Journal of the Korean Institute of Industrial Engineers
Vol. 51, No. 1, pp. 95-107, February 2025.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2025.51.1.095

머신러닝 기반의 방산육성지원 수혜기업 예측모형 개발

전고운[1] · 유동희[2] · 전정환[3†]

[1]경상국립대학교 기술경영학과 · 국방기술진흥연구소 / [2]경상국립대학교 경영정보학과
[3]경상국립대학교 산업시스템공학부

Development of a Funded Defense Companies Prediction Model based on Machine Learning

G. W. Jeon[1] · D. H. Yoo[2] · J. H. Jeon[3]

[1]Department of Management of Technology, Gyeongsang National University · KRIT
[2]Department of Management Information Systems (Bus & Econ Res Inst.), Gyeongsang National University
[3]Department of Industrial System Engineering, Gyeongsang National University

This research focused on the needs of the defense industry to understand what companies should preemptively prepare from the perspective of their current management status in order to increase the probability of receiving benefits from being selected for defense industry development support projects. An experiment was conducted to build a prediction model for beneficiaries of the parts localization development support project and the weapon system modification development support project using corporate information. To compensate for the imbalance problem of data classes of variables, random sampling methods such as oversampling, undersampling, and hybrid sampling methods were applied. The backward elimination technique was applied as a variable selection technique, and the ensemble technique was additionally applied to improve the performance of the prediction model. What differentiates this study is that it analyzes corporate characteristics that determine beneficiary and non-beneficiary companies and builds a prediction model for beneficiary companies.

Keywords: Defense Industry, Government Support Project, Machine Learning, Prediction Model

1. 서 론

방위산업은 국가수호 및 영토방위에서 필요한 군사력 건설을 지탱하는 산업으로, 국가가 자국안보를 보장하고 영토를 수호하는 국가적 전략 목표하에 군을 운영하는 데 소요되는 물자들을 연구개발 및 생산 · 공급하는 모든 활동이 산업을 구성한다. 2023년 방위사업청이 발표한 '23-27 방위산업발전 기본계획'에는 우리 정부가 2027년까지 달성할 목표로서 '국방과학기술 5대 강국' 도약과 '글로벌 4대 방산수출 국가' 진입이 명시되어 있다.

정부의 방위산업 육성 지원사업은 방위산업에 종사하는 기업들의 성장단계별 맞춤형 지원을 목표로 하며, 매년 지원예산과 지원기업의 수를 늘리고 있다. 최근 전례 없는 방산수출 호조세에 힘입어 기존에 민수산업에만 종사하던 일반업체들 또한 방위산업에 진입하기를 희망하고 정부의 방산육성 지원사업에 참여하길 희망한다. 정부가 보조금의 형태로 이러한 기업을 지원하는 궁극적인 목표는 기업의 역량을 증진시켜 기업을 성장시키고 나아가서는 국가 경제발전의 토대인 우량기업을 육성하고자 함에 있다(Oh *et al.*, 2020). 그러나 정부의 방산육성 지원사업에 참여하고 싶은 기업의 입장에서는 기업선정단계에서 육성이 필요한 기업이 아닌 이미 우량한 기업을 지원대

이 연구는 2022년도 경상국립대학교 발전기금재단 재원으로 수행되었음.
† 연락저자 : 전정환 교수, 52828 경상남도 진주시 가좌길29번길 63 국립경상대학교 산업시스템공학부, Tel : 055-772-1704, Fax : 055-772-1699, E-mail : jhjeon@gnu.ac.kr
2024년 8월 13일 접수; 2024년 9월 8일 수정본 접수; 2024년 9월 11일 게재 확정.

상으로 선정하는 것이 아닌가 하는 우려와 아쉬움을 가질 수 있다. 2024년 5월에 '2024년도 방위산업 실태조사'의 일환으로 한국방위산업진흥회가 조사한 '잠재적 방산기업 지원정책 수립연구' 설문 결과, 수혜기업으로 최종 선정되지 못한 기업 중 일부가 실제로 이러한 아쉬움을 표현한 바 있다.("기 수혜기업만 계속 선정되는 것 같다" 등) 그렇지만 탈락한 대부분의 기업이 앞으로도 지원사업을 계속 신청할 의사가 있다고 응답하여, 사후에라도 선정기준에 미달한 역량분야가 무엇인지 식별할 수만 있다면 이를 보완하여 도전할 기업이 다수임을 확인하였다. 정부지원사업의 대상기업으로 선정하는 과정에서는 지원사업의 목적에 맞는 선정기준이 있고, 선정계획 공고 당시부터 공표되어 있다. 그러나 지원받고 싶은 기업의 입장에서는 제안서의 심사결과 외에도 수혜기업들이 공통적으로 갖는 객관적인 경영상태의 특징을 알고 싶고, 기업의 특징 중에서도 무엇이 중요한 결정요인인지 아는 것이 선정 단계의 경쟁력을 높이는 데 도움이 된다. 이러한 정보를 토대로 사전에 지원사업 선정 가능성을 높이는 노력을 할 수 있기 때문이다.

그러나 정부지원사업 선정단계에서 탈락된 기업에게 탈락의 원인이나 평가결과의 세부내용을 공개하지 않는 것이 일반적이며, 선정된 기업과 탈락된 기업의 목록정보 조차도 접근할 수 없다. 현재까지 두 그룹간의 두드러지는 기업 경영상태를 비교 분석하기 위한 정보가 제한적이었기 때문이다. 정보의 비대칭성에서 기인한 한계를 지적하고자 Kim *et al.*(2018)은 민간 벤처기업을 대상으로 한 정부지원사업 연구에서는 수혜기업 선정 시 정성적인 기술평가 외 재무성과와 같은 경영현황 평가비중을 높여야 한다는 결론을 주장했다. 그러나 지원사업 선정 여부를 예측하는 수혜기업들의 경영상태 공통점 도출 연구는 부족한 실정이다.

본 연구에서는 앞서 언급한 문제점을 해결하고 근거기반(evidence-based) 실증분석과 기업전략 도출을 위해 머신러닝 기반 예측모형을 설계하고자 한다. 구체적으로는 국방기술진흥연구소에서 보유하고 있는 수혜기업 목록과 방위산업 실태조사의 기업 경영현황 등의 방대한 정보를 활용하여 방위산업육성 지원사업 수혜기업 예측모형을 구축하고 한다. 이를 토대로 수혜기업을 결정하는 주요 요인을 식별하고, 방산업체 관점에서 정부지원사업 수혜확률을 높일 수 있는 전략을 제안하고자 한다.

본 연구에서는 머신러닝 분야에서 널리 사용되는 다양한 알고리즘을 학습에 활용하여 방산육성 지원사업 수혜기업 예측모형을 구축하는 연구를 진행하고자 한다. 방산육성 지원사업의 종류가 다양하고 목적별 지원규모도 상이하기 때문에, 본 연구에서는 '24년 정부지원금 예산 배정이 가장 큰 부품국산화개발 지원사업과 무기체계 개조개발 지원사업으로 한정하여 데이터셋을 구성하였다. 전처리 단계에서는 방산업체를 각 지원사업의 수혜기업과 비수혜기업으로 구분하여 사례 비율을 동일하게 구성하기 위해 3가지 데이터 균형화 방법을 적용했다. 기업의 경영현황 중 수혜기업을 결정하는 중요도 높은 변수로 수정된 예측모형을 구축하기 위해 후진제거 기법의 변수선택 과정을 거쳤다. 그 후 5가지의 머신러닝 알고리즘을 적용하여 우수한 성능을 보이는 데이터셋 샘플링 방법과 알고리즘을 찾고, 수혜기업 예측모형의 성능을 한 단계 더 개선하는 실험을 진행하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 이론적 배경과 선행연구를 살펴보고, 제3장에서는 본 연구에서 제시하는 연구모형에 대해 간략히 기술한다. 제4장에서는 예측모형 구축 결과와 적중률 개선 시도의 결과를 서술하며 방산업체들의 정부지원사업 수혜확률을 향상시킬 수 있는 전략을 제안한다. 마지막으로 제5장에서는 연구의 결과와 시사점, 한계점을 제시한다.

2. 이론적 배경 및 선행연구

2.1 방위산업 육성 정부지원사업에 대한 선행연구

2027년까지 '국방과학기술 5대 강국' 도약과 '글로벌 4대 방

Table 1. Government Support Project List

No.	Government support project		2024 Budget (Billion ₩)
1	Program for components localization		123.7
2	Defense Innovation Technology Leaders-100		39.0
3	Program for developing Global defense industry small but strong enterprise		14.6
4	Program for defense venturing		11.0
5	Defense professional manpower training project		1.7
6	Program for providing consulting services to SME		0.7
7	Defense Venture business incubation		0.2
8	Regional basis	Defense innovation cluster project	12.5
9		Support to increase productivity	2.3
10	Export related	Program for supporting weapons modification	53.6
11		Management of foreign certification/export consulting programs	1.3
Budget Sum			260.6

산수출 국가' 진입이라는 목표를 달성하기 위한 추진과제의 일환으로 정부는 방위산업에 종사하는 기업들의 성장과 수출 경쟁력 강화를 위해 전년 대비 높은 사업예산을 배정하였다. <Table 1>은 한국 정부가 국내 방위산업을 육성하고 업체들의 글로벌 경쟁력을 강화하기 위해 추진하는 2024년 정부지원 사업의 목록과 배정 예산 현황이다.

2024년 2,600억 원이라는 막대한 재정투입이 이루어지고 있으나, 방위산업 분야를 육성시키기 위한 정부지원사업에 관한 기존 연구들은 대부분 지원 후의 효과나 성과를 분석하는 데 중점을 두고 있다.

방위산업 분야 정부지원사업에 대한 선행연구로서, 먼저 Rho(2021)는 방위산업 중소기업들의 경영성과 제고에 영향을 미칠 수 있는 결정요인인 내재적 핵심역량, 경쟁전략, 경영성과 간의 인과관계과 상관관계를 살펴보았다. 그리고 원청업체와의 공급사슬 파트너십과 정부지원제도의 활용도가 이들 변수 간 어떤 조절효과를 나타내는지 통계적 분석을 수행하였다. Lee *et al.*(2020)은 현재의 국내 방위산업 실태와 미비점을 진단하고, 방위산업에 대한 주요 선진국 분석을 통해 정책적 발전방안을 모색하였다. Kong *et al.*(2020)은 산업혁신 관점에서 정부지원이 방위산업 종사 기업의 성과에 미치는 영향을 살펴보기 위하여 국내 방산업체 자료를 바탕으로 이중차분법(DID)을 활용한 실증분석을 수행하였다. Choi *et al.*(2018)은 의사결정 방법론 중 하나인 자료포락분석기법(DEA)을 활용하여 국내 방산업체들의 경영 효율성을 측정하였다. 투입 변수로 품질부서 종업원수, R&D 현황을 설정하였고 산출 변수로는 매출액과 영업이익을 설정하여 CCR 모형 및 BCC 모형을 각각 사용하였다. Lim *et al.*(2019)은 국방핵심기술 연구를 통해 도출되는 연구성과와 이를 활용하여 실제 무기체계 적용으로 창출되는 실질성과의 효율성을 분석하기 위해 DEA 모형을 활용하였다. Hwang *et al.*(2022)은 국방핵심기술 연구개발사업의 성과평가 요인을 도출하기 위하여 탐색적 요인분석 및 신뢰도분석(PCA/FA)을 활용하였다.

방위산업이 아닌 타 산업과 관련한 정부지원사업 연구 현황으로서, Lee *et al.*(2024)은 스마트공장 보급・확산 정부지원사업에 참여한 16,325개 기업을 대상으로 국내 GDP 성장에 어느 정도 기여하였는지, 그리고 산업연관분석을 적용하여 스마트공장 도입기업의 생산유발효과, 부가가치 유발효과, 고용 유발효과를 확인하였다. Yoon *et al.*(2020)은 기업 성장지원센터 사업의 집중컨설팅 지원을 받은 수혜기업을 대상으로 효율적 집단과 비효율적 집단을 구분하여 투입변수의 효과를 분석하기 위해 자료포락분석을 시행하였다. Kang *et al.*(2018)은 정부지원사업에 참여한 중소기업들을 대상으로 기업의 성장성, 수익성, 생산성, 혁신성, 안정성의 5대 영역에 적합한 성과지표를 선정하여 정부에서 지원하는 자동화 및 스마트화 지원사업의 수혜 여부에 따른 중소기업의 기업성과 영향 분석과 성향점수 매칭 방법을 이용한 실증분석을 실시하였다. Oh *et al.*(2020)은 과학기술정책연구원(STEPI)에서 매년 수행하고 있는 기업혁신조사와 KISTEP의 국가연구개발사업 조사분석데이터를 활용하여 기업에 대전 정부의 R&D 지원이 제조기업의 혁신활동과 혁신성과에 미치는 영향을 성향점수매칭법을

Table 2. Previous Studies on the Government Support Projects

Industry	Researcher	Methodology	Research Topic
Defense Industry	Rho (2021)	Literature research, Statistical analysis	Analysis of correlations between factors of core competencies, government support systems that affect the management performance of defense companies
	Lee *et al.* (2020)	Literature research	Derive policy development measures applicable to Korea through research on the defense industry policies of major developed countries
	Kong *et al.* (2020)	Difference-in-Difference	Analyzing the impact of government support on defense companies and industrial development from the perspective of industrial innovation
	Choi *et al.* (2018)	Data Envelopment Analysis	Analysis of management efficiency targeting major defense companies with more than 30 employees
	Lim *et al.* (2019)	Data Envelopment Analysis	Discussion of implications after analyzing the efficiency of each stage of the defense core technology project
	Hwang *et al.* (2022)	PCA/FA	Derive performance evaluation factors for defense core technology research and development projects
Other Industries	Lee *et al.* (2024)	Industry correlation analysis	Analysis of production inducement coefficient and inter-industry correlation for companies that introduced smart factories through government support
	Yoon *et al.* (2020)	Data Envelopment Analysis	Analysis of efficiency of small and medium-sized consulting support projects and presentation of optimal input elements
	Kang *et al.* (2018)	Statistical analysis	Analysis of the impact of receiving government support projects on the business performance of small and medium-sized businesses
	Oh *et al.* (2020)	Propensity Score Matching	Analyzing the impact of government R&D support on the innovation activities and innovation performance of manufacturing companies

통해 분석하였다. 그 결과 정부 R&D 지원이 수혜기업의 혁신활동을 유인하는 것으로 분석되었고, 제품혁신성과 또한 높게 발생하는 것으로 나타났다.

지금까지 설명한 정부지원사업에 관한 기존 연구들의 분석방법과 연구내용을 요약하면 <Table 2>와 같다.

2.2 머신러닝 기반 예측에 대한 선행연구

머신러닝은 다양한 연구 분야에서 예측력 및 속도가 뛰어나 활용되고 있다. 특히, 머신러닝은 여러개의 변수 데이터를 대상으로 하는 예측분석에서 탁월한 성과를 내어 활발히 사용되고 있다.

먼저 국방분야 및 방산 분야에서는 Lim *et al.*(2021)가 폐렴, 결핵, 횡문근융해 증 등 3가지 주요 국방 의료 질병을 진단하기 위해 국방의료 데이터에 딥러닝 기술을 적용하였다. 건강검진, 신체검사 등의 검진 정보와 특정 질병간의 인공신경망(Artificial Neural Network, ANN)을 구성하여 병력감소 및 군의료진 부족현상에 대비한 무증상 환자 조기발견 및 선제 대응을 목표로 연구하였다. Lee *et al.*(2020)은 저장탄약 신뢰성 평가 데이터 특성을 고려하여 입력 변수를 줄이는 정규화 기법으로 제안하였으며, 이를 통해 저장탄약 신뢰성 평가를 위한 인공신경망 모델의 학습 능률을 향상시키는 연구를 수행하였다. Han *et al.*(2011)은 인공신경망 모형을 이용하여 국방조달에 있어서 부정당업자 발생 예측을 분석하였다. 15개의 입력변수 중 13개의 변수가 입력변수로 도출되었고, 학습용 데이터를 통한 분석결과 96.61%, 검증용 데이터를 통한 분석결과 92.5%의 높은 예측력을 나타냈다.

국방·방산분야가 아닌 일반 분야에서는 Oh *et al.*(2017)가 사회보장 빅데이터 분석을 실시하였으며, 구체적으로는 의료정보, 생활여건 등의 변수를 활용한 보건복지 정책에 대한 영향 분석에 머신러닝을 활용하였다. 로지스틱 회귀분석(Logistic Regression, LR), 의사결정나무(Decision Tree, DT), 랜덤포레스트(Random Forest, RF), 부스팅, SVM을 적용하여 예측력이 가장 우수한 모형을 비교분석 하였다. Choo *et al.*(2023)는 부산, 가덕도, 거제도 이상조위 발생여부를 예측하기 위해 로지스틱 회귀분석을 활용하여 데이터를 분석하였다. Choi *et al.*(2023)은 의사결정나무 기법을 활용하여 초중고 학생의 사교육과 학업성적에 대한 패턴 규칙을 추정하였다. 국가통계포털에서 제공하는 초등·중등·고등학생 사교육비 조사 데이터를 활용하여, 학업성적에 따라 3개의 클래스(상·중·하위권)로 분류하였으며 총 58개의 변수를 사용하였다. 이외에도 Kim *et al.*(2022), Lee(2023), Joo *et al.*(2023)가 해당분야의 문제를 미리 감지하고 조치를 취하는 사전 예방전략을 지원하기 위해 머신러닝 기반의 예측 모델을 제안하였다.

연구에서 활용된 머신러닝 알고리즘별 선행연구를 <Table 3>으로 정리하였다.

머신러닝 기반의 예측모형을 구축할 때, 데이터의 학습 전 불균형 문제로 인한 예측편향을 완화시키기 위한 데이터 균형화 과정도 자세히 작성한 선행연구가 존재하였다.

Kim *et al.*(2016)은 사회경제를 확장기와 수축기로 구분하여 KOSPI, KOSDAQ에 상장된 제조업 및 비제조업 1,691개의 기업 재무 데이터에 의사결정나무, 인공신경망, 앙상블 알고리즘을 적용하여 부실기업화 예측모형을 개발하였다. 이때 오버샘플링 기법(SMOTE)을 활용하여 데이터 불균형 문제를 해결하였으며

Table 3. Previous Studies on the Machine Learning Prediction Model

Algorithm	Researcher	Research Topic
ANN	Lim *et al.* (2021)	Diagnose, classify, and predict major diseases using Ministry of Defense physical examination information
ANN	Lee *et al.* (2020)	Proposal to build a human neural network model and improve learning speed based on storage ammunition reliability evaluation data
ANN	Han *et al.* (2011)	Proposal of a prediction model for early detection of the possibility of fraudulent contractors when defense companies participate in procurement bidding.
LR, DT, RF, Boosting, SVM	Oh *et al.* (2017)	Discover and confirm welfare recipients in welfare blind spots through simulation analysis of welfare supply and demand forecasts
LR	Choo *et al.* (2023)	Proposal of a method for distinguishing between normal and abnormal tide levels and predicting damage through a tide level prediction function
Decision Tree	Choi *et al.* (2023)	Propose evidence for educational fields and policies by analyzing differences between school and academic levels
BN	Kim *et al.* (2022)	Identify areas with high sensitivity to drought influencing factors and propose measures to prepare for drought
LR, SVM, ANN, DT, Ensemble	Lee(2023)	Predictive analysis of limitations of non-performing hotels and classification of sustainable companies
ANN	Joo *et al.*(2023)	Proposal of sewage water quality prediction model based on water quality data

Table 4. Previous studies on the data balancing

Data balancing	Researcher	Variable	Algorithm	Data set	prediction Rate	improvement
Over-sampling (SMOTE)	Kim *et al.* (2016)	78	Boosting (AdaBoost-J48)	2216 (86 : 2,130)	96%	10%
	Hwang (2022)	39	Light GBM	100 (59 : 41)	80%	50%
Over-sampling (DCGAN)	Yoo *et al.* (2022)	7	DNN	1,262,426 (Undisclosed)	98%	24%
Over-sampling + Under-sampling (SMOTE+TOMEK)	Batista *et al.* (2003)	-	DT	120,000 (Undisclosed)	91%	4%
Over-sampling + Under-sampling (SMOTE+ENN)	Batista *et al.* (2004)	-	DT	13 (Undisclosed)	76%	19%

표본 비율을 1:1로 변화시켜 최종 78개의 변수를 구성하여 부실기업화 예측 모델을 개발하는 실험을 수행하였다. Hwang(2022)과 Yoo *et al.*(2022) 또한 예측모형에서 데이터 불균형을 해소하는 방법으로 오버샘플링 기법을 채택하여 예측성능을 향상시킨 연구모형을 제안하였다. 한편 Batista *et al.* (2003)은 DNA결합 키워드와 비결합 키워드 예측에 대한 훈련데이터의 불균형을 해결하는 방법으로 언더샘플링기법과 오버샘플링기법 중 불균형 해소율이 더 높은 복합샘플링 기법에 대한 결과값을 제시하였다. 그러나 데이터 균형화 적용시 변수값의 증가에 따라 결과가 변경될 수 있고, 다른 불균형 해소 기법을 사용 시 다른 결과값이 도출될 수 있어 이후 연구를 통해 비교 연구가 필요하다는 시사점을 남겼다. Batista *et al.*(2004) 연구에서는 데이터 불균형을 해소하는 가장 좋은 성능을 보이는 복합샘플링기법을 도출하였다. 실험 결과 랜덤 오버샘플링기법보다 SMOTE+TOMEK 및 SMOTE+ENN 기법이 적은수의 데이터 셋에서 높은 예측률을 보이는 것을 확인하였다.

예측모형의 성능향상을 위한 데이터 불균형 완화에 중점을 둔 선행연구와 데이터 균형화 기법을 <Table 4>로 정리하였다.

3. 연구방법

3.1 연구 데이터

먼저 연구에 활용할 데이터 수집 대상기업 범위는 「방위사업법」 제35조(방산업체의 지정)에 따라 무기체계와 주요 구성품을 생산하는 지정된 방산업체 85개사(2023년 기준)로 한정한다. 방위산업은 국가 안보에 직결되는 산업이기 때문에 시장경제원리에 의해 조달이 곤란한 군수물자의 안정적인 조달원 확보와 엄격한 품질보증을 위해 방산물자와 그 물자를 생산하는 방산업체 지정제도를 운영하고 있다. 방산업체는 매년 산업통산자원부 장관이 방위사업청장과 협의하여 지정하며, 정기적으로 재무・매출・고용 등의 경영현황에 대해 응답하고 있다.

방산업체의 경영상태 데이터는 방위산업 실태조사(Fact-finding survey on Defense Industry)에 응답한 항목을 기준으로 수집한다. 방위산업 실태조사는 방위산업발전 기본계획 수립을 위해 방위사업청에서 매년 집계하며, 조사항목에는 기업의 일반현황, 재무제표, 매출 및 생산, 수출・입 및 연구개발 현황, 고용현황, 지식재산권 등이 포함되어 있다.

3.2 연구 설계

이후 전처리(preprocessing) 단계에서 개별 방산업체들의 2022년 말까지 부품국산화개발 지원사업과 무기체계 개조개발 지원사업 수혜 이력을 조회하여 지원사업별 새로운 데이터 셋을 구성한다. 지원사업 수혜 이력은 국방기술진흥연구소의 과제관리 데이터를 참조하고, 사업별 수혜 및 비수혜기업 그룹을 표기한 데이터 셋을 각 1개씩 총 2개 생성한다.

이때 정부지원사업 수혜기업과 비수혜기업 샘플에서 발생되는 데이터의 불균형 문제를 해결하기 위해 데이터 균형화(Data Balancing) 단계로서 오버샘플링 3회씩, 언더샘플링 3회씩을 실시하였다. 그리고 오버샘플링과 언더샘플링을 결합한 하이브리드 방식의 샘플링을 3회씩 실시하여 최종적으로 18개의 데이터셋을 만든다.

각 데이터셋의 변수 데이터 값을 Min-Max 정규화(normalization)로 최소 0, 최대 1로 설정한 후 변수 중요도를 기준으로 후진제거법(Backward Elimination)을 적용하여 변수를 선택(feature selection)하고, 예측모형 개발을 위해 대표적인 머신러닝 알고리즘 5가지를 적용한다. 로지스틱 회귀분석(Logistic regression), 의사결정나무(Decision tree), 베이지안 네트워크(Bayesian network), 인공신경망(Artificial neural network), 최근접이웃(K-nearest neighbor)을 활용하여 총 90개의 수혜기업 예측모형을 만들고 적중률이 가장 높은 샘플링 방법과 알고리즘을 알아본다.

마지막으로 예측모형의 적중률을 개선시키는 방법으로서 앙상블 기법인 배깅과 부스팅을 적용하여, 적중률에 양(+)의 변화가 있는지 살펴본다. 그리고 수혜기업 예측모형 변수를 기반으로 한 방산업체의 정부지원사업 수혜확률 향상 방안을 제언한다.

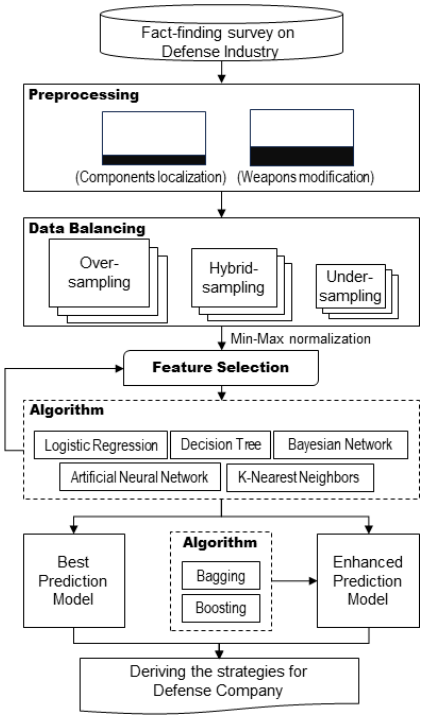

Figure 1. Research Process

본 연구에서 제안하는 분석 프로세스는 데이터 전처리(Preprocessing), 데이터 균형화(Data Balancing), 변수 선정(Feature Selection), 머신러닝 알고리즘 적용을 통한 예측모형 구축과 적중률 개선 단계로 구분할 수 있고, <Figure 1>과 같이 도식화하였다.

4. 예측모형 구축 결과

4.1 데이터 전처리

본 연구에서는 산업통상자원부가 매년 고시하는 '방위산업체 지정 목록'의 2022년 말 기준 방산업체 85개사를 연구대상 기업으로 한정하였다. 각 기업이 부품국산화개발 지원사업 및 무기체계 개조개발 지원사업이 시작된 최초시점(2010년, 2015년)부터 2022년 12월 말일까지의 사업참여를 위한 협약서 작성 여부를 기준으로 수혜기업과 비수혜기업 그룹으로 구분하였다. 그리고 '방위산업 실태조사'라는 방위산업 분야 종사기업의 경영통계 자료를 결합하여 이종간 데이터를 통합한 지원사업별 데이터셋을 구성했다.

Table 5. (Tentative) Independent Variables and Basic Statistics

Division	Category	Variable	Result	Average	Standard Deviation
Target variable	government support	Companies that received support	{Yes, No}	-	-
Independent variable	Company	Size	{S, M, L}	-	-
		Weapon Items Produced	numeric data	2.8	1.3
		Number of research institutes		0.4	0.6
		Number of factories		0.5	0.9
	R&D/facility	R&D Government-invested		2,390,241.3	14,524,594.0
		R&D Self-invested		7,703,306.5	36,516,087.5
		Facility Government-invested		590,201.4	3,666,674.0
		Facility Self-invested		4,282,279.4	12,951,470.4
	Employee	employee		1,412.3	4,211.1
		researcher		93.5	297.3
		blue collar		162.1	316.6
		Overseas worker		34.1	214.6
		Bachelor's degree		42.8	135.0
		Master's degree		39.2	128.6
		Ph.D.		6.9	22.4 6
	Management indicators (Defense sector)	Total assets		690,440,486.4	1,756,384,153.3
		Gross profit		22,318,618.3	54,551,066.0
		Net Income		8,576,196.8	24,724,386.8
		Net income to sales		0.0	0.1
		Sales		186,825,287.0	411,117,316.7
		Import amount		22,542,112.1	61,170,801.1
		Export amount		19,191,169.8	64,474,499.2
	performance (non-economic)	Domestic Patents/IPR		71.9	264.7
		Overseas Patents/IPR		2.2	9.7

산업통상부와 방위사업청은 ‘방위산업물자 및 방위산업체 지정 규정’에 따라 방산물자를 생산하는 업체를 방산업체로 지정 공개하고 있다. 이 업체들은 방위산업 발전법(방위산업 발전 및 지원에 관한 법률) 제6조에 따라 방위산업 실태조사에서 기업의 인적 자원, 설비투자・기술수준 및 연구개발에 관한 사항, 생산・매출・수출입에 관한 사항을 조사하고 있다. 본 연구에서는 방산업체의 지원사업 수혜여부 값(Yes, No)을 목표변수로 사용하고, 이를 예측하기 위해 관련성이 있는 24개의 조사항목을 후보 독립변수로 사용하였다.

후보 변수는 업체특징과 관계된 ‘회사규모’, ‘무기체계’, ‘방산전담 연구소 보유수’, ‘방산전담 공장 보유수’, R&D와 시설부문 투자와 관계된 ‘방산부문 정부지원 R&D 투자금(천원)’, ‘방산부문 자체 R&D 투자금(천원)’, ‘방산부문 정부지원 설비투자금(천원)’, ‘방산부문 자체 설비투자금(천원)’, 업체 인력과 관계된 ‘종업원수’, ‘연구원수’, ‘생산직수’, ‘해외지사 근무자수’, ‘방산부문 학사 졸업자수’, ‘방산부문 석사 졸업자수’, ‘방산부문 박사 졸업자수’, 업체 경영지표인 ‘방산부문 자산총계(천원)’, ‘방산부문 매출총이익(천원)’, ‘방산부문 당기순이익(천원)’, ‘매출액 대비 당기순이익(천원)’, ‘방산매출액(천원)’, ‘방산부문 수입액(천원)’, ‘방산부문 수출액(천원)’, 비경제적 성과와 관계된 ‘국내 특허/지재권수’, ‘해외 특허/지재권수’ 등이다. 후보 변수들의 데이터 유형은 대부분 수치형 데이터로, 기초통계는 아래 <Table 5>와 같다.

4.2 데이터 균형화

방산업체 85개사 목록 중 방위사업청 산하 공공기관인 국방기술진흥연구소의 ‘부품국산화 개발지원사업’, ‘무기체계 개조개발 지원사업’에 2022년 12월까지 참여한 이력이 있는 수혜기업은 각각 21개사(24.7%), 25개사(29.4%)로 식별하였다. 머신러닝 과정에서 학습데이터를 통해 예측모형을 구축할 때는 목표변수의 클래스 분포가 5:5로 균등하여야 알고리즘 과적합을 피할 수 있다. 그렇지 않으면 학습과정에서 클래스가 많은 비수혜기업 위주로 학습이 이루어지기 때문에 특정 클래스만 우수하게 예측하는 편향된 예측모형 구축으로 이어질 수 있기 때문이다. 목표변수에서 수혜기업(Yes)의 표본 개수가 비수혜기업(No)에 비해 현저히 적기 때문에, 이러한 과적합을 피하기 위해 지원사업별로 데이터 균형화 작업을 수행하였다.

본 연구에서는 데이터 균형화 기법으로 널리 쓰이는 오버샘플링과 언더샘플링, 그리고 널리 쓰이는 방법은 아니지만 오버샘플링과 언더샘플링을 혼합한 방식을 통해 데이터의 균형을 맞추었다. 오버샘플링은 다수인 클래스 샘플의 수에 맞추어 소수인 클래스 샘플의 수를 랜덤으로 증가시키는 방법이고, 언더샘플링은 소수 클래스의 샘플 크기에 맞추어 다수 클래스의 샘플 크기를 줄이는 방법을 의미한다(Mathew *et al.*, 2018). 본 연구에서는 데이터의 클래스가 85개라는 소수인 점을 감안하여 오버샘플링, 언더샘플링, 그리고 이를 혼합하여 클래스 수를 중간값으로 설정한 샘플링 방식을 하이브리드샘플링으로 명명하고 3가지 랜덤샘플링 방법을 연구에 적용해 보았다.

랜덤샘플링 과정에서 발생할 수 있는 오류를 줄이기 위해 각 샘플링 방법을 3회 실시하여 총 19개의 데이터셋을 분석을 위해 구성하였다. 오버샘플링된 부품국산화(128개) 및 개조개발(120개), 하이브리드샘플링된 부품국산화(84개) 및 개조개발(84개), 언더샘플링된 부품국산화(42개) 및 개조개발(50개) 데이터셋 19개를 <Table 6>으로 나타내었다.

4.3 변수 선정

정규화 등 전처리 후 선택된 독립변수 중에서 예측모형 구축에 도움을 주지 않는 변수는 분류의 정확성을 낮출 수 있기 때문에 제거한 후에 예측모형을 구축한다.(Dash *et al.*, 1997) 이처럼 전처리 단계에서 독립변수 중 목표변수 분류에 필요없는 변수를 제거하고 중요도가 높은 변수만 선택하는 과정을

Table 6. Data Set after Data Balancing

Data balancing	Components localization			Weapons modification		
	Data set	Supported : Non-supported	Instance	Data set	Supported : Non-supported	Instance
Over-sampling	①	64 : 64	128	④	60 : 60	120
	②			⑤		
	③			⑥		
Hybrid-sampling	⑦	42 : 42	84	⑩	42 : 42	84
	⑧			⑪		
	⑨			⑫		
Under-sampling	⑬	21 : 21	42	⑯	25 : 25	50
	⑭			⑰		
	⑮			⑱		

변수선택(feature selection)이라고 한다. 본 연구에서는 예측모형 구축시 유의미한 핵심적 변수를 선택하기 위하여 이득비(Gain ratio) 개념을 사용하여 독립변수들의 중요성을 평가하였으며, 아래 식 (1)과 같이 산출방법이 정의된다.

$$Gainratio = IG/Intrinsicvalue \quad (1)$$

여기서 IG는 정보 이득(information gain)을 의미하며 아래 식 (2)와 같이 계산된다.

$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{|\, x \in T \mid x_a = v \,|}{|\, T \,|} \times H(x \in T \mid x_a = v) \quad (2)$$

T는 학습데이터의 집합을 의미하고, x_a는 변수 a에 대한 샘플 x가 가지는 값이며, $vals(a)$는 앞의 변수 a가 가질 수 있는 값들의 집합을 의미한다. 위 수식의 변수 a에 대한 정보이득(IG)는 엔트로피 H로 정의되며, H는 다시 아래의 식 (3)과 같이 정의된다. 이때 p_i는 한 집합 내에 클래스 i의 차지 비율을 의미한다.

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \quad (3)$$

정보이득(IG)를 분류기준으로 사용할 경우 아이디, 일련번호 변수처럼 다양한 값을 가진 변수일수록 큰 값을 갖게 되기 때문에 중요도가 높은 변수로 선택될 수 있다는 문제가 발생한다. 이득비(gain ratio)는 위의 IG 문제를 개선하기 위해 IG를 내재 가치(intrinsic value)로 나눈 값으로 계산된다. 내재 가치는 분할된 집합의 수와 집합 내에 있는 샘플 수를 반영하여 계산한 엔트로피 값이다.

본 연구에서는 래퍼 방식(wrapper)을 사용하여 변수를 선택하였다. 래퍼 방식은 크게 후진 제거기법과 정방향 탐색 방식으로 구분된다. 역방향 제거기법으로도 불리는 후진 제거기법은 매회 가장 중요도가 낮은 변수를 제거하여 예측모델을 개발하고, 독립변수들의 중요도에 따라 순위를 다시 산정하기 때문에, 이전 단계에서 탈락된 변수의 영향도가 없어진 상태에서 현재 변수들의 중요도를 산정한다. 이와는 반대로, 정방향 탐색 방식은 처음 한 번만 변수들의 중요도 순위를 산정한 후, 이 순위에 따라 중요한 변수부터 하나씩 모형에 추가하면서 모형 성능을 평가하는 방식이다. 일반적으로 정방향 탐색 방식보다 후진 제거방식의 성능이 더 우수한 것으로 알려져 있다(Witten *et al.*, 2005).

따라서 본 연구에서는 후진 제거기법을 사용하여 예측모형을 구성할 가장 이상적인 변수 조합을 찾고자 하며, 이때의 이득비 값이 낮은 변수 순부터 독립변수를 제거하면서 예측모형을 구축한다.

4.4 머신러닝 기반의 예측모형 구축

본 연구에서는 방산육성 정부지원사업의 수혜기업을 예측하기 위해 데이터 마이닝 분석 도구인 웨카(Weka) 버전 3.8.6을 사용하였다. 예측모형을 구축하기 위해 수집한 데이터는 70%의 학습데이터와 30%의 검증데이터로 분할하여 사용하였고, 학습데이터로 예측모형을 학습시킨 후 검증데이터로 기구축된 예측모형의 예측성능을 평가하였다.

예측모형 구축에는 로지스틱 회귀분석(LR), 의사결정나무(DT), 인공신경망(ANN), 베이지안 네트워크(BN), 최근접이웃(K-NN) 알고리즘 총 5가지를 사용하였다. 각 학습모형의 파라미터는 Weka에 설정된 기본값을 적용하였으며, ANN(Multilayer Perceptron)으로 예를 들면 설정환경은 Epoch 500, Batch Size 100, Learning Rate 0.3, momentum 0.2이다.

전처리 후 정부지원사업별로 선택된 독립변수의 중요도를 계산한 후, 중요도 값이 낮은 순서대로 변수를 하나씩 제거하면서 예측모형의 적중률(hit rate)을 살펴보았다. 적중률은 예측한 전체 클래스 중에서 정확히 예측한 클래스 수의 비율, 정확도(total accuracy)를 의미한다.

본 연구는 여기서 더 나아가 하이브리드샘플링을 적용한 데이터셋4, 데이터셋5, 데이터셋6 중에서 방산육성 정부지원사업별 적중률이 가장 높았던 예측모형 각 1개, 총 2개에 앙상블 기법을 적용하여 예측성능을 향상시키는 연구를 진행하였다. 앙상블 기법에는 배깅(bagging), 부스팅(boosting), 랜덤포레스트(Random Forest) 등이 있다. 배깅의 샘플링은 복원 추출 방식으로 복원 추출 집단을 여러 개 만들어 결합하여 알고리즘 성능을 향상시키는 기법이고, 부스팅은 데이터를 여러 군으로 나어 중복 선택이 가능하도록 하지 않고, 학습 데이터의 예측모형에 대한 중요도를 이용하여 선택 혹은 제거의 방법으로 데이터 군을 구성하고 요소 예측모형을 학습하여 조합하는 방법이다.

4.5 샘플링 방법별 예측모형 성능 결과

18개의 데이터셋을 대상으로 5가지 알고리즘을 적용하여 구축한 예측모델 90가지의 적중률은 각각 상이하였으며, 각각의 데이터셋 알고리즘마다 가장 높은 적중률을 아래 <Table 7>과 같이 정리하였다.

예측모델 90가지의 최고적중률 평균은 76.18%로 도출되었다. 랜덤샘플링 방법을 기준으로 구분하여 보면, 오버샘플링 데이터셋 ①~⑥ 평균 77.63%과 언더샘플링 데이터셋 ⑬~⑱ 평균 70.24%보다 하이브리드샘플링 데이터셋 ⑦~⑫의 평균이 80.67%로 가장 높은 적중률을 나타냈다. 샘플링 전처리 방법에 따라 최대 10.43%의 개선 효과가 나타났다. 알고리즘별 적중률 평균은 로지스틱 회귀분석(LR), 의사결정나무(DT), 인공신경망(ANN), 베이지안 네트워크(BN), 최근접이웃(K-NN) 알고리즘 중에서 최근접이웃(K-NN) 알고리즘을 활용한 예측모

Table 7. Data Set and Total Accuracy

		Over-sampling			Hybrid-sampling			Under-sampling			Average
Components localization	Data set	①	②	③	⑦	⑧	⑨	⑬	⑭	⑮	
	LR	63.16%	73.68%	78.95%	68.00%	68.00%	80.00%	69.23%	61.54%	69.23%	70.20%
	DT	71.05%	89.47%	78.91%	80.00%	68.00%	76.00%	46.15%	69.23%	76.92%	72.86%
	ANN	73.68%	81.58%	73.44%	68.00%	64.00%	84.00%	61.54%	61.54%	69.23%	70.78%
	BN	65.79%	55.26%	71.05%	60.00%	64.00%	80.00%	46.15%	46.15%	38.46%	58.54%
	K-NN	84.21%	84.21%	81.58%	84.00%	72.00%	92.00%	76.92%	76.92%	84.62%	81.83%
Weapons modification	Data set	④	⑤	⑥	⑩	⑪	⑫	⑯	⑰	⑱	
	LR	80.56%	86.11%	66.67%	88.00%	84.00%	84.00%	73.33%	66.67%	73.33%	78.07%
	DT	88.89%	88.89%	72.22%	92.00%	92.00%	88.00%	86.67%	73.33%	73.33%	83.93%
	ANN	80.56%	77.78%	69.44%	88.00%	76.00%	80.00%	66.67%	80.00%	80.00%	77.60%
	BN	86.11%	77.78%	69.44%	92.00%	92.00%	72.00%	73.33%	73.33%	73.33%	78.81%
	K-NN	91.67%	88.89%	77.78%	96.00%	96.00%	92.00%	100.0%	86.67%	73.33%	89.15%
Average		77.63%			80.67%			70.24%			76.18%

형의 평균 적중률이 부품국산화개발 지원사업과 무기체계 개조개발 지원사업 모두 높은 것으로 나타났다. 이 결과를 통해 클래스 수가 적은 데이터셋의 예측모형 구축 시, 데이터 불균형 문제를 해결하기 위해 흔히 쓰이는 오버샘플링, 언더샘플링 보다 더 나은 랜덤샘플링 방법이 있는 것으로 판단하였다.

하이브리드샘플링을 적용한 데이터셋 ⑦~⑫ 중에서 지원사업별 모든 알고리즘에서 비교적 높은 적중률을 보인 데이터셋 ⑨와 ⑩의 후진제거법 기반의 변수 수별 적중률을 아래 <Figure 2>와 <Figure 3>에 나타내었다. x축은 후진제거 기법을 통해 변수 중요도가 낮은 순부터 하나씩 제거되어 예측모

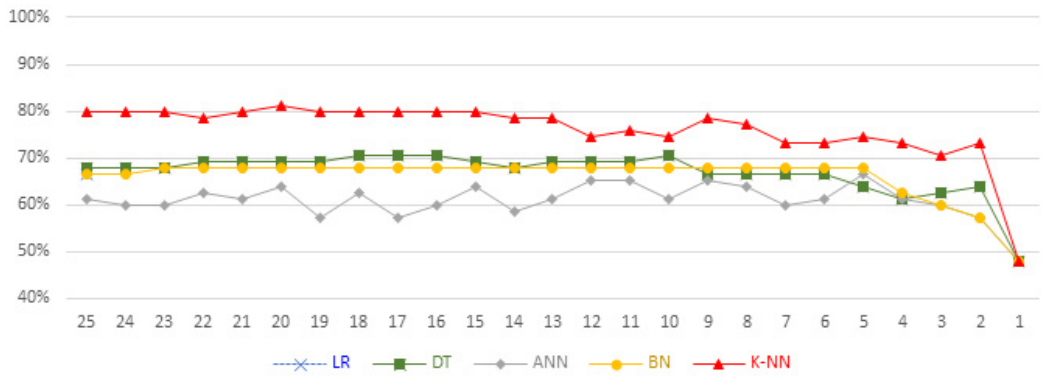

Figure 2. Data Set ⑨ Total Accuracy

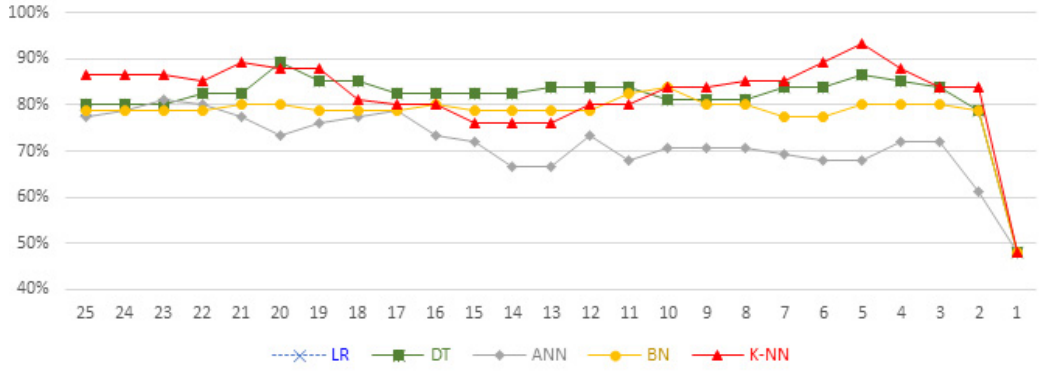

Figure 3. Data Set ⑩ Total Accuracy

Table 8. Attribute Selection Output

Data set ⑨			Data set ⑩		
Ranked attributes		Total accuracy	Ranked attributes		Total accuracy
Feature	Weight		Feature	Weight	
researcher	0.24894	84.00%	R&D Self-invested	0.2592	88.00%
Bachelor's degree	0.24894	84.00%	Total assets	0.2582	80.00%
R&D Government-invested	0.21021	88.00%	Domestic Patents/IPR	0.2426	80.00%
blue collar	0.20184	80.00%	Facility Self-invested	0.2259	96.00%
Master's degree	0.16785	80.00%			
Ph.D.	0.14979	80.00%			
Company Size	0.00857	84.00%			
Number of research institutes	0	92.00%			

형을 구성한 변수의 수가 줄어드는 것을 의미하며, y축은 해당 변수로 만든 예측모형의 적중률을 나타낸다.

부품국산화개발 지원사업 수혜기업 예측을 위한 하이브리드 샘플링 데이터셋 ⑨의 알고리즘별 적중률은 변수의 수와 관계없이 처음부터 끝까지 최근접이웃(K-NN) 알고리즘의 적중률이 70% 이상을 유지하며 가장 높았다. 반면 무기체계 개조개발 지원사업 수혜기업 예측을 위한 하이브리드 샘플링 데이터셋 ⑩의 경우에는 최고 적중률인 96%를 기록한 알고리즘이 최근접이웃(K-NN) 알고리즘이기는 하지만, 구성하는 변수의 수에 따라 최근접이웃(K-NN) 또는 의사결정나무(DT)가 번갈아가며 80% 이상의 높은 적중률을 보였다. 이로써 방위산업 육성 지원사업인 부품국산화개발 지원사업과 무기체계 개조개발 지원사업의 수혜기업을 가장 잘 예측하는 머신러닝 알고리즘은 로지스틱 회귀분석(LR), 의사결정나무(DT), 인공신경망(ANN), 베이지안 네트워크(BN), 최근접이웃(K-NN) 알고리즘 중 최근접이웃(K-NN) 알고리즘인 것으로 판단된다.

위의 <Table 8>은 각 사업별 대표 데이터셋의 예측모형을 구성하는 변수와 변수별 가중치를 비교한 표이다. 부품국산화개발 지원사업의 경우 최고의 예측모형을 구성하는 변수는 '연구원수', '학사졸업자 수', '정부지원 R&D 투자금', '생산직 수', '석사졸업자 수', '박사졸업자수', '업체규모', '방산전담 연구소 수'였다. 반면 무기체계 개조개발 지원사업의 경우 최고의 예측모형을 구성하는 변수는 '업체 자체 R&D투자금', '방산부문 자산총계', '국내 지식재산권/특허 수', '업체 자체 시설투자금'으로 차이가 있었다.

4.6 앙상블 기법을 적용한 예측모형 성능 개선

추가적으로 배깅, 부스팅 등을 포함하는 앙상블 기법을 적용하여 예측모형의 적중률을 향상시키기 위해 하이브리드샘플링 예측모형의 성능이 90%에 미치지 못하는 데이터셋 ⑦과 ⑧을 앙상블 기법을 적용할 데이터셋으로 선정하였다. 그 이유는 인스턴스 수가 84개인 소수의 하이브리드 샘플링 데이터셋에서 이미 90% 이상의 적중률을 보이는 예측모형은, 검증데이터(30%)의 90% 이상을 예측에 성공했다는 뜻이기 때문이다. 이는 수혜기업과 비수혜기업 여부를 데이터셋에서 1~2개를 제외하곤 다 맞췄다는 의미이기 때문에 그 이상의 적중률을 기대하며 성능 개선을 시도하기에는 데이터셋의 규모가 충분하지 않다. 따라서 하이브리드샘플링 데이터셋 중 예측성능이 비교적 낮은 부품국산화개발지원사업의 데이터셋⑦, ⑧에 앙상블 기법을 적용하여 적중률이 향상되는지 살펴보았다.

배깅은 복원 추출 방식을 사용하며, 여러 개의 복원 추출 그룹을 만들고, 이들 그룹을 결합하여 알고리즘 성능을 향상시킨다. 데이터가 N개인 데이터셋에 N의 값보다 작은 랜덤 그룹을 중복 가능한 상태로 추출하여 만든 학습데이터를 사용한다. 여러 학습 데이터셋으로 구축된 여러 기본 예측모형을 결합한 후 다중 기본 예측모형들의 다수결 원칙으로 최종 모형을 결정한다. 다시 말하자면, 여러 예측모형의 조합으로 전체 예측모형의 효과를 평준화 시켜 일정 수준 이상의 예측성능 모형을 설계하는 것이다. 그러나 배깅은 기본 예측모형의 가중치를 고려하지 않고 다수결을 이용해 결정한다는 점 탓에 한계가 있다. 일례로 정확도가 높은 1개와 정확도가 낮은 2개의 기본 예측모형 조합이 다수결 전략 탓에 새로운 샘플을 예측한다면, 부정확한 예측 결과가 나올 것이다.

한편 부스팅은 데이터를 여러 그룹으로 나누지 않고, 학습데이터에서 예측모형의 중요도를 이용하여 변수선택 혹은 제거의 방법으로 데이터 군을 구성하고 요소 예측모형을 학습 및 조합하는 방식이다. 그 중 아다부스트(Adaptive Boosting) 알고리즘은 Freund *et al.*(1997)이 고안하였으며, 부스팅 사용자가 원하는 성능이 나올 때까지 예측모형을 추가해도 과적합 문제가 적다는 장점이 있다.

아래 <Table 9>는 각 데이터셋에 배깅과 부스팅을 적용하기 전과 후의 적중률을 비교한 표이다. 각각의 앙상블 기법에서 최근접이웃(K-NN) 단일 알고리즘을 적용할 때와 동일하게 학습데이터와 검증데이터의 비율을 70:30으로 하였고, 예측모형을 구성하는 변수의 수도 동일하게 설정하였다. 그 결과, 데이터셋⑦의 경우에는 배깅을 적용했을 때 적중률이 4.0% 상승하였고, 데이터셋⑧의 경우에는 배깅과 부스팅을 적용했을 때

적중률의 상승이 없었다. 이로써 앙상블 기법 적용을 통해 방산육성 지원사업 수혜기업 예측모형의 성능을 일부 향상시킬 수 있다고 판단하였다.

Table 9. Total accuracy after applying Ensemble

Data set	Single algorithm	Ensemble	
		Bagging	Boosting
⑦	84.00%	88.00%	84.00%
⑧	72.00%	72.00%	72.00%

지금까지의 연구 결과를 종합하면 다음과 같다. 방산업체의 방위산업육성 지원사업의 수혜기업 여부를 예측하는 방법으로 방위산업 실태조사의 기업 경영정보 변수가 활용될 수 있음을 알 수 있다. 방산업체를 운영하는 기업 경영자들은 정부의 방산육성 지원정책의 수혜기업이 되기 위해서 지원사업 유형별로 수혜기업을 결정하는 요인이 서로 상이하다는 것을 인식하고, 기업별 자사의 역량 중 부족한 분야를 선제적으로 찾아 개선하는 활동을 통해 수혜기업이 될 확률을 높일 수 있다. 정부의 정책입안자 입장에서는 그간의 방산육성 지원사업 예산이 방위산업을 구성하는 업체 모두에게 고르게 분배되었는지, 특정 경영상태를 띄는 기업을 집중적으로 지원하였는지를 파악할 수 있다. 한편 데이터 분석의 관점에서는 수혜기업 예측모형의 성능 향상을 추구하기 위해 데이터셋의 클래스 균형화와 변수 선택과 같은 전처리 기법과 더불어 추가적으로 앙상블 기법을 적용해볼 수 있다.

4.7 방산육성지원 수혜확률 향상 방안

부품국산화개발 지원사업과 무기체계 개조개발 지원사업의 수혜기업 예측모형 중, 적중률이 가장 우수했던 모형을 구성하는 변수에는 사업별 약간의 차이가 있었다. 먼저 두 사업 예측모형에 공통적으로 사용된 변수는 일정 수준 이상의 '방산부문 자체 R&D 투자금'과 일정수준 이상의 '방산전담 연구소 수', '연구원수', '학/석/박사 직원수'이다. 이는 방위산업분야 정부지원사업에 선정되기 위해서는 기본적으로 기업의 연구전문 종사인력 규모와 연구시설 유무가 중요한 고려 요소임을 알 수 있다.

이를 바탕으로 기업이 할 수 있는 방산육성지원 수혜확률을 높이는 방안은 첫째, 업체자체 R&D 투자금이 높은 기업이 정부 지원사업 수혜기업이 될 확률이 높으므로 R&D에 투자하려는 자체적인 노력을 확대할 필요가 있다. 기업의 자발적 R&D 투자는 장기적으로 기업의 사업모델을 R&D에 집중하게 하고, 이는 기업의 연구인력을 질적・양적으로 향상시키는 노력과 같은 맥락으로 기업의 혁신역량을 높이는 방법이다. 둘째, 기업은 연구인력을 확대하고 가능하다면 R&D 인력 보유나 R&D 부서 운영을 넘어서서 방산 R&D 전담 연구소 규모로 확대 운영하는 것이 수혜기업이 될 확률을 높일 수 있다. 이는 업체자체 R&D 투자금을 확대하는 것과 마찬가지로 정부 지원사업 수혜기업의 공통적인 특징이자, 대상 기업에 개발비 지원의 필요성과 당위성을 뒷받침해줄 수 있는 필수적 수혜요건이 될 수 있다. 종합하자면 기업의 비즈니스 모델의 일환으로 R&D 활동을 장려하는 것이 수혜기업이 될 확률을 높이고, 또 고용시장에서 R&D 사업을 왕성하게 하는 우량기업이라는 긍정적인 평가가 고기술・고숙련 연구개발인력의 채용을 수월하게 만드는 선순환 구조의 시작점이다. 기업은 지속가능한 성장을 위해서 연구장려 전략과 R&D 및 시설투자, 고급인력 확충을 선제적으로 준비하여 정부지원사업의 수혜확률을 높일 수 있다.

한편 지원사업별 수혜 예측모형에서 차별성을 갖는 변수는 부품국산화개발 지원사업의 경우 '업체규모', '생산직 수' 변수이다. 부품국산화개발 지원사업은 무기체계의 안정적 운용을 보장하기 위해 국외도입 부품을 국내 기업의 기술과 인력을 이용하여 개발 및 생산해야 하기 때문에 '생산직 수'와 '업체규모'가 일정수준 이상에 도달한 업체여야 한다. 즉, 부품국산화개발 지원사업에 선정되길 희망하는 기업일 경우, R&D 인력과 R&D 자체투자를 확대할 뿐만 아니라 '생산직 수'와 '업체규모' 또한 일정수준 이상으로 높여서 기업의 기술과 인력를 이용하여 국외도입 부품을 국산으로 개발 생산할 수 있음을 어필할 필요가 있다.

무기체계 개조개발 지원사업의 경우 차별성을 갖는 변수는 '방산부문 자산총계', '국내 지식재산권/특허 수', '업체 자체 시설투자금'이다. 무기체계 개조개발 지원사업에 선정되길 희망하는 기업일 경우, R&D 인력과 R&D 자체투자를 확대할 뿐만 아니라 업체의 '자산총계'가 어느 정도 규모에 있는 안정적 기업이며, '자체 시설투자'가 활발하고, IRP 관리가 잘 되고 있음을 어필하여야 한다. 즉, 해외 구매국이 요구하는 사양에 따라 기업의 자체 설비를 이용하여 개조개발을 할 수 있음을 증명하는 방법으로 개조개발 지원사업의 선정 확률을 높일 수 있다.

5. 결 론

본 연구는 방위산업육성 지원사업에 선정되는 수혜확률을 높이기 위해 기업이 경영현황 측면에서 선제적으로 준비해야 할 것이 무엇인지 파악하고자 하는 방산업체의 니즈에 착안하여 수행하였다. 이에 방위산업 실태조사에서 확보한 기업의 방산부문 경영현황 정보를 활용하여 부품국산화개발 지원사업과 무기체계 개조개발 지원사업의 수혜기업 예측모형을 구축하는 실험을 진행하였다. 이때 변수의 데이터 클래스 불균형 문제를 보완하기 위하여 랜덤샘플링 방법인 오버샘플링, 언더샘플링, 하이브리드샘플링 방법을 적용하였다. 변수 선택 기법으로는 후진제거 기법을 적용하여 중요도 높은 변수로 구성된

적중률 높은 수혜기업 예측모형을 구축하였다. 또한 예측모형의 성능을 향상시키는 방법으로 앙상블 기법을 적용하는 시도를 하였다.

본 연구의 의의는 정부지원사업 수혜기업의 경영성과를 분석한 수많은 연구와는 달리, 수혜기업과 비수혜기업을 결정하는 기업의 특징적 요인을 분석하고, 수혜기업 예측모형을 구축한 점이다. 연구 과정에서 발견한 시사점은 첫째, 소수의 데이터셋에서 데이터의 불균형 문제를 해결할 때 다양한 랜덤 샘플링 시도가 예측모형의 성능을 높일 수 있음을 확인하였다. 본 연구의 데이터셋에서는 오버샘플링(77.63%), 언더샘플링(70.24%)보다 이를 혼합한 하이브리드샘플링(80.67%) 방법의 평균 성능이 더 우수했다. 이는 국내 데이터 예측모형 연구 분야에서 특히 샘플링이라는 전처리를 다각화하여 성능을 높이려는 시도에서 학술적 기여점이 있다. 둘째, 5가지 알고리즘 중 최근접이웃(K-NN) 알고리즘의 예측모형이 두 지원사업 데이터셋별 평균 81.83%, 89.15% 수준으로 예측성능이 가장 우수하였다. 배깅 및 부스팅과 같은 앙상블 기법을 적용하여 예측모형의 성능을 한 단계 더 개선할 수 있음도 확인하였다. 셋째, 예측모형의 중요 변수들을 토대로 방산업체가 정부의 지원사업 수혜확률을 높일 수 있는 선제적 전략 방안을 제시하였다. 방산부문 자체 R&D 투자금 확대와 방산전담 연구소 운영, 양질의 연구원 확보 등 인사정책을 통해 수혜기업에 선정될 가능성을 높일 수 있음을 확인하였다. 해당 전략들은 특히 부품국산화 사업이나 무기체계 개조개발 사업에 참여를 준비중인 방산업체에게 실무적으로도 도움이 될 수 있는 기여점이라고 할 수 있다.

본 연구의 한계점은 방산육성 지원사업 수혜기업이 선정 당시 제안했던 R&D 주제와 내용의 질적 수준은 예측모델 구축 시 고려되지 않았다는 점이다. 정부지원사업은 정책적 목표에 맞게 지원대상과 과제선정 기준을 갖고, 선정기준에 부합하는 가장 우수한 R&D 계획을 제안한 기업이 수혜기업이 되는 것이 일반적인 선정평가 과정이다. 그러나 본 연구에서는 업체별 지원사업 수행계획 제안정보를 열람 및 확보하기에 제한되는 탓에 기업의 정량적 경영정보만을 변수로서 분석하였다. 향후 연구에서는 수혜기업 선정단계에서 확보한 기업의 개발계획서 자료를 추가적으로 활용하여 기술의 활용성, 경영현황 등을 다각적으로 분석할 수 있는 모형 설계가 필요하다. 또한, 향후 정부의 지원사업 선정시 수혜기업이 되길 희망하는 방산업체라면, 본 연구내용의 수혜기업 결정요인 중 중요도가 높은 항목을 선제적으로 개선하는 활동을 통해 지원사업 수혜확률을 높이는 데 활용할 수 있을 것으로 기대한다.

참고문헌

Batista, G. E., Bazzan, A. L., and Monard, M. C. (2003), Balancing training data for automated annotation of keywords: A case study, *Wob*, **3**, 8-10.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004), A study of the behavior of several methods for balancing machine learning training data, *ACM SIGKDD Explorations Newsletter*, **6**(1), 20-29.

Choi, J., Lee, Y., Jang, M., and Seo, S. (2018), Efficiency Evaluation of Defence Industry Firms by Utilizing DEA, *Journal of the Korea Academia-Industrial cooperation Society*, **19**(9), 501-507.

Choi, K., Narangarav Bathhuyag, Kim, G., and Lee, J. (2023), Private Tutoring and Academic Performance of Elementary, Middle, and High School Students: A Decision Tree Algorithm, *Journal of Digital Contents Society*, **24**(4), 847-860.

Choo, T., Kim, J., Park, W., and Choi, H. (2023), A Study on the Evaluation of Tidal Prediction Capacity of Busan, Gadeokdo, and Geoje Island using Logistic Regression Analysis and Multiple Regression Analysis, *Journal of the Korea Academia-Industrial cooperation Society*, **24**(7), 466-473.

Dash, M. and Liu, H. (1997), Feature selection for classification, *Intelligent Data Analysis*, **1**(3), 131-156.

Freund, Y. and Schapire, R. E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**(1), 119-139.

Han, H. and Choi, S. (2011), An Artificial Neural Network Approach for the Prediction of Unlawful Company in Defense Procurement, *Journal of the Korea Management Engineers Society*, **16**(3), 29-38.

Hwang, C. (2022), Improvement of early prediction performance of under-performing students using anomaly data, *Journal of the Korea Institute of Information and Communication Engineering*, **26**(11), 1608-1614.

Hwang, H., Lim, Y., and Jeon, J. (2022), Analyzing of Factors for Performance Evaluation of Defense R&D Projects by Using PCA/FA, *Journal of Korea Technology Innovation Society*, **25**(2), 275-289.

Joo, H. and Lim, J. (2023), Machine Learning Prediction Model of Water Quality Factors in Sew age Treatment Facilities, *Journal of the Korean Institute of Industrial Engineers*, **49**(1), 95-106.

Kang, J. and Cho, K. (2018), An Analysis of the Effect of Government Support on Automation and Smart Factory, *Journal of Korea Technology Innovation Society*, **21**(2), 738-766.

Kim G. and Kwak K. (2018), Adverse Selection in the Government R&D Support for Venture Business : Evidence from the Managerial Efficiency Comparison of the Recipient and Non-recipient of R&D Grants, *Journal of Technology Innovation*, **21**(4), 1366-1385.

Kim, J., Kim, M., Yoo, J., Jung, S., and Kim, T. (2022), Sensitivity Analysis of Drought Factors to Drought Risk Using Structural Equation Model and Bayesian Network, *Journal of the Korean Society of Civil Engineers*, **42**(1), 11-21.

Kim, R., Lee, D., and Kim, G. (2016), Development of Prediction Model of Financial Distress and Improvement of Prediction Performance Using Data Mining Techniques, *Information Systems Review*, **18**(2), 173-198.

Kong, H., Bong, K., and Park, J. (2020), A Study on the Effect of Government Support on the Innovation of Defense Industry: Evidence from Korean firms, *Journal of Digital Convergence*, **18**(1), 1-10.

Lee, D., Yoon, K., and Noh, Y. (2020), Study on Improving Learning Speed of Artificial Neural Network Model for Ammunition Stockpile Reliability Classification, *Journal of the Korea Academia-Industrial cooperation Society*, **21**(6), 374-382.

Lee, G. (2023), A study on the predictive models of marginal companies based on machine learning: Focused on hotels' financial ratio for three years, *International Journal of Tourism and Hospitality Research*,

37(1), 123-135.
Lee, K., Kim, M., and Yoon, B. (2024), Analysis of the Economic Effect of the Smart Factory-supporting Government Programs, *Journal of the Korean Institute of Industrial Engineers*, **50**(1), 23-35.
Lee, S. and Cha, D. (2020), A Study on the Devolopment of Defense Industry in the Republic of Korea Based on Main Forein Countris, *The Journal of Social Convergence Studies*, **4**(5), 113-133.
Lim, T., Lim, K. Chung, S., and Han, S. (2021), Disease diagnosis research using deep learning based on military medical data, *Journal of Digital Contents Society*, **22**(9), 1359-1367.
Lim, Y. and Jeon, J. (2019), Analyzing the Performance of Defense R&D Projects based on DEA, *Journal of the Korea Institute of Military Science and Technology*, **22**(1), 106-123.
Mathew, J., Pang, C. K., Luo, M., and Leong, W. H. (2018), Classification of imbalanced data by oversampling in kernel space of support vector machines, *IEEE Transactions on Neural Networks and Learning Systems*, **29**(9), 4065-4076.
Oh, M., Choi, H., Kim, S., Jang, J., Jin, J., and Chun, M. (2017), Machine learning-based social security big data analysis and prediction model research, *Korea Institute for Health and Social Affairs*.
Oh, S. and Jang, P. (2020), The Effect of Government R&D Support on Manufacturing Firms' Innovation Activities and Innovation Performance, *Journal of Korea Technology Innovation Society*, **23**(5), 941-966.
Rho, D. (2021), A Study on the Effect of Key Competency and Competitive Strategy Factors on Management Performance of Small and Medium Businesses in the Defense Industry ; Adjustment Effect of Supply Chain Partnership and Government support systems, *Journal of the Korea Association of Defense Industry Studies*, **28**(3), 1-16.
Witten, I. H. and Frank, E. (2005), *Data mining: Practical machine learning tools and techniques*, San Francisco: Morgan Kaufmann Publishers.
Yoo, S. and Kim, K. (2022), Comparison of Anomaly Detection Performance Based on GRU Model Applying Various Data Preprocessing Techniques and Data Oversampli, *Journal of The Korea Institute of Information Security & Cryptology*, **32**(2), 201-211.
Yoon, D., Jung, H., Park, K., Lee, S., and Lee, J. (2020), The Effect of Government SME Support Programs (2008-2017): Focused on Consulting Support Programs, *Korean Journal of Business Administration*, **33**(9), 1597-1623.

저자소개

전고운 : 한국외국어대학교 스페인어과에서 2012년 학사를 취득하고, 경상국립대학교 2017년 산업시스템공학 석사학위 취득 후, 기술경영학과 박사과정에 재학 중이다. 국방기술진흥연구소 재직 중이며 연구분야는 방위산업 육성지원정책, 데이터마이닝이다.

유동희: 고려대학교 경영학과에서 MS/IS 전공으로 2009년 경영학박사를 취득하였다. 육군사관학교와 연세대학교에서 근무하였고, 현재 경상국립대학교 경영정보학과 교수 재직 중이다. 연구 분야는 지능형 정보시스템, 인공지능, 빅데이터 분석, 지식 그래프, 온톨로지 등이다.

전정환 : 서울대학교 산업공학과에서 기술경영/정책 전공으로 공학박사를 취득하였다. 삼성전자, 국가과학기술위원회 등에서 근무하였고, 현재 경상국립대학교 산업시스템공학부/기술경영학과/과학기술정책학과 교수로 재직 중이다. 연구분야는 개방형 혁신, 기술로드맵, 기술기획, 특허분석, 의사결정기법 등이다.