붙임 2 **정책제안(아이디어) 작성 양식**

접 수 번 호	2076	※ A4용지 5장 이내로 작성
응모자(팀장)명	황인영	
응 모 주 제	중소기업 특징별 분류 및 우수기업 선별 모델	
제안정책명	'TED'를 활용한 기업 선별 모형	
활 용 자 료	① 중소기업중앙회 제공자료(1개 이상 체크) ■기술통계조사 ■실태조사 □제조업직종별임금조사 □경기전망조사	
	② 공공데이터( )	
	③ 민간데이터( )	
	④ 기타(온라인 기사 크롤링)	
제 안 정 책 분 석 요 약	'TED'를 활용하여 중소기업들을 분류한 후, 각 군집별로 적정(우수)한 기업들을 선정하는 것이 아이디어의 핵심 내용이다. 이때, 중소기업을 분류하는 기준인 TED는 'Type of industry(산업군)', 'Enterprise growth(기업성장단계)', 'Data character(기업특성)'을 의미한다. 산업군은 제조/비제조로 분류되며, 기업성장단계는 창업초기& 안정성장/ 성장/ 성숙 단계로 분류된다. 마지막으로 기업의 특성은 군집분석(Spectral Clustering)을 활용하여 취약형(Weak)/ 안정형(Stable)/ 도전형(Challenge)으로 분류하였다. 이렇게 총 18개(2*3*3=18) 그룹으로 분류한 후, '예측분석(XGBoost)'을 실시하여 각 군집별로 우수 기업을 선정한다. 즉, 각 기업의 성장여부를 예측하여 지원 정책을 실행하기에 우수한 기업들을 선별하는 것이다.	

1. 주제 선정 및 자료분석 배경

■ 주제 선정 배경 및 필요성

- 중소기업이 필요로 하는 지원 정책 분석 1)

위 이미지는 중소기업 관련 이슈 및 현황을 분석하기 위해 2017~2019년도의 '중소기업 정부지원' 관련 뉴스를 크롤링하여 워드클라우드를 생성한 결과이다. 텍스트 마이닝을 통해 '기술', '원금', '긴급', '대출', '컨설팅', '고용', '해외' 등 다양한 단어가 추출되었다. 이를 통해 중소기업이 지원받고자 하는 분야가 다양하다고 볼 수 있다. 따라서 기업 특성에 따른 분류와 선별을 통한 맞춤 지원이 필요하고 판단하였다.

- 현 중소기업 관련 지원 정책의 문제점

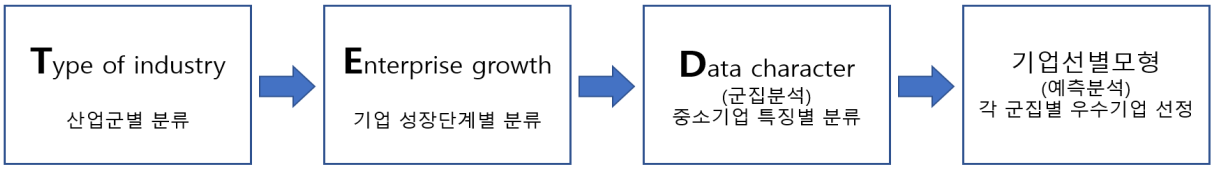
중소벤처기업부는 중소기업의 성장환경 구축과 대/중소기업간 격차해소를 위해 다양한 지원정책을 펼치고 있다. 그러나 일각에서는 이러한 정책이 '자금지원'에 초점이 맞춰져 있으며, 기업의 산업구조를 반영하지 못하고 있다는 의견[2]이 나오고 있다. 또한 퍼주기식 지원으로 기술력 있는 기업의 지원보다, 좀비기업을 늘리는 정책이라는 비판 또한 생겨나고 있다. [3]그러한 문제점을 해결하기 위한 데이터 분석 기반의 'TED를 활용한 기업 선별 모형'을 제안한다.

1) 분석도구 : Python 의 BeautifulSoup, random, time 등

2) 김선우, 김재원 (2020). 혁신성장을 위한 중소기업 R&D지원 개선방안, p.21

3) 양종곤, "돈 주는 정책, 마약과 같다"...중기 전문가들의 쓴소리, <서울경제>, 2019.10.26

■ 제안 아이디어



앞서 설명한 문제점을 해결하기 위해 'TED를 활용한 기업 선별 모형'을 제안한다.

TED는 인간의 성격을 4단계 과정으로 16가지의 그룹으로 분류하는 'MBTI 성격유형 검사'에서 아이디어를 착안해내었다. 이에 'TED'는 3단계의 분류 과정을 거쳐 기업을 특성별로 분류한다.

① 산업군은 제조/비제조로 분류되며, ② 기업성장단계는 창업초기&안정성장/ 성장/ 성숙 단계로 분류된다. ③ 기업의 특징은 군집분석을 활용하여 취약형(Weak)/ 안정형(Stable)/ 도전형(Challenge)으로 분류하였다.

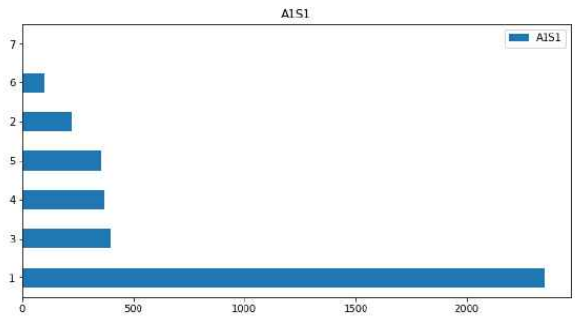
이렇게 총 18개(2*3*3=18) 그룹으로 분류한 후, 예측분석을 실시하여 각 군집별로 우수기업을 선정한다.

2. 분석 내용

■ 활용 데이터 설명

- 기술 통계조사

기술 통계조사 자료에 대한 충분한 탐색을 위해 EDA(탐색적 자료 분석)를 진행하였다. EDA 과정을 통해 기업의 현 상황, 기업의 목표, 기업이 원하는 지원 정책을 반영하는 변수 12개를 선정하였다. 또한 실태조사 자료와 함께 사용하기 위해 global_id 변수를 활용하여 [기술통계조사]와 [실태조사_서비스업, 실태조사_제조업]의 중복 데이터만을 추출하여, 업종(line) 변수를 추가하였다. 이러한 과정을 통해 추출된 변수 12개 데이터는 추가 전처리과정(변수 문항 축소) 후 군집분석을 실시하였다. (Data개수 : 제조업834개, 서비스업195개, 총 1029개)



<기술통계조사 자료 _ 데이터 시각화 A1S1 >

- 기술 통계조사 _ 변수 문항 축소

기술 통계조사 자료에서 추출한 변수의 신뢰도를 높이기 위해 12개 변수의 문항을 축소하는 데이터 전처리 과정을 거쳤다. 유사한 항목끼리 묶어서 같은 항목으로 분류하였으며, 결측치는 최빈값으로 대체하였다.

활용 변수	기존 문항	변경 문항 (포함된 기존문항)
기술개발동기 1순위 (A1S1)	1.내수시장 점유율 확대 유지 ... 7.기타	1.시장확대형 (1,2) 2.다양성확대형(3) 3.개선형 (4,5,6) 4.기타
기술개발목적 1순위 (A1N1)	1.시장점유율 확대 ... 15.기타	1.시장확대형 (1,2) 2.다양성확대형(10,11) 3.개선형(4,5,6,8,9,12) 4.인력형 (3,7) 5.기타 (13,14,15)
향후 1년간 중점 투자계획 분야 (C2S2)	신제품 개발 ... 4.기존공정 개선	1.신제품 개발 (1) 2.기존 개선 (2,4) 3.신 공정 개발 (3)
자체기술개발 애로요인1 (H1_1)	1.기술개발 인력확보 곤란 및 잦은 이직 ... 9.기타	1.자금 문제 (2,3) 2.기술 부족 (4,5,6,7) 3.인력 부족 (1) 4.기타 (8,9)
사업화추진 애로요인 (H4_1)	1.사업화 자금부족 ... 10.기타	1.자금 문제 (1, 3) 2.기술 문제 (4) 3.경쟁 문제 (5,6) 4.원료 문제 (2) 5.인력 문제 (7,9) 6.기타 (8,10)
기술개발 지원 필요성 (I4Q1~7)	1.매우 필요함 ... 5.전혀 필요하지 않음	1. 필요성을 거의 느끼지 않음(3,4,5) 2. 보통 (2) 3. 필요성을 느낌 (1)

- 실태조사 (제조업, 서비스업)

군집분석 결과별로 분류된 중소기업 중 지원 정책을 실행하기 우수한 기업을 찾고자 실태조사 자료를 활용하였다. 우수한 기업의 기준으로 '2017년 대비 2018년의 영업 손익 증가율'을 종속변수로 설정하였다. 실태조사 자료 손익계산서의 2017년 기준 33개 항목 중 19개를 변수를 추출하였다. 변수 제거 기준은 1)항목끼리 연산되어 구해진 변수, 2)종속변수와 높은 공선성을 가진 영업 손익 변수, 3)결측치가 75%이상인 변수이다. 위의 기준으로 총 14개의 항목을 제거, 19개의 항목을 예측분석의 독립변수로 설정하였다. 또한 결측치는 평균으로 대체하였다. 그 후, 모델 설명력 확인을 위해 실태조사 데이터를 Train_set 과 Test_set 으로 나누었다. 이 때 데이터의 손실을 줄이고자 기술통계 자료와 global_id가 중복되지 않는 데이터는 Train_set 으로 설정하고, 중복되는 데이터는 Test_set 으로 설정하였다. 따라서 이러한 전처리를 거친 19개의 독립변수와 1개의 종속변수를 활용하여 기업의 손익증가 여부를 예측하였다.

(*2017년 대비 2018년의 영업 손익 증가율 계산식
= (INC_A_21 – INC_A_61) / (INC_A_61))

■ 분석 기법 및 분석 결과

(1) Clustering Analysis (군집분석)[4]

1-1) 데이터 변환

- 전처리를 완료한 기술통계 자료를 효율적으로 활용하기 위해 다양한 데이터 변환 과정을 진행하였다.
- Original : 데이터 전처리 과정을 거친 기본 데이터
- Dummy : 기본 데이터의 범주형 변수를 더미변수 처리한 더미 데이터
- PCA : 기본 데이터의 모습을 최대한 유지하며 차원을 축소, 즉 PCA(주성분분석)를 실행 한 차원 축소 데이터
- Dummy + PCA : 더미 변수 처리와 PCA (주성분분석)을 모두 실행 한 이중 변환 데이터

1-2) Clustering 모델[5]

- 기업 특징별 분류를 위해 Scikit-Learn에서 제공하는 다양한 모델을 사용하였다. 그 중 최적의 군집결과를 보인 3가지 알고리즘을 소개한다.

Spectral Clustering (Spectral, 스펙트럼군집화 알고리즘)

- 유사성 매트릭스의 스팩트럼(고유값)을 사용한 데이터 차원 축소 후, 군집화 하는 알고리즘

DBSCAN Clustering (DBSCAN, 밀도기반 군집분석 알고리즘)

- 데이터가 밀집한 정도 즉 밀도기반의 군집화 알고리즘

Hierarchical Agglomerative Clustering (AGG, 합체 군집화 알고리즘)

- 모든 데이터에 군집 생성 후 유사도가 높은 군집을 합쳐가며 군집 개수를 줄여가는 알고리즘

즉, 위 3가지 종류의 군집화 알고리즘을 1) Original, 2) Dummy, 3) PCA, 4) Dummy + PCA 데이터로 각 4번씩 총 12번의 군집화를 진행하였다.

1-3) Silhouette coefficient (실루엣 계수)[6]

- 군집분석 3가지 알고리즘의 결과 중 가장 적절하게 군집화가 이루어진 알고리즘을 채택하기 위해 실루엣 계수를 측정하였다.
- 군집화 성능을 판단을 위한 실루엣 계수는 값이 클수록 좋은 군집화라고 할 수 있다. 같은 군집의 데이터가 다른 군집의 데이터보다 더 유사하다면 실루엣 계수가 커지기 때문이다.

	Original	Dummy	PCA	Dummy + PCA
Spectral	0.636	**0.659**	0.588	0.637
DBSCAN	0.550	0.480	0.609	0.573
AGG	0.601	0.567	0.627	0.542

- 그 결과, Dummy데이터를 활용한 Spectral 군집화모델의 실루엣 계수가 가장 큰 것으로 드러났다. 이는 총 12개의 군집분석 결과 중 가장 적절하게 군집화가 이루어졌다고 볼 수 있다. 따라서 Dummy 변수 처리를 한 Spectral 군집분석을 1029개의 중소기업을 각각의 유사한 군집으로 분류하기 위한 적절한 알고리즘으로 채택하였다.

1-4) 분석 결과[7]

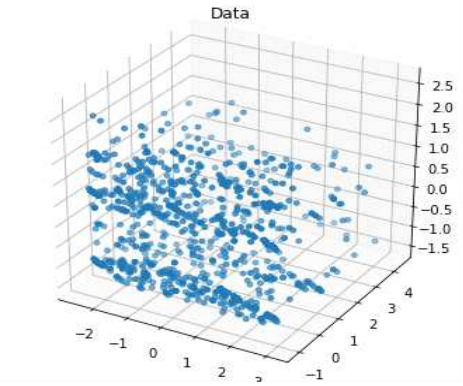
- 아래 <첫번째 이미지>는 기술통계 자료 중 전처리 과정을 마친 1029개의 중소기업을 3차원 시각화로 나타낸 모습이다.
- 아래 <두번째 이미지>는 1029개의 중소기업을 Spectral 군집화를 통해 3개의 군집으로 나눠 3차원 시각화한 모습이다.

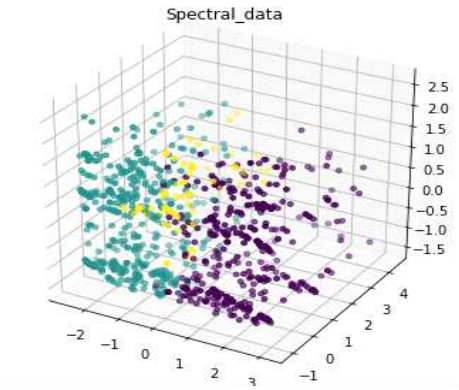
4) 분석도구 : Python의 Pandas. Numpy 패키지
Scikit-learn의 PCA, Preprocessing 패키지 등

5) 분석도구 :
Scikit-learn 라이브러리의 SpectralClustering, DBSCAN, AgglomerativeClustering, NearestNeighbors, Kmeans, GaussianMixture, Meanshift, estimate_bandwith, PCA 등

6) 분석도구 : Scikit-learn 라이브러리의 silhouette_score 등

7) 분석도구 : Python의 matplotlib, Axes3D 패키지 등





군집분석을 통해 형성된 각 군집을 EDA(탐색적 자료분석)를 통해 관찰한 결과는 다음과 같다

- 그룹1(취약형, Weak) : 자금지원과 기술개발에 대한 지원 필요
- 그룹2(안정형, Stable) : 시장확대와 개선에 대한 지원 필요
- 그룹3(도전형, Challenge) : 다양성 확대를 위한 자금 지원 필요

따라서 1) 산업군 2) 기업성장단계 3) 기업의 특징을 활용한 분류 과정은 아래 그림과 같다.

(2) Predictive Analysis (예측 분석)

2-1) Predictive 모델[8)]

산업군, 성장단계 및 기업의 특징으로 분류된 18개의 군집별 우수한 기업 발굴을 위해 기업 성장 예측분석을 하였다. 이에 기업 성장도를 파악할 지표로 '2017년 대비, 2018년의 영업 손익 증가율'을 활용하였다. 예측분석의 종속변수를 영업 손익의 증가여부(0: 증가하지 못함, 1: 증가함)로 설정하였고 이를 예측하기 위한 독립변수로 실태조사 자료에서 전처리를 거친 19개 변수를 채택하였다. 즉, 독립변수들을 활용하여 영업 손익 증가여부를 예측한 것이다. 이에 Scikit-Learn에서 제공하는 다양한 예측모델을 활용하였다. 특히 여러 약한 학습을 순차적으로 학습하며 오차에 가중치를 부여, 오류를 개선하는 '부스팅 알고리즘'에 중점을 두었다. 그 중 최적의 예측 결과를 보인 3가지 알고리즘을 소개한다.

XGBoost Classifier

- 오차를 학습하는 부스팅 계열 중, 트리 기반의 알고리즘 학습에서 우수한 성능을 보이는 알고리즘

Gradient Boosting Classifier (GBM)

- 오차를 학습하는 부스팅 계열 중, 가중치 업데이트를 경사하강법을 이용해 최적의 결과를 얻는 알고리즘

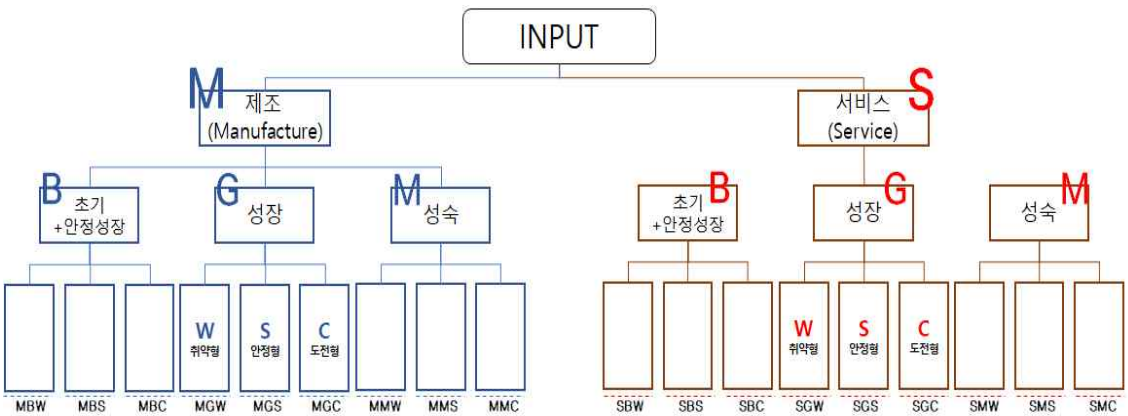
Logistic Regression

- 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 알고리즘

2-2) Accuracy score (정확도)[9)]

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

- 3가지 모델의 예측 결과 중 가장 높은 정확도를 가진 알고리즘을 채택하기 위한 정확도 계산식은 위와 같다.
- 정확도(Accuracy)는 전체 샘플 중 올바르게 예측한 샘플을 'Confusion matrix' 에 나타내 계산한 값이다.



	Accuracy (정확도)
XGBoost Classifier	**0.7318944**
GBM Classifier	0.6211031
Logistic Regression	0.5827998

<Accuracy(정확도)>

	예측 퇴화기업 (N)	예측 성장기업 (P)
실제 퇴화기업 (N)	True Negatives (TN)	False Positives (FP)
실제 성장기업 (P)	False Negatives (FN)	True Positives (TP)

< Confusion Matrix >

- 그 결과, 'XGBoost Classifier'의 정확도가 가장 높은 것으로 드러났다. 이는 중소기업의 성장 여부를 (0: 2017년 대비 2018년에 성장하지 않음, 1:2017년 대비 2018년에 성장함) 약 73%의 확률로 예측함을 의미한다. 따라서 우수 기업을 선별하는 가장 적합한 알고리즘으로 XGBoost Classifier를 채택하였다.

3. 분석 결과 활용

■ 아이디어의 핵심내용

'TED'를 활용하여 중소기업들을 분류한 후, 각 군집별로 우수한 기업들을 선정하는 것이 아이디어의 핵심 내용이다. 이때, 중소기업을 분류하는 기준인 TED는 'Type of industry 산업군', 'Enterprise growth 기업성장단계', 'Data character 기업특성'을 의미한다. 산업군은 제조/비제조로 분류되며, 기업성장단계는 창업초기&안정성장/ 성장/ 성숙 단계로 분류된다. 마지막으로 기업의 특성은 '군집분석(Spectral Clustering)'을 활용하여 취약형(Weak)/ 안정형(Stable)/ 도전형(Challenge)으로 분류하였다. 이렇게 총 18개(2*3*3=18) 그룹으로 분류한 후, '예측분석(XGBoost)'을 실시하여 각 군집별로 우수 기업을 선정한다. 즉, 각 기업의 성장여부를 예측하여 지원 정책을 실행하기에 우수한 기업들을 선별하는 것이다.

8) 분석도구 : 다음페이지 각주 참고
8) 분석도구
Scikit-learn 라이브러리의 LogisticRegression, Ridge, Lasso, ElasticNet, GradientBoostingClassifier, LGBMClassifier, RandomForestClassifer 패키지
Xgboost 라이브러리의 XGBClassifier 패키지 등
9) 분석도구
Scikit-learn 라이브러리의 confusion_matrix, accuracy_score, f1_score

■ 기존 정책과의 차이

맞춤 지원

- '자금지원'에 초점이 맞춰져있던 기존 정책과는 달리 'TED'는 각 기업의 특징과 요구사항을 반영한 18개의 그룹으로 나누어 맞춤 지원 정책을 시행할 수 있다.

지속가능성

- 'TED'를 활용한 기업 성장여부 예측 모형은 기존의 단기 지향적이고 즉흥적인 정책에서 벗어나, 기업의 지속 가능성 평가를 통해 기업을 선정, 지원한다. 따라서 미래 성장 가능성이 높은 기업들을 발굴해내는 데 도움을 줄 수 있다.

정확한 기준 정의

- 기존에는 정부의 중소기업 분류 및 선별 기준이 모호하지만, TED를 활용한 기업 선정은 데이터를 분석을 활용하였기에 정확한 기준 정의가 가능하다.

■ 실현가능성

해당 아이디어의 핵심 알고리즘인 군집분석(Spectral model)의 실루엣 계수는 0.659, 예측분석(XGboost)의 정확도는 0.73이다. 이는 'TED'모형을 신뢰할 충분한 수치이다. 또한 모든 과정은 공모전 내 제공된 데이터를 통해 진행된 것이므로, 같은 형식의 더욱 많은 데이터가 수집된다면 아이디어의 실현가능성이 더욱 높아질 것이다.

최근 제주시의 '중소기업 빅데이터 지원 사업'이 일자리 창출과 매출 증대에 효과를 보이고 있다. 해당 사업은 도 내 중소기업들의 기업 내부 데이터 및 사회 데이터 분석을 통해 기업 혁신과 새로운 비즈니스 창출을 지원한다. 이처럼 빅데이터 분석 기반의 기업 지원 사업은 현실적으로 도움주고 있으며, 본 문서의 TED모델 실현 가능성을 기대한다.

■ 기대효과

자원 낭비 절감

- 중소기업의 무조건적인 자금지원의 요구에서 벗어나, 데이터 분석을 바탕으로 기업에게 실질적으로 필요한 정책을 파악할 수 있다. 이를 통해 과거 문제로 다루어졌던 '퍼주기식 지원'으로 인한 자원 낭비를 막을 수 있다.

효율성

- 기업의 성장여부를 예측하는 예측모형을 통해 기존의 까다로웠던 기업 심사 절차를 축소시켜, 빠르고 정확한 기업 선정에 도움을 줄 수 있다.

적시성 효과

- 데이터를 활용하여 중소기업의 특성과 요구사항을 찾아내고 성장 가능성을 예측함으로써, 기업에 대한 실시간 평가가 가능해지므로 적시성 효과를 기대한다.