#### 2024학년도 2학기

## 문제해결프로그래밍 강의 1주차



2024.09.05



### 강의 목표

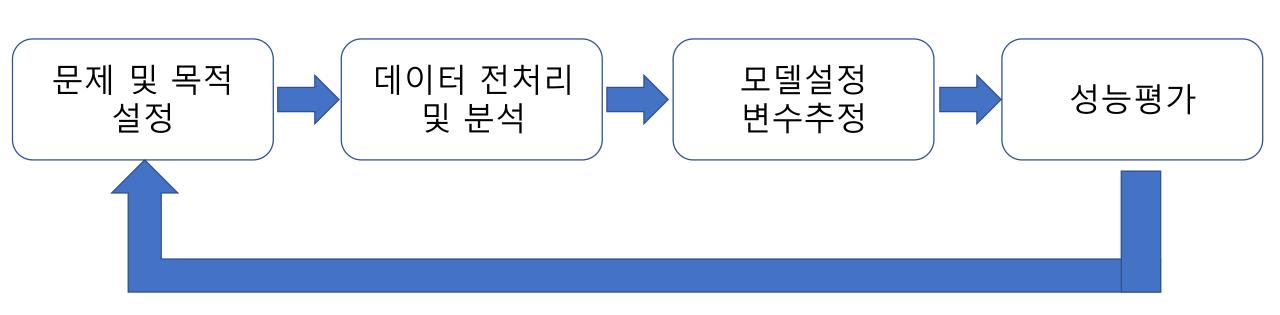
#### 강의 목표

다양한 분야에서 발생하는 문제를 해결하는 컴퓨팅 실습을 통한 프로그래밍 응용능력의 향상

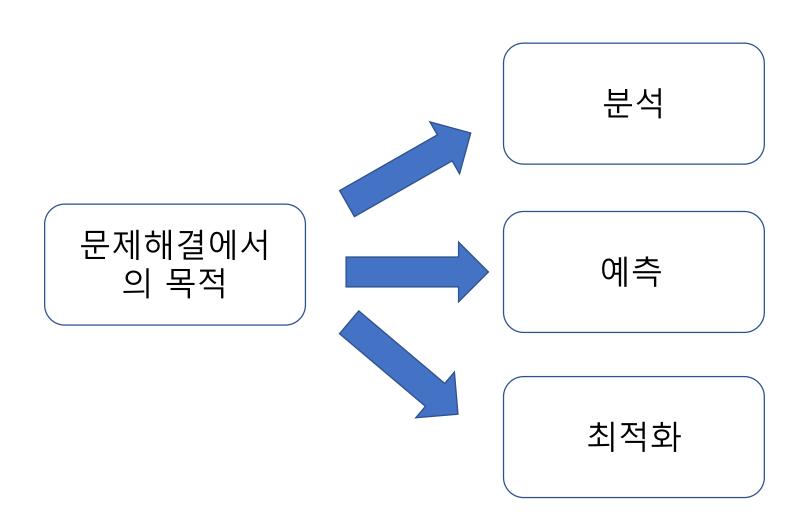
문제를 해결하기 위한 단계적 활동

- 1. 문제의 특성 이해
- 2. 문제해결 알고리즘의 이론적 탐색
- 3. 프로그래밍 실습
- 4. 알고리즘의 효율성 평가

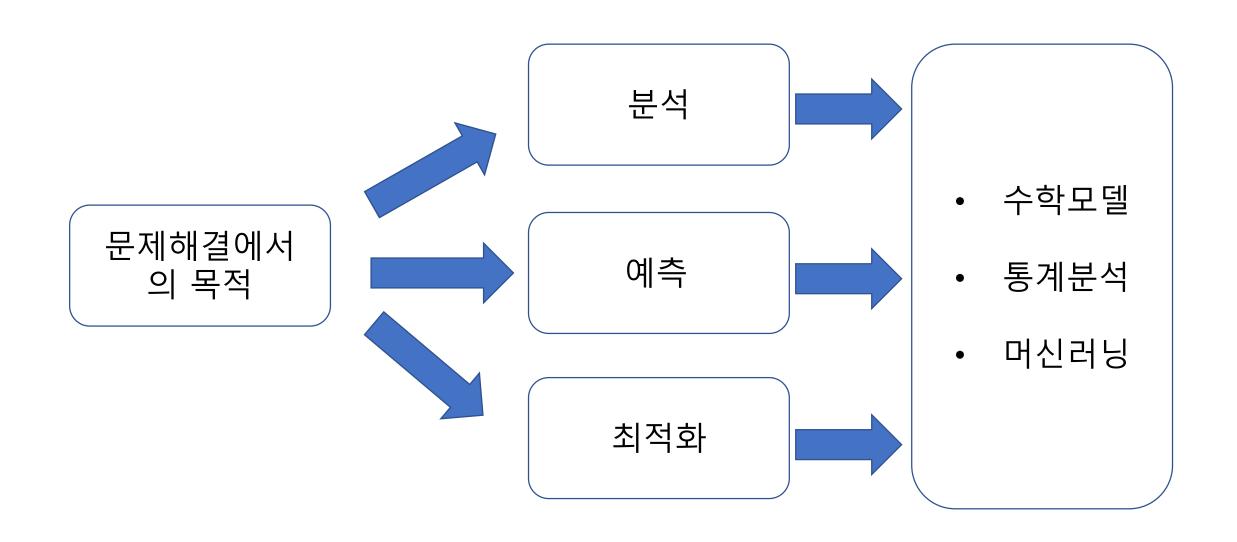
## 문제 해결 과정



## 문제 해결 종류 및 방법



## 문제 해결 종류 및 방법



### 문제해결 프로그래밍 소개

모델 특징 및 선택

원리 이해 (수리 모델링) 통계적 추론 (통계 모델)

응용지향형 모델 (머신러닝)

해석성이 좋다

적합성 or 정확도가 낮다

과소적합되기 쉽다

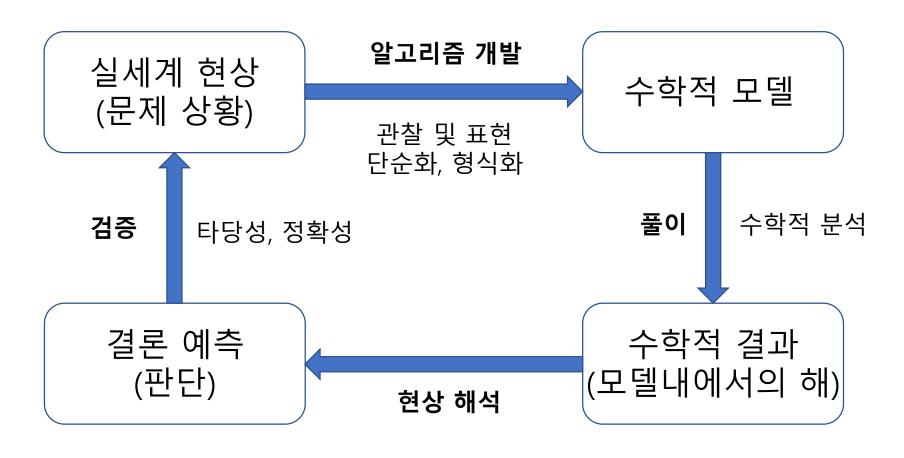
해석성이 나쁘다

적합성 or 정확도가 높다

과대적합되기 쉽다

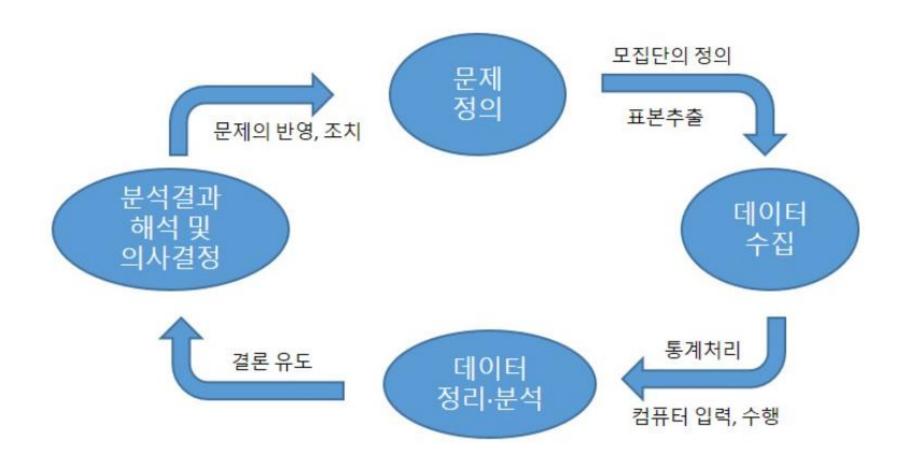
### 문제해결 프로그래밍 소개

수학적 해결 방법 : 수리 모델링



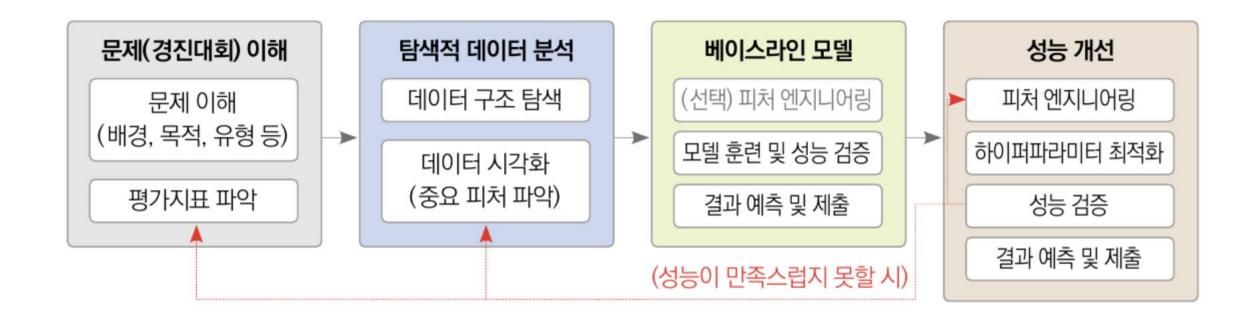
### 문제해결 프로그래밍 소개

통계적 해결 방법 : 가설 검정, 통계분석



### 문제해결 과정

#### 머신러닝을 이용한 문제해결 과정



### 머신러닝을 이용한 문제해결 과정에서 다룰 내용



# 강의 계획

주차	내용	
1주차	문제해결프로그래밍 강의 소개	
2주차	캐글을 통한 문제해결 (자전거 대요 수요 예측 : 문제의 이해 및 파이썬 기초)	
3주차	캐글을 통한 문제해결 (자전거 대요 수요 예측 : 데이터 분석 및 시각화)	
4주차	캐글을 통한 문제해결 (자전거 대요 수요 예측 : 다양한 예측 모델)	
5주차	캐글을 통한 문제해결 (자전거 대요 수요 예측 : 성능 개선 방법)	
6주차	캐글을 통한 문제해결 (안전 운전자 예측 : 데이터 분석 및 시각화)	
7주차	캐글을 통한 문제해결 (안전 운전자 예측 : 예측 모델 및 성능 개선)	
8주차	중간 고사 (객관식 + 코딩)	

# 강의 계획

주차	내용	
9주차	다양한 사회적 문제와 수리 모델링 (팀 구성 및 주제 선정)	
10주차	코로나-19 감염자 예측 문제와 거리두기 전략 (감염병 수리 모델링 및 수치해석 기법)	
11주차	코로나-19 감염자 예측 문제와 거리두기 전략 (데이터를 통한 파라메터 추 정)	
12주차	코로나-19 감염자 예측 문제와 거리두기 전략 (최적화 전략 이란?)	
13주차	코로나-19 감염자 예측 문제와 거리두기 전략 (자신만의 거리두기 전략 제시)	
14주차	팀별 과제 발표	
15주차	기말 고사 (객관식 + 코딩)	

### 평가 방법

• 출석 (10%)

• 과제 및 발표 (15%) – 기말고사 1주전 1회

• 중간고사 (35%)

• 기말고사 (40%)

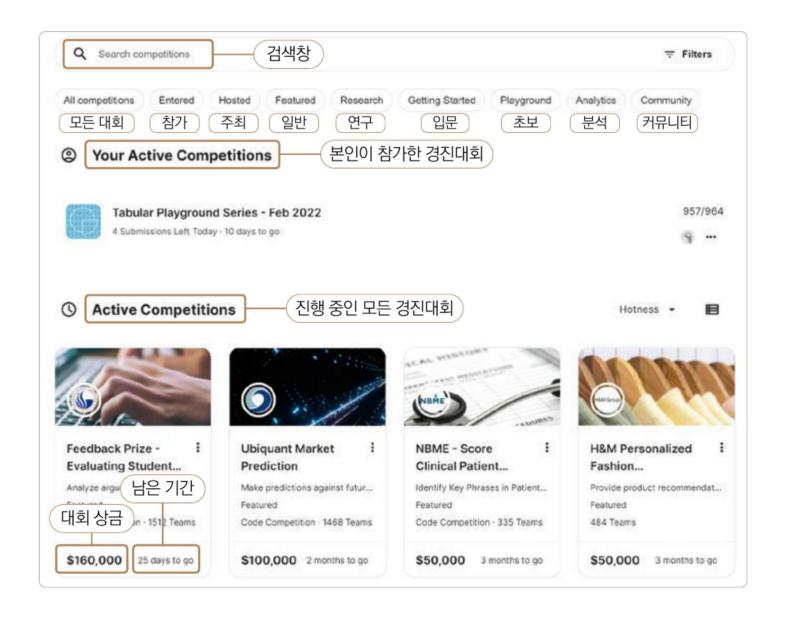
캐글 (Kaggle)

데이터 과학 및 머신러닝 경진대회를 주최하는 온라인 커뮤니티

일부 회사에서 취업시 캐글 경진대회 (competitions)의 실적 우대

홈페이지 https://www.kaggle.com/

캐글 경진대회 창

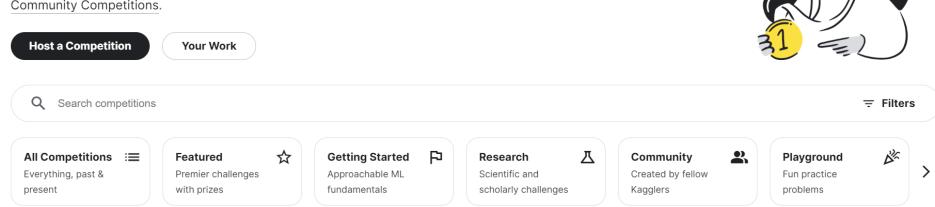


#### 캐글에서 Competitions 로 들어가보자.

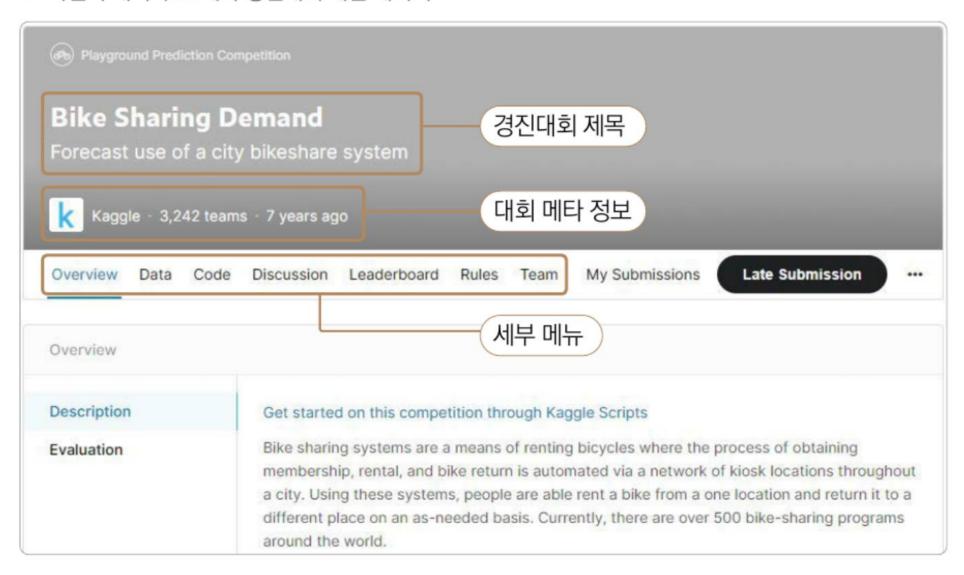
https://www.kaggle.com/competitions

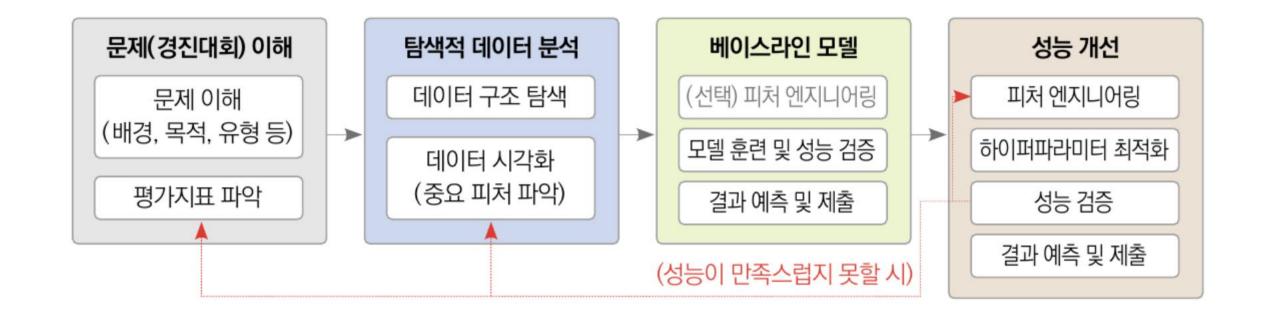
#### **Competitions**

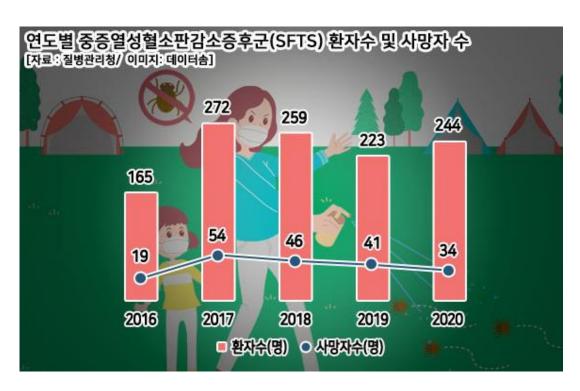
Grow your data science skills by competing in our exciting competitions. Find help in the  $\underline{\text{documentation}}$  or learn about Community Competitions.

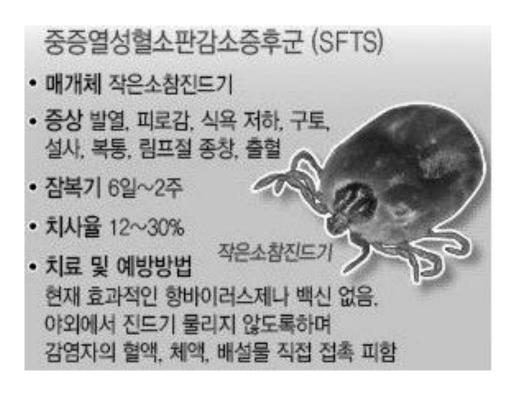


▼ 자전거 대여 수요 예측 경진대회 메인 페이지









문제 설정

위험한 SFTS, 언제 어디서 얼마나 감염자가 발생할지 예측할 수 있을까??

감염자를 예측할 수 있는 요인이 무엇이 있을까? 또 얼마나 영향을 미칠까?

문제 설정

시도 단위의 월별 SFTS 환자수를 예측해보자.

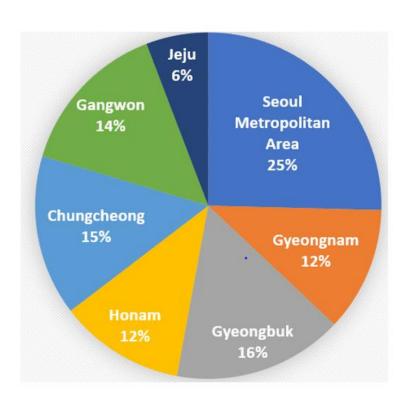
데이터 수집

Data: SFTS cases from 2016 to 2019 in Korea.

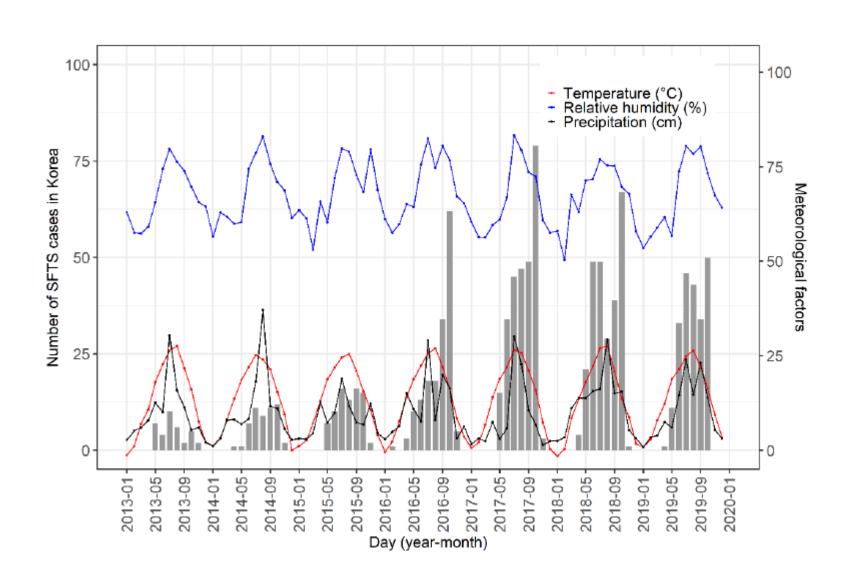
#### Cumulative number of cases from 2016 to 2019

Total: 913



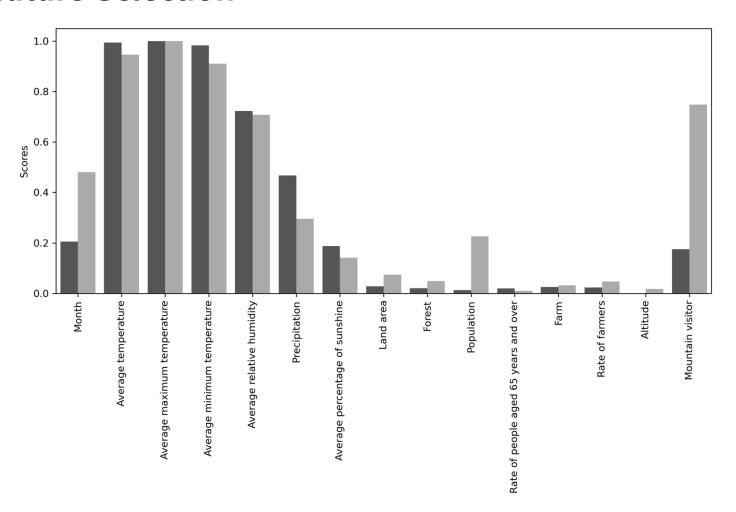


데이터 수집

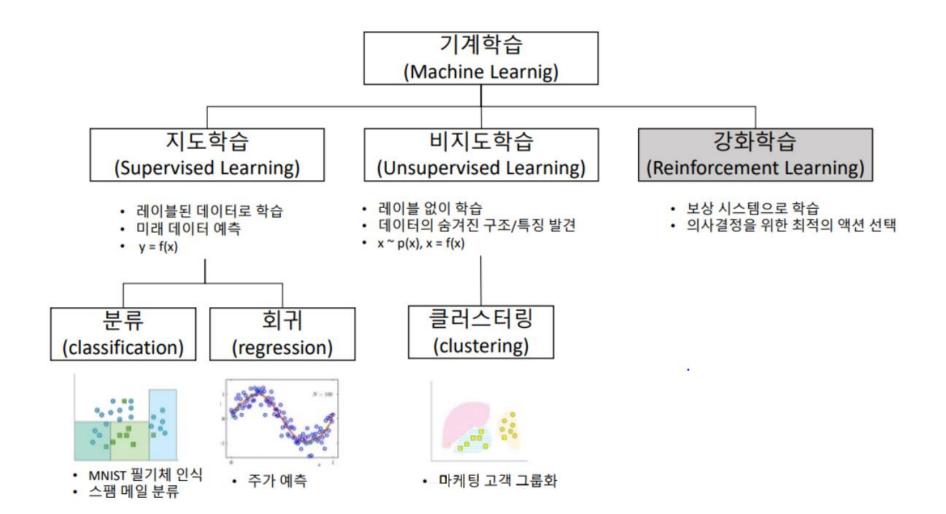


데이터 분석

#### **Univariate Feature Selection**



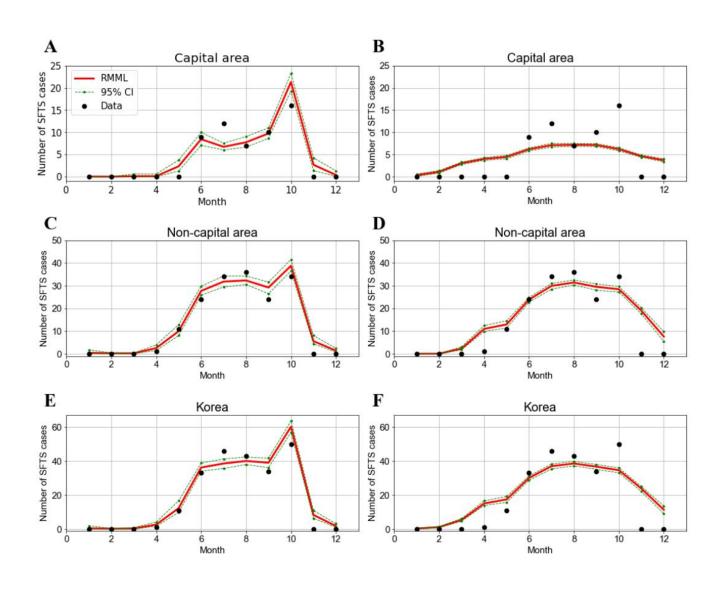
#### 모델 선정



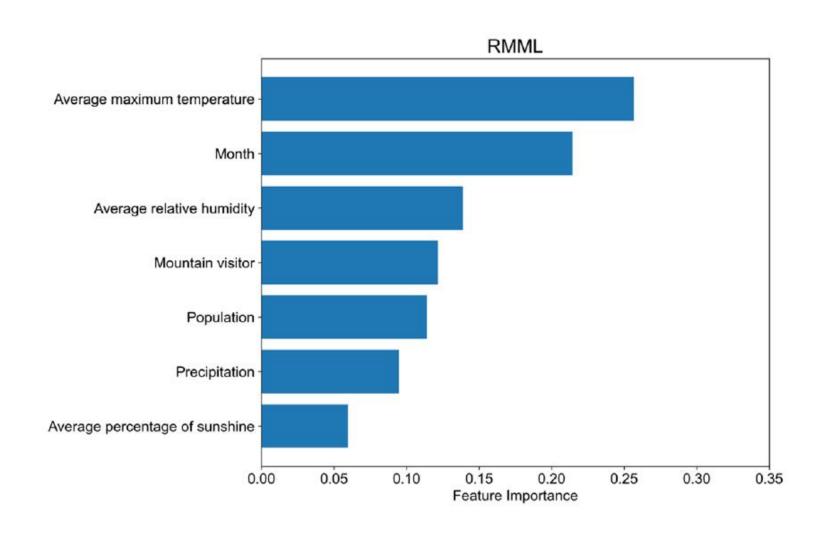
#### 모델 평가 및 검증

	Modified-RMML						
	Train		Test	Test			
Methods	MSE	R <sup>2</sup>	MSE	R <sup>2</sup>			
LR	6.558 (-40.6)	0.636 (64.5)	4.175 (-35.3)	0.687 (33.0)			
Ridge	6.562 (-40.6)	0.636 (64.6)	4.162 (-35.5)	0.689 (33.2)			
GBR	0.187 (-54.8)	0.990 (1.29)	3.127 (-14.2)	0.766 (5.31)			
BTR	1.373 (-46.6)	0.924 (7.75)	2.937 (-12.6)	0.780 (4.24)			
MLPR	2.329 (-79.2)	0.871 (130)	3.190 (-52.2)	0.761 (52.0)			

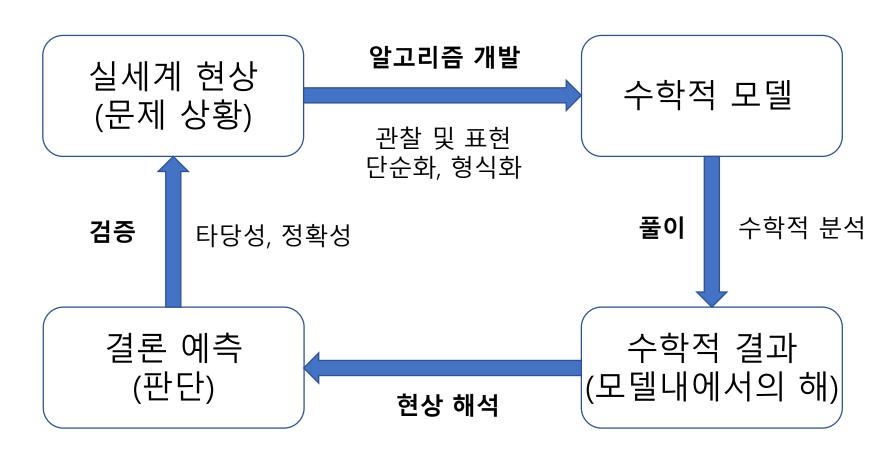
#### 모델 평가 및 검증



#### 모델 평가 및 검증



수학적 해결 방법 : 수리 모델링



#### 문제상황

문제 8%의 소금물 150g이 있다. 여기에 물을 더 넣어 6%의 소금물을 만들고 싶다. 몇 g의 물을 넣으면 될지 구하여라.

계산	풀이 과정
물의 양을 $x$ 라 두자. 물에 들어 있는 소금의 양에 대한 관계식을 세우자. $\frac{8}{100} = (150 + x) \times \frac{6}{100}$	수학적 모델
$150 \times 8 = 6(150 + x), \qquad -6x = -300 \qquad \therefore x = 50$	수학적 결과
따라서 $50g$ 물을 넣으면 $6\%$ 의 소금물이 된다.	결론 예측
처음 소금물에 들어 있는 소금의 양은 $150 \times \frac{8}{100} = 12(g)$ 이고, 의 소금물 에 들어 있는 소금의 양은 $200 \times \frac{6}{100} = 12(g)$ 이므로 문제의 뜻에 맞다.	검증

현실에서는 시간의 따른 변화를 나타내는 수리 모델링이 가장 많이 사용 된다.

- 차분 방정식 (Difference Equation)
- 미분 방정식 (Differential Equation)

일별 코로나-19 감염자 양상을 예측하고 싶다. 어떻게 해야 할까?

COVID-19 감염자 데이터 (2020년 8월)

날짜	1일	2일	3일	4일	5일	6일	7일	8일	9일	10일	11일
감염된											
사람 수	6	6	15	14	22	19	34	17	17	29	45
(명)											



#### 코로나-19 감염자 수에 대한 데이터

• 전세계 데이터

구글: 코로나 현황판

https://coronaboard.kr/

• 한국 데이터

구글: 질병관리청 코로나-19

https://ncov.kdca.go.kr/

#### 폐 그물, 폐 로프에 의한 바다자원 문제



플라스틱 그물 대신에 시간이 지나면 바다에 녹은 재질의 그물 Or 친환경 재질 그물로 대체할 수 있을까?

어민들에게 어떻게 그물의 재질을 바꾸도록 할까?

문제 설정

플라스틱 그물보다 친환경 재질의 그물이 어획이 더 잘되거나 못하지 않다는 것을 보이자.

문제 설정

플라스틱 그물보다 친환경 재질의 그물이 어획이 더 잘되거나 못하지 않다는 것을 보이자.

데이터 수집

1번그물	2번그물	3번그물	4번그물	5번그물
0	0	0	585	0
0	291	0	0	442
178	410	223	301	258
595	440	376	0	928
701	0	420	970	0
630	0	255	472	1244
655	494	457	300	217
472	0	380	0	144
340	240	0	1404	786
234	201	326	1878	2626

#### 데이터 요약

	1번 그물	2번 그물	3번 그물	4번 그물	5번 그물
총시행횟수	134	134	134	134	134
총어획중량	33709	35976	42614	54612	57943
평균중량	251.6	268.5	318	407.6	432.4

# 문제해결 프로그래밍 소개 (통계 모델)

#### 데이터 요약

	1번 그물	2번 그물	3번 그물	4번 그물	5번 그물
총시행횟수	134	134	134	134	134
총어획중량	33709	35976	42614	54612	57943
평균중량	251.6	268.5	318	407.6	432.4

#### 모델 설정

집단간 비교가 가장 쉬운 방법은? --→ 평균



집단간 평균이 차이가 있는지 없는지 어떻게 알 수 있지?



통계의 가설 검정 방법

# 문제해결 프로그래밍 소개 (통계 모델)

#### 모델 설정

집단간의 평균이 같은지를 분석하는 통계 기법

	모수 검정	비모수 검정
2개 집단	T-test Paired T-test	Mann-Whitney U 검정
3개 이상의 집단	ANOVA	Kruskal-Wallis H 검정

# 문제해결 프로그래밍 소개 (통계 모델)

#### 모델 평가 및 검증

집단1	집단2	그룹간평균차이	통계량	p-value	
1	2	-18.82835821	42.05938	0.916994	
1	3	-35.82462687	25.06312	0.494135	
1	4	-43.05970149	17.82804	0.301762	
1	5	-65.38432836	-4.49659	0.02807	
2	3	-16.99626866	43.89147	0.94156	
2	4	-24.23134328	36.6564	0.814097	
2	5	-46.55597015	14.33177	0.226174	
3	4	-7.235074627	53.65267	0.997618	
3	5	-29.55970149	31.32804	0.676042	
4	5	-22.32462687	38.56312	0.855409	

만약 1번이 일반 플라스틱 그물이고, 5번이 친환경 그물이면 이 분석자료로 어민들을 설득 가능

문제해결 프로그래밍으로 얻을 수 있는 것은?

연구자 : 논문

기업:특허

## 코로나에 가려진 조류독감… 피해규모 역대 2위

4개월여간 2800만마리 살처분 피해 보상 4000억 이상 필요할듯

고병원성 조류인플루엔자(AI)의 전국적인 확산으로 농가의 가축 살처분 규모가 최악의 피해를 입었던 2016~2017년에 이은 역대 두 번째가 된 것으로 나타났다.

국회 농림축산식품해양수산위원회 소속 국민의힘 안병길 의원실이 15일 농림축산식품부로부터 제출받은 자료에 따르면 지난해 10월부터 시작된 AI의 확산이 138일째 이어지며 2808만1000마리의 가축이 살처분됐다. 하루 평

(총 3807만6000마리)가 살처분됐던 역대 최악의 상황에 육박한다. 2016~2017년 당시 살처분 보상금, 생계소득안정자금 지원 등에 투입된예산은 3621억 원이었다. 안 의원은 당시 사례를 기반으로 이번 AI 피해 보상에 4000억 원이상이 필요할 것으로 추산했다.

이날 AI 중앙사고수습본부는 가금농장에서 AI 발생 시 시행했던 예방적 살처분의 대상을 기존 반경 3km에서 1km로 축소 조정하는 내용 등의 '고병원성 조류인플루엔자 방역대책 추진 방안'을 발표했다. 안 의원은 "정부가 충분한 역학조사 없이 무조건 '반경 3km 이내 살처분'이라는 기계적인 대응으로 일관했다"고 지적했다.

고병원성조류인플루엔자가 확산되기 전에 조기에 제어할 수 있을까?

추가 확산이 일어나기 전에 미리 위험한 농 가를 판단하여 방역을 할 수 없을까?

## 고병원성 조류 인플루엔자란?

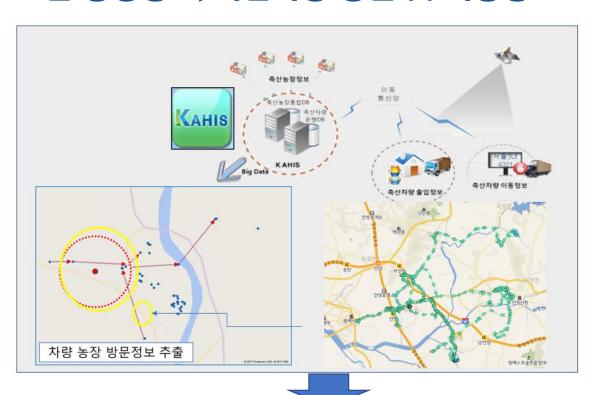
- 잠복기:닭에서의 HPAI 바이러스 잠복기는 수 시간에서 3일 정도
- 병원성: 닭의경우 HPAI에 감염되면 90%에 가까운 치사율, 오리의 경우 10%정도의 폐사가 발생
- 생존력: 농림축산검역본부 자료에 의하면 대략 30일정도의 생존기간을 갖음
- 영향력: HPAI의 항원 발견으로부터 주의해야하는 반경은 아직 알려지지않음, 현재 HPAI가 발견되면 그 농장중심 3km 이내의 모든농장의 가금을 예방적으로 폐사

문제 설정

추가 확산을 막기위해 감염 확률이 높은 위험한 농가를 판단하고, 확산의 원인을 파악하자.

데이터 수집

## 모든 농장정보, 축산차량 방문 및 이동정보



차량번호	방문목적	방문일자	방문농장번호	방문농장주소	축종
02하5245	컨설팅	2019.02.13	00269555	충청남도 공주시 탄천면 남산리 690	메추리(200000)
03저4172	진료.예방접종	2019.02.15,2019.02.14	00924018	충청남도 천안시 동남구 병천면 용두리 227-2	닭-토종닭(5),소-한우(19)
03저4172	진료.예방접종	2019.02.19,2019.02.13	00939283	충청남도 천안시 동남구 병천면 매성리 79	닭-토종닭(0),소-한우(320)
04조0798	인공수정	2019.02.14	00290987	충청남도 아산시 염치읍 쌍죽리 0041-0011	닭-토종닭(0),소-한우(24)

#### 모델 설정

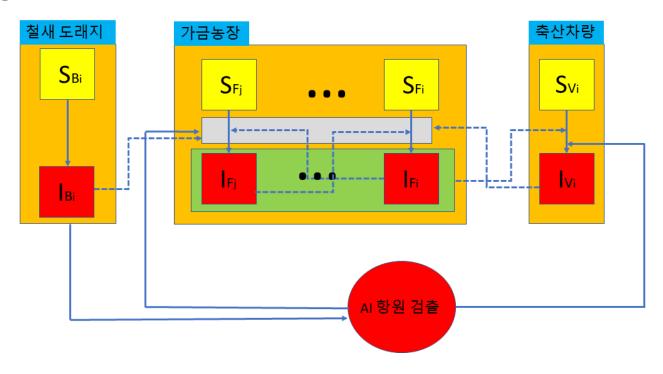


표 2.1 가금농장과 차량의 감염 사건별 감염 확률

사건(Event)	감염 경로	감염 확률
I .	$S_{F_i}, I_{V_i} \rightarrow I_{F_i}, I_{V_i}$	$eta_1 I_{V_i}$
인근 감염된 철새에 의한 가금 농장 간접 감염		$eta_3 w_{ij}$
인근 감염된 가금농장에 의한 감염		$\sum_{j} \beta_2 F_j(f_h) I_{F_j}$
감염된 가금농장에 방문한 차량 감염	$I_{F_i}, S_{V_i} \rightarrow I_{F_i}, I_{V_i}$	$\sum_{j=$ ৬৮৮ ১ ১ $\beta_4 I_{F_j}$

#### 변수 정의

 $\mathrm{t}$  시간에 감염되지 않은 가금농장 :  $S_{\!E}(t)$ 

 $\mathrm{t}$  시간에 감염된 가금농장 :  $I_{\!F}(t)$ 

채취지점 반경  $3 \mathrm{km}$ 를 지나는 차량에서 t 시간에 AI 바이러스에 오염되지 않은 차량 :  $S_{V}(t)$ 

 $\mathrm{t}$  시간에 바이러스에 오염된 차량 :  $I_{V}(t)$ 

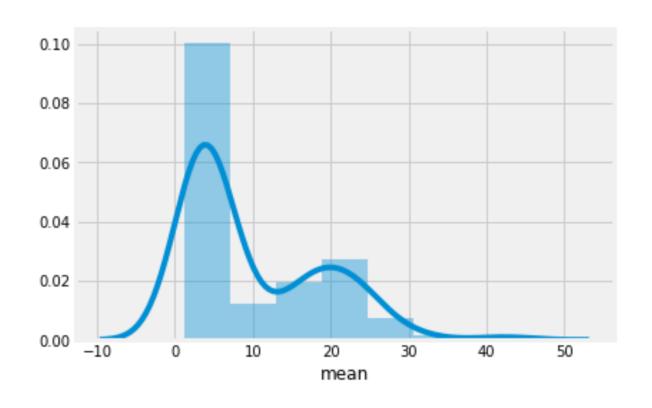
 ${
m t}$  시간에 감염되지 않은 철새도래지:  $S_{\!B}(t)$ 

 $\mathrm{t}$  시간에 바이러스에 감염된 철새도래지 :  $I_{\!B}(t)$ 

모델 평가 및 검증

농장별 위험도 산출

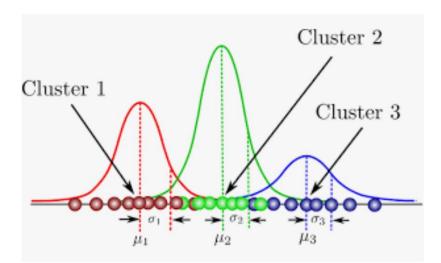
## 총 M회를 N번 시뮬레이션 결과 GMM(Gaussian Mixture Model)로 위험도 계산



## 농장별 위험도 산출 & 지역(시군구)단위 위험도 계산 방법

- 1. 총 M회를 N번 시뮬레이션 결과 GMM(Gaussian Mixture Model)로 위험도 계산
- 2. 군집화(Clustering)를 통한 위험 수준 결정
- 3. 지역(시군구)안의 개별농장 위험도 점수로 지역단위 위험도 계산 및 군집화(Clustering)를 통한 위험 수준 결정

#### GMM 군집화 예시



#### 위험수준별 점수

표 3.1 개별 농장 위험 수준에 따른 점수표							
위험 수준	판단 기준	위험 수준 점수					
관심	Clustering	1					
주의	Clustering	2					
경계	Clustering	4					
심각	Clustering	8					

그림 출처:https://intellipaat.com/community/10664/understanding-concept-of-gaussian-mixture-models

### 모델 평가 및 검증

	farm_index	sub_way	mean	std	conf_interb	conf_interb++	risk	rank	rank-
0	846743	[1028927, , 846743, 20042072]	16.600000	5.207138	6.39401	26.805990	0.748493	4	심각
1	20003421	[인천80바4567, 20003421]	3.066667	1.279881	0.55810	5.575233	0.244017	1	관심
2	20003430	[20005114, 전북84사2940, 20003430]	4.133333	2.416215	0.00000	8.869115	0.365271	2	주의
3	158952	[경북86아7074]	3.733333	2.374467	0.00000	8.387289	0.320627	2	주의
4	870038	[82가5870, 1406636, 870038]	4.000000	2.171241	0.00000	8.255632	0.350647	2	주의



## 군집화 결과



위험 수준	판단 기준
관심	0~0.258
주의	0.258~0.456
경계	0.456~0.712
심각	0.712~1

## 모델 평가 및 검증

	광역시도	시군구	위험도(점수)	순위	rank	rank-
0	경상북도	김천시	100.000000	1.0	4	심각
1	경상북도	문경시	5.187320	2.0	3	경계
2	충청남도	공주시	1.729107	3.0	2	주의
3	경상북도	구미시	1.729107	3.0	2	주의
4	전라북도	무주군	1.729107	3.0	2	주의
5	경상북도	상주시	1.440922	6.0	2	주의
6	경상북도	의성군	0.576369	7.0	1	관심
7	경상북도	칠곡군	0.576369	7.0	1	관심
8	경상남도	함안군	0.288184	9.0	1	관심
9	충청북도	영동군	0.288184	9.0	1	관심
10	경상남도	양산시	0.288184	9.0	1	관심
11	충청남도	논산시	0.288184	9.0	1	관심
12	경상북도	성주군	0.288184	9.0	1	관심

실제 감염지역과 비교







#### 모델 평가 및 검증

#### [경북 김천 김천] 이생조류 분변 H5 항원 검출 관련 빅데이터 기반 기금사육농장 AI 위험도 분석 결과

#### 1. AI 바이러스 항원(H5) 검출현황

시료체취일	검출지역	시료종류
2019.10.23,	경상복도 김천시 양천동 1774-1 (감천)	분변

#### 2. AI 위험도 분석 방법

○ 야생조류 분변시료 채취 전일(10.22)~위험도 분석일(10.25일) 기간 동안 항원 검출 지점 반경 3km 내 지역을 통행한 축산차량의 가금 사육농장 방문내역 추적 ※ 방경 3km 내 닭 45높가 6587수 사용

#### 3. AI 위험도 분석 결과

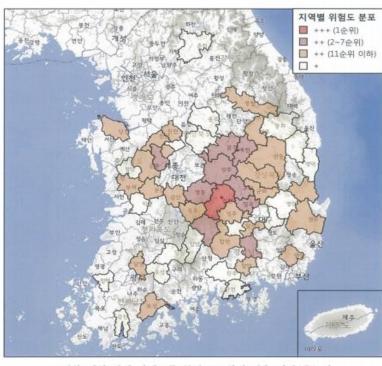
- □ 농장별 위험도
- 항원 검출 지점 반경 3km 내 지역 통행차량 373대 중 202대가 가금사육농장 277개소(9개 서도 52개 서군구) 방문
- ※ 위험도 분석 농장 277개소 및 차량 202대 목록은 붙임(액셀) 참조
- □ 지역(시군구)별 위험도
- O 지역(시군구)별 위험도 순위

① 경상북도 김천시	② 경삼북도 삼주시	③ 충청북도 영동군
③ 겸상북도 구미시	⑤ 경상남도 거창군	③ 경상남도 창녕군
① 경상북도 문경시	① 세종특별자치시	① 경상북도 칠곡군
① 경상북도 예천군		

※ 지역(시군구)별 위험도 분석 결과는 붙임(액셀) 참조

#### O 지역(시군구)별 위험도 분포

야생조류에서 AI 항원 검출지인 경북 김천시에서 최상위 위험이 예측 되며,
 인근 지역인 경북경남충북 소재 시군에 상위위험 분포



<경복 김천 감천 야생조류 분변 H5 항원 검출 지점(●) 및 지역별(시군구) 위험도 분포도>



#### 등록사항

특 허 제 10-2431814 호 Patent Number

#### 특허권자 Patentees

부산대학교 산학협력단(184771-\*\*\*\*\*\*\*) 부산광역시 금정구 부산대학로63번길 2 (장전동, 부산대학교)

대한민국(농림축산식품부 농림축산검역본부장) 경상북도 김천시 혁신8로 177 (율곡동)

#### 박명자 Inventors

조기필(830221-\*\*\*\*\*\*) 부산광역시 금정구 금강로321번길 98-11, 405호

# 강의 일정

• 10월 3일 (개천절) <del>→</del> 12월 19일 (기말고사)

