<머신러닝을 활용한 프로젝트 요약>

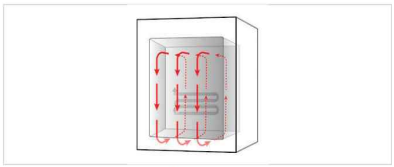
※ 프로젝트명: 공정 데이터를 활용한 설비 오류 발생 예측

1. 분석용 데이터 개요

1) 데이터 수집

- 열풍건조기 공정 데이터 (수집기간: 2022.09.06.~2022.10.27. 수집주기: 5sec)

[그림 1] 열풍건조



[그림 2] 열풍건조

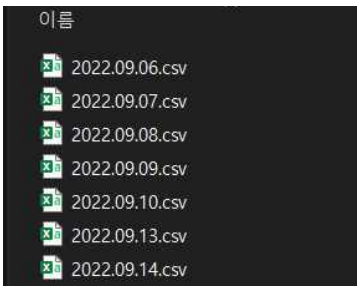
2) 데이터 출처

: Ai-hub(http://ai-hub.co.kr)에서 다운로드

3) 데이터 형태

- 공정 데이터 : 프로세스(1~43), 날짜, 시간, 공정 온도, 공정 전압
  (날짜별로 총 33개의 csv파일, 총 row: 51,084개)

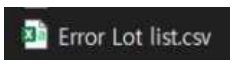
	A	B	C	D	E	F
1	Index	Process	Time	Temp	Current	Date
2	1	1	오후 4:24:03.0	75.13914228	1.61	2022-09-06
3	2	1	오후 4:24:08.0	76.66042142	1.53	2022-09-06
4	3	1	오후 4:24:13.0	77.17766014	1.701	2022-09-06
5	4	1	오후 4:24:18.0	76.58643441	1.736	2022-09-06
6	5	1	오후 4:24:23.0	77.87710396	1.748	2022-09-06
7	6	1	오후 4:24:28.0	76.01353467	1.734	2022-09-06
8	7	1	오후 4:24:33.0	71.00784573	1.763	2022-09-06



[공정 데이터의 csv 파일 내부(왼쪽)와 전체 리스트 중의 일부(오른쪽)]

- 에러 발생 리스트 : 날짜별 설비 에러 발생 프로세스(1~43) 및 에러 유형(1~11)
  (1개의 csv 파일)

	A	B	C	D	E	F	G	H	I	J	K	L
1	0	1	2	3	4	5	6	7	8	9	10	11
2	2022-09-06	32	33	20	21	22	31					
3	2022-09-07	32	33	34								
4	2022-09-08											
5	2022-09-09	15	16	17	21	22	23	29	30	31		
6	2022-09-10	32	28	29	30	31						
7	2022-09-13	27	28	29								
8	2022-09-14											
9	2022-09-15	40	41	39								
10	2022-09-16	2	35	3	34	36						
11	2022-09-17	12	13	14	16	17	18	28	29			



[에러 리스트의 csv 파일 내부(왼쪽)와 해당 파일(오른쪽)]

- 라벨링 작업: 데이터를 간단하게 바꾸기 위해 에러 유형에 관계없이 모든 에러 발생 시,

해당 프로세스는 1로 라벨링하고, 그 외에는 0으로 라벨링하여 수정하도록 함

	A	B	C	D	E	F	G
1	Index	Process	Time	Temp	Current	Date	Label
2	1	1	오후 4:24:(	75.13914	1.61	2022-09-06	0
3	2	1	오후 4:24:(	76.66042	1.53	2022-09-06	0
4	3	1	오후 4:24:'	77.17766	1.701	2022-09-06	0
5	4	1	오후 4:24:'	76.58643	1.736	2022-09-06	0
6	5	1	오후 4:24:2	77.8771	1.748	2022-09-06	0
7	6	1	오후 4:24:2	76.01353	1.734	2022-09-06	0
8	7	1	오후 4:24::	71.00785	1.763	2022-09-06	0
9	8	1	오후 4:24::	71.11938	1.721	2022-09-06	0

680	19	오후 5:27:	65.1194	1.596	2022-09-06	0
681	19	오후 5:27:	64.1218	1.448	2022-09-06	0
682	19	오후 5:27:	66.062	1.473	2022-09-06	0
683	19	오후 5:27:	65.0865	1.451	2022-09-06	0
684	19	오후 5:27:	67.2568	1.495	2022-09-06	0
685	20	오후 5:27:	123.927	0.92035	2022-09-06	1
686	20	오후 5:27:	122.738	0.91714	2022-09-06	1
687	20	오후 5:28:	121.825	1.19303	2022-09-06	1
688	20	오후 5:28:	117.936	1.10014	2022-09-06	1
689	20	오후 5:28:	125.562	1.11335	2022-09-06	1
690	20	오후 5:28:	125.037	1.18635	2022-09-06	1

[라벨링 작업을 완료한 공정 데이터]

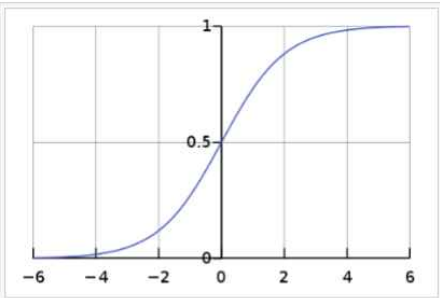
2. 분석 방법

: 학습에 필요한 Label이 0과 1 두 가지이므로, 일반적인 통계 기법인 로지스틱 회귀 분석 방법과 인공신경망을 활용한 딥러닝 분석(시계열 데이터이므로, LSTM을 활용)을 둘 다 수행하여 아래의 성능 지표들로 각 분석 방법들의 성능을 비교해 봄

1) 로지스틱 회귀 (Logistic Regression)[1)]

: 영국의 통계학자인 D. R. Cox가 1958년에 제안한 확률 모델로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 데 사용되는 통계 기법. 로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이지만, 로지스틱 회귀는 선형 회귀 분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 (classification) 기법으로도 볼 수 있음

흔히 로지스틱 회귀는 종속변수가 이항형 문제(즉, 유효한 범주의 개수가 두개인 경우)를 지칭할 때 사용되며, 의료, 통신, 데이터마이닝과 같은 다양한 분야에서 분류 및 예측을 위한 모델로서 폭넓게 사용되고 있음



표준 로지스틱 함수 $\sigma(t)$; 모든 t에 있어서 $\sigma(t) \in [0,1]$이다.

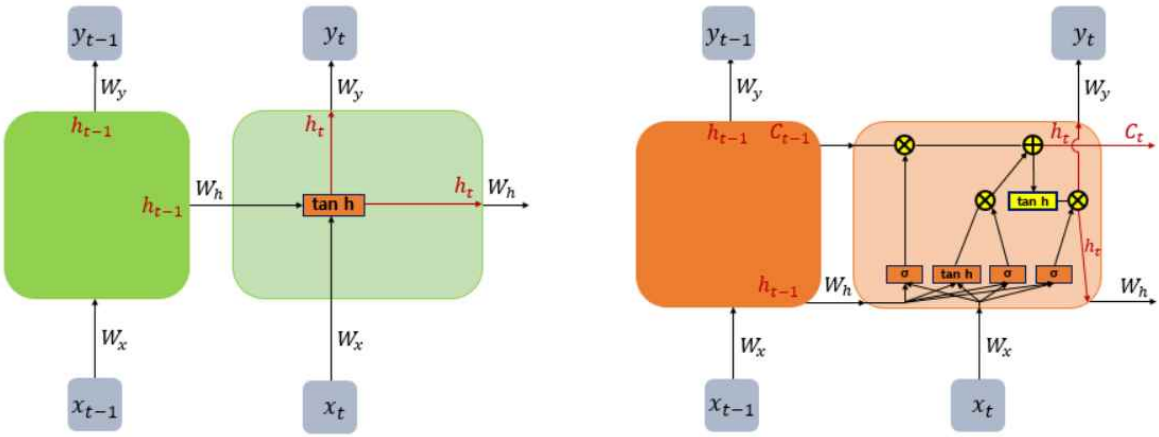
$$\text{logistic function} = \frac{e^{\beta \cdot X_i}}{1 + e^{\beta \cdot X_i}}$$

2) 딥러닝 - LSTM (Long Short Term Memory)[2)]

: 기존의 RNN은 출력 결과가 이전의 계산 결과에 의존하지만, 비교적 짧은 시퀀스(sequence)에

1) 출처: https://ko.wikipedia.org/wiki/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1_%ED%9A%8C%EA%B7%80
2) 출처: https://wikidocs.net/22888

대해서만 효과를 보이는 단점이 있으며, 시점(time step)이 길어질수록 앞의 정보가 뒤로 충분히 전달되지 못하는 현상이 발생함(장기 의존성 문제). RNN의 이러한 단점을 보완한 LSTM은 은닉층의 메모리 셀에 입력 게이트, 망각 게이트, 출력 게이트를 추가하여 불필요한 기억을 지우고, 기억해야할 것들을 정함으로써, 은닉 상태(hidden state)를 계산하는 식이 조금 더 복잡해졌으며 셀 상태(cell state)라는 값을 추가하여, 긴 시퀀스의 입력을 처리하는데 탁월한 성능을 보임. 주로 시계열 처리나 자연어 처리에 사용



[RNN 내부구조(왼쪽)와 LSTM의 내부구조(오른쪽)]

3. 성능 지표

: 위의 두 방법으로 학습 후, 각자 검증용 데이터로 예측한 결과에 대하여, TP(True Positive), FP(False Positive), FN(False Negative), TN(True Negative)을 오차 행렬로 나타내고, 정확도, 예측률, 재현율을 계산하여 비교함. 또한 ROC Curve, AUC, F1 Score를 추가로 산출하여 각 분석모델의 성능을 평가함

1) 오차 행렬 (Confusion Matrix)

: 오차 행렬은 학습된 분류 모델이 예측을 수행하면서 얼마나 헷갈리고(confused) 있는지를 보여주는 지표로, 이진 분류의 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표임

		예측 클래스 (Predicted Class)	
		Negative(0)	Positive(1)
실제 클래스 (Actual Class)	Negative(0)	TN (True Negative)	FP (False Positive)
	Positive(1)	FN (False Negative)	TP (True Positive)

① 정확도 (Accuracy)

: (TP+TN)/(TP+TN+FP+FN). 예측을 긍정으로 했든 부정으로 했든 실제로 참이었는지에 중점을 둔 지표로, 어떤 방식으로 예측을 하였든 실제로 그러한 예측이 참이었는지를 확인

② 정밀도 (Precision)

: TP/(TP+FP). 긍정적으로 예측하였을 때 그러한 예측이 실제로 참이었을 확률을 계산

③ 민감도(Sensitivity), 재현율(Recall)

: TP/(TP+FN). 실제로 참일 때 참으로 예측했을 확률

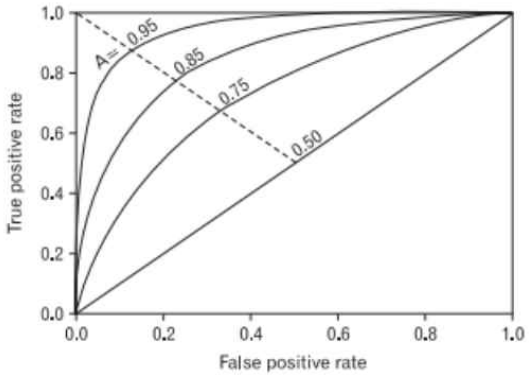
④ 특이도(Specificity)

: 특이도는 실제로 거짓일 때 거짓으로 예측했을 확률. 민감도와 대칭을 이루는 지표

2) ROC Curve 및 AUC, F1 Score

- ROC curve

: 민감도와 특이도의 공통적인 문제점인 한 쪽으로 찍어서 높은 점수를 맞을 수 있다는 문제점을 크게 보완한 지표로, ROC 커브의 x축은 FPR(False Positive Rate, 실제 거짓일 때 참이라 판단할 확률) y축은 민감도(Sensitivity, 실제 참일 때 참으로 예측할 확률)로 되어 있음. 곡선 위의 면적이 작을수록 민감도(TPR)는 1에 가까워지고, FPR은 0에 가까워지므로, 실제로 참일 때 참으로 판단하고, 실제로 거짓일 때 거짓이라 판단하는 예측력이 높아짐.



[ROC Curve와 AUC(A) 산출 값]

- F1 score

: F1 score = 2 * (정밀도 * 민감도) / (정밀도 + 민감도). 즉, 정밀도와 민감도를 조화평균으로 구한 것

(조화평균의 특징은 분모의 값이 일정할 때 분모의 두 값이 이질적일 수록 전체적으로 더 낮은 값을 갖게 된다는 것으로, 예를 들어, 정밀도와 민감도가 각각 0.9과 0.1을 갖는 경우와, 각각 0.5와 0.5를 갖는 경우, 분모의 값은 1로 동일하지만, 전자의 조화평균은 0.09, 후자의 조화평균은 0.25가 되어 후자의 값이 더 크게 됨)