붙임 4 **2023 경영경제대학 학술제 참가보고서**

2023 경영경제대학 학술제 참가보고서

제 목	화재 위험성 지표 개발 및 활용 (팀명 : 푸앙가디언즈)				
신청팀 인적사항	성명	학번	생년월일	연락처	E-mail
	이호용	20203374	19960531	010-5599-6677	bkupshoot@gmail.com
	곽수민	20180942	19991211	010-3181-3803	ytnals3803@naver.com
	권동구	20196130	20000318	010-6698-4492	donggu5654@naver.com
	박정현	20205602	20020121	010-6866-3569	petite0121@naver.com
	유태권	20196746	20000629	010-5388-6038	ytg000629@naver.com
선정주제	건물 안전성 데이터 마이닝 및 분석을 통한 위험성 지표 개발				
주제 선정 이유 (사회적 현상과 관련지어 서술)	우리나라의 지난 10년간(2012년~2021년) 연평균 화재 건수는 41,571건, 인명피해는 2,242명(사망 310명, 부상 1,933명), 재산 피해는 5,607억 원이다. 대한민국 공식 전자정부 누리집에 따르면 지난 10년간 화재에 의한 인명피해는 평균적으로 매년 2,300여 명이 발생하고 있다. 또한 재산피해액은 2012년 2,894억 원에서 2017년 5,069억, 2021년 1조 991억 원으로 해를 거듭할수록 큰 폭으로 증가하고 있음을 알 수 있다. 2021년도에는 **건축/구조물 화재의 발생 건수**가 23,997건으로 전체 화재 건수의 66.2%를 차지하고 있으며, 사망자 240명(87.0%), 부상자 1,589명(85.7%), 1,046,802백만 원(95.2%)의 재산 피해를 발생시킴으로써 유형별 화재 중 건축/구조물 화재가 **매년 가장 큰 인명, 재산피해 규모를 차지**하고 있다. 따라서 건물의 **화재 위험도를 예측 및 활용**할 수 있다면, 화재로 인한 물적/인적 피해를 줄이고 예방하는 데 큰 도움이 될 것이다. 이를 위해 '**건물 안전성 데이터 마이닝 및 분석을 통해 건물의 위험성 지표 및 응용 가능한 프레임워크를 개발**'했다. 이러한 지표를 건물 소유자, 시설 관리자, 정부 기관 등이 이용한다면 더욱더 안전한 사회를 만드는 데 기여할 수 있을 것으로 기대된다.				
주제 분석 및 연구	화재에 의한 피해는 크게 재산 피해와 인명피해로 구분된다. 화재 발생 시 발생하는 재산피해액(재산 피해)과 인명피해 발생여부(인명피해)를 예측하여 **화재 위험성 지표**(이하 **화재피해등급**)를 개발했다. 기존의 국내 연구를 살펴보면, 화재와 관련된 데이터 기반의 분석은 시도-시군구별로 화재위험지수를 산정해 오고 있다. 하지만 이러한 분석 결과는 지역 단위로 제공되기 때문에 실제 정책이나 화재안전점검 우선순위 선정 등 건물 단위의 예방 활동에 활용하기에는 어려움이 클 것으로 예상된다. 국외 사례를 살펴보면, 미국 아틀란타 소방청 (Atlanta Fire Rescue Department, AFRD)에서 구축한 Firebird가 대표적이다. Firebird는 건물 단위로 화재위험도를 예측하여 화재 발생 위험도가 높은 건물의 점검 우선순위를 결정하는 것이 주목적이다. 해당 모델은 랜덤포레스트(Random Forest) 기반으로 개발되어 약 71%의 정확도를 보였으나, 건물에 관한 속성도 면적, 층수, 필지 등 한정적인 변수만을 활용했다는 한계점이 존재한다. 따라서 본 연구에서는 소방재난본부, 국가공간정보포탈 등에서 관리되고 있는 **건물 데이터를 중심**으로 다양한 변수들을 고려한 화재 관련 융합 데이터 셋을 생성하고 이를 통해 **화재피해등급**을 산정하는 연구를 진행했다. 또한 실시간으로 동적인 데이터를 수집하고 처리하기 위한 소켓 프로그래밍				

<table>
<tr><td colspan="2"></td><td>및 분산 처리 시스템을 통하여 선제적인 대응이 가능한 시스템을 구축했다.</td></tr>
<tr><td rowspan="2">현실성
및
지속가능성</td><td>주제의
현실
적용 가능성</td><td>화재 발생 시 재산 피해액 및 인명피해 예측을 위해 사용한 변수들로 현장안전센터거리, 기온, 풍속, 가시거리, 실효습도, 전기사용량, 가스사용량, 화재발생시각 등이 고려되었다.

위에 나타난 대부분 변수는 화재가 발생하기 이전에 실시간으로 수집하고 업데이트할 수 있다. 즉, 화재피해등급 변화의 실시간 모니터링이 가능하며 대용량 데이터 처리에 견딜 수 있는 시스템을 구축했다.
이를 통해 건물 화재 발생에 대한 사전 대비가 가능하며 화재피해등급에 따라 화재의 피해 규모를 사전에 예측하고 인지할 수 있다.

화재 예방과 더불어 화재 발생 시 건물 소유자, 시설 관리자, 정부 기관 등이 신속하고 선제적인 대처를 하는 데 도움을 줄 수 있다.</td></tr>
<tr><td>주제의
발전 가능성
및 타당성</td><td>추후 건물 이용 연령층, 건물 유동 인구 등 더 많은 데이터를 확보하여 다양한 변수를 고려한다면 모델의 정확도를 향상할 수 있다. 이를 통해 화재피해등급을 좀 더 세분화하는 것이 가능하며 중요한 의사결정 및 정책 등 활용 가능 범위를 넓힐 수 있을 것이다.

또한 데이터가 동적으로 수집되고 더 나은 지표를 개발함에 따라 모델 확장 및 개선이 용이하다. 이를 통해 계속 변화하는 상황 속에서 최적화된 대처 전략 모델을 바탕으로 화재피해등급 업데이트를 한다면 효과적인 대응 전략을 수립할 수 있을 것이다.</td></tr>
<tr><td rowspan="2">창의성
및 전문성</td><td>주제에 대한
분석의
차별성 및
독창성</td><td>화재 위험도 지표는 현재 '화재 위험 지수'라는 명칭으로 사용 중에 있다. 하지만 이러한 위험도 지표는 특정 지역에 한정하여 일정 기간의 상대적인 화재 위험 수치를 나타냈을 뿐 건물 단위로 실시간으로 모니터링 할 수 있는 지표는 아직 존재하지 않는다.

또한 국외 사례까지 확장하여 살펴보았을 때도 화재가 발생할 확률을 예측하는 모델은 있지만 화재 발생으로 인한 피해액 혹은 사상자 여부를 예측한 모델은 없다.

이러한 상황 속에서 '화재피해등급'은 실시간으로 건물 단위 모니터링을 할 수 있는 지표로서 새로이 활용될 수 있을 것이다.</td></tr>
<tr><td>의견 도출의
논리 및
과정</td><td>건물 단위로 데이터를 융합하여 화재피해등급을 예측하는 것이 목적이기 때문에 전기/가스 사용량, 날씨 데이터, 화재 출동 현황 등 다양한 데이터를 건물별로 수집했다.
<table>
<tr><th>데이터명</th><th>설명</th><th>출처</th></tr>
<tr><td>전기에너지</td><td>2020년 건물별
전기에너지 사용량</td><td rowspan="2">건축데이터 개방</td></tr>
<tr><td>가스에너지</td><td>2020년 건물별
가스에너지 사용량</td></tr>
<tr><td>화재출동현황</td><td>사망인명피해수,
화재발생 시각,
현장안전센터거리,
시간단위날씨
(기온, 풍속),
가시거리,
재산피해액</td><td>서울소방재난본부</td></tr>
<tr><td>상대습도</td><td>2017~2020년 시간별
상대 습도 데이터</td><td>기상청</td></tr>
<tr><td>서울시 건축물대장 법정동 코드정보</td><td>법정동 코드, 시군구코드, 법정동명, 시군구명 등</td><td>서울 열린데이터 광장</td></tr>
</table></td></tr>
</table>

데이터 전처리

- **변수 선택** : 결측치가 1% 이상인 칼럼을 제거하고 분석 목적에 맞지 않는 데이터를 분류했다.
- **실효습도 변수 생성** : 실효습도란 화재 예방을 목적으로 사용하기 위해 고안된 것으로 건조도를 나타낸다. 평균습도에 지나온 시간에 따른 가중치를 두어 산출하게 되고 건조한 날이 연속되는 경우 실효습도는 낮아지고 불이 날 가능성은 커지게 된다. 기상청에서 얻은 상대습도 데이터를 기준으로 계산하여 2017년~2020년의 실효습도 변수를 생성했다.

결측치

- 4개의 변수(**현장소방지역대거리**, **기온**, **풍속**, **풍향**)에서 각각 4, 3, 34, 34개의 결측치를 발견했다.
- **현장소방지역대거리** : 인접 주소의 거리로 대체했다.
- **기온** : 일별 평균 기온으로 대체했다.
- **풍속, 풍향** : 풍속과 풍향은 시간 별로 변화하는 폭이 크기 때문에 일별 평균 풍속과 풍향으로 대체하지 않고 삭제했다.

이상치

- 변수별 기술 통계량을 확인해 보고 이상치로 의심되는 값들이 존재하는지 확인했다.
- 총 6개의 변수(**현장소방서거리**, **현장안전센터거리**, **시간단위가시거리**, **재산피해액**, **전기사용량**, **가스사용량**)에서 이상치로 의심되는 값들을 확인했다.
- 분포 형태에 따라 이상치 확인에 유리한 방법들이 다르다. 변수들의 분포가 다양했기에 여러 가지 방법을 활용하여 이상치를 탐지했다.
  - **Graphical method** : Box-plot, Scatter plot을 통해 시각적으로 이상치를 확인했다.
  - **z-score** : 각 변수를 표준화하여 z-score 값을 통해 3표준편차 이상의 값들을 확인했다.
  - **DBSCAN**(Density-based spatial clustering of applications with noise) 알고리즘 : 클러스터의 수를 지정하기 어려운 상황이었기 때문에 K-Means의 대안으로 이상치 탐지에 적합한 밀도 기반의 DBSCAN 알고리즘을 사용했다.
  - 추가적으로 변수별 특징을 고려하여 결측치 대체 및 제거했다.
- **현장소방서거리, 현장안전센터거리, 시간단위가시거리**
  - Box-plot의 1.5*IQR 범위를 벗어나고 z-score의 값이 3표준편차 이상인 값을 이상치로 판별하고 동별 평균 거리로 대체했다.
- **재산피해액**
  - 변수 특성상 제거하지 않는 것이 맞지만 분석의 정확도를 위해 극단값에 있는 한 개의 이상치를 제거했다.
- **전기사용량(KWh), 가스사용량(KWh)**
  - Box-plot을 확인한 결과 max(Q3 + (1.5 * IQR)) 이상의 값들이 각각 246개, 270개가 확인됐다.
  - z-score 값을 확인한 결과 각각 3표준편차 이상의 216개, 261개의 값들을 확인했다.
  - 이상치로 의심되는 값들 중 분석의 정확도를 위해 총 59개의 값들을 제거했다.

데이터 병합

- 화재출동현황 데이터셋을 기준으로 전처리한 데이터들을 병합했다.
- 전기에너지와 가스에너지는 주소를 기준으로 병합했고 실효습도 변수는 화재발생일자를 기준으로 병합했다.

	사망인명피해수	화재발생시	현장안전센터거리	시간단위기온	시간단위풍속	시간단위가시거리	재산피해액	실효습도	전기사용량(KWh)	가스사용량(KWh)
0	0	0	3	0.2	2.2	491	13391	54.273750	18096.210463	11774.10017
1	0	15	2	9.3	1.9	703	35	49.554054	18096.210463	11774.10017
2	0	20	1	-4.2	1.8	2000	308	31.607682	18096.210463	11774.10017
3	0	10	1	-4.4	4.5	558	184	50.355768	18096.210463	11774.10017
4	0	3	3	1.5	5.9	634	877	43.742292	18096.210463	11774.10017

모델링 기법

- 인명피해발생여부 - 로지스틱 회귀분석
  - 총 8개의 변수(현장안전센터거리, 시간단위 (기온, 풍속, 가시거리), 실효습도, 전기사용량(KWh), 가스사용량(KWh), 화재발생시각)를 바탕으로 인명피해발생여부에 대해 모델링을 진행했다.
  - 각 변수별 결과는 다음과 같았으며, 예측 정확도는 92.7%를 기록했다.

	coef	std err	z	P>|z|	[0.025	0.975]
현장안전센터거리	0.0038	0.031	0.121	0.903	-0.058	0.066
시간단위기온	0.0801	0.037	2.177	0.029	0.008	0.152
시간단위풍속	0.1648	0.032	5.176	0.000	0.102	0.227
시간단위가시거리	-0.0140	0.032	-0.434	0.664	-0.077	0.049
실효습도	-0.1159	0.037	-3.129	0.002	-0.188	-0.043
전기사용량(KWh)	-0.0240	0.041	-0.587	0.557	-0.104	0.056
가스사용량(KWh)	0.0685	0.040	1.733	0.083	-0.009	0.146
화재발생시각_새벽	-2.7225	0.080	-34.033	0.000	-2.879	-2.566
화재발생시각_저녁	-3.4376	0.096	-35.714	0.000	-3.626	-3.249
화재발생시각_주간	-3.0664	0.051	-59.747	0.000	-3.167	-2.966

  - 추정된 로지스틱 회귀식을 바탕으로 인명피해발생 확률을 계산하여 지표 산출에 활용했다.

- 재산 피해액 – 다중선형 회귀분석
  - 인명피해발생여부 예측을 위해 사용한 독립변수 그대로 사용하여 재산피해액에 대해 모델링을 진행했다.
  - 종속변수로 사용된 재산피해액의 경우 편차가 심해 분산 안정화 변환 방법 중 하나인 로그 변환을 진행한 뒤 사용했다.

	coef	std err	t	P>|t|	[0.025	0.975]
현장안전센터거리	0.1964	0.020	9.969	0.000	0.158	0.235
시간단위기온	-0.0288	0.002	-16.452	0.000	-0.032	-0.025
시간단위풍속	0.1095	0.016	6.979	0.000	0.079	0.140
시간단위가시거리	0.0005	2.85e-05	18.965	0.000	0.000	0.001
실효습도	0.0536	0.001	39.122	0.000	0.051	0.056
전기사용량(KWh)	1.615e-05	2.03e-06	7.950	0.000	1.22e-05	2.01e-05
가스사용량(KWh)	-1.84e-06	1.3e-06	-1.411	0.158	-4.4e-06	7.16e-07
화재발생시각_새벽	1.5449	0.069	22.545	0.000	1.411	1.679
화재발생시각_저녁	1.0991	0.068	16.068	0.000	0.965	1.233
화재발생시각_주간	1.4202	0.060	23.504	0.000	1.302	1.539

Dep. Variable:	재산피해액	**R-squared (uncentered):**	0.831
Model:	OLS	**Adj. R-squared (uncentered):**	0.831
Method:	Least Squares	**F-statistic:**	8342.
Date:	Fri, 05 May 2023	**Prob (F-statistic):**	0.00
Time:	19:48:36	**Log-Likelihood:**	-38466.
No. Observations:	17001	**AIC:**	7.695e+04
Df Residuals:	16991	**BIC:**	7.703e+04
Df Model:	10		
Covariance Type:	nonrobust		

Omnibus:	431.122	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	464.087
Skew:	0.403	**Prob(JB):**	1.68e-101
Kurtosis:	3.063	**Cond. No.**	1.84e+05

  - 각 변수별 결과는 다음과 같으며, MAE(Mean Absolute Error)는 1.8287을 기록했다.
  - 모델의 설명력을 나타내는 지표인 R-Squared의 값이 83.1%로 높게 나와 설명력이 높은 모델임을 확인할 수 있었다.
  - 추정된 선형회귀식을 바탕으로 예상 ln(재산피해액)을 계산하여 지표 산출에 활용했다.

결과

- 모델링 결과 인명피해 발생여부는 정확도 92.7%, 재산 피해액은 R-squared 83.1%, MAE 1.8287을 기록하며 준수한 성능을 보여주었다. 이러한 모델을 바탕으로 인명피해 발생 확률과 예상피해액 값을 추출했고, 화재피해등급을 도출하는 데에 활용했다.

<화재피해등급 도출>

- 재산 피해액 모델이 예측한 값에 MinMax Scaler를 적용하여 0~1 사이의 값(이하 재산피해지수)으로 표현했다.
- 사상자 발생 여부 모델은 0~1 사이의 확률값을 출력할 수 있다. 이 예측 확률에 MinMax Scaler을 적용하여 마찬가지로 0~1 사이의 값(이하 인명피해지수)으로 표현했다.
- 재산 피해액 지수와 인명피해지수를 곱한 값에 제곱근을 취하고 100을 곱한 값으로 얻은 화재피해지수를 통해 최종적으로 화재 피해 등급을 도출했다.

$$X_{재산피해지수} = MinMax(\hat{y}_{예상피해액})$$

$$X_{인명피해지수} = MinMax(\hat{y}_{인명피해확률})$$

$$Y_{화재피해지수} = \sqrt{X_{재산피해지수} * X_{인명피해지수}} * 100$$

$$(0 \leq Y_{화재피해지수} \leq 100)$$

화재피해등급	화재피해지수	건물 수
1	66 이상	26개
2	33 이상 66 미만	1,744개
3	33 미만	19,482개
		총 21,252개

전공융합 적절성

화재 출동 데이터, 기후 데이터, 건물 정보 데이터 등 화재와 관련된 통계 데이터를 수집했고, **기술 통계량**을 통해 이상치가 존재할 확률이 높은 변수를 뽑아 그래프 적 방법을 통해 이상치를 탐지했다.

이상치와 결측치에 적절히 대처한 최종 데이터를 통계적 기법인 **로지스틱 회귀분석**을 통해 인명피해 확률을, **다중 선형 회귀분석**을 통해 재산 피해액을 예측했다.

위 모델을 통해 화재위험지수를 계산했으며 **실시간으로 데이터 스트림 및 처리**가 가능한 **프레임워크**를 구성했다.

의의 및 기대효과

건물 화재에 대한 화재 피해(인명피해, 재산 피해)를 예측하는 모델을 활용하여 '**화재피해등급**'을 산출할 수 있다.

화재피해등급의 변수 특성상 실시간 모니터링이 가능하며 건물 단위로 모델링을 진행하였기에 **실시간 건물 단위 모니터링**이 가능하다.

이를 활용한다면 건물 화재 발생에 대한 **피해의 크기와 범위를 미리 파악**할 수 있으며 화재 발생 시 건물 소유자, 시설 관리자, 정부 기관 등이 신속하고 선제적인 대처를 할 수 있기에 상황에 맞는 **최적화된 대응 전략을 수립**하여 화재 발생에 의한 **피해를 최소화**할 수 있을 것이다.