1. 철강산업 안전사고 데이터를 이용한 안전사고 분류 모델

1.1. 데이터 확인

Data Shape: 4844 rows x 16 columns

순서	컬럼명	설명
1	date_time	사고 발생 시간
2	employee_id	직원 ID
3	risk_level	담당 업무의 위험도
4	work_type	작업 유형
5	project_type	산업군 유형
6	part_of_body	상해 부위
7	injury_type	상해 유형
8	accident_type	사고 종류
9	warning1	작업시 주의사항 대분류
10	warning2	작업시 주의사항 소분류
11	human_factor1	사고발생 인적요소 대분류
12	human_factor2	사고발생 인적요소 소분류
13	task_assigned	정기적 작업 여부
14	project_cost	프로젝트 비용
15	safety_edu	안전교육 받은 횟수
16	label	사고발생 결과

1.2. 데이터 전처리 방법

1.2.1. 데이터 불균형 처리

"심각한 골절과 절단 위험" 데이터는 7개만 있기 때문에 "골절과 절단 위험" 데이터와 동일한 label_no를 부여하여 데이터 불균형 해소

label	label_no	데이터개수
골절 위험	0	1272
골절과 절단 위험	1	533
심각한 골절과 절단 위험	2	7
심각한 골절 위험	3	648
사망 위험	4	2384

label	label_no	데이터개수
골절 위험	0	1272
골절과 절단 위험	1	540
심각한 골절과 절단 위험	1	
심각한 골절 위험	3	648
사망 위험	4	2384

1.2.2. Feature Extraction

기존 컬럼 데이터를 가지고 추가적으로 컬럼 생성. Date_time 데이터를 datetime64 형식으로 변경한 이후 월, 일, 시, 분 데이터 추출

date_time	month	day	hour	minute
2015-07-01 08:30:00	07	01	08	30
2015-08-02 13:15:00	08	02	13	15

1.2.3. Feature Selection

모델에 학습시키고자 하는 데이터는 프로젝트 목표에 맞게 선택되어야 함. 목표는 사고 발생 이전의 기록들을 통해 어떤 사고가 발생할 수 있는지 경고를 주는 것으로 필요없는 컬럼들은 삭제하고, 필요한 컬럼만 선택하여 학습

- ⑩ 사고 발생 이후 데이터는 사용할 수 없음: part_of_body, injury_type, accident_type
- ⑩ 결측값이 많거나 상관관계가 적은 컬럼은 선택하지 않음: date_time, employee_id, work_type, project_type, project_cost

1.2.4. 결측치 처리

위에서 선택한 컬럼들 중 결측값이 있다면 drop을 하거나 통계값으로 결측치를 대체할 수 있음. 결측치가 많이 없기 때문에 drop하여 24개의 데이터 행 버림

```
data.isna().sum()

risk_level        0
warning1          0
warning2          7
human_factor1     0
human_factor2     7
task_assigned     0
safety_edu       22
month             0
day               0
hour              0
minute            0
label_no          0
dtype: int64
```

```
# 결측치가 많이 없기 때문에 drop
print('결측치 drop 전', data.shape)
data = data.dropna()
print('결측치 drop 후', data.shape)

결측치 drop 전 (4844, 12)
결측치 drop 후 (4820, 12)
```

1.2.5. Label Encoding

모델학습을 위해서는 모든 데이터가 숫자형으로 변환되어야 함. 사용하는 컬럼들 중 범주형으로 값이 들어있는 warning1, warning2, human_factor1, human_factor2, task_assigned 컬럼들에 대해 Label Encoding 수행

1.3. 사용 모델

- ⑩ Random Forest Classification
- ⑩ XGBoost Classification

1.4. 산출물

- ⑩ 정제된 학습용 데이터.csv
- ⑩ Python 모델 학습 및 실행파일.ipynb
- ⑩ Orange3 모델 학습 및 실행파일.ows

2. 철강산업 모터센서 데이터를 이용한 모터 고장유형 분류 모델

2.1. 데이터 확인

⑩ 데이터 크기: 929.9MB

⑩ 데이터 구성: 82개 디렉토리, 9996개 파일

순서	컬럼명	설명
1	Date	데이터 수집일
2	Filename	데이터 파일 이름
3	Data_Label	고장 유형
4	Label_No	고장 유형 고유 번호
5	Motor_spec	모터 rpm, 정격 출력, 정격 전류
6	Period	수집 시간
7	Sample_Rate	수집 신호 샘플 주파수
8	RMS	고장 유형에 따른 실효값
9	Data_Length	데이터 길이
10	1 – 12000	진동 데이터

2.2. 데이터 전처리 방법

2.2.1. 데이터프레임화

RAY 프레임워크 활용해 9996개의 개별 csv파일을 하나의 데이터프레임으로 병합

병합한 데이터프레임 shape: 9996 rows x 12012 columns

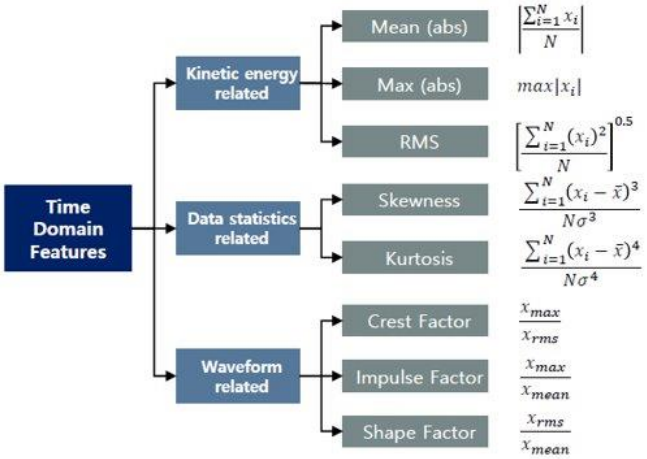
1	Date	2020-12-02 12:44:26
2	Filename	STFMK-20201105-LW19-1935_20201202_124426_004.dat
3	Data Label	측정멀불량
4	Label_No	03
5	Motor Spec	L-CAHU-01R, 1765,11, 21.8,
6	Period	3SEC
7	Sample Rate	4000
8	RMS	0.009861,
9	Data Length	12000,
10	0	0.0081095211,
11	0.00025	-0.0030722495,
12	0.0005	0.0117186196,
13	0.00075	0.0095781777,
14	0.001	0.0001816829,
15	0.00125	0.0104843080,
16	0.0015	0.0055182031,
17	0.00175	-0.0081522893,
18	0.002	0.0063686399,
19	0.00225	0.0036040884,
20	0.0025	-0.0096390676,
21	0.00275	0.0052996846,

	Date	Filename	Data Label	Label_No	Motor Spec	Period	Sample Rate	RMS	Data Length	RPM	KW	VR	0	1	2
0	2020-11-27 11:40:02	STFMK-20201105-LW19-2249_20201127_114002_004.dat	정상	0	R-CAHU-03R	3SEC	4000	0.006673	12000	1760.0	7.5	15.0	-0.000375	-0.000173	0.002411
1	2020-11-27 11:52:47	STFMK-20201105-LW19-2249_20201127_115247_004.dat	정상	0	R-CAHU-03R	3SEC	4000	0.006160	12000	1760.0	7.5	15.0	-0.006714	-0.008092	-0.002692
2	2020-11-27 10:50:56	STFMK-20201105-LW19-2249_20201127_105056_004.dat	정상	0	R-CAHU-03R	3SEC	4000	0.006777	12000	1760.0	7.5	15.0	-0.018027	-0.011090	-0.012713
3	2020-11-27 12:46:11	STFMK-20201105-LW19-2249_20201127_124611_004.dat	정상	0	R-CAHU-03R	3SEC	4000	0.006601	12000	1760.0	7.5	15.0	0.007564	0.012273	0.010702
4	2020-11-27 15:26:44	STFMK-20201105-LW19-2249_20201127_152644_004.dat	정상	0	R-CAHU-03R	3SEC	4000	0.005785	12000	1760.0	7.5	15.0	0.005307	-0.000428	0.006611
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9991	2020-12-19 14:16:43	STFMK-20201105-LW19-2328_20201219_141643_004.dat	회전체 불평형	2	L-CAHU-01R	3SEC	4000	0.015048	12000	1750.0	11.0	22.0	0.012692	0.021566	0.011371
9992	2020-12-21 16:30:21	STFMK-20201105-LW19-2328_20201221_163021_004.dat	회전체 불평형	2	L-CAHU-01R	3SEC	4000	0.015621	12000	1750.0	11.0	22.0	0.001178	0.011818	0.009177
9993	2020-12-01 06:57:22	STFMK-20201105-LW19-2328_20201201_065722_004.dat	회전체 불평형	2	L-CAHU-01R	3SEC	4000	0.016222	12000	1750.0	11.0	22.0	-0.026664	-0.017909	-0.002067
9994	2020-12-01 07:14:01	STFMK-20201105-LW19-2328_20201201_071401_004.dat	회전체 불평형	2	L-CAHU-01R	3SEC	4000	0.014851	12000	1750.0	11.0	22.0	0.013695	0.024506	0.016657
9995	2020-12-23 15:36:39	STFMK-20201105-LW19-2328_20201223_153639_004.dat	회전체 불평형	2	L-CAHU-01R	3SEC	4000	0.015240	12000	1750.0	11.0	22.0	0.024740	0.001284	-0.014873

9996 rows × 12012 columns

2.2.2. 시간 특성 추출

12000개의 센서 데이터에서 시간 특성 추출



2.2.3. Feture Selection

목표는 모터 센서 데이터 시간 특성을 추출해 모터의 고장 유형을 분류하는 모델로, 학습에 사용할 컬럼 선택

- ⑩ 전체 컬럼: Date, Filename, Data_Label, Motor_Spec, Period, Sample_Rate, RMS, Data_Length, RPM, kW, VR, abs_max, abs_mean, skewness, kurtosis, creset, ptp, rms, impulse, shape
- ⑩ 선택한 컬럼: kW, abs_max, abs_mean, skewness, kurtosis, creset, ptp, rms, impulse, shape

2.3. 사용 모델

- ⑩ GMB(Gradient Boosting Machine)
- ⑩ AdaBoost
- ⑩ CatBoost(Categorical Boosting)
- ⑩ Decision Tree
- ⑩ Random Forest Classification
- ⑩ Extra Trees

2.4. 산출물

- ⑩ Python 학습용 dataset
- ⑩ Orange3 학습용 dataset
- ⑩ Python 모델 학습 및 실행 파일.ipynb
- ⑩ Orange3 모델 학습 및 실행 파일.ows

3. 철강산업 현장 이미지 데이터를 이용한 헬멧 탐지 모델

3.1. 데이터 확인

- ⑩ 이미지 객체탐지 데이터

  Helmet, Vest, Head 3가지 라벨 존재

  안전모, 안전조끼, 안전모 없는 머리에 대한 라벨을 가지고 있는 현장 이미지와 일상 이미지를 포함한 png, jpg 이미지 파일 4427개 + 라벨 정보를 담고 있는 txt 파일 4427개

- ⑩ 이미지 분류 데이터

  Helmet, Head 2가지 라벨 존재

  안전모를 착용한 사진과 안전모를 착용하지 않은 png, jpg 이미지 파일 2242개, 라벨값은

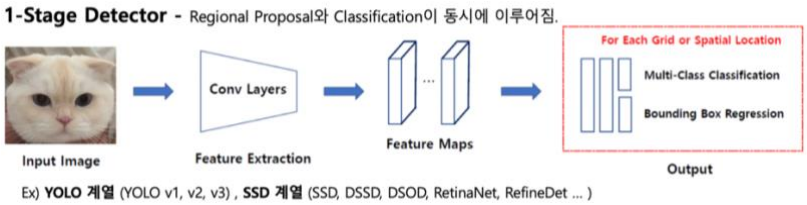
폴더명으로 구분

3.2. Object Detection

- ⑩ 영상처리와 컴퓨터 비전 분야에서 많이 활용되는 기술
- ⑩ 특정 이미지에서 물체를 정확하게 탐색하고 분류

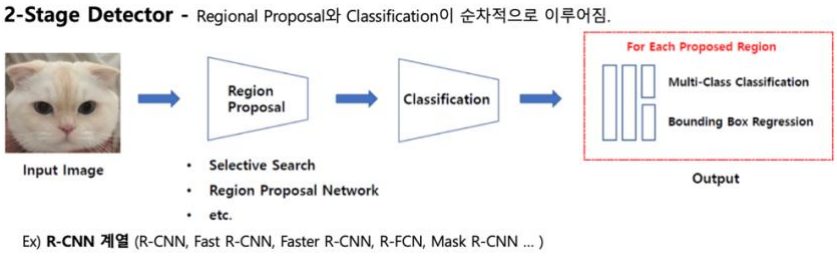
3.2.1. 1-stage Detector

- ⑩ 입력 이미지를 한번에 처리하여 객체의 위치와 클래스 예측
- ⑩ 빠른 처리속도와 간단한 구조를 가짐



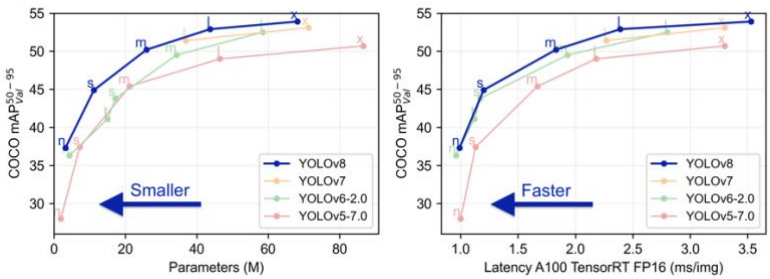
3.2.2. 2-stage Detector

- ⑩ 입력 이미지를 두 단계로 처리하여 객체의 위치와 클래스 예측
- ⑩ 1단계에서 Region Proposal 생성
- ⑩ 2단계에서 Region Proposal을 통해 객체 탐지
- ⑩ 정확한 객체 탐지와 성능



3.3. YOLOv8

1-stage Detector인 YOLO의 최신 버전(2023.01)



3.4. 산출물

⑩ Python 학습용 dataset

⑩ Orange3 학습용 dataset

⑩ Python 모델 학습 및 실행 파일.ipynb

⑩ Orange3 모델 학습 및 실행 파일.ows

4. 철강산업 에너지사용 데이터를 이용한 에너지사용 예측 모델

4.1. 데이터 확인

⑩ 데이터 크기: 2.7MB

⑩ Data Shape: 35040 rows x 11 columns

순서	컬럼명	설명
1	data	매월 1일에 수집된 연속 시간 데이터
2	Usage_kWh	산업 에너지 소비량 연속 kWh
3	Lagging_Current_Reactive_Power	후행 전류 무효 전력 연속 kVarh
4	Leading_Current_Reacitve_Power	선행 전류 무효 전력 연속 kVarh
5	CO2	이산화탄소 ppm
6	Lagging_Current_Power_Factor	후행 전류 파워 요소
7	Leading_Current_Power_Factor	선행 전류 파워 요소
8	NSM	자정부터 초 단위 연속 S
9	Day_of_week	범주형 (일요일~토요일)
10	Week_Status	범주형 (주말, 평일)
11	Load_Type	범주형 (경부하, 중부하, 과부하)

4.2. 데이터 전처리 방법

4.2.1. OneHot Encoding

⑩ 범주형 데이터를 분석에 용이하게 변경하는 작업 중 하나

⑩ 컬럼을 범주별로 이진 형태 변환을 하기 때문에 정보 손실이 없다는 장점

⑩ 차원이 증가하여 데이터가 복잡해지는 단점

4.2.2. 상관관계 분석

예측하고자 하는 컬럼값인 에너지 사용량을 기준으로 상관관계가 가장 높은 컬럼 순서대

로 확인 후 상관관계가 낮은 컬럼은 모델에 학습시키지 않을 수 있음

4.3. 사용 모델

- Decision Tree Regressor
- Random Forest Regressor
- Extra Tree Regressor
- Grid Search: 모델은 아니고 상기 모델들의 하이퍼 파라미터 최적값을 찾는데 사용되는 하이퍼 파라미터 튜닝 기술

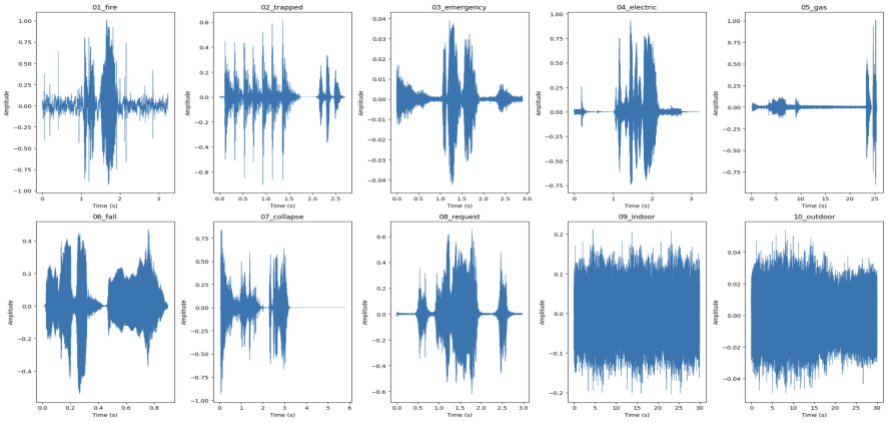
4.4. 산출물

- 정제된 학습용 데이터.csv
- Python 모델 학습 및 실행 파일.ipynb
- Orange3 모델 학습 및 실행 파일.ows

5. 철강산업 현장 음성 데이터를 이용한 위험상황 분류 모델

5.1. 데이터 확인

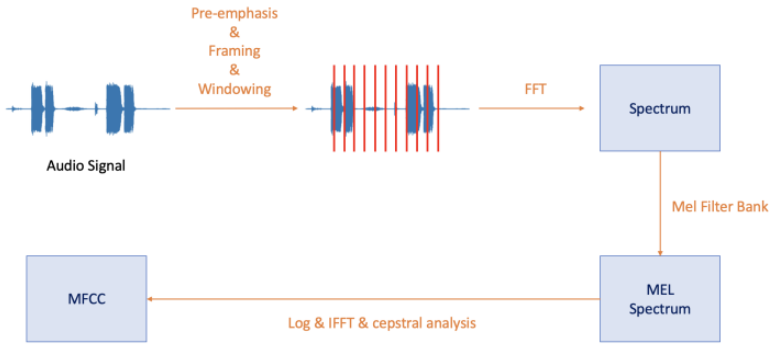
- 데이터 크기: 3.74GB
- Label: 화재, 갇힘, 응급의료, 전기, 가스, 낙상, 붕괴, 도움요청, 실내(정상), 실외(정상)
- Python 실습용 데이터: 음성데이터.wav 파일 2000개, 라벨.json 파일 2000개
- Orange3 실습용 데이터: 음성데이터 특징을 추출한 하나의 csv파일(2000 rows x 31 columns)



5.2. 데이터 전처리 방법

5.2.1. MFCC 변환

- MFCC는 Mel-Frequency Cepstral Coefficient의 약자로 오디오 데이터를 특징 벡터로 변환하는 알고리즘
- 오디오 데이터 중에서도 특히 사람의 음성 추출에 특화된 알고리즘



5.3. 사용 모델

- ⑩ Decision Tree Classification
- ⑩ Random Forest Classification
- ⑩ Extra Trees Classification
- ⑩ GradientBosst
- ⑩ CatBoost
- ⑩ AdaBoost

5.4. 산출물

- ⑩ Python 학습용 dataset
- ⑩ Orange3 학습용 dataset
- ⑩ Python 모델 학습 및 실행 파일.ipynb
- ⑩ Orange3 모델 학습 및 실행 파일.ows