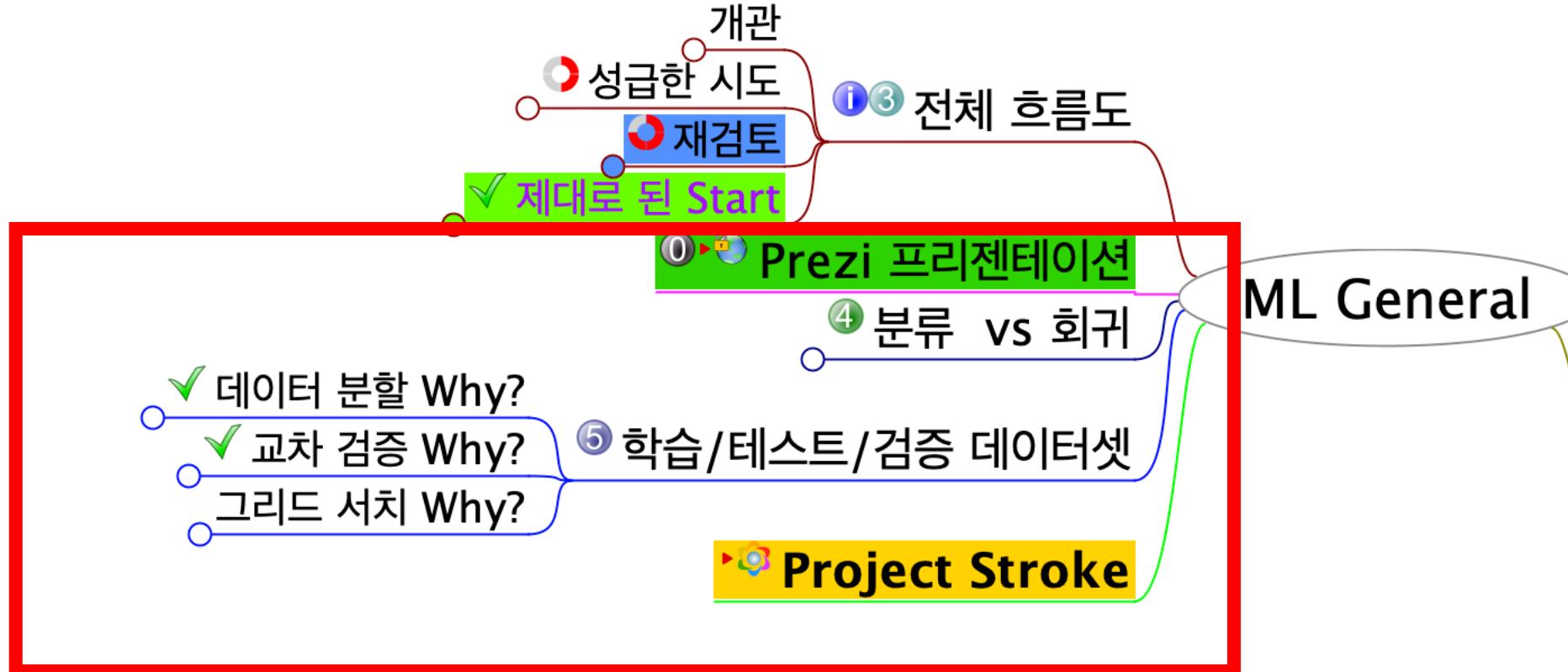


The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** Stroke Data Cleaning 2 v2.ipynb
- Code Cell:**

```
[ ] import seaborn as sns
sns.histplot(data=df, x="age", hue="stroke", bins=20);
```
- Output:** A histogram titled "Count" versus "age". The x-axis ranges from 20 to 80, and the y-axis ranges from 0 to 250. The legend indicates two categories: "stroke 0" (blue bars) and "stroke 1" (grey bar at the end of each bin).
- Header:** File Edit View Insert Runtime Tools Help Last edited on January 7
- Sidebar:** Includes buttons for + Code, + Text, and a search bar.





지금부터 코랩 및 오렌지3 실습(Lab)

실습 작업은 온라인 강의 화면 참조

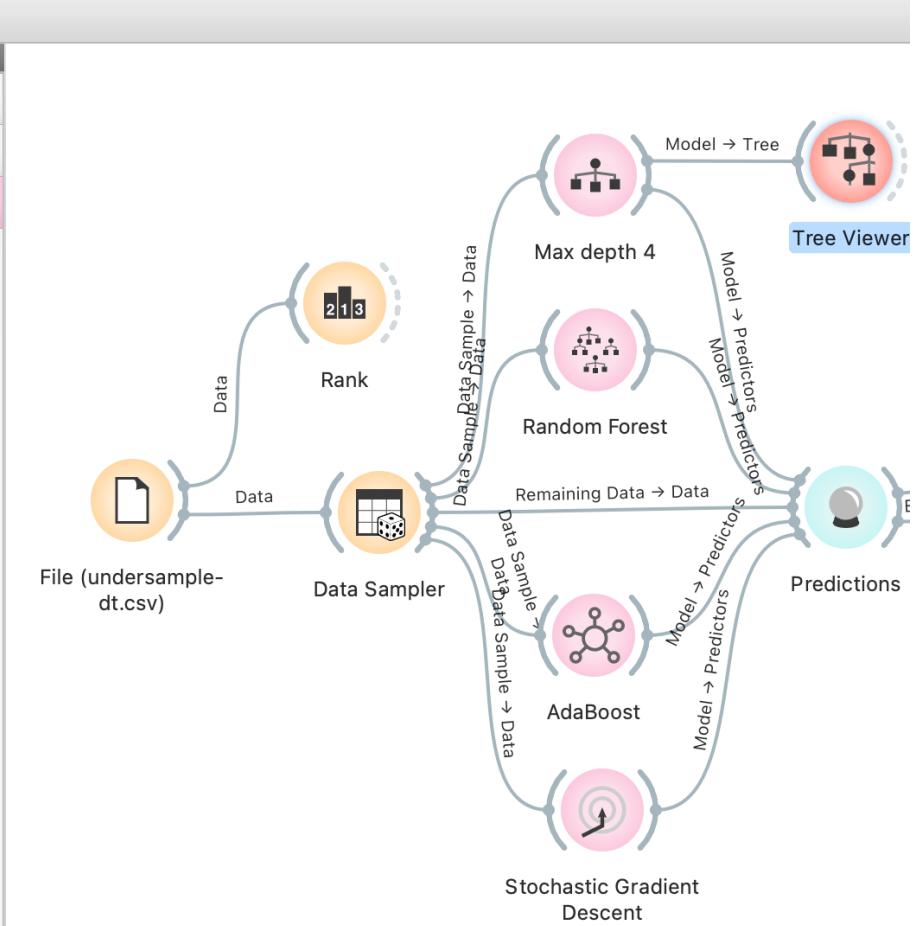
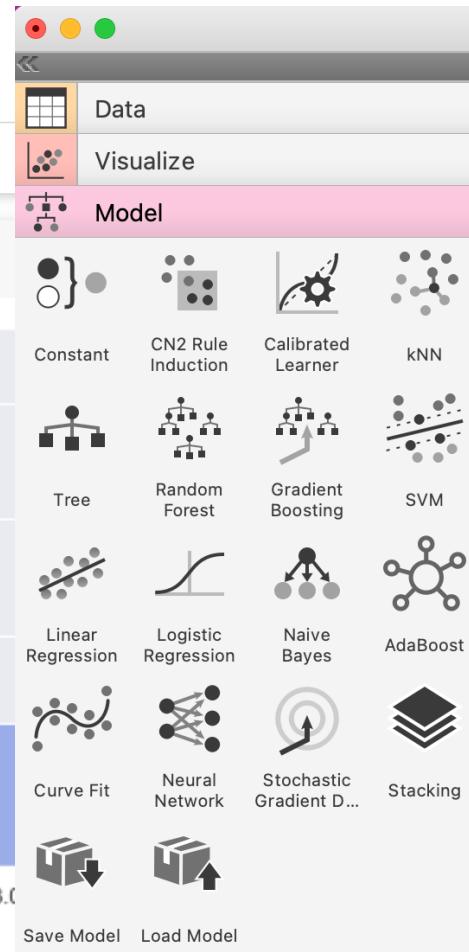
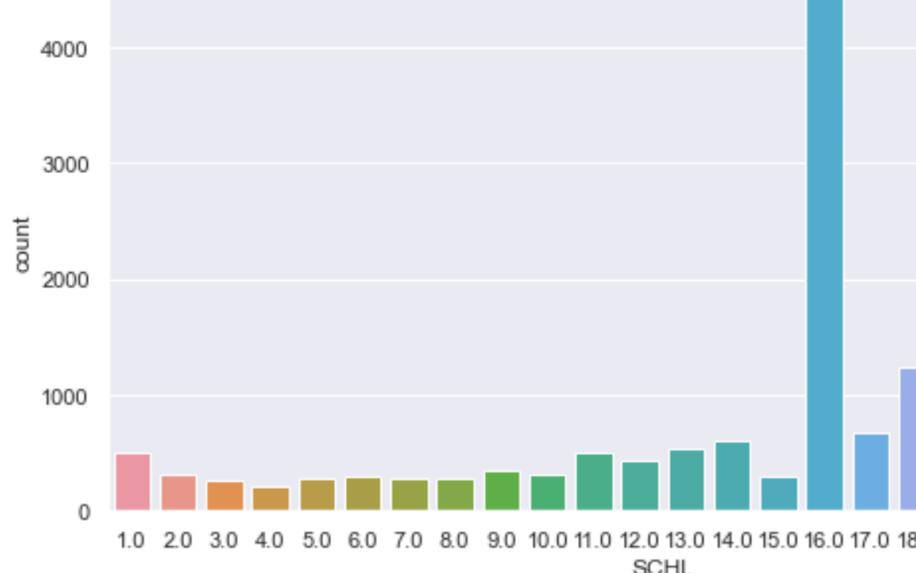


House Data Cleaning (2017) 3 v2.ipynb

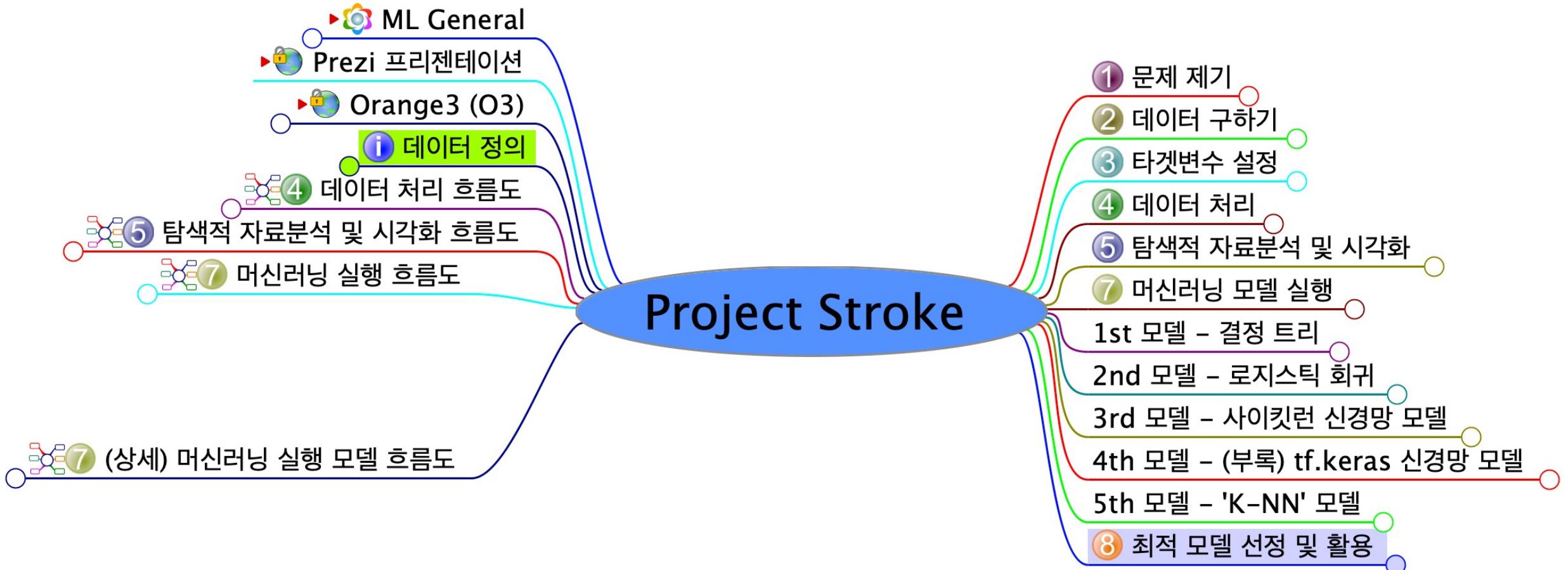
File Edit View Insert Runtime Tools Help Last edited on January 7

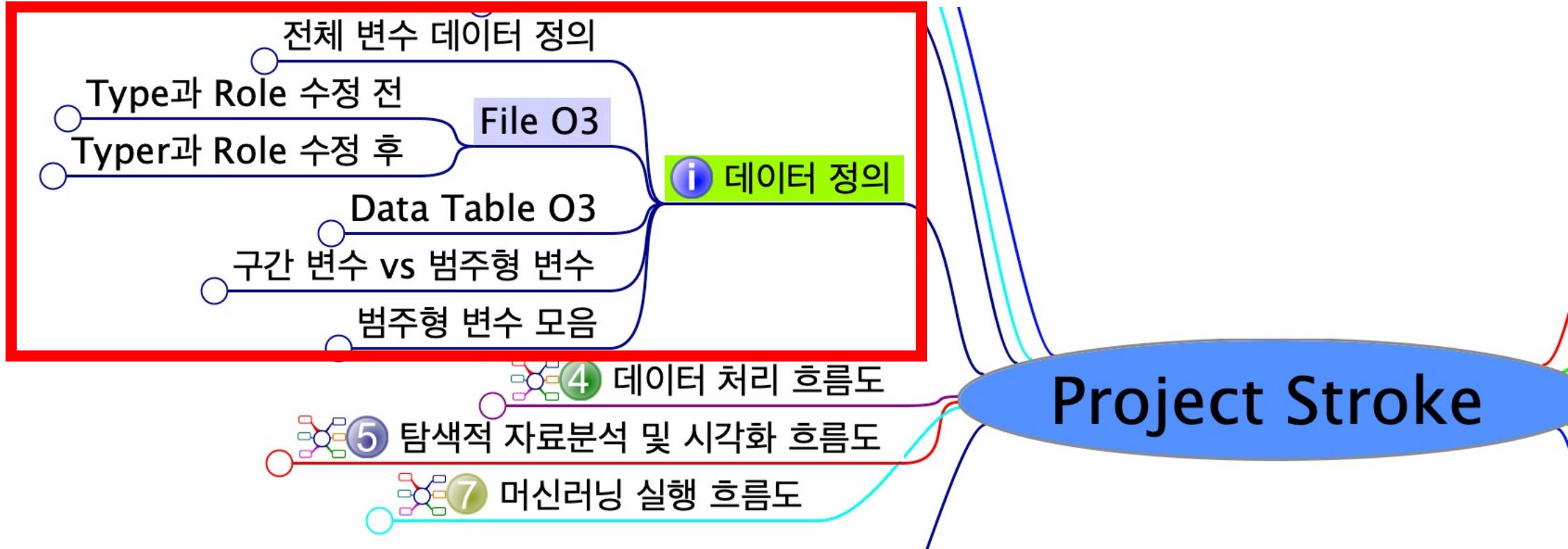
+ Code + Text

```
sns.countplot(x='SCHL', data=df);
```



Project Stroke





Info

4194 instance(s)

12 feature(s) (0.4% missing values)

Data has no target variable.

0 meta attribute(s)

Columns (Double click to edit)

구간 변수	범주형 변수
age, avg-glucose-level, bmi	gender, hypertension, heart_disease, ever_married, work_type, residence_type, smoking_status

	Name	Type	Role	Values
1	id	N numeric	skip	
2	gender	C categorical	feature	Female, Male, Other
3	age	N numeric	feature	
4	hypertension	C categorical	feature	0, 1
5	heart_disease	C categorical	feature	0, 1
6	ever_married	C categorical	feature	No, Yes
7	work_type	C categorical	feature	Govt_job, Never_worked, Private, Self-employed
8	residence_ty...	C categorical	feature	Rural, Urban
9	avg_glucose...	N numeric	feature	
10	bmi	N numeric	feature	
11	smoking_sta...	C categorical	feature	Unknown, formerly smoked, never smoked, smokes
12	stroke	C categori...	target	0, 1

① 문제 제기

Project Stroke

- ① 문제 제기
- ② 데이터 구하기
- ③ 타겟변수 설정
- ④ 데이터 처리
- ⑤ 탐색적 자료분석 및 시각화
- ⑦ 머신러닝 모델 실행
- 1st 모델 - 결정 트리
- 2nd 모델 - 로지스틱 회귀
- 3rd 모델 - 사이킷런 신경망 모델
- 4th 모델 - (부록) tf.keras 신경망 모델
- 5th 모델 - 'K-NN' 모델
- ⑧ 최적 모델 선정 및 활용

연구주제: 뇌졸증(stroke) 발병 요인 중 중요한 변수는 무엇인가?

별로 흥미롭지 않다고요? 그럼 이런 연구주제는 어떻습니까?

프리미어 리그 축구 선수의 랭킹을 결정하는 요인은 무엇인가?

혹은 NBA 팀의 선수 구성을 어떻게 해야 우승 확률을 높아지는가?

① 문제 제기

연구 질문

② 데이터 구하기

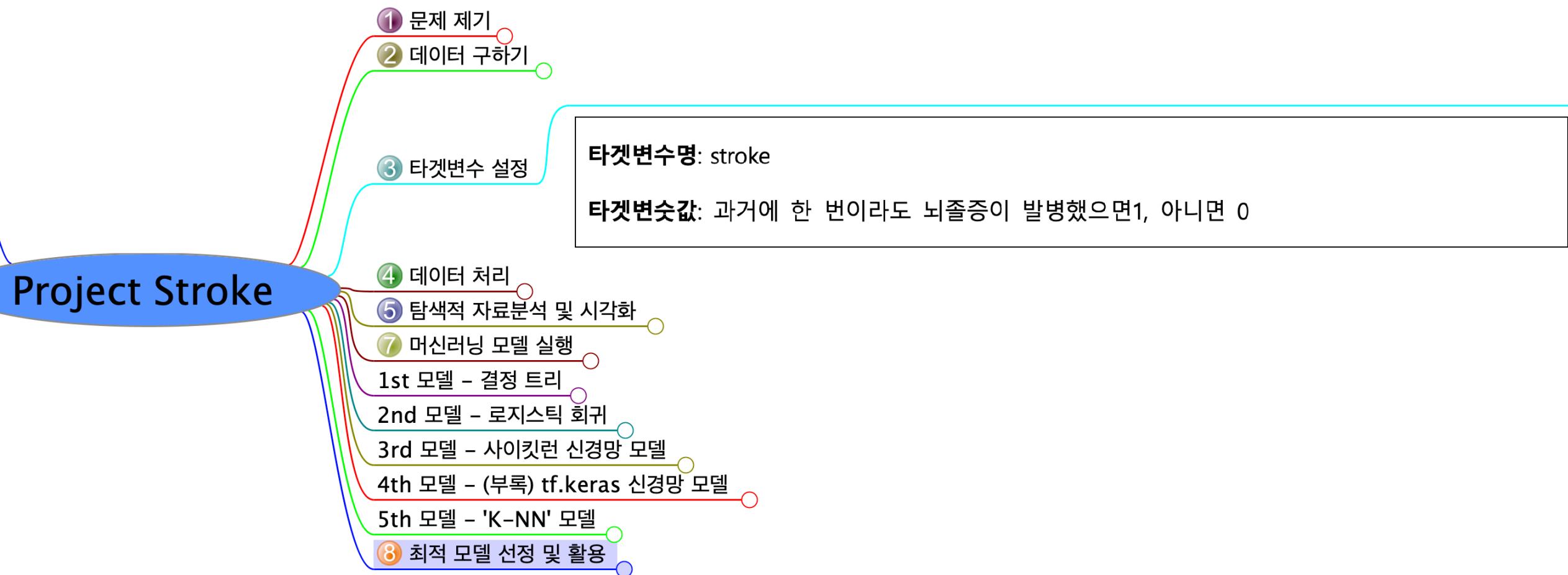
Kaggle 데이터셋

Stroke-prediction_dataset

<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

The screenshot shows a Kaggle dataset page for 'Stroke Prediction Dataset'. The page features a large image of a brain with red highlights. At the top left is a 'Dataset' icon. On the right, there's a yellow profile picture, an upward arrow, and the number '1073'. Below the title, it says 'Stroke Prediction Dataset' and '11 clinical features for predicting stroke events'. Underneath, it shows the author 'fedesoriano' with a small profile pic, and the text 'updated 3 months ago (Version 1)'. A navigation bar below includes 'Data' (which is underlined), 'Tasks (1)', 'Code (345)', 'Discussion (18)', 'Activity', and 'Metadata'. To the right of these are 'Download (310 KB)' (which is highlighted with a red box), 'New Notebook', and a three-dot menu. At the bottom, there are sections for 'Usability 10.0', 'License Data files © Original Authors', and 'Tags' (which include 'health', 'health conditions', 'public health', 'healthcare', and 'binary classification').

③ 타겟변수 설정

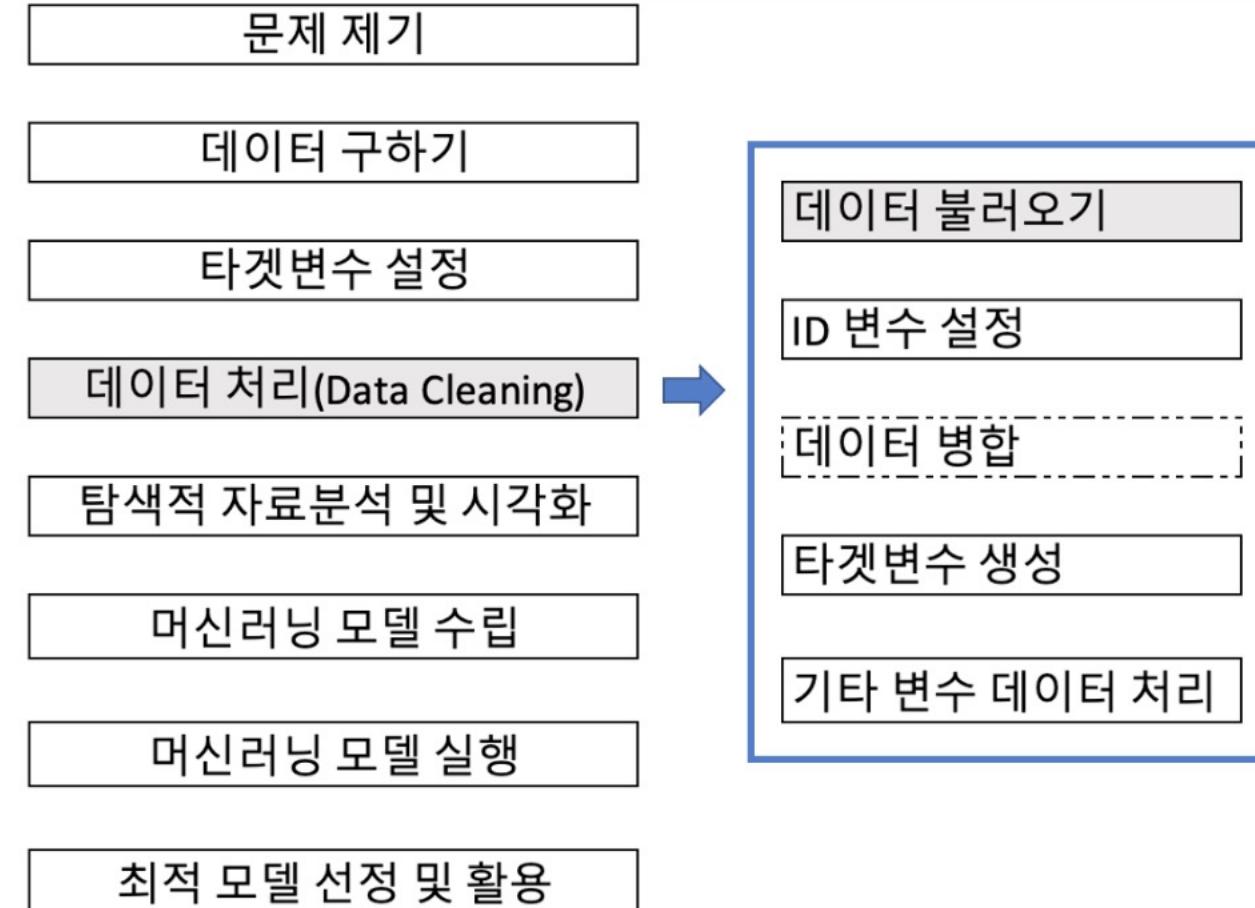


4 데이터 처리



4 데이터 처리 흐름도

데이터 처리 상세 흐름도



Project Stroke

① 문제 제기

② 데이터 구하기

③ 타겟변수 설정

4-1 데이터 불러오기

4-2 ID 변수 설정

4-3 타겟변수 생성

4-4 기타 변수 데이터 처리

청소년 vs 성인 Stroke

⑤ 탐색적 자료분석 및 시각화

⑦ 머신러닝 모델 실행

1st 모델 - 결정 트리

2nd 모델 - 로지스틱 회귀

3rd 모델 - 사이킷런 신경망 모델

4th 모델 - (부록) tf.keras 신경망 모델

5th 모델 - 'K-NN' 모델

⑧ 최적 모델 선정 및 활용

⑤

탐색적 자료분석 및 시각화

5 탐색적 자료분석 및 시각화 흐름도

문제 제기
데이터 구하기
타겟변수 설정
데이터 처리(Data Cleaning)
탐색적 자료분석 및 시각화
머신러닝 모델 수립
머신러닝 모델 실행
최적 모델 선정 및 활용

결측값 50% 초과 변수 제거
구간(Interval) 변수
요약통계 검토
이상값 제거
상관관계 검토
시각화
t 검정
범주형 변수
도수분포표 검토
시각화

Project Stroke

- ① 문제 제기
- ② 데이터 구하기
- ③ 타겟변수 설정
- ④ 데이터 처리

5 탐색적 자료분석 및 시각화

- 5-1 결측값 50% 초과 변수 제거 `isnull()`
- 5-2 요약 통계 검토
- 5-2 (계속) 도수분포표 검토 `pd.crosstab()`
- 5-3 이상값 제거
- 5-4 상관관계 검토 `corr()`
`sns.heatmap()`
- 5-5 시각화
- 5-6 t-검정 `stats.ttest_ind()`
`E notation`

7 머신러닝 모델 실행

- 1st 모델 - 결정 트리
- 2nd 모델 - 로지스틱 회귀
- 3rd 모델 - 사이킷런 신경망 모델
- 4th 모델 - (부록) `tf.keras` 신경망 모델
- 5th 모델 - 'K-NN' 모델
- 8 최적 모델 선정 및 활용

⑤ 탐색적 자료분석 및 시각화

5-1 결측값 50% 초과 변수 제거

`isnull()`

bmi의 결측값 비율 확인

```
df['bmi'].isnull().mean()
```

5-2 요약 통계 검토

`describe()`

`value_counts()`

5-2 (계속) 도수분포표 검토

`pd.crosstab()`

5-3 이상값 제거

5-4 상관관계 검토

`corr()`

`sns.heatmap()`

5-5 시각화

구간 변수

범주형 변수

5-6 t-검정

`stats.ttest_ind()`

E notation

5-1 결측값 50% 초과 변수 제거

5-2 요약 통계 검토

5-2 (계속) 도수분포표 검토

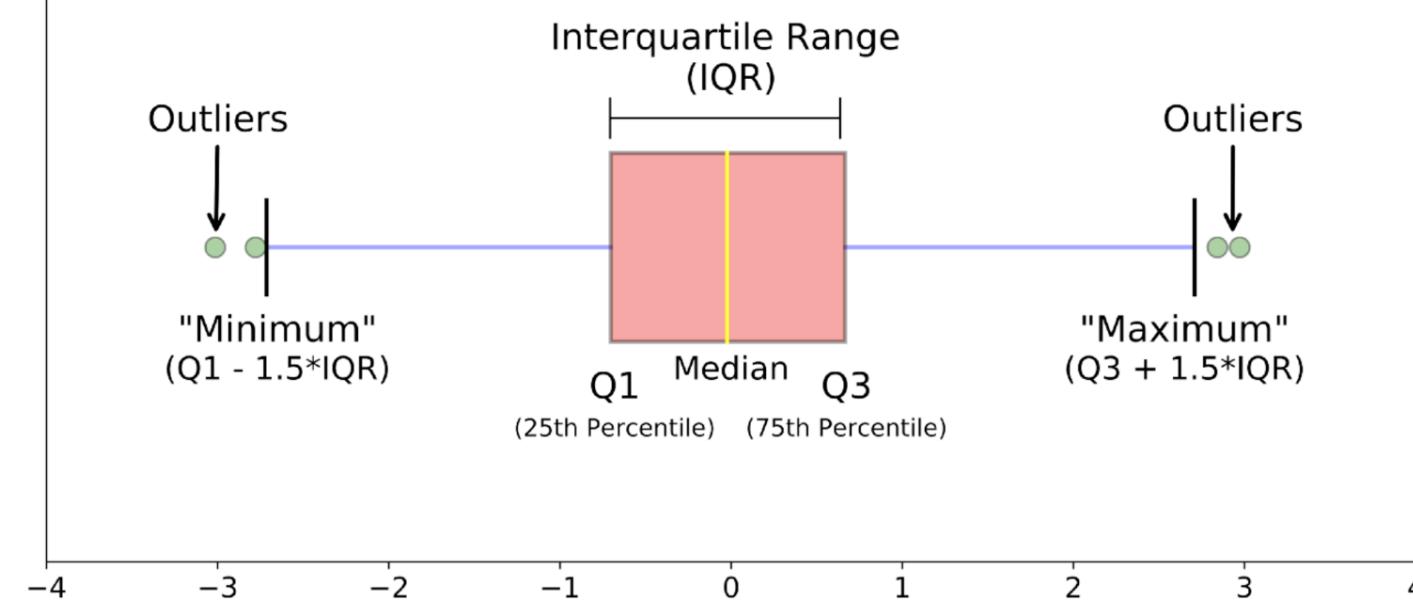
pd.crosstab()

URL

상자그램
5-3 이상값 제거

IQR 규칙

로그 변환



1.5*IQR 규칙: ($Q1 - 1.5 \times IQR$) 미만 값 혹은 ($Q3 + 1.5 \times IQR$) 초과 값 제거

3.0*IQR 규칙: ($Q1 - 3.0 \times IQR$) 미만 값 혹은 ($Q3 + 3.0 \times IQR$) 초과 값 제거

5-1 결측값 50% 초과 변수 제거

5-2 요약 통계 검토

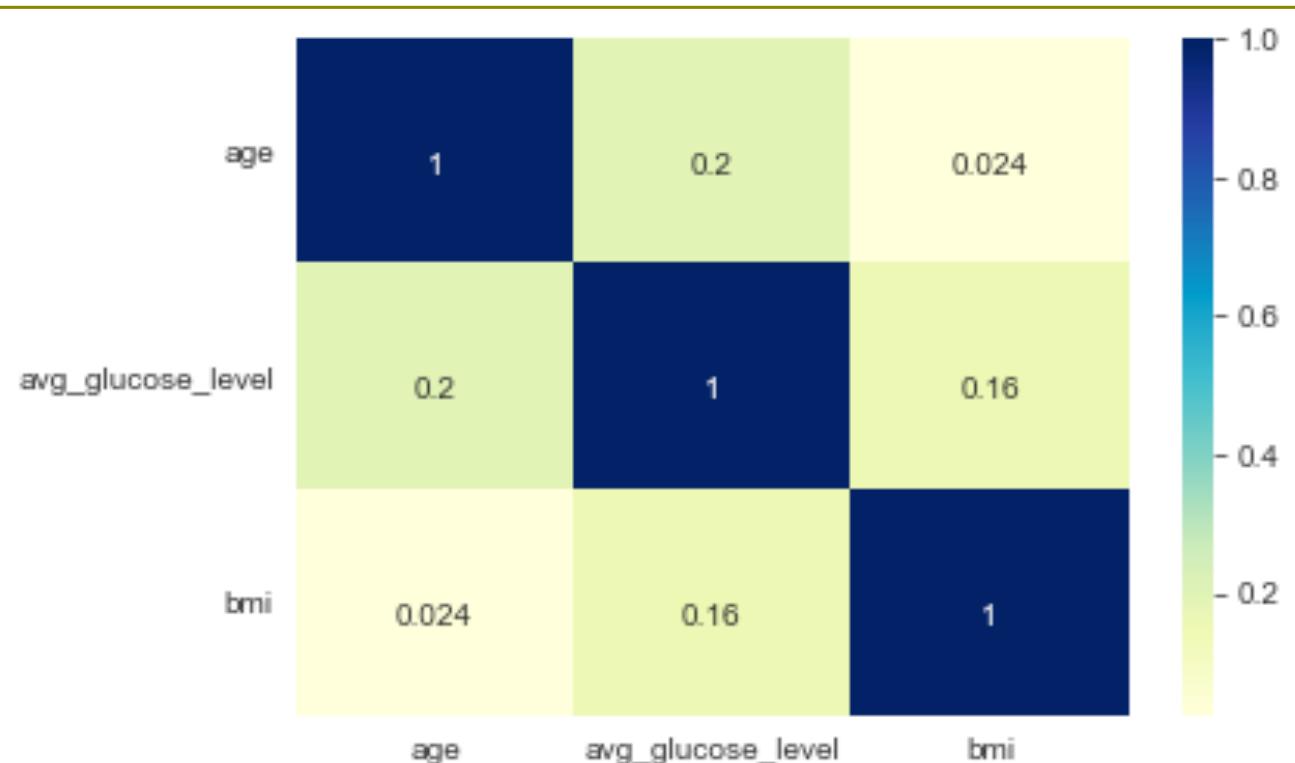
5-2 (계속) 도수분포표 검토

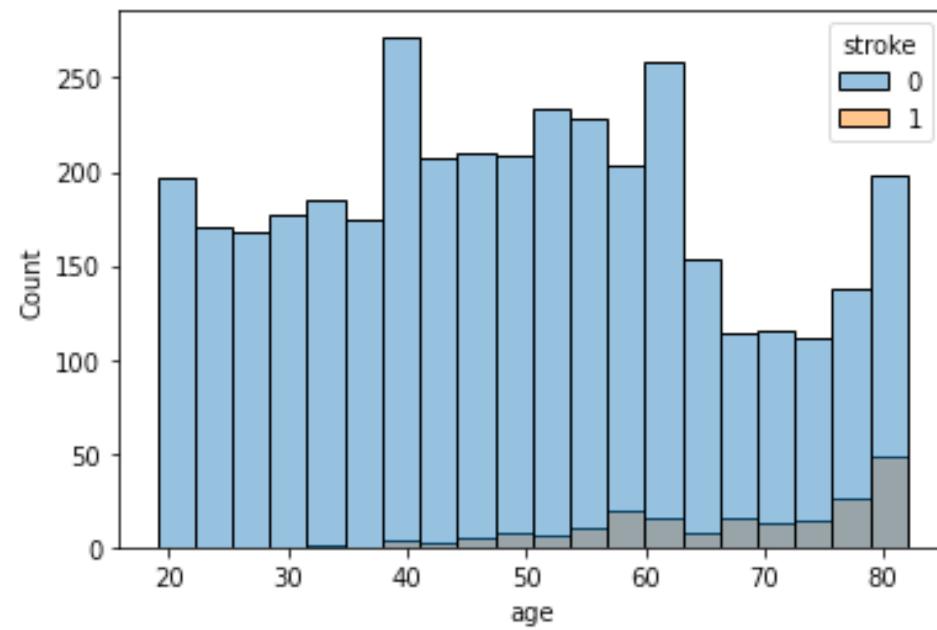
`pd.crosstab()`

5-3 이상값 제거

`corr()`

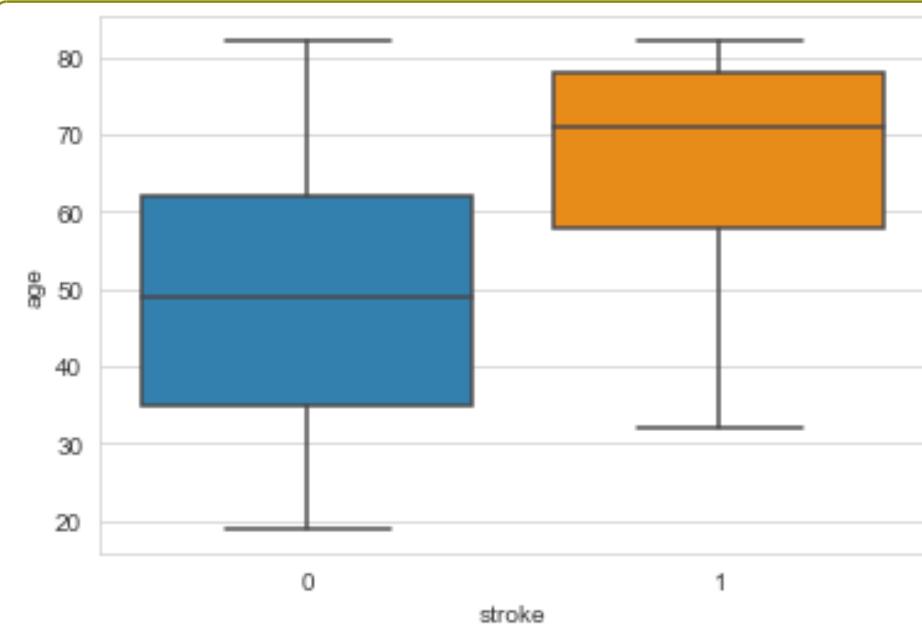
5-4 상관관계 검토

`sns.heatmap()`



sns.histplot()

5-5 시각화
구간 변수



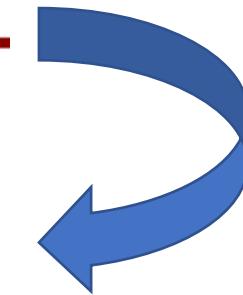
sns.boxplot()

범주형 변수

sns.countplot()

6 머신러닝 모델 수립

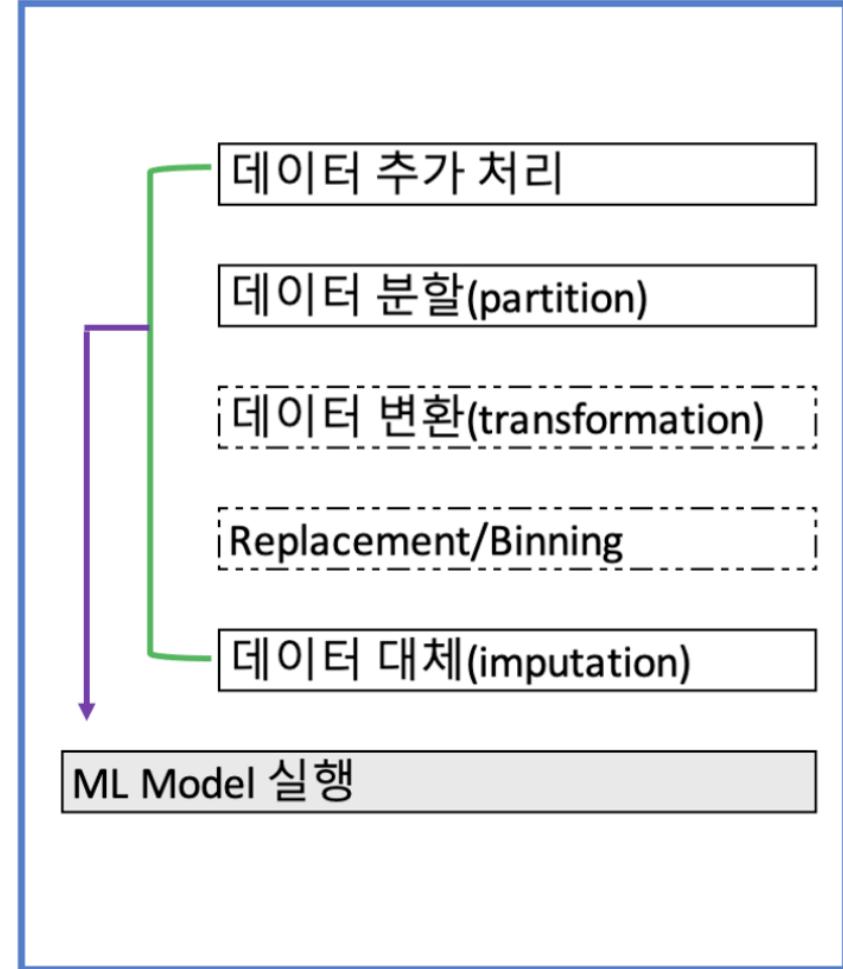
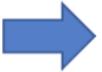
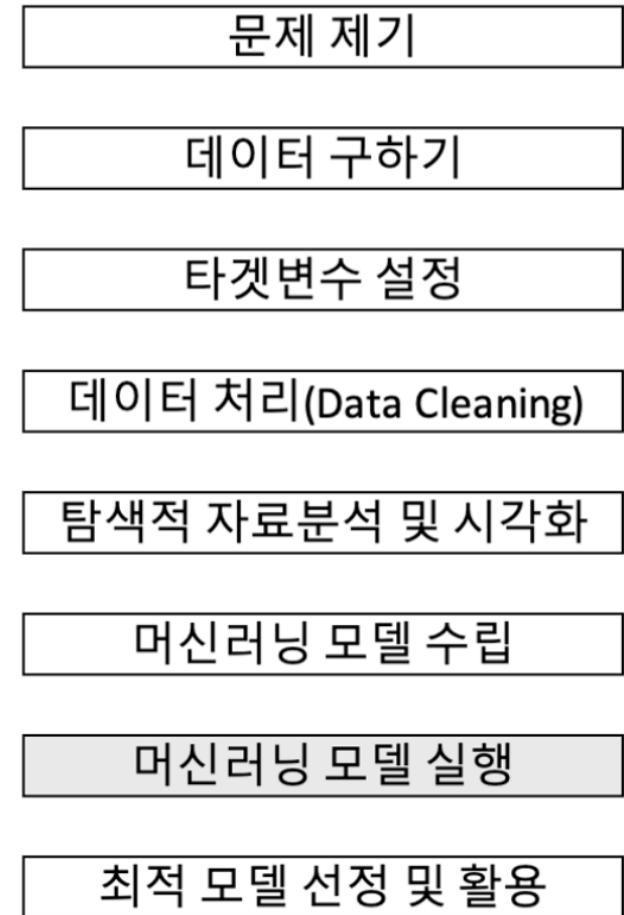
7 머신러닝 모델 실행



과정 6은
과정 7에
포함



7 머신러닝 실행 흐름도



OrdinalEncoder()

변환 예 1

	ever_married	ever_married_encoded	
No		0.0	804
Yes		1.0	3111

7-1 데이터 추가 처리 (문자형 데이터)

변환 예 2

	smoking_status	smoking_status_encoded	
Unknown		0.0	780
formerly smoked		1.0	788
never smoked		2.0	1643
smokes		3.0	704

⑦ 머신러닝 모델 실행

Train vs Test

Available Data

Training

Testing

(holdout sample)

New Available Data

Training

Validation

Testing

(validation holdout sample)

(testing holdout sample)

7-2 데이터 분할 및 대체

데이터 분할

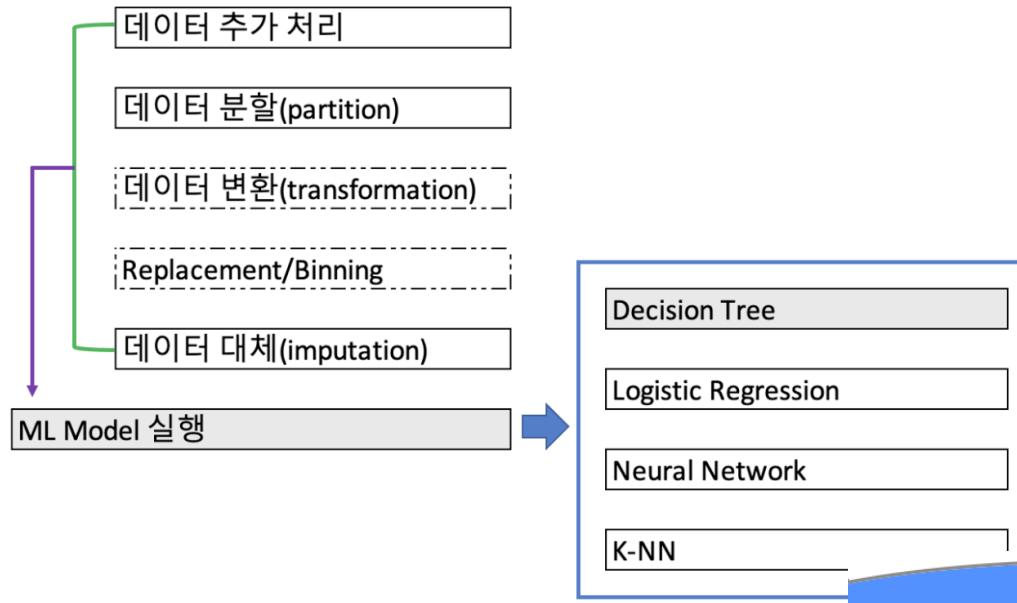
타겟변수 불균형 분포

타겟변수 분포 확인

불균형 분포 문제 Why?

▶ 해결 방법

value_counts()



Project Stroke

7 머신러닝 모델 실행

1st 모델 - 결정 트리

2nd 모델 - 로지스틱 회귀

3rd 모델 - 사이킷런 신경망 모델

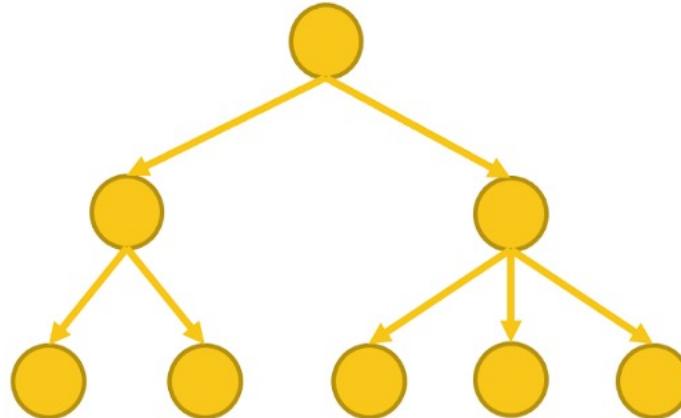
4th 모델 - (부록) tf.keras 신경망 모델

5th 모델 - 'K-NN' 모델

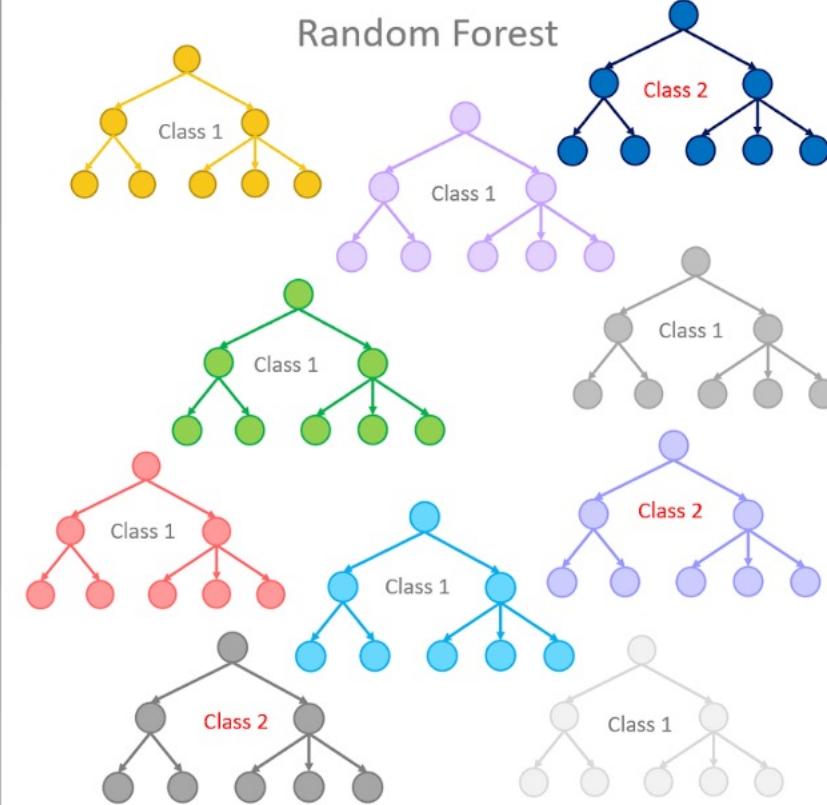
8 최적 모델 선정 및 활용

결정 트리 모델

Single Decision Tree

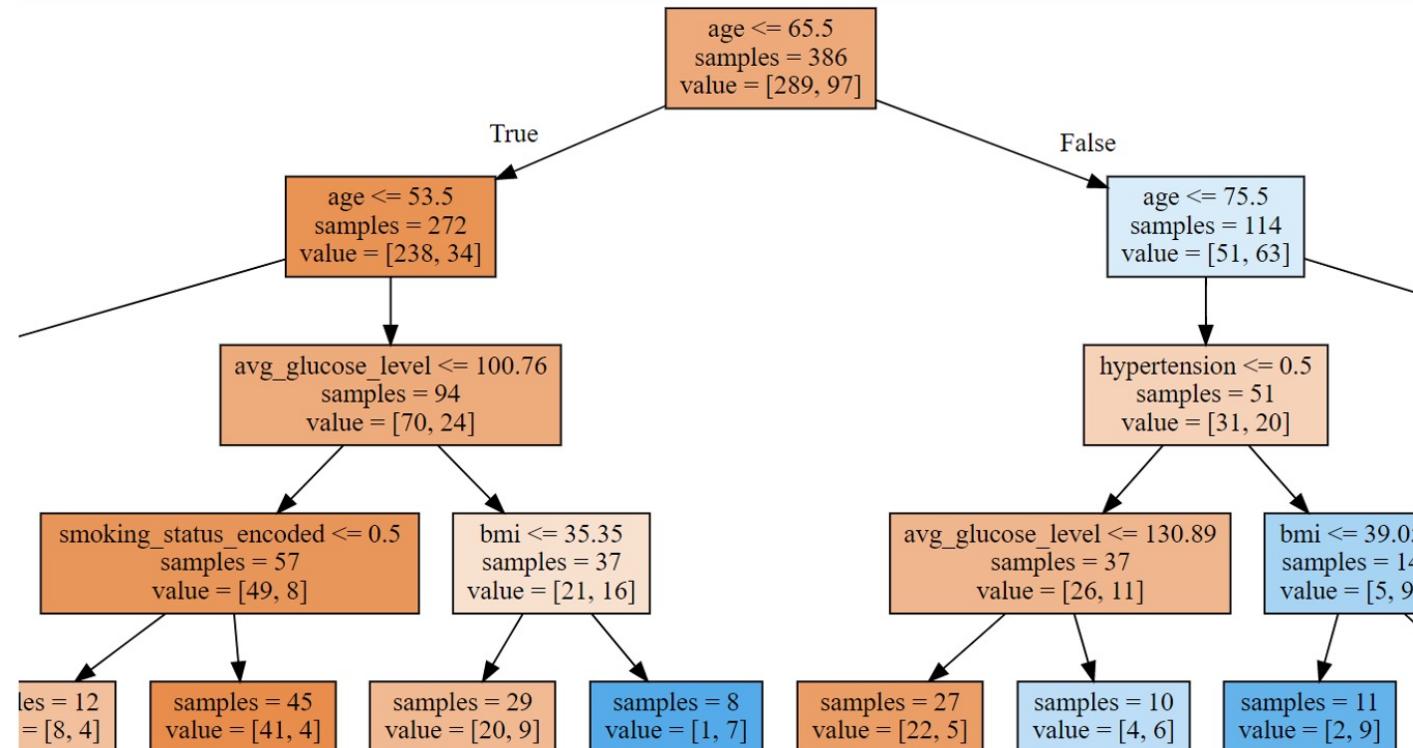


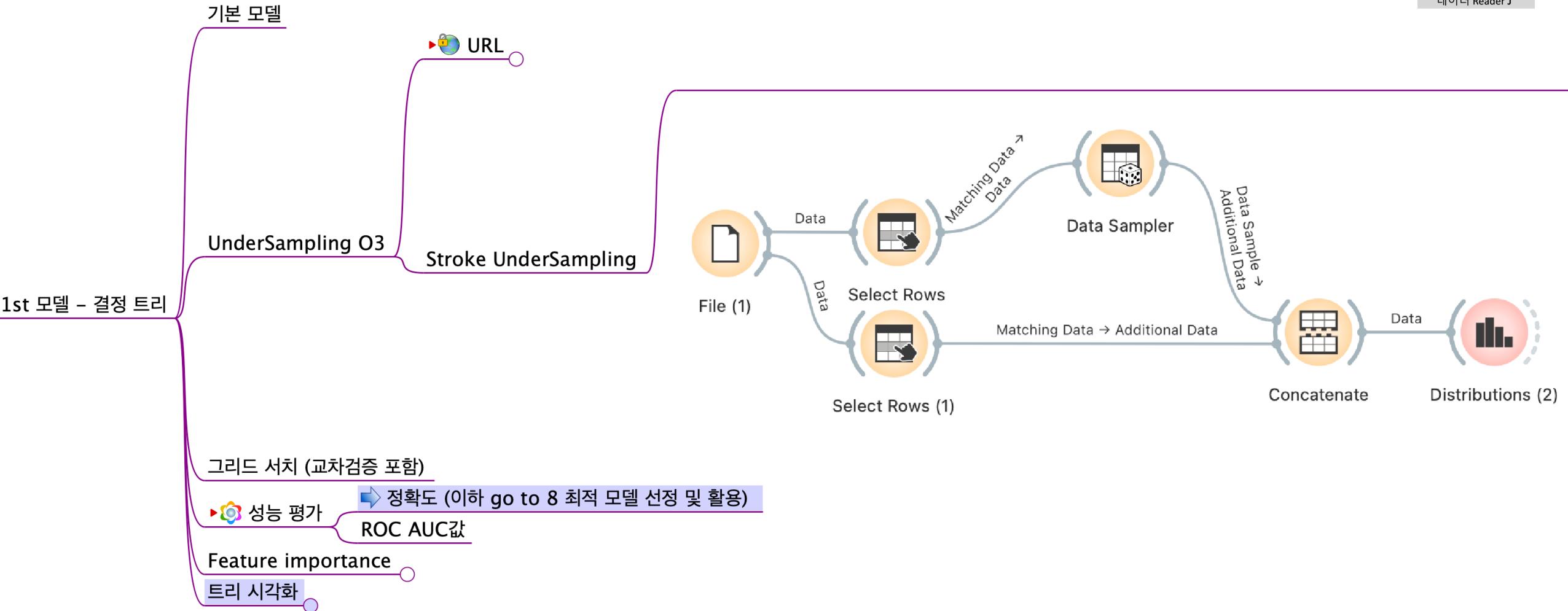
Random Forest



결정 트리 모델

결정 트리(Decision Tree)는 지도학습 중 하나이며 분류와 회귀에 모두 사용할 수 있습니다. 이 모델은 주어진 학습 데이터셋 특성으로부터 유추할 수 있는 의사결정 규칙(Decision Rules)을 학습해서, 이 규칙을 적용해 테스트 데이터셋의 타겟변수 값을 예측하는데 사용됩니다. 이렇게 데이터를 나누는 규칙(rule)은 상자 모양의 노드로 표기되고 이들 노드는 선으로 연결되어 있어서 전체적으로는 트리(tree) 모양을 형성합니다. 통상적인 트리 모양의 예시로서 이 절의 결과물을 여기에 미리 소개합니다.





기본 모델

UnderSampling O3

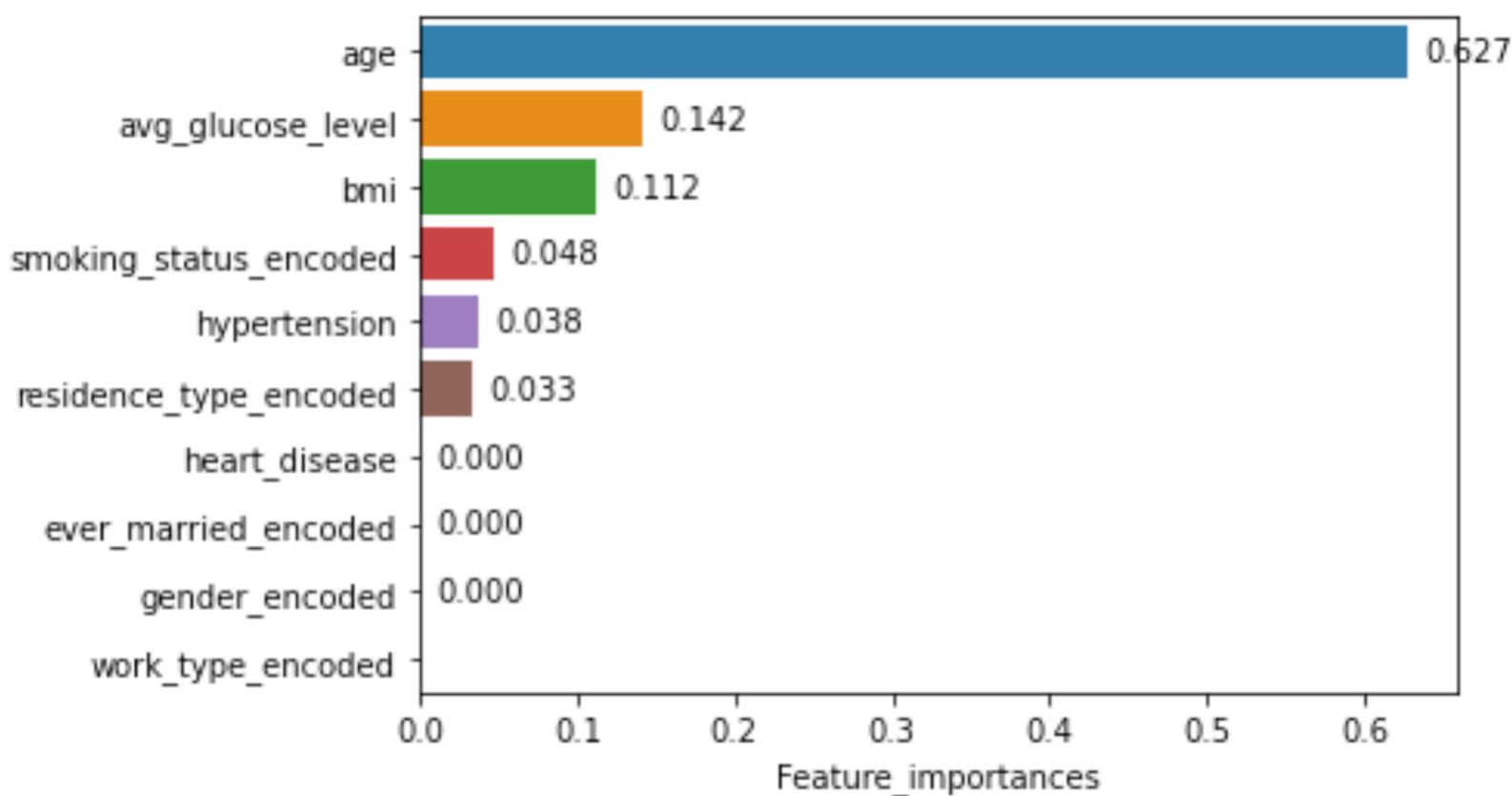
그리드 서치 (교차검증 포함)

▶ 성능 평가

1st 모델 - 결정 트리

Feature importance

트리 시각화



기본 모델

UnderSampling O3

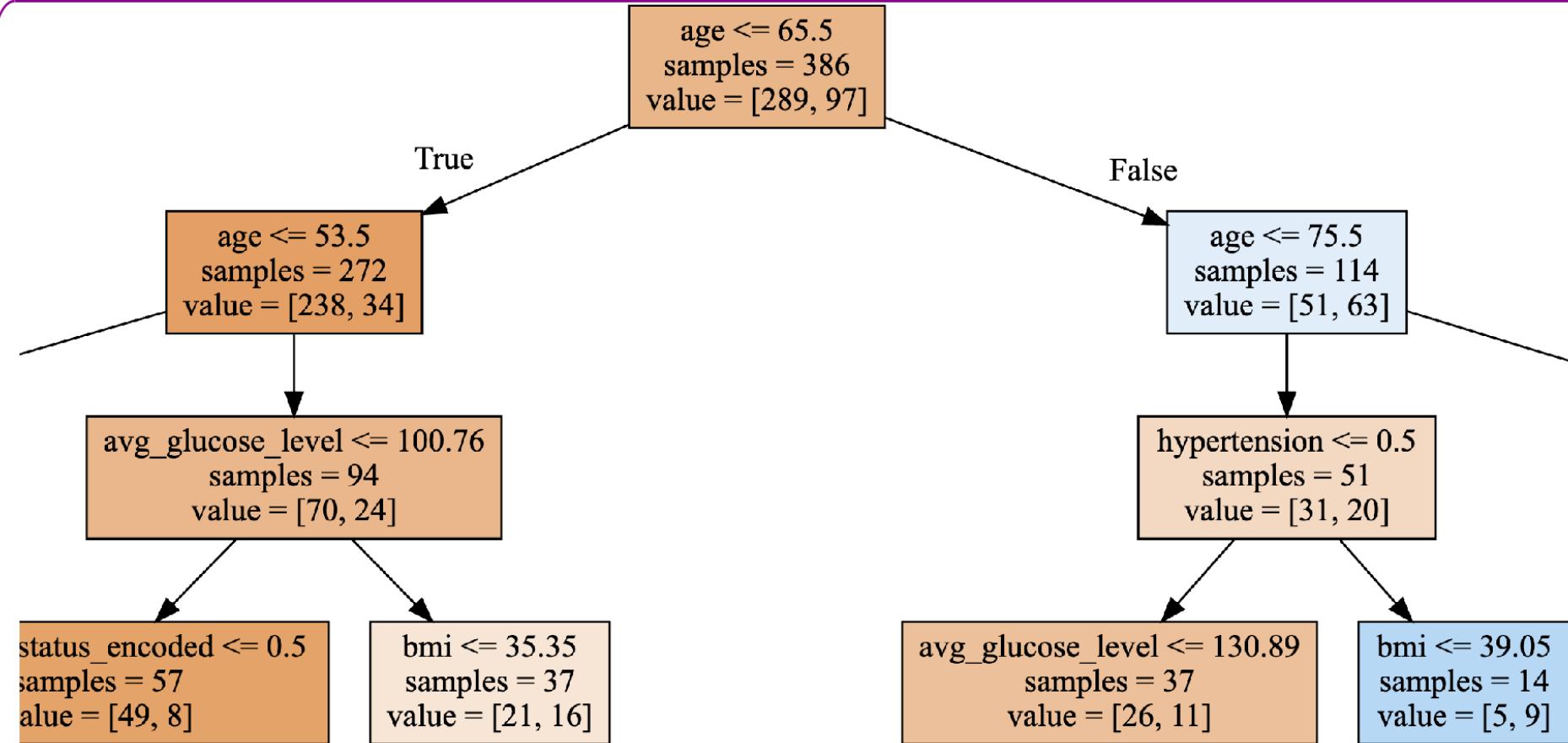
그리드 서치 (교차검증 포함)

▶ 성능 평가

Feature importance

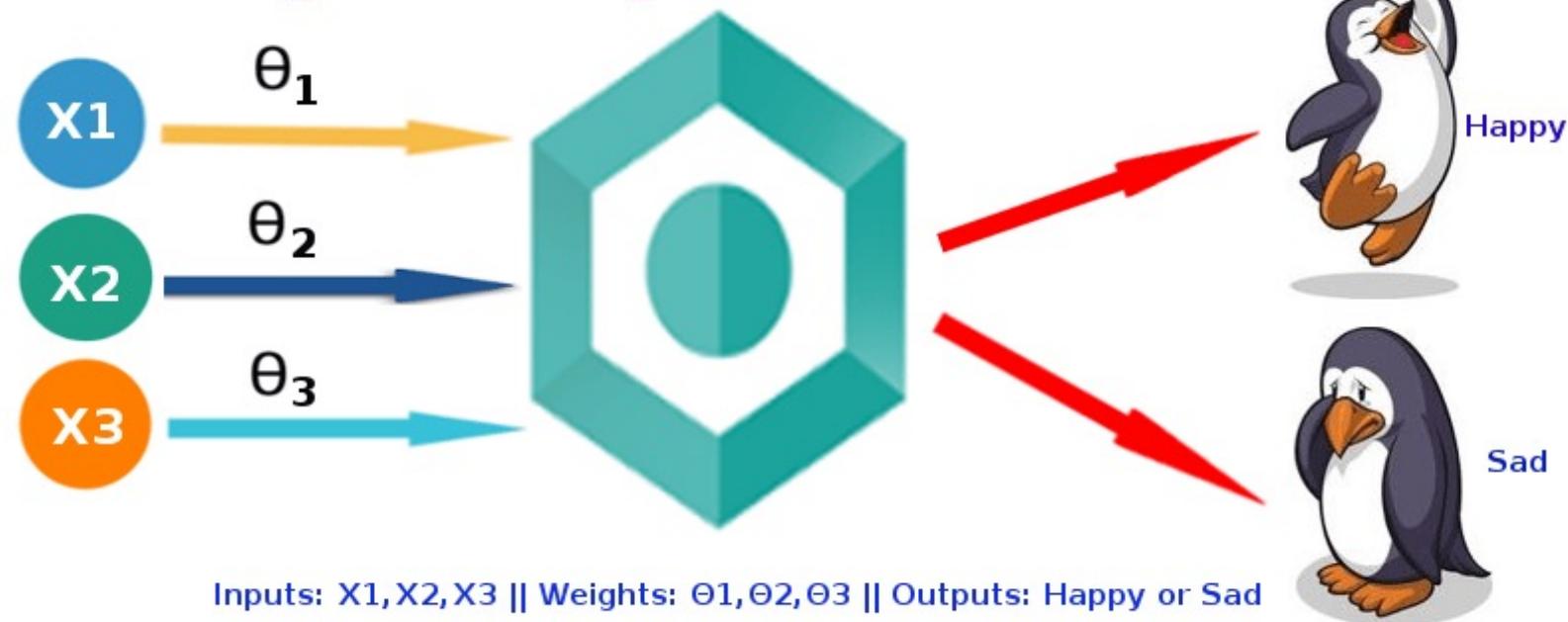
1st 모델 - 결정 트리

트리 시각화



로지스틱 회귀 모델

Logistic Regression Model



@dataaspirant.com

회귀 모델

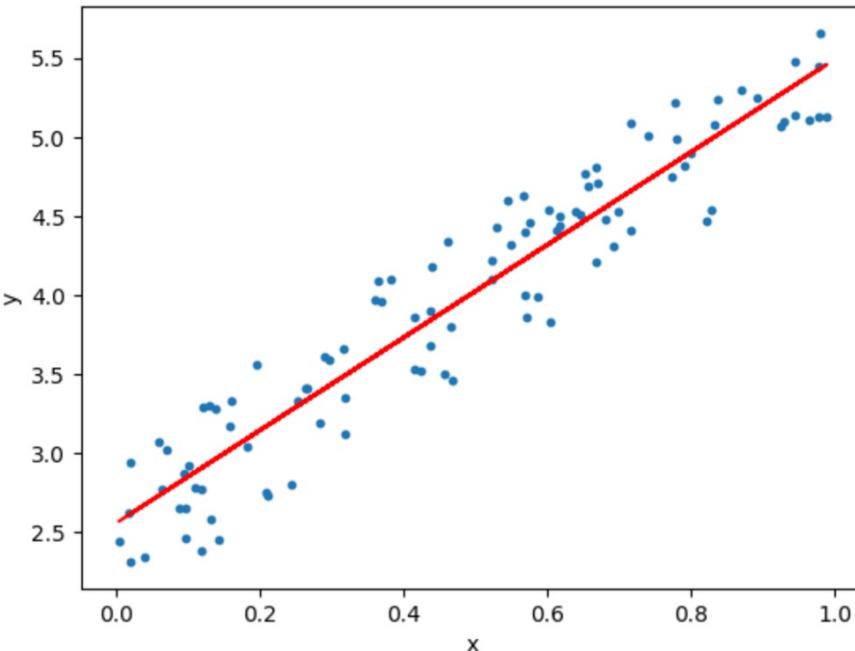
회귀 모형은 다음과 같이 나타낼 수 있습니다.

$$Y = a*X + b$$

여기서 X 는 입력변수, Y 는 출력변수입니다. a 는 계수, b 는 절편입니다. 가장 쉽게 X 입력변수가 하나인 경우를 상정해서 설명하겠습니다. X 를 자동차 고유의 최고속도라 하고 Y 를 자동차 가격이라고 설정합니다. 회귀 모형의 목표는 다음과 같습니다.

1단계: X , Y 변수가 흩뿌려진 공간에 이들을 대표해서 가로지는 선을 구합니다.

2단계: 추가로 입력변수를 주면, 1단계에서 구한 선의 수식(즉 모델)을 이용해서 출력변수를 예측합니다.



로지스틱 회귀 모델

로지스틱 회귀 모델

로지스틱 회귀(Logistic Regression)는 단순 선형 회귀 모델에서 출발했습니다. 로지스틱 회귀가 회귀 모델과 결정적으로 다른 점은 출력변수값이 이진 값(0 혹은 1)이라는 것입니다. 이 모델에서는 입력변수를 활성화 함수인 로지스틱(Logistic) 함수²⁰에 넣으면 중간결과값이 산출되는데 이 값이 임계치(threshold, 예:0.5) 이상이면 출력변수값을 1, 임계치 미만이면 출력변수값을 0으로 출력합니다.

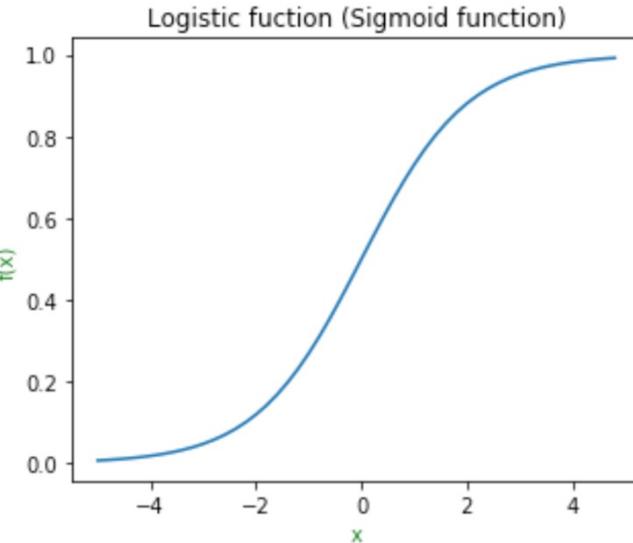


그림 로지스틱 함수

예를 들어서 위와 같은 로지스틱 함수가 주어진 상황에서, 입력 값이 $x=4$ 인 경우에 $f(x)$ 값은 1에 가까운 값이 나옵니다. $x=-2$ 인 경우에 $f(x)$ 는 0에 가까운 값이 나옵니다. 임계치가 0.5라면 $x=4$ 인 경우 $f(x)$ 값이 임계치를 넘기에 최종결과로서 출력변수값은 1이 됩니다. $x=-2$ 인 경우 $f(x)$ 값이 임계치보다 작기에 출력변수값은 0이 됩니다.

2nd 모델 – 로지스틱 회귀

1st step. 더미 변수 생성

범주형 변수에 한함

더미 변수화 불필요 변수

그렇지 않은 범주형 변수

`pd.get_dummies()`

Why 제거?

제거 기준: 비활동/부재

Variable Name	Description
gender_encoded (성별)	0. Female 1. Male 2. Other
work_type_encoded (근무 형태)	0. Govt-job 1. Never_worked 2. Private 3. Self_employed 4. Children
smoking_status_encoded (흡연 습관)	0. Unknown 1. Formerly smoked 2. Never smoked 3. Smokes

3rd step 로지스틱 회귀 모델 실행

4th step 데이터 표준화

5th step 모델 재실행 with 표준화 데이터셋

2nd 모델 - 로지스틱 회귀

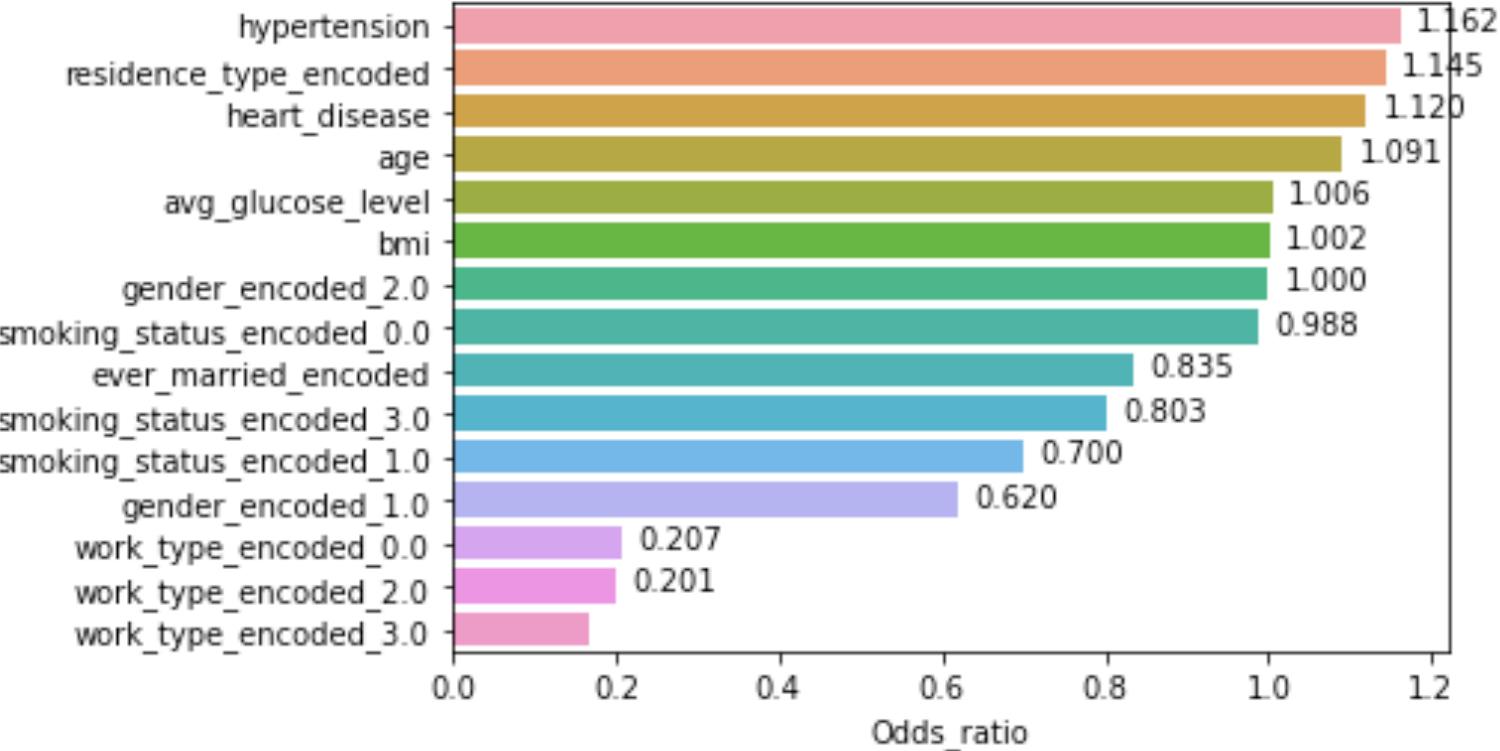
- 1st step. 더미 변수 생성
- 2nd step. 기준 더미 변수 제거
- 3rd step 로지스틱 회귀 모델 실행
- 4th step 데이터 표준화
- 5th step 모델 재실행 with 표준화 데이터셋

성능 평가 (go to 8)

3rd step 로지스틱 회귀 모델 실행

Odds ratio 그래프

Odds ratio 해석



2nd 모델 - 로지스틱 회귀

- 1st step. 더미 변수 생성
- 2nd step. 기준 더미 변수 제거
- 3rd step 로지스틱 회귀 모델 실행
- 4th step 데이터 표준화
- 5th step 모델 재실행 with 표준화 데이터셋

Odds ratio 해석

구간 변수 Odds ratio

odds ratio의 해석 (구간 변수의 경우)

나이 (age), 평균 혈당치(avg_glucose_level)의 odds ratio는 각각 1.091, 1.006입니다. 구간 변수의 odds ratio를 다음과 같이 해석합니다.

- 나이(age)가 1단위(살) 증가할 경우, 뇌졸증 경험(stroke=1)이 있을 가능성은 1.091만큼 변합니다. 즉 9.1% 증가합니다.
- 평균 혈당치(age)가 1단위 증가할 경우, 뇌졸증 경험(stroke=1)이 있을 가능성은 1.006만큼 변합니다. 즉 0.6% 증가합니다.

구간 변수의 odds ratio가 1보다 작게 나오는 경우의 해석을 알려드리기 위해, 체질량 지수(bmi)의 odds ratio가 0.992가 나오는 경우를 가정하고 해석해 보겠습니다.

- 체질량 지수(bmi)가 1 단위 증가할 경우, 뇌졸증 경험(stroke=1)이 있을 가능성은 0.992만큼 변합니다. 즉 0.8% 감소합니다.

범주형 변수 Odds ratio

2nd 모델 - 로지스틱 회귀

- 1st step. 더미 변수 생성
- 2nd step. 기준 더미 변수 제거
- 3rd step 로지스틱 회귀 모델 실행
- 4th step 데이터 표준화
- 5th step 모델 재실행 with 표준화 데이터셋

구간 변수 Odds ratio

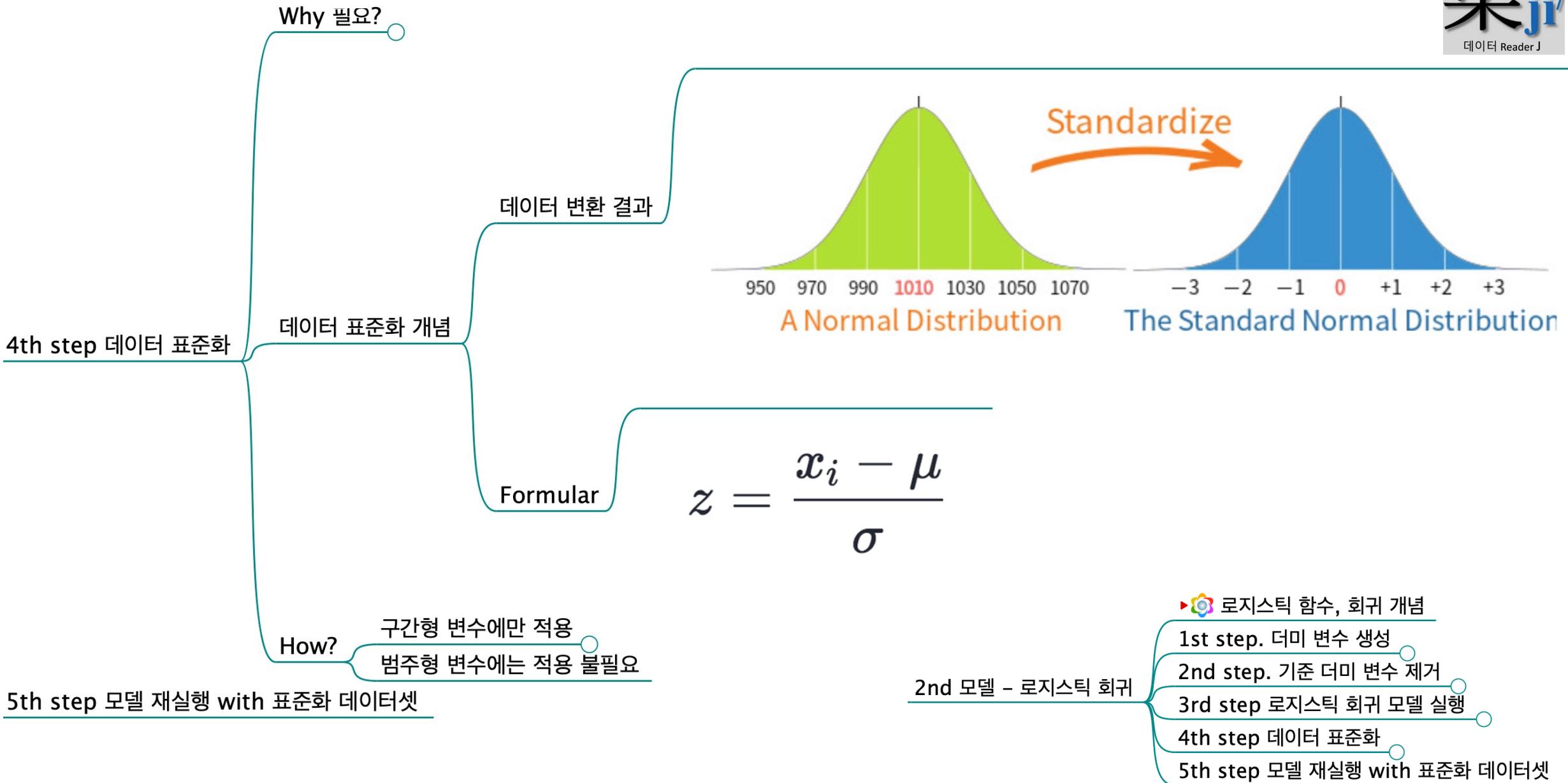
Odds ratio 해석

범주형 변수 Odds ratio

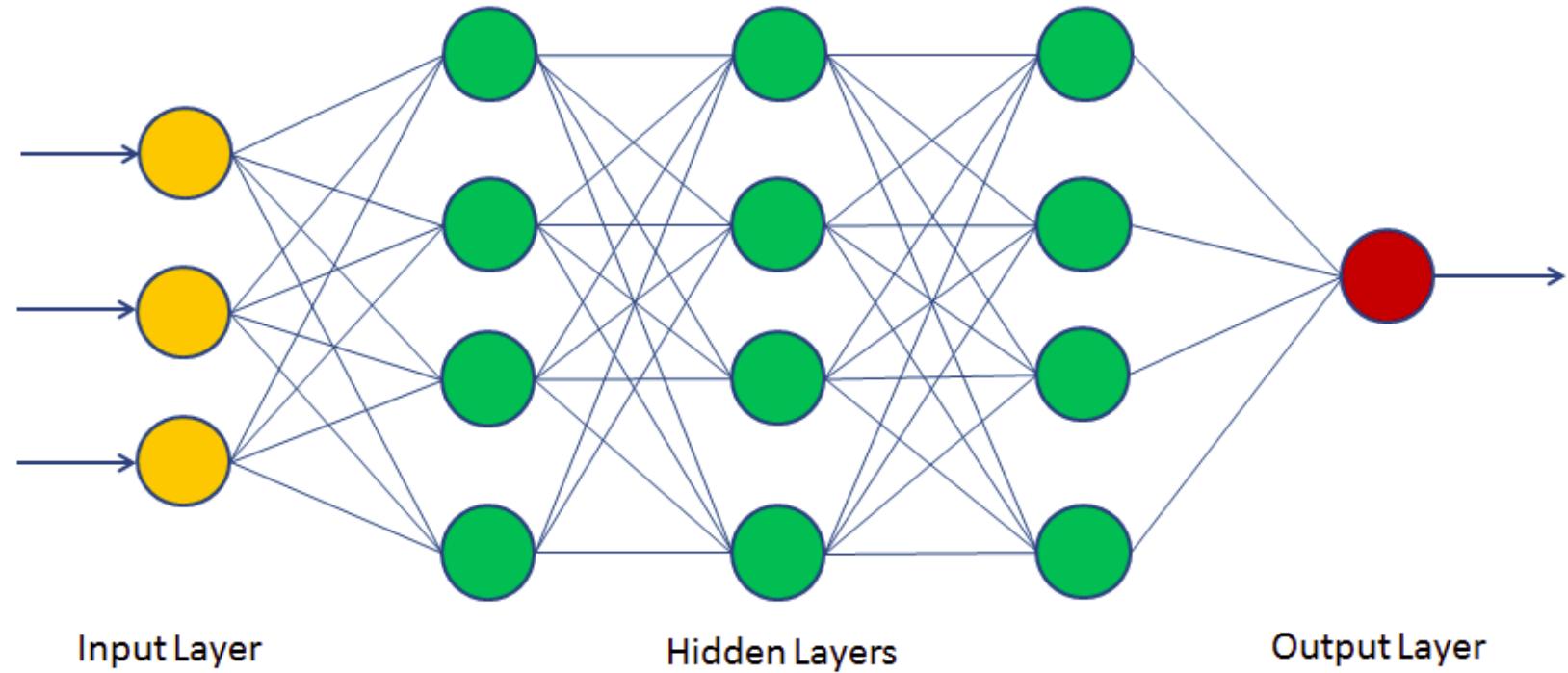
odds ratio의 해석 (범주형 변수의 경우)

고혈압 (hypertension)의 odds ratio는 1.162이고, 결혼한 경험이 있는 사람 (ever_married_encoded)의 odds ratio는 0.835입니다. 참고로 비교대상이 되는 기준(base) 고혈압 상태는 고혈압이 없는 상태이며, 기준 결혼 상태는 결혼 경험이 없는 상태입니다. odds ratio 해석은 다음과 같습니다.

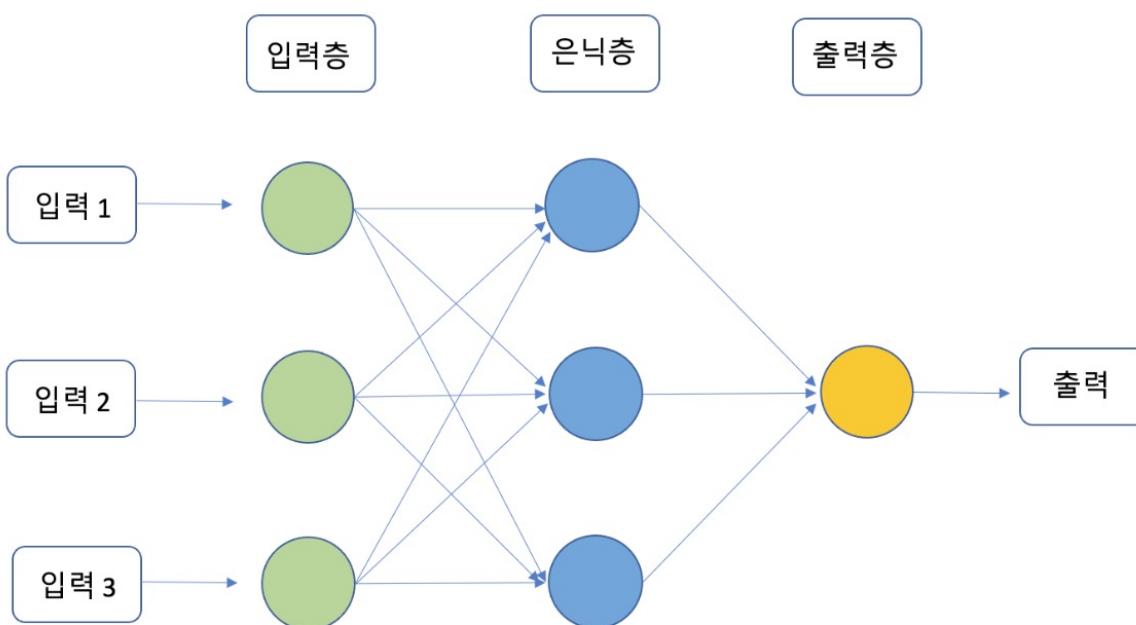
- 고혈압이 없는 경우와 비교하여, 고혈압이 있는 경우가 뇌졸증 경험(stroke=1)이 있을 가능성은 1.162배 높습니다(1.162 times higher).
- 결혼하지 않은 미혼자에 대비해서, 결혼한 적이 있거나 현재 결혼 중인 사람 (ever_married_encoded)이 뇌졸증 경험(stroke=1)이 있을 가능성은 0.835배 낮습니다(0.835 times lower).



신경망 모델

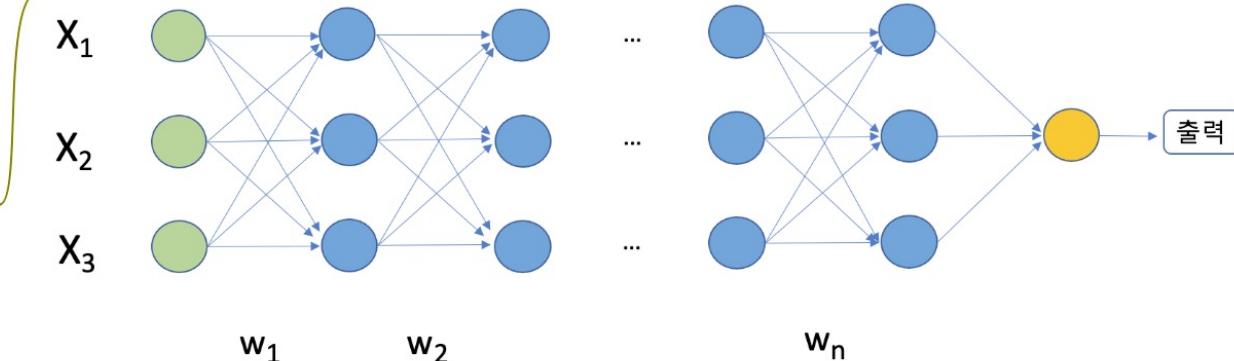


단일 신경망



⑥ 신경망

딥러닝 신경망 (완전 연결)



신경망 노드 작동 원리

이미지 분석 (합성곱 신경망)

텍스트 분석 (순환 신경망)

오토인코더

생성적 적대 신경망

단일 신경망

딥러닝 신경망 (완전 연결)



x_1

x_2

x_i

x_i

w_1

w_2

w_i

w_0

Σ

x_0

가중 총합(Weighted Sum)

신경망 노드 작동 원리

⑥ 신경망

① 활성화 함수(비선형)

가중치 조정
손실 함수

x_1

x_2

x_i

w_1

w_2

w_i

w_0

Σ

x_0

합계
Weighted Sum

활성화 함수
(비선형 함수)



출력값
(0 또는 1)

ML General

신경망 노드 작동 원리

가중 총합(Weighted Sum)
활성화 함수(비선형)

초기 가중치 부여

가중치 조정

- 가중치 조정 1
- 가중치 조정 2
- 가중치 조정 3

분류 혹은 회귀 유형

이진 분류

단일 레이블 다중 분류

다중 레이블 다중 분류

임의의 값에 대한 회귀

0과 1 사이 값에 대한 회귀

손실 함수

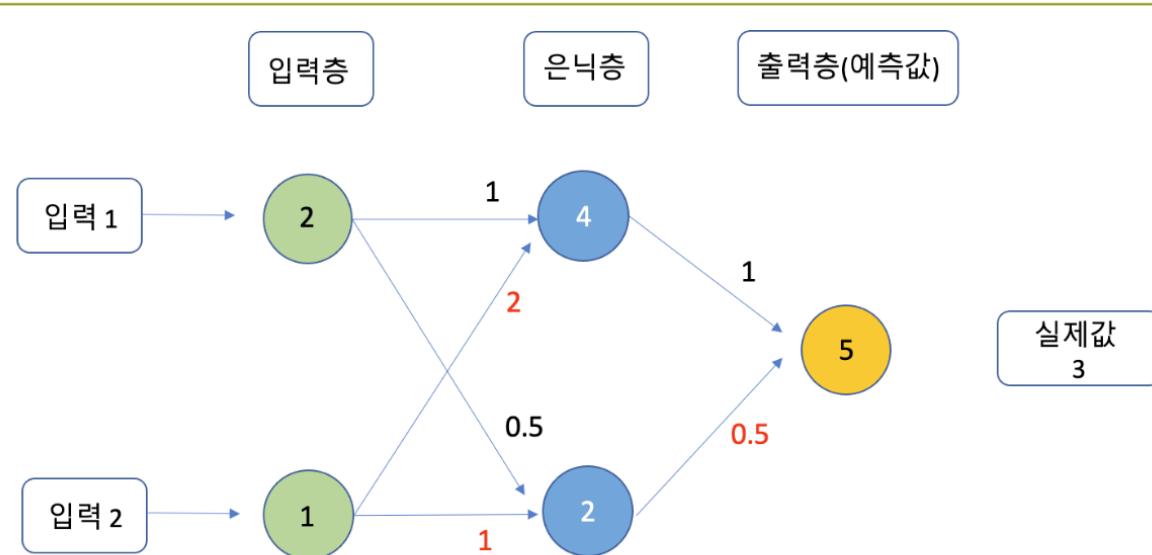
Binary_crossentropy

Categorical_crossentropy

Binary_crossentropy

Mse

Mse 또는 Binary_crossentropy





Project Stroke

1st 모델 - 결정 트리

2nd 모델 - 로지스틱 회귀

3rd 모델 - 사이킷런 신경망 모델

4th 모델 - (부록) tf.keras 신경망 모델

5th 모델 - 'K-NN' 모델

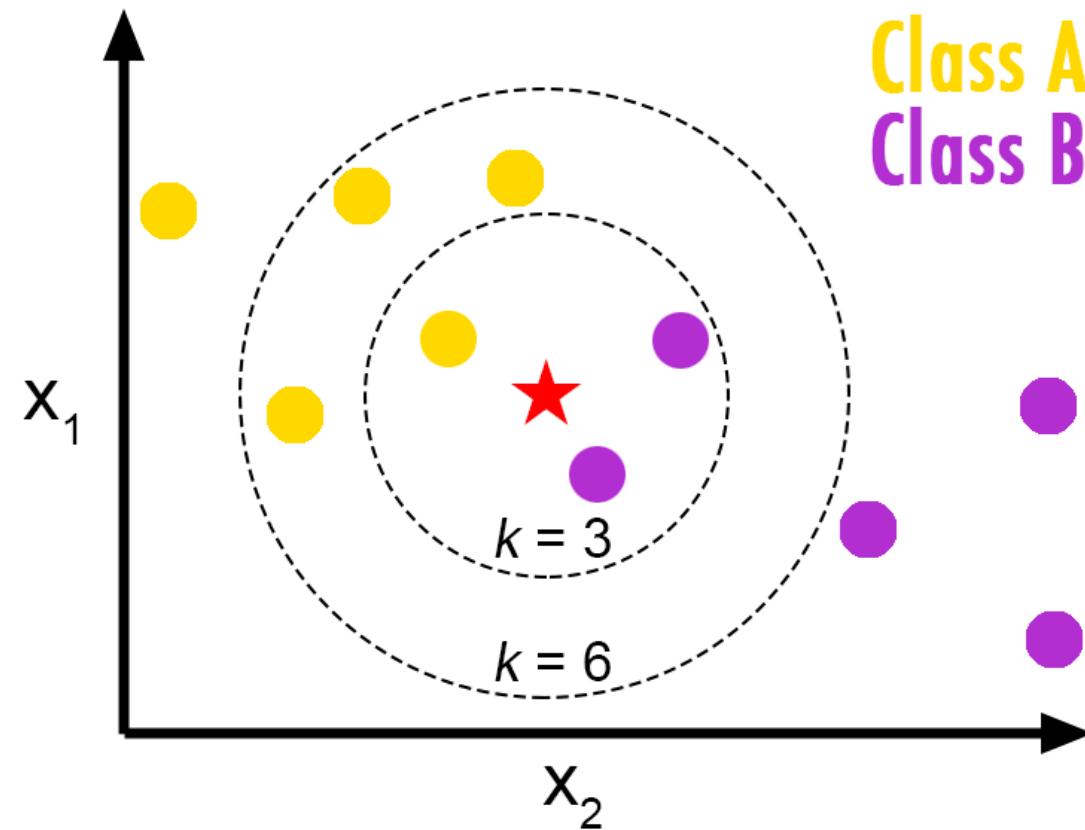
 신경망 알고리듬

표준화 데이터셋 사용

기본 모델 및 그리드 서치 (은닉층 추가)

성능 평가 (go to 8)

K-최근접 이웃(K-NN) 모델



K-최근접 이웃 모델

이 모델은 K-최근접 이웃(K-nearest neighbors, KNN) 알고리즘을 사용하여 학습합니다. K-최근접 이웃 알고리즘은 분류 및 회귀 모두에 사용할 수 있습니다. 여기서는 분류를 기준으로 설명하겠습니다. 이 알고리즘은 유유상종(Birds of a feather flock together) 원리를 이용합니다. 즉 비슷한 특성(깃털)을 지닌 데이터(새)는 서로 모인다는 원리이며 구체적인 작동 원리는 다음과 같습니다.

- 이웃(neighbors)의 개수(예: K개)를 설정합니다.
- 특정한 입력 데이터와 가장 가까운 K개의 학습 데이터를 찾습니다.
- 이들 이웃 K개의 학습 데이터의 출력변수 값(레이블)을 보고 다수결로 특정 입력 데이터의 레이블을 결정합니다.

위의 작동 원리에서 거리에 대한 정의가 필요합니다. 보통은 통상적인 거리 개념인 유클리디안 거리 개념으로 거리를 측정합니다. 그래프를 통해 시각적으로 K-최근접 이웃 알고리즘을 예시하겠습니다.

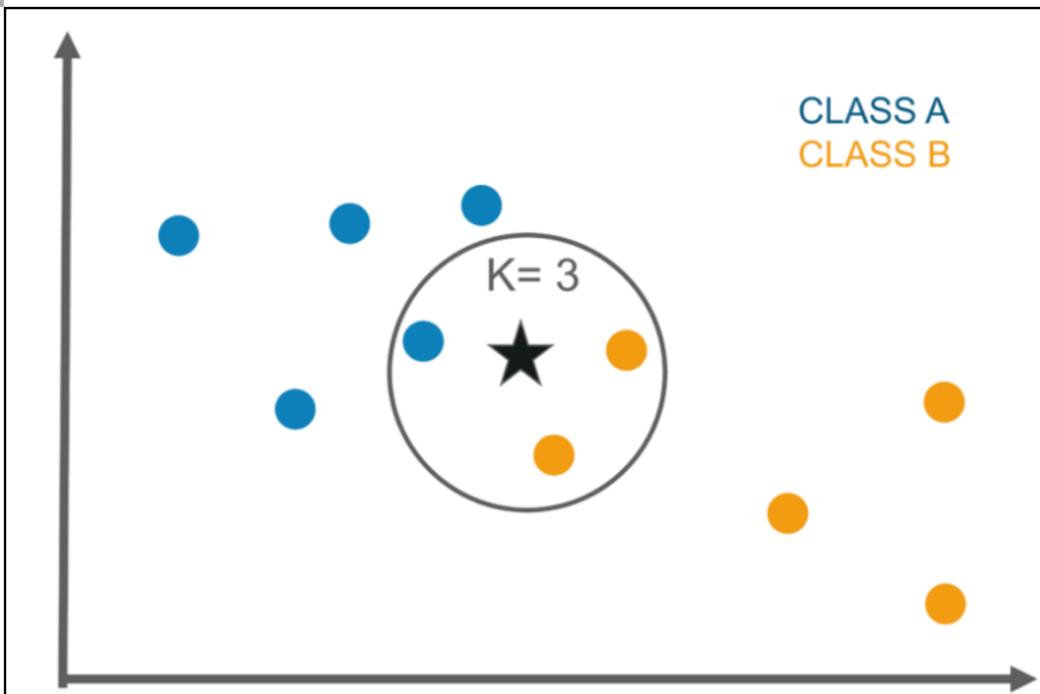
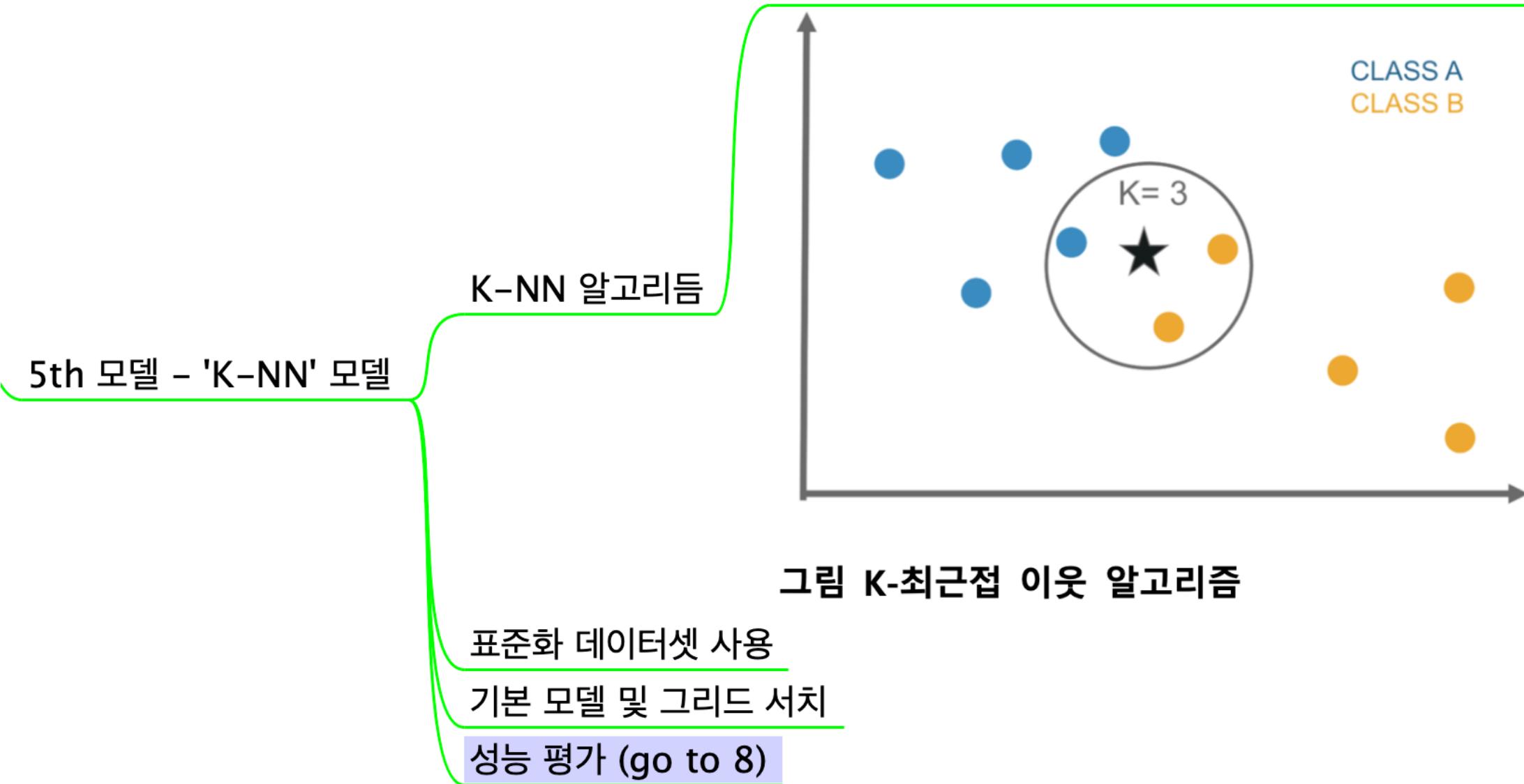


그림 K-최근접 이웃 알고리즘

출처: <https://bit.ly/3HkWHwt>

가운데 검은 별이 레이블이 결정되지 않은 분류 대상 데이터입니다. 레이블 값을 비교할 이웃의 개수를 3개라고 하겠습니다. 그럼 가장 가까운 거리 범위 안에 3개의 학습 데이터의 레이블을 비교하면 됩니다. 이들 세 개의 점 중에 파란색 점이 1개, 주황색 점이 2개입니다. 다수결의 원칙에 의해 검은 별의 레이블은 주황색으로 결정합니다. 이러한 방식으로 K-최근접 이웃 모델은 입력변수의 레이블을 판별합니다. 이 과정을 반복해서 분류를 완성합니다.



8

최적 모델 선정 및 활용

8 최적 모델 선정 및 활용

모델 성능 비교표

활용 방안

최적 모델 보고

모델명	Accuracy	순위
Decision Tree	0.76943	
Logistic Regression	0.75130	
Logistic Regression*	0.75389	
Neural Network*	0.75648	
K-NN*	0.77720	1

*StandardScaler로 표준화된 데이터셋 사용

⑧ 최적 모델 선정 및 활용

모델 성능 비교표

활용 방안

최적 모델 보고

활용 방안

머신러닝 결과의 활용

- **예측(Scoring)**: 향후에 최적 모델을 이용해서 새로운 환자의 뇌졸증 가능성을 예측하는 용도
- **기존 문헌연구 확인(Confirmation)**: 이번 프로젝트에서 발견된 중요한 변수들이 기존의 문헌에서 나오는 중요변수들과 일치하는지를 확인하는 용도
- **탐색적 연구(Exploratory study)**: 여러가지 모델 중에서 공통적으로 도출되는 중요한 변수를 살펴서 기존문헌에서 알려지지 않았던 변수를 찾아내는 탐색적인 용도

활용 예 1

활용 예 2

⑧ 최적 모델 선정 및 활용

모델 성능 비교표

활용 방안

최적 모델 보고

머신러닝 결과의 활용

연구주제: 뇌졸증 발병 요인 중 중요한 요인은 무엇인가?

활용 방안

활용 예 1

우리의 결론은 무엇일까요? 결정 트리 모델과 로지스틱 회귀 모델에 따르면 구간 변수 중에서는 특히 나이(age)가 중요한 역할을 합니다. 나이가 들면 뇌졸증을 겪을 가능성이 높아지니 무리한 상황을 만들지 않는 것이 좋겠습니다. 또한 평균 혈당치(avg_glucose_level)가 높으면 뇌졸증 발병 확률이 약간 증가하고, 체질량 지수(bmi)가 높으면 뇌졸증 발병 확률이 약간 떨어집니다. 이 결과를 알고 있으면 식이요법이나 운동으로 평균 혈당치를 줄이거나, 일정량의 체질량 지수를 유지하기 위한 예방적인 조치를 취할 수 있습니다. 여러 범주형 변수 중에서는 특히 고혈압이 뇌졸증과 관련이 깊은 것으로 보입니다.

활용 예 2

⑧ 최적 모델 선정 및 활용

모델 성능 비교표

활용 방안

최적 모델 보고

머신러닝 결과의 활용

활용 예 1

아울러 이 장의 분석 프로젝트는 추가적인 머신러닝 모델 추가를 통해 퍼포먼스를 최대로 끌어올릴 수 여지가 있습니다. 이를 통해 도출한 최종 모델을 통해 병원에 오는 환자들의 진단 예측이 더 정확해지고, 교육이나 홍보를 통해 뇌졸증을 예방하는 올바른 생활 습관을 보급할 수 있습니다. 수만 명의 동일 인물에 대한 수십 년간의 관찰 자료를 수집하여 분석할 있다면, 현재 젊은 세대가 노년이 될 때 뇌졸증에 걸릴 확률을 계산할 수 있을 것입니다. 그 결과로서 의료/생명보험산업의 요율표 책정, 국가의 의료/연금 예산 계획 수립 시 활용할 수 있게 됩니다. 제약 업체와 병원 등은 특정 연령대의 타겟 고객군을 대상으로 광고를 집행하거나 치료제 개발을 염두에 둘 수 있습니다. 때문에 머신러닝 기법은 학계와 공공분야 뿐만 아니라 산업계에도 막대한 파급효과를 불러옵니다.

활용 방안

활용 예 2

⑧ 최적 모델 선정 및 활용

모델 성능 비교표

활용 방안

최적 모델 보고

성능지표 정확도의 보고 시 해석

정확도는 이진값 분류 모델에서 모델이 타겟변수 예측한 값과 실제 타겟변수값과 일치할 확률입니다.

성능지표 ROC AUC값 보고 시 해석

이진값 분류 모델에서 ROC AUC 값이 1에 가까이 갈수록 우수한 모델입니다.

성능지표 결정계수 R^2 의 해석

연속형 타겟을 사용하는 회귀모델에서 결정계수 R^2 는 타겟변수 전체의 분산 중에서 모델이 설명하는 분산 비율을 의미합니다.

보고 예 1

최적 모델 보고

보고 예 2

보고 예 3

⑧ 최적 모델 선정 및 활용

모델 성능 비교표

활용 방안

최적 모델 보고

보고 예 1

결정 트리 모델 계열은 변수의 Feature importance 그래프 보고

보고 예 2

분류 모델의 로지스틱 회귀(릿지, 라소 포함) 모델은 odds ratio 보고

연속형 타겟의 회귀(릿지, 라소 포함) 모델은 계수 보고

최적 모델 보고

보고 예 3

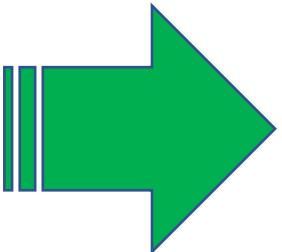
신경망 모델과 SVM 모델은 퍼포먼스 지표 해석 이외에 다른 해석은 용이치 않음



다시 한 번 강조!

Orange3는 빠른 review용 or ML 학습용
단, 실전은 꼭 Python으로!!!

The screenshot shows the Orange3 interface with a central workspace containing a workflow diagram. The diagram consists of several nodes connected by arrows: a 'File (health csv)' node connected to a 'Feature Statistics (3)' node; a 'CSV File Import (health csv,bmi OK)' node connected to a 'Feature Statistics (2)' node; and a 'Distributions (1)' node. The 'Feature Statistics' nodes have descriptive text below them: 'File (health csv) bmi meta -- type role change needed' and 'bmi not showing'. The 'Feature Statistics (2)' node also has the text 'Data' above it. The 'Distributions (1)' node has a yellow warning icon above it. To the left of the workspace is a sidebar with tabs for 'Data', 'Visualize', and 'Model', and a list of available models including Constant, CN2 Rule Induction, Calibrated Learner, kNN, Tree, Random Forest, Gradient Boosting, and SVM.



The screenshot shows a Jupyter Notebook titled 'Stroke Data Cleaning 2 v2.ipynb'. The code cell contains the following Python code:

```
[ ] import seaborn as sns
sns.histplot(data=df, x="age", hue="stroke", bins=20);
```

Below the code is a histogram plot titled 'Count' vs 'age'. The x-axis ranges from 20 to 80, and the y-axis ranges from 0 to 250. The plot shows two distributions: one for 'stroke = 0' (blue bars) and one for 'stroke = 1' (grey bar). The 'stroke = 0' distribution peaks around age 40, while the 'stroke = 1' distribution is much lower and appears to be centered around age 60.

파이썬 머신러닝 학습 팁!

이론보다는 실습!

실습을 최대한 많이!

다루는 데이터를 다양하게 많이!

코딩 에러에 Don't panic! – 코딩 에러는 학습의 동반자!

어려운 코딩 에러는 웹 혹은 전문가에 문의하여 해결

파이썬 머신러닝 학습 팁!

Orange3로 최대한 직관적으로 학습 시작!

그러나 본격적인 머신러닝은 반드시 파이썬으로!!!

파이썬은 구글 코랩에서 실행!

캐글과 국내 데이터 경진대회 데이터 분석 프로젝트 학습 병행!

데이터 분석 프로젝트에 참여하면 실력이 급격히 향상!

강사 연락처 및 추가 정보

강사 이메일: jasonyimg@gmail.com, jasonyim@naver.com 카카오톡 아이디: [jasonyimkakao](#)

2022년 6월 경 "(가제) 미국 수업에서 배운 파이썬 머신러닝 " 출간 예정

Orange3 사용법 강의

마인드맵 S/W FreePlane 강의

중상급 파이썬 머신러닝 프로젝트 추가 강의 다수



Lyrics: Love Actually Scene - Christmas Is All Around

Billy Mack

• TSMA_JR_

Machine Learning is all around.

All you need is Python.

