분석 23-10

Fiscal Information Service

머신러닝 기법을 적용한 재정데이터 활용

김 민 경

2023. 12.

FIS 한국재정정보원

머신러닝 기법을 적용한 재정데이터 활용

김 민 경 부연구위원

이 보고서는 한국재정정보원 홈페이지(www.fis.kr)을 통해 보실 수 있습니다.
한국재정정보원은 발간 보고서에 관한 여러분의 의견을 기다리고 있습니다.
제안하신 소중한 의견이 향후 재정관리시스템 개선과 재정운영에 활용될 수 있도록
최선을 다하겠습니다.

☎ 02-6908-8200 ✉ mkkim@fis.kr

요 약

1. 분석의 필요성 및 목적

□ AI 기술이 급속도로 발전하고 다양한 분야에 걸쳐 관련 기술이 활용되면서 공공분야 및 분석 영역에서도 적용에 대한 관심이 증대

- AI 기술은 여러 분야로 다시 세분화 되는데, 대표적으로 머신러닝과 딥러닝 등을 들 수 있음
  - 다차원 변수 및 빅데이터 처리가 가능하다는 점에서 머신러닝 알고리즘은 복잡한 사회문제를 분석하고 극복할 방안을 제시할 새로운 메커니즘으로 기대
- 공공영역에서는 예측 및 전망, 이상거래 탐지 및 선별, 의사결정 등 머신러닝 알고리즘이 도입 및 활용
  - 다만, 재정데이터를 대상으로 한 머신러닝 기법이 적용된 분석은 소수이며 이는 재정데이터의 접근성과 결합데이터의 제약에 기인
  - 이러한 제약을 극복하기 위한 대안으로 빅데이터 기반의 머신러닝 분석 기법이 도입되었고 학습모델에 따른 예측이나 전망, 분류 등과 같은 분석 가능

□ 본 연구는 dBrain+에 탑재되어 있는 재정 데이터를 대상으로 머신러닝 기법을 활용한 다양한 분석기법의 적용가능성 탐색

- 활용 가능한 머신러닝 기법의 알고리즘을 살펴보고, 국내·외 공공부문을 중심으로 머신러닝 기법 활용 가능성 및 도입사례 분석
  - 다양한 머신러닝 기법을 적용 가능한 사례와 접목해 가능성 제시
  - 머신러닝 알고리즘을 적용하기 위해 전제되어야 할 데이터 측면에서의 선결조건도 제시
- 실제 dBrain+ 불용사유 관련 데이터를 활용한 텍스트 분석 결과에 기초해 가능성 탐색 및 유용한 분석이 되기 위한 개선사항 제시

2. 머신러닝 이론 및 적용사례

□ 머신러닝은 일반적으로 AI[1] 기술의 알고리즘을 기반으로 프로그래밍하지 않아도 스스로 데이터로부터 규칙을 학습하는 기술

- 머신러닝(merchine learning)이란 “컴퓨터에게 명시적으로 프로그래밍 하지 않고도 학습을 할 수 있는 능력을 주는 연구 분야”로 정의
  - 인간의 학습에 비해 머신러닝이 가지는 가장 큰 장점은 매우 방대한 양의 데이터를 대상으로 학습을 통해 최적의 모델링 제시가 가능
- 전통적인 통계적 모델링에 기반한 추정 방식은 이론을 토대로 연구자가 설정한 변수를 모형에 채택하는 방식인데 비해 다양한 변수들의 조합 가능한 경우를 모두 계산하여 가장 최적화된 모델을 찾아주는 방식

□ 머신러닝은 일반적으로 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 구분

- 지도 학습은 데이터셋에서 일부를 학습데이터(training data)로 지정하여 학습을 진행시키고 테스트샘플로 하여 결과를 예측하여 최적의 모델을 추정
  - 지도학습 알고리즘은 금융, 경영, 의학 등 예측이 필요한 영역에서 활발히 활용되고 있으며 최근에는 공공정책 분야에서 전망, 효과성 예측에 활용
- 비지도학습은 학습데이터 설정 없이 알고리즘이 스스로 각 관측치끼리의 동질성과 이질성을 기준으로 유사한 집단으로 분류
  - 주로 이미지 분류나 손글씨 판별 등 비정형데이터 중심으로 활용

□ 우리나라에서도 정부부처 및 유관기관을 중심으로 탐지, 예측, 분석 등의 영역에서 활용

- 한국재정정보원에서는 국고보조금의 부정수급을 방지하고자 하는 목적으로 보조금의 부정수급을 탐지하는 시스템을 구축하여 운영하고 있는데 일부 머신러닝 알고리즘 활용

1) 인공지능(Artificial Intelligence)이란 사람의 지능을 모방하여 사람이 하는 것과 같은 복잡한 일을 할 수 있는 기술을 의미함

- 부정수급 관리 데이터 현황은 ’23년 7월 30일 기준으로 총 36.6만여 건이며 데이터 속성 정보로는 소관_회계_계정_분야_부문_프로그램_단위_세부_내역_세목_보조사업이며 사업연도, 예산액, 예산현액, 사업금액, 수행기관, 부정수급자 유형(1,2,3호), 대, 중, 소분류(60 유형), 환수결정액, 환수액 등이 포함
- e나라도움의 부정수급 탐지 시스템을 활용한 의심사례는 4,603건이고 금액은 2,379억 원으로 집계, 최종 부정수급 사례 적발 실적은 260건, 98.1억 원 수준

• 건강보험심사평가원에서는 2018년부터 시범적으로 이미지 정보 중심의 의료영상심사판독시스템 도입 및 활용을 시작으로 분석심사를 위한 특성기관 예측, 급여정보 분석시스템 등을 구축 활용

- 의료영상심사판독 시스템은 개별 급여기관에 저장되어 있는 영상정보를 활용하여 판독대상을 선별하고 심사단계에서는 심사연계 과정 중에 청구명세서 정보, 영상정보 등을 활용하여 심사업무의 효율성과 객관성에 기여
- 집중 심사가 필요한 요양기관을 선별 및 특성기관 예측을 위해 요양기관을 대상으로 한 임상결과, 임상 및 비용에 대한 진료행태 및 진료추세 등을 파악하여 회귀분석을 통해 예측, 집중관리 기관 선별에 활용
- 급여정보 분석시스템은 진료비와 보험자부담금을 대상으로 구축한 패널 데이터셋을 대상으로 머신러닝 예측 알고리즘을 적용해 진료비 예측정보 산출

□ 해외에서는 미국 보건후생부의 아동학대예측시스템, 국세청의 전자부정적발시스템(EFDS), 영국 감사원의 부적정 거래 위험요인 탐지 시스템

• 아동학대 예측 시스템은 위험요인을 수집하고 질문에 대한 응답 데이터를 축적해 패턴인식 및 확률모델 기법을 활용해 사망사건과의 연관성을 분석하고 위험군을 선별

- 시스템 도입을 통해 연 6,550만 명의 아동을 관리 할 수 있었으며, 가정방문의 질적 향상 등 서비스의 총체적 질이 23% 향상

• 미국 국세청의 전자부정적발시스템(EFDS)은 기계학습 등을 활용하여 세금 탈루자에 대한 분류 및 패턴 분석도 병행

- 또, EFDS를 구축하여 납세자의 은행계정 인적정보 등을 분석하여 부정 위험을

탐지함

- 2015년 기준, 부정한 환급 의심사례 69만 건을 조사한 결과 62%인 약 43만 건이 실제 부정하게 세금을 환급받은 것으로 판명

• 영국 감사원은 대상기관 재정자료를 분석하여 부정한 사용이 의심되는 사례를 추출하는 부적정거래 위험요인 탐지 시스템을 구축 운영

- 대상기관 재정DB와 연계한 data warehouse를 구축하여 감사담당자가 해당 분야의 원시자료 추출을 통해 분석 가능하도록 지원
- 지출과정에서 발생할 수 있는 위험의 정도를 점수화하여 부적절한 거래 유형 및 내역을 분석하여 제공하고 있으며 거래 세부 정보를 분석하여 과다 집행, 부적정 집행 여부를 파악

3. dBrain+와 머신러닝 기법 적용 가능성

□ 지도학습의 예측과 관련된 알고리즘을 정량적 dBrain+ 재정데이터에 적용하여 효과 분석이나 전망 분석 가능

• 선행연구에서는 랜덤 포레스트 기법 등을 활용하여 특정 재정사업의 정책효과를 수혜자 집단을 세분화하여 가장 효과성이 높은 집단을 규명하는 방식으로 평가

• dBrain+의 세부사업 단위로 연간 재정데이터가 존재하고 있으므로 여기에 다른 DB(통계청의 SBM, 국세청 소득DB 등)와의 변수 결합 등을 통해 특정 재정사업에 대한 효과 분석 및 재정사업의 경제적 파급효과 분석도 가능

• 미시, 거시 경제 지표(GDP, 무역수지, 환율 등등)와 재정 데이터를 결합시킨 머신러닝 기법으로 구축된 재정전망 모형은 이미 해외에서도 활발히 연구가 이루어지고 있는바, dBrain+의 재정데이터를 활용하여 추후 모형 설계 가능

• 머신러닝의 분류 알고리즘 활용 시, 국고보조금 시스템에서 구축된 부정수급 의심사례 선별이나 실시간 집행에서의 이상치 탐지 등으로 활용되고 있고, 추후 알고리즘의 고도화를 위한 새로운 유형의 패턴 발굴에 활용 가능

□ 비지도학습 알고리즘을 적용하면 한글 문서를 포함한 텍스트 데이터로 대표되는 비정형데이터의 활용 분석 가능

- 텍스트 마이닝을 활용하여 각 재정사업별 사업 설명자료, 불용사유, 성과정보 등에서 정성적인 측면의 재정사업의 특성 또는 성과 정보와 결합된 정책효과 측정 가능
- 텍스트마이닝 기법을 재정정보 대상으로 적용하면 재정정보 말뭉치 구축, 재정 집행 이상 탐지, 통계 서비스 고도화, 정책 사업 평가, 정책 사각지대 발굴 등이 가능

□ 실제 불용사유 데이터를 대상으로 텍스트 빈도분석을 실시한 결과 입력내용의 충실성 미흡과 오분류가 다수 나타나 이에 대한 시스템적 개선과 재정 용어 말뭉치 구축을 통한 분석환경 조성 필요

- 분류 오류를 줄이고 텍스트 분석 결과의 품질을 높이기 위해서 텍스트 raw 데이터 입력자들의 입력 오류를 줄이려는 노력과 더불어 불용사유 입력에 있어 상세한 사유를 입력하려는 독려가 필요
- 오분류가 일어나지 않도록 '재정용어' 중심으로 텍스트 분석을 할 수 있도록 재정용어 말뭉치 및 재정토큰에 기반한 사전을 활용할 수 있도록 전처리 환경을 구축할 필요
- 추후 불용유형별, 사업유형별 텍스트 빈도 분석이나, 불용금액과 텍스트 간의 관계를 규명하는 데에도 활용 가능

4. 머신러닝 기법 분석 기반 확충을 위한 선결조건

□ 머신러닝 기법은 학습데이터에 있어 변수가 많을수록 데이터의 수가 많을수록 모형의 예측 성능이 향상되기 때문에 재정 데이터에 결합할 다른 DB와의 연계 필요

- 예측 및 전망에 있어 다양한 변수를 활용할수록 일반적으로 예측력이 높아지는 머신러닝 기법의 특성에서 기인
- 유관 타 DB와의 연동을 통한 수혜자 집단 중심의 변수 획득을 통해 분석의 품질을 높이고 보다 고차원적 분석에 활용할 데이터를 확보 필요

- 이를 통해, 수혜자 개인단위, 사업자 단위에서의 재정집행의 경제적 파급효과를 실집행 관점에서의 분석 및 수혜집단별 정책효과 분석도 가능

□ 비정형데이터를 활용하기 위해서는 전처리 과정을 통한 데이터 정제 수준에 따라 분석의 품질이 결정되는데 말뭉치-토큰화-재정용어사전 으로 이어지는 유기적인 분석 인프라 구축 필요

- 열린재정과 dBrain+상에는 성과계획서, 조세지출, 예·결산 문서 등 다수의 텍스트 문서가 시계열로 탑재되어 있음에도 이러한 텍스트 자료의 활용은 미진한데 언어의 의미에 따른 전처리 과정에 필요한 기반이 조성되어 있지 않은 데서 기인
- 텍스트 마이닝 분석은 데이터 정제작업에 따라 분석의 난이도와 품질이 결정되는데 재정 용어 및 문서에 특화된 재정말뭉치, 재정 용어사전 구축이 선행될 필요

□ 재정사업속성 정보가 양적, 질적으로 확충됨과 동시에 잘 관리 되어야 재정사업 효과분석이나 성과분석에 활용 가능

- 재정사업은 분석에 있어 변수 및 수혜 집단, 사업 특성 등을 분류 인자로 작용할 특성을 속성화하여 메타 정보화하는 작업이 필요하며 그 자체로 데이터 변수로 활용
- 머신러닝 기법을 적용해 실질적인 효과를 얻기 위해서는 구조화된 속성정보 설계 및 확충이 관건, 즉 재정데이터 분석에서 수요가 높은 실집행 분석을 위한 수혜자 및 정책유형에 따른 분석 등에 활용

목 차

목 차

표 목차

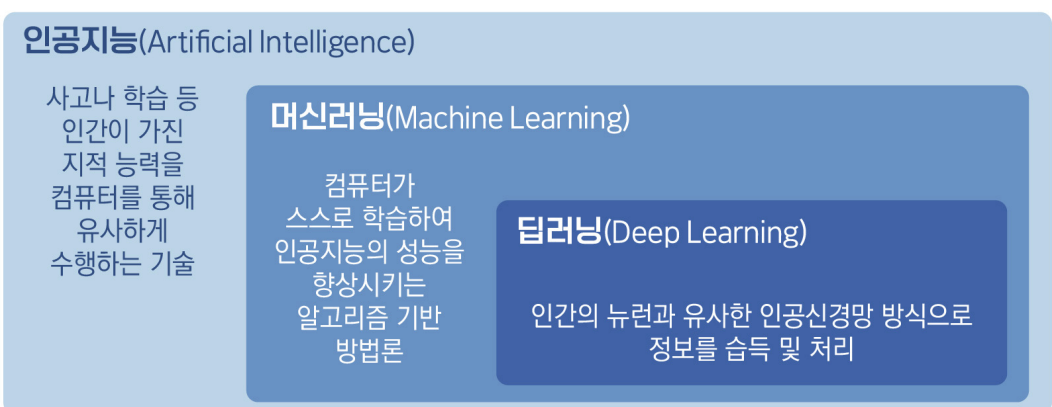
그림 목차

Ⅰ. 서론

1. 분석 배경 및 목적

AI 기술이 급속도로 발전하고 다양한 분야에 걸쳐 관련 기술이 적용되면서 연구분석 영역에서도 그 활용에 대한 관심이 높아지고 있다. 초기의 단순반복적인 업무 보조 수준을 넘어 최근의 챗GPT와 같은 생성형 AI 기술까지 발달하게 되면서 행정 및 각 산업 분야에서의 활용은 거스를 수 없는 대세로 자리 잡아 가고 있다. AI 기술은 여러 분야로 다시 세분화되는데 대표적으로 머신러닝과 딥러닝 등을 들 수 있다. 그 중 다차원 변수 및 빅데이터 처리가 가능하다는 장점으로 인해 머신러닝 알고리즘은 복잡한 사회문제를 분석하고 극복할 방안을 제시할 새로운 메커니즘으로 기대되고 있다(정일환, 2021).

〈그림 Ⅰ-1〉 인공지능, 머신러닝, 딥러닝 관계



자료 : 저자 작성.

행정, 정책 분야에서도 이러한 흐름에 따라 관련 기술의 적용을 통한 빅데이터 분석을 중심으로 머신러닝 알고리즘을 활용한 예측 및 전망, 이상거래 탐지 및 선별, 의사결정 등의 다방면에서 실무적으로 이미 활용되고 있거나 도입되고 있는 실정이다.

해외에서는 이미 IT 산업을 중심으로 알파고나 메타버스, 챗GPT로 대변되는 민간에서의 기술 발전을 공공부문에서 정책의사결정을 중심으로 적용 및 활용이 이루어지고 있다. 영국 국가 감사원(NAO)에서는 감사 주제 및 대상을 선정하는데 머신러닝 기법을 통한 빅데이터 분석을 활용하고 있으며 미국 회계감사원(GAO)에서는 공공부문의 부적절한 지출사례를 적발하거나 정책 및 사업의 모니터링에도 활용하고 있다(차경엽, 2022).

이러한 점에 비추어 볼 때, 국가 재정관리 측면에서도 이러한 머신러닝 기법이 활용 가능한 부분이 존재한다. 머신러닝 분석 기법이 빅데이터 기반으로 한 학습모델에 따른 예측이나 전망, 분류 등과 같은 분석에 강점이라는 점이 그것이다. 현재 우리나라의 국가 재정 정보는 디지털예산회계시스템(이하 dBrain+)에 적재 및 관리되고 있는데 예산순기에 따른 재정정보가 탑재되어 있는 시스템이다. dBrain+에 탑재된 정형, 비정형데이터가 곧 재정 빅데이터로 볼 수 있다면 이를 머신러닝 기법을 적용하여 당면한 재정현상 분석이나 각종 전망, 비정형데이터의 분류를 통하여 데이터기반 정책결정에 활용할 수 있는 방안을 모색해 볼 수 있을 것이다.

본 논의에서는 dBrain+에 탑재되어 있는 재정 데이터의 활용성 증진 및 다양성을 위해 머신러닝 기법을 중심으로 살펴보고자 한다. 특히 AI의 여러 분야 중에서도 머신러닝 기법의 적용 가능성을 탐색해 보고자 한다. 이를 위해 머신러닝 기법의 이론적 논의 및 국내외 공공부문의 데이터를 중심으로 머신러닝 기법 활용사례를 분석하고 디브레인 시스템상의 활용 가능한 데이터 현황을 분석해보고자 한다. 그리고 분석적용 사례로 사업별 불용유형 데이터를 텍스트 마이닝 알고리즘을 예시적으로 적용하여 정책적 시사점 및 활용 가능성을 분석해보고 분석기법의 유용성을 논의해 보고자 한다.

2. 분석 내용

본 보고서는 dBrain+에 탑재된 방대한 분량의 재정데이터 활용의 확장가능성을 모색해 보고자 하는 목적을 가지고 있다. 재정데이터 활용의 확장 가능성 모색을 위해서 민간 및 공공영역에서 적극적으로 채택이 이루어지고 있는 인공지능 기술 중 머신러닝 기법을 중심으로 활용 가능한 분석을 논의해 보고자 한다.

이를 위해 본 보고서에서는 먼저 인공지능-머신러닝-딥러닝으로 이어지는 개념을 살펴보고 그중에서도 본 분석에서 활용될 머신러닝의 여러 알고리즘에 관한 이론적 논의를 분석적 측면을 중심으로 논의한다. 이론적 논의 이후에는 실제로 공공부문을 중심으로 국내외 머신러닝 기법 실제 활용사례를 분석하여 재정데이터에 활용될 수 있는 방향에 대해 논의한다. 다음으로 분석 대상인 dBrain+ 데이터의 특성에 대해 논의하고 앞서 살펴본 머신러닝 알고리즘을 적용하기 위해 전제되어야 할 데이터 측면에서의 선결조건을 제시한다. 그리고 논의를 바탕으로 활용 가능성 제시를 위해 실제 dBrain+ 데이터를 대상으로 텍스트 분석을 예시적으로 실시하여 가능성을 탐색해 보고자 한다. 이러한 과정을 통해서 향후 머신러닝 기법을 적용한 dBrain+ 데이터 분석이 활용될 수 있는 가능성과 dBrain+ 시스템 발전에도 기여할 수 있는 시사점을 제시해보고자 한다.

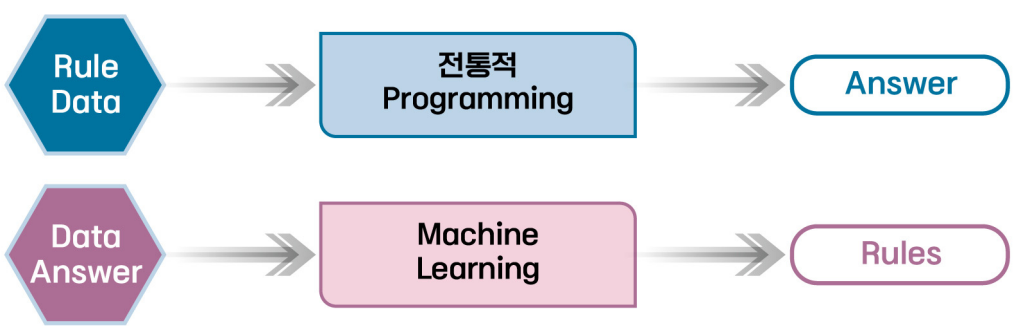
Ⅱ. 머신러닝 기법의 공공부문에서의 활용

1. 머신러닝 기법 개요

가. 머신러닝 기법의 정의

머신러닝은 일반적으로 AI[2] 기술의 알고리즘을 기반으로 프로그래밍하지 않아도 스스로 데이터로부터 규칙을 학습하는 기술이다. 여기에 더 나아가 신경망 학습을 기반으로 제공되는 학습 데이터 없이도 스스로 학습을 하여 모델링하는 기법은 딥러닝이라 받아들여지고 있다. 최초의 머신러닝 시스템이라고 언급되고 있는 체커스(Checkers)의 개발자인 Arthur Samuel은 머신러닝(merchine learning)이란 "컴퓨터에게 명시적으로 프로그래밍하지 않고도 학습을 할 수 있는 능력을 주는 연구 분야"로 정의 하였다(차경엽, 2022).

〈그림 Ⅱ-1〉 머신러닝과 전통적 방법론과의 차이



자료 : 저자작성.

머신러닝에서 분석 모델은 데이터와 학습 알고리즘으로 구성되어 있는데 모델은 데이터로 학습이 완료된 알고리즘이라 할 수 있으며 학습이 완료된 상태로 계수가 채워진 방정식이 되므로 의사결정 정보로 활용이 가능하다. 그리고 알고리즘이란 수학에

2) 인공지능(Artificial Intelligence)이란 사람의 지능을 모방하여 사람이 하는 것과 같은 복잡한 일을 할 수 있는 기술을 의미함

서의 수식을 의미하는데 수식 그 자체만으로는 의사결정을 할 수는 없기 때문에 모델의 데이터를 입력하여 활용하는 것이다. 그러므로 모델은 입력에 대한 출력이 피드백이 되는 방식이고 그 출력값을 의사결정에 사용하게 되는 원리이다. 머신러닝은 개별적 요인이 복합적으로 결합되어 나타나는 현상이 각각 어떤 확률을 가지고 미래에 발생할 수 있는지를 계산한다. 즉 Bayes 이론에 따라 각 사건이 함께 발생했을 때의 확률과 각 요소의 조건이 충족되었을 때의 조건부 확률을 계산한 후, Markov Decision Process 등과 같은 확률집합함수를 통해 미래에 발생할 확률을 계산한다. 인간의 학습에 비해 머신러닝이 가지는 가장 큰 장점은 매우 방대한 양의 데이터를 대상으로 학습이 가능하다는 점이다.

전통적인 통계적 모델링에 기반한 추정 방식은 이론을 토대로 연구자가 설정한 변수를 모형에 채택하는 방식이다. 이러한 기존의 방식은 표본을 분석대상으로 하여 가장 모수에 가깝게 추정하기 위한 변수를 이론 및 경험칙에 의해 구성하여 추정하는 것으로 방법론의 신뢰성이 오랜 시간 축적되어 있다는 점에서는 장점이다. 반면 머신러닝 기법은 추론에 필요한 다양한 변수들의 조합이 가능한 경우를 모두 계산하여 가장 적합한 모델을 찾아주는 방식으로 볼 수 있다. 물론 머신러닝 기법도 알고리즘의 기반은 전통적 통계학에서 유래된 것이 다수이지만 모수추정에 있어 가장 적합한 모델을 가능한 모든 경우의 수에서 채택하는 것이다.

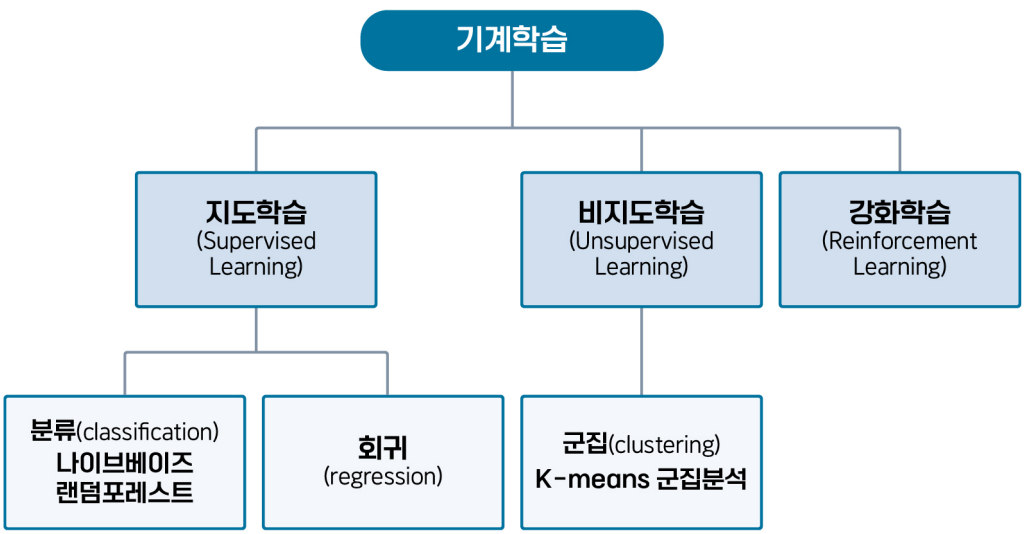
나. 머신러닝 알고리즘 분류

머신러닝은 일반적으로 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 크게 나누어진다[3]. 지도학습은 복수의 입력값을 이용하여 결과값을 예측하는 모델을 개발하는 것이다. 데이터셋에서 일부를 학습데이터(training data)로 지정하여 학습을 진행시키고 나머지 데이터를 테스트샘플로 하여 결과를 예측하여 최적의 모델을 추정하는 방식이다. 예를 들어 은행에서 여신 관련 데이터를 활용하여 대출연체 위험 확률을 계산하여 고위험 연체자를 선별해보고자 한다면 우선 해당

3) 일부에서는 보상과 강화라는 딥러닝 중심의 강화 학습으로 세 가지로 분류하기도 하지만 일반적으로는 지도학습 비지도학습 두 가지 분류로 보아도 무방함

확률 함수는 $f(X_i)$로 수식화할 수 있다. 여기서 X_i는 대출실행자 i의 특성변수로 나이, 직업, 소득 등 인적변수 및 과거 대출연체건수, 소득대비 대출금액비, 자산규모 등의 개인 금융정보 등이 해당된다. 지도학습 알고리즘은 금융, 경영, 의학 등 예측이 필요한 산업 영역에서 활발히 활용되고 있으며 최근에는 공공정책의 전망, 예측에 활용이 시작되고 있다. 지도 학습은 결과값이라고 할 수 있는 라벨(label) 데이

〈그림 Ⅱ-2〉 머신러닝 알고리즘 분류



자료 : 저자 작성.

터가 존재하는 것으로 분석대상 중 일부 데이터를 학습데이터로 하여 모형을 학습시킨다는 점이 비지도 학습과의 가장 큰 차이점이 된다. 대표적인 지도학습을 활용한 것은 이상 거래 탐지, 질병 구분 등에서 활용될 수 있는데 사전적으로 정의된 질병 및 이상거래 유무를 '그렇다', '아니다'로 라벨링 하여 부여한 것이다. 그리고 분석 대상 데이터에서 '그렇다' 집단과 '아니다' 집단의 각각의 특성과 패턴 등을 후술할 알고리즘을 적용하여 질병 및 이상거래 위험도가 높은 사례를 구분해 내는 활동이다.

비지도학습은 목표값은 없고 데이터만 입력되었을 때 스스로 연결 가중치들을 패턴화하여 분류하는 방법이다. 주어진 입력 패턴 자체를 기억하거나 유사한 패턴을 군집화하는 방식으로 각 특성변수별 거리 유사도나 패턴 유사도로 선을 그어 군집화하는

것이다. 비지도 학습은 지도학습과는 다르게 학습데이터가 없기 때문에 테스트데이터에서의 정답이 존재하지 않고 그 활용 목적도 알고리즘을 통해 데이터를 탐색을 통해 패턴 및 특성을 파악하는 것이 목적이라 할 수 있다.

강화학습(reinforcement learning)은 "학습을 시키는 과정에서 기계가 취하는 행동에 서로 다른 보상을 제시하여 보상을 가장 많이 받을 수 있는 방식이 무엇인지 스스로 학습하도록 하는 방식이다[4]". 흔히 알고 있는 강화학습 사례로는 딥마인드의 알파고를 들 수 있다. 바둑이나 체스의 경우처럼 주어진 상태에서 다음 스텝으로 가능한 경우의 수가 너무 많은 상황에서 정해진 답이 없이 스스로 가장 최적의 답을 구해보는 알고리즘을 의미한다. 강화학습은 이처럼 특정 분야를 중심으로 논의되고 있고 학습이 이루어진다는 점에서 일종의 지도학습 영역이기도 하기 때문에 머신러닝 알고리즘의 대분류에서는 흔히 지도학습/비지도학습 으로 한다. 본 논의에서도 지도학습과 비지도학습 등 두 영역으로 구분하여 살펴보고자 한다.

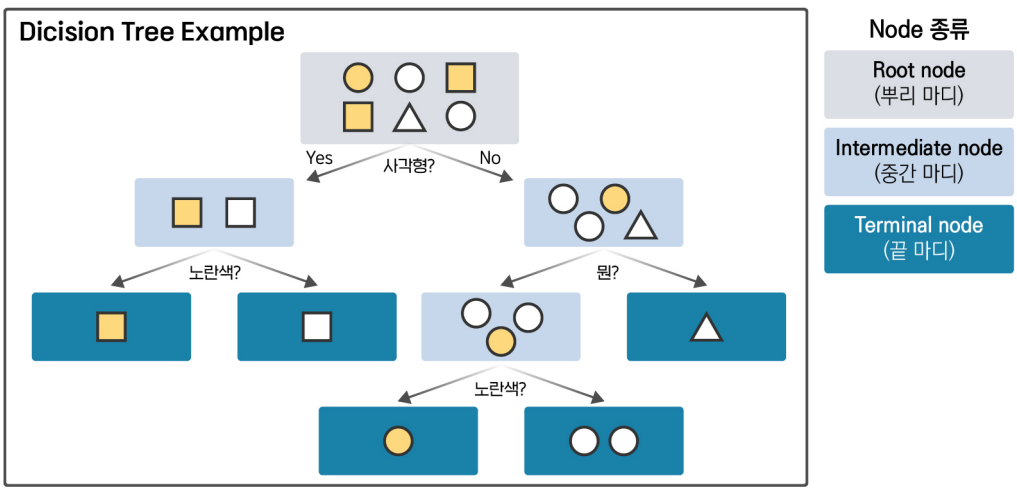
2. 지도학습

가. 의사결정나무

의사결정나무 알고리즘은 지도학습 중 분류에 속하는 방법론으로 기본적인 구조는 전체 집단을 노드로 계속 yes or no로 양분하여 유형화하는 방식이다. 분류모형 이란 몇 가지 대상의 카테고리 별로 라벨링이 정의되어 있을 때 제시된 대상을 어느 카테고리에 해당하는지 분류하는 것을 의미한다(이재호 외, 2019). 이는 다시 알고리즘에 따라 kNN(k-Nearest Neighbor), 서포트 벡터 머신(Support Vector Machine), 의사결정나무(Decision tree) 등이 해당한다.

4) https://terms.naver.com/entry.naver?docId=6653727&cid=69974&categoryId=69974

〈그림 Ⅱ-3〉 의사결정나무 모형 개념도



자료 : 건강보험심사평가원(2021). p.86.

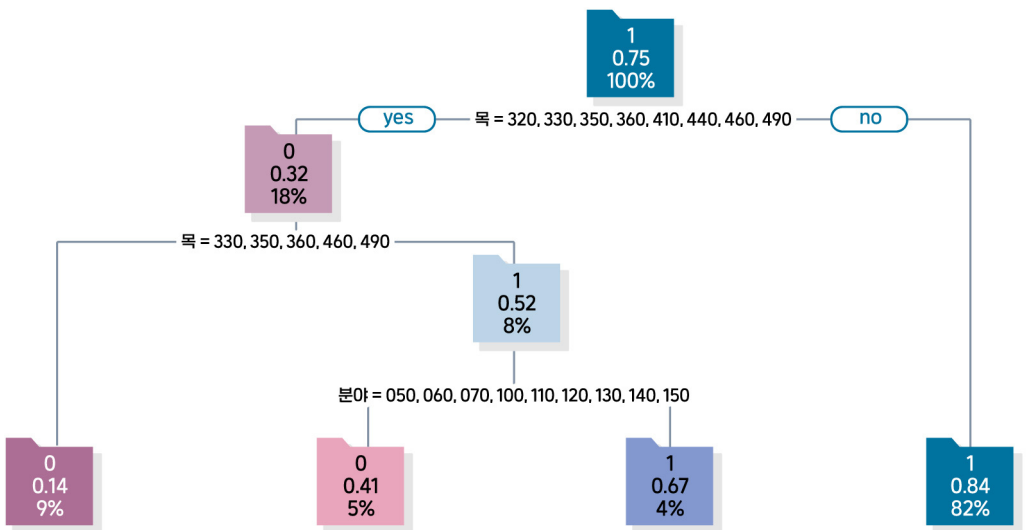
〈그림 Ⅱ-3〉에서 볼 수 있듯이 분기가 발생하는 노드에는 기준이 되는 조건문이 있으며 조건에 부합 여부에 따라서 다음 노드 이동방향이 정해지게 된다. 의사결정나무 모형은 분류와 회귀예측 모두 가능한 알고리즘인데 분류나무 모형인 경우에는 이산형 자료 값을 예측하므로 주로 범주형 레이블을 가진 데이터에 적합하다. 각 노드별 조건문에 대해 yes or no의 방식으로 설정된 집단의 수까지 도달하거나 충분히 분류가 이루어질 때까지 반복된다. 이렇게 분류될 각 그룹은 그 특성을 규명함에 있어 역으로 노드들의 기준 질문을 추적하면 알 수 있게 된다. 이를 응용하여 구성요인이나 특성을 규명할 수 있다. 그리고 회귀나무모형은 연속적인 값을 목적으로 하는 분석으로 특정 노드에 속한 데이터들의 평균값으로 예측값을 산출하게 되는데 이 경우에는 연속적인 수치형 변수의 특성을 가진 데이터 분석에 적합하다. 이러한 회귀형 의사결정나무 알고리즘을 활용하여 주택가격 예측, 환율예측, 고객신용정보 분석, 각종 질병 예후 예측 등에서 활용된다.

〈그림 Ⅱ-4〉는 재정 데이터에 예시적으로 의사결정 알고리즘을 활용한 것으로 집행잔액이 발생한 사업의 특성을 분류한 것이다[5]. 간략한 분석 모델링에 기반한 결과로 불용사업의 발생을 결정하는 데에는 첫 번째 노드인 특정 “목” 항목일 경우에서 해당

5) 예시적으로 간략한 형태로 분석한 것이기 때문에 실제 결과와는 차이가 있음

하면 다음 노드로 분기하며 집행잔액이 발생할 확률이 높은지 낮은지를 판별하는 것을 가능하게 한다. 이 모형에서는 상위노드에서는 특정 “목”에 주요 요인으로 하위 노드에서는 특정 “분야”가 주요 요인으로 나온다는 것을 의미한다.

〈그림 Ⅱ-4〉 의사결정나무 알고리즘 적용 예



자료 : 저자작성.

다만 의사결정나무 알고리즘은 가장 기본적인 머신러닝 형태로 단독으로 활용할 경우 학습 데이터 이외의 실제 검증 데이터에서 성능이 떨어지는 과적합(over-fitting) 문제가 나타날 수도 있다. 이를 극복하기 위한 방안으로 분석에서 하나의 알고리즘만 활용하는 것이 아닌 다른 알고리즘과 결합하여 활용하는 앙상블 모델을 결합하는 방식도 있다. 후술할 랜덤포레스트 알고리즘 등이 대표적인 앙상블 모델이다.

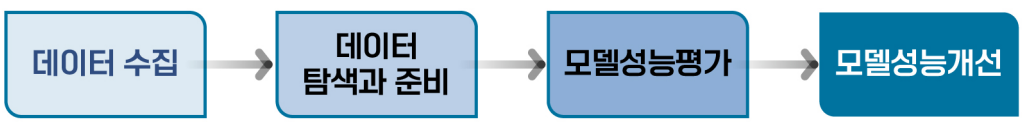
나. 나이브베이즈

나이브 베이즈 분류 기법은 지도학습 기법 중 분류에 해당하는 알고리즘 기법으로 전통적 통계학인 베이지안 기법[6]에 기반하고 있다. 베이지안 기법 기반의 분류 알고리즘은

6) 18세기 수학자 토마스 베이즈의 연구에서 유래한 것으로 사건의 확률과 추가 정보를 고려했을 때 확률에 어떤 식으로 영향을 미치는지를 설명한 원칙.

훈련데이터에서 관측된 값을 기반으로 결과가 나타날 확률을 계산하는 방식으로 한다. 이러한 분류 알고리즘에 따라 라벨이 부여되지 않은 데이터에 적용했을 때, 결과가 관측될 확률을 이용해서 새로운 특징에 가장 적절한 클래스를 예측하여 분류하는 방식인 것이다. 가장 쉬운 예로는 스팸 이메일 분류 기능이나 전산시스템 관리에서 비정상 접근 시도를 구분하는 것이 그것이다. 즉 이 알고리즘은 결과에 대한 전체 확률을 추정하고자 하는 동시에 여러 속성 정보를 고려해야 하는 문제에 적합도가 높다고 할 수 있다.

〈그림 Ⅱ-5〉 나이브베이즈 알고리즘 분석 과정



자료 : 저자 작성.

나이브베이즈 기법은 텍스트 분류도 가능하다는 점에서 텍스트 마이닝과도 이어진다. 텍스트 형태의 원시자료는 비정형데이터이지만 전처리와 알고리즘 등을 활용하여 기계학습에 활용할 수 있는 정형데이터화하여 분석에 활용할 수 있는데 나이브 베이즈 기법은 그 중 텍스트 데이터를 분류에 활용할 수 있는 기법이다.

〈표 Ⅱ-1〉 나이브 베이즈 기법의 장단점

	내용
장점	개념이 단순하고 계산 속도가 빠름 고차원의 데이터셋에 적합함 데이터에 노이즈 및 결측값이 포함되어 있어도 분석가능 알고리즘의 단순함에도 복잡한 분류 알고리즘보다 예측 결과가 더 효과적일 수 있음
단점	범주 분류 문제에 적합하지만 예측된 범주의 확률값을 활용해야 할 경우에는 적합하지 않음 독립변수들이 범주 형태가 아닌 수치 형태일 경우에는 정확성이 떨어짐 독립변수가 서로 독립적이고 중요도가 같다는 가정이 충족되지 않을 경우에는 분석결과에 오류가 발생할 수 있음

자료 : 장원중 이정인(2022). p316.

나이브 베이즈의 확률적 추론 방법은 어떤 가설의 확률을 평가하기 위해 임의적으로

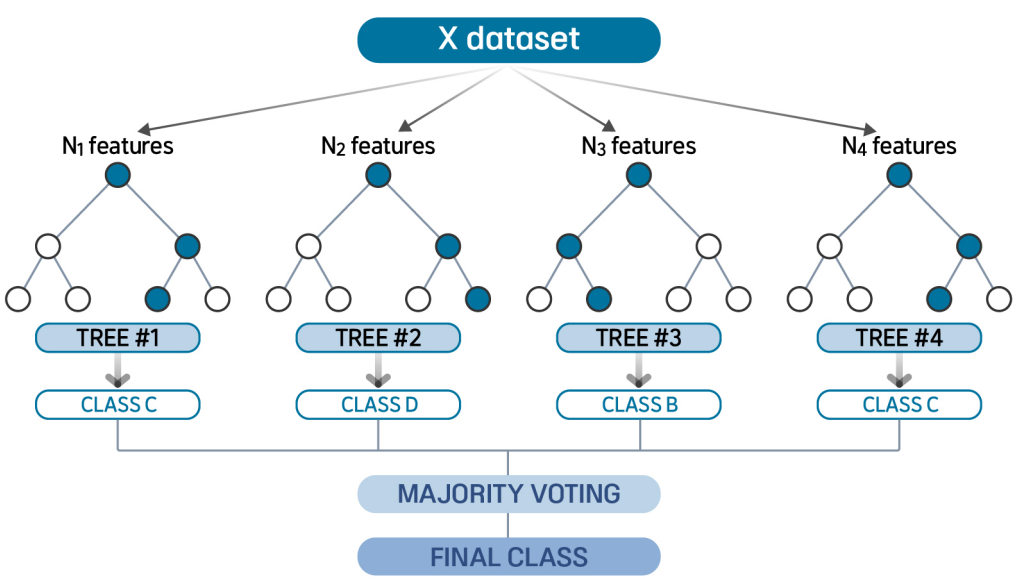
사전확률을 먼저 정하고 관찰된 데이터를 기반으로 하는 가능도를 계산해서 처음에 설정한 임의적 확률을 보정하는 방법인 것이다.

다. 랜덤포레스트

앞서 논의한 의사결정나무 알고리즘은 단독으로 활용하면 학습데이터 이외의 데이터에서는 그 적합성 및 성능이 떨어지는 과적합 문제가 발생할 수도 있다. 이러한 과적합 현상을 줄이기 위해서 나온 것이 앙상블 모델인데 배깅(bagging)과 더불어 랜덤포레스트 알고리즘이 그것이다.

배깅은 학습 데이터의 일부를 복원 및 추출하여 복수의 학습 데이터 샘플을 생성하여 각각의 샘플링을 통해 다수의 의사결정나무 알고리즘을 독립적으로 학습시키는 알고리즘을 의미한다. 예측결과를 결합함에 있어서는 학습이 완료된 여러 개의 의사결정 나무들을 대상으로 회귀 문제의 경우에는 평균값을 사용하고 분류 문제의 경우에는 투표 방식을 적용하여 최종적인 결과를 도출하는 방식이다.

〈그림 Ⅱ-6〉 랜덤포레스트 알고리즘 개념도

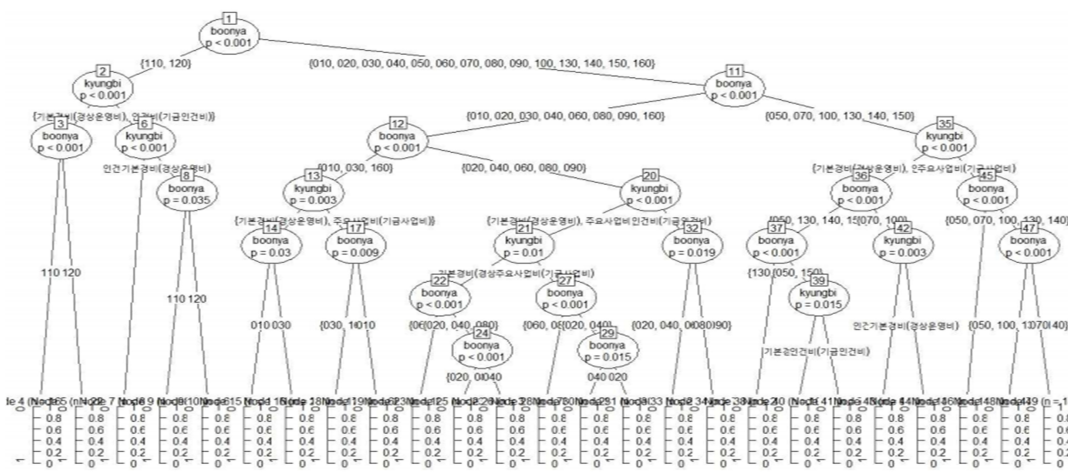


자료 : https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/

랜덤 포레스트(Random forest) 알고리즘은 배깅을 더욱더 확장시킨 알고리즘으

로 주어진 데이터 셋에서 무작위 샘플을 추출하고 추출된 모든 샘플에 대한 의사결정 트리를 구축하여 각각의 모든 의사결정나무로부터 예측값을 계산한 뒤, 〈그림 Ⅱ-6〉에서 볼 수 있듯이 예측 결과 값에 대한 평균값들과 비교하여 최종 예측결과 중 최적의 결과를 선택할 수 있게 하는 방법론이다. 이를 통해 데이터의 분포를 보다 다차원의 관점에서 학습하는 과정에서 과적합을 방지하게 된다. 의사결정나무를 결합해서 랜덤 포레스트를 만들고 각각의 나무에 대한 예측을 생성하는 것으로 다른 방법에 비해 적용 및 훈련시간이 단축되고 많은 양의 데이터가 누락된 경우에도 정확도가 유지된다는 장점이 있다(최은진, 2021). 해당 알고리즘 방법론은 다수결의 원칙에 따르는 방식으로 의견의 통합 또는 여러 다수의 결과를 취합하는 방식으로 하나의 거대한 결정 트리를 생성하는 것이 아니라 다수의 하부 결정 트리를 여러 개로 만드는 것이다(최은진, 2021).

〈그림 Ⅱ-7〉 랜덤포레스트를 적용한 분석 예



자료 : 저자작성.

〈그림 Ⅱ-7〉은 랜덤포레스트 알고리즘을 적용하여 재정데이터를 활용한 분석을 예시적으로 보여주는 것으로 가장 높은 노드에서 끊임없이 분기하여 최종 결정요인이 무엇인지에 관한 회귀와 분류분석을 모두 적용한 것이다. 이를 활용하면 분석목적에 따른 사업분류가 가능하며 결정요인의 우선순위 등도 분석이 가능해진다. 다만 프로그

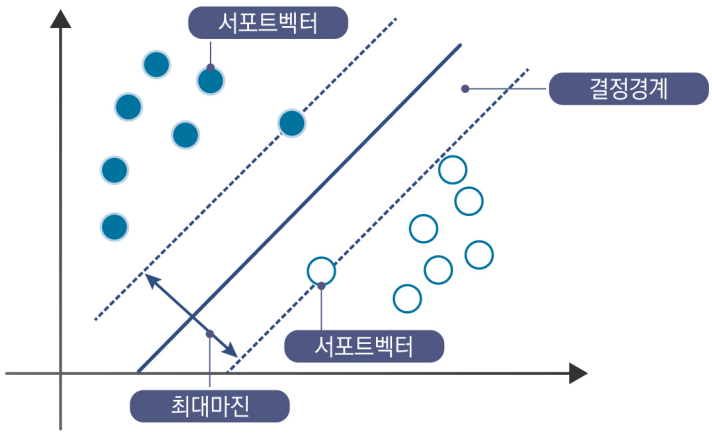
램 예산 체제하에서의 활용 시에는 분야-부문-프로그램-단위사업-세부사업으로 위계적 구조를 가지고 있는 데이터의 특성을 고려해야 하며 각각의 범주 개수가 최소한으로 요구되는 케이스 개수 이상으로 나타나야 알고리즘 적용이 가능해야 함을 유념해야 한다.

라. 서포트벡터머신(SVM)

서포트벡터머신 기법은 앞서 논의한 거리 기반의 최근접 이웃 분류와 유사하지만 최근접 이웃분류가 가까운 정도를 기준으로 분류하는 것에 비해 서포트 벡터 머신 기법은 서로 다른 분류에 속하는 데이터 간의 거리(margin)가 최대가 되는 선을 기준으로 데이터를 분류하는 모델이다. 즉 데이터를 선형으로 분리하는 최적의 선형 결정 경계를 찾는 기법이다.

두 범주 간의 데이터를 같은 간격으로 최대로 멀리 떨어진 선 또는 평면을 찾는 것으로 두 범주 간의 데이터를 나누는 직성 혹은 평면은 여러 개가 있을 수 있지만 제시된 학습된 데이터가 아닌 미래의 데이터를 분류 예측하는데 최대한 일반화 할 수 있도록 하는 최대 여백 초평면을 찾고자 하는 것이다. 여기서 서포트 벡터는 이 경계선과 가장 가까운 각 분류에 속한 점들을 의미한다. 각 분류는 최소 하나 이상의 서포트 벡터를 가지고 있다. 〈그림 Ⅱ-8〉은 서포트 벡터 머신 개념을 도식화한 것으로 결정경계에 가장 인접한 흰점과 검은점이 각 집단의 서포트 벡터이며 이를 기준으로 두 집단을 가르는 결정 경계를 지지(support)한다는 것을 의미한다.

〈그림 Ⅱ-8〉 서포트벡터머신 개념도



자료 : 저자작성.

서포트 벡터머신 기법은 분류와 수치 예측 문제에 모두 활용할 수 있는데 분류 성능이 우수하면서 과적합(overfitting)이 잘 발생되지 않는다. 그리고 일반화 능력이 우수해 정교한 분류 성능이 필요한 유전체 데이터 분류, 보안결함, 언어식별, 이상거래 탐지 등 다양한 분야에 활용되고 있다. 서포트 벡터머신 기법은 최대 마진을 찾는 과정을 통해 일반화할 수 있는 성능 향상을 스스로 한다는 장점이 있으며 이를 통해 과적합을 줄이면서 새로운 데이터에 대한 예측 성능을 높일 수 있다. 그리고 이 알고리즘은 비선형 문제를 선형 문제로 변환하여 해결할 수 있는 장점이 있는데 대부분의 현실을 반영한 데이터가 선형 구조가 아닌 복잡한 형태인 비선형 구조에 기인한다는 점을 보면 그 장점이 명확하다.

〈표 Ⅱ-2〉 서포트벡터 머신 기법의 장단점

	내용
장점	범주 분류나 수치 예측 문제에 모두 활용이 가능함 노이즈 데이터에 영향을 많이 받지 않아 과적합이 잘 나타나지 않음 일반적으로 분류 분야에서 타 알고리즘에 비해 성능이 우수한 것으로 알려짐 분류 경계가 복잡한 비선형 문제일 경우 타 기법 대비 성능이 우수
단점	최적 분류를 위해 커널 함수와 매개변수 등에 대한 반복적인 조합 테스트가 필요함 입력 데이터가 대량이거나 변수가 많은 경우 오랜 학습시간이 소요 배경이 되는 이론과 알고리즘 구현 시 타 기법에 비해 상대적으로 난해함 결과 해석이나 설명에 있어 어려움 존재

자료 : 장원중 이정인(2022). p316.

반면 서포트 벡터 머신 알고리즘이 가지는 단점은 첫 번째로는 하이퍼 파라미터 튜닝으로 최적 분류를 위해서는 커널 함수, 매개 변수 등의 하이퍼파라미터에 민감하게 반응하기 때문에 초기에 이를 설정하는 튜닝과정이 복잡하다는 것을 의미한다. 초기에 튜닝이 제대로 이루어지지 않으면 그로 인한 분석결과 역시 적합하다고 말하기 어렵다. 두 번째로는 느린 학습속도와 그로 인한 분석에 소요되는 시간이 길다는 점이다. 이 알고리즘은 학습 과정에서 최적화 문제를 해결해야 하므로 실제로 대용량의 데이터셋이나 실시간 및 다 채널에서 유입되는 데이터에서 활용할 경우에는 속도가 느리며 시간이 오래 소요된다. 세 번째로는 3차원 공간에서의 결정경계를 찾는 과정도 가능하기 때문에 여기서 도출되는 결과를 직관적으로 해석하기 어려울 수도 있다.

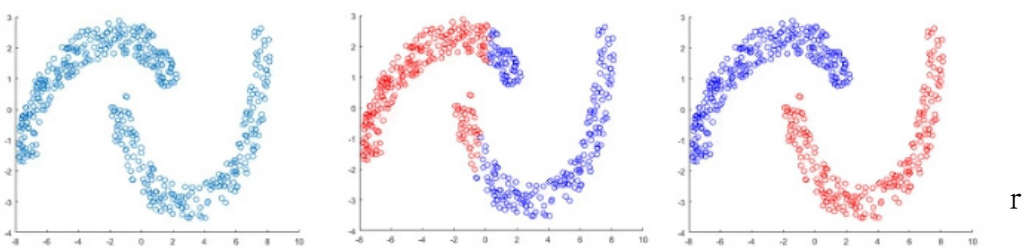
서포트 벡터 머신 알고리즘은 머신러닝 기법 중 비선형 문제에도 특화되어 있다는 점과 분류와 회귀 문제 모두에 적합한 알고리즘이라는 점에서 활용성이 높은 기법 중 하나이다. 앞서 논의된 알고리즘의 장단점을 고려하되 적절한 상황에 활용 시 이미지 인식, 텍스트 분류, 손글씨 인식 등의 분야에서 활용이 이루어지고 있다. 이러한 점에서 재정 데이터에서도 재정사업설명자료의 텍스트 분석을 통한 수혜집단별 분류와 과거의 데이터를 학습시켜 집행 실적을 테스트하여 회계연도 말에 집행잔액 발생의 유무를 예측해 볼 수 있을 것으로 생각해 볼 수 있다. 다만 복잡한 해석의 문제를 극복하기 위해서는 앞서 논의한 의사결정나무와 같은 다소 직관적인 머신러닝 알고리즘과 결합한 앙상블 기법을 활용하는 방식을 활용할 수 있다.

3. 비지도학습

앞서 논의한 지도학습과는 다르게 비지도학습은 학습시킬 정답이라고 할 데이터가 존재하지 않으며 해당 데이터의 패턴, 특성 및 구조를 스스로 파악하여 새로운 규칙성을 규명해가는 기법을 의미한다. 즉 입력값에 대한 목푯값이 없다. 이러한 비지도학습은 데이터들 속에서 유의미한 변수를 추출하거나 데이터의 숨겨진 구조 및 특성을 규명하는데 적합하다.

비지도학습은 데이터를 군집화하는 방법론을 의미하는데 지도학습과의 가장 큰 차이는 직관적으로도 지정된 학습데이터가 불필요하며 알고리즘에 따라 각 샘플별로 거리 및 유사도에 따라 분류하는 것이다. 비지도학습의 목적은 무엇을 예측하는 것이 아니라 입력 데이터에서 어떤 특징을 찾기 위함이다. 그러므로 미리 설정된 label이 없다는 점을 제외하면 군집모델은 지도학습의 분류 모델 알고리즘과 그 목적이 동일하다고 볼 수 있다(이재호 외, 2019). 분류 알고리즘에 따라 중심기반 알고리즘

〈그림 Ⅱ-9〉 분류 알고리즘 결과 비교



원본데이터 K-means clustering 결과 밀도 기반 알고리즘 결과

자료 : https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/

(Center based algorithm)과 밀도 기반 알고리즘(Density based algorithm)으로 크게 볼 수 있다. 중심기반 알고리즘은 동일 군집에 속하는 데이터는 어떠한 중심을 기준으로 분포하게 된다는 가정 아래, 데이터와 군집 간의 거리를 기초로 군집을 형성하는 알고리즘이다. 이러한 중심기반 알고리즘의 대표적인 방법론으로는 K-means, 군집 알고리즘 등이 있다. 여기에서 응용된 연관패턴 분석은 비지도학습의 하나의 알고리즘인데 우리가 가장 쉽게 접할 수 있는 예시로는 온라인 쇼핑에서 특정 상품을 골랐을 때 연관된 상품을 추천해주는 알고리즘이나 유튜브와 같은 콘텐츠 이용 시, 그와 연관된 혹은 다른 이용자가 함께 이용한 콘텐츠를 제안해서 보여주는 것을 들 수 있다.

군집 알고리즘은 비슷한 샘플끼리 그룹으로 모으는 작업을 군집(Clustering)이라고 하며 밀도 기반 알고리즘은 동일한 군집에 속하는 데이터는 서로 근접하게 분포한다는 가정에 기반한다. 중심 기반 알고리즘은 클러스터의 중심을 기준으로 가까운 거리에 위치한 데이터들은 그 클러스터에 분류시키기 때문에 원의 형태로 클러스터가 형성된

다. 그리고 밀도 기반 알고리즘은 서로 이웃한 샘플들을 같은 클러스터로 분류하기 때문에 불특정한 형태의 클러스터로 군집되게 된다. 일반적으로 가우시안 분포나 정규분포가 아니라 불특정한 분포를 가지는 데이터에 관해서는 밀도기반 클러스터링 분류가 적합하다. K-means 군집은 가장 군집분류 알고리즘의 대표적인 것으로 사전에 결정된 군집 수 K에 기초하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 구분하는 방법이다(안찬식·오상엽, 2021). K-means 군집은 클러스터 개수 K값 결정-데이터 공간에 클러스터 중심 K개 할당- 각 클러스터 중심을 해당 클러스터 데이터의 평균값으로 조정 – 클러스터 중심이 변하지 않을 때까지 해당 과정 반복으로 수행된다(안찬식·오상엽, 2021).

비지도학습 알고리즘의 재정분야에서의 적용 가능성은 연관규칙 알고리즘과 분류에 기반한 알고리즘을 활용하는 것에서 살펴볼 수 있을 것이다. 연관규칙 알고리즘을 적용해 본다면 특정한 복지 사업을 택한 사람이 다음에 수혜받을 확률이 높은 다른 복지사업을 살펴보도록 제안하여 복지 사각지대를 줄이고자 하는 방식으로 접근해 볼 수 있을 것이다.

가. 시사점

지금까지 머신러닝의 주요 알고리즘을 지도학습, 비지도학습으로 구분하여 각각 분류되는 알고리즘에 대해 살펴보았다. 지도학습은 학습데이터와 테스트 데이터로 분리하여 투입값과 라벨 값이라고도 하는 결과 값을 모형으로 학습시켜서 최종적으로 테스트 데이터에 적용하여 분류분석이나 회귀분석을 수행하는 방식이다. 비지도학습은 학습데이터가 존재하지 않는 것으로 데이터에서 유사도와 이질성에 따라 유사한 집단의 데이터끼리 군집하도록 하는 알고리즘을 의미하는 것으로써 주로 이산적 분류와 같은 분류분석에 활용된다. 머신러닝에는 무수히 많은 알고리즘이 있고 지금도 생성되고 있지만 활용빈도가 높은 몇몇의 알고리즘을 장단점을 중심으로 서술하였는데 지도학습에는 의사결정나무, 나이브 베이즈, 서포트벡터머신(SVM), 랜덤포레스트 알고리즘에 혼합형인 앙상블 기법을 살펴보았다. 〈표 Ⅱ-3〉은 이러한 머신러닝의 알고리즘의 특성과 재정분야에서 적용 가능할 예를 정리한 것이다.

〈표 Ⅱ-3〉 재정분야에서의 알고리즘별 적용 가능성

	알고리즘 특성	재정분야에서 적용가능 예
지도학습	투입값에 대한 목표값을 학습시켜 예측이나 분류 수행	
의사결정나무	학습을 통한 분류 모형, 노드 별로 분기하며 같은 경로의 데이터끼리 분류	재정 사업 집행 패턴 유형 등
나이브베이즈	베이지안 확률기법에 기반한 알고리즘으로 조건에 따른 각각의 경우에 따른 예측 및 분류	텍스트 분석, 이상집행 감지, 시스템 이상 접근 탐지
서포트벡터머신(SVM)	복수의 집단이 차이를 보이는 선형 분리에 따른 분류와 회귀분석에 모두 활용 가능	재정문서 텍스트 분석, 부정수급 등 이상거래 탐지
랜덤포레스트	의사결정나무 알고리즘에서 확장한 것으로 각각의 예측결과 중 최적의 결과를 채택하는 알고리즘	수혜자 집단별 재정정책효과분석, 국고잔액전망, 세입 예측 등
앙상블(혼합형)	하나의 알고리즘이 아닌 복수의 알고리즘을 결합하여 설명의 직관성과 예측과 분류의 타당도 높이는 방법론	수혜자 집단별 재정정책효과분석, 재정위험요인 판별 등
비지도학습	데이터를 유사도와 이질성에 따라 학습 데이터 없이 군집화하여 분류	

자료 : 저자작성.

지도학습이나 비지도학습 모두 공통적으로 방대한 데이터셋을 대상으로 한 '분류'에 강점으로 가지고 있는 것을 볼 수 있다. 재정분야에 패턴 분류 방식이 적합한 것으로 재정집행 패턴 분석이 가능할 것으로 보이며 지도학습으로는 기존의 집행패턴을 결과값으로 놓고 분류하고 싶은 집행 추이를 패턴화 분류하는 방식으로 적용이 될 것이고 비지도학습으로는 기존 패턴에 대한 학습 없이 각 재정사업집행추이를 각각 유사도에 따라 비슷한 몇몇의 패턴으로 군집화하여 분류할 수 있을 것이다. 머신러닝이 가지는 분석 유형 중 분류에 이어 예측과 전망의 목적을 가지는 '회귀'에서는 랜덤포레스트나 앙상블모형, 서포트 벡터 머신 기법 등이 활용된다. 이를 활용하면 부정수급이나 이상집행과 같은 탐지, 감지 등이 예측 가능해지고 국고잔액예측, 세입예측 등의 전망에서

도 활용이 가능해 보인다. 이러한 활용을 위해서 다양한 변수와 속성정보, 시점별 재정 데이터 등이 확보된다면 충분히 적용 가능할 것으로 보인다.

4. 머신러닝 기법의 공공부문 활용사례

가. 국내 적용 사례

1) 공공부문 AI 활용 현황 실태조사

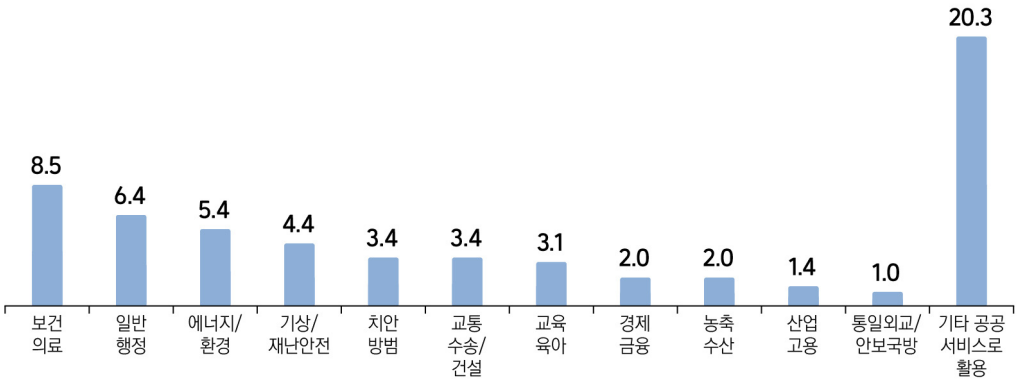
우리나라는 소프트웨어정책연구소에서 2022년도에 처음으로 “공공부문 AI 활용 현황 실태 조사”를 시행하였다. 머신러닝 기법이 AI 기술 중 하나임을 감안한다면 본 조사 결과는 국내 공공부문에서의 머신러닝 기술 활용에 관한 현황을 파악할 수 있는 자료로 볼 수 있다. 이 조사는 41개 정부부처, 17개 광역지방자치단체, 350개 공공기관을 대상으로 서베이 방식으로 수행되었으며 인공지능 기술, 활용현황, 인공지능 기술 도입 효과, 인공지능 기술 도입 관련 애로사항 및 정부지원, 인공지능 관련 인력, 인공지능 기술 도입 사례 및 파급효과 등의 항목이 포함되어 있다(남현숙 외, 2023).

조사 결과 응답기관의 55% (295개 기관)이 인공지능을 도입하여 활용 중인 것으로 나타났는데 정부부처가 80.0%, 광역지방자치단체 82.4%, 공공기관은 50.7%로 나타났다(남현숙 외, 2023). 활용분야로는 〈그림Ⅱ-4〉에서 볼 수 있듯이 기타 공공서비스로 활용이 가장 응답비율이 높았는데 여기에는 챗봇, 업무자동화(RPA[7]), 기관 내부 데이터 분석을 통한 조직진단 등이 포함된다.

7) 특정한 업무의 관리를 위해 구조화된 작업을 수행하는 소프트웨어 로봇을 의미하는데 노동자원의 행동을 벤치마킹하는 알고리즘과 규칙 배치 작업 등을 의미하기도 한다. RPA는 다른 형태의 인공지능에 비해 상대적으로 저렴하고 프로그래밍의 수월성 등으로 인하여 워크플로우 비즈니스 규칙, 정보 시스템과의 표현 계층의 통합을 통해 사람과 같이 작동한다. 활용 사례로는 각종 기록물의 주기적 갱신, 청구 및 지불결제 등의 단순하면서도 반복적인 업무 등이 있다. 여기에 이미지 인식 기술을 접목하여 택배 송장이나 상품 이미지와 같은 데이터화 하여 거래 시스템에 입력하여 자동으로 발부하는 업무 등에서도 활용 가능하다(강송희, 2021).

〈그림 Ⅱ-10〉 국내 공공부문 인공지능 도입 분야

[BASE: 인공지능 도입 및 도입 예정 기관(n=295), 단위: %, 중복응답]



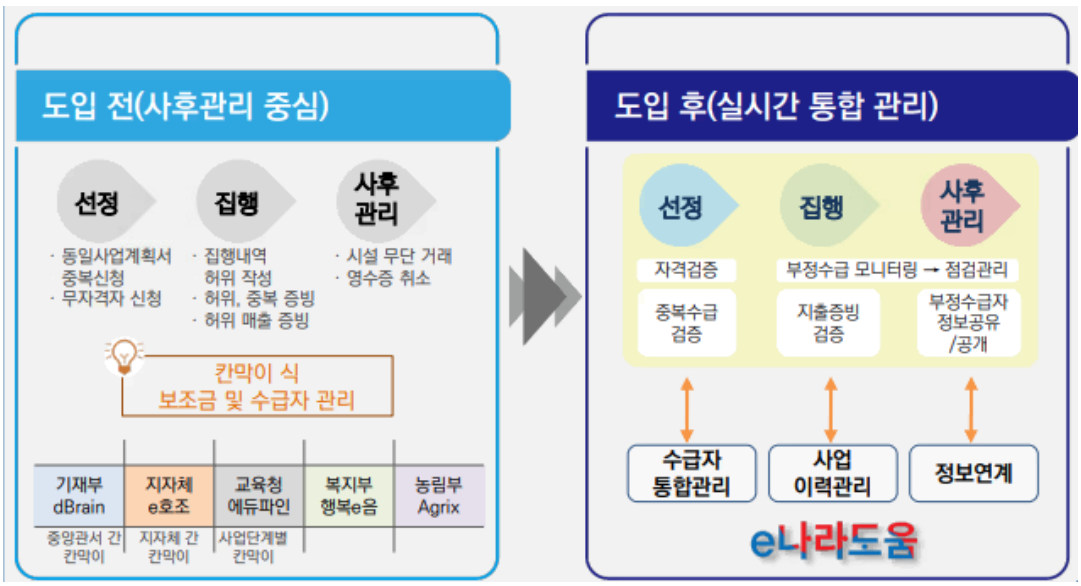
자료 : 남현숙 외(2023), p14

보건의료, 일반행정 ,에너지·환경 분야 순으로 도입이 이루어지거나 예정으로 응답된 것을 볼 수 있는데 세부적으로는 텍스트 분석과 같은 온톨로지 기반 기술을 통한 동향 분석이나 반복 응대, 음성인식, 수요예측과 같은 기술을 활용하는 것으로 보인다. 이 설문을 통해 국내의 공공분야에서도 머신러닝을 포함한 인공지능 기술의 도입의 활용은 이미 이루어지고 있으며 도입 분야도 거의 대부분 영역에서 이루어거나 예정인 것으로 시사점을 주고 있다.

2) 한국재정정보원 부정수급 탐지

한국재정정보원에서는 국고보조금의 전 처리 과정을 시스템으로 관리하여 보조사업을 원활하게 수행하고 보조금의 부정수급을 방지하고자 하는 목적으로 보조금 관리 시스템인 e나라도움을 운영하고 있다(보조금 관리에 관한 법 제26조의2). 이에 따라 보조금의 부정수급을 탐지하는 시스템을 구축하여 운영하고 있는데 여기에 일부 머신러닝 알고리즘을 활용하였다.

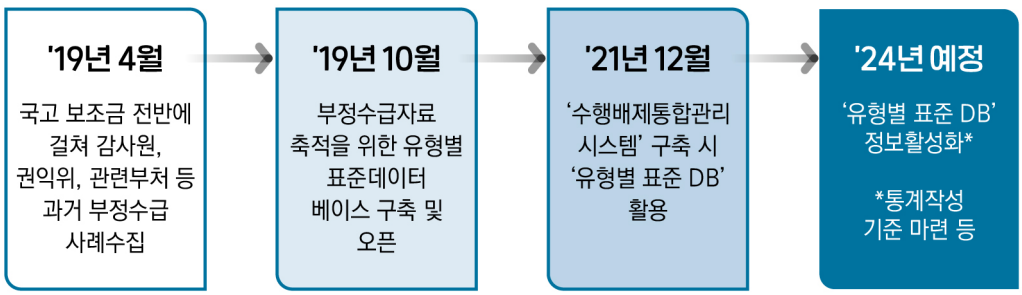
〈그림Ⅱ-11〉 부정수급 방지 프로세스 도입 전후 비교



자료 : e나라도움 사용자 매뉴얼(한국재정정보원, 2022). p17.

한국재정정보원에서는 부정수급 유형의 표준화 프로세스를 정립하기 위해서 〈그림Ⅱ-11〉과 같이 제도를 운영하고 있다. 부정수급과 관련한 정확하고 활용 가능한 통계의 양이 절대적으로 부족한 실정인데, 부정수급 탐지에 있어 머신러닝 기법을 활용하기 위해서는 이에 대한 자료 축적이 필수이다. 부정수급 탐지 시스템 구축은 다양한 형태의 부정수급 관련 정보가 각 중앙관서에 광범위하게 산재되어 부처별 대응방식이 상이하게 나타나게 되는 데에서 문제의식을 가지고 출발하였다. 2019년 4월에 감사원, 권익위, 관련 부처 등에서 부정수급 사례 수집을 하면서 본격적으로 구축되었다. 여기에서 최종 적발된 개인이나 기관은 차후 국고 보조금 사업에 참여할 수 없도록 수행배제 시스템을 구축하는 데까지 이어지고 있다.

〈그림 Ⅱ-12〉 국고보조금 부정수급 유형 표준화 추진 프로세스



자료 : 한국재정정보원 내부 세미나 자료. 저자 재작성.

국고보조금 부정수급 통계 자료는 e나라도움의 부정수급관리시스템에 등록되어 있다. 부정수급 데이터 현황은 '23년 7월 30일 기준으로 총 36.6만여 건이며 데이터 속성 정보로는 소관_회계_계정_분야_부문_프로그램_단위_세부_내역_세목_보조사업이며 사업연도, 예산액, 예산현액, 사업금액, 수행기관, 부정수급자 유형(1,2,3호), 대, 중, 소분류(60 유형), 환수결정액, 환수액 등이 포함되어 있다(한국재정정보원, 2023).

〈표 Ⅱ-4〉 부정수급 유형 분류

유형	내용
주요 빈발 유형	허위신청, 허위 출결, 가격 부풀리기, 정산서류 조작, 부적정 계약 거래, 자부담 회피·대납, 바우처카드 허위 결제, 목적 외 사용, 중요재산 임의처분, 기타, 행정착오 등
단계별 유형	선정, 집행, 사후관리
보조금법상 유형	허위수급, 다른 용도 사용, 법령 등 위반 및 요건 미충족, 중요재산 임의처분, 광의의 부정수급
행위주체별 유형	공무원, 보조사업자, 보조금 수령자, 제 3자 등

자료 : 기획재정부(2022). 저자 재작성. pp5-6.

부정수급 유형은 기획재정부에서 각 부처에 배포한 '국고보조금 부정수급 관리 가이드라인'에 따라서 〈표 Ⅱ-4〉에서 분류하고 있다. 이러한 부정수급 '유형별 표준DB' 시스템의 활용을 통해 얻고자 하는 목적은 부정수급 방지 정책 및 제도 개선 등에 활용하고자 하는 것이다. 〈표 Ⅱ-5〉는 2018년도부터 2022년도까지의 e나라도움에서 추출된 국고보조금 부정수급 환수[8], 의심[9], 적발 사례 건수 및 금액을 집계한 것이다.

〈표 Ⅱ-5〉 국고보조금 부정수급 관리 현황 *

(단위 : 건, 억 원)

		2018년	2019년	2020년	2021년	2022년
환수	건수	132,758	103,428	65,477	66,966	46,856
	금액	594.7	209.4	286.6	345.6	245.0
의심 (e나라도움)	건수	4,291	7,175	3,853	4,243	4,603
	금액	2,319	2,734	2,662	1,757	2,379
적발 (e나라도움)	건수	18	154	132	231	260
	금액	1.7	25.0	31.5	34.8	98.1

자료 : 국회예산정책처(2023). p14.

* 사업연도 기준으로 23년도 9월 추출 기준이며 기획재정부 원자료임

주무관청 등에서 인지 혹은 적발한 지급오류 사례가 상당수 포함된 2022년도 부정수급 환수 현황은 46,856건이고 금액으로는 245.0억 원으로 나타나고 있다. 여기에 e나라도움의 부정수급 탐지 시스템을 활용한 의심사례는 4,603건이고 금액으로는 2,379억원으로 집계되었다. 그 중, 최종적으로 부정수급 사례로 적발된 실적은 260건, 98.1억 원으로 나타났다. 의심 금액 대비 적발 금액은 약 4% 수준으로 이는 탐지시스템이 지나치게 과적합되어 민감도가 높거나 실제 선별된 의심사례를 대상으로 한 현장실사나 심층 조사와 같은 보완으로 타당성을 확보한 것이라고도 볼 수 있다. 향후 탐지 시스템 개선에 의심과 적발 사이의 괴리율을 좁힐 수 있도록 고도화된 탐지시스템 알고리즘 개선 및 패턴분석, 분석전문인력 배치 등이 필요할 것이다.

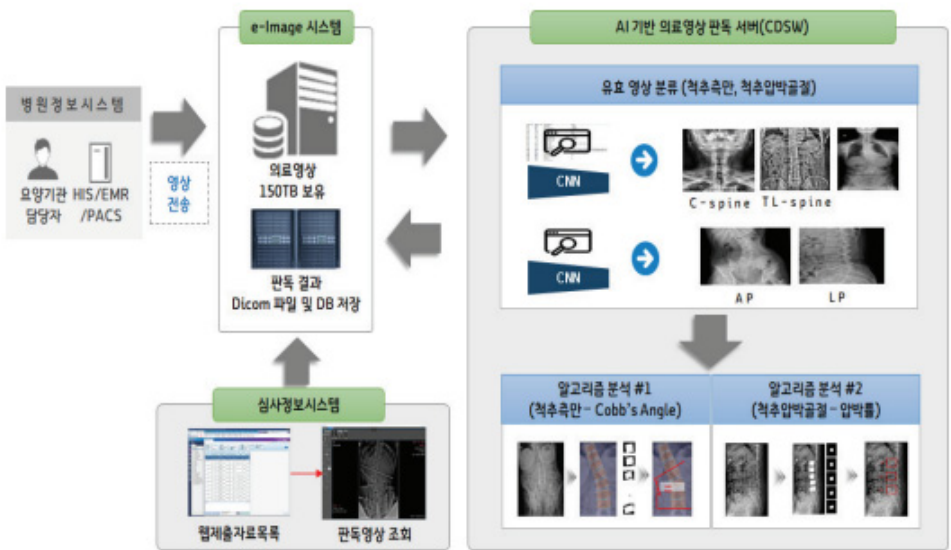
8) 주무관청 등에서 적발하여 환수한 전체 부정수급 집계치이며 지급오류(광의의 부정수급)를 포함한 통계자료임(예산정책처, 2023).

9) 의심 및 적발 사례는 기획재정부가 최근 5년간 e나라도움의 부정수급관리시스템을 통해 추출한 부정수급 의심 또는 적발사업의 건수 및 금액(예산정책처, 2023).

3) 건강보험심사평가원 판독활용 부당청구 탐지

건강보험심사평가원에서는 2018년부터 시범적으로 이미징 정보 중심의 의료영상심사판독시스템 도입 및 활용을 시작으로 분석심사를 위한 특성기관 예측, 급여정보 분석시스템 등으로 적극적으로 활용하고 있다.

〈그림 Ⅱ-13〉 의료영상심사판독시스템 흐름도

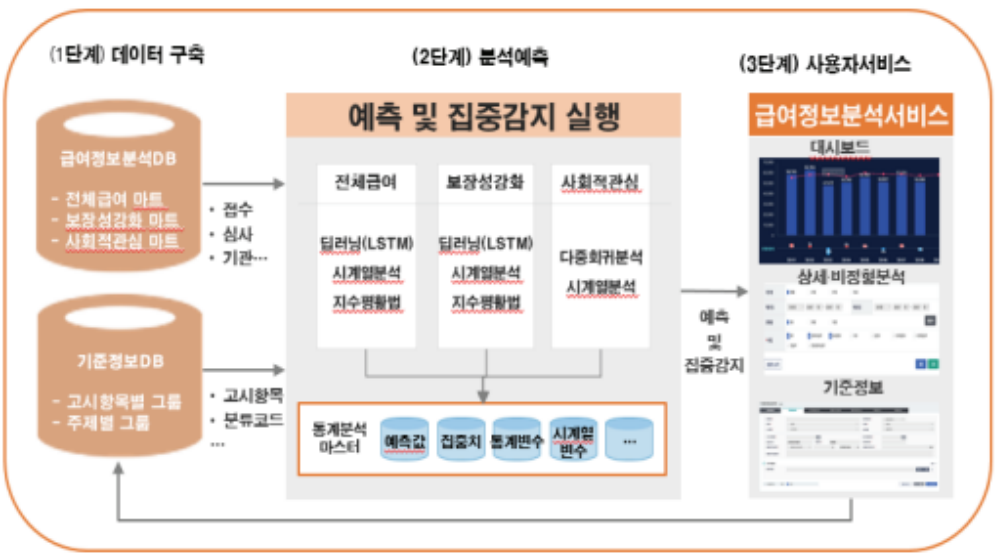


자료 : 건강보험심사평가원(2021), p141.

의료영상심사판독 시스템은 개별 급여기관에 저장되어 있는 영상정보를 활용하여 판독대상을 선별하고 심사단계에서는 심사연계 과정 중에 청구명세서 정보, 영상정보 등을 활용하여 상병 이미지를 인식하게 된다. 이를 통해 심사자들은 적정한 판독 여부와 급여 심사에 활용하게 되는데 완전한 활용은 아니고 심사자들에게 보조적 판단의 수단으로 작동하고 있다. 인공신경망 방법론을 주로 활용하고 있으며 영상정보와 같은 이미지 중심의 비정형데이터를 처리하는데 적합한 방식이다. 이를 통해 심사판독 업무의 효율성과 객관성을 기대할 수 있다(건강보험심사평가원, 2021).

두 번째로는 집중 심사가 필요한 요양기관을 선별 및 특성기관 예측에 활용하고 있는 것으로 요양기관을 대상으로 한 임상결과, 임상 및 비용에 대한 진료행태 및 진료추세 등을 파악하여 회귀분석을 통해 예측, 선별하는 모형이다. 총 진료비, 보험자 부담금 등을 대상으로 데이터 셋을 구성하고 학습하여 사용자에게 진료비 예측 정보 제공하는 알고리즘이다. 분석방법은 로지스틱 회귀분석으로 Hadoop 기반의 모델을 개발하여 급여정보 및 기준 정보를 활용하여 적정 급여 수준을 예측하고 실제 지급액과 차이가 클수록 관리대상으로 선별하는 방식이다(건강보험심사평가원, 2021).

〈그림 Ⅱ-14〉 급여정보분석 시스템 흐름도



자료 : 건강보험심사평가원 (2021), p154.

세 번째는 진료비와 보험자부담금을 대상으로 구축한 패널 데이터셋을 활용하여 진료비 예측정보를 제공하는 급여정보시스템이다. 기본적인 알고리즘은 시계열 분석 기법으로 1단계에서는 급여정보분석DB, 기준정보DB 등의 데이터 웨어 하우스에 축적된 데이터를 기반으로 예측모델에 활용할 입력값으로 확보한다. 2단계로는 보장성 강화, 전체 급여 등 업무 관점별로 예측을 실행하고 집중 탐지를 수행한다. 3단계로는 급여정보 포털에서 대시보드 및 상세 분석과 같은 분석 및 예측 결과를 조회할 수 있도록

하는 단계로 이루어진다. 급여정보 분석 시스템의 구현은 현재 한국재정정보원에서 dBrain+ 데이터를 기반으로 운영 중인 열린재정 포털에서의 통계분석 기능 제공과 유사한 방식인 것으로 보인다.

나. 해외 적용 사례

1) 미국 보건후생부 아동 학대 예측 시스템

미국 보건후생부(The Department of Health and Human Services)에서는 자국에서의 아동 학대에 의한 사망 및 부상과 같은 희생이 증가하면서 이를 미리 예측하고 스크리닝하여 선제적 개입이 가능한 시스템을 개발하여 운영 중이다. 위험요인을 수집하고 질문에 대한 응답을 통한 데이터 축적을 바탕으로 패턴인식 및 확률모델을 통해 사망사건과의 연관성을 분석하여 위험군을 선별하는 시스템을 의미한다(차경엽·정근주, 2020). 그리고 해당 시스템에 축적된 데이터를 공개하여 여러 연구자들로 하여금 효과 분석 및 활용하도록 한다. 보건후생부는 시스템 도입을 통해 연 6,500만 명의 아동을 대상으로 한 가정방문과 같은 서비스 품질이 23% 향상된 것으로 평가되고 있다.

2) 미국 GAO의 정부업무카드 부당 지출 선별

미국 GAO는 공무원의 업무상 구매 및 지출의 수월성을 위해 도입된 정부업무 카드 사용에서의 부적절한 지출 선별을 위해 해당 거래 원장을 부적절 거래 유형으로 학습시켜 부적절 지출을 선별하고 있다. 공무원 구매 및 업무카드의 거래 원장 데이터 및 활동패턴과 관련된 변수를 선정 후 모형화하여 학습시킨 후, 부정가능성 있는 케이스를 패턴화 분류하여 적발하는 것이다. 이를 위한 데이터로 거래 원장 데이터에 포함되어 있는 거래성격, 거래업체, 거래금액, 거래시기, 이와 연계된 예산 항목, 정부업무 카드 사용자 정보 등을 DB화하여 활용하고 있다. 미국 국방부 사례로는 2002년 9월 말 감사 결과, 약 68억 달러 규모의 정부업무카드 사용에서의 부적합 지출이 적발된 것으로 나타났다.

〈표 Ⅱ-6〉 GAO의 정부업무카드 부당지출 선별 알고리즘 활용 지표

	내용
거래성격	공무수행에 부적격한 용역 및 재화(귀금속류, 카지노, 유흥업소 등)
	개인적 용도(식료품, 의류, 사치품, 완구류 등)
	여행 관련 거래(출장을 제외한 항공료, 숙박, 렌트비)
거래액수	고액지출 승인, 동일 업체와 비정상적 반복 승인 등
거래업체	취미 등 개인 활동과 연관성이 높은 업체코드를 입력한 후 선별
거래시기	휴일, 주말, 야간 등

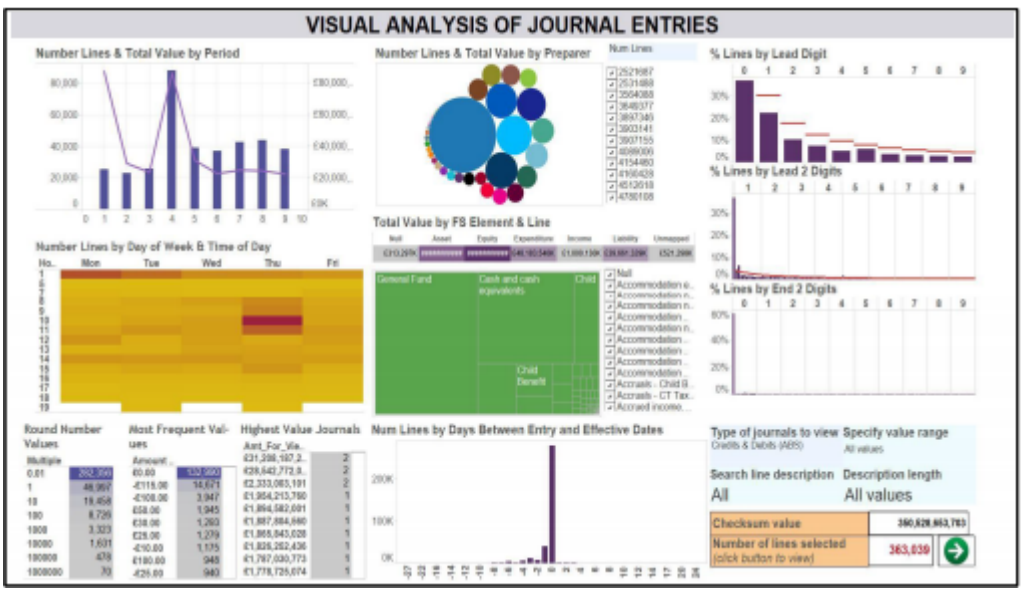
자료 : GAO(2003). pp.4-5. 저자 재구성.

〈표 Ⅱ-6〉에는 미국 GAO의 정부업무카드 부당지출 선별 알고리즘에 활용된 지표들의 예를 정리한 것이다. 이와 같은 지표들을 활용하여 부당지출 거래 선별에 활용하고 있으며 하나의 요인에만 의존하는 것이 아닌 다양한 조건을 중첩시키고 예외 로직을 구성하는 등의 알고리즘을 설계하여 활용하고 있으며 선별 후, 최종 적발은 담당 부서의 검토 후에 이루어지게 된다.

3) 영국 감사원 부적정 거래 위험요인 탐지

영국 감사원은 감사 대상기관 재정DB와 연계된 Data warehouse를 구축하여 감사자가 해당 분야의 원시자료를 추출하여 분석할 수 있도록 지원하는 NAO data system을 구축하여 운영하고 있다. 이 시스템을 활용하여 감사 대상 기관의 재정자료를 분석하여 부정 의심사례를 추출하고자 함이다. 지출거래 내역 데이터를 중심으로 지출과정에서 위험 정도를 점수화하여 부적절한 거래 수와 유형을 분석하여 제공하고 있으며 거래 세부 정보를 분석하여 과다 집행, 부적정 집행 여부를 파악하여 탐지 및 적발에 활용하고 있다(차경엽·정근주, 2021).

〈그림 Ⅱ-15〉 NAO의 텍스트 마이닝 시각화



자료 : INTOSAI(2017).

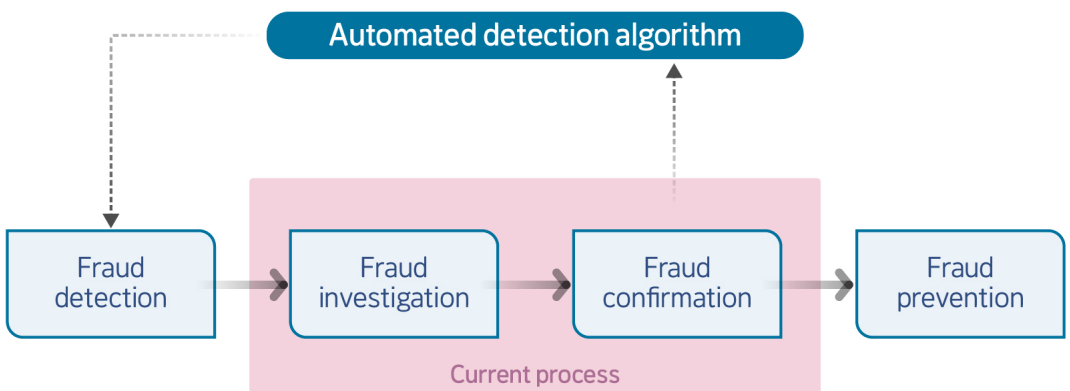
그리고 NAO 에서는 감사 이슈 및 대상 선별을 위해 정형데이터에서 추출되지 않는 정성적, 비정형 데이터를 활용하여 분석시스템에 탑재하였다. 여기에는 문서, 보고서, SNS, 웹상 자료를 활용한 텍스트 분석을 통해 감사대상 업무로 활용될만한 주요 이슈를 추출하는 것으로 〈그림 Ⅱ-15〉에서 볼 수 있듯이 대시보드 및 다양한 방식의 시각화로 직관적으로 보여주고자 하는 것이 특징이다. 문서나 언론에 국한하지 않고 크롤링과 스크래핑을 통해 웹상 및 텍스트 자료를 추출한 후 특정 주제어를 중심으로 상호 연관성 분석을 통해 정부와 공공부문에 대한 여론을 시각화하여 보여준다.

4) 미국 국세청 전자부정적발시스템(EFDS)

머신러닝으로 비교적 활발히 실용화되어 활용되고 있는 분야가 부정적발, 탐지 영역이라고 할 수 있는데 민간 금융 영역에서 이상 거래 탐지 및 부정 거래 선별의 목적에서 시작되어 발전을 거듭하여 공공부문에서 세무행정, 감사 영역에서도 활용되고 있다. 미국 국세청에서는 머신러닝 기법을 활용하여 세금 탈루자에 대한 은행거래 내역 및

기본 정보 등을 벡터화하여 수집한 데이터를 구축하여 분류 및 패턴 분석을 실시하여 탐지하게 된다. 이러한 시스템을 전자부정적발시스템 EFDS(Electronic Fraud Detection System)이라 하는데 이를 구축하여 납세자의 은행계정 인적정보 등을 분석하여 부정 위험을 탐지하는 것이다. 탐지 결과 2015년 기준, 부정 의심사례 69만 건이 나타났고 그 중, 62%가 실제 부적절하게 세금을 환급받은 것으로 판명되었다.

〈그림 Ⅱ-16〉 부정적발 흐름도



자료 : Bart et al, 2015. p23.

5) 뉴욕 주 세무국의 체납세액 징수 최적화 시스템

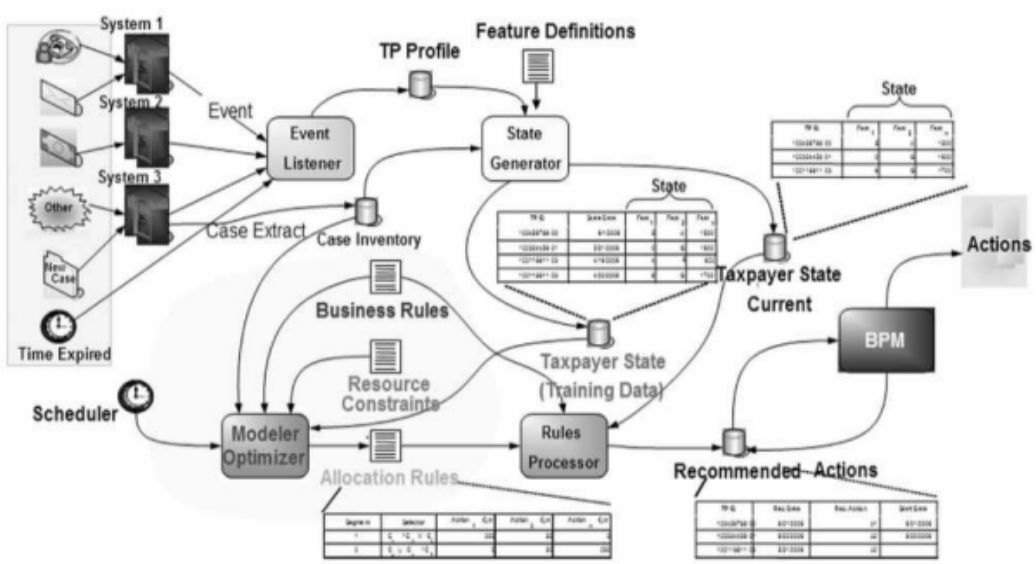
미국 뉴욕 주 세무국(The New York Department of Taxation and Finance) 차원에서 체납세액 징수를 최대한 실행 하면서도 체납자의 경제적, 환경적 여건에 따라 유연하게 대응할 수 있도록 하는 최적화 시스템을 구축하여 운영하고 있다. 세무당국은 제한된 자원으로 세무 행정을 실현해야 하므로 미국의 IBM사와의 계약을 통해 세금징수 최적화 솔루션 시스템을 구축하여 2009년 12월부터 운영하고 있다(행정안전부, 2019).

해당 시스템은 머신러닝 기법 중 강화학습 알고리즘을 채택하여 운영하고 있는데 그 중에서도 제한된 마코프 결정 프로세스[10](constrained Markov decision process

10) 확률 이론 중 하나로 시간의 경과에 따라 상태가 확률적으로 변화하는 과정과 결과에 대하여 파악하는 기법이다. 미래는 과거 상황과는 독립사건이며 현재 상태에 의해서만 결정된다는 논리.

es)의 통합 프레임 워크를 활용하여 체납자 기초 정보 및 은행 거래 내역 등의 빅데이터와 최적화를 결합한 방식이다(행정안전부, 2019). 해석의 여지와 다툼이 있을 수 있는 광범위한 규칙 대신 사용자 정의 징수 정책을 생성하여 징수 프로세스의 효율성과 적응성을 향상시키고자 하는 목적을 갖고 설계되었다. 이 시스템에서는 납세자 배경 정보, 납세자 거래내역 및 뉴욕 주 세무국에 의해 취해진 행정처분 등으로 구성된 데이터로 각 납세자 사례에 대한 다중 샘플링 단계에서 특성 벡터를 생성하게 된다. 그리고 이 데이터가 머신러닝의 훈련데이터로 활용된다.

〈그림 Ⅱ-17〉 뉴욕 주 세무국의 체납징수 최적화 시스템 처리 과정



자료 : 행정안전부(2019). p26.

이 시스템에서는 이를 활용하여 약 200개의 모델링 기능이 탑재되었다. 운영 성과로는 2009년부터 뉴욕 주 세무당국은 시스템 및 동일한 인적자원을 활용하여 체납세액의 징수금액이 약 8,300만 달러 증가한 것으로 나타났다.

다. 시사점

본 장에서는 머신러닝 기법이 활용되고 있는 국내 및 해외 공공부문에서의 사례를 살펴보았다. 우리나라에 비해 해외에서는 이미 2000년대부터 활발하게 머신러닝 기업이 도입이 되어 공공부문에서도 다양한 방면에서 활용되고 있음을 볼 수 있었다. 민간에서의 인공지능 및 머신러닝 기술 발전을 토대로 공공부문에서도 머신러닝 기법을 적용하여 정책 및 제도에 활용하고 있는데 행정영역별로는 국세 징수와 관련된 세정, 방대한 자료를 활용하고 정확도를 높이기 위한 감사 영역, 각종 보조금의 부정수급 탐지 등으로 크게 나누어 활용되고 있었다. 활용 분야로는 이상거래 탐지, 관리대상 선별, 예측 등으로 나누어 볼 수 있었다. 데이터 유형 및 알고리즘은 문서나 신문과 같은 언론, 감사 관련 기록, SNS와 같은 텍스트 자료기반 텍스트 마이닝과 GIS와 결합된 지리적 표시가 적용된 지도와 같은 이미지, 수치로 기록된 정형데이터가 분류와 예측에 활용되는 머신러닝 기법이 적용된 것을 사례를 통해 알 수 있었다.

우리나라에서도 선진 수준의 전자정부 구축을 배경으로 하여 다양한 DB들이 구축되어 운영되고 있다. 그리고 이러한 시스템들은 데이터의 양적인 축적과 더불어 세계적으로 선진적 사례로 벤치마킹의 대상이 될 정도로 질적인 우수성도 갖추고 있다. 국세 행정이나 국가재정시스템 dBrain+, 지방재정시스템 e호조, 국고보조금관리시스템 e나라도움, 건강보험 및 각종 사회보장성 보험들의 DB 등이 그 예가 될 수 있다. 데이터기반 행정법과 디지털 플랫폼 정부 기조에 따라서 머신러닝과 같은 인공지능 기술을 데이터 분석에 활용하여 실제 정책이나 제도에 활용하고자 하는 흐름이 일어나고 있음을 사례를 통해 볼 수 있었다. 그러므로 추후 활용에 있어서도 데이터 DB 간의 활발한 merge 작업을 통한 부가적인 분석이 이루어질 수 있도록 하는 정책적 제도적 뒷받침이 수반되어야 할 것이다.

〈표 Ⅱ-7〉 해외 공공부문 머신러닝 활용 사례 유형별 특징

활용분야	국가 및 기관명	활용 알고리즘 및 데이터	내용
세정 영역	뉴욕 주 세무국	- 세금징수 최적화 솔루션 활용 - 머신러닝의 강화학습 활용 - 납세자 기초정보, 거래내역, 세무국에 의해 취해진 조치로 구성된 데이터를 훈련데이터로 활용	- 2009년 12월에 운영 시작하여 운영 후, 체납세액에 대해 8,300만 달러(8%) 증가 - 징수 프로세스의 효율성과 적응성 향상 평가
	미국 국세청	- 전자부정적발시스템(EFDS)를 구축하여 납세자의 거래내역 이상치 감지를 통해 부정적발	- 2015년 부정 의심사례 69만 건을 조사한 결과 그 중, 62%가 실제 부정하게 세금을 환급받은 것으로 판명
감사 영역	영국 감사원(NAO)	- 대상기관 재정DB와 연계한 데이터 웨어 하우스를 구축(NAO data system) - 지출과정에서 위험 정도를 점수화하여 부적정한 거래 수와 유형을 분석하여 제공 - 거래 세부 정보를 분석하여 과다 집행, 부적정 집행 여부를 파악 - 문서, 보고서, SNS, 웹상 자료를 활용한 텍스트 분석을 통해 감사대상 업무로 활용될 만한 주요 이슈 추출	- 감사 이슈 및 대상 선별을 위해 정형데이터에서 추출되지 않는 정성적, 비정형 데이터를 활용하여 분석시스템에 탑재 - 웹상 및 텍스트 자료를 추출한 후 특정 주제어를 중심으로 상호 연관성 분석을 통해 정부와 공공부문에 대한 여론을 시각화
	미국 감사원(GAO)	- 공무원 구매 및 업무카드의 거래 원장 데이터 및 활동패턴과 관련된 변수를 선정 후 모형화하여 부정가능성 있는 케이스 선별 - 거래성격, 거래업체, 거래금액, 거래시기 등	- 공무원 구매카드 및 업무 카드 관련 비리 적발을 위해 도입 - 미국 국방부 2002년 9월 말 감사 결과, 약 68억 달러 부당 지출 적발
정책의사결정 지원	미국 보건후생부(HHS)	- 위험요인 변수 데이터 축적 및 질문에 대한 응답을 통한 데이터 축적 - 패턴인식 및 확률모델을 통해 위험군 선별	- 시스템 도입을 통해 연 6,550만 명의 아동을 관리 가능 - 가정방문의 질적 향상 등 서비스 질이 23% 향상

자료 : 저자작성.

Ⅲ. 머신러닝 기법 적용 가능성 탐색

3장에서는 dBrain+ 재정데이터를 대상으로 머신러닝 분석 기법을 활용할 수 있는 구체적 가능성 및 분석기반 마련을 위한 개선방안을 탐색해 보고자 한다.

1. 머신러닝 기법의 재정 데이터 활용 가능성

가. dBrain+ 데이터 구성

디지털예산회계시스템은 예산의 편성과 집행, 자금이체, 국유재산 관리, 결산, 성과관리 등 국가재정업무 전반을 전산화한 재정정보시스템(Fiscal Information Management System : FIMS)이다. 시스템에서 수행하고 있는 재정업무는 국가재정법, 국고금관리법 등에 근거하고 있으며 2022년에는 고도화 사업이 완료되어 dBrain+로 개통되어 운영되고 있다. “dBrain+ 시스템 하부 모듈은 개별 업무 담당자가 해당 모듈의 관련 화면에만 접근할 수 있는 폐쇄 형태의 데이터 거버넌스를 가지고 있다”(박정수·나원희, 2020).

dBrain+의 시스템은 〈그림 Ⅲ-1〉에서 볼 수 있듯이 중앙정부, 지방정부, 공공기관, 금융기관 등의 외부 기관 시스템에서의 재정연계 데이터를 바탕으로 16개 카테고리의 통합재정정보시스템을 구성하고 KOFIS(재정통계서비스), KORAHS(정책상황관리시스템), KODAS(데이터분석서비스), 열린재정 등의 4개의 재정데이터시스템으로 구현하고 있는 형태로 볼 수 있다(김민경, 2022).

〈그림 Ⅲ-1〉 dBrain+ 시스템 구성

자료 : dBrain+ 설명자료(한국재정정보원).

재정 투입 및 집행에 따른 경제·사회적 파급효과는 정부의 역할이 점증적으로 확대 및 다양한 기능의 수행 등으로 인하여 분석의 중요성을 갖는다. 한정된 재정자원을 형평성을 확보하면서도 효과적으로 분배하기 위해서는 재정정보를 활용하여 투입과 산출, 결과에 대한 분석이 이루어져야 하기 때문이다. 재정당국이 재정통계를 포함한 재정정보를 생산하고 관리하는 것은 효과적이면서 효율적인 재정운용과 정책 의사결정에 필요한 정보를 체계적으로 제공하는데 그 목적이 있다고 할 수 있다(옥동석·배근호, 2001). 재정정보를 활용한 대표적인 것이 정부에서 공식적으로 발표하는 예산서나 결산서를 들 수 있지만 이것만으로는 정부의 재정활동의 집계적 수치결과를 볼 수는 있지만 그 내용을 파악하기는 상당히 어렵다. 예산 통계나 결산 통계의 경우에는 집행 이전과 이후의 결과를 보여주는 통계로서 실시간 정보로서의 성격은 집행통계에 비해서는 어려운 편이다. 또한 예산과 결산의 경우에는 국가재정법과 예·결산 작성 지침이 기획재정부 주도하에 배포되고 있기 때문에 장-관-항으로 이어지는 분야-부문-프로그램은 소관별, 회계·기금별로 작성되는 예산, 결산 통계정보는 입법부 심사의 대상

이므로 명확한 관리 지침과 법적 근거가 있다고 볼 수 있다. 하지만 세부사업을 수준으로는 통계의 명확한 기준이 불분명하며 관리와 공개를 위한 법적 근거[11]가 미약하여 사업속성 정보도 충실하게 입력되지 않은 실태이며 이에 대한 관심도 부족한 상황으로 이어진 것으로 보인다. 그리고 각 사업별 사업설명자료 역시 이러한 기준과 지침의 부재로 인하여 표준화 작업이 선행되지 않았기 때문에 각 부처별, 사업담당자별 사업설명자료의 내용의 편차가 존재하고 있는 실정이다.

나. 비정형데이터의 활용

정형데이터는 일정한 규칙이나 형태를 지닌 숫자 데이터를 의미한다. 비정형데이터는 정형데이터와 대비되는 것으로 신문기사, sns, 인터넷 웹페이지와 같은 텍스트 형태나 유튜브, 방송과 같은 동영상 형태, 지도, 그림과 같은 이미지 형태를 의미한다. 이러한 원천에서 생산된 데이터는 정제되지 않은 형태로 정형데이터처럼 특정한 컴퓨터가 처리할 수 있는 특정한 값을 가지지 않는 것이 특징이다. 그중에서도 텍스트 분석은 이미 신문기사에서 추출한 텍스트 자료를 활용한 여론 분석, sns나 인터넷상 관련 게시물들을 분석하여 다양한 방면의 반응을 살피는 감성분석을 넘어 이를 정성적 정책효과 분석으로 확장하여 활용하려는 시도도 이어지고 있다. 1950년대 인공지능을 필두로 한 기계학습 영역의 발전에서 텍스트 마이닝과 같은 자연어 처리를 제외하고 논의할 수는 없는데 인간의 언어를 기계로 얼마나 이해하고 그로부터 의미 있는 부분을 추출하는 것을 목표로 한다. 여기에는 음성인식, 텍스트 분석과 같은 분석 작업에 국한된 것이 아니고 음성 인식 후 텍스트로 변환 출력, 텍스트 인식 후 음성으로 변환 출력과 같은 언어와 관련된 다양한 응용을 포함한다(강송희, 2021). 통계적인 자연어 처리는 이후 설명할 머신러닝에 기초하고 있는데 이는 학습을 위한 말뭉치(corpus) 벡터, 집합이 분석의 기초로 활용된다.

dBrain+에는 국유재산과 관련된 이미지 자료들도 존재한다. 이미지 데이터는 인공지능 기술 중 딥러닝에 보다 최적화된 비정형 데이터로 해외에서는 GIS 기술과 결합하

11) 물론 국가재정법 시행령에 따르면 “구분하여 매일 공개”한다고 규정되어 있으나 이것이 반드시 “세부사업” 단위로 강제된 사항은 아니다.

여 비슷한 유형의 국공유지를 분류하여 가치 측정에 활용하거나 메타버스나 시뮬레이션에 활용하고 있다. 국유재산 관리에는 e나라재산 포털이 활용되는데 고해상도 항공영상을 촬영해 이를 '디지털 트윈 국토'로 구축한다는 구상이 발표된 바가 있다(국토지리정보원, 2021). 국토연구원(2022)의 「국유재산 관리혁신을 위한 프롭테크 활용방안」 보고서에 따르면 이미지 분류 및 인공지능 기술 활용을 통해 국유재산의 유휴 면적 예측시스템을 구축할 수 있을 것으로 제안했다. 그리고 방대한 국유지 관리 측면에서는 해당 국유지 이미지 비교 알고리즘을 통해 무단 점유나 불법 이용 여부를 판별 및 선별할 수 있으며 앞서 언급한 디지털 트윈 국토 실현을 통해 개발 시뮬레이션도 국유재산 관리의 측면에서 가능할 것으로 보인다.

다. 정형데이터 활용을 통한 예측

머신러닝 기법은 앞서 이론적 분석에서 언급한 바와 같이 학습을 통한 분류, 예측 및 전망으로 주로 활용되고 있다. 기존 전통적 추정 방법론에서는 인간의 능력이나 정보의 한계로 인한 제한된 합리성 등으로 분석에 활용되는 변수가 제한적이라면 머신러닝을 활용하면 이론적으로는 수백 개 이상의 변수도 분석에 활용할 수 있다는 점이 특징이다. 앞서 논의한 바와 같이 머신러닝 기법은 학습데이터를 바탕으로 스스로 예측하고자 하는 변수를 가장 잘 대표할 수 있는 요인이나 변수를 추출하여 구성한다. 즉, 스스로 모형을 구축하고 그 모형을 토대로 특정 정책을 통해 가장 큰 편익을 가져올 수 있는 수혜집단을 구별하는 등의 사전적 정책효과 분석이 가능하다(정재현 외, 2020). 김민호, 한재필(2020)은 머신러닝 기법의 정책에서의 활용은 정책수혜집단의 선정에서 가장 효과적으로 이루어질 수 있는데 이는 데이터를 기반으로 한 예측이 정책의 효과성을 높일 수 있기 때문이다.

머신러닝을 활용한 정책효과 분석한 기존 선행연구들을 살펴보면 재정데이터를 활용한 정책효과 분석에의 방향을 잡을 수 있을 것이다. 정재현 외(2020)은 "비과세 종합저축 과세특례 제도[12]"를 대상으로 정책 효과분석을 머신러닝 기법을 활용하여 기존의 선행연구와 비교하였다. 기존 정책효과분석은 전통적인 선형모형을 활용한 것으

12) 해당 제도는 노인, 장애인, 생활보호대상자 등 취약계층의 재산형성을 지원하고자 하는 목적을 가진 제도임

로 선행연구에 따르면 근로와 사업 소득의 유무 모두 상관없이 해당 정책이 취약계층 자산형성에 기여하지 못하는 것으로 나타났다. 이처럼 특정 정책에 대한 효과를 데이터 기반으로 분석하여 기존에 비해 다양한 경우를 상정하여 분석이 가능하게 한다.

이와 같은 데이터를 활용한 정책효과 분석 모델은 데이터에 근거를 두는 과학적 기반 논리와 함께 실질적으로 정책목표를 실현하고, 결과적으로는 성과를 높인다는 의미에서 성과기반(performance-based) 정책결정 모델을 지향한다고 할 수 있다(장우영, 2020). 향후 정책 결정 프로세스 에서는 데이터와 함께 정책 수요자의 반응(response) 데이터를 분석하여 정책의 실질적 효과성을 증대 할 수 있는 방향으로 진행되어야 할 것이다. 이를 위해 부처 사업 담당자, 데이터 사이언티스트, 민간영역 전문가, 정책에 관련된 이해당사자가 참여하는 활용 거버넌스의 범위의 확대도 반드시 확보되어야 한다(장우영, 2020).

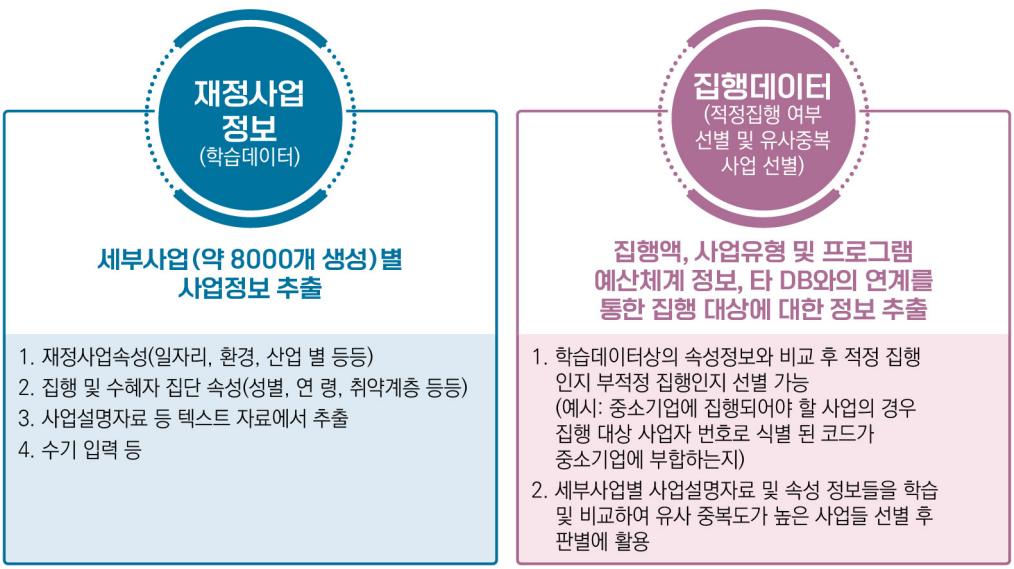
기존 머신러닝을 활용한 정책효과분석에서 볼 수 있듯이 수혜자 집단과 관련된 데이터와 사업집행 데이터가 확보된다면 충분히 dBrain+ 데이터를 활용해서도 정책효과 분석이 가능하다. dBrain+에는 세부사업 단위까지 2007년도부터 2022년도 결산자료로 탑재되어 있는데 여기에는 일부지만 사업속성정보가 포함되어 있으며 집행 대상자 정보도 일부 포함되어 있다. 집행데이터 중 집행대상이 개인이 아닌 사업자를 대상으로 이를 통계청의 SBM(통계기업등록부)과 결합하여 집행의 경제적 파급효과를 분석하거나 국고보조금의 수혜자 정보를 매칭하여 파급효과 및 수혜효과를 정책효과 분석 측면으로 볼 수도 있을 것이다.

라. 부정 집행 탐지

민간의 은행, 카드사와 같은 금융분야에서는 가장 활발하게 머신러닝 기법이 활용되고 있는 부분이 '이상거래 탐지' 이다. 금융분석원(FIU)에서는 민간 은행 거래에 있어서 금액, 대상, 주기성에 따라 알고리즘을 설정하여 이체성 거래에서 이러한 조건이 충족되는 거래가 탐지될 때에는 즉각 시스템에 기록하고 통보되도록 하고 있다. 또한 카드사에서도 카드 사용자의 출입국 기록과 평소 사용패턴 등을 학습시킨 알고리즘을 활용하여 이상 승인거래가 탐지되면 즉각 통보하여 피해를 방지하고자 노력하고 있다.

이러한 이상거래 탐지의 영역은 비지도학습의 분류에서 특이치(out-lier)를 선별하는 방식일 수도 있고 지도학습 중에서 분류영역의 다양한 알고리즘을 활용한 방식이 될 수도 있다. 한국재정정보원에서도 이미 국고보조금 부정수급 의심사례 선별에 앞서 2장에서 소개된 사례와 같이 머신러닝 기법을 활용하고 있다.

〈그림 Ⅲ-2〉 머신러닝을 활용한 재정사업정보 활용 예시



자료 : 저자작성.

현재 재정사업은 매년 약 8000여 개의 세부사업이 재정정보시스템으로 관리되는 가장 하부의 사업 단위로 관리되고 있다. 부정수급 선별과 더불어 머신러닝 기법을 활용해 볼 수 있는 영역으로 사업정보 관리의 확장으로 유사중복 사업 선별 및 부적정 집행 탐지 기능을 생각해 볼 수 있다. 학습데이터로 세부사업별 사업정보를 재정사업고유의 속성, 집행대상 혹은 수혜자 집단에 대한 속성, 사업설명자료 등의 텍스트 자료에서 추출 등으로 각각의 사업 정보를 데이터화하여 활용한다. 이와 같은 사업정보 베이스의 학습데이터를 활용하여 두 가지로 이상탐지 분류로 활용을 생각해 볼 수 있는데 첫 번째는 세부사업 간의 유사중복 여부 선별과 두 번째는 각 세부사업별로 학습된 속성에 따른 집행이 이루어짐에 있어 집행 대상에 적정한 수혜 대상 혹은 집행 대상에 적

정집행이 이루어지는지를 판별하는 것이 그것이다.

첫 번째 세부사업 간의 유사중복 여부는 주로 사업목적, 사업 대상 수혜집단 등을 중심으로 유사·중복 여부를 판별해 볼 수 있다. 기존 연구에서는 해마다 8000여 개 이상의 세부사업이 계속사업, 신규사업 등으로 구성되는바, 그 방대한 양으로 인하여 시범적으로 세부사업명을 기준으로 토픽모델링으로 유사 중복 사업 여부를 기초적인 수준에서 분류가 시도된 적이 있다. 하지만 사업명은 대단히 함축적이며 때로는 사업명과는 다소 상이한 사업내용을 가지고 있는 경우도 많다. 그러므로 각 세부사업별 설명자료 및 속성정보를 학습시켜 사업 간의 유사성과 이질성을 분류하여 유사·중복 사업 판별에 선별적으로 활용할 수 있을 것이다. 다만 기계학습을 통한 분류 결과만으로 판별에 바로 활용하기보다는 선별적으로 하되 최종 판단은 사업담당자가 다양한 요소를 고려하는 것이 타당할 것으로 보인다.

두 번째로는 적정집행 여부 선별인데 학습데이터의 사업정보 및 목적 수혜자 속성이 주로 활용되는 것으로 사업의 집행이 사업에서 정의되는 적정한 대상자에게 이루어지는 것을 판별하는 것이 골자이다. 이 기능이 활용되기 위해서는 거래상대방, 집행대상에 대한 정보가 디브레인 상의 집행정보에 연계되어 있어야 할 것이다. 예를 들어 특정 업종에 집행이 이루어져야 하는 데 사업자 정보에는 다른 업종으로 등록된 곳으로 집행이 이루어진 것으로 선별되면 이를 부적정 집행으로 보고 담당자가 심층적으로 파악하는 것이다.

앞서 논의한 디브레인 집행과 머신러닝 알고리즘을 통한 유사중복 사업의 선별 및 부적정거래 탐지 등은 실현되기 위해서는 후에 서술할 다양한 선결조건들이 수반되어야 할 것이다. 현재로서는 재정데이터, 재정사업속성데이터, 수혜자 데이터가 모두 각각 기획재정부, 국세청, 통계청 등에서 배타적으로 관리되고 있는 실정이기 때문이다. 그리고 머신러닝에 의한 결과만으로는 최종 판단이 이루어져서는 안 되며 선별이라는 점을 감안하여 최종 판단은 담당자와 당국이 다양한 요소들을 고려하여 이루어져야 할 것이다.

2. dBrain+ 데이터를 활용한 텍스트 분석

가. 머신러닝과 텍스트 분석

텍스트 마이닝 혹은 분석은 텍스트 데이터를 사용해 패턴이나 관계를 추출하고 텍스트에서 의미 있는 정보나 가치를 발견 및 해석하는 일련의 과정으로 일반적으로 받아들여지고 있다(장원중·이정인, 2021). 좀 더 자세히 살펴보면 다양한 형식의 텍스트 자료 원천에서 문서형태로 만들고 여기에서 데이터를 추출하여 이를 문서-단어 행렬(document-term matrix) 형태화하여 비정형 데이터에서 정형데이터로 전환한다. 이를 데이터 분석 알고리즘을 적용하여 인사이트를 얻거나 의사결정을 지원하는 하나의 방식이라 볼 수 있다. 여기서 텍스트에 대해서는 여러 정의가 가능하지만 텍스트 데이터 접근에서 말하는 텍스트는 일반적으로 디지털 형태로 저장된 상징들(symbols)이라고 정의되고 있다(Gentzkow et, al, 2019).

〈표 Ⅲ-1〉 텍스트 마이닝의 종류

구분	정의
텍스트 마이닝	인간의 언어로 이루어진 비정형 텍스트 데이터들을 자연어 처리(Natural Language Processing)방식을 이용하여 대규모 문서에서 정보 추출, 연계성 파악, 분류 및 군집화, 요약 등을 통해 데이터의 숨겨진 의미를 유추하는 방법
웹 마이닝	데이터마이닝 기술의 응용분야로서 인터넷을 통해 웹 서비스를 이용하면서 웹에서 패턴을 발견하는 기법
오피니언 마이닝	어떤 사안이나 인물, 이슈, 사건 등에 관련 원천 데이터에서 의견이나 평가, 태도, 감정 등과 같은 주관적인 정보를 식별하고 추출하는 것
소셜 마이닝	SNS 사용자의 로그, 관심사, 정보를 분석하여 트렌드를 읽는 것

자료 : https://datafreedom.tistory.com/entry/R-%ED%85%8D%EC%8A%A4%ED%8A%B8-%EB%A7%88%EC%9D%B4%EB%8B%9D-4-R%EB%A1%9C-%ED%95%98%EB%8A%94-%ED%85%8D%EC%8A%A4%ED%8A%B8-%EB%A7%88%EC%9D%B4%EB%8B%9D-1.

텍스트 데이터와 같은 비정형데이터를 수치형 자료와 같은 정형데이터로 전환해서 분석에 적합한 형태로 변환하는 과정이 분석에 앞서 선행되어야 한다. 텍스트 마이닝에 사용되는 전체 텍스트 데이터를 말뭉치라고 하며 이는 용량의 정형화된 텍스트 집합(large and structured set texts)라고 정의 된다(Miner et al, 2012). 텍스트 데이터를 벡터 행렬 등의 단어표현 형태로 바꾸면 텍스트 마이닝을 활용하여 보다 쉽게 분석이 가능한데 텍스트 마이닝의 목적에 따라 다양한 데이터 전처리 방법이 존재한다. 데이터 전처리 방법에는 토큰화(tokenization), 정제(cleaning), 정규화(normalization), 문서단어행렬(document-term matrix) 등이 있다. 토큰화는 주어진 말뭉치에서 토큰이라 불리는 단위로 나누는 작업을 의미한다. 텍스트 마이닝에서 분석단위로 사용되는 토큰은 단어를 의미하며 이를 단어 토큰화라고 한다. 다만 한글의 경우에는 단어는 여러 개의 형태소로 구성되거나 두 단어 이상으로 이루어진 구절이 의미를 표현하는데 더 적합한 경우도 있다(장원중·이정인, 2021). 토큰화 작업은 단순히 구두점이나 특수문자를 제거하는 정제 작업만으로 해결되지 않으며 의미를 잃어버리지 않도록 정교한 분리 작업이 필요하다.

나. 재정정보에서의 텍스트 마이닝

재정 분야에서는 상대적으로 머신러닝이나 텍스트 마이닝 기술을 활용한 분석의 빈도가 그 중요성과 필요성에 비해서 아직은 활발히 언급되지 않는 분야이다. 하지만 공공부문의 재정 부분은 예산 순기인 예산-집행-결산이라는 과정을 통틀어서 정형, 비정형데이터가 생산되고 있으며 dBrain+ 상에는 집행데이터가 실시간으로 쌓이고 있는데 여기에는 거래상대방 정보를 포함한 시계열적으로도 광범위한 데이터가 존재한다. 그리고 정형데이터 외에도 dBrain+에는 성과정보시스템도 하부 시스템으로 존재하고 있는데 여기에는 각 사업별 설명자료를 포함하여 성과계획서 등 텍스트 문서들이 탑재되어 있다. 넓게 보면 오픈형으로 입력되는 사업속성정보나 프로그램명, 단위사업명, 세부사업명 등도 텍스트 데이터라고 볼 수 있다. 그리고 불용액이 발생했을 때 콤보형으로 입력하는 불용유형의 하부 수기입력 오픈형 데이터인 불용 사유도 어절로 치환되는 말뭉치에서 사유나 함의를 끌어낼 수 있다는 점에서 비정형데이터 중 텍스트

마이닝의 가치를 가진다고 볼 수 있다. 이러한 재정정보에서의 텍스트 마이닝을 통해서 정형데이터와는 다른 관점에서 리스크 관리, 정책효과 분석, 감성분석 등으로 활용될 수 있다. 공공부문의 재정 분야와 데이터 측면에서 일정 부분 유사성을 가진 민간의 은행이나 카드 영역에서는 이미 빅데이터 분석의 한 부분으로 텍스트 분석을 활용하고 있는데 뉴스 기사, 고객 의견, 전문가 인터뷰 등 자연어로 이루어진 텍스트를 통해 새로운 마케팅 타겟팅이나 니치 수요 발굴에 활용하고 있다.

〈표 Ⅲ-2〉 텍스트 마이닝을 적용할 수 있는 재정정보 활용 분야

분야	내용
재정정보 말뭉치 구축	재정정보분야 텍스트마이닝 연구 수행을 위한 말뭉치(corpus) 구축 재정정보 말뭉치를 구축하여 공개하면 재정정보와 관련된 다양한 텍스트마이닝 연구가 수행될 수 있음
재정 집행 이상 탐지	뉴스 기사 또는 SNS 등으로부터 지자체를 포함한 정부 기관의 예산 낭비 사례와 계획과 다른 형태의 정책 집행 사례를 도출하여 올바르지 않은 재정 집행을 탐지
통계 서비스 고도화	dBrain에서 제공하는 KOFIS와 같은 재정정보시스템에 텍스트마이닝 분석 결과를 포함하여 제공 정량적 데이터와 더불어 설문조사, 웹데이터 수집 등을 통해 정성적 텍스트 데이터를 추가로 수집하여 제공
정책 사업 평가	텍스트 분석 결과와 부처별 데이터를 결합한 정책 부처별 데이터와 텍스트 데이터 및 분석결과가 통합된 재정사업 성과관리시스템 구축 텍스트 분석결과와 부처별 데이터를 결합한 정책 사업 평가 정보 제공으로 정책의사결정과 향후 발전 방향에 대한 지원 언론이나 SNS 상의 재정사업에 대한 반응을 탐지해서 부정적, 긍정적 감성분석으로 해석 가능
정책 사각지대 발굴	다양한 매체로부터 여론을 분석하여 사회적 이슈에서 소외된 핵심 정책을 발굴 여론 형성이 어려운 사회적 약자를 위한 재정 및 정책적 지원책 마련

자료 : 강주영(2019), p91. 저자재구성.

〈표 Ⅲ-2〉에서는 재정정보에서의 텍스트 데이터를 활용한 텍스트 마이닝 기법이 적용 가능할 분야를 제시하고 있는데 재정정보 말뭉치 구축, 재정 집행 이상 탐지, 통계서비스 고도화, 정책 사업 평가, 정책 사각지대 발굴 등이 그것이다. 이 중에서도 재정정보 말뭉치 구축은 그 중요성과 실현 가능성이 높은 분야라고 볼 수 있다. 말뭉치 구축은 일종의 질적 분석 및 텍스트 마이닝에 있어 기초 작업이라고 할 수 있는데 말뭉치가 텍스트 활용 분석에 있어서 기초 단위이기 때문이다. 수치와 같은 정형데이터는 그 자체로 통계적 분석을 통해 가공이 가능하고 여기에서 함의를 이끌어 낼 수 있지만 텍스트 데이터의 경우에는 글자 자체는 통계적 분석에 활용하기 어려우며 글자가 조합된 단어에 의미를 기계가 인식할 수 있게끔 분리하는 작업이 필요한데 일정규모 이상의 크기를 갖추고 내용적으로 다양성과 균형성이 확보된 자료의 집합체가 곧 말뭉치(corpus)인 것이다[13]. 말뭉치를 기초로 하여 토큰들을 분류하고 생성형 AI 구축도 가능한 것이며 텍스트 정보를 기계적 분석기법으로 분석할 수 있는 것이다.

다. dBrain+ 텍스트 분석 예시 - 불용사유 텍스트를 중심으로

재정사업의 불용사유는 dBrain 자료상으로도 매우 다양한 것으로 나타나고 있다. 재정사업은 8,000여 개 이상의 세부사업이 있으며, 예산은 지출 성격에 따라 목·세목으로 구분하여 편성되어 사업계획에 따라 지출할 수 있도록 예산편성지침으로 정해져 있다. 또한, 사업을 추진하는 과정에서 발생하는 수 많은 요인들로 인해 예산을 집행하지 못하게 된다. 이것이 미집행 즉, 불용이다. 따라서 불용사유는 정말 다양하며, 하나의 사업이라고 하더라도 원인은 여러 개일 수 있다.

13) https://ko.wikipedia.org/wiki/%EB%A7%90%EB%AD%89%EC%B9%98

〈표 Ⅲ-3〉 불용유형 코드

2007~2013년		2014~2021년	
유형코드	**불용유형**	**유형코드**	**불용유형**
601	계획변경 및 취소	611	지급사유 미발생
602	정원 및 기준호봉 미달 운영	612	계획변경 또는 취소
603	지급사유 미발생	613	예산 절감
604	예산 절감	614	환율 변동
605	집행잔액	615	정원 및 기준호봉 미달
606	환차로 인한 세출감소	616	자금부족
607	세수결함	617	내부거래
608	기타	618	집행잔액 등

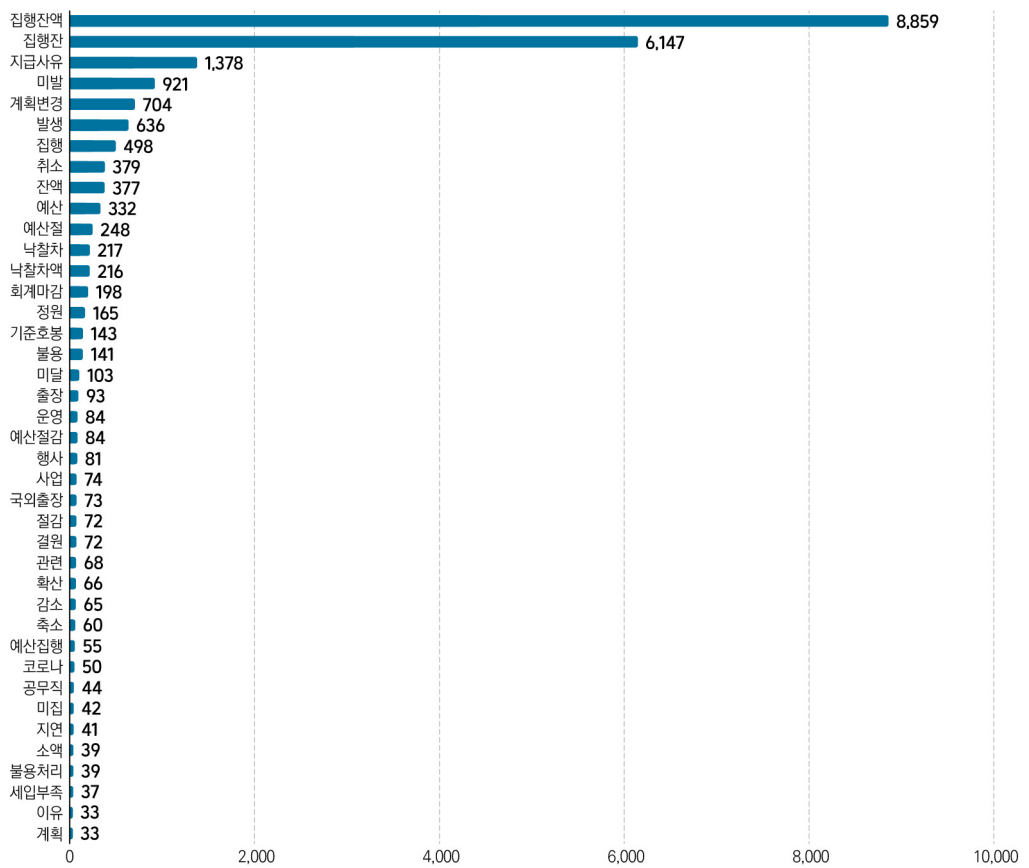
자료: dBrain+

dBrain 사용자는 결산과정에서 집행 후 사용하지 않은 예산에 대해 불용사유를 입력한다. 불용사유는 2021년 결산 시점에서는 불용사유를 8개 유형으로 분류하고 있다. 8개의 불용유형은 '지급사유 미발생', '계획변경 또는 취소', '예산 절감', '환율 변동', '정원 및 기준호봉 미달', '자금부족', '내부거래', '집행잔액 등'이다. dBrain에서는 불용사유를 매년 8개 유형으로 분류하고 있으며, 2007년부터 사용하던 불용유형은 2014년 이후 불용코드의 변경이 있었다. 불용유형코드의 변경과 사유 명칭에 대한 변화는 다소 있었으나, 기존의 사유는 동일하게 유지하고 있다. 불용유형이 불용이 발생한 이유를 8개의 유형으로 분류하여 선택 입력하는 형태라면 불용사유는 오픈형 문항으로 담당자가 직접 상세한 사유를 부가적으로 입력하는 형태이다.

사업의 불용사유는 개별 사업에도 여러 사유가 있게 마련이다. 재정사업의 가장 작은 단위인 세부사업의 경우에도 여러 부서에서 사용하게 되면, 불용사유도 여러 부서에서 입력하게 되므로, 입력 유형이 다양하다. 예를 들어, 2021년 '경찰청 경찰서 청사시설 신축' 세부사업의 경우, 불용사유가 설계용역기간 지연으로 감리용역 발주 계획 변경, 불용, 감리 미계약으로 불용 처리한 것으로 나타났다. 이는 사업의 '612 계획변경 또는 취소'와 '611 지급사유 미발생', '618 집행잔액 등'의 복합적인 불용유형으로 나타난다. 이렇게 세부사업별·세목별로 복합적인 유형이 있는 경우 불용유형 분석이 어렵게 된다.

불용사유를 입력하지 않는 경우, 즉 불용사유를 '미입력'한 불용액은 전 기간 4.6%나 되며, 2020년에는 10% 가까이 되는 상황이다. 이는 dBrain 사용자가 결산업무 처리 과정에서 불용유형을 입력하지 않아 발생하는 문제로 성실한 정보입력의 개선이 필요함을 의미한다. 결산정보가 단순히 결산과정에서 끝나지 않고, 예산편성 시 환류될 수 있음을 더욱 명확하게 알림으로써 다소 개선할 수 있으리라 기대해본다. 또한, 불용사유를 입력하지 않음으로 인해 모든 불용 비목에 대한 세부분석이 불가능하다는 점은 이하 불용사유에 대한 해석 시 유의해야 할 것이다.

〈그림 Ⅲ-3〉 불용사유 빈도 분석 결과

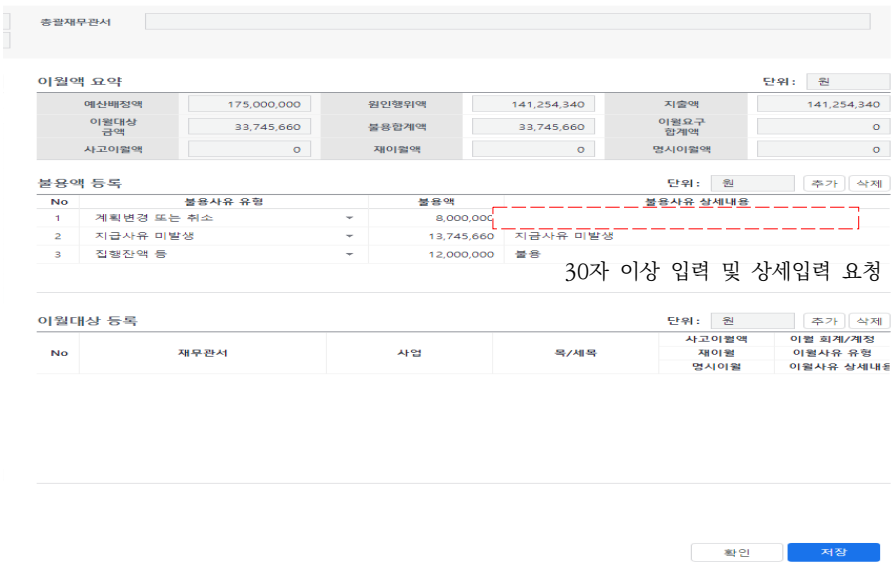


2021년도 재정사업 비목 별 데이터를 대상으로 불용이 발생한 비목을 중심으로 불용사유 입력 항목을 텍스트 분석하였다. 21년도 재정사업 비목별 raw 데이터는 총 41,901건이며 이 중 불용액이 발생한 불용사업은 14,621건으로 집계되었다[14]. 여기에서 불용사유 항목에 미입력이나 의미 없는 구두점만 남긴 경우가 475건으로 나타났으므로 최종적으로 텍스트 분석에 활용될 데이터는 14,146건이다.

불용사유란의 오픈 항목 입력 텍스트를 분석한 결과는 〈그림 Ⅲ-3〉에 나타나 있다. 공백과 특수문자 등을 제거하는 등의 전처리 과정을 거친 것으로 1,577개의 어절로 분류 되었으나 본 분석에서는 편의상 가장 빈도수가 높은 40개를 선정하여 보여주고 있다. 분석결과 불용사유 입력 텍스트 벡터이기 때문에 '집행잔액'이 가장 많은 빈도인 8,857개로 나타났다. 그리고 이와 유사한 '집행잔', '잔액', '예산집행' 등이 집계되었다. '집행잔', '미발', '예산절', '낙찰차' 등등은 아마도 '집행잔(액)', '미발(생)', '예산절(감)', '낙찰차(액)' 의 어절 탈락으로 인한 것으로 보인다. 이러한 오분류가 발생한 원인은 우선 입력자들이 오입력 한 경우가 상당수 존재하고 가장 많은 경우는 '재정 토큰' 기반 사전으로 분류를 한 것이 아닌 범용적인 한글 토큰 기반 사전을 활용하였기 때문에 '집행잔액', '집행 잔액'의 형태를 '집행잔액', '집행'+'잔액', '집행잔'+'액' 등으로 오분류 하기도 하기 때문인 것으로 보인다.

14) 많은 논문에서 불용사업을 논의할 때, 분석대상 불용사업을 선정함에 있어 불용액을 특정 금액 이상일 경우로 한정하여 분석하는 경우가 많다. 이는 회계상 1원이라도 불용액이 발생하면 회계연도 내에 집행되지 않고 남은 금액이라는 불용액 정의에 따르면 지나치게 금액 분포가 넓어지기 때문으로 분석의 편의 및 효율성을 높이기 위해서이다. 다만 본 분석은 불용비목 발생 시 입력된 불용사유에 대한 시범적인 텍스트 분석이기 때문에 1원 이상 불용액이 발생한 비목 모두를 분석대상으로 설정하였다.

〈그림 Ⅲ-4〉 불용사업 등록 화면 개선 예시



자료 : dBrain+

이와 같은 분류 오류를 줄이고 텍스트 분석 결과의 품질을 높이기 위해서는 우선 텍스트 raw 데이터 입력자들의 입력 오류를 줄이려는 노력과 더불어 불용사유 입력에 있어 상세한 사유를 입력하려는 독려가 필요하다. 〈그림 Ⅲ-4〉은 dBrain+상의 불용액 등록 화면에서의 개선방안을 예시적으로 제시한 것이다. 현행은 불용사유 유형 선택 후 입력자의 선택에 의해 상세내용을 입력하게 되어 있는데 그러다 보니 앞서 분석에서처럼 의미 없는 내용 입력 혹은 무응답이 상당 부분을 차지하고 있다. 물론 불용사유 유형 선택항목이 대분류 적으로 사유를 담고 있고 불용사업의 경우 특히 소액일수록 단순 집행잔액일 확률이 클 것이다. 하지만 금액이 크다면 단순히 사업 미집행, 집행잔액 발생으로 단순히 기록하기보다는 예를 들어 “코로나 상황으로 인한 사업수요 저조”처럼 보다 상세히 남겨둔다면 차후에 사업관리에 있어 참고자료 및 분석 데이터로 활용할 수 있을 것이다. 이를 위해서 예시적으로 불용사유 항목 입력을 강제하고 최소 30자 이상으로 입력하도록 폼 형태를 설정하고 각 유형별로 예시나 입력 가이드를 제시하는 것도 대안이 될 수 있을 것이다. 이를 통해 입력 데이터의 품질을 높이고 향후 사업관리에도 활용될 수 있을 것이다.

그리고 가장 중요한 것은 오분류가 일어나지 않도록 '재정용어' 중심으로 재정용어 말뭉치를 활용한 재정토큰에 기반한 사전을 활용할 수 있도록 이러한 환경을 구축할 필요가 있다. 이를 통해 불용유형별, 사업유형별 텍스트 빈도 분석이나 불용금액과 텍스트 간의 관계를 규명하는 데에도 활용될 수 있을 것이다.

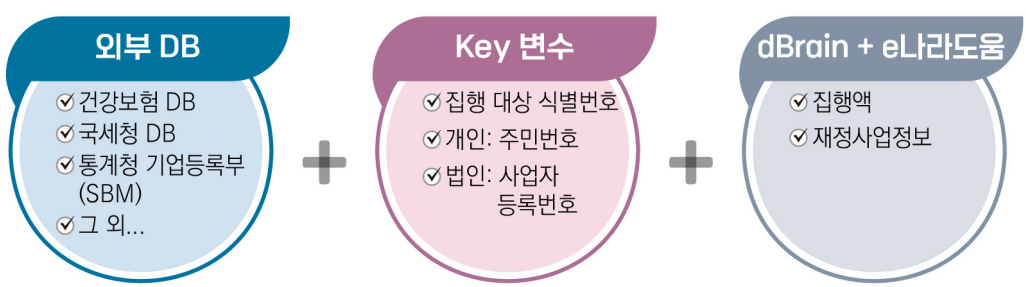
Ⅳ. 머신러닝 분석기반 확충을 위한 선결조건

본 장에서는 앞서 논의한 재정데이터를 대상으로 머신러닝 기법을 적용하기 위해 확보되어야 하는 데이터나 선결조건 및 고려사항에 대해 논의해 보고자 한다.

1. 타 DB와의 연계를 통한 데이터 변수 추가

기존의 전통적 통계 추론 방법과 비교하여 머신러닝 분석기법이 가지는 최대 장점은 이론적으로는 수백 개 이상의 변수를 동시에 종속변수에 영향을 미치는 정도를 분석할 수 있다는 점이다. 알고리즘에 따라서 앞서 논의한 비지도학습 기법을 활용한 군집별로 유사도에 따른 분류분석도 가능한데 보다 정확도를 높이기 위해서는 집단의 성격을 정의함에 있어 유사도와 불순도를 구분할 수 있도록 다양한 변수를 활용하는 편이 효과적이다. 현재 dBrain+와 e나라도움 시스템에 탑재된 변수들도 프로그램 예산 체계

〈그림 Ⅳ-1〉 재정시스템과 외부 DB와의 연계 방안



자료 : 저자작성.

에 따른 재정사업 정보 변수 및 국고보조 사업 정보 변수들이 다수 구성되어 있다. 하지만 집행 정보 위주라는 특성에 비추어 볼 때 이것이 분석에 활용하여 의미 있는 결과를 도출하기 위해서는 집행 대상자 혹은 수혜자 특성에 관한 변수 추가가 필요하다. 특히 이것은 머신러닝 기법을 활용한 분야 중에서 수혜집단별 정책효과 분석이나 전망

및 예측에 있어서 가장 필요한 선결조건인데 앞서 논의한 바와 같이 dBrain+나 e나라도움의 재정사업 및 국고보조사업 데이터만으로는 정책 수혜자 집단에 대한 변수가 부재하기 때문에 유의미한 정책효과 분석이 이루어지기 어렵다.

〈표 Ⅳ-1〉 dBrain+와 기업통계등록부 매칭 변수

dBrain+	내용	기업통계등록부	내용
소관명	집행 부처 및 기관	시도	거래 상대방 소재주소지
목	지출목	사업자구분	개인, 법인 사업자
세목	지출세목	사업체구분	개인사업체, 회사법인, 회사이외법인, 비법인단체, 국가지방자치단체
사업자등록번호	식별번호 Key 변수	공공기관유형	중앙정부기관, 중앙정부비영리단체, 지방정부기관, 지방정부비영리단체, 사회보장기금, 공기업
거래일	집행시점	기업유형	상출기업, 기타대기업, 중견기업, 중기업, 소기업
지출액	집행금액	중소기업구분	중소기업제외, 중기업, 소기업, 소상공인
		기업체 매출액	매출액
		기업체 근로자	상용근로자
		사업자 산업분류	대분류, 중분류

자료 : 김민경(2019) 내부용 보고서. p.14.

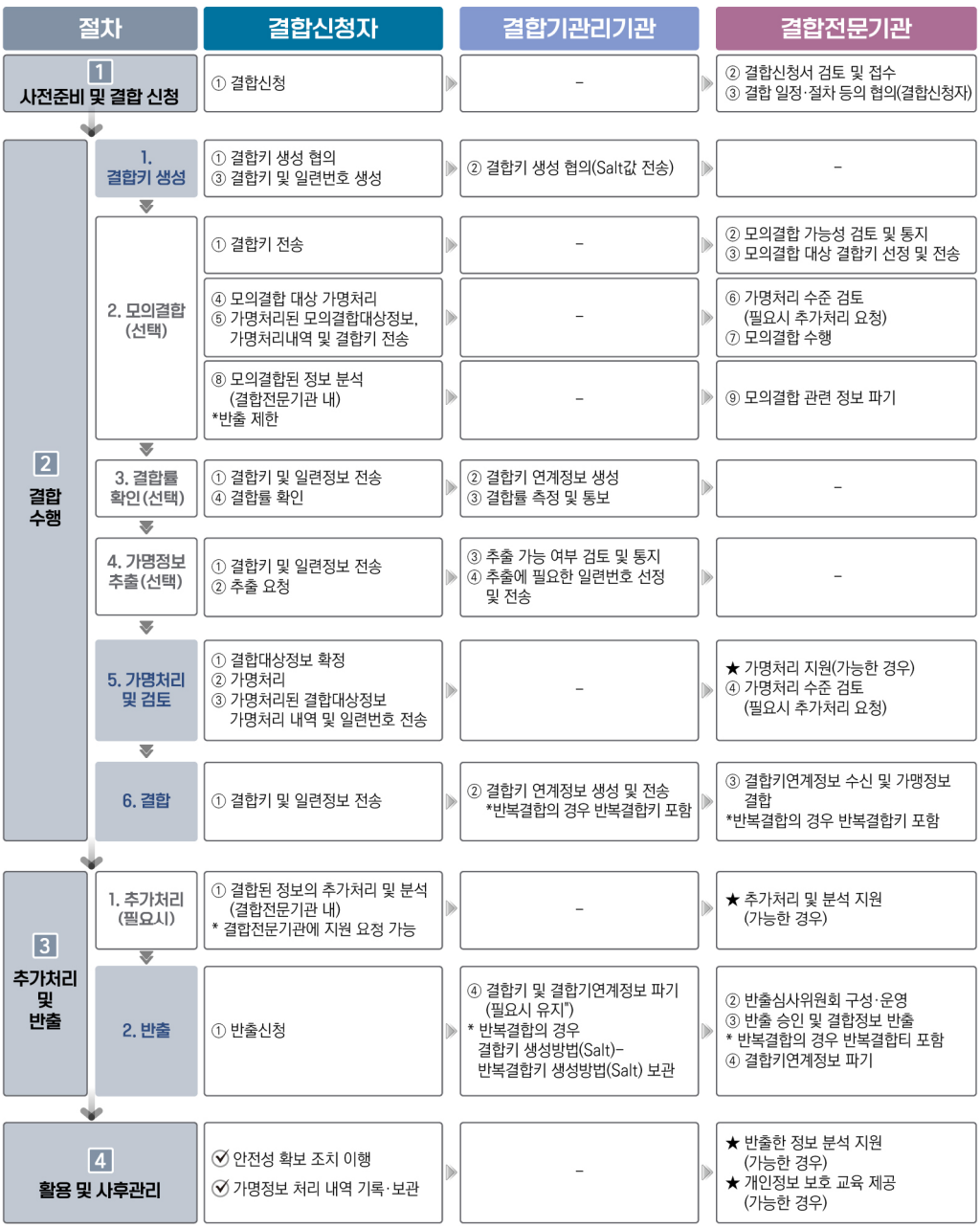
이미 dBrain과 통계청의 기업통계등록부와의 연계를 통한 경제적 파급효과와 관련된 분석 사례가 존재하고 있으므로 기술적으로 불가능한 것은 아니라고 할 수 있다. 〈표 Ⅳ-1〉은 이전에 수행된 dBrain과 기업통계등록부의 결합사례를 변수 중심으로 정리한 것으로 두 DB를 결합한 key 변수는 사업자등록번호로 설정하여 진행한 것을 볼 수 있다. 다만 개인식별번호 및 민감한 재정 DB 데이터와의 결합이다 보니 지금까지의 결합은 1회성에 그쳤고 분석 장소도 한정되는 등 절차상의 복잡성으로 상시적으로 이루어지진 않고 있다. 각각의 기관의 DB 끼리 결합하여 활용하고자 함에 있어 기관별로 개별적 협의로 이루어지고 있었는데 그때마다 기존에 진행된 key 값이나 작업이력이 관리되지 않아 매번 새롭게 과정이 이루어지는 등 효율성이 낮은 문제점도 드러났다. 그렇지만 서로 다른 DB의 데이터를 개인정보 나 식별번호 등을 key 값으로 하

여 결합이 가능하고 분석할 경우에는 더욱 가치 있는 데이터가 생성될 수 있다는 점에서 타 DB와의 연계는 시계열적으로 지속적인 분석을 위해 필요하다. dBrain 상의 집행데이터, 통계청의 기업통계등록부 데이터만으로는 재정집행의 경제적 파급효과를 실집행수준으로 분석할 수는 없으며 결합된 데이터로 분석이 가능했던 것이다.

현재 디브레인과의 연계에 있어 가장 큰 이슈는 재정집행정보라는 본질적 민감성과 기록되어 있는 수혜자 식별정보로 인한 것으로 관련 유관 부처 및 데이터 소유권자들의 복잡한 협의를 요하는 작업이며 개인정보 보호라는 이슈를 고려해야 하기 때문에 법적, 규정별로 준수하여 진행해야 할 것이다. 이러한 과정에서의 걸림돌을 해결하기 위해 개인정보보호위원회에서 수행하는 '가명정보결합지원제도[15]' 프로세스를 이용하는 것도 방안이 될 수 있을 것이다. 개인정보보호법 제28조의2에 근거하여 서로 다른 DB 간의 식별정보를 활용한 결합에 관한 당위성은 있지만 동법 제28조의3에서는 그럼에도 불구하고 서로 다른 개인정보처리자 간의 가명정보의 결합은 개인정보보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관이 수행하도록 함으로써 DB 결합에서 야기될 수 있는 문제들을 상당 부분 절차적 타당성을 확보할 수 있을 것으로 보인다. 〈그림 Ⅳ-2〉는 개인정보보호위원회에서 발간한 가명정보결합 가이드라인에서 제시한 결합절차 진행 과정을 도식화한 것이다. 기존의 두 기관에서 자체적으로 하던 과정을 법적 근거를 가진 가명정보결합지원기관을 통해서 수행함으로써 개인정보보호라는 신뢰성을 확보할 수 있을 것으로 기대되고 시간과 자원의 단축을 통해 분석의 효율성도 확보할 수 있을 것으로 보인다.

15) 가명정보의 결합은 개인정보처리자의 정당한 처리 범위 내에서 통계작성, 과학적 연구, 공익적 기록보존 등의 목적으로 정보주체의 동의 없이 처리할 수 있음(개인정보보호법 제28조의2).

〈그림 Ⅳ-2〉 가명정보결합 절차 진행 과정



자료 : 가명정보처리 가이드라인(2022).

머신러닝을 활용한 분석이나 데이터 활용을 위해서는 재정데이터 이외의 타 DB와의 연계를 통한 변수 및 데이터의 추가 작업이 반드시 필요하다. 이는 분석의 다양성 및 실용적인 정책지원을 위해서 선결되어야 하는 작업이다. 이를 통해서 재정사업의 최종 수혜자 정보와 매칭이 이루어지면 실질적인 수혜자 집단별 정책 효과분석이 가능할 것이며 지방재정집행DB가 연계된다면 중앙-지방-수혜자로 이어지는 실집행 현황 및 파급효과 분석도 가능할 것으로 기대된다.

2. 재정 말뭉치를 통한 텍스트 정보 활용 기반 조성

dBrain+에는 성과 관련 문서자료를 중심으로 한 텍스트 자료와 국유재산 지도 및 항공사진과 같은 이미지 자료가 탑재되어 있다. 그중에서도 앞서 논의한 바와 같이 텍스트 정보와 같은 비정형데이터를 유의미하게 활용하기 위해서는 정제하여 활용할 수 있도록 분석기반 조성이 선행되어야 할 것이다. 좋은 분석은 결국 좋은 데이터로부터 얻을 수 있는 결과라고 본다면 분석기반과 같은 텍스트 분석에 필요한 분석 인프라를 조성하는 것이 중요하다. 텍스트 분석에서의 인프라 조성은 결국 전처리과정에서 분야별로 의미 있는 덩어리로만 분리하여 분석에 활용할 수 있도록 하는 작업이 핵심인데 텍스트 분석에서의 인프라는 곧 말뭉치-토큰화-사전 구축으로 유기적인 작업이 선행되어야 함을 볼 수 있다.

앞서 시험적으로 실시했던 불용사유 텍스트 데이터[16]를 대상으로 한 분석결과에서 볼 수 있듯이 재정 토큰 사전이 구축되지 않아 범용으로 활용되는 한글 사전을 분석에 활용한 결과 '불용액'이라는 재정한 토큰이 '불용'+'액'이라는 형태로 분리되어 인식되는 경우가 발생하는 등 현시점에서는 여러 단계로 전처리가 이루어질 필요가 있음을 시사했다. 텍스트마이닝에 있어서 전처리 과정은 단순히 수치데이터에서의 공백제거, 특이치 제거와 같은 정량적 성격의 데이터 클리닝 수준을 넘어 정성적인 측면에서 언어의 분야별 정제 작업의 수준으로 이루어져야 함을 의미한다.

16) 분석에 활용한 불용사유 텍스트 데이터는 상위 콤보 항목인 불용유형의 하위 변수로써 이미 상위에서 유형화가 이루어진 상태에서 주관식으로 입력된 항목이라는 특수성이 있음에도 바로 활용하기에는 상당 부분 전처리와 같은 정제가 필요한 것으로 나타났다.

현재 dBrain+에는 AI 기술을 활용한 챗봇 서비스가 운영 중인데 현재는 자연어 처리 기반 챗봇으로 생성형 기술이 아니기 때문에 미리 설정된 답변 범주를 넘는 질문은 처리하기 어려운 한계가 존재한다. 현시점에서 공공부문을 포함한 민간부문에서의 챗봇의 대부분은 정해진 질문과 답변 범주 내에서 인식하여 처리하는 자연어 처리 방식이지만 챗gpt의 출현 이후, 생성형 AI 방식의 챗봇이 구글[17]과 같은 대형 포털을 중심으로 발표되거나 도입되고 있다. 기술의 발전과 상용화 시기가 가속화되고 있다는 점을 상기해 본다면 곧 챗봇 영역에서도 기존의 자연어 처리를 대체하여 생성형 챗봇이 대세를 이룰 것이라 볼 수 있다. 생성형 AI 기술이 활용되는데 가장 기본은 텍스트 데이터와 같은 비정형데이터를 정형화하는 것이다. 그러므로 범용적인 형태소 분석이나 토큰 활용을 넘어서 재정 분야에 맞는 형태소 구축기와 토큰 사전 등을 만드는 작업이 필요하고 그러기 위해서는 재정 말뭉치를 구축하는 방안을 우선 추진해야 할 것이다. 〈표 Ⅳ-2〉은 조세지출예산서에 쓰인 일부를 예시적으로 재정 토큰화로 분류한 것을 보여준다.

〈표 Ⅳ-2〉 재정 형태소 분석 도입 예시

기존 한글 토큰 분석 시	재정 토큰 분석 시
‘조세’ ‘지출’ ‘예산’ ‘제도’	‘조세지출예산제도’
‘재정’ ‘지원’	‘재정지원’
‘재정’ ‘정보’	‘재정정보’
‘조세’ ‘특례’	‘조세특례’
‘소득’ ‘세법상’	‘소득세법상’
‘특별’ ‘세액’ ‘공제’	‘특별세액공제’
‘과세’ ‘이연’	‘과세이연’
‘고정’ ‘자산’	‘고정자산’

자료 : 열린재정 내 조세지출예산서 pp1-3.

17) 구글의 바드(Bard), 마이크로소프트 포털서비스 Bing에 생성형 챗봇 투입, 네이버의 큐(QUE) 등이 현재 발표되었고 도입 중이다.

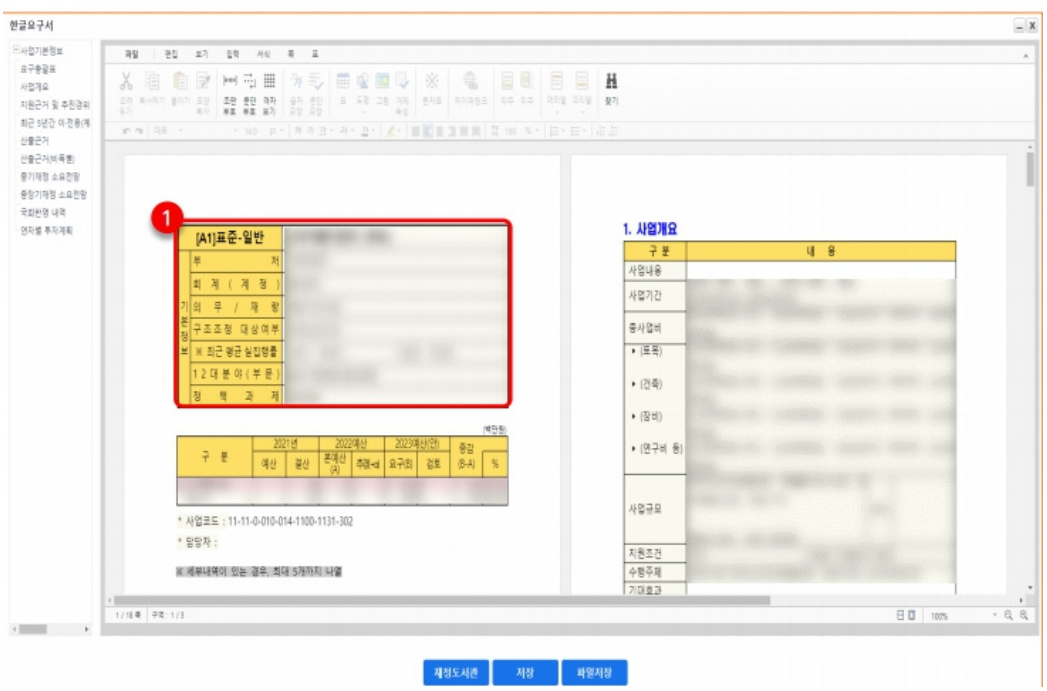
기존 회계용어를 대상으로 한 형태소 분석기 구축 관련 선행연구[18]에서와 유사하게 복합명사로 이루어진 고유 학술용어의 경우는 〈표 Ⅳ-2〉에서 볼 수 있듯이 기존 한글 형태소 토큰 기반의 사전에서 대부분 어절 단위로 분리가 되는 경향을 보였다. 〈표 Ⅳ-2〉은 예시적으로 열린재정 내에 탑재된 조세지출예산서의 일부 페이지를 (가칭) '재정토큰'화를 보여준 것으로 단순하게는 어절로 분리 가능한 복합명사적 용어를 하나의 용어화 하여 편찬하는 작업으로 이해할 수 있다. 이처럼 텍스트 마이닝 작업의 기초 작업은 시간과 인력이 투입이 전제되어야 하는 작업이지만 학술용어의 경우에는 초기 말뭉치 구축 작업 대비 추후 이를 자원으로 이루어질 텍스트 분석 전처리 작업은 점점 단축될 것이다. 현재 열린재정 내의 재정문서자료 중 '재정 말뭉치' 구축에 활용될 만한 문서로는 성과계획서, 조세지출예산서, 한국재정정보원 발간 재정 관련 통계집 및 책자, 기획재정부 발간 책자 등이 그 대상이 될 수 있을 것으로 보인다. 이렇게 구축된 '재정 말뭉치' 및 이를 기초로 한 '재정 토큰 사전' 등은 재정원에서 구축하는 재정용어 사전과 함께 향후 텍스트 마이닝 분석에 유용하게 활용될 것으로 기대된다. 나아가 텍스트 분석에서만의 활용이 아니라 향후 도입될 생성형 AI 기술 기반 대화형 상담과 같은 텍스트 기반 서비스 향상에도 활용 효과성이 높다고 볼 수 있다.

3. 속성정보 정비를 통한 데이터 확충

머신러닝 기법을 활용한 수혜집단별 정책효과 분석, 파급효과 분석 또는 재정 전망 예측에서 보다 나은 결과를 도출하기 위해서는 알고리즘을 활용한 모델링에 투입될 수 있는 변수가 많을수록, 관측치가 많을수록 시점이 다양할수록 예측 품질이 올라가는 경향이 있다. 특히 특정 재정사업 효과 분석을 위해서는 수치형 자료뿐만 아니라 재정사업이 어떤 대상으로 어떤 정책목적을 가지고 있는지의 정보와 관련된 사업 속성 정보도 중요하다. 이러한 기초 분류 정보를 학습시켜 재정사업 및 지출을 유형화하거나 패턴분류 등을 통한 효과분석을 할 수 있기 때문이다.

〈그림 Ⅳ-3〉에서 볼 수 있듯이 프로그램 예산의 계층 구조는 필수입력 사항으로 선택형 콤보 항목으로 필수 지정으로 입력이 의무적으로 이루어지고 있으나 정책사업이나 국정과제 등의 항목은 사업담당자의 재량에 따라 선택적으로 오픈형 항목으로 입력이 이루어지는 방식으로 시스템상에서 구현되고 있기 때문이다. 그렇기 때문에 입력이 이루어진 항목에서도 오탈자, 적합하지 않은 분류항목 등 오류가 존재하고 있으며 데이터 활용을 위한 전산화 작업도 대부분 이루어지지 않고 있는 실정이다. 하지만 확장재정에서 건전재정으로의 재정운용 기조의 변경과 이로 인한 재정성과관리의 중요성이 대두되고 있는 시점에서 프로그램 단위별 재정사업 관리도 중요하지만 정부가 집중적으로 집행하고자 하는 국정과제별 혹은 정책별 재정 소요 관리가 중요하므로 이에 대한 메타 데이터 관리 및 생성이 중요하다(김민경, 2022).

〈그림 Ⅳ-3〉 사업정보 입력 관련 한글 요구서 항목



자료 : 디지털예산회계시스템.

정책사업에 대한 사업정보관리는 각 사업 담당자가 직접 입력하는 방식으로 관리되고 있는데 의무 사항이 아니기 때문에 극히 일부의 사업만이 입력되어 관리되고 있다.

박정수(2020)에 따르면 디지털예산회계시스템에는 지침과 법률에서 정의하고 있는 공식 재정사업 분류체계와 연동되는 원리로 세부사업별로 공통 색인을 부여하고 각종 텍스트 정보를 관리할 수 있는 사업관리 시스템이 별도로 구성되어 있다고 보고 있다. 이를 바탕으로 각 세부사업별로 상세정보와 집행 현황을 모니터링 할 수 있다는 것인데 다만 텍스트 정보가 전산화가 된 자료가 아닌 한글문서 기반의 텍스트 자료이기 때문에 그 활용에 한계가 존재하고 있다고 언급하고 있다. 일반적으로 정보시스템에서 그 정보를 가공하여 활용하기 위해서는 활용하고자 하는 목적에 맞게 필요한 데이터를 추출할 수 있어야 하는데 이는 데이터가 문서 형태가 아닌 데이터베이스화된 형태로 되어 있어야 활용가치가 높다고 할 수 있다. 사업정보 입력이 이루어진 뒤에는 예산요구서 작성을 하게 되는데 여기에는 기본적인 재정사업 정보 이외에도 구조조정 대상 여부, 정책과제 여부 지원조건, 수행 주체, 기대효과 등 더 상세한 정보가 입력되게 된다(김민경, 2022). 그러나 이러한 정보는 최근까지 데이터베이스화가 이루어지기 쉬운 형태가 아닌 텍스트 기반의 한글 파일 형식(*.hwp)으로 부속파일로 기록되고 있어 그 활용이 어려운 단점이 존재한다(김민경, 2022). 향후 예산순기부터는 한글파일 형식에서 데이터베이스 입력 방식으로의 전환을 고려할 필요성이 존재하며 이미 한글요구서로 탑재된 파일도 데이터베이스 방식으로의 전환이 이루어져야 그 활용도가 높아질 것이다. 부차적으로는 사업설명자료에서의 텍스트 마이닝 작업을 통한 사업속성 추출 등의 작업 등을 통해 더 보완할 수 있을 것이다.

재정사업 속성정보의 확충은 머신러닝 기반 분석에 있어 분석활용 변수 및 데이터 추가라는 측면에서는 앞서 논의한 첫 번째 선결조건과 유사한 개념이다. 하지만 속성정보의 확충은 예를 들어 재정사업 수혜자 집단 속성별 분류 분석 및 패턴분석을 앞서 논의한 서포트벡터머신이나 앙상블 기법을 수행함에 있어 분석타겟을 명확하게 보여줄 수 있을 것이다. 관련 선행 연구 등에서 꾸준히 논의된 바처럼 dBrain+의 하부시스템으로 볼 수 있는 사업관리시스템을 좀 더 활용하여 속성정보 체계 정비 및 이를 관리할 수 있는 거버넌스를 구축하여 현재의 극히 일부만 입력이 이루어지면서 사업담당자의 재량에만 의존하는 방식을 탈피하여 속성정보의 양과 질을 고도화하는 노력이 선행되어야 할 것이다. 현재 시스템상에는 사업 속성 정보에 대해 아주 일부에 관해서 입력이 이루어져 있기는 하지만 정책 단위로 재정정책 분석과 수립을 위한 지원기능을 수

행하기 위해서는 단순히 프로그램 예산제도 기반의 분류가 아닌 수혜자 속성 정보 등으로 보다 세분화한 사업속성체계 구축으로 데이터 보완이 필요하다. 이를 통해 재정데이터 분석에서 수요가 높은 실집행 분석을 위한 수혜자 및 정책유형에 따른 분석이 머신러닝 기법을 활용하여 이루어질 수 있다.

Ⅴ. 결론

본 논의에서는 최근 AI 기술의 발달로 산업 및 경제, 사회 전반에 많은 변화가 이루어지고 있는 현시점에서 공공부문에의 AI기술의 적용, 그중에서도 머신러닝 기법의 재정정보에의 적용 가능성을 탐색해 보고자 하였다. 현재 공공부문에서의 머신러닝 기법을 포함한 AI 기술은 챗봇, 단순 정보제공 업무 보조 등 한정적 분야를 중심으로 도입이 이루어지고 있는데 빅데이터 처리 및 알고리즘을 활용한 고도의 모델링을 활용하는 것이 머신러닝을 비롯한 인공지능 기술의 장점임을 고려해 봤을 때, 공공부문에서의 그 활용 방안을 고도화할 수 있는 방안 탐색의 필요성이 존재한다.

그 중에서도 머신러닝 기법은 정형데이터 및 비정형데이터, 다변수 고려에 특화되어 있다는 점에서 재정데이터 분석에 활용하기에 적합하다고 볼 수 있다. 재정데이터에의 머신러닝 기법 적용 가능성을 살펴보기 위하여 논의에서는 머신러닝의 여러 알고리즘 및 기법에 대해 소개하고 머신러닝 기법이 공공부문에 적용된 사례를 해외와 국내로 나누어 살펴보았다. 여기에서 예측과 전망, 분류, 이상거래 탐지 등이 세정영역, 감사영역, 일반행정에서의 정책의사결정 지원 등의 영역에서 활용되고 있음을 사례 및 이론을 통해 확인할 수 있었다. 이러한 사례를 바탕으로 재정정보에 머신러닝 기법을 적용할 시 기대할 수 있는 활용 가능성을 제시하였다.

첫 번째로는 dBrain+에 탑재되어 있는 텍스트 정보와 같은 비정형데이터의 활용도를 높이는 것이다. 기본적으로 머신러닝을 활용한 텍스트마이닝은 문서요약, 감성분석, 텍스트 분류, 텍스트 유사도 분석 등 네 가지 영역으로 활용될 수 있다. 현재 dBrain+에는 재정사업성과정보와 관련된 문서 정보와 평가결과와 같은 텍스트 문서가 축적이 되어 있다. 하지만 한글문서로 존재하고 있어 그 활용은 많이 이루어지고 있지 않은 편인데 이를 머신러닝의 텍스트 분석 알고리즘을 활용하여 정성적인 정보와 텍스트가 내포하고 있는 사업 정보 등을 데이터화하여 활용하는 것이다. 여기에 재정수치 데이터와 결합시키면 활용도는 높아질 것으로 기대된다.

두 번째로는 예측을 통한 정책지원으로 머신러닝 기법이 갖고 있는 강점인 다변수와

빅데이터를 활용한 예측 모델링으로 재정분야에서 전망 기능을 수행하는 것이다. 머신러닝 알고리즘은 앞서 논의한 바와 같이 전통적인 인과관계의 추정이 필수적이지 않은 예측문제에 있어 기존 전통적 모델링과 차별점을 갖는 전망 모형을 구축할 수 있다. 재정분야에서는 세입예측, 재정지출소요 전망, 재정 집행 효과 분석 등에서 활용될 수 있을 것이다.

세 번째로는 부적정 집행 탐지와 같은 이상치 탐지 영역이다. 금융권을 중심으로 이미 민간영역에서는 이상거래 탐지 및 선별에서 의사결정나무와 같은 알고리즘을 적용하여 가장 활발하게 활용하고 있는 분야이다. 공공부문에서도 각종 부정수급 탐지 및 부적정 지출 스크리닝 및 적발에 활용 가능할 것으로 보인다.

실증적인 분석으로는 예시적으로 dBrain+ 상의 불용이 발생한 사업을 대상으로 불용사유 텍스트 빈도 분석을 실시하였다. 분석 결과 불용액이 발생한 사업은 불용 유형을 콤보박스로 선택하고 불용사유를 주관식으로 입력하게 되어 있지만 상당 부분은 공란이거나 구두점과 같은 의미 없는 입력, 선행된 불용유형과 동일내용 입력 등으로 불용사유 입력의 실익이 없는 것으로 나타났다. 분석결과에 따라 불용사유 입력에 있어 10자 이상 입력 폼으로의 수정 등을 통해 입력 텍스트의 유의미성을 높여 향후 불용사업 관리에 있어 의미 있는 분석 및 데이터로 활용될 수 있도록 해야 할 것이다. 그리고 머신러닝 활용 차원에서는 분석결과에서 재정용어의 복합 명사 형태 인식이 되지 않아 불용발생 사유에 대한 의미 있는 함의를 이끌어내기 어려웠기 때문에 재정용어 분석에 특화된 재정용어 사전 및 이를 위한 재정 말뭉치 구축이 선행되어야 할 것이다.

이처럼 재정정보 활용에 있어 머신러닝 기법을 적용하기 위해서는 먼저 선결되어야 할 조건들이 존재한다. 우선 예측 및 전망에 있어 다양한 변수를 활용할수록 일반적으로 예측력이 높아지는 머신러닝 기법의 특성상, 유관 타 DB와의 연동을 통한 수혜자 속성 변수 확충 및 재정사업 정보 보완을 통해 분석에 활용할 데이터를 확보하는 것이 필요하다. 이를 통해 사전적 정책 효과 예측, 사후적 정책 분석에 활용이 가능할 것이며 예측 및 전망 모형 구축에도 고도화된 모델링이 가능하게 된다. 두 번째는 텍스트 분석의 기초 작업인 '재정 말뭉치', '재정분석 사전' 구축 작업이다. 열린재정과 dBrain+ 상에는 성과계획서, 조세지출, 예결산 문서 등이 탑재되어 있다. 그러나 이러한 문서에

서 텍스트 마이닝 작업을 하기 위해서 하는 전처리 작업에 따라 분석의 난이도와 품질이 결정되는데 고품질 및 의미 있는 분석 작업이 되기 위해서는 재정 용어 및 문서에 특화된 재정말뭉치 구축과 여기에서 재정 용어를 중심으로 사전 구축 작업이 선행되어야 할 것이다. 세 번째로는 재정사업 속성정보 정비 및 확충이다. 이는 앞서 언급한 수혜자 정보 중심의 타 DB와의 연계와 유사한데, 다만 재정사업은 분석에 있어 변수 및 분류 인자로 작용할 특성을 속성화하여 메타 정보화하는 작업이 필요하다. 현재 시스템상에는 사업 속성 정보에 대해 아주 일부에 관해서 입력이 이루어져 있기는 하지만 사업설명 자료에서의 텍스트 마이닝 작업을 통한 사업속성 추출 등의 작업 등을 통해 더 보완할 수 있을 것이다.

최근 데이터를 활용한 분석 결과를 정책수립과 의사결정에 활용하는 데이터 기반 행정이 본격 등장하면서 데이터의 생산 및 관리 분석을 넘어서 그 활용에 대한 관심도 높아지고 있다. 한계로는 머신러닝 기법은 탐색적인 특성을 갖고 있는데 전체 대상 데이터를 대상으로 일종의 공통적인 요인을 분류하는 방식에 가깝기 때문이다. 또한 머신러닝 기법은 논리적으로는 귀납적 추론 방식을 따른다고 볼 수 있는데 원인변수를 설정해서 그에 따른 결과를 예측하는 방식이 아닌 유사한 결과를 가져온 케이스들을 대상으로 그 특성을 파악하여 유사성과 차이점을 분류해서 대상의 성격을 예측하는 방법론적 특성을 가지고 있기 때문이다. 그렇기 때문에 특정 현상을 데이터를 바탕으로 도출하더라도 그 현상이 “왜” 혹은 어떤 환경에서 기인하여 발생한 것인지와 같은 인과관계를 보여주기는 어렵다. 결국 인과관계 및 원인을 규명하기 위해서는 데이터에 근거하는 단서를 따라 담당자 면담과 같은 실지조사나 문헌 조사 등 연구자의 정성적인 분석이 뒷받침되어야 한다. 현재 일부 재정사업을 대상으로 구축되어있는 사업정보시스템 내의 사업속성정보 등을 결합하여 분석에 활용한다면 더 정교한 모델구성이 머신러닝을 통해 구현될 수 있을 것이다.

그럼에도 불구하고 머신러닝과 같은 데이터 기법은 학습대상의 양과 계산 능력치가 인간을 넘어서 변수가 수백 개가 될수록 예측의 정확성이 높아질 수 있다. 또한 전통적인 통계적 방식을 넘어선 데이터 처리 및 조합 분석이 가능하다는 점에서 머신러닝 기법이 가진 독보적 분석의 힘이라고 볼 수 있다.

본 연구는 머신러닝 기법에 대한 전반적 소개 및 논의를 바탕으로 이를 재정정보에 적용할 수 있는 방안을 모색해보고자 하는 탐색적 논의라고 할 수 있다. 향후 후속 연구에서는 탐색적 논의를 넘어서 본 논의에서 제시한 비정형데이터 활용, 수혜집단 정보와 같은 변수 추가 및 재정사업 속성정보 시계열 및 연동을 통한 정책 효과 분석 등으로 실증적 분석이 이루어져야 할 것이다. 그리고 머신러닝을 활용한 분석은 데이터를 기반으로 학습과 분석이 이루어지는 방식이기 때문에 이미지 분석 및 분류에는 머신러닝에서 더 나아간 딥러닝 방식이 효과성이 더 검증되고 있기 때문에 이에 대한 후속연구가 필요할 것이다.

〈참고문헌〉

강송희. (2021). “미국 공공부문 인공지능 기술 활용동향 및 시사점”, 소프트웨어 정책연구소.

강주영. (2019). 「재정정보 분야 텍스트마이닝 활용 방안 연구」. 한국재정정보원.

개인정보보호위원회. (2022). 「가명정보 처리 가이드라인」.

건강보험심사평가원. (2021). 「파이썬을 활용한 데이터·AI 분석사례」.

국토연구원(2022). 「국유재산 관리혁신을 위한 프롭테크 활용방안」.

국토지리정보원. (2021). “디지털 트윈국토 실현 속도낸다”. 보도자료.

국회예산정책처(2023). 「2024년도 예산안 주요 사업 분석 (2) 국고보조사업 분석」.

기획재정부. (2022). 「국고보조금 부정수급 관리 가이드라인」.

김명규·김민경·지은초. (2022). 재정지출 효율화를 위한 불용유형분석 및 집행관리방안. 한국재정정보원.

김민경.(2022).「집행통계 작성을 위한 재정사업유형분류」. 한국재정정보원.

김민호·한재필. (2020). 「AI기반 정부지원 통합체계 구축방안」. 한국개발연구원

남현숙·안미소·장진철·이동현. (2023). 「국내·외 공공부문 AI 활용현황 분석 및 시사점」. 소프트웨어정책연구소.

박정수. (2020). 「dBrain 기반 재정성과관리 지원 강화 방안」. 한국재정정보원.

박정수·나원희. (2020), 「재정사업 시계열자료 구축을 위한 분류체계 분석」, 한국재정정보원.

안찬식·오상엽. (2021). "개선된 k-means 알고리즘을 적용한 사용자 특성 선호도 추천 시스템", 한국컴퓨터정보학회논문지, 15(8), pp.141-148.

옥동석·배근호. (2001). “재정정보 개선을 위한 한국의 정책과제”, 「한국행정학회학술발표논문집」, 2001, pp.461-479.

이재호·정소윤·강정석. (2019). 「인공지능 기술의 행정분야 활용에 관한 탐색적 연구」. 한국행정연구원.

장우영. (2020). 「4차 산업혁명 시대 빅데이터 기반 정책결정모델 구축 방안」, 정책기획위원회.

장원중·이정인. (2022). 「데이터 분석의 모든 것」. 아이리포.

정일환. (2021). “머신러닝 접근의 재정관리: 세입추세 예측모형 연구”, 「국정관리연

구」, 16(4), pp.1-28.
정재현·이환웅. (2020).「머신러닝 (Machine Learning) 을 활용한 조세· 재정정책의 평가와 설계」. 한국조세재정연구원.
차경엽·정근주(2021).「감사 및 조사 분야의 빅데이터 활용사례 조사」. 감사연구원.
차경엽(2022).「머신러닝 기법의 감사적용방안 연구」. 감사연구원.
최은진. (2021). “머신러닝 기반의 중소기업 대출 약정해지 예측모델에 관한 연구”,
한국재정정보원. (2022).「e나라도움 사용자 매뉴얼」.
한국재정정보원. (2023). 내부 세미나 자료.
행정안전부(2019). 클라우드·빅데이터·인공지능 융합에 대한 선진사례 연구.
Bart et al, (2015). Fraud anaytics using descriptive, predictive and social network techniques.
GAO(2003). Data Mining:Results and Challenges for Government Program Audit and Investigations.
Gentzkow, M., Kely, B., & Taddy, M. (2019). Text as data. Journal of
Economic Literature, 57(3), 535-74.
INTOSAI(2017). Big data Working Group Seminar paper.
Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., and Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6): 275-285.

머신러닝 기법을 적용한 재정정보 활용

발행연월 2023년 12월

발 행 인 박용주 한국재정정보원장

편 집 재정정보분석센터

발 행 처 한국재정정보원

(04637) 서울특별시 중구 퇴계로 10, 메트로타워

(Tel. 02-6908-8200)

인 쇄 처 경성문화사

I S B N 979-11-91094-33-6

FIS 한국재정정보원

www.fis.kr

한국재정정보원

[04637] 서울특별시 중구 퇴계로 10, 메트로타워
Tel. 02-6908-8200

93320
791191094336
ISBN 979-11-91094-33-6