「2022년 빅데이터 아이디어 오디션」 아이디어 기획서

1. 참가자 정보	
개인·팀·기관명	요들벡스
연락처	(전화) 010-9972-2974 (전자우편) ysh03634@naver.com
아이디어명	기업의 소셜 데이터 정보를 정교화해 기업의 평판지수 예측

2. 기획서 작성

1) 아이디어 요약

○ 추진배경 및 필요성

[대한민국 실업률 , 2012~2021]

- 시간이 지날수록 한국의 실업률은 감소하는 추세보다 증가하는 추세로 보이고 특히 실업률에서도 청년 실업률의 상승 속도가 빠르다.
- 청년 실업률이 높은 이유는 청년들이 바라는 양질의 일자리가 부족한 것과 안정적인 직업을 갖는 것이 낫다는 인식으로 공무원시험 수험생이 많아지고 있는 점들이 실업률을 높이는 원인이 된다.
- 이를 해결하기 위해 마구잡이 식으로 일자리 창출만 앞세우다 보니 비정규직, 인턴사원 등 늘어나 앞으로 안정된 기반이 마련되지 않아 일시적인 현상으로 감소될 뿐 근본적인 문제는 해결되지 않는다.
- 다양한 기업들의 데이터 기반의 객관적인 지표로 기업들의 평판을 보여주며 구직자들에게 기업의 장점을 어필하며 기업에 맞는 사람들을 구하는 것이 중요해짐
- 다양한 범주의 데이터를 활용해 기업 성장성, 기업추천 분석 결과를 통해 B2C, 혹은 B2B 운영을 통한 실업률 감소가 필요해짐

○ 아이디어 요약
- 최종 목표 : 실업률을 줄이기 위한 기업의 추천성, 성장성 데이터 구축
- 목적 : 기업의 소셜 데이터 수집 및 구축을 통한 기업 성장성, 추천성 예측
기업이 ESG 경영을 할 수 있도록 컨설팅

○ 추진방안
1. 기업의 소셜데이터 정보 DB 구축
2. DB에 적재된 데이터를 기반으로 고부가가치를 가진 데이터의 생산하여 빅데이터 거래시장을 활성화
3. 구축된 빅데이터 플랫폼을 활용하여 기존 컨설팅 서비스 대비 차별성을 확보

2) 활용 DB 목록 (최소 3개 이상 활용)

□	1. 기업 경쟁력 정보 DB	□	7. 국내 시장 동향 DB
□	2. 기술기반 보유 역량 DB	□	8. 경쟁사 현황 DB
□	3. 인적자원 보유 역량 DB	■	9. 온라인뉴스 DB
□	4. 마케팅 보유 역량 DB	□	10. 비즈니스모델 정보 DB
■	5. 기업 정보 요약 DB	■	11. 취업포탈뉴스 DB
□	6. 글로벌 시장 동향 DB	□	12. RND 정보 DB

3) 추가 융합 DB

1. 잡플래닛(https://www.jobplanet.co.kr/welcome/index) : 기업 평판 리뷰 크롤링
2. 블라인드(https://www.teamblind.com/kr/) : 기업 평판 리뷰 크롤링

4) 분석 기법

■	Regression	□	lightGBM	□	Gradient Boost	□	AdaBoost
□	DNN	■	Random Forest	■	Decision Tree	■	Naive Bayes
□	SVM	□	LSTM	■	K-means	□	DBSCAN
■	기타:				)		

5) 분석 내용

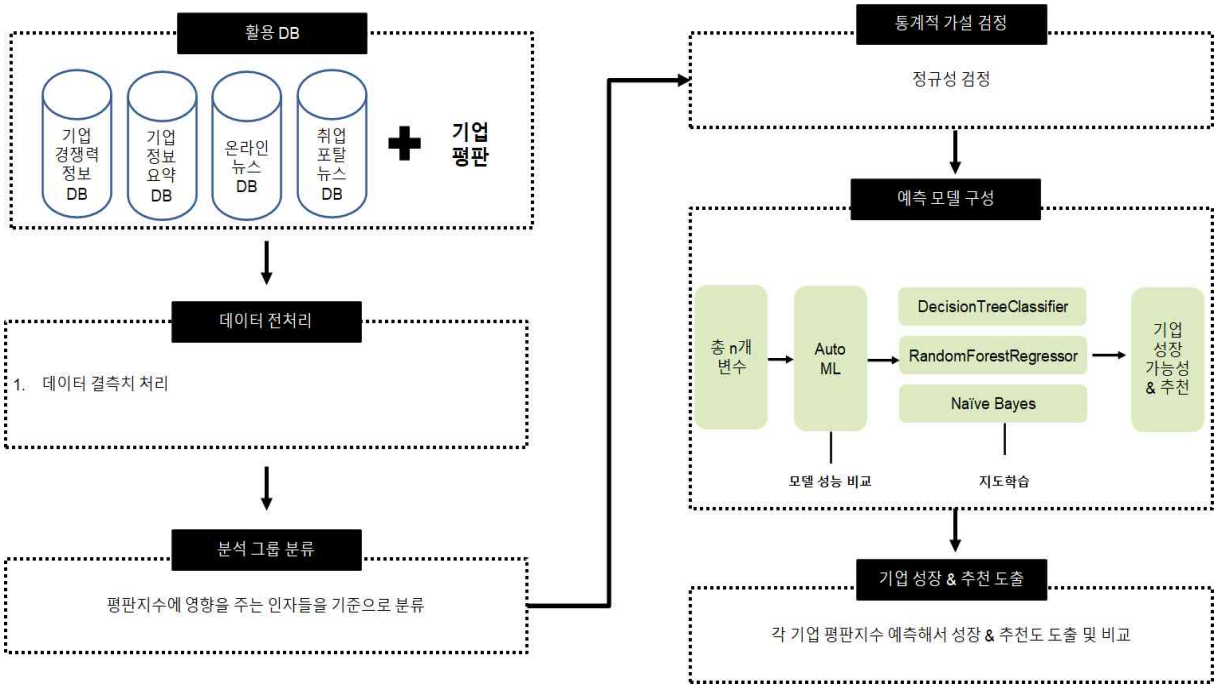
1. 융합데이터 생성
- 활용 데이터

활용 데이터	
대회용 데이터	기업 경쟁력 정보 DB
	기업 정보 요약 DB
	온라인뉴스 DB
추가 데이터	기업 평판 댓글

3개의 DB를 사업자 번호를 기준으로 융합했다.
또한 제공된 사업자번호를 조회해 기업 확인 후 취업포탈에서 평판 댓글을 추가해 활용함

2. 분석 목적
기업 평판지수가 낮은 기업들이 높은 기업들을 벤치마킹해 여러 구직자들을 불러모을 수 있는 방법에 대해 인사이트를 도출한다.

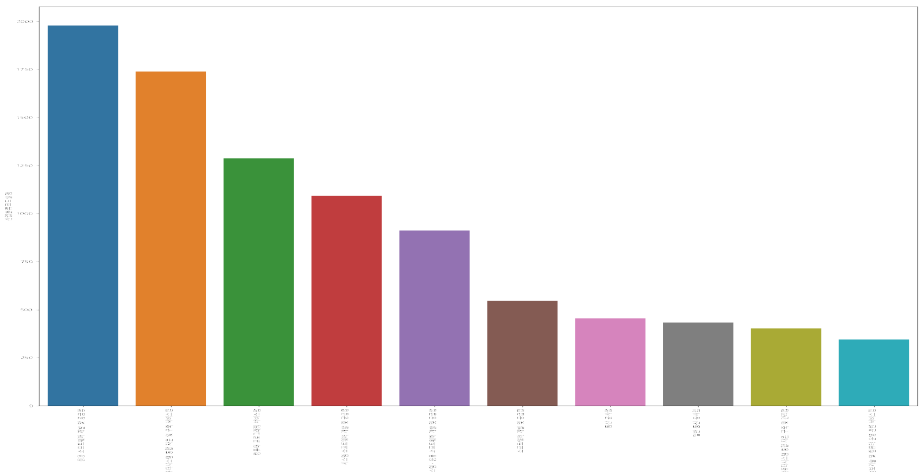
3. 분석방법



1) 데이터 전처리

- 기업 정보 요약 DB, 온라인뉴스 DB, 취업포탈뉴스 DB 3가지 데이터 융합 후 평판지수에 반영될 컬럼에 null 값이 들어간 결측치 처리

2) 분석 그룹 분류



- 기업 수 순으로 상위 10개 업종 중 6개 선정해서 사용
- '정보서비스업', '도매 및 상품 중개업', '실내건축 및 건축마무리 공사업', '토목 건설업', '전기 공사업', '게임 소프트웨어 개발 및 공급업'

3) 통계적 가설 검정

- 기업 평판지수에 이용할 연속형, 범주형 데이터들을 가설검정을 시행하였고 통계적으로 유의한지 확인해보았을 때 다음과 같은 결과가 나왔다.

1. 정규성 검정
   - 방법 : Normal Test, ranksums, kruskal, chi2_contingency
   - 사업자등록번호 : pvalue=5.403558898263674e-07
   - 종합점수 : pvalue=0.034614727574078 97
   - 승진기회점수 : pvalue=0.005190154789704387
   - 복지및급여점수 : pvalue=0.0007489287229882414
   - 기업문화점수 : pvalue=52.1995497773508805e-05
   - 기업규모, 기업형태 : pvalue= 2.0385644140415173e-27

결과 : 비정규분포가 대부분임을 알 수 있다.

4. 예측 모델 구성

<변수 중요도 추출>

DecisionTreeRegressor

	Feature	Importance
0	종합점수	0.380404
1	승진기회점수	0.143477
2	복지및급여점수	0.122217
10	x0_정보서비스업	0.070697
3	업무생명균형점수	0.060140
5	경영진점수	0.051557
11	x0_토목 건설업	0.043995
13	x1_일반	0.038044
7	x0_도매 및 상품 중개업	0.029654
9	x0_전기 공사업	0.023831
4	기업문화점수	0.019993
8	x0_실내건축 및 건축마무리 공사업	0.015992
6	x0_게임 소프트웨어 개발 및 공급업	0.000000
12	x1_외감	0.000000
14	x1_코스닥등록	0.000000

XGBRegressor

	Feature	Importance
0	종합점수	0.227923
10	x0_정보서비스업	0.127668
7	x0_도매 및 상품 중개업	0.110482
1	승진기회점수	0.106387
2	복지및급여점수	0.101570
11	x0_토목 건설업	0.091245
3	업무생명균형점수	0.061798
5	경영진점수	0.045316
13	x1_일반	0.041110
14	x1_코스닥등록	0.028741
4	기업문화점수	0.021803
8	x0_실내건축 및 건축마무리 공사업	0.020167
9	x0_전기 공사업	0.011384
12	x1_외감	0.003847
6	x0_게임 소프트웨어 개발 및 공급업	0.000560

RandomForestRegressor

	Feature	Importance
0	종합점수	0.285635
1	승진기회점수	0.188942
2	복지및급여점수	0.106446
5	경영진점수	0.097703
4	기업문화점수	0.094460
3	업무생명균형점수	0.072980
7	x0_도매 및 상품 중개업	0.038132
11	x0_토목 건설업	0.027084
8	x0_실내건축 및 건축마무리 공사업	0.024898
9	x0_전기 공사업	0.024893
10	x0_정보서비스업	0.013675
13	x1_일반	0.009609
14	x1_코스닥등록	0.008292
12	x1_외감	0.007252
6	x0_게임 소프트웨어 개발 및 공급업	0.000000

<분류 모델 평가 Accuracy >

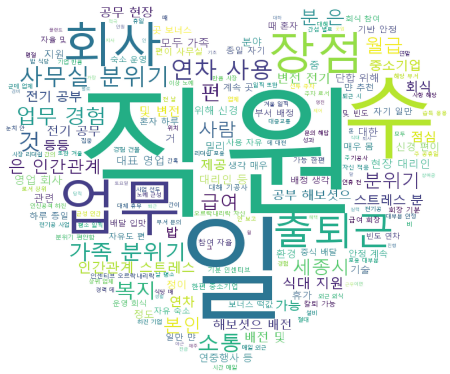
모델 종류	학습능력평가	일반화능력평가
DecisionClassfier	0.93	0.88
RandomForestClassifier	0.94	0.90
KNeighborsClassifier	0.93	0.89
GaussianNB	0.82	0.72

5. 분석 결과

1) 모델 성능 및 변수 중요도

RandomForestClassifier 이용해 기업 평판지수로 기업의 성장 가능성, 추천을 예측하는 모델 중 가장 성능이 좋은 모델을 사용했다.

2) 각 기업의 장단점, 및 기업 평판 댓글을 워드클라우드로 시각화해서 표현했다.



6) 아이디어 활용방안 및 기대효과

○ 아이디어 활용방안

1. '기업 추천, 성장 서비스'를 제공하는 플랫폼

- 플랫폼에 적재된 양질의 기업의 데이터를 바탕으로 기업이 셀프로 자가진단 할 수 있도록 하거나 높은 추천,성장을 가진 기업들을 벤치마킹해 기업의 취약점을 파악하고 개선할 수 있도록 한다.

2. 구직자들을 위한 취업 알선 서비스 제공

- 기업의 성장, 추천 서비스로 구직자나, 이직자, 대학교 4학년 취업준비생들을 타겟으로 구직자들의 성향에 따른 기업을 추천할 수 있도록 한다.
- 대학교들과 연계해서 졸업준비생들을 타겟으로 한 취업 알선 서비스 제공

3. 기업 정보를 활용해 기업을 평가하는 기관들에게 인사이트 제공

- 양질의 기업정보를 이용할 수 있는 구독 서비스를 만들어 보증기관, 은행, 투자자들에게 수익을 창출 할 수 있도록 한다.

○ 기대효과

- 경제적 관점

구직자들에게 알맞은 기업을 추천해주기 때문에 업무역량도 좋을것이라고 기대되며
실업률을 줄여서 경제활동을 하는 사람들을 증가시켜주기 때문에 경제적으로도
많은 이점이 있을 수 있다는 기대가 된다.

- 사회적 관점

중소기업은 별로라는 사회적 관점이 있지만 기업의 추천과 성장 가능성을 보여주면서
사람들에게 중소기업이 별로라는 관점을 바꿔줄 수 있다고 기대된다.

- 기술적 관점

대기업, 중견기업, 중소기업, 스타트업 등등 여러 사람들이 분포해 취직을 하게된다면
산업구조에 선순환이 이루어지면서 기술적으로도 많은 교류가 있을거라고 기대되며
중소,스타트업에도 기술력이 많이 센 기업들이 나타날것이라고 기대된다.

7) 에로사항

제공된 데이터에대한 모수데이터가 부족해서 어떤식으로 분석을 해야하는지 어려웠고
부족한 데이터는 크롤링으로 수집시도를 하고 일부 데이터를 수집했지만
평판에 대한 데이터 크롤링은 오픈 API가 적어 시간적으로 문제점이 있었다.
그리고 응용분야에서 하는 평판이 이에 당사자에 따라서 원하는 평판정보가 다
틀리기 때문에 정형화시키는데 어려움이 있었습니다.
데이터 결측치에 대한 고민과 검토가 부족한점도 아쉬웠습니다.