





실무 프로젝트 기반 빅데이터 전략 마에스트로 과정

BigData분석을 통한 막걸리 활성화 방안

# 맛걸리 프로젝트

김윤진 | 변수현 | 이병민 | 최유림



# CONTENTS



막걸리 시장 현황 및 문제점



데이터 전처리

데이터 수집 데이터 정제 파생속성 생성 데이터셋 생성



데이터 분석

군집 분석(K-Medoids) 분류 분석(Gradient Boosting)



서비스 구현



### 막걸리 시장 현황 및 문제점



찹쌀·멥쌀·보리·밀가루 등을 쪄서 누룩과 물을 섞어 발효시킨 한국 고유의 술

1964 막걸리 쌀 사용 금지로 주질 저하

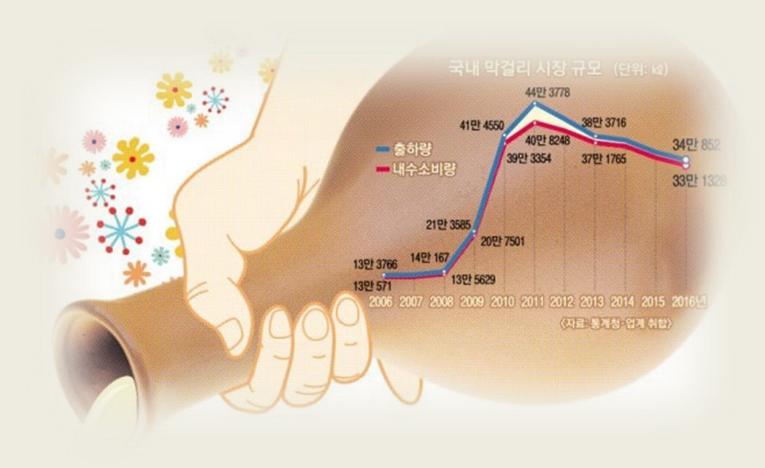
1971 쌀막걸리 재허가, 막걸리 주조방법 규격화

1972 쌀 가격 상승으로 쌀 막걸리 사라짐

2009 막걸리 열풍, 다양한 제품이 출시

출처: 한국민족문화대백과(한국학중앙연구원)

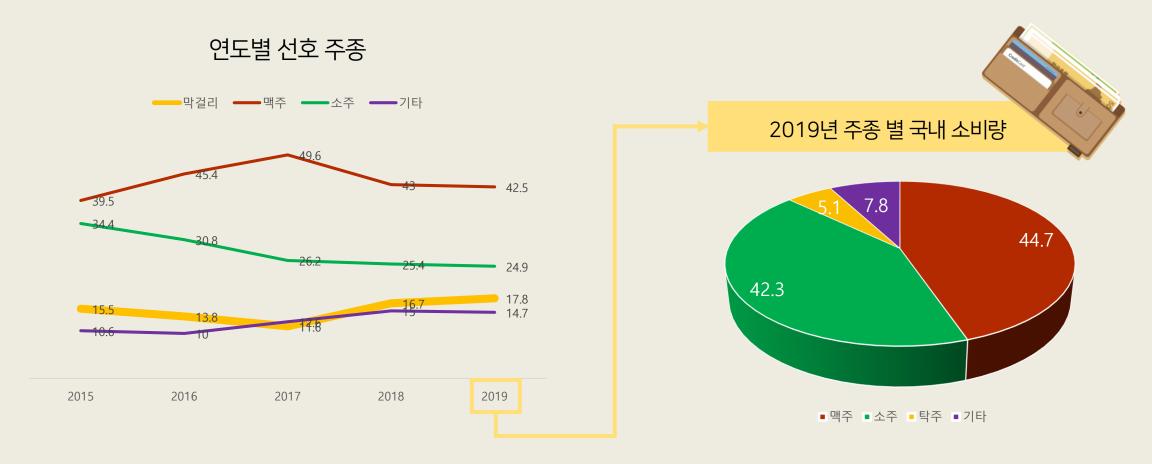
### 막걸리 시장 현황 및 문제점



2011년 이후로 막걸리 출하량 및 내수소비량 지속적 감소

출처: 한국민족문화대백과(한국학중앙연구원)

### 막걸리 시장 현황 및 문제점



막걸리보다 맥주와 소주를 각 2배, 1.5배 선호

2019년 기준 막걸리는 주류 <mark>소비 중 5%</mark>만 차지

출처: 2019 주류 시장 트렌드 보고서



### 막걸리 시장 현황 및 문제점



#### 주류시장트렌드보고서 전통주 맛 속성 질문





막걸리 소믈리에 직업이 있을 정도로 막걸리는 다양한 맛을 가지고 있음

출처: 네이버 쇼핑, 트래비매거진, : 2019 주류 시장 트렌드 보고서

### 막걸리 시장 현황 및 문제점

#### 제조법/발효법에 따른 분류



#### 색상/알코올 함량/당도에 따른 분류

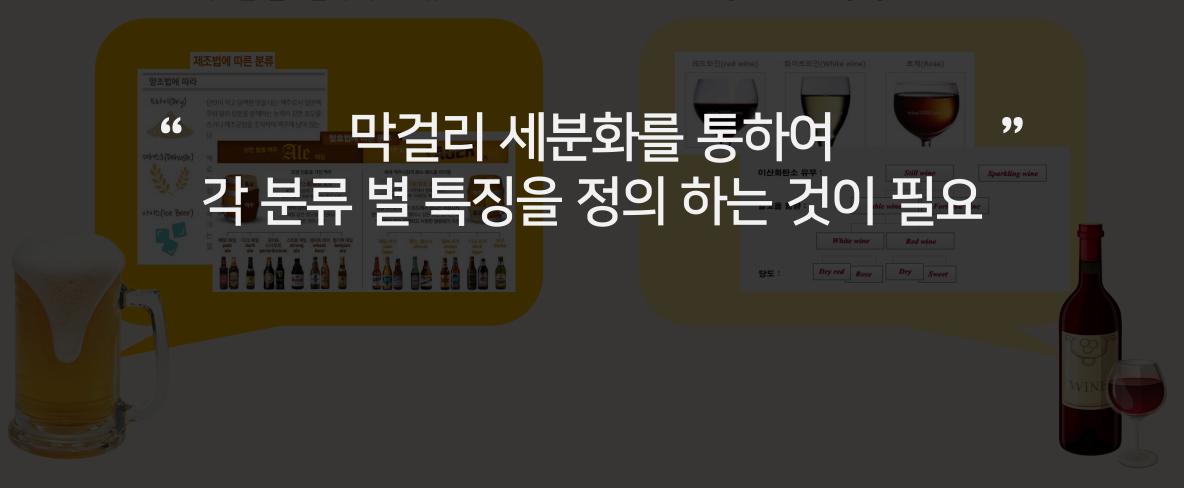


출처: 한국맥주문화진흥회, 한국주류산업협회

### 막걸리 시장 현황 및 문제점

제조법/발효법에 따른 분류

색상/알코올 함량/당도에 따른 분류



출처: 한국맥주문화진흥회, 한국주류산업협

# 데이터 전처리 | 데이터 수집



#### 식품안전나라 막걸리 제품 확인



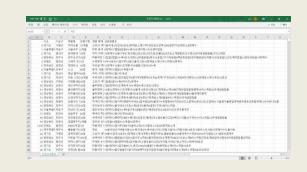
식품유형 : 탁주 소재지, 유통기한, 제품명, 제품형태, 성분 및 원료 등

#### 원본 데이터셋 생성



**네이터 크롤링** 

속성 : 소재지, 제품명, 유통기한, 제품형태, 성분 및 원료 행 : 2251개



# 2 데이터 전처리 I데이터 정제

#### 중복 행 및 결측치 제거

제품명, 유통기한, 소재지, 성분이 같은 행 제거



유통기한, 성분 등 결측치가 있는 행 제거

	시도	시군구 .	제품명 교 유통기한	- 제품 형( - 성분및 [ - ]
	경상북도		강황막걸리 30일	노란색으로1정제수2평화미3쌀4입국발입국5울리고당6곡자7강황뿌리줄기8아스파탐9정제효소제10치자기
1114	경상북도	경주시	강황막걸리 30일	노란색의 (1정제수2쌀3입국업국(쌀)4과당5감황뿌리중기6치자엘매7효모8정제효소제9아세설팤칼룕10아
980	중청북도	정주시	검은콩 막: 365일	1소맥분소맥분5.082조제중국3효모4백미백미10.45전분당6정제효소제7검은공능축액검은공능
1009	충정북도	단양군	검은콩 막:30일	거무스름한1밀가루2쌀3입국4검정콩5검정깨(흑임자)6효모7아스파팀8아세설땀칼륨9정제수
683	경기도	포천시	검은콩막길10개월	우유색의 (33정제수34발35소맥분36검정콩분말37아스파탑38중국39젖산40정제효소제41효모
1939	충청북도	정주시	검은콩막집365일	연회색의 51소맥분2중국조제중국3효모분말4앱쌀백미수입산5평화미분말팽화미분(수입산)6물명전분당7.
1144	광주광역시	광산구	공주 생 쌀 15일	1정제수2쌀(수입쌀100)쌀(수입쌀100)10.23구연산4효모5아스파팀6구연산7입국입국0.0258물
2247	광주광역시	광산구	공주 생 쌀 20일	87쌀88물엿전분당89소맥분90정제수
735	전라북도	진안군	급산인삼성30일	1정제수2팽화미팽화미9.8053고과당고과당3.3354입국입국2.6685누록6효모7첯산8효소처리스
2133	충청남도	금산군	급산인삼산30일	연한 황갈/1소맥분2미분3누룩4아스파람5효모6정제수7총국8수삼(고시형)10불댓11아세설팜칼륨아세설
66	충정북도	정주시	느린마음 115일	고유의 향(9정제수10쌀가루11엣류14입국국15조효소제국16건조효모
200	광주광역시	동구	느린마음 115일	고유의 향(7정제수8쌀가루9입국쌀10조효소제11건조효모활성12물엿예주32,쌀
359	경기도	포천시	느린마을 130일	미황색의 - 45발46국47효모48정제수
934	광주광역시	서구	느린마을 115일	고유의 향(1정제수2쌀가루3물몃예주324입국쌀입국5조효소제6건조효모(활성)
1085	경기도	성남시	느린마을 115일	고유의 향(1정제수2쌀가루3입국쌀입국4조효소제5건조효모활성건조효모,유화제6물엿예주32(Yeju32)
1170	경기도	포천시	느린마을 5	0 1보리2국3호프4효모5정제수
1283	경기도	시흥시	느린마을 115일	고유의 향(31정제수32쌀가루33조효소제34건조효모35정제효소제36입국
2001	경기도	포천시	느린마을 115일	고유의 향(1정제수2쌀가루3입국4조효소제5건조효모6정제효소제
			느린마을 115일	고유의 향(9정제수10쌀가루11엣류14입국15조효소제16건조효모
	대구광역시		느린마을 막15일	교유의 향(13정제수14쌀가루15입국쌀(쌀입국)16조효소제17건조효모18물엿예주32(쌀가루,정제수,정제3
			느린마을 15일	고유의 향(1정제수2쌀가루3입국쌀입국4조효소제5건조효모6정제효소제
	충청남도		더덕 막걸: 30일	1소맥분소맥분7.92팽화미팽화미7.93더덕4누룩5효모6조제종국7젖산8아스파탐9정제수
	경기도	연천군	더덕 막걸:1년	액상제품 1정제수2쌀백미3일가루4입국밀입국5더덕농축액6아세설팜칼륨7정제효소제8아스파탐9효모
		가평군	더덕막걸리10개월	1정제수2쌀(수입산)쌀(수입산)13.3113쌀입국쌀입국3.5844검은콩(수입산)검은콩(수입산)0.491
	경기도	포천시	더덕막걸리 10개월	우유색의 [1정제수2백미백미8.293소맥분소맥분7.644더덕분말5아스파탐6중국7젖산8정제효소정제효소(
		청주시	더덕막걸리 365일	1소맥분소맥분4.6732조제중국3효모4백미백미9.3475전분전분1.5356정제효소정제효소0.0097
2219	경삼북도	상주시	더덕막걸리1년	약간 황색:1정제수2쌀3밀가루4평화미5효모6젖산7물엿8입국밀9정제효소제10아스파탐11더덕추출물12

#### 성분 컬럼 정리



#### 1. 성분 컬럼 나누기

성분및원료			
1쌀국내산2입국3효모	4정제효소	제5기타과당	상포도당
6정제수7멥쌀8찹쌀9-	누룩10오렌	지제스트오	렌지껍?
19정제수20쌀가루21	입국쌀22포	도당23건조	효모(활
1찹쌀2멥쌀3누룩4건	조국화5쇠두	무릅분말6하	수오분 <sup>다</sup>
9국10효모11밀가루1	2쌀13올리그	고당14정제.	효소저
13정제수14쌀15조제	종국16벌꿀	17효모18누	-룩
1정제수2멥쌀3누룩밀	누룩		
1정제수2쌀3국7효모			
1정제수2쌀3입국밀입	국4밀가루	5팽화미6사	과농축 <sup>ç</sup>
1찹쌀2쌀3누룩4미나	리5정제수		
7쌀8정제수9조제종국	10누룩밀누	-룩11효모1	2젖산

0					
정제수	쌀가루	입국쌀	포도당	건조효모(활	조.
차·쓰	멥쌀	누룩	건조국화	쇠무릅분밀	하
	효모	밀가루	쌀	올리고당	정.
01	쌀	조제종국	벌꿀	효모	누
정제수	멥쌀	누룩밀누룩	ŧ		
정제수	쌀	국	효모		
정제수	쌀	입국밀입국	밀가루	팽화미	사.
찹쌀	쌀	누룩	미나리	정제수	
쌀	정제수	조제종국	누룩밀누룩	효모	젖.

#### 2. 성분명 정리하기

마(산약)마 마 마가목마기마가목

마가목마기 마개량누룩 및 밀누룩 마가목마기 마개량누룩 및 밀누룩 마가목열미마 개량누룩 및 밀누룩 마케 량누룩 이 누룩 말토을리고 올개량누룩 지수룩 매실과즙 매상 등 연한 방풍

당군1	성군2	성군3	성군4	성군5	성군6	ĉ
ţ	입국	효모	효소	포도당	설탕	틴
성제수	멥쌀	찹쌀	누룩	오렌지		
성제수	쌀	입국	포도당	효모	효소	구
쌀	멥쌀	누룩	국화	쇠무릅	하수오	7
1	立모	밀	쌀	올리고당	효소	0
제	쌀	종국	꿀	효모	누룩	
	멥쌀	밀누룩				
	쌀	국	立모			
성제	밀입국	밀	쌀	사과	구연산	0
쌀	쌀	누룩	미나리	정제수		
ţ	정제수	종국	밀누룩	立모	젖산	
ŀ	정제수	종국	물엿	효모	젖산	豆
ŀ	정제수	종국	물엿	효모	젖산	豆

# 2 데이터 전처리 | 파생속성 생성



#### 파생속성 생성

1. 성분

누룩 : 곡, 입국, 국, 종국

탄산 : 탄산, 일반

종류: 막걸리, 동동주, 전통주, 청주

과일 : 오렌지,약재,율무...

주원료: 쌀, 밀, 쌀&밀, 기타

2. 제품형태

색 : 백색, 미황색, 연미색, 유백색...

3. 유통기한

살균: 생, 살균



#### 제품 별 맛(단맛,신맛,고소한맛,드라이) 속성 생성

- 1. 성분 별 맛(단맛, 신맛, 고소한맛, 드라이) 확인
- 2. 제품 별 맛 컬럼 생성

성분	맛			드라이	신맛	고소한맛	단맛	탄
개암가루	고소한맛			1	1	0	2	
아스파탐	단맛 녹기	사	드라이	0	1	0		
솔잎	드라(라	베더	드라이	1	1	0		
밀	고소 <sup>(</sup> 옥-		고소한맛	1	0	0	1	
백미				0	1	0	2	
정제수	711.	보마일	꽃드라이	1	0	0	0	1
立 " I 克모	후	추	드라이	0	0	0	-	
스테비아	단맛 히브	비스커	스시마	1	1	0	_	
				0	1	0		
배추	고소한건	포도	단맛	1	0	0		
복분자	신맛, 검	ᄋᆍ	고소한맛	0	0	0	2	
블루베리			고조안짓	0	0	0	2	1
2117	<sup>^</sup> 검	정깨	고소한맛	0	0	0	_	
	흑[	ור	고소한맛	1	1	0	_	
	₹,	-1	고모인것	0	1	0	2	L

국	탄산	살균/생	종류(동동	색	과일	단맛	주원료
입국	탄산	살균	막걸리	백색	0	설탕, 기타.	쌀
누룩	일반	살균	막걸리	미황색	오렌지	0	쌀
입국	일반	생	막걸리	미황색	0	포도당, 아	쌀
누룩	일반	생	전통주	미황색	약재	0	쌀
국	일반	살균	막걸리	백색	0	올리고당,	씰,밀
종국	일반	생	막걸리	우유색	0	벌꿀	쌀
누룩	일반	살균	막걸리	백색	0	0	쌀
국	일반	살균	막걸리	유백색	0	0	쌀
입국	탄산	살균	막걸리	우유색	0	아스파탐	밀
누룩	일반	생	전통주	미색	0	0	쌀
종국	일반	생	막걸리	백색	0	0	쌀
종국	일반	생	막걸리	백색	0	물엿, 아스	쌀
종국	일반	생	막걸리	백색	0	아스파탐,	쌀
종국	일반	생	막걸리	백색	0	물엿, 아스	쌀
입국	일반	생	막걸리	미색	0	물엿, 사카	밀
국	일반	생	막걸리	미색	0	물엿, 설탕,	쌀

# 2 데이터 전처리 | 데이터셋 생성



#### 데이터 범주화

누룩: 곡 0/ 입국 1/ 종국 2/ 국 3

탄산: 없음0/있음1

종류: 막걸리 0/동동주1/전통주2/청주 3

주원료: 쌀0/밀1/쌀밀2/기타3

살균: 없음0/있음1

과일: 없음0/있음1

단맛: 단맛 없음()~아주 단맛(4)

신맛: 없음0/있음1

드라이: 없음0/있음1

고소한맛: 없음0/있음1



#### 데이터 셋 생성

1. 군집분석 : 주원료, 누룩, 살균, 탄산

2. 분류분석: 단맛, 과일, 신맛, 드라이, 고소한맛

제품명	국	살균	주원료	탄산
막이오름	1	1	0	1
서울곡주요	0	1	0	C
동네방네수	1	0	0	C
청주신선주	0	0	0	C
고세이프리	3	1	2	C
대대포13	2	0	0	C
노크	0	1	0	C
옛날할머니	3	1	0	C
이동스파를	1	1	1	1
의령황새골	0	0	0	C
세오녀탁주	2	0	0	C
홀인원탁주	2	0	0	C
탁주블라썯	2	0	0	C
보경사탁주	2	0	0	C
강철구의미	1	0	1	C
아리스타싱	3	0	0	C

제품명	과일	드라이	신맛	고소한맛	단맛
막이오름	0	1	1	0	2
서울곡주요	1	0	1	0	0
동네방네수	0	1	1	0	1
청주신선주	1	1	1	0	0
고세이프리	0	1	0	0	1
대대포13	0	0	1	0	2
노크	0	1	0	0	0
옛날할머니	0	0	0	0	0
이동스파를	0	1	1	0	2
의령황새골	0	0	1	0	0
세오녀탁주	0	1	0	0	0
홀인원탁주	0	0	0	0	2
탁주블라썯	0	0	0	0	2
보경사탁주	0	0	0	0	2
강철구의미	0	1	1	0	2
아리스타싱	0	0	1	0	2

# 데이터 분석 | 군집 분석

#### K-Means

#### 연속형 자료

샘플을 K개의 부분 집합으로 분리

현재의 분할이 군집이 되고 이 분할의 중심을 계산(<mark>평균</mark>)

각 객체를 가장 가까운 중심에 할당(<mark>거리</mark>)

다시 시작하여 더 이상 개체의 움직임이 없으면 마침



### K-Medoids

#### 범주형 자료

샘플을 K개의 부분 집합으로 분리

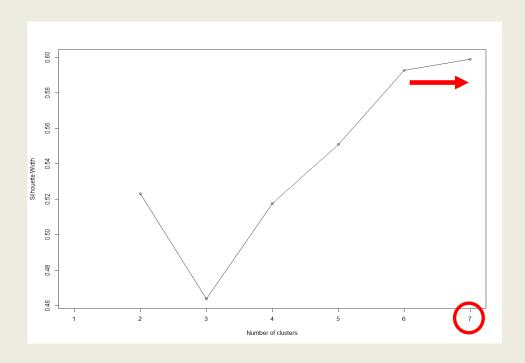
현재의 분할이 군집이 되고 이 분할의 중심을 계산(<mark>최빈값</mark>)

각 객체를 가장 가까운 중심에 할당(<mark>빈도</mark>)

다시 시작하여 더 이상 개체의 움직임이 없으면 마침

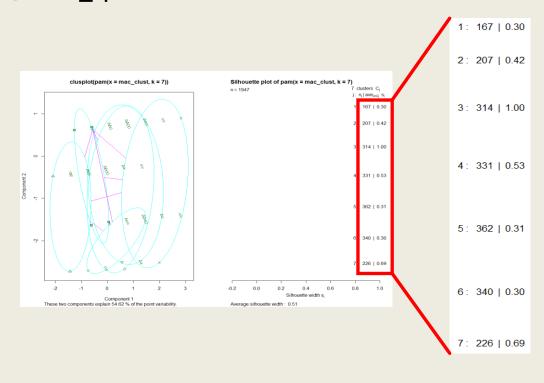
## 데이터 분석 | 군집 분석

#### 1. 군집의 수 설정



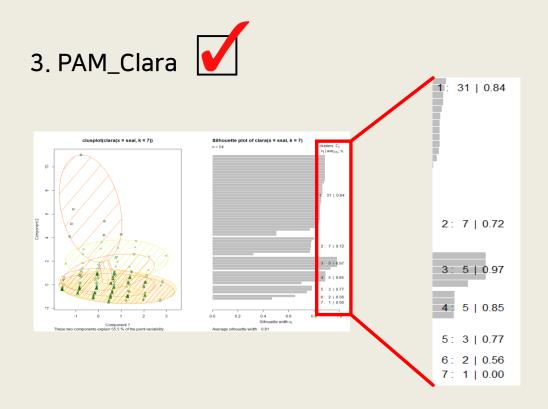
dasiy함수를 통해 최적의 군집 개수 도출 : <mark>7개</mark> (Ward연결법 기준)

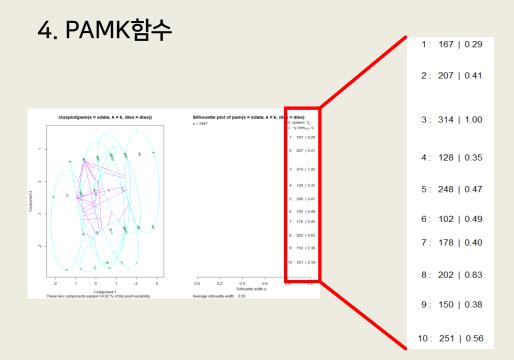
#### 2. PAM함수



PAM함수의 경우 <mark>큰 데이터셋에 비효율적</mark> 군집 설명력이 좋지 않음

### 데이터 분석 | 군집 분석





큰 데이터의 샘플을 가져와서 <mark>각 샘플에 PAM함수를 적용</mark>해 최선의 결과를 반환하는 CLARA 알고리즘

신뢰할 수 있는 군집 설명력

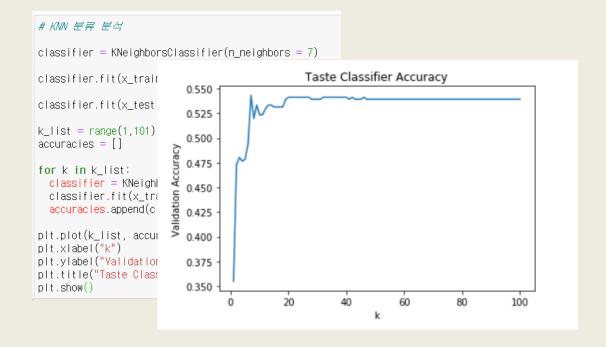
군집의 수를 자동적으로 <mark>설정</mark>하여 군집분석을 실행 Clara보다 설명력이 낮음



# 3 데이터 분석 | 분류 분석

#### sparse\_categorical\_crossentropy

#### KNeighborsClassifier(KNN)



정확도 0.49

정확도 0.54

### 데이터 분석 | 분류 분석

#### **Gradient Boosting**



```
# 사전가지치기와 학습률 설정하여 학습 실시
gbrt = GradientBoostingClassifier(random_state=0,
                              max_depth=3,
                              learning_rate=0.1,
                              n_estimators=100,
                         # 사전가지치기와 학습률 설정하여 학습 실시
                         gbrt = GradientBoostingClassifier(random state=0,
                                                        max depth=3.
gbrt.fit(x_train, y_train)
                                                        learning_rate=0.1,
                                                       n_estimators=100,
|print("훈련 세트 정확도 : -
                                                        validation fraction=0.1
print("테스트 세트 정확도
C:\BigData\anaconda3\lib\si gbrt.fit(x_train, y_train)
y was expected. Please char
 y = column or 1d(y, warn=|print("훈련 세트 정확도 : {:.2f}".format(gbrt.score(x_train,y_train)))
훈련 세트 정확도 : 0.58
                         print("테스트 세트 정확도 : {:.2f}".format(gbrt.score(x_test,y_test)))
테스트 세트 정확도: 0.54
                         C:\BigData\anaconda3\lib\site-packages\sklearn\ensemble\_gb.py:1454: Data
                         y was expected. Please change the shape of y to (n_samples, ), for exar
                           y = column or 1d(y, warn=True)
                         훈련 세트 정확도: 0.67
                         테스트 세트 정확도 : 0.65
```

정확도 0.65

#### 동적프로그램 생성

```
드라이 한 것을 좋아하십니까?(YES:1/N0:0) : 0
신맛을 좋아하십니까?(YES:1/N0:0) : 0
고소한맛을 좋아하십니까?(YES:1/N0:0) : 1
단맛을 얼마나 좋아하십니까?(0~4) : 1
과일 막걸리를 좋아하십니까?(YES:1/N0:0) : 1
```

```
드라이 한 것을 좋아하십니까?(YES:1/N0:0) : 0
신맛을 좋아하십니까?(YES:1/N0:0) : 0
고소한맛을 좋아하십니까?(YES:1/N0:0) : 1
단맛을 얼마나 좋아하십니까?(0~4) : 1
과일 막걸리를 좋아하십니까?(YES:1/N0:0) : 1
[3]번 군집 막걸리 입니다
['이동마가리미주(중국수총전용)' '제주한라봉막걸리' '포천일동쌀막걸리(미국수총용)']막걸리를 추천합니다.
```

### 서비스 구현



식품유형	탁주
에탄올 함량	10%
원재료명 및 함량	쌀(국내산), 물, 국, 효모, 젖산 (산도조절제), 밀 함유
내용량	750ml
업소명	(주)배혜정도가
소재지	경기도 화성시 정남면 서봉로 835
유통기한	병 어깨 표기일까지
보관방법	10℃이하 냉장 보관
품목보고번호	2013001702836

19세 미만 판매 금지 부정·불량식품 신고는 국번 없이 1399

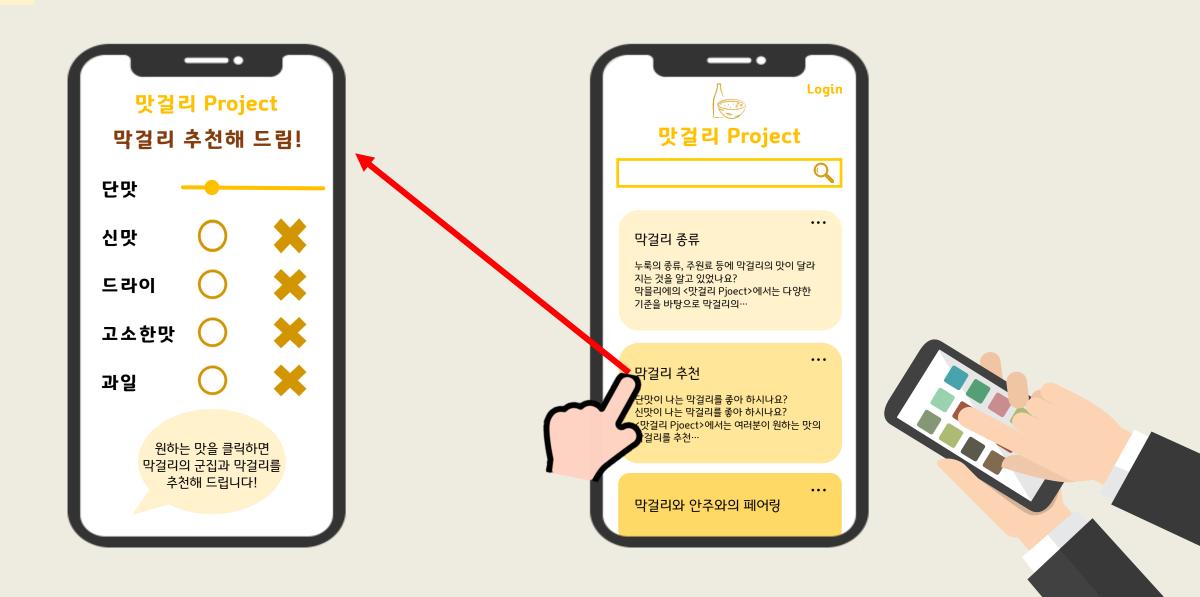
#### 경고

지나친 음주는 뇌졸중, 기억력 손상이나 치매를 유발합니다. 임신 중 음주는 기형아 출생 위험을 높입니다. 제품 파손의 우려가 있으니 고온의 밀폐 공간에 놓아두지 마십시오.

A 막걸리(쌀)



성분표 하단에 막걸리의 군집을 표기해 막걸리 정보를 제공



# Q&A