공공부문 데이터 분석
업무 운영 가이드

행정안전부

NIA 한국지능정보사회진흥원

CONTENTS / 목차

데이터 분석 업무 추진 흐름도 ····· 6

01. 개요

1. 목적 ····· 9
2. 구성과 활용 범위 ····· 10
3. 데이터 분석 및 활용의 의의 ····· 11

02. 데이터 분석 추진 프로세스

1. 사전 준비 단계 ····· 16
   - 1.1 추진체계 구성 ····· 17
   - 1.2 데이터 분석과제 발굴 ····· 22
   - 1.3 분석 추진계획 수립 ····· 34
   - 1.4 분석 결과 활용계획 수립 ····· 54

2. 분석 추진 단계 ····· 58
   - 2.1 요건 정의 ····· 59
   - 2.2 데이터 수집 ····· 67
   - 2.3 데이터 전처리 ····· 75
   - 2.4 분석모델 설계 ····· 89
   - 2.5 모델 구현 ····· 96
   - 2.6 시각화 ····· 107
   - 2.7 모듈 개발 ····· 110
   - 2.8 테스트 및 안정화 ····· 116

03.

분석모델 활용 및 고도화

1. 분석 결과 활용 ····· 120
2. 분석모델 고도화 및 유지관리 ····· 121
2.1 분석모델 고도화 ····· 121
2.2 유지관리 ····· 123

04.

첨부

데이터 분석과제 단계별 주요 점검 항목 ····· 126
데이터 분석 주요 용어 ····· 132

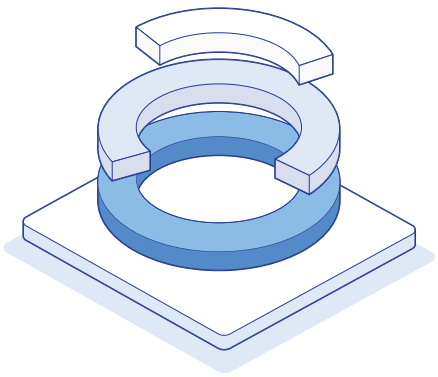


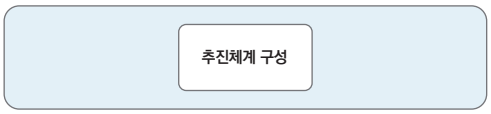
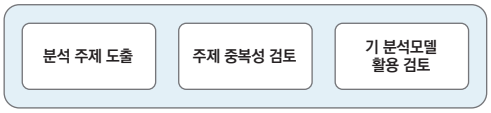
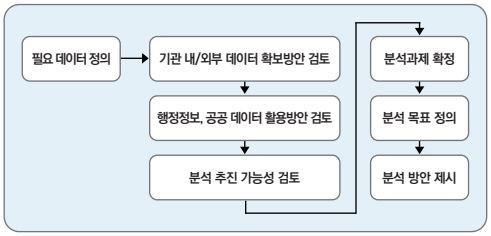
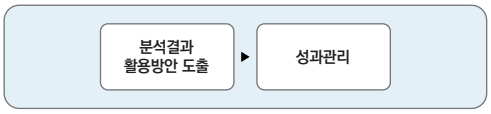
표 목차

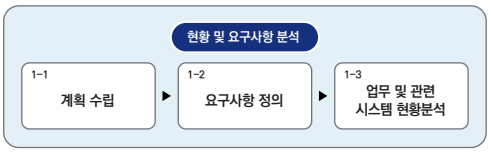
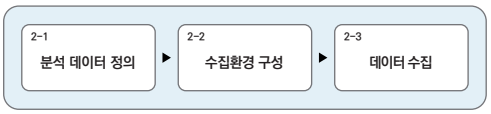
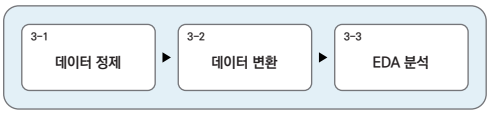
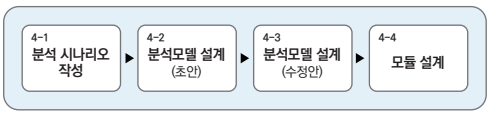
[표 1] 이해관계자의 점검 및 고려사항 ····· 20
[표 2] 데이터 분석과제 추진을 위한 역할과 책임(R&R) 예시 ····· 21
[표 3] 주제 도출 시 고려사항 ····· 24
[표 4] 2024년 행안부 추진 표준분석모델 예시 ····· 26
[표 5] 과제 평가 기준표 예시 ····· 30
[표 6] 국내 주요 데이터 현황 ····· 38
[표 7] 공공데이터포털 보유데이터 건수(2025. 2. 기준) ····· 41
[표 8] 국내 공공분야 주요 오픈 API ····· 44
[표 9] 사업비 산정 단계별 주요 내용 ····· 49
[표 10] ITSQF(IT Sectoral Qualifications Framework) 직무체계 및 정의 ····· 50
[표 11] 품질관리 목표 ····· 80
[표 12] 비식별화 전문기관 ····· 81
[표 13] 유지관리 점검항목 ····· 123
[표 14] 유지관리 대상별 관리 포인트 ····· 124

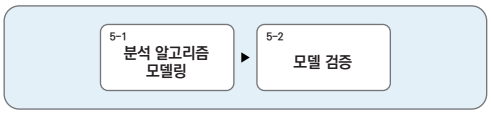
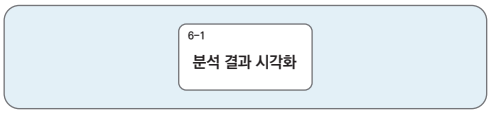
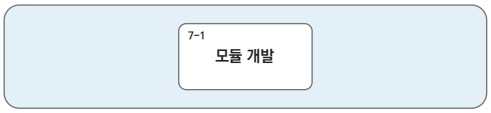
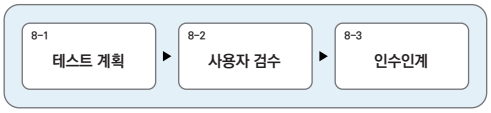
그림 목차

[그림 1] 공공데이터 분석업무 추진 프로세스 ····· 10
[그림 2] 데이터 분석 추진 프로세스 ····· 16
[그림 3] 분석과제 추진체계 예시 ····· 18
[그림 4] 데이터 이해력 ····· 21
[그림 5] 분석 주제 후보 도출 절차 예시 ····· 23
[그림 6] 분석 주제 시급성/난이도별 순위 선정 예시 ····· 32
[그림 7] 추진단계별 점검항목 ····· 58
[그림 8] 요건 정의 단계의 주요 활동 ····· 59
[그림 9] 데이터 수집 단계의 주요 활동 ····· 67
[그림 10] 데이터 전처리 단계의 주요 활동 ····· 76
[그림 11] 탐색적 데이터 분석(Exploratory Data Analysis) ····· 83
[그림 12] EDA 수행 과정 ····· 85
[그림 13] 분석 모델링 단계의 주요 활동 ····· 89
[그림 14] 모델구현 단계의 주요활동 ····· 96
[그림 15] 시각화 설계서 ····· 109
[그림 16] 테스트 및 안정화 단계의 주요 활동 ····· 116
[그림 17] 통합 테스트 절차 예시 ····· 117
[그림 18] 분석모델 활용 및 고도화 추진 절차 ····· 119

업무추진 흐름도: 추진 절차별 주요 활동

사전준비 – 추진 절차	주요 내용
1. 추진체계 구성 	• 추진체계 구성 – 현업부서, 정보화부서, 데이터 전문가 (분석사업 관리자)
2. 데이터 분석과제 발굴 	• 분석 주제 도출 – 분석 필요성, 분석 가능성 • 주제 중복성 검토 • 기 분석모델 활용 검토 – 범정부데이터분석시스템 (www.insight.go.kr) 등
3. 데이터 확보 방안 검토 	• 필요 데이터 정의 • 기관 내/외부 데이터 확보방안 검토 • 행정정보, 공공 데이터 활용방안 검토 • 분석 추진 가능성 검토 • 분석과제 확정 • 분석 목표 정의 • 분석 방안 제시
4. 분석결과 활용계획 수립 	• 분석결과 활용방안 도출 • 성과관리

사업추진 – 추진 절차	주요 내용
1. 요건 정의  	• **계획 수립** – 일정계획수립, 인터뷰 계획 수립 • **요구사항 정의** – 인터뷰 수행, 선생사업 분석 및 자료조사 • **업무 및 관련 시스템 현황 분석** – 수집대상 데이터 현황분석, 수집환경 분석
2. 데이터 수집  	• **분석 데이터 정의** – 데이터 수집 대상 목록 및 정리 • **데이터 수집 환경 구성** • **데이터 수집** – 데이터 보유기관 / 기업 및 부서 협의
3. 데이터 전처리  	• **데이터 정제** – 데이터 표준화 및 융합 • **데이터 변환** • **데이터 기초 통계 분석** • **탐색적데이터분석(EDA)**
4. 분석모델 설계  	• **분석 시나리오 작성** – 분석 대상, 범위, 데이터 정의 – 분석 목표별 구현모델 정의 – 분석 방법론 정의 • **분석모델 설계(초안, 수정안)** • **모듈 설계** – 모듈의 기능 정의 및 설계 – 표준화 모듈 설계 – 데이터 정제 모듈 설계

사업추진 – 추진 절차	주요 내용
5. 모델 구현 	• 분석 알고리즘 모델링 • 알고리즘 적용 • 모델 검증
6. 시각화 	• 시각화 설계 및 개발 • 분석 결과 활용 • 결과 공유
7. 모듈 개발 	• 모듈 개발 – 데이터 저장 – 시스템 연계 • 데이터 관리 및 수집 자동화
8. 테스트 및 안정화 	• 테스트 계획 수립 • 통합 테스트 • 사용자 검수 • 인수인계 및 교육 • 분석모델 고도화 및 유지관리

1. 개요

1.1 목적

> 본 가이드는 공공데이터의 다양한 가치에 대한 인식을 제고하고 효율적인 데이터 분석과 활용으로 국가, 지자체 및 공공기관 업무의 효과성 극대화를 도모할 수 있도록 도움을 주고자 제작되었다.

본 가이드는 데이터 분석을 통한 공공정책의 의제발굴 및 정책결정 그리고 추진전략 수립의 방향을 이해하고 올바르게 적용할 수 있는 지침을 제공한다.

빠르게 변화하는 인공지능 기술, 데이터 분석 관련 신기술 및 정책환경을 반영하여 최신 데이터 분석 업무 운영 가이드를 마련하였다.

본 가이드는 과제 발굴 및 기획, 역할분담, 데이터 품질 및 보안 관리, 분석모델 설계 및 검증, 성과 검증 등 데이터 분석업무 전반을 모두 다루고 있으며, 최신 인공지능 및 데이터 분석 기술 동향을 반영하여 실무자가 효과적으로 활용할 수 있도록 제작했다.

특히, 공공데이터 분석업무 담당자가 업무 추진 시 점검해야 할 항목과 점검항목을 단계별로 제시함으로써, 실제 업무에 쉽게 활용할 수 있도록 작성하였다.

- 본 가이드를 통해 정보화 담당자는 물론 데이터 분석 경험이 적은 현업 실무담당자도 데이터 분석과제를 기획하고 관리할 수 있다.
- 과제 목표와 방향 설정, 과제 추진 및 관리를 위한 체계적인 절차와 점검 항목 등을 검토할 수 있다.
- 데이터 분석 시의 추진 프로세스를 고려한 실무지침을 현장 업무에 직접 적용할 수 있다.

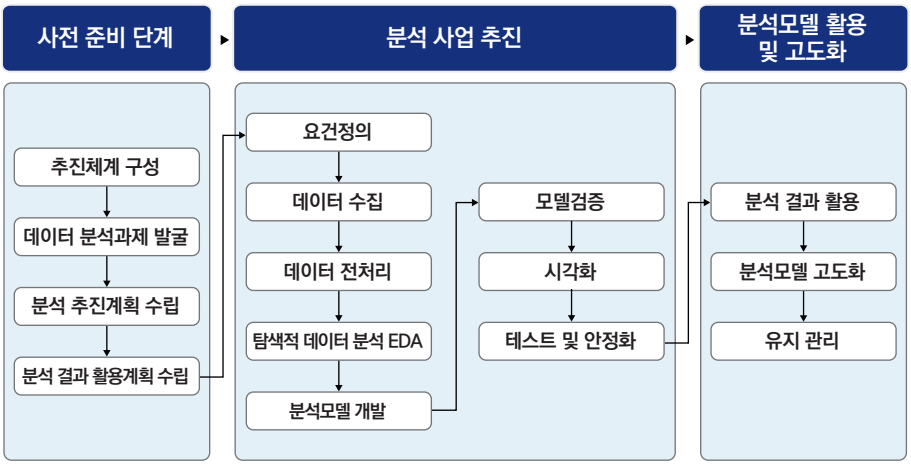
1.2 구성과 활용 범위

본 가이드는 공공기관이 데이터 분석 프로젝트를 계획, 발주, 실행, 평가하는 데 있어 실질적인 기준과 방향을 제공한다. 특히 과업지시서 작성, 사업 관리, 분석결과 활용방안 마련 등 분석 업무의 여러 부분에 활용할 수 있도록 제작되었다.

본 가이드의 구성은 사전 준비 단계, 분석 사업 추진 단계, 분석모델 활용 및 고도화의 3단계로 구성되어 있다.

'사전 준비 단계'에서는 가이드를 활용하여 분석과제 추진을 준비할 때의 필수 활동들을 상세하게 설명하였다. **'분석사업 추진단계'**에는 실제 분석 업무 과정에서 업무 담당자로서 체크할 점검사항을 중심으로 작성했고, **'분석모델 활용 및 고도화 단계'**에서는 분석모델의 활용, 유지보수, 고도화 업무 과정에서 활용할 수 있도록 기술했다.

[그림 1] 공공데이터 분석업무 추진 프로세스



본 가이드는 주로 데이터 분석 수행 업무를 대상으로 작성되었으며, 데이터 및 시스템 구축 사업에 대한 내용은 포함하지 않았다.

본 가이드는 공공부문에서의 데이터 분석과 활용에 대한 이해를 높이고, 과제 타당성 검토와 방향성 확보 등의 분석 기획 업무 및 사업 관리 등에 활용할 수 있다.

본 가이드에서 제시하는 업무 절차는 필수 적용 절차가 아닌 참고용으로, 각 기관의 데이터 활용 목표와 운영 환경 및 분석 대상 데이터의 특성에 따라 유연하게 적용할 수 있다.

1.3 데이터 분석 및 활용의 의의

공공부문의 데이터 분석은 궁극적으로 객관적 데이터를 기반으로 한 정책 의사결정의 타당성을 확보하기 위한 과정이다. 이에 따라 본 가이드의 주요 대상자인 분석 실무자와 관리자 역시 데이터 기반 행정의 정착과 고도화에 있어 핵심적인 역할을 수행하게 되며, 정책 기획과 집행 전반에 있어 데이터에 기반한 행정 운영은 정책 수립의 선택이 아닌 필수 조건이 되고 있다.

1.3.1 데이터 분석이란

데이터 분석은 원시 데이터를 실행 가능한 인사이트로 변환하는 과정이다. 이는 특정 대상의 패턴과 트렌드 등 유용한 인사이트를 발견하기 위해 데이터를 수집·정제·가공·분석하는 전 과정을 포함한다. 이때 통계 분석, 머신러닝 및 데이터 마이닝과 같은 다양한 방법을 사용한다.

1.3.2 데이터 분석의 필요성

데이터 분석을 활용하면 정책 및 업무 프로세스를 체계적으로 구성하고, 보다 효과적인 의사결정을 지원할 수 있다. 데이터 분석은 프로세스와 서비스의 가시성을 높여 행정 운영과 문제에 대해 더 깊이 이해할 수 있는 인사이트를 제공한다. 이를 바탕으로 단순한 데이터 활용을 넘어 패러다임을 전환함으로써 정책 운영을 최적화하고 업무의 생산성을 높일 수 있다. 뿐만 아니라, 데이터 분석을 통해 재난이나 위기 상황에 신속하게 대응할 수 있다.

1.3.3 공공데이터 분석의 효과

데이터는 과거에 어떤 일이 일어났는지, 그 원인은 무엇인지, 앞으로 무엇이 발생할지를 파악하는 단서를 제공한다. 따라서 데이터 분석을 통해 문제를 명확히 이해하고, 예측을 기반으로 합리적인 의사결정을 내릴 수 있다.

공공데이터 분석은 공공기관이 생성하거나 다른 공공기관 및 법인, 단체 등으로부터 취득하여 관리하고 있는 데이터를 수집·가공·분석·시각화 등의 방법을 통해 정책수립 및 의사결정에 활용한다. 다음은 공공기관 업무에 데이터 분석을 활용한 일부 사례들이다.

정책 발굴 분야

(시민 목소리 분석) 특정 주제에 대한 시민의 목소리를 이해하고 그 추이를 분석함.
민원 접수 데이터와 소셜 데이터를 기반으로 정책의제 발굴과 전략 방향 설정 등에 활용할 수 있음.

- (예시) 중학교 입학배정 민원 분석을 통한 입학배정 제도 개선 **[2023, 국민권익위원회]**
- (예시) 숲길 방문객 데이터 분석을 통한 지리산 둘레길 산촌 체험마을 지정 운영 **[2024, 산림청]**

(사회 이슈 분석) 사회 이슈의 자동 감지와 연관 주제의 동향 분석을 통해 잠재 정책수요를 발굴함. 주요 일간지, 소셜 데이터, 민원 접수 데이터 등을 분석하여, 지역별 이슈를 도출하고 지역별 맞춤형 대국민 서비스 전략을 수립할 수 있음.

- (예시) 상권변화 요인을 활용한 상권 부실 징후 예측 **[2024, 소상공인시장진흥공단]**

(맞춤형 민원 서비스) 지역별, 기관별 민원인의 행동 및 요구사항 분석을 통해 민원 행정의 개인 맞춤형 서비스를 구현함. 지역별 기관별 서비스 사용로그, 게시판 및 민원 접수 데이터, 포털 게시판과 질의응답 데이터 등을 기반으로 분석함.

- (예시) 코로나19 긴급지원금 사각지대 해소 **[2021, 보건복지부]**

국민 안전 분야

(범죄예방과 대응) 지역별, 시간별, 이벤트별, 유형별, 범죄 패턴 분석과 이를 활용한 지역별, 시기별 맞춤형 범죄예방 전략 도출이 가능함. 범죄 및 경찰업무기록, 보고서, 뉴스 및 소셜 데이터 등이 주요 분석 대상임.

- 보이스피싱, 사기범 식별 등 범죄조직 잡는 한국형 음성분석모델
  [2023, 행안부 통합데이터분석센터/국과수]

(도시관제 및 재난대응) 도심 내의 문제를 조기 파악하거나 재난을 조기에 감지하여 대응할 수 있음. 도심 및 재난지역 시민 목소리를 바르게 이해하여 관련 정책의제 발굴에 활용함.
센서 데이터, CCTV, 소셜 데이터 등이 주요 분석 대상임.

- 해양안전 세이프존 확대를 위한 빅데이터 분석 **[2021, 해양경찰청]**
- 피해절감을 위한 소방용수시설 취약지수 개발 **[2021, 소방청]**
- AI 기반 적재불량 자동 판단시스템 **[2023, 한국도로공사]**

조세 분야

(체납유형별 징수활동) 체납자별 빅데이터 분석, 등급과 유형분석을 통해 납부 가능성을 예측 하여 체납유형에 따른 맞춤형 징수 활동을 수행함.

- (예시) 데이터 분석으로 지능형 지방세정 실현 **[2021, 행안부]**

의료·복지 분야

(의료 및 복지서비스 강화) 의료보험 비용 분석을 통한 사업 최적화, 부당 청구 방지, 복지 정책 입안과 만족도 분석, 지역별 복지 불균형 해소 등에 활용됨. 의료 및 복지 지출데이터, 민원센터 로그, 소셜 데이터, 서비스 기관 홈페이지 및 포털 게시판 데이터, 주요 일간지 등이 주요 분석 대상임.

- (예시) 교통사고 환자 이상패턴 탐지모델 개발 **[2021, 건강보험심사평가원]**
- 주요 질병별 의약품품절예측모델 개발 **[2025, 행안부 공공지능데이터분석과/식약처]**
- 데이터 기반 안성맞춤형 복지정책 수립 **[2021, 안성시]**

(질병 및 전염병 관리) 유행 전염병, 질병에 대한 예측 및 대응과 지역별 분포 분석에 활용됨. 연도별 매크로 분석, 가축 전염병과 환경 데이터, 이동 경로 데이터 등이 주요 분석 대상임.

- (예시) 2급 감염병 발생위험도 예측 분석 **[2020, 서울시·경북]**
- 맞춤형 치매 진단 검사로 다각적인 치매 관리 **[2021, 김해시]**

교통·환경 분야

(환경감시 및 대응) 환경오염과 변화 상황의 모니터링을 통한 대응 및 환경정책 관련 중장기적 수행 전략 수립 등에 활용됨. 국가 및 도심 센서 네트워크로부터 수집된 환경 데이터, 관련 보고서 등이 주요 분석 대상임.

- (예시) 광역상수도 관로 누수감지 모델 개발 **[2023, 한국수자원공사]**
- 국·공유지 무단 점유 감시 **[2021, 통영시]**

(교통상황관리 및 최적화) 교통흐름 모델링과 예측, 최적화 시스템 구현, 교통 신호체계와 유지보수 정책에 활용됨. 도로 센서 네트워크, 사건·사고 로그, 날씨, 명절 및 스포츠 등 이벤트 데이터 등이 주요 분석 대상임.

- (예시) CCTV 영상데이터를 활용한 도로 교통량 분석모델 **[2023, 행안부]**
- 교통 데이터 로드맵으로 맞춤형 교통안전 지원 모델 개발 **[2021, 한국교통안전공사]**

교육 분야

(교육정책 및 현안분석) 교육환경 개선과 교육 민원 처리, 합리적 교육 예산 집행과 절감 등을 위해 활용됨. 교육 예산 집행 데이터, 각종 보고서 및 소셜 데이터, 민원 로그데이터 등이 주요 분석 대상임.

- (예시) 신규 공동주택 내 초등돌봄 수요예측 모델 개발 **[2021, 복지부·교육부·국토부]**

1.3.4 AI 및 공공데이터 분석의 최근 현황

산업 전반에서 AI와 데이터 분석 기술은 자동화와 효율화를 이끄는 핵심 도구로 자리잡고 있다. 철강 업계에서는 열연공장에서 발생하는 방대한 생산 데이터를 AI로 분석해 슬래브 절단 최적화를 이뤘고, 자동차 분야에서는 AI 기반 예지보전 시스템을 통해 설비 고장을 사전에 감지하여 보수함으로써 생산 중단 일수를 줄였다. 의료 분야에서는 수많은 영상 데이터를 학습한 AI 진단 솔루션을 통해 암 진단의 정확도를 높이며 국내외 병원에 적용되고 있다. 금융권에서는 고객 데이터 분석을 바탕으로 AI 업무비서를 도입해 내부 업무 자동화와 맞춤형 금융상품 제안에 활용하고 있으며, 유통과 물류에서는 택배 물량, 도로 상황, 배송 시간 등의 실시간 데이터를 분석해 배송 경로를 최적화하고 자율주행 트럭과 연계한 스마트 물류 시스템을 추진 중이다. 이처럼 데이터 기반의 AI 활용은 산업별로 특화된 방식으로 정착되며, 고도화된 분석을 통해 현장 문제 해결과 경쟁력 향상을 동시에 이끌고 있다.

AI가 미래 30년을 좌우할 핵심 기술이자 데이터 기반 혁신의 중심으로 부상하면서, 정부는 데이터 분석과 AI 기술을 적극적으로 도입하여 정책 결정의 정확성을 높이고, 행정 효율성을 극대화하기 위해 노력하고 있다. 또한, 공공부문 AI 도입률 95%를 목표로 설정하면서, 공공기관에서는 AI를 활용해 행정 프로세스를 최적화하고 정책 수립을 정교하게 지원할 것이다.

정부의 각 부문에서도 데이터 분석을 통한 정책 의사결정 및 디지털 행정은 적극 추진되고 있다. 특히, 행정안전부는 행정·공공기관 간 데이터 공동활용을 촉진하고, 정책 수립에 데이터 활용을 활성화할 수 있도록 한 「**데이터기반행정 활성화에 관한 법률**」을 2021년 시행했고, 부처 간 정보 칸막이를 해소하고 데이터기반행정 촉진을 위해 본 법령을 2024년 개정하였다. 이는 공공부문의 데이터 활용도를 극대화하고, 과학적 행정 정책의 정밀도를 높이기 위해 **'공유데이터'** 개념을 도입하고 데이터 공유 체계를 강화하는 것을 핵심 목표로 삼고 있다.

과거에는 공공기관별로 데이터를 개별적으로 관리하면서 기관 간 연계가 원활하지 않고, 동일한 데이터를 중복으로 구축하는 비효율적인 문제가 자주 발생했다.
이번 개정안은 이러한 문제를 해결하기 위해 공공기관이 보유한 데이터를 보다 효과적으로 공유하고 활용할 수 있도록 법적 기반을 정비하였다. 이는 데이터 분석 담당자에게 중요한 의미가 있는데, 기존에 다소 폐쇄적으로 개방되던 공공데이터가 유기적으로 연결되면서 보다 포괄적인 데이터 분석이 가능해지고, 실질적인 정책 지원을 위한 데이터 활용성이 대폭 증가할 것으로 기대되기 때문이다.

행정안전부는 제2차 데이터기반행정 활성화 기본계획으로 **'데이터에 기반한 과학적 행정 추진으로 똑똑하게 일 잘하는 정부 구현'**의 비전을 수립하고, 2026년까지 다음과 같은 3개 기본 추진전략을 수행한다.

❶ 범정부 데이터공유플랫폼을 통한 기관 간 데이터 칸막이 해소
❷ 정책 맞춤형 데이터 분석으로 과학적 행정 추진 가속화
❸ 데이터 공유·분석·활용 일상화로 데이터기반행정 문화 정착

최근 정부는 2조 원 규모의 **'국가 AI컴퓨팅 센터'**를 구축하여 방대한 데이터를 효과적으로 처리하고 분석할 수 있는 환경을 조성할 계획이다.
여기에 더해 2024~2027년 동안 민간 AI 투자 65조 원을 유도하여 AI 및 데이터 분석 기술을 활용한 행정 혁신을 가속하고 있다. 이러한 변화는 데이터 분석 담당자들이 더 강력한 연산 자원과 최신 AI 기술을 활용하여 복잡한 행정 문제를 해결하고, 정책 결정을 뒷받침하는 데 있어 중요한 기회를 제공한다.

이러한 추세에서 기관의 데이터 분석 담당자는 단순히 데이터를 분석하는 역할을 넘어, 정책 결정 과정에서 보다 전략적인 분석과 데이터 기반 인사이트를 제공하는 핵심적인 역할을 수행해야 한다. 데이터 공유가 원활해짐으로써 다양한 데이터를 결합하여 정책의 효과성을 예측하고, 맞춤형 분석모델을 설계하는 등 보다 정교한 분석이 가능해지기 때문이다.

또한, 기관 간 협업이 강화되면서 데이터 분석 담당자는 공공기관 내·외부에서 데이터를 연계하여 실질적인 정책 효과를 높이는 방향으로 분석 역량을 확장해야 한다.
이를 위해 AI 및 데이터 기술을 적극적으로 활용하는 것은 물론, 데이터의 품질과 신뢰성을 유지하면서도 기관 간 협력을 통한 종합적인 분석을 수행할 수 있는 전문성을 갖추어야 한다.

본 가이드의 독자인 데이터 분석 담당자와 관리자의 역할 또한 이에 맞춰 더욱 확대되고 있으며, 데이터 기반의 과학적 행정이 정책 수립의 핵심 요소로 자리 잡고 있다.

2. 데이터 분석 추진 프로세스

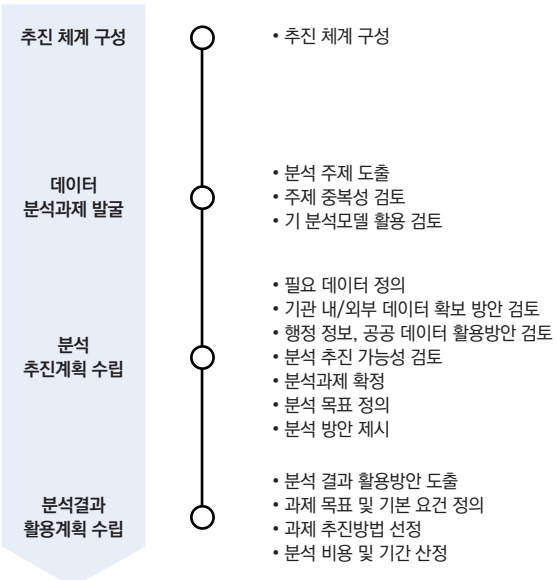
1. 사전 준비 단계

사전 준비 단계에서는 데이터 분석과제를 선정하는 과정을 다룬다. 이는 분석업무 담당자가 가장 어려워하는 작업이기에, 어떤 과정을 거쳐 분석과제를 발굴하고 과제 후보들을 어떤 기준으로 선별하여 최종 추진 과제를 선정할지에 대한 절차를 소개한다.

과제가 확정되면, 과제의 필요성과 문제를 정의하며 데이터 분석을 통해 어떻게 문제를 해결할 수 있는지에 대한 방안을 정리하고 분석 결과를 어떤 목적으로 활용할 수 있을지 정의해야 한다.

따라서 이 장에서는 추진체계 구성부터 데이터 분석과제 발굴, 분석 추진계획 수립까지의 업무를 처리할 수 있도록 안내한다.

[그림 2] 데이터 분석 추진 프로세스



1.1 추진체계 구성

분석 추진체계는 데이터 분석을 효과적으로 수행하기 위한 조직 구조와 프로세스를 의미한다. 데이터 분석과제는 복잡한 문제를 해결하기 위해 수행되는 경우가 많으며, 이는 다양한 부서와의 협력을 필요로 한다. 따라서 분석 추진체계 구성을 통해 공공기관 내·외의 역할과 책임을 명확히 정의하고, 협업 체계를 마련하는 것이 중요하다.
이를 통해 다양한 이해관계자 간 협력에 기반해 원활한 의사결정을 지원할 수 있다.

분석 이해관계자 구성

데이터 분석 추진체계의 조직은 통상적으로 데이터 분석결과를 활용하는 현업부서 담당자, IT 분야의 정보화부서 담당자, 내부의 데이터 전문가 또는 외부 전문가로 구성된다. 그리고 데이터 분석업무를 수행·관리하는 분석과제 총괄 관리자가 필요하다.

정보화부서는 데이터, 정보, 기술 등에 대한 전문성을 갖추고 있고, 현업부서는 국민의 요구를 직접 파악하고 의사결정을 수행하는 역할을 한다. 따라서, 두 부서 간 협업을 통해 시너지를 창출하고, 데이터 분석이 실질적인 정책 및 업무개선으로 이어질 수 있도록 해야한다.

또한, 창의성을 갖춘 데이터 분석과제의 총괄 관리자가 함께 참여하여 아이디어를 발굴하고 현업 및 정보화 부서를 지원할 수 있는 체계가 이상적이다. 데이터 전문가인 데이터 분석과제 총괄 관리자는 현업 부서와 정보화 부서를 지원하는 역할로서 상시 참여하도록 구성한다.

데이터 분석과제의 총괄 관리자는 조직 내에서 정보화 사업과 도메인 업무를 두루 경험한 자를 선정하여, 프로젝트를 총괄하고 공정관리와 실적 점검, 산출물 관리 및 검수, 데이터 현황조사 지원, 업무 분석 지원 등이 원활히 진행될 수 있도록 하여야 한다.
분석과제 총괄 관리자는 IT부서와 현업부서의 다양한 이해관계자들과 협업을 위해 원활한 의사소통 능력을 갖추어야 한다.

또한, 외부 전문가를 자문조직으로 구성하여 분석과제에 대한 자문 등 과업 수행에 관한 지원을 받고, 사업이 효율적으로 추진될 수 있도록 정책적·기술적 조언을 받는 것이 필요하다.

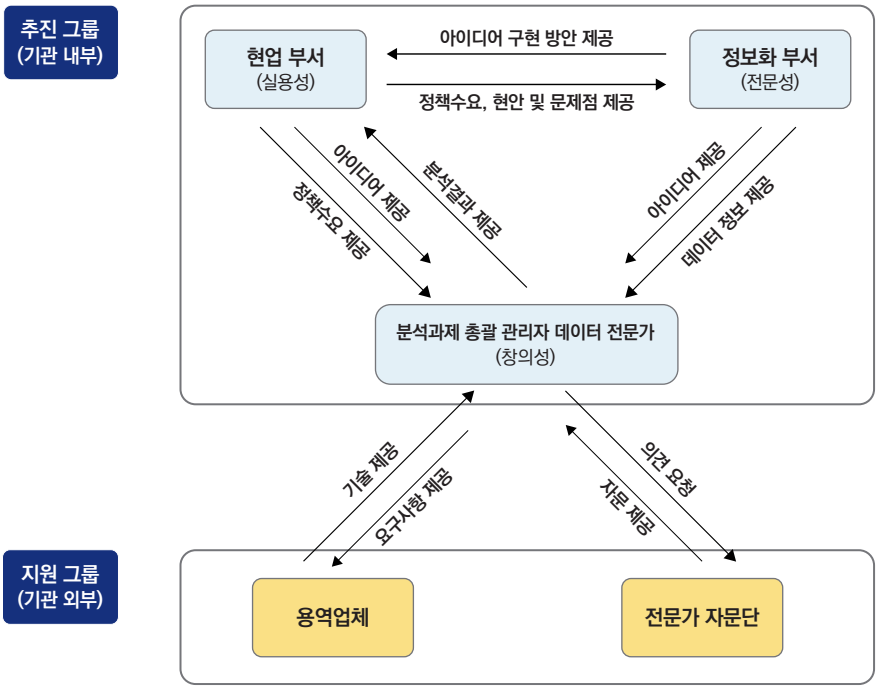
자문조직과 별도로 기술지원이 필요할 때는 외부의 용역업체 활용을 검토할 수 있다.

과제 추진 및 분석 활용에 대한 기관 내·외 다양한 이해관계자간 효율적인 협업 체계를 갖추어야 하며 또한 조직 내 다양한 분야에서 과제를 발굴하여 전 기관의 데이터 활용 관심도를 제고해야 한다.

- 국민 또는 사용자와 직접 현장에서 업무를 수행하며 현안을 가장 잘 이해하는 현업 부서가 과제 발굴 단계부터 참여하는 것이 중요하다.

따라서, 데이터 분석과제를 추진하기 위한 추진체계의 구성이 우선되어야 하며, 데이터, 정보, 기술 등에 대한 전문성을 지닌 정보화 부서와 정책 분야별로 국민의 요구를 가까이에서 파악할 수 있는 현업 부서와의 시너지 효과가 창출될 수 있도록 해야 한다.

[그림 3] 분석과제 추진체계 예시



- 현업 부서에서 분석 수요가 제기되면, 현업 부서는 이를 정보화 부서와 협의해야 한다.
- 정보화 부서는 분석수요를 해결할 수 있는 데이터의 유무를 확인하고, 관련 정보를 데이터 전문가에게 제공하여 문제 해결 방안에 대한 지원을 받을 수 있다.

- 분석과제 총괄 관리자는 내부의 데이터 전문가로서 분석과제를 총괄적으로 진행한다. 분석을 내부적으로 추진할 수 있는 경우는 내부에서 추진한다. 용역 등 외부의 기술적 도움이 필요할 경우에는 발주 등 관련 절차를 진행한다.
- 전문가 자문단은 내부 추진과 외부 추진 두 경우 모두 수행하는 것이 바람직하다. 분석과제 총괄 관리자나 정보화 부서는 전문가의 자문 내용을 종합하여 추진방향 및 아이디어를 도출하고, 실제 어떻게 구현할 수 있는지 등에 관한 전문적인 지식과 정보를 현업 부서에 제공한다.
- 현업 부서는 제공받은 분석 결과를 활용하여 정책을 최종적으로 기획, 조정하여 수요 맞춤별 정책을 제공한다.
- 이상의 과정이 유기적으로 연계되고 각 주체 간의 교류와 협력이 활발히 이루어질 때 분석과제가 효과적으로 수행되고 그 의미를 극대화할 수 있다.

현업부서에서는 수요에 따른 분석과제 후보들을 평가하여 최종적으로 수행할 데이터 분석 과제를 결정하고, 정보화 부서와 데이터 전문가는 이를 지원한다.
[표1]은 데이터 분석 관련 이해관계자가 점검하거나 고려해야 해야 할 사항을 정리한 것이다.

[표 1] 이해관계자의 점검 및 고려사항

이해관계자		점검 및 고려사항
추진그룹 (기관내부)	현업 부서 (분석결과 활용자)	• **요구사항 정의:** 실제 업무 및 정책에 활용하기 위한 데이터 분석 요구사항 마련 • **성과지표 설정 참여:** 데이터 분석결과 활용의 효과성 평가를 위한 KPI 정의 • **교육 및 역량 강화:** 데이터 활용 능력 향상을 위한 교육 프로그램 참여
	정보화 부서 (데이터, 정보, 기술 제공자)	• **분석 수요를 해결할 수 있는 공공데이터의 유무 확인 데이터 제공 가능성 판단** • **데이터 분석과제 추진을 위한 기술 제공**
	분석과제 총괄 관리자 (내부 데이터 전문가)	• **프로젝트 관리:** 단계별 진행 상황을 모니터링하고 목표 달성을 위한 조정 역할 수행 • **분석 수행과정에서의 의사결정 지원** • **수요 조사 프로세스 설계:** 분석과제 발굴을 위한 체계적인 수요 파악 방안 마련 • **분석 결과 활용계획 수립:** 수요 조사 데이터를 기반으로 결과 활용방안 도출
지원그룹 (기관외부)	용역업체 (필요시)	• **분석 실무 수행:** 데이터 전처리, 모델링, 시각화 등 분석의 실제 기술 작업 수행 • **성과물 납품:** 요구된 분석 결과 및 보고서를 프로젝트 일정에 맞춰 산출 • **기술 도입 및 구현:** 최신 분석 기법 및 도구를 실제 프로젝트 환경에 구현하고 적용 • **계약 기반 성과 관리:** 계약 범위와 일정, 산출물 품질에 대한 성과 책임 이행 • **전문성 보완:** 내부 역량 부족 시 외부 전문가를 초빙해 분석 기술 지원
전문가 자문단		• **의사결정 지원:** 분석 방향성, 방법론 선정, 분석결과 해석 등 전략적 판단에 대한 조언 제공 • **품질 및 타당성 검토:** 분석 과정과 결과의 객관성·신뢰성 확보를 위한 외부 검증 역할 수행 • **지식 공유 및 역량 강화:** 최신 기술·동향·정책을 공유하여 내부 역량 강화 및 교육적 효과 부여 • **독립적 자문:** 프로젝트 외부 전문가로서 독립적이고 객관적인 시각 제공

분석과제 추진을 위한 역할과 책임

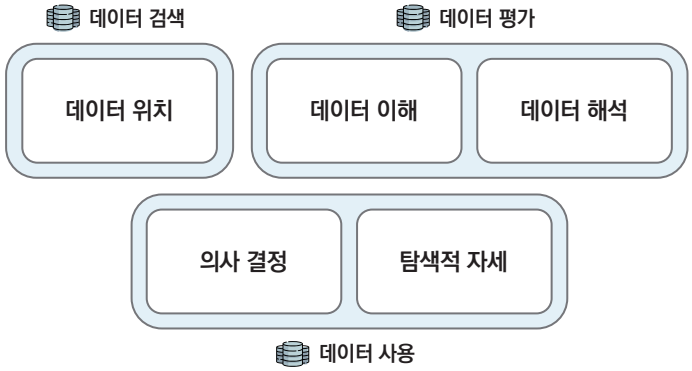
성공적인 데이터 활용을 위해서는 기관의 데이터 분석 역량 내재화와 분석 업무의 체계성이 중요하다.

데이터를 통해 기관이 혁신적인 결과를 이뤄내려면 데이터 과학자나 분석가들이 데이터를 잘 활용하는 것만으로는 충분하지 않으며, 기관 내부의 구성원들이 반드시 데이터 활용에 관련된 데이터 이해력(Data Literacy)를 갖추는 게 바람직하다.

- **'데이터 이해력을 갖춘 사람'**이란 의사결정이 필요할 때, 주어진 데이터를 이해하고 분석하여, 분석 결과를 근거로 명확하게 소통하고 합리적인 판단을 하는 사람을 말한다.

데이터 이해력을 높이려면 가장 먼저 비즈니스 영역에 어떤 데이터가 존재하고, 어떤 데이터가 중요하며, 이 데이터들이 어떻게 사용되는지 알아야 한다. 더 나아가 기초적인 통계 지식을 익혀 데이터 간의 연관 관계와 정량적, 정성적 데이터의 의미를 이해할 필요가 있다.

[그림 4] 데이터 이해력



데이터 분석 프로젝트에 적합한 팀을 구성하거나 업무 분담을 효율적으로 계획하는 데 필요한 데이터 분석의 직무별 기술과 필요 역량은 아래와 같다.

[표 2] 데이터 분석과제 추진을 위한 역할과 책임(R&R) 예시

역할	책임	필요 역량
데이터 기획	• 데이터 활용 전략 수립 • 데이터 수집 및 관리 계획 수립 • 데이터 요구사항 정의	• 비즈니스 분석 능력 • 데이터 관리 및 품질 보증 지식 • 프로젝트 관리 및 커뮤니케이션 능력
데이터 분석	• 데이터 탐색 및 패턴 분석 • 통계 모델 및 데이터 시각화 • 인사이트 도출 및 보고서 작성	• 통계 및 데이터 분석 툴 사용 능력 • Python, R, SQL 등의 프로그래밍 능력 • 데이터 시각화 스킬
분석 적용 시스템 개발	• 데이터 기반 서비스 및 응용 프로그램 개발 • 분석 결과를 서비스에 통합 • 실시간 데이터 처리	• 소프트웨어 개발 능력 • API 설계 및 개발 지식 • 시스템 최적화 및 성능 모니터링 능력
데이터 엔지니어링	• 데이터 파이프라인 구축 및 관리 • 데이터 정제 및 가공 • 데이터 저장소 설계 및 관리	• 데이터베이스 및 클라우드 플랫폼 지식 • ETL 도구 사용 능력 • 대규모 데이터 처리 기술
AI/ML 엔지니어링	• AI/ML 모델 설계 및 학습 • 모델 성능 최적화 • 모델 배포 및 운영	• 머신러닝 알고리즘 및 프레임워크 지식 • 딥러닝 및 자연어 처리 경험 • 모델 배포 및 MLOps 지식

데이터 분석과제를 성공적으로 수행하기 위해서는 명확한 전담 부서 및 역할 설정이 중요하다.

단계별 명확한 역할분담에 대한 점검은 원활한 협력과 의사소통을 촉진하며, 과제 진행 중 발생할 수 있는 혼선을 최소화한다.

검토 필요사항

Key Question : 과제 추진을 위한 역할분담·지원체계 방안이 마련되어 있는가?

과제 역할분담

- 과제 관리자, 데이터 전문가, 도메인 전문가 등 과제참여자의 역할과 책임, 참여자별 성과 측정방법 등 정의
- 참여자 간의 원활한 의사소통 및 공유를 위한 정기회의 계획 마련
- 데이터 활용에 대한 기관장과 관리자의 관심과 추진 의지 확인
- 추진체계에 현업 담당자 포함

과제 지원체계

- 자문단, 협의체, 이해관계자 등의 적절한 참여
- 데이터 전문가가 참여하여 필요시 지원을 받을 수 있는 체제 구축

사전·사후 컨설팅

- 필요시, 과제 전반의 이해 및 과제의 추진 가능성 검토 등을 위한 예산 총괄부서의 지원제도(컨설팅 예산 배정 등)를 적절히 활용

1.2 데이터 분석과제 발굴

데이터 분석과제 발굴 단계에서는 공공기관이 직면한 문제를 해결하거나 개선할 수 있는 구체적이고 실현 가능한 분석과제를 정의한다. 이를 위해 우선 분석 후보가 되는 주제들을 도출해야 한다. 분석 주제 후보 도출 절차와 방법은 다음과 같다.

1.2.1 분석 주제 후보 도출

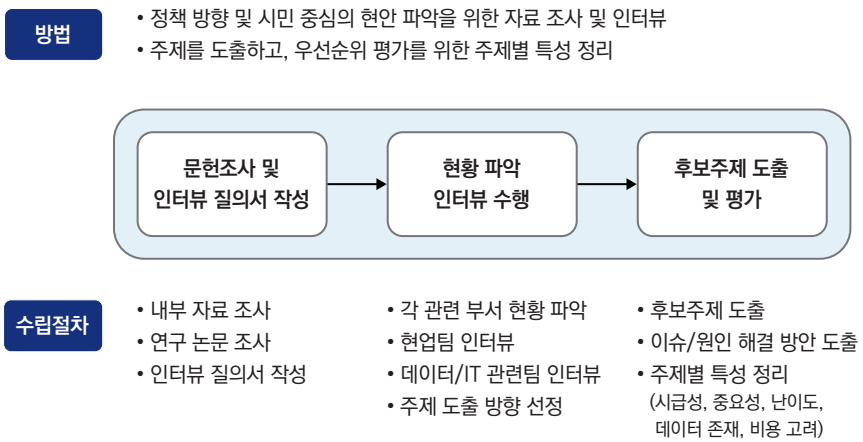
추진체계가 구성되고 IT 담당자, 현업 담당자, 데이터 전문가가 함께 참여하여 문제를 정의하고 분석 주제 후보를 도출하는 과정은 프로젝트의 방향성과 목표를 정의하는 근본적인 단계라 할 수 있다.

데이터 분석 주제를 도출하는 방법은 내부의 당면한 문제와 현안을 파악하여 해결하는 방안을 찾는 방법과 기관 외부로부터 아이디어를 수집해서 주제를 도출하는 방법이 있다.

- 현업 담당자는 조직이 당면한 문제와 해결 방안, 국민 만족도 제고 방안을 가장 잘 파악하고, 이를 위해 내·외부 자료 조사, 연구 논문 조사, 내·외부 관련 담당자나 전문가 인터뷰를 진행 후 정리하여 주제 후보를 도출한다.
- 분석 주제는 때로는 아이디어 공모전이나 외부의 분석 우수사례를 통해서 인사이트를 얻기도 한다. 특히 공모전은 전 국민의 관심을 유도하고 기관의 정책을 홍보할 수 있는 이벤트로서의 효과도 있다.

분석 주제 선정 절차는 문헌 조사 및 인터뷰 질의서 작성, 현황 파악 인터뷰 수행, 주제 후보 도출 및 평가 순으로 진행된다.

[그림 5] 분석 주제 후보 도출 절차 예시



분석 주제는 다음과 같은 기준에 적합한 것들을 도출하며, 도출된 후보 주제는 추진 가능성을 최종 검토하여 분석 주제를 확정한다.

[표 3] 주제 도출 시 고려사항

주제 도출 기준	고려 사항
분석 필요성	• **분석주제가 기관의 운영방향 및 목표에 부합하는가?** • **분석 주제는 어떤 효과를 기대할 수 있는가?** ex. 대국민 서비스 개선, 예산 절감, 정책 수립 또는 개선 등
분석 가능성	• **분석에 필요한 데이터가 있는가?** ex. • 내부에 존재하는 데이터를 활용하는가? • 필요한 외부 데이터는 확보 가능한가? • 분석 대상 데이터의 품질은 좋은가? • **분석에 필요한 인프라가 있는가?** • **분석이 기술적으로 가능한가?** ex. 외부 데이터를 내부 데이터와 **매시업***할 수 있는가? ***매시업(Mashup):** 데이터를 서로 융합하여 새로운 서비스를 만들어 내는 것 • **분석에 필요한 예산 및 인력을 확보할 수 있는가?**

검토 필요사항

Key Question : 데이터 분석 주제는 어떻게 도출하는가?

데이터 분석 주제를 도출하기 위해, 조직 내부에서 해결해야 하는 문제를 데이터 관점에서 정리하는 것이 필요

- **부서에서 현재의 문제점을 적는다.**
  현재 운영이나 정책에서 발생하는 문제를 리스트업한다.
- **문제를 데이터 주제로 변환한다.**
  문제를 데이터로 표현할 수 있도록 정리한다.
  ex. **"교통 체증이 심하다" ▸ "출퇴근 시간대 차량 속도 변화 분석"**
- **문제 해결을 위한 목표를 수치화한다.**
  목표를 수치로 정의할 수 있는지 확인한다.
  ex. **"사고 발생률을 10% 줄이는 것"**
- **필요한 데이터를 파악한다.**
  해당 문제를 해결하기 위해 어떤 데이터가 필요한지 정리한다.
  ex. 교통량, 차량 속도, 사고 데이터
- **해결 가능한 문제인지 검토한다.**
  데이터가 충분한지, 분석이 현실적으로 가능한지 평가한다.

데이터 분석과제 발굴 시 실제 고려하는 점검항목

공공기관 데이터분석 담당자들의 FGI 요약, 2024. 12

데이터 활용 환경과 지원 관점

- 분석에 걸리는 기간
- 분석 결과 및 대시보드 구축 시 지속적 운영, 검토, 지원 가능 여부
- 데이터 활용 담당 부서의 의지, 역점 과제 여부
- 부서 간 협업의 용이성(데이터 제공 및 분석 요구사항 정의, 인터뷰 협조)
- 분석 데이터 활용 수집을 위한 정보제공 사전 협의

분석 결과의 기대효과 및 효과성 관점

- 분석의 효익(인력 절감, 비용 감소, 시간 단축, 대민 서비스 질 향상 등)
- 업무 개선 기대효과

조직 및 정책적 연계 관점

- 분석과제 보유 부서와의 협조 용이성, 협력 가능성
- 기관 주요 사업과의 연계성
- 정책 적용 가능 여부 및 활용 부서의 적극성
- 기관장의 관심 분야, 비예산으로 가능한 분석 범위
- 분석 업무에 대한 조직내 인센티브 부여

분석과제의 중요도 및 시의성 관점

- 시급성
- 이슈 발생에 따른 해결을 위한 시의성
- 분석 결과의 정책적 활용 가능성

기술 및 자원 관점

- 구현 가능성(기술, 전산 자원, 인력, 예산 등)
- 부서 활용 가능성 필요 여부

1.2.2 분석 주제 중복성 검토

범정부 데이터분석시스템에는 각 정부 부처, 공공기관에서 수행한 분석과제에 대한 정보들이 정리되어 있다. 추진하고자 하는 분석 주제가 유관 기관에서 이미 실행한 것인지에 대한 중복성 여부를 사전 검토하여 불필요한 자원 투자를 최소화해야 한다.
유사하거나 중복인 경우 범정부 데이터분석시스템에 공유된 분석사례를 참조한다.

- 범정부 데이터분석시스템(http://www.insight.go.kr)에 접속하여 인증서 또는 모바일 공무원증으로 로그인하여 활용할 수 있다.

1.2.3 기존 분석모델 활용 검토

주제 중복성 검토 결과 유사하거나 중복된 데이터 분석 서비스가 이미 존재하는 경우, 담당자는 기존 분석모델 활용을 우선 고려한다.

- 분석 주제가 유사할 경우 사용하는 데이터도 유사하게 사용될 가능성이 크기에, 비슷한 기 분석모델을 활용함으로써 효율성을 높이고 시행착오를 줄일 수 있다.

(예시) 행안부는 기 추진 데이터 분석모델 중 기관의 수요가 높고 공통적으로 적용이 필요한 모델을 표준분석모델로 개발하여 범정부 데이터분석시스템을 통해 제공하고 있음.

[표 4] 2024년 행안부 추진 표준분석모델 예시

모델명	활용 영역	분석모델 상세 내용
주거환경	청년 1인 가구를 위한 주택 입지 선정	• 청년 1인 가구주들은 새로운 주거지를 찾을 때 통근조건, 교통시설, 그리고 주거시설 등을 고려하여 입지를 선정함 • 서울시는 이를 고려하여 양질의 주거환경을 개선하는 청년 주택 정책 시행 • 분석 절차는 1. 전체가구 수 대비 청년 1인 가구 수에 CCTV 구축 밀도 와 노후 건축물 비율을 고려하여 대상지 선정 2. 용도지역, 경찰관서와의 거리, 지하철역 및 대학교와의 거리를 반영하여 가중치를 계산한 입지 적합도 분석 3. 선정된 대상지와 입지 적합도 분석을 종합 고려하여청년 1인 가구를 위한 주택 입지 선정

모델명	활용 영역	분석모델 상세 내용
모빌리티	교통카드 데이터 기반 퍼스널 모빌리티 노선 선정	• 포스트 코로나 시대에 1.3인의 단거리 주행에 알맞은 **'퍼스널 모빌리티'**의 이용량이 꾸준히 늘어날 것을 예상하여 인프라 구축이 필요하다고 판단 • 또한, 전기에너지를 사용하기 때문에 친환경 교통수단으로서 가치가 기대됨 • 분석 절차는 1. 도로 폭 25m 이상, 지하철 역사반경 500m 이내, 대학교 반경 400m 이내, 경사도 15도 이내로 추출한 도시 환경 기반 분석 2. 출발지 좌표, 승차 정류장, 출근 시간대(5~10시)를 가공하여 300m 격자화하여 알뜰 교통카드 데이터 기반 분석 3. 기 분석한 도시 환경 기반 분석과 알뜰 교통카드 데이터 기반 분석을 종합 고려하여 퍼스널 모빌리티 전용도로 선정
교통	효율적인 대중교통(버스) 정책 수립	• 교통카드 이용 현황, 유동인구 및 주거인구 데이터 등을 분석하여 이동수요, 저상버스 필요 노선, 환승 지점 등 파악 • OD 분석, 운행기록 분석 등을 통한 탄력배차제 적용 노선 결정, 배차간격 조정·결정 등
관광	관광 정책 수립	• 지역행사·축제 방문인구, 전년 대비 증감분석, 축제장소별·시간추이별 관람객 증감분석을 통한 관광객 수요예측 및 관광 시설 공급 판단 등 관광 정책 수립 • 상권 통계 기반 지역상권 매출 분석을 통한 최적화된 상권 관리 및 지역 경제 활성화 기여도 평가 • 관광 통계 기반 외국인 관광객 유치 활성화 방안 마련 및 외국인 전용 관광 코스 개발 활용 등
근로감독	효율적인 근로감독	• 사업장별 5대 취약 지수(최저임금, 서면계약, 임금체불, 근로시간 불이행, 약자 보호) 도출 및 이를 활용하여 반복적으로 위법 행태를 보이는 근로사업장을 선정하여 우선 감독 • 근로감독 분야와 유사한 형태의 관리 감독 업무가 필요한 분야에 대한 확대 적용 가능한 개념적 분석모델 제공
CCTV	CCTV 설치지역 분석	• 범죄, 사고 다발지역 등 보안이 취약한 안전 사각지대 도출 및 CCTV를 설치할 우선 지역 선정을 위한 분석 • CCTV 취약 지수, 범죄 취약 지수, 유동인구 취약 지수 등을 활용한 단위구역별 가중치 생성 및 설치 포인트 도출

검토 필요사항

Key Question : 데이터 분석 주제 점검 체크리스트

- 주제 도출 과정에 현업 담당자가 참여하여 적극적으로 의견을 제시하였나?
- 도출된 주제의 서비스 대상, 데이터 획득 가능성, 구현 가능성은 고려하였나?
- 해결하고자 하는 문제가 국가·지역·사회적으로 시급한 것인가?
- 수행하고자 하는 주제가 무엇인지 구체적이고 명료하게 정의할 수 있는가?
- 수행하고자 하는 주제에 대해 범정부 데이터분석시스템을 통한 중복성 검토를 하였나?

1.2.4 분석과제 선정

분석과제 선정은 데이터 분석을 통해 해결하고자 하는 문제나 주제를 분석 주제 후보 리스트에서 실현 가능성과 우선순위를 고려해서 최종 분석과제를 정하는 과정이다. 이 과정에서 고려해야 할 주요 요소는 다음과 같다.

분석과제 선정 시 고려사항

- **목표 및 필요성:** 분석이 필요한 이유와 목표를 명확히 하여 우선순위를 정한다.
- **데이터 가용성:** 필요한 데이터를 확보할 수 있는지를 평가한다.
- **문제의 중요성:** 해결하고자 하는 문제가 조직이나 정책에 얼마나 중요한지를 고려한다.
- **자원 및 시간:** 분석을 수행하는 데 필요한 자원과 시간을 고려하여 현실적인 과제를 선정한다.
- **기대효과:** 분석 결과가 가져올 수 있는 긍정적인 영향을 사용자, 기획자, 운영자 및 시스템 등 다양한 측면으로 평가한다.
- **전문성:** 해당 과제를 수행할 수 있는 주제 관련 도메인과 분석 기술 그리고 시스템 관련 인력의 전문성을 고려한다.
- **법적 및 윤리적 고려:** 분석과정에서 법적, 윤리적 문제가 발생하지 않도록 검토한다.

공공기관에서 데이터 분석과제를 기획할 때는 단순한 프로젝트 수행이 아니라 기관의 전략적 방향과 연계된 장기적인 접근이 필요하다.

- 이를 위해 분석 난이도와 복잡성을 고려하고, 내부 역량과 외부 전문성을 적절히 조합하는 방식이 요구된다.
- 공공기관에서 데이터 분석을 추진할 때, 내부적으로 수행할 부분과 외부 전문가의 지원이 필요한 부분을 명확히 구분하는 것이 필요하다.

데이터 범위가 넓거나 기관 간 연계가 필요한 경우, 내부에서 단독으로 해결하기 어려울 경우에는 외부 전문가의 지원이 필요하다.

- 특히, 초기 단계에서 데이터 분석 로드맵을 수립하고 전략적 방향을 설정하는 과정에서는 외부 전문성을 적극적으로 도입하는 것이 바람직하다.
- 다만, 외부 용역을 활용하기 위해서는 공공부문의 특성상 사전에 예산을 확보하고 사업 발주를 진행해야 하므로 장기적인 시각에서 기획하는 것이 중요하다.

반대로, 기관 내 특정 주제나 분야에 대한 높은 이해도가 필요한 분석과제의 경우 직접 수행하는 것이 효율적이다.

- 우선, 전사적 차원의 데이터 분석 필요성을 도출하고, 기관이 다루어야 할 다양한 분석 주제를 리스트업하여 연도별 추진계획을 마련한다.
- 이후, 정책적 파급력이 높은 우선 실행 과제를 선정하여 기관의 역량에 맞춰 내부적으로 수행하거나 외부 지원을 받는 방식으로 진행하는 것이 바람직하다.
- 이러한 과정에서 초기 소규모 프로젝트(PoC)를 수행한 후 점진적으로 확장하는 전략을 적용하면 데이터 간 연계성이 강화되고 분석 결과의 활용도가 높아진다.
- 초기 소규모 데이터 분석 프로젝트를 수행한 후 점진적으로 분석 주제를 확대하면서 장기적인 데이터 분석 로드맵을 구축한다.
- 이를 통해 데이터 분석이 개별 프로젝트로 단절되지 않고 지속적으로 확장될 수 있는 기반을 마련한다.

궁극적으로, 공공기관이 데이터 기반 행정을 구현하기 위해서는 **정보화전략계획(ISP)***과 유사한 개념으로 데이터 분석을 위한 전략적 접근이 필요하다.

* **정보화전략계획**(Information Strategy Planning, ISP): 기관의 경영 목표와 IT 전략을 일치시키고, 효과적인 정보 시스템을 구축하기 위한 중장기적 계획

이를 데이터 기반 전략기획(Data-Driven Strategy Planning ,DSP) 개념으로 정립하여 추진하면, 개별 프로젝트가 단절되는 것이 아니라 기관 전체 차원에서 데이터 활용의 연속성을 유지할 수 있다.

내부 역량과 외부 전문성을 조화롭게 활용하고, 데이터 기반 전략계획을 기반으로 체계적인 실행 전략을 마련하는 것이 데이터 분석의 시너지를 극대화하는 핵심 요소가 될 수 있다.

분석 추진 가능성 및 우선순위

분석과제의 추진 가능성 및 우선순위를 검토하기 위해서는 기관이 자신의 상황에 맞는 평가 항목으로 구성한 평가 기준표를 활용하여 항목별로 점수를 부여하고, 총점을 기준으로 평가하는 방법이 있다.

- 평가 항목별 점수 배점은 기관의 상황에 따라서 달리 적용한다. 예를 들어, 추진 시급성이 가장 중요한 항목이라면 높은 점수를 할당한다.
- 지금까지 조사한 내용을 바탕으로 기관 자체 평가 기준을 적용한 과제 평가기준표를 만들고 분석과제별 타당성을 평가한다.

[표 5]의 예시는 분석과제 후보를 평가할 때 사용하는 평가 기준표로, 분석과제의 추진 필요성, 파급효과, 추진 시급성, 구현 가능성, 데이터 수집 가능성, 모델 확장성을 검토하고 평가 항목별 점수를 부여하여 추진 가능성을 검토한다.

[표 5] 과제 평가 기준표 예시

평가 항목	기준 설명
필요성	공공정책 결정이나 공공서비스 측면에서 정성적, 정량적 기대효과 등 분석과제가 필요한지 판단
추진 시급성	당장 해소되어야 할 사회 현안 여부 판단: 장기과제 성격 분리
구현 가능성	과제를 구현 시 어려움이 없는지 현실성 판단 (표준분석모델을 활용하거나, 범정부 데이터분석시스템에서 구현 가능한 과제는 구현 가능성이 높을 수 있음)
데이터 수집 가능성	기관 협조나 데이터 확보, 데이터 구매비용 등 제약사항 판단 (표준분석모델을 활용하거나, 범정부 데이터분석시스템에서 구현 가능한 과제는 데이터 수집 가능성이 높을 수 있음)
모델 확장성	과제가 시범과제로 끝나지 않고 전국 모델로 확장 가능한지 판단

과제 선정에 있어서 중요한 평가요소는 과제의 **'추진 시급성'**과 데이터 분석모델의 **'구현 가능성'**으로, 경우에 따라서는 추진 시급성과 구현 가능성만을 고려하여 주제를 선정할 수도 있다. 평가 기준은 기관 특성에 따라 다를 수 있다.

참고 : 과제의 추진 시급성/구현 가능성을 고려한 주제 선정 방안

> 기관이 분석과제 선정을 위해 평가 기준표를 활용하여 과제 후보별로 평가하는 방식이 일반적이지만, [그림 6]과 같이 과제의 추진 시급성과 구현 가능성(난이도)만으로 평가하여 주제를 선정하는 방안도 있다.

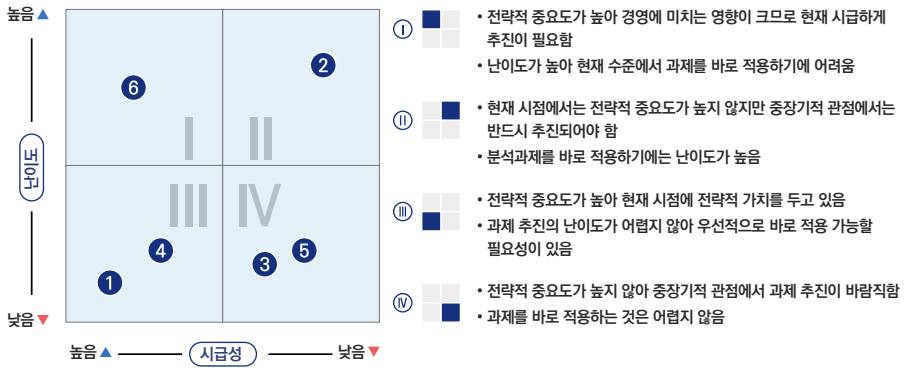
분석과제를 추진할 때 전략적 중요도에 따른 시급성을 우선 고려해야 하며, 데이터를 생성, 저장, 가공, 분석하는 비용과 과제에 대한 분석 수준을 고려한 난이도 또한 우선순위 선정에 중요한 기준이 된다.

우선순위 선정 기준을 토대로 우선 추진해야 하는 분석과제와 제한된 자원을 고려해 단기적 또는 중장기적으로 추진해야 하는 분석과제 등 4가지 유형으로 구분해 분석과제의 적용 우선순위를 결정한다.

[그림 6]의 사분면의 번호와 같이 각 과제별 시급성과 난이도를 구분하여 사분면 안에 과제를 위치하고 우선순위를 식별한다. 사분면 영역에서 가장 우선해서 분석과제 적용이 필요한 영역은 Ⅲ영역이다.

또한, 전략적 중요도가 현재 시점에는 상대적으로 낮은 편이지만 중장기적으로는 현안에 미치는 영향도가 높고, 분석과제를 바로 적용하기 어려워 우선순위가 낮은 영역은 Ⅱ영역으로 볼 수 있다.

[그림 6] 분석 주제 시급성/난이도별 순위 선정 예시



적은 비용으로 추진하여 빠른 시일 내에 효과를 거둘 수 있는 주제를 선정하는 것이 좋지만, 기관이 처한 상황에 따라 비용이 많이 투입되어도 꼭 해야만 하는 과제의 경우에는 시급성이 매우 높게 측정될 것이므로, 난이도를 기관의 상황에 따라 조율하여 추진한다.

시급성이 높고 난이도가 낮은 과제는 성공했을 때의 홍보 효과도 상당하기에, 이에 따라 기관의 빅데이터 추진도 탄력받을 가능성이 커진다.

* (예시) 서울시의 심야버스 사례의 경우, 심야버스 노선을 신규로 만들기 위해 빠른 분석이 필요했고 비교적 적은 비용으로 분석을 완료하여 우수사례로 소개되고 있으며, 이후 서울시의 데이터 예산이 크게 증액되고 본격적으로 데이터 프로젝트를 늘린 계기가 됨.

과제 선정 시에는 다양한 접근 방식을 균형 있게 추진해야 한다.

- 내부 논의가 중심이 되면 실무적 관점은 충분히 반영되지만, 정책적 방향 설정이 미흡할 수 있다.
- 외부 컨설팅이 주도하는 경우 전문성이 강화되지만, 내부 소통이 부족해질 우려가 있다.
- 또한, 기관장의 지시가 중심이 되는 경우 상향식 소통 채널을 확보하여 현장의 의견이 반영될 수 있도록 해야 한다.

분석과제 확정

과제 평가표에서 전체 총점이 가장 높은 과제를 최종 후보로 선정한다.

- 단, 전체 총점이 가장 높을지라도 구현 가능성 점수가 현저히 낮은 경우는 실제 추진 시 실패할 가능성이 높다.
- 전체 총점은 다소 낮을지라도 추진 시급성에서 가장 높은 점수를 받은 과제는 제약 요소를 해소할 수 있는 방안을 찾고 과제의 범위를 보완하여 추진할 수 있다.

검토 필요사항

Key Question : 분석과제 선정 점검 체크리스트

- ✔ 과제 후보를 평가할 수 있는 평가 기준표를 만들고 객관적으로 평가하였나?
- ✔ 평가 기준표에 적절한 가중치를 적용하여 타당한 결과를 도출하였나?
- ✔ 확정된 분석과제는 추진 시급성이 높고 구현 가능한 주제인가?
- ✔ 난이도가 높은 주제를 선정하였다면, 기간 내에 분석 결과를 도출할 수 있는 방안을 가지고 있는가?

1.3 분석 추진계획 수립

사전 준비 단계에서 분석 추진계획은 데이터를 수집, 분석, 활용하기 위한 구체적인 전략과 목표를 제시한다. 주로 데이터 확보방안, 분석 비용 및 기간 산정을 계획한다. 먼저 분석 추진계획 시 주요 고려사항을 점검하여 분석 추진계획을 수립한다.

분석 추진계획의 주요 고려사항

분석 추진계획을 수립할 때는 목표와 기대효과를 설정하고 이에 필요한 기간 범위, 지속가능성, 주제 세분화, 난이도 및 프로젝트 규모 등을 체계적으로 검토하는 과정이 필요하다.

데이터 분석의 목표와 기대효과를 명확히 설정한다.

- 분석의 주요 목표로서 수혜 대상, 목표서비스, 성과 측정지표 등을 구체적으로 정의하여 분석과정에서 발생할 수 있는 혼란을 줄이고 일관된 방향으로 프로젝트를 진행할 수 있어야 한다.
- 데이터 분석을 통해 얻을 수 있는 비용 절감, 효율성 향상, 정책 개선 등의 기대효과를 구체적으로 명시한다.

과제의 단계적 실행계획을 수립한다.

- 단계적 실행계획 및 목표설정을 통해 실효성을 극대화하고 점진적 개선을 통해 목표 완성도에 도달할 수 있도록 한다.
- 예를 들어, 교통량 예측 모델을 구축할 때, 초기에는 단순 통계 분석을 수행하고, 이후 머신러닝 모델을 적용한 후 최종적으로 실시간 데이터를 활용한 예측 시스템으로 발전시키는 단계별 접근이 효과적이다.

단계별 목표 구체화

- 큰 주제를 그대로 분석하기보다 작은 단위로 나누어 단계적으로 접근하는 것이 데이터 분석 간의 시너지 효과가 크다.
- 예를 들어, 교통 혼잡도 분석을 진행할 경우, 전체 도시 교통 흐름을 분석하기보다 특정 시간대(출퇴근 시간) 또는 특정 지역(혼잡 구간)으로 세분화하여 분석하는 것이 더 실용적일 수 있다.

- 목표서비스 구현을 위해 도입·적용 가능한 AI 기술은 언어모델, 분석모델, 자동 시각화 기술, 생성형 AI 기술 등 여러 가지가 있으며, 현재도 기술이 계속 진화되고 있다. 특히 1~3년 전에 예산작업을 통해 계획된 AI 사업의 경우 신기술의 등장 및 기술의 업데이트가 필요할 수 있다. 적용하고자 하는 AI 기술에 대한 선정사유, 유사·중복 서비스 사례, 전문가 의견 등에 대한 검토는 목표서비스를 구현하고자 할 때 도움이 된다.

분석과제의 시간적 단기, 중기, 장기 범위를 결정하는 것이 필요하다.

- 단기적인 분석은 당면한 과제를 빠르게 해결하기 위한 과제 중심적인 접근방식이다.
  ex. 버스 노선 최적화를 위해 특정 구간의 승·하차 데이터를 분석 등
- 중기 분석은 주로 PoC(Proof of Concept) 프로젝트로 실효성을 검증하는 단계에 해당한다.
  ex. 복지 사기 탐지 모델을 시범 운영하는 방식 등
- 장기 분석(2년 이상)은 정책 및 행정 혁신을 위한 전사적, 장기적 관점의 과제로 체계적이며 지속적인 분석이 요구된다.
  ex. 의료 빅데이터 플랫폼 구축과 같은 대규모 프로젝트 등

분석의 지속가능성을 검토해야 한다.

- 분석이 일회성보다는 주기적으로 활용될 수 있는지, 필요한 데이터를 지속적으로 확보할 수 있는지 검토해야 한다.
- 예를 들어, 재난 예측 모델을 개발할 경우, 기상 데이터와 재난 발생 이력이 정기적으로 업데이트되는지 확인하는 것이 중요하다.

분석 난이도와 프로젝트 규모를 고려해야 한다.

- 단순 통계 분석인지, 머신러닝 적용이 필요한지, 대규모 AI 기반 분석인지에 따라 실행 전략이 달라진다.
- 예를 들어, 단순 민원 유형 분석은 통계 분석으로 충분하지만, 공공안전을 위한 범죄 예측 시스템 구축은 고도화된 AI 모델과 데이터 통합 플랫폼이 필요할 수 있다.

검토 필요사항

Key Question : 성공적인 데이터 분석 계획 수립을 위한 고려사항은?

최종목표를 설정한 후, 단계적 실행 계획 수립

- ex. 교통량 예측 모델을 구축할 때, 초기에는 단순 통계 분석을 수행하고, 이후 머신러닝 모델을 적용한 후, 최종적으로 실시간 데이터를 활용한 예측 시스템으로 발전

필요시, 분석 주제를 세분화

- Divide&Conquer(작은 단위로 나누어 단계적으로 접근)
- 과제를 하위 과제로 세분화하고 이를 각각 독립적으로 분석할 수 있는지 검토

  ex. 교통 혼잡도를 분석하기 위하여 도시 전체 교통 흐름 분석하는 방안(큰 단위) vs 혼잡 구간 / 특정 시간대 분석(작은 단위)

분석의 시간적 범위 결정

- ex. 단기 분석(개별 과제 중심), 중기 분석(PoC / 실효성 검증), 장기 분석(2년 이상, 정책 및 행정 혁신을 위한 체계적인 분석)

분석의 지속가능성 평가

- 일회성이 아니라 주기적으로 활용 가능한지, 필요 데이터를 지속적으로 확보할 수 있는지 확인

  ex. 재난예측 모델 개발한 경우, 기상 데이터와 재난발생 이력이 정기적으로 업데이트할 수 있는지

분석 난이도와 프로젝트 규모 결정

- 단순 통계 분석인지, 머신러닝 적용이 필요한지, 대규모 AI 기반 분석인지 등에 따라 프로젝트 규모 산정

  ex. 단순 민원 유형 분석은 전통적 통계 분석 vs 공공안전 범죄예측은 고도화된 AI 모델 필요

검토 필요사항

Key Question : 구현하고자 하는 목표서비스에 대해 명확히 설명할 수 있는가?

[목표 설정]

수혜 대상

- 분석 서비스 수혜 대상에 따라 예산/기술 난이도 큰 편차 발생, 수혜 대상을 명확히 정의하여 AI용 인프라 구축 비용, 서비스 사용 비용, 보안사항, 분석 기술 난이도 등 경제적, 기술적 타당성 검토

목표서비스

- 최종 서비스 채널(웹/앱/내부 시스템) 및 형태(챗봇/시각화) 결정
- PoC(개념 정의), 분석모델 개발, 적용 시스템 구축 등 사업의 목표 범위를 결정하여 실현가능한 장기적인 로드맵 검토
- 데이터 분석은 유사영역의 확장가능성이 높아 서비스 목표 설정시 확장 가능한 영역을 함께 고려

성과측정지표

- **분석 기술:** 자체 보유 성과지표 솔루션, BLEU(Bilingual Evaluation Understudy) 지표, 분석모델 정확도·F1-score 등
- **AI 서비스:** 질의/응답건수, 서비스 만족도 등

검토 필요사항

Key Question : 적용하고자 하는 분석 기술이 목표서비스 구현을 위한 최선의 방법인가?

선정사유

- 분석 기술을 도입·적용할 분야 및 기술을 특정하고 그 **분야***에 해당 **기술****을 선정한 사유 제시

  ***분야:** 사전학습 / 검증·평가 / 분석 등

  ****기술:** 생성형AI(LLM, LMM) / 분석모델 / 자연어처리(NLP, BERT) / 검색증강생성(RAG) 등

유사·중복 서비스 사례

- 범정부·기관 유사·중복 서비스 검토
- 유사사례 및 기술간 장단점 분석 등을 통한 본 기술이 목표서비스를 구현하기 위한 최선의 방법인지 검토

전문가 의견

- 기술 및 서비스, 현업에 대한 이해도 있는 전문가의 의견 또는 자문을 통해 필요성 검토

1.3.1 데이터 확보방안

분석 목적의 달성 가능성을 판단하기 위해서는 내부 보유데이터의 현황을 확인하고, 데이터 간의 연계 가능성을 검토하며, 외부 데이터와의 연결 방안을 마련하는 것이 중요하다. 이는 주제 선정 단계와 데이터 수집 단계에서 모두 고려해야 할 핵심 요소이다.

[표 6] 국내 주요 데이터 현황

분야	기관명	주요 데이터	활용 사례
보건의료	심사평가원, 건강보험공단	수진자 인적 사항, 진료과목, 병명, 급여비용, 투약 정보 등	환자별 맞춤형 진단·치료 서비스 제공 및 정밀의료 솔루션 개발 등
	국립암센터	환자 일반정보, 유형별 암 영상정보, 진료/처방 정보 등	
교통	교통연구원, 교통안전공단	도로 현황, 도로시설물 관리, 교통량, 사고지점, 피해 상황 등	교통 혼잡 완화를 위한 분석 서비스 및 교통사고 원인 분석 서비스 제공 등
	한국스마트카드	시간별, 지역별 이동수단, 승/하차 인원 등	
금융	한국은행, 신용정보원	경제·금융 통계, 개인·법인 대출, 세금체납, 채무불이행, 부도 등	맞춤형 금융 서비스 개발 및 보험사기 분석, 연체자 예측 모델 개발 등
	은행, 보험, 신용카드	계좌정보, 대출, 상품거래, 인터넷 뱅킹 이용내역, 민원, 가맹점 등	
통신 · 미디어	통신사, IPTV 업체	가입자/위치정보, 유동인구, 서비스별 트래픽/구매 내역 등	감염병 차단 서비스, 상권 분석, 콘텐츠 추천 서비스, 광고 전략 개발 등
	언론진흥재단, 방송광고진흥공사	종사자 정보, 광고시장 현황, 구독/시청패턴, 디지털 콘텐츠, 매출 등	
도시 · 공간	지자체 (광역/기초)	도로/가로등 위치, 전력/가스 공급 체계, CCTV, 주차/횡단보도 정보 등	지능형 도시 서비스, CCTV 기반 보안 서비스, 입지 분석, 부동산 가격 예측 등
	국토연구원, 한국토지주택공사	지형, 산업입지, 택지, 부동산 거래, 3차원 공간정보 등	
에너지 · 환경	한전, 지역가스, 한국에너지공단	전력 판매, 전력 시설물, 전력 계량, 과금, 태양광 정보, 소비 패턴 등	에너지 공급 제어·관리서비스, 전력 소비 패턴 분석, 자연재해 예측 등
	기상청, 수자원공사	기상정보, 도로 날씨, 기후변화, 자연재해, 가뭄/지하수 정보 등	
연구	과학기술정보 연구원	바이오, 소재 등 기초과학 데이터 및 대형 연구장비, 연구노트 등	신약 후보 물질 발굴, 신소재 연구 등
	참조표준 데이터센터	건강지수, 수질, 인체치수 등	

문화 · 관광	문화정보원	예술작품, 문화재, 역사 자료, 문화산업, 도서, 체육 정보 등	문화재 위험관리 분석, AI 기반관광안내 서비스 등
	한국관광공사	관광지, 다국어 관광 정보, 숙박, 음식점, 축제 정보 등	
제조	삼성전자, LG전자, 현대자동차, 기아자동차 등	자재관리·구매/생산/설비관리/품질 관리 정보 등	제조공정 자동화, 재고관리, 신상품 개발, 자원관리 등
유통	백화점, 할인마트, 홈쇼핑, 외식업체 등	고객현황, 구매이력, 배송이력, 식자재 내역 등	상품 추천 등 고객마케팅, 반품률 예측 등

필요 데이터의 정의

관련 기관, 지자체, 실무자 및 외부 전문가 등 이해관계자들과 업무 해결을 위한 인터뷰 등을 통하여 데이터 분석 목적에 적합한 데이터 목록을 작성하고 데이터별로 확보 가능 여부를 점검한다.

기관 내 데이터 확보방안 검토

필요 데이터에 대해 데이터명, 데이터 설명, 데이터 형태, 용량 등 기관 보유데이터의 현황을 조사한다.

- 기관의 보유데이터 중 분석을 위해 필요한 데이터를 도출한다.
- 내부 데이터라도 데이터가 잘 정제되어 품질이 좋은지, 데이터가 지속적으로 업데이트 되고 있는지, 분석에 필요한 기간만큼의 데이터가 충분히 적재되어 있는지, 코드 데이터의 과거 변경 이력이 있는지 등도 파악한다.

분석 대상 데이터의 관리 권한이 주관부서 이외의 타 부서에 속하는 경우, 관련 부서 간 협의를 통하여 데이터의 공유가 가능한지 확인한다.

내부 데이터이지만, 관련 법, 보안, 개인정보보호 문제 등으로 제약사항은 없는지 확인이 필요하며, 개인정보의 경우 비식별화를 통해 사용하는 방안도 고려한다.

기관 외 데이터 수집 가능성 검토

데이터 보유기관, 데이터명, 데이터 설명, 데이터 형태, 용량 및 데이터 제공형태 등 현황을 검토한다.

- 분석 대상 데이터 수집에 법률상 제약사항은 없는지 확인한다.
  * (예시) 국세기본법 제81조의13에 명시된 사유 외에는 국세청의 과세정보 활용이 곤란함.

법률상 제약이 없는 데이터의 경우 보유기관과의 협의를 통하여 데이터의 공유가 가능한지 확인한다.

공공데이터 수집 가능성 검토

공공데이터포털(http://www.data.go.kr)에서는 행정·공공기관이 보유한 데이터를 개방하고 있다. 공공데이터포털에서 개방된 데이터셋 명칭과 설명, 보유기관, 원천정보시스템명, 데이터 유형 등의 내용 조회가 가능하다.

[표 7] 공공데이터포털 보유데이터 건수(2025. 2. 기준)

순서	분류기준		공공데이터포털 데이터 건수
	대분류	소분류	
1	공공행정	일반행정	9,521
		국가통계	1,246
		국가행정/재정지원	866
		법제	429
		정부자원관리	404
		국정홍보	170
		국정운영	113
		공정거래	101
		국민권익/인권	61
		정부조달	49
		소계	**12,960**
2	문화관광	관광	3,181
		문화예술	2,433
		문화체육관광 일반	2,065
		문화재	1,954
		체육	839
		소계	**10,472**
3	산업고용	산업/중소기업일반	3,396
		에너지및자원개발	2,922
		고용노동	986
		산업기술 지원	309
		산업진흥 고도화	298
		원자력 기술	104
		통상	92
		소계	**8,107**

4	교통물류	도로	2,954
		철로	1,719
		물류 등 기타	1,698
		해운/항만	447
		항공/공항	356
		소계	**7,174**
5	환경기상	환경 일반	1,974
		상하수도/수질	1,257
		폐기물	1,247
		자연	996
		대기	945
		해양환경	336
		소계	**6,755**
6	국토관리	지역 및 도시	4,709
		주택	855
		수자원	367
		산업기술지원	173
		소계	**6,104**
7	농축산	농업/농촌	2,647
		임업/산촌	1,924
		해양수산/어촌	1,256
		소계	**5,827**
8	재정금융	재정/금융	1,651
		세제	1,540
		산업금융	1,009
		무역 및 투자유치	415
		금융	360
		기획재정	181
		소계	**5,156**
9	사회복지	사회복지 일반	1,758
		보육/가정 및 여성	1,036
		노인/청소년	966
		취약계층 지원	483
		공적연금	392
		기초생활보장	117
		소계	**4,752**

10	보건의료	보건의료	3,907
		건강보험	542
		소계	**4,449**
11	재난안전	안전관리	3,368
		경찰	648
		해경	407
		소계	**4,423**
12	교육	교육 일반	2,059
		평생/직업교육	1,026
		유아 및 초/중등교육	650
		고등교육	501
		소계	**4,236**
13	식품건강	식품 의약 안전	2,392
14	과학기술	과학기술 진흥	876
		과학기술 연구	878
		방송통신	526
		우정	84
		소계	**4,756**
15	통일외교안보	외교	374
		보훈	350
		병무 행정	262
		통일	139
		방위력개선	85
		병력 운영	68
		전력 유지	31
		소계	**1,309**
16	법률	법무및검찰	483
합계			**89,355**

공동활용데이터 수집 가능성 검토

공동활용데이터는 상업적 활용이 가능하도록 민간에 개방하는 공공데이터와 달리 행정·공공기관에서 행정업무를 위해 기관 간 공유하는 데이터이므로, 공공데이터보다 더 많은 정보를 기대할 수 있다.

공동활용데이터 등록관리시스템(sharedata.insight.go.kr)에서는 행정·공공기관이 보유한 데이터를 등록하고, 이를 기관 간 공동활용 할 수 있도록 제공하고 있다.

이 외에도, 각 공공기관이 자체 홈페이지 등을 통해 데이터를 독립적으로 지원하는 공공분야 오픈 API를 활용할 수 있다.

[표 8] 국내 공공분야 주요 오픈 API

제공 포털	분류 체계	주요 오픈 API(활용 순)
서울 열린데이터 광장 data.seoul.go.kr	일반행정, 문화관광, 보건, 환경, 산업경제, 도시관리, 교통, 복지, 안전, 교육	1. 공공와이파이 위치정보
		2. 시장마트 현황
		3. 공중화장실 위치정보
		4. 개인서비스 요금정보
		5. 지하철 호선별 첫차와 막차정보
공간정보 오픈플랫폼 www.vworld.kr	행정주제도, 3차원 영상, 기타 메타데이터	1. 2D 지도
		2. 3D 지도
		3. 배경지도
		4. WMS/WFS
		5. 데이터
		6. 지도검색
		7. Static Map
		8. 3D 모바일
보건의료데이터개방 시스템 opendata.hira.or.kr	요양기관, 의료인력, 의약품, 진료비정보, 진료정보, 질병정보, 치료재료, 환자표본, 기타	1. 병원정보 서비스
		2. 질병정보 서비스
		3. 병원코드정보 서비스
		4. 병원약국찾기 정보
		5. 약국코드정보 서비스

기상자료개방 포털 data.kma.go.kr/cmmn/main.do	지상, 고층, 해양, 황사, 레이더, 위성, 수치예보, 역사기후, 간행물, 기상자원지도, 지점정보	1. 동네예보정보 조회서비스
		2. 예보구역정보 조회
		3. 생활기상지수 조회
		4. 중기예보정보 조회서비스
		5. 보건기상지수
		6. 항공기상정보
재난안전데이터 포털 data.mpss.go.kr	안전정책, 생활안전, 비상대비, 재난예방, 재난대응·재난복구, 소방정책, 119구조구급, 해양경비안전, 해양오염방제	1. 교통사고정보
		2. 낙뢰관측정보
		3. 재난상황 전파정보
		4. 자연재해상황 및 복구비 지원내역
		5. 습득물정보 조회서비스
농림축산식품 공공데이터 포털 data.mafra.go.kr	농산, 산림, 농지용수, 교육, 말산업, 검역, 식품안전, 축산, 유통, 시설장비, 농촌복지	1. 전국 도매시장 일별 경락가격
		2. 전국 도매시장 일별 실시간
		3. 경락가격
	농축산기술, 식품산업, 품질관리	1. 낙농체험 목장정보
		2. 농촌마을별 자원정보
		3. 농업마이스터대학 현황
교통정보공개서비스 openapi.its.go.kr	교통소통정보, 공사정보, CCTV정보, 돌발상황정보	1. 교통소통정보
		2. 공사정보
		3. 사고정보
		4. CCTV정보
		5. VMS정보
KOSIS 공유서비스 kosis.kr/openapi	국내통계(주제별), 국내통계(기관별), 광복이전통계, 작성중지통계, 지역통계(주제별), 지역통계(기관별), e-지방투표, 대상별통계, 이슈별통계, 인기통계, 영문 KOSIS, 국제통계	1. 추계인구
		2. 경제활동참가율
		3. 경제성장률
		4. 전산업 생산지수
		5. 기대수명
		6. 사교육 참여율

민간 데이터 구매 검토 및 비용 산정

공공데이터 및 공동활용데이터만으로 분석 목적 달성이 어려울 경우, 민간 데이터 활용을 적극적으로 고려할 수 있다.

공공부문에서 주로 활용되는 민간 데이터는 다음과 같다.

- **유동인구 데이터(통신사):** 특정 지역·시간대 인구 밀도 및 이동 패턴 분석
- **카드매출 데이터(카드사):** 상권 분석, 지역 경제 활성화 정책 효과 측정
- **모바일/내비게이션 데이터(IT/모빌리티 기업):** 교통량 분석, 관광객 이동 경로 파악

데이터 구매 비용 산정

민간 데이터 구매 비용은 정가(定價)가 없고 활용 범위와 조건에 따라 비용이 천차만별이므로, 사전에 정확한 비용을 예측하기는 어렵다. 실제 비용은 일반적인 소프트웨어 도입 시 적용되는 할인율 정책과 유사하게, 대부분 발주기관의 정해진 예산에 맞춰 협의가 이루어지는 구조이다. 따라서 객관적인 **'정가'** 기준을 일괄 제시하는 것은 현실적으로 어렵고, 사업별로 유연하게 조정되는 관행이 일반적임을 인지해야 한다.

그럼에도 불구하고, 신규 사업으로 예산 기획부터 시작해야 할 경우, 아래 절차에 따라 데이터 구매 비용의 객관적인 근거를 마련하여 예산 확보를 할 수 있다.

첫째, 구매할 데이터의 활용 범위(종류, 기간, 지역, 사용 목적 등)를 명확히 정의한다.

둘째, 정의된 요구사항을 기반으로 데이터 공급 기업에 구매 견적을 요청한다.(가능하다면 2~3곳 이상에 요청하여 비교 견적을 확보한다.)

셋째, 확보한 비교 견적서와 유사 사업의 실거래(RFP, 계약서 등)를 근거로 합리적인 예산 계획을 수립하고, 이를 통해 예산을 확보한다.

***민간 데이터 구매시, 한국데이터산업진흥원 발간 데이터 거래 및 가격책정 절차안내서 참고 ('20.05.)**

검토 필요사항

데이터 확보방안 검토 점검사항

데이터 분석을 위한 다양한 데이터 원천을 파악하고 있는가?

- 데이터가 분석 목표 달성을 위해 적절한지 검토
- 신뢰할 만한 품질의 데이터가 존재하며, 적당한 시기에 입수 가능한지 검토
- 내·외부 협력체계와 같은 구체적인 데이터 확보방안 및 절차를 갖추고 있는지 확인

분석 대상 외부 데이터 확보를 위한 기관 간 협의가 되어 있는가?

- 외부 데이터(공공데이터, 민간데이터, 협력 기관 데이터 등)의 활용을 위해 데이터 접근, 활용 범위, 법적·기술적 제약사항을 검토하여 협의해야 함
- 분석 대상 외부 데이터 목록을 정리하고 확보 필요성 검토: 분석을 위해 반드시 필요한 데이터인지, 대체 가능한 내부 데이터가 있는지 확인
- 데이터 제공 기관 및 담당 부서(담당자)를 확인
- 데이터 제공 방식 및 범위 협의(ex. 실시간 제공/배치 제공 등), 데이터 제공 형식(API, CSV, JSON 등), 데이터 활용 범위(연구목적, 상업적 활용 가능 여부 등)
- 법적·윤리적 제약사항 검토 및 계약 여부 확인: 데이터가 개인정보를 포함하는지, 데이터 공유 계약 필요 여부, 데이터 제공기관의 정책 및 가이드라인 준수 여부 등
- **데이터 제공 일정 및 주기 확인:** 최초데이터 제공과 정기적 업데이트 일정

외부 데이터는 내부 데이터와 통합하여 분석 가능한가?

- 외부 데이터와 내부 데이터 간 통합을 위해서는 데이터 매칭(Join)이 가능한지, 데이터 형식(정형/비정형, 파일 포맷 등)이 호환되는지, 데이터의 정합성(일관성, 중복성, 결측치 등)이 확보되어 품질이 보장되는 데이터인지, 데이터 갱신 주기(실시간, 배치 등)가 일치하는지, 보안 및 법적 요건을 충족하여 데이터 활용이 가능한지 검토해야 함
- 이를 위하여 내부 데이터와 외부 데이터의 속성을 비교하여 데이터 구조 분석(데이터 유형, 파일 형식, 컬럼 비교 등) 수행 필요
- 공통 키(Key) 또는 매칭 기준을 정의하여 데이터 연결 가능성 확인
  ex. 민원인 ID, 사업자 등록번호 등 고유 식별자(Unique Identifier)가 일치하는지 확인
  공통 키가 없을 경우, 대체 키(Secondary Key) 방안 고려
- 데이터 통합 후, 데이터 품질을 보장하기 위해 데이터 정합성(Completeness, Consistency, Accuracy) 분석(결측치, 중복 데이터 등)
- 데이터 처리 방식(API 연동, 배치 적재 등)을 검토하여 외부 데이터를 어떻게 반영할지 검토
- **데이터 갱신 주기 및 보안 규정 검토:** 외부 데이터와 내부 데이터 간 갱신주기(업데이트 빈도)를 검토

수집 예정 데이터에 개인정보가 포함되어 있는가?

- **개인정보 침해 가능성을 검토했는지 확인**
  ex. 직접 식별 정보(이름, 주민 번호) 포함 여부, 간접 식별 가능 정보 포함 여부, 데이터 조합 시 개인 식별 가능 여부, 데이터 암호화/비식별화 처리 여부, 법적 요건 준수 여부(동의서 확보 등), 수집 목적과의 적합성(필요 데이터만 포함)

1.3.2 분석 비용 및 기간 산정

데이터 분석과제에 전문성을 가진 사업자를 선정하여 정보화 사업 형태로 추진할 수 있다.

- 분석과제의 추진체계가 갖추어져 있고, 추진 예산이 확보되어 있다면 요구사항을 상세화하여 사업 발주한다.

프로젝트 일정 및 자원 계획 수립: 과제 분석에 필요한 인력, 기술, 추진 예산 등 자원계획을 수립하고 요구사항에 반영한다.

공공 데이터 분석 사업비 대가산정 가이드

비용은 한국소프트웨어산업협회에서 발표하는 SW사업 대가산정 가이드(2024년 개정판)를 참고할 수 있으며 기관의 상황에 맞게 적용할 수 있다. 대가산정의 주요 내용은 아래와 같다.

[표 9] 사업비 산정 단계별 주요 내용

단계	주요 내용
사전 준비	• 분석 대상 업무와 요구사항을 명확히 정의하고, 단계별 주요활동과 태스크 도출
단계별 투입공수 산정	• 요구사항에 근거하여 분석 대상 업무의 태스크별로 난이도, 작업자 수, 작업일수 산출
단계별 보정계수 산정	• 분석사업의 특성을 고려하여 보정요소별로 복잡도(보정계수) 산정 **보정요소:** 데이터 규모 보정, 데이터 처리 기술 보정, 분석모델 수립의 난이도 보정, 빅데이터 분석 기술 적용에 따른 복잡도 보정
단계별 수행률 설정	• **사업 단계별 수행률 설정** 예를 들어, 데이터 수집, 전처리를 사업자가 아닌 기관 내부에서 해당 단계를 100% 수행할 경우에는 사업비 산정에 포함하지 않음
보정 후 소요비용 산정	• **식별된 보정계수에 따라 소요공수 보정** 소요공수 = 보정 전 분석규모 × 보정계수 • 소요 공수에서 도출된 기술등급별 투입공수로 등급별 소요비용 산정 • 등급별 소요비용에 제경비와 기술료를 더해 소요비용 합계 산정
직접경비 산정	• 해당 분석사업에 관련된 직접경비 계산
분석 사업비 산정	• 분석 사업비 = 개발원가 + 직접경비

- 분석사업의 세부 태스크별로 난이도와 작업자 수, 작업일수를 산정하고 작업 단계별 난이도를 고려한 보정계수와 단계별 수행률 설정을 통해 전체 투입공수를 계산하여 개발원가*를 산정한다.
  *** 개발원가는 인건비, 제경비, 기술료를 포함**
- 데이터의 구매료, 특정 소프트웨어 도구 사용료, 특정 기술 도입 관련 전문가 비용, 홍보물 제작 및 배포, 컨퍼런스 개최 등의 홍보비용을 포함하는 직접경비를 계산한다.
- 개발원가와 직접경비를 더해 공공 빅데이터 분석 사업비를 산정한다.

분석모델을 지속적으로 활용하고 개선해 나가기 위해서 과제 완료 이후 유지관리를 위한 계획도 고려되어야 하며, 이에 필요한 인력과 예산도 별도로 수립해야 한다.

ITSQF 직무체계 및 정의

한국소프트웨어산업협회에서 고시하는 ITSQF 직무체계 및 정의를 참고하여 필요 인력의 분석비용 및 기간 산정 시 활용할 수 있다.

[표 10] ITSQF(IT Sectoral Qualifications Framework) 직무체계 및 정의

번호	직종	조사용 직무	ITSQF 직무	직무 정의
1	IT 컨설팅 및 기획	IT기획자	정보기술 기획	조직의 경영목표를 달성하기 위하여 IT전략을 기획하고, 거버넌스, 투자성과분석, 운영 정책, R&D, 프로세스, 아키텍처 등 분야별 전략을 수립하는 일
2		IT컨설턴트	정보기술 컨설팅	조직의 목표를 달성하는데 도움이 될 수 있도록, 객관적인 시각에서 조직 경영 환경을 이해하고 대상 업무 및 정보시스템을 분석하여 개선 방안을 지도, 자문 및 상담을 수행하는 일
			정보보호 컨설팅	주요 정보자산을 보호하기 위한 관리적, 물리적, 기술적 영역의 보안 요구사항과 사전 정의된 프로세스에 대해 객관적인 충족 여부를 검증하고 자문하는 일
3		업무분석가	업무분석	조직의 비전과 목표, 구조, 정책 등의 이해를 바탕으로 업무 요구사항을 도출하고 **분석***하여, 목적에 부합하는 대응전략을 수립하는 일 *** 분석:** 조직 내·외부의 경영 환경에 영향을 주는 고객과 경쟁기업, 산업동향, 내부 역량을 분석하는 능력
4		데이터분석가	데이터분석	데이터 이해 및 처리 기술에 대한 기본지식을 바탕으로 데이터 분석 기획, 데이터 분석, 데이터 시각화 업무를 수행하고 이를 통해 프로세스 혁신 및 마케팅 전략 결정 등의 과학적 의사결정을 지원하는 일

번호	직종	조사용 직무	ITSQF 직무	직무 정의
5	IT 프로젝트 관리	IT PM	IT프로젝트 관리	IT프로젝트 인도물의 납기 준수를 위하여 프로젝트를 기획하고, 범위, 일정, 원가, 인적자원, 품질, 위험, 의사소통, 조달, 변경, 보안, 정보시스템 성과 등을 통합 관리하는 일
		제외	IT프로젝트 사업관리	명확한 의사결정과 방향 설정이 가능토록 지표를 제공하고 사업 관리 지침 및 표준화 방안 제시, 주요 이슈, 위험, 자원, 일정/문서, 범위관리를 통하여 프로젝트 수행을 지원하는 일
6	IT 아키텍처	IT아키텍트	SW 아키텍처	SW의 기능, 성능, 보안 등의 품질을 보장하고 SW를 구성하는 요소와 관계를 분석, 설계하여 전체적인 SW 구조를 체계화하는 일
			Infrastructure 아키텍처	하드웨어, 미들웨어, 네트워크, 클라우드를 포함하는 인프라를 설계, 구성하여 모든 지원들의 적합성 및 신뢰성 있는 서비스를 제공할 수 있도록 체계화하는 일
			데이터 아키텍처	데이터를 구조적 관점에서 설계, 생성, 배치, 관리하며, 다양한 데이터 엔터티뿐만 아니라 해당 데이터를 처리하는 애플리케이션에 의해 데이터가 저장, 소비, 통합 및 관리될 수 있도록 체계화하는 일
7	SW 개발	UI/UX 기획 · 개발자	UI/UX 기획	서비스의 본질적 특성에 대한 이해를 기반으로 트렌드 분석, 사용자 이용 행태 분석 등을 통해 이해관계자 및 사용자의 요구를 발굴하고 사용성을 극대화 할 수 있는 UI/UX를 설계 및 검증하여 서비스의 목적과 용도에 맞게 최적화된 UI를 제공하는 일
			UI/UX 개발	사용자의 이용형태 및 기술환경을 분석하여, 사용자 인터페이스(UI/UX)의 기획 및 아키텍처를 구축하고, 프로토타입 검증, 설계 및 구현 과정을 통해 효과적인 UI/UX를 개발하는 일
8		UI/UX 디자이너	UI/UX 디자인	UI/UX 디자인의 매체별 트렌드, 사용자 경험 분석을 통해 디자인 전략 및 콘셉트를 도출하고 UI 디자인 요소를 다양한 기법을 활용해 시각화하여 사용자 요구를 검증하고 매체별 최적화된 디자인과 사용성을 제공하는 일
9		응용SW 개발자	응용SW 개발	컴퓨터 프로그래밍 언어로 응용소프트웨어의 **분석***, **설계****, 구현 및 테스트, 배포 등을 통해 제품의 기능을 개발하고 개선하는 일 ***분석:** 구현하고자 하는 애플리케이션의 요구고려사항을 도출, 분석, 명세화 및 요구사항 검증을 수행하는 능력 ****설계:** 요구사항 확인을 통한 상세분석 결과, SW아키텍쳐 가이드라인 및 SW 아키텍처 산출물에 의거하여 이에 따른 애플리케이션 구현을 수행하기 위해 공통 모듈 설계, 타 시스템 연동에 대하여 상세 설계하는 능력
10		시스템SW 개발자	시스템SW 개발	운영체제 환경에서 시스템 자원을 제어 및 관리하는 소프트웨어와 응용프로그램의 동작을 위한 시스템 플랫폼의 요구사항 분석 및 설계, 구현, 배포를 수행하는 일
			임베디드 SW개발	하드웨어 플랫폼에 대한 이해를 바탕으로 플랫폼별 운영체제 이식과 펌웨어, 디바이스 드라이버, 애플리케이션 등의 SW를 개발하고, 하드웨어 플랫폼 최적화를 수행하는 일

11	시스템구축및운영	정보시스템 운용자	데이터베이스 관리	데이터에 대한 요구사항으로부터 데이터베이스를 설계, 구축, 전환하고, 최적의 성능과 품질을 확보하도록 추이 분석 등을 통하여 데이터베이스를 수정, 개선, 백업하는 등의 업무를 수행하는 일
			NW 엔지니어링	네트워크 환경을 분석하고 네트워크에 대한 토폴로지, 자원 관리, 품질관리를 설계하고 구성하는 일
			IT시스템 관리	시스템 요구사항을 분석하고 클라우드와 가상화, 시스템과 네트워크 및 스토리지 자원의 HW, SW 서비스 플랫폼을 구축, 운영, 관리하여 안정적 컴퓨팅 인프라 및 정보시스템의 운용을 담당하는 일
12		IT지원 기술자	IT시스템 기술지원	정보기술 인프라에 대한 이해를 바탕으로 컴퓨터 하드웨어, 스토리지, 클라우드와 가상화, 네트워크 등 IT자원을 이용한 시스템의 구성과 장애처리를 지원하며 시스템 개선 및 정기 점검 등을 통해 안정적인 컴퓨팅 인프라 운영을 지원하는 일
13	IT마케팅	IT마케터	SW제품 기획	기업의 경영전략을 바탕으로, SW 활용 분야에 대한 기업 내/외부 환경, 요구 기술, 시장성 등을 분석하여 제품 전략을 수립하고, SW제품의 개발, 지원, 판매, 마케팅 계획을 수립, 운용하는 일
			IT서비스 기획	정보기술 환경 분석을 통해 고객과 시장의 니즈에 맞는 IT 서비스를 발굴하고, 제품 및 솔루션 융합으로 새로운 서비스를 기획하는 일
			IT기술영업	정보기술 지식을 바탕으로 고객 관리 및 영업 전략을 수립, 사업기회를 창출하고 요구사항에 적합한 솔루션 제안으로 협상, 계약, 판매 및 사후 관리 등 IT 영업을 수행하는 일
14	IT품질관리	IT품질 관리자	IT품질관리	IT품질목표를 달성하기 위하여 전사적인 품질정책 및 관리체계를 수립하고 품질향상을 위해 교육 및 관리 활동 등을 수행하며, 프로젝트 차원에서의 품질보증 활동을 수행하는 일
15		IT테스터	IT테스트	테스트를 효과적으로 수행하기 위해 필요한 기획, 진단 컨설팅, 계획, 환경구축, 실행, 결함관리, 문서화를 수행하고 관리하는 일
16		IT감리	IT감리	감리발주자 및 피감리원의 이해관계로부터 독립된 자가 정보시스템의 효율성을 향상시키고 안전성을 확보하기 위하여 제2자의 관점에서 정보시스템의 기획, 구축 및 운영 등에 관한 사항을 종합적으로 점검하고 문제점이 개선되도록 시정조치 사항을 도출하고 확인하는 일
		제외	IT감사	컴퓨터 시스템의 유효성과 효율, 신뢰성, 안전성을 확보하기 위해 독립적인 입장에서 일정한 시스템 감사 기준에 의거하여 시스템을 종합적으로 점검·평가하고, 관계자에게 조언 및 권고하는 작업을 수행하는 일

17	정보보호	정보보안 안전전문가	정보보호 관리	조직의 비전과 미션을 수행하기 위하여 정보 자산을 안정적으로 운영하는 데 필요한 보안정책을 수립하고 관련 법제도 준수, 보호관리 활동을 수행하며, 위험관리에 기반한 정보보호 대책을 도출하여 실행토록 관리하는 일
			보안사고 대응	침해사고의 피해확산 방지를 위해 위협정보를 탐지하고, 시스템 복구와 예방 전략을 수립하는 일과 업무 및 서비스에 영향을 준 증거를 확보 후 분석하여 신속하게 대응하는 일
	IT기술교육	제외	IT기술교육	IT분야의 기술교육을 체계적이고 효과적으로 수행하기 위하여 IT기술교육 방향 수립과 IT기술교육 환경조성, IT기술교육 교과개발 및 자료개발, IT기술교육 성과평가 등을 통해 성과향상을 수행하는 일

[출처] 한국소프트웨어산업협회 IT분야역량체계(ITSQF)
https://www.sw.or.kr/site/sw/02/10204020000002017070310.jsp

검토 필요사항

Key Question : 분석 비용 및 기간 산정

분석비용 및 기간 산정이 최적화될 수 있도록 프로젝트의 범위가 명확히 정의되었나?

- 프로젝트의 범위(Scope)를 명확히 설정하기 위해서는 요구사항을 체계적으로 수집, 분류, 검증해야 함
- 기능 요구사항(시스템이 제공해야 하는 기능)과 비기능 요구사항(성능, 보안, 확장성 관련 요구사항(API 응답시간 2초 이내 등)을 반영
- 요구사항 상세 정의 및 누락 여부 검토
- 수집된 요구사항을 과제 이해관계자와 공유하고 검토

예산수립 및 기간산정의 적정성 여부는 적절히 검토되었는가?

- 프로젝트 목표 및 범위를 확인하여 예산과 기간의 적절성을 검토
  ex. 분석할 데이터의 양과 복잡도 등
- 주요 비용 항목을 식별하고 예상 비용을 산정
  ex. 인건비, 인프라 비용, 소프트웨어 비용, 데이터 구매비용, 교육 및 컨설팅 비용 등
- 비용-효과분석을 통해 우선순위가 높은 항목을 우선순위에 두고 예산이 배분되었는지 검토
- 프로젝트 일정 및 단계별 소요 시간을 산정
  ex. 각 단계별로 필요한 기간을 산정하고, 프로젝트 전체 일정이 현실적인지 검토
- 유사 프로젝트와 비교하여 예산 및 기간의 적절성 분석
  ex. 과거 유사한 데이터 분석 프로젝트와 비교하여 예산과 일정이 적정한지 검토
- **리스크 분석 및 예산/일정 조정:** 예산과 일정 산정 시 고려해야 할 리스크를 식별하고 대비책을 마련
  ex. 데이터 품질 문제: 사전 데이터 검증 수행, 클라우드 비용 초과 등

1.4 분석 결과 활용계획 수립

이 단계에서는 데이터 분석 결과를 어떻게 업무에 반영할 것인지에 대한 활용 계획을 만들고 업무 성과를 지속적으로 모니터링할 수 있는 방안을 수립한다.

분석 결과 활용방안은 과제 수행 과정에서도 수립할 수 있지만, 과제 착수 전 해당 방안을 선제적으로 검토하고 분석 목적을 명확히 함으로써 과제 실패 가능성을 줄일 수 있다.

왜 데이터 분석을 하려고 하는지, 데이터 분석을 통해 어떤 결과를 도출하고자 하는지, 도출된 결과를 어떻게 활용할 것인지, 활용을 통해 어떤 효과를 얻을 것인지를 고려하여 활용계획을 수립해야 한다.

- 단순히 데이터를 도입하거나 분석하는 것만으로는 국민 만족도 향상과 같은 즉각적인 성과로 이어지기 어려우므로, 분석 결과를 통해 어떤 효과를 얻고자 하는지 명확한 목표를 설정하고, 이를 달성하기 위한 구체적인 활용계획을 수립하여 지속적으로 실행하는 것이 필수적이다.

과제 수행 후 분석 결과를 즉시 적용할 수 있는 단기 활용계획을 제시할 수 있어야 하며, 중장기적인 활용계획을 수립하고 상세화해야 한다.

데이터 분석이 끝나면 각 현업 부서에서 분석 결과를 활용해야 하며, 이때 이전에 수립한 활용방안이 계획대로 잘 수행되고 있는지 모니터링이 필요하다.

- 이를 위해, 데이터 분석 및 활용을 위한 구체적인 업무 분담, 책임 및 권한 체계를 미리 정의한다.

실질적인 정책 개선을 위한 분석 목적 구체화

공공기관이 데이터 분석과제를 추진할 때는 단순한 모델 개발에 그치는 것이 아니라, 분석 결과를 어떻게 활용할 것인지에 대한 명확한 계획을 수립해야 한다. 분석 목적이 구체적일수록 프로젝트의 방향성이 유지되고, 실질적인 정책 개선으로 이어질 수 있다.

분석 목적이 명확하지 않으면 프로젝트가 진행되면서 요구사항이 추가되고, 결국 특정 문제 해결을 위한 모델이 범용적인 형태로 변질되면서 실질적인 정책 개선에 기여하지 못할 가능성이 높아진다.

예를 들어, **'버스 하차 정보 예측 모델'**을 개발할 때 단순히 데이터를 분석하는 것에서 끝나는 것이 아니라, 이를 통해 **❶노선 개편을 위한 정책적 의사결정에 활용할 것인지, ❷버스 배차 간격 조정을 위한 기준으로 사용할 것인지** 등을 명확하게 정의해야 한다.
분석 결과를 어떻게 활용할지에 대한 구체적인 계획이 없으면, 처음에는 단순한 예측 모델을 만들려 했던 것이 점점 확장되면서 분석 목적이 흐려지고, 결국 다목적 모델이 되어버리는 문제가 발생할 수 있다.

과제 초기에 목표를 구체적으로 명확히 설정해야, 분석 결과를 어떻게 활용할 것인지에 대한 계획에 변동이 없게 된다.

따라서 공공기관이 데이터 분석과제를 추진할 때는 단순한 모델 개발에 그치는 것이 아니라, 분석 결과를 어떻게 활용할 것인지에 대한 명확한 계획을 수립해야 한다.

활용 성과에 대한 점검은 수립한 활용 계획에 대해 6개월이나 1년 단위로 측정하여 정량화하는 관리가 필요하다.

분석모델을 기관 내·외부에서 지속적으로 활용하고 발전시킬 수 있도록 기관 간 데이터 연계, 데이터 통합, 분석, 결과 활용 내용을 포함한 확산 계획도 고려하여 방안을 수립한다.

검토 필요사항

Key Question : 분석 결과 활용계획 수립 점검사항
데이터 분석 및 활용에 대한 방안을 수립할 때 전체 이해관계자들의 아이디어와 의견이 반영되었나? • 수요조사, 워크샵 등 구성원들의 의견수렴 과정 과정 필요
분석 결과를 활용하고, 적용하고, 활성화하려는 계획을 갖고 있나? • 결과를 즉시 적용하기 위한 단기 활용계획과 중장기 활용계획 수립 및 상세화
데이터 분석 종료 후 결과 활용을 위한 관련 부서 협조와 예산 확보가 가능한가? • 데이터 분석 및 활용방안은 기술적, 경제적으로 충분히 실현가능해야 함
분석모델을 지속적으로 유지보수하고 업그레이드할 수 있나? • ex. 유지보수 및 고도화 예산을 과제 기획 시 선제적으로 포함하여 추진 등

성과관리와 연계하여 활용계획 수립

분석 결과 활용 계획 수립 시, 과제 수행 결과를 업무에 적용하여 평가할 기준 및 계획을 세우고, 결과를 실제 업무에 적용하기 위한 단계적 활용방안과 확산을 위한 방안을 계획한다.

검토 필요사항

Key Question : 성과관리 계획

[성과관리 계획]

과제목적에 적합한 성과목표를 세웠는가?

- 성과목표가 분석과제의 사업 목적에 합당한지 검토
- 분석과제에 대하여 성과목표가 현실성이 있는지 파악
- 성과목표가 각각의 이해관계 당사자들 사이에 공유하고 협의

성과목표에 대하여 대상, 기간, 측정 항목들이 명확하고 구체적으로 작성되었는가?

- 성과지표를 구체화하는 방식인 SMART(구체적 Specific, 측정 가능 Measurable, 달성 가능 Attainable, 관련성 높은 Relevant, 시간 제한 있는 Time-bound) 방식을 활용해 목표 구체화

성과 측정 시기별 계획안이 마련되었는가?

- 측정 시기 간격이 성과를 측정하기에 적절한지 파악
- 성과 측정이 단기, 장기로 계획안이 마련되었는지 파악

[성과관리 지표 정의]

성과관리 지표가 분석과제를 평가하기 위해 타당한가?

- 지표의 측정 주체가 정의되어 있는지 파악
- 지표의 측정 단위가 적절한지 검토
- 지표 측정 산식의 근거가 타당한지 검토
- 측정 시기별 목표치 설정 근거가 타당한지 검토

성과관리 운영방안을 수립하였는가?

- 성과점검의 기관별 역할이 명확한지 파악
- 기관 간의 성과평가 범위를 정의하였는지 파악
- 성과점검 및 성과평가의 절차를 정의하였는지 파악
- 성과평가의 주기적 보고서 생성이 가능한지 파악
- 가능하면 주기적 생성을 위한 보고서를 위한 설계 및 개발 검토

[성과관리 활용 계획]

성과지표의 단기 / 중기 / 장기 활용계획을 수립하였나?

과제 수행 결과 활용계획을 합리적으로 세웠는가?

- 활용계획의 목표가 현실성이 있는지 파악
- 활용계획 설정 근거가 타당한지 점검

분석모형의 일반화, 확산 및 운영방안 수립 등의 활용방안을 제시하였는가?

- 단기 및 중·장기 활용계획의 적용대상 업무 정의
- 과제 확산으로 얻을 수 있는 기대효과 명시
- 과제 확산 계획의 타당성 검토

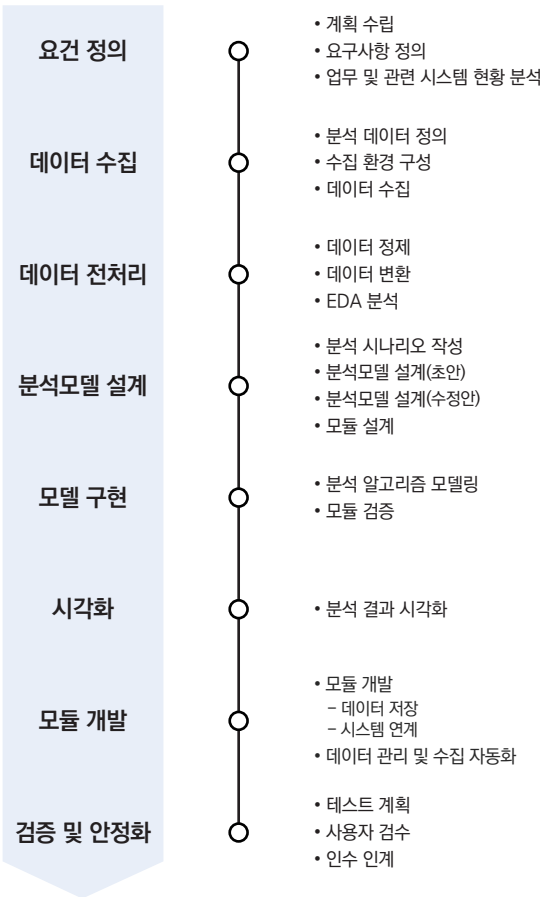
2. 분석 추진 단계

분석 프로젝트를 추진하게 되면, 분석과제의 요건 정의, 데이터 수집, 데이터 전처리, 분석 모델링, 모델구현, 시각화, 모듈 개발, 검증 및 안정화 등 체계적인 분석 프로젝트 절차를 수행해야 한다.

분석모델을 시스템에 탑재하거나 서비스로 개발하는 경우에는 모델 구현, 모듈 개발, 검증 및 안정화와 같은 후속 개발 과정이 필수적으로 요구된다. 반면, 정책 인사이트 발굴을 위한 단발성 현황 분석의 경우 요건 정의부터 시각화 단계까지를 핵심범위로 하고, 현황분석 보고서 작성으로 마무리 한다.

분석 프로젝트의 각 절차별로 점검사항을 미리 숙지하고 자칫 간과할 수 있는 수많은 시행착오를 줄일 수 있다.

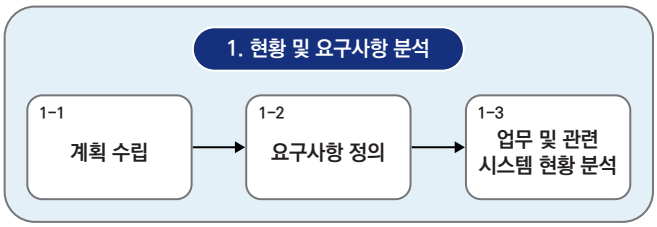
[그림 7] 추진단계별 점검항목



2.1 요건 정의

요건 정의 단계에서는 데이터 분석과 활용을 통해 해결하고자 하는 현안에 대해 명확히 이해하고 정의하며, 분석 추진을 위해 필요한 환경을 분석하고 프로젝트의 추진일정 및 과업을 수립하는 과정을 포함한다.

[그림 8] 요건 정의 단계의 주요 활동



2.1.1 계획 수립

계획 수립 단계에서는 요구사항을 정의하기 위한 실무자 및 관계자와의 인터뷰 계획을 포함하여 분석과제 전체 수행 일정계획을 수립한다.

검토 필요사항

Key Question : 일정계획 수립

일정계획을 수립할 때 아래의 사항을 고려하였는가?

- 프로젝트 추진을 위한 행정 절차 및 내부 검토 등 승인 일정
- 데이터 확보 가능 시점
  - 분석에 필요한 데이터가 내부 시스템에 존재하더라도, 개인정보 검토, 부서 협의, 추출 작업 요청 등으로 시간 소요가 크므로 고려해야 함

인터뷰 일정, 자료 협조 요청 시기 등 업무 담당자와의 협업일정 사전 검토

중간보고 및 자문회의 등의 피드백 반영을 위한 일정 고려

- 1·2차 보고 및 자문에 따른 분석방향 조정 및 최종 보고 등

예상치 못한 리스크 대비

- 담당자 변경, 데이터 오류, 내부 이슈(감사, 예산 등)을 고려한 버퍼기간 확보

2.1.2 요구사항 정의

요구사항 정의 단계에서는 과제의 기본적인 이해를 바탕으로 현업 및 IT 담당자와 인터뷰를 수행하고 현안을 분석하며 관련 자료를 수집, 정리하여 요구사항을 정의한다.

검토 필요사항

Key Question : 분석과제의 요구사항을 잘 정의하였는가?

요구사항 항목을 잘 포함하는가?

- **업무 목적 및 배경**: 왜 이 분석이 필요한가
- **업무 범위:** 어떤 데이터를, 어떤 기간, 어떤 지역/대상으로
- **데이터 요건:** 어떤 데이터가 사용되어야 하는가
  ex. 공공기관 내부 행정데이터 + 공공데이터포털에서 수집한 외부자료 포함
- **기대 성과:** 어떤 인사이트/산출물을 도출해야 하는가
  ex. 정책제안 3건 이상 도출, 시각화 리포트 포함
- **기술적 요구:** 어떤 분석 방법/기술을 써야하는가
  ex. 기계학습 기반 예측모델(랜덤포레스트 등)을 활용하여 수요예측 수행
- **시스템/리포트 요건:** 결과를 어떤 형태로 받아야 하는가
  ex. 웹 대시보드 제공 or Power BI 기반 리포트 제출
- **제약조건:** 개인정보보호, 보안, 업무협조 등

요구사항이 잘 정의되었는가를 판단하는 기준(예시)

- **"이 분석이 왜 필요한가?"**에 대한 문제의식이 명확한지 체크
- 분석범위가 되는 시간, 공간, 대상 범위가 명확한지 체크
- 데이터 가용성이 고려되었는지 체크
- 산출물이 정책적 시사점이나 개선안 도출로 쓸 수 있는지 체크
- 기술 수준이 현실적인지, 예산과 일정안에 실현 가능한지 체크

인터뷰 수행

분석 프로젝트에서 초기 인터뷰는 단순한 자료 요청이 아니라, 문제 정의, 데이터 맥락 이해, 실무 관점 수집을 위한 핵심적인 절차이므로, 인터뷰를 잘 설계해야 요구사항이 구체화되고 분석의 방향성과 실효성을 높일 수 있다.

주요 이해관계자와의 인터뷰 또는 워크숍 등을 통해 분석 목표 및 요구사항을 구체화하고 의견을 조율할 수 있다.

검토 필요사항

Key Question : 인터뷰 수행
[인터뷰를 통해 어떤 내용을 구체화해야 하는가]
업무/정책의 맥락 파악 • 분석이 어떤 행정업무/정책과 연결되는지 이해
실제 문제 인식 확인 • 담당자 입장에서 느끼는 **"실제 문제"**를 듣는 것
데이터의 의미와 맥락 확보 • 데이터 컬럼의 의미, 수집 방식, 왜곡 가능성 등 파악
활용 기대사항 확인 • 분석 결과를 어디에, 어떻게 쓸 것인지 명확히
기존 시도 및 한계 • 유사 시도가 있었는지, 무엇이 부족했는지
협조 가능 범위 확인 • 추가 인터뷰, 자료 협조, 검토 시간 등 행정 여건 파악

선행사업 분석 및 자료조사

선행 사례를 분석하여 요구사항 구체화에 반영한다. 선행사례 및 유사사례 분석은 유사한 목적과 범위를 가진 프로젝트에서 도출된 성공요인과 실패요인을 파악함으로써, 프로젝트의 요구사항을 구체화하고 현실적이며 효과적인 방향을 설정하는 데 목적이 있다.

검토 필요사항

Key Question : 선행사업 분석 및 자료 조사

선행 사례 분석 대상 예

- **국내 유사 과제:** 동일 분야(ex. 교통, 환경, 보건 등) 또는 동일 유형(ex. 정책 수립, 민원 개선 등)의 데이터 분석 사례
- **해외 우수 사례:** 유사한 공공 문제 해결을 위해 데이터 분석이 활용된 국제 사례
- **민간 벤치마킹 사례:** 공공영역에도 적용 가능성이 있는 민간 분석 사례

선행 사례 분석 시 분석 항목 예시

- **사업 목적 및 배경:** 어떤 문제 해결을 위해 분석이 수행되었는가
- **분석 데이터 및 수집 방식:** 어떤 데이터를 활용했는가, 어떻게 수집했는가
- **분석 기법 및 도구:** 어떤 분석 방법론을 사용했는가(ex. 시계열, 분류, 텍스트 마이닝 등)
- **주요 성과 및 활용 결과:** 분석 결과가 어떻게 정책이나 서비스에 반영되었는가
- **문제점 및 개선사항:** 시행착오나 한계는 무엇이었는가
- **운영 및 유지보수 체계:** 분석 이후의 데이터 갱신, 시스템 관리 방안 등

2.1.3 업무 및 관련 시스템 현황 분석

기존 시스템에 새로운 기술을 도입할 때는 기존 시스템과 호환되는지 확인하는 것이 중요하다. 그렇지 않으면 데이터가 손실되고, AI 서비스 오류 문제가 발생하게 되면, 심한 경우 도입 자체가 물거품이 될 수 있다. 그러므로 데이터, AI 모델 구축, 서비스 운영 등 과제를 진행하는 단계마다 기술과 서비스의 호환성을 검토해야 한다.

검토 필요사항

Key Question : 기술과 서비스 인프라가 호환 가능한가?

[호환 가능성 검토 사항 예시]

개발 환경 및 구현 언어의 호환성

- 기존 시스템이 사용하는 언어(Java, Python, R 등)와 새 기술의 언어가 호환되는가?

데이터 포맷 및 인터페이스 호환성

- 기존 시스템의 데이터 포맷(CSV, JSON, XML 등)과 새 기술이 요구하는 데이터 포맷이 호환되는가?

인프라 자원 및 하드웨어 요건

- AI 모델 학습/추론에 필요한 GPU/CPU/메모리/스토리지 자원이 확보되어 있는가?

검토 필요사항

Key Question : 수집 환경 분석

내부망 또는 외부망에서 데이터 수집이 가능한지 점검했는가?

- **내부망/외부망 여부 점검**
  ex. 수집 대상 데이터가 내부망에서 접근 가능한지, 외부망인가 확인
- **외부망 사용 계획 점검**
  ex. 외부망 활용 시 접근 경로(IP, VPN 등)와 보안 정책을 마련했는가 확인
- **망 분리 정책 점검**
  ex. 망 분리 환경에서 수집이 가능한 방식(망 연계 장비, 보안 게이트웨이 등)을 정의했는지 점검

외부망 활용 시 데이터 수집 루트를 정하였는가?

- **수집 대상의 위치 점검**
  ex. 데이터가 저장된 위치(서버, DB, API, 파일 등)를 명확히 파악했는지 확인
- **수집 방식 정의**
  ex. 크롤링, API, DB 연동 등 어떤 방식으로 수집할지 계획되었는지 확인

데이터 수집을 위한 개발 환경 계획이 잘 되었는가?

- **개발 환경 확보 점검**
  ex. 데이터 수집을 위한 개발 및 실행 환경(서버, 언어, 툴 등)이 구축되었는지 확인
- **권한 및 접근성 점검**
  ex. 계정, 인증, 접근 권한 등 필요한 접근 설정이 완료되었는지 확인

데이터 수집을 위한 환경 분석이 완료되었는가?

- **보안 및 법적 검토**
  ex. 개인정보 포함 여부, 내부 정책 또는 외부 규정에 따라 수집 가능한 데이터인지 확인

2.1.4 주요 산출물

데이터 분석과제의 관리자 관점에서, 본 단계에서 마련해야 할 주요 산출물은 아래와 같다.

요구사항 정의서
활용부서와 협의된 요구사항을 정리한 문서

# 현황 분석 보고서	
데이터 분석을 위한 조직의 시스템 및 데이터 현황을 분석한 문서	
시스템 현황 분석	데이터 정의서

요구사항 정의서(예시)

제안서 요구사항(기존)	주관기관 협의된 요구사항(현재)	비고
업무 프로세스, 데이터 현황, 인프라 현황 등 현황 및 요구사항을 분석	• 업무 프로세스, 데이터 현황, 인프라 현황 등 현황 및 요구사항을 분석	-
최근 5년 이내 불만 승인 및 종결 고객 데이터 수집 (고객정보, 접수일자, 불만 유형 등)	• 19.01.01 ~ 23.12.31 이전에 고객 불만 데이터(최근 5년 이내)를 수집하여 학습용 데이터 구축 • 불만유형은 1.제품의 파손 2.배송지연 3.반품으로 정의 • 종결 건은 제외 • 고객 불만정보, 접수일자, 불만유형 등 고객정보 테이블의 데이터를 수집하여 활용	• 회의를 통해 수집 및 분석 대상 고객정보 기준을 체적으로 정의 • 실무자 의견을 반영하여 수집 테이블 확정 ***참고: v1.0.고객정보.hwp**
고객정보에서 반품일수 예측	1. 반품 일수는 최초 배송완료일~반품 도착일 까지의 일수 2. 재반품 일수를 구분하여 예측	회의를 통해 반품일수 정의를 구체화

시스템 현황 분석 보고서(예시)

*** 연관 시스템 보유 현황**

시스템명	네트워/위치	용도	설명
고객 시스템 (운영)	내부망 (본사)	업무처리용	• 실제 고객 업무상에서 활용되는 시스템으로, Oracle DB 활용 중
고객 시스템 (개발)	내부망 (본사)	백업용	• 데이터 백업을 위해 활용되는 시스템으로 매월 데이터가 백업되고 있으며 운영시스템과 동일 스펙(Oracle DB)으로 구성되어 추가적인 개발사항이나 점검 및 테스트 용도로 활용 중
지능형 시스템	내부망 (A지사)	분석 및 연구용	• 데이터 연구 및 분석 용도로 활용되고 있으며, 실제 데이터 분석/AI 서비스를 시스템 내 개발하여 고객 시스템으로 서비스하고 있음 • 현재 고객시스템(운영)과 연계되어 매일 주요 테이블 및 데이터를 주기적으로 수집 및 업데이트하고 있음. 〈SW현황〉 CentOS7.564bit JDK1.8 Python3.8 TIBERO6Standard(8Core) SecureTOSV5.0 SECUREGUARDAMNodeLic Zenius-EMSAgentv7.0(Server)

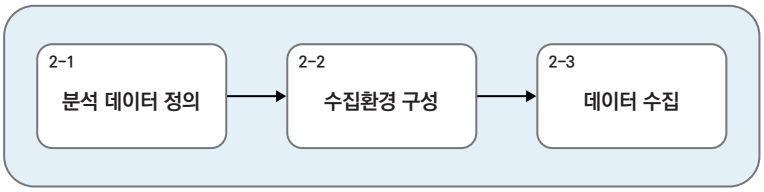
데이터정의서(예시)

순번	파일명	항목명 (영문)	항목명 (한글)	데이터 타입	데이터 수	NON-NULL 수	NOT-NULL (%)	고유값 수	고유ID 수	고유ID (%)	명목형 변수					
											최빈값명 (RANK 1)	최빈값명 (RANK 2)	최빈값명 (RANK 3)	최빈값수 (RANK 1)	최빈값수 (RANK 2)	최빈값수 (RANK 3)
1	AB	SAMPLE_ID	식별자	VARCHAR (10)	590,750	590,750	100%	590,750	590,750	100%						
2	AB	JAEHAE_DT	점검일자	CHAR (8)	590,750	590,750	100%	9,113	590,750	100%	20210204	20210114	20181217	548	500	498
3	AB	SANGBYEONG_JONGRYU_CD	유형코드	VARCHAR (10)	590,750	586,155	99%	16	586,155	99%	1	8	3	263,605	112,560	37,098

2.2 데이터 수집

데이터 수집 단계는 분석에 필요한 데이터를 내부 및 외부에서 확보하며, 분석 데이터를 정의하고 수집 환경을 구성하여 데이터를 수집하는 단계이다.

[그림 9] 데이터 수집 단계의 주요 활동



2.2.1 분석 데이터 정의

분석 데이터 정의 단계에서는 분석 범위에 맞는 데이터 수집 대상 목록을 정리하고 보유기관 및 담당자를 조사한다.

분석 데이터 정의에 앞서 보유데이터 파악을 위한 메타데이터 보유 여부를 확인해야 한다. 메타데이터는 데이터의 구조와 의미를 기술하는 정보로, 사용자가 원하는 데이터를 신속하고 정확하게 탐색할 수 있도록 지원한다. 또한, 데이터의 신뢰성과 품질을 평가하기 위해 필요한 정보를 제공한다.

검토 필요사항

Key Question : 필요 데이터 정의

데이터 수집 대상 목록을 작성하였는가?

- 사전 조사 및 인터뷰를 바탕으로 수집 대상 목록 작성
- 추가 수집 대상 데이터가 없는지 파악
- 데이터 종류, 양, 보관 방식, 수집 주기와 분석 주기, 백업과 이중화 등 원시(Raw) 데이터에 대한 요구사항 도출 및 분석용 범위를 구체적으로 파악

검토 필요사항

Key Question : 데이터 확보 방안 수립 및 점검

조사한 데이터는 사용 가능한 데이터인가?

- 데이터가 활용 가능한 형태로 제공되는지(ex. 비구조화된 데이터는 전처리 필요), 데이터 품질 검토(결측치, 오류 데이터, 최신성 등), 데이터 제공방식(API 등), 데이터 라이센스 및 법적 제한(개인정보 포함 여부 등), 데이터 활용 목적과의 적합성(데이터 구조가 분석 목적과 일치하는지), 데이터 규모(건수, 크기, 길이 등) 점검

조사한 데이터는 수집 가능한가?

- 데이터 제공방식(API 등), 자동화된 수집이 가능한지 확인
- 수집시의 데이터 크기 및 저장가능 여부(GB 단위 이하, 클라우드 저장 가능 등), 데이터 수집 비용 및 인프라(API 사용요금, 서버 및 저장소 비용 등), 데이터 제공기관의 사용승인 여부 및 데이터 업데이트 주기 확인 등

데이터 수집 항목 중 민간 기업에서 수집되는 경우가 있다면, 데이터 수집 비용은 적절한가?

- 공공데이터로 대체 가능한지 검토하고 민간데이터가 공공데이터에 비해 추가적인 가치를 제공하는지 평가
- 데이터가 프로젝트 목적에 부합하고 신뢰성과 최신성을 확보하고 있는지 확인
- 데이터제공자가 신뢰할 수 있는지 확인

수집 대상 데이터 보유기관 및 담당자를 조사하였는가?

- 데이터 보유 가능성이 높은 기관 리스트업(정부, 연구기관, 기업, 비영리기관 등)
- 기관 대표 홈페이지, 공공데이터포털, 범정부 데이터분석시스템 및 오픈데이터 플랫폼 활용, 기관에 공식문의, 전문가 및 관계자 활용, 정보공개포털 활용
- 제공방식 데이터 포맷, 제공 주기(일회성 제공 또는 정기적 업데이트 가능 여부), 법적, 윤리적 고려사항(기관 약관, 개인정보보호법 등 확인) 등 협의

보유기관이 데이터를 제공하는 형식을 조사하였는가?

- **조사 대상:** 어떤 기관에서 데이터를 제공하는지 (ex. 정부 기관, 연구소 등)
- **데이터 형식:** CSV, JSON, XML, API, DB 연동 등 다양한 형태로 제공되는지 여부

보유기관으로부터 데이터 메타데이터(칼럼, 건수, 크기, 길이)를 입수하였는가?

- 기관에서 제공가능한 데이터 형태 확인(API, CSV 등)
- 메타데이터와 함께 샘플 데이터를 요청하여 데이터의 구조 및 품질을 확인
- 데이터 보안 및 약관을 검토(개인정보포함여부, 데이터 사용약관 및 비밀유지협약(NDA) 필요 여부 검토 등)

검토 필요사항

Key Question : 개인정보에 대한 점검

수집 예정 데이터에 개인정보가 포함되어 있는가?

- **내부 개인정보보호 및 보안 관련 가이드가 있는지 우선 확인**
- **개인정보 침해 가능성 검토를 하였는지 확인**

  ex. 직접 식별 정보(이름, 주민 번호) 포함 여부, 간접 식별 가능 정보 포함 여부, 데이터 조합 시 개인 식별 가능 여부, 데이터 암호화/비식별화 처리 여부, 법적 요건 준수 여부(동의서 확보 등), 수집 목적과의 적합성(필요 데이터만 포함)

수집 예정 데이터에 개인정보가 포함되어 있다면, 유출될 문제에 대하여 방안을 세웠는가?

- **개인정보 데이터 제공 범위 협의**

  ex. **최소한의 데이터만 제공하는 원칙 적용(Data Minimization):** 데이터 제공기관과 협의하여 최소한의 필드(칼럼)만 포함하도록 조정. 개인정보 필터링 및 마스킹 등 개인정보 필터링 및 비식별화 수행

- **개인정보 보안 방안 협의**

  ex. 데이터 저장 및 접근 통제 강화, 데이터 전송 보안, 개인정보보호법 준수 등

기술의 안전성(위험요소 진단)

데이터 분석과정에서 AI 기술을 적용할 경우, 치명적 피해·사고, 법적·경제적·기술적·전문인력 문제 등 위험요소에 대해 사전에 인식하여 분석 및 대응 방안을 마련하는 것이 중요하다.

검토 필요사항

Key Question : AI 적용시 예상되는 위험요소를 분석하였는가?

[위험요소]

할루시네이션* (Hallucination)

*** AI 모델이 실제로 존재하지 않는 정보를 생성하거나 왜곡된 데이터를 출력하는 현상.**
주요 원인으로는 모델의 편향성, 오류나 잡음(NOISE)이 많은 학습데이터, 사람의 악의적 개입 등이 있음

- AI의 잘못된 예측으로 심각한 문제 발생(사망사고, 인체 위해성 등) 가능성 검토 및 이로 인해 발생할 수 있는 문제 검토
  ex. 허위정보 확산, 신뢰도 저하, 책임문제, 모델활용 제약 등
- AI의 실수가 치명적이지 않은 분야 선택
  ex. 특히 의료 및 법 분야는 개인 피해에 가장 밀접하므로 보다 신중한 접근 필요

법적·윤리적 이슈

- 데이터의 보안 문제, 저작권 문제 등 법적 요구사항 위배 여부 검토 및 AI 기술 활용 서비스의 윤리적 문제 고려

비용-효과(편익) 이슈

- 비용-편익 또는 비용-효과 분석으로 투자 대비 수익이 낮거나 회수 기간이 긴 기술인지, 서비스 활용·유지보수 비용 등 검토

기술적 호환성

- AI 서비스가 기존 시스템과 기술적으로 호환이 잘되는지 검토

전문인력 유지 가능성

- 작업자 이탈 및 변경 시 AI 기술을 유지관리 할 수 있는지 검토

데이터의 보안성 및 투명성 확보

데이터를 외부 위협이나 내부의 실수로부터 보호하고 데이터의 기밀성, 무결성, 가용성을 확보하기 위한 방안은 프로젝트 계획 단계에서부터 필수적으로 고려되어야 한다. 또한 데이터 처리 전체 과정에 대한 명확하고 투명한 정보제공을 통해 신뢰성과 적법성을 확보할 수 있다.

검토 필요사항

Key Question : 데이터 보안등급 설정 및 관리 절차를 투명하게 하였는가?

데이터 보안성

- 데이터 공개·비공개 등급 기준 설정 및 등급별 접근자의 권한 설정·사용자 진위 확인 등으로 기밀성(인가된 사람에게만 공개) 확보
- 원래의 데이터 내용을 훼손하여 활용하는 것을 금지하여 무결성(데이터가 변형 없이 전달) 확보
- 가용성(필요시 원하는 데이터에 접근 또는 사용 가능) 확보

데이터 관리의 투명성

- 데이터 보안을 위해 데이터 관리 명세서 작성

2.2.2 수집 환경 구성

수집 환경 구성 단계에서는 기관이 보유한 수집 환경을 활용하여 데이터를 수집하고 이를 위한 환경을 구성한다.

검토 필요사항

Key Question : 데이터 수집 환경 구성

타 기관 데이터 수집 환경을 구성하였는가?

- 외부 기관이 보유하고 있는 데이터의 경우 데이터 이관 및 연동 시스템 구현 필요
- 데이터 연계 방안에 대하여 협의하였는지 검토

수집할 데이터의 크기 및 최대 저장기간 등을 고려하여 저장공간을 설계하였는가?

- 데이터의 크기 및 저장공간을 계산하고 DB를 선택(RDBMS, NoSQL, Data Warehouse 등)
- 핫 데이터 / 콜드 데이터를 분리 설계하고, 최대 저장기간 및 자동 삭제 정책을 설정
- **핫 데이터(Hot data):** 자주 접근되거나 실시간 처리가 필요한 데이터
  ex. 스마트시티의 실시간 교통 데이터 등
- **콜드 데이터(Cold data):** 거의 접근하지 않지만 보관은 필요한 데이터
  ex. 3년 전 고객 구매 이력 등
- 백업 및 보안 강화(암호화, 접근제어, 복제 설정 등)

데이터 수집 기술이 웹 크롤링일 경우 수집 환경 구성에 아래와 같은 사항을 검토하였는가?

- 소셜 데이터와 같은 외부 공개 데이터를 자체 수집하는 경우, 범정부 데이터분석시스템의 수집 환경 활용 또는 데이터 크롤링 시스템 별도 구현

데이터 수집을 위한 사전테스트를 진행하였는가?

- 데이터 소스 및 접근성 확인
  (데이터가 API, 웹 크롤링, 파일 다운로드, 데이터베이스 연결 등 어떤 방식으로 제공되는지 확인)
- 샘플 데이터 수집 및 품질 점검
  (샘플 데이터를 수집한 후, 결측치, 중복 데이터, 이상치 여부를 점검)
- 데이터 저장 및 처리 성능 테스트
  (수집된 데이터를 저장할 DB 또는 파일 저장소에 대량으로 입력할 때, 성능이 저하되지 않는지 확인 ex. Batch Insert vs. Single Insert 성능 비교 등)
- 데이터 암호화 및 접근 제어 테스트(데이터 수집 중 개인정보나 민감한 데이터가 포함될 가능성이 있는 경우, 암호화 및 접근 제어 정책을 사전테스트)
- 데이터 수집 자동화 및 오류 처리 테스트(크롤러/스크립트 실행 자동화 및 오류 처리 등)

2.2.3 데이터 수집

데이터 수집 단계에서는 수집 대상을 명확하게 정의하여 보유기관에 요청하여 수집한다.

검토 필요사항

Key Question : 데이터 수집

데이터를 수집하기 위한 보유기관과의 협의 일정을 세웠는가?

- 데이터 제공목적, 요청할 데이터 범위, 법적/정책적 문제(사전 검토), 데이터 제공방식(기관의 기술적 지원 가능여부 확인) 등 협의

보유기관에 요청할 데이터 리스트는 구체적으로 작성하였는가?

- 대상 데이터의 시간, 공간의 범위가 명확한지 파악
- 모든 보유기관에 요청 시 대상 데이터의 시간, 공간의 범위가 동일한지 파악

기관 내·외부 데이터 수집을 완료하였는가?

- 수집한 데이터가 내부 데이터(기관 내부 보유데이터)와 외부 데이터(API, 웹 크롤링, 제휴 데이터 등)에서 모두 확보되었는지 확인
- 데이터가 수집되었더라도 결측치, 중복 데이터, 이상치 존재 여부 점검
- 데이터가 지정된 저장소(DB, 파일 시스템 등)에 정상적으로 저장되었는지 검토
- 수집된 데이터가 보안 정책에 맞게 저장되고, 접근 권한이 적절하게 설정되었는지 확인(DB 접근 권한, 데이터 암호화, API 접근 제한 등)
- 모든 수집 과정이 정상적으로 완료되었는지 로그 및 문서화를 점검(API 응답코드, 크롤링 실행로그, DB 저장로그 등)

데이터 수집 시 주의사항

- 데이터의 품질, 수집 기술, 데이터 보안 및 개인정보보호 문제 등 고려할 부분이 많으므로 전문가의 조언이 필요
- 데이터 수집 활동은 분석 결과의 품질을 좌우하는 중요한 과정이므로 분석에 필요한 데이터 항목들이 반드시 포함될 수 있도록 사전 점검 필요

데이터 수집 및 연계 방안의 주요 고려사항

데이터 분석 프로젝트를 수행할 때, 데이터 수집 단계에서 내부 데이터와 외부 데이터를 어떻게 연계할 것인지에 대한 계획을 사전에 수립하는 것이 필수적이다.

특히, 프로젝트가 진행된 후 데이터 연계 문제가 발생하면 이를 해결하기 어려워지므로, 초기 단계에서 데이터의 연결 가능성과 오류를 점검하는 과정이 반드시 필요하다.

데이터 연계를 고려할 때 첫 번째로 해야 할 일은 내부 데이터의 구조와 연계 가능성을 확인하는 것이다.

- 내부에서 보유한 데이터를 분석하여 필요한 변수들이 존재하는지, 각 데이터 간의 키값이 올바르게 설정되어 있는지, 연결 시 인터벌(시간 간격)이나 단위 차이로 인해 연계가 불가능한 문제가 없는지 등을 검토해야 한다.
- 이러한 점검이 이루어지지 않으면, 프로젝트 후반부에서 데이터가 분석에 활용될 수 없다는 사실을 뒤늦게 발견하는 문제가 발생할 수 있다.

또한, 외부 데이터를 활용할 경우, 내부 데이터와의 연계 가능성을 미리 검토하는 것이 중요하다.

- 외부 데이터를 확보하는 것 자체도 중요한 작업이지만, 확보한 데이터가 내부 데이터와 제대로 연결될 수 있는지를 먼저 판단해야 한다.
- 외부 데이터가 내부 데이터와의 연결 키를 제공하는지, 만약 제공하지 않는다면 대체할 수 있는 변수를 활용할 수 있는지 등을 사전에 확인해야 한다.
- 이러한 연계 방안이 마련되지 않으면, 외부 데이터를 확보한 이후에도 제대로 활용하지 못하는 상황이 발생할 수 있다.

이러한 데이터 연계 문제는 데이터 수집 단계에서 고려할 사항이 아니라, 프로젝트 기획 단계에서부터 검토해야 하는 요소이다.

분석 주제를 선정할 때, 내부 데이터만으로 분석이 가능한지, 외부 데이터가 추가로 필요한지 판단하고, 필요하다면 연계 방안을 사전에 마련해야 한다.

주제 선정 단계에서 이러한 고려가 이루어지지 않으면, 분석 진행 과정에서 데이터가 부족하거나 연계할 수 없는 문제로 인해 프로젝트가 실패할 가능성이 높아진다.

결국, 프로젝트 기획 단계에서 내부 데이터 간의 연계 가능성과 외부 데이터 활용방안을 사전에 검토하고 이를 체계적으로 정리하는 과정이 필수적이다.

데이터 연계가 제대로 이루어지지 않으면 분석의 정확성과 활용성이 떨어지기 때문에, 이를 방지하기 위해 기획 단계에서부터 데이터의 연계 가능성을 철저히 점검해야 한다.

이러한 부분들은 주제 선정과 데이터 수집 단계에서 모두 확인할 필요가 있다.

검토 필요사항

Key Question : 데이터 수집 및 연계 방안 점검사항
내부 데이터의 구조와 연계 가능성을 확인하였는가? ex. 내부에서 보유한 데이터를 분석하여 필요한 변수들이 존재하는지, 각 데이터 간의 키값이 올바르게 설정되어 있는지, 연결 시 인터벌(시간 간격)이나 단위 차이로 인해 연계가 불가능한 문제가 없는지 등을 검토
외부 데이터를 활용할 경우, 내부 데이터와의 연계 가능성을 미리 검토하였는가? ex. 외부 데이터가 내부 데이터와의 연결 키를 제공하는지, 만약 제공하지 않는다면 대체할 수 있는 변수를 활용할 수 있는지 등을 사전에 확인

2.3 데이터 전처리

데이터 전처리 단계는 수집된 데이터를 파악하여 데이터의 오류를 점검하고, 데이터를 정제함과 동시에 매칭키(Matching Key) 값을 기준으로 여러 데이터를 융합하고 기초 통계 분석을 통해 데이터의 고유 패턴을 파악한다. 매칭키는 데이터 분석이나 데이터베이스에서 서로 다른 데이터 세트를 연결하거나 비교하기 위해 사용하는 고유한 식별자다. 주로 두 개 이상의 데이터 소스에서 관련된 정보를 결합하기 위해 사용된다. 매칭키를 잘 설정하면 데이터 통합이 쉬워지고, 데이터 분석의 정확성이 높아진다.

예를 들어, 고객 데이터베이스와 주문 데이터베이스가 있을 때, 고객의 ID나 이메일 주소가 매칭키가 될 수 있다. 이렇게 매칭키를 사용하면 각 고객이 어떤 주문을 했는지 쉽게 연결할 수 있다.

데이터 전처리는 데이터 정제, 변환, 통합, 축소 등 다양한 전처리 단계를 수행한다.

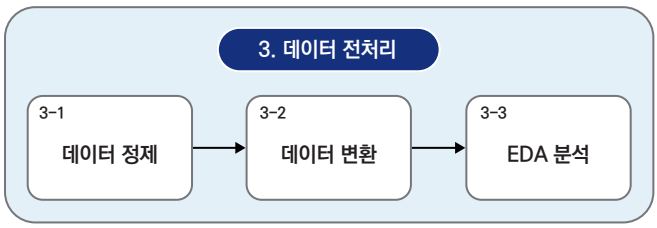
- **데이터 정제(Cleansing):** 데이터에서 결측값, 중복값, 이상값 등을 제거하거나 수정하는 과정
  ex. 설문조사 데이터에서 응답하지 않은 항목을 처리하는 등
- **데이터 변환(Transformation):** 데이터를 분석에 적합한 형태로 변환하는 과정
  ex. 범주형 데이터를 숫자형 데이터로 변환하거나, 데이터의 단위를 맞추는 등
- **데이터 통합(Integration):** 여러 출처에서 수집된 데이터를 하나의 일관된 데이터셋으로 통합하는 과정
  ex. 여러 시스템에서 수집된 고객 데이터를 하나의 데이터베이스로 통합
- **데이터 축소(Reduction):** 분석의 효율성을 높이기 위해 데이터를 요약하거나 축소하는 과정
  ex. 주요 변수만을 선택하거나, 차원 축소 기법을 사용

데이터 전처리는 데이터를 정제·가공하여 데이터의 품질을 높이고 결과의 신뢰성을 높이는 과정으로, 단계별 품질관리 활동을 수행하기 위해 품질 기준을 정의하고, 데이터를 정제·가공하는 것은 매우 중요하다.

지속적으로 데이터 분석을 수행할 경우의 전처리

분석 대상 업무가 일회성이 아닌, 지속적으로 서비스를 수행해야 하는 경우 API로 연계하기 위한 작업과 기관 자체 인프라를 활용하여 데이터 수집, 정제, 표준화, 융합이 자동화되도록 시스템 연계를 추진하고, 모듈을 설계하고 구현해야 한다.

[그림 10] 데이터 전처리 단계의 주요 활동



2.3.1 데이터 정제

데이터 정제 단계에서는 다양한 매체로부터 수집된 데이터를 기초 분석하여 오류가 있는지 파악하고, 원하는 형태로 변환하여 데이터화한 후 저장하고, 데이터를 활용할 수 있도록 품질을 확인하고 관리한다.

검토 필요사항

Key Quuestion : 데이터 정제

데이터 기초분석을 하였는가?

- 데이터를 정제하기 전에 탐색적 데이터 분석(Exploratory Data Analysis, EDA)을 수행해 데이터의 구조와 품질을 파악하고, 정제 과정에서 필요한 조치를 결정
- 이를 위해 데이터 구조 분석, 결측치 및 이상치 탐색, 데이터 분포 확인, 변수 간 관계 분석 등

오류 데이터 제거, 캐릭터 셋 통일, 길이 체크, 날짜 포맷 통일 등 데이터 검증 및 정제 과정을 거쳤는가?

- 논리적으로 불가능한 값(ex. 음수 나이 등) 제거, 허용되지 않는 카테고리값 제거, 미래 날짜 또는 잘못된 날짜 제거 등
- 데이터 파일 또는 DB에서 인코딩 방식이 다르면 문자 깨짐(Encoding Issue)발생하므로 특정 표준 인코딩으로 통일
- 각 필드(컬럼)의 데이터 길이가 허용된 범위를 초과하는지 또는 너무 짧은지 확인하는 등

데이터에 포함된 개인정보는 모두 비식별화 하였는가?

- 개인정보 항목 식별 및 탐색(데이터셋 내에서 개인정보가 포함될 가능성이 있는 필드(컬럼)를 식별. 직접, 간접 식별정보, 기기 정보, 건강 및 금융정보 등)
- 직접 식별 가능한 정보는 삭제하거나 대체(마스킹, 해싱, 암호화 등)하여 비식별화(이름, 주민번호, 전화번호, 이메일 등의 데이터 마스킹 처리, 암호화 적용 여부 확인)
- 간접 식별 가능 정보(Quasi-Identifiers) 점검(다른 데이터와 결합하면 특정 개인을 식별할 수 있는 정보)
  ex. 대학명 + 입학연도 + 전공 ▸ 특정 개인 유추 가능
- 비식별화 효과 검증(K-익명성 기준, L-다양성 기준, 비식별화 후 데이터 활용 가능성 검토 등)
- 최종 개인정보 비식별화 검증 테스트
  ex. 직접 식별 정보(이름, 주민등록번호 등) 제거 또는 마스킹, 간접 식별 정보(연령, 지역 등) 범주화 및 가명처리, 데이터 암호화 또는 해싱 적용 여부 확인, 재식별 위험 평가(K-익명성, L-다양성) 수행

사용하는 모든 데이터가 개인정보보호법에 따른 제약사항을 준수하고 있나?

- 현재 사용 중인 데이터가 개인정보보호법상 개인정보(Personally Identifiable Information,PII)에 해당하는지 점검(직접 식별정보(이름, 주민등록번호 등), 간접 식별정보가 조합될 경우 식별 가능성 분석, 민감정보(의료, 신용카드 등), 기기정보 등
- 데이터 수집 시 데이터 수집 전 사용자의 동의 획득 여부 확인, 목적 제한 원칙(데이터가 수집 목적 외에 사용되지 않는지), 최소 수집 원칙(불필요한 데이터는 수집하지 않았는지) 확인
- 데이터 저장 및 암호화 여부 적용 여부 확인(비밀번호 및 중요 데이터 암호화 적용, 저장된 데이터 접근권한 제한 등)
- 데이터 접근 권한 및 보안 점검(개인정보 접근 권한이 제한되었는지, 관리자 및 일반 사용자 접근 구분 여부 등)
- 데이터 공유 및 제3자 제공 여부 점검(제3자 제공 동의, 데이처 처리계약(NDA)-외부 업체가 데이터를 처리하는 경우 계약 필요)
- 데이터 보관 및 파기 절차 준수 여부 확인(보관 기간이 초과된 데이터 자동 삭제 여부 확인 ex. 금융거래 기록 5년, 로그데이터 1년, 사용자 계정정보 등)

데이터 품질 확보

분석을 위해 필요한 데이터를 정의하고 데이터의 속성을 제대로 알아야 하며, 편향성 없는 데이터를 활용해야 한다. 이를 위해 데이터의 출처 및 확보방안을 마련해야 한다.

그리고 데이터의 품질을 확보하기 위해서는 데이터의 정제 및 가공에 대한 방법과 기준을 문서화하여 현행화하고 결측치, 이상치 점검 등 데이터 탐색 결과를 제시하여야 한다.

검토 필요사항

Key Question : 데이터 품질관리 방안을 검토하였는가?

품질관리 기준

- 데이터 정제, 가공, 학습 등 품질관리를 위한 지표 기준 정의
- **지표 기준에 따른 품질 자가점검 대상범위를 설정하여 단계별로 자가 점검***
  * **시작**(모수의1%) ▸ **수집**(10%) ▸ **DB구축**(50%) ▸ **AI구현**(100%)

데이터 정제

- 정제 방법 및 기준 현행화(정제 기준 변경 관리 등)를 위한 기준점검, 전처리를 통한 오류 제거(결측치/이상치 식별 및 처리)
- 개인정보/민감정보 등 비식별화 기준 및 방안 마련
- 정제단계에 사용될 정제도구 및 저장환경 검토
- 원천데이터 품질검사

데이터 가공

- 데이터 가공(**라벨링***, **인코딩**** 등) 방법 및 기준 현행화(가공 기준 변경 관리 등)
  * **라벨링:** 데이터에 대한 설명 추가
  ** **인코딩:** 데이터 변환작업 (ex. 문자 ▸ 수치화, 정규화, 차원축소 등)
- 가공 도구 및 저장환경 검토
- 가공데이터 품질검사

[표 11] 품질관리 목표

구분	품질요소	품질기준	검사내용
데이터 품질	완전성	개별완전성	필수 컬럼의 값 누락 여부 확인
	유효성	범위유효성	컬럼값이 주어진 범위 내에 존재하는지 확인
	일관성	형식표준화	날짜, 시간, 등의 정보가 동일 형식으로 표준화 되었는지 확인
	보완성	오류조치율	오류 및 누락 데이터의 조치 및 처리 여부 확인

비식별화 전문기관 활용

데이터 결합 전문기관은 서로 다른 개인정보처리자가 보유한 가명정보를 안전하게 결합하기 위해 지정된 기관이다. 이러한 기관은 개인정보보호위원회 또는 관계 중앙행정기관의 장에 의해 지정되며, 가명정보의 결합, 폐쇄공간 제공 및 처리 지원, 반출심사 등의 역할을 수행한다.

- 일례로, 보건복지부는 국민건강보험공단, 건강보험심사평가원, 국립암센터 등을 보건의료분야 결합전문기관으로 지정하여 보건의료 데이터의 안전한 결합과 활용을 지원한다. 과기부의 경우 한국지능정보사회진흥원(NIA)을 제1호 가명정보 결합전문기관으로 지정하여 다양한 분야의 데이터 결합을 지원한다.

데이터 가명처리 지원센터는 개인정보를 안전하게 활용하기 위해 가명처리 및 결합을 지원하는 전문 기관이다. 이러한 센터들은 가명처리 인프라 제공, 가명처리 솔루션 제공, 가명처리 적정성 검토 및 컨설팅, 가명처리 관련 교육 및 실습 등을 통해 데이터를 안전하고 유용하게 활용할 수 있도록 지원한다.

- 서울, 강원, 부산, 인천, 대전, 대구, 전북 등 여러 지역에 가명정보 활용 지원센터가 운영되고 있으며, 각 센터는 지역 내 데이터 중소기업 및 새싹기업 등을 대상으로 가명처리 및 데이터 결합을 위한 상담, 기술지원, 자문(컨설팅), 가명처리 실습을 위한 시설 및 시스템 제공, 가명처리 적정성 검토 지원, 가명처리 및 가명정보 결합 등 활용 교육 등을 제공한다.

[표 12] 비식별화 전문기관

구분	데이터 결합전문기관	가명정보 활용지원센터
주요 역할	가명정보 간 안전한 데이터 결합	기관이 가명처리를 효과적으로 수행할 수 있도록 지원
주요 대상	가명정보를 보유한 기관(공공·민간)	가명정보 활용이 필요한 기업·연구소·기관 등
주요 기능	데이터 결합, 반출심사, 폐쇄망 분석 지원	가명처리 방법 컨설팅, 솔루션 제공, 교육 및 실습 지원
운영 주체	개인정보보호위원회, 관계 중앙행정기관	과학기술정보통신부, 개인정보보호위원회 등
예시 기관	국민건강보험공단, 통계청, 한국지능정보사회진흥원(NIA)	서울·부산·대구·대전 지역별 지원센터

데이터 보안성 및 투명성 확보

데이터의 보안성을 확보하기 위해서는 데이터 항목별로 등급을 정하고(공개/비공개), 필요시 비식별화 조치 및 접근 권한 부여 등을 검토해야 하며, 이를 위한 관리 절차를 투명하게 해야한다.

검토 필요사항

Key Question : 데이터 보안등급 설정 및 관리 절차를 투명하게 하였는가?

데이터 보안성

- 데이터 공개/비공개 등급 기준 설정 및 등급별 접근자의 권한 설정·사용자 진위확인 등으로 기밀성(인가된 사람에게만 공개) 확보
- 원래의 데이터 내용을 훼손하여 활용하는 것을 금지하여 무결성(데이터가 변형 없이 전달) 확보
- 가용성(필요시 원하는 데이터에 접근 또는 사용 가능) 확보

데이터 관리의 투명성

- 데이터 보안을 위해 데이터 관리 명세서 작성

2.3.2 데이터 변환

검토 필요사항

Key Quuestion : 데이터 변환

코드 데이터가 정의된 코드값에 맞게 변환되었나?

- 코드 데이터(코드값, Category, Enum 등)는 특정한 규칙에 맞게 변환되어야 하며, 정의된 코드값 목록과 비교하여 일치하는지 검증하는 과정이 필요. 이를 통해 데이터 정합성을 유지하고, 분석 오류를 방지함

  ex. 코드값이 기준목록과 일치하는지 확인, 정의되지 않은 코드값 탐색 및 제거, 코드값 변환이 기준목록과 올바르게 매핑되었는지 확인, 잘못된 코드값을 변환하거나 오류 데이터 처리, 코드값 표준화 및 유지보수 계획 수립 등

2.3.3 탐색적 데이터 분석(Exploratory Data Analysis, EDA)

EDA는 데이터를 이해하고 특성을 파악하기 위해 수행하는 과정으로서, 이를 통해 데이터의 분포, 패턴, 이상치, 결측치, 상관관계 등을 확인할 수 있다.

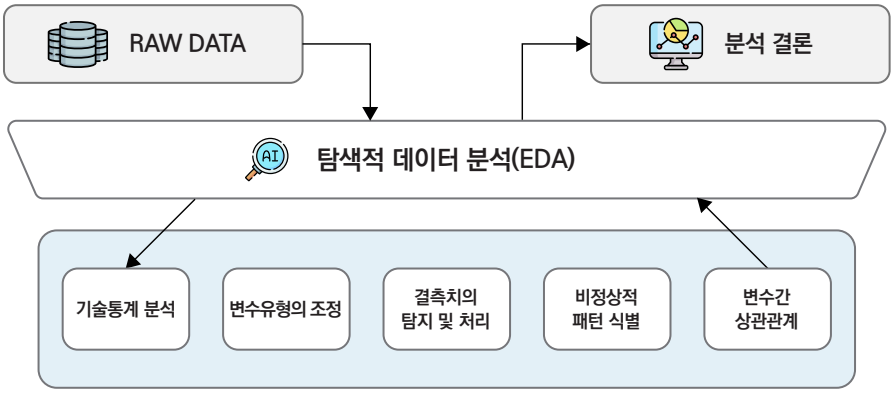
EDA는 데이터 분석에서 단순한 사전 점검이 아니라, 데이터의 본질을 이해하고 분석 방향을 설정하는 핵심 과정이다.

- 데이터의 전반적인 분포, 패턴, 이상치 등을 파악하여 분석 방향을 설정하는 과정이다.
- EDA는 단순한 데이터 확인을 넘어, 데이터의 특성을 심층적으로 이해하고, 모델링을 위한 주요 변수를 선정하며, 데이터 전처리 방향을 설정하는 중요한 과정이다.

EDA 과정에서 데이터 품질을 검토하고, 변수 간 관계를 분석하며, 모델링에 적합한 데이터 변환 방법을 설정해야 한다.

분석과제 진행 시, EDA 결과를 기반으로 분석 기획을 구체화하고, 도메인 전문가의 자문을 통해 분석모델의 타당성을 검증하는 과정이 필수적이다.

[그림 11] 탐색적 데이터 분석(Exploratory Data Analysis)



EDA의 주요 목적

- 데이터의 전반적인 특징 탐색(기초 통계 분석, 시각화)
- 데이터 정제 및 전처리 필요성 확인(결측치, 이상치 탐색)
- 데이터 패턴 및 관계 분석(변수 간 상관관계, 트렌드 파악)
- 모델링을 위한 데이터 준비

EDA 분석을 하는 이유

EDA가 철저히 수행될수록 데이터 분석의 신뢰성과 예측 모델의 성능이 향상될 수 있다.

- 따라서, 분석 수행 전에 충분한 EDA를 수행하고, 이를 바탕으로 최적의 분석 전략을 수립하는 것이 바람직하다.

EDA는 분석 모델링 전에 반드시 수행해야 하는 단계이며, 이 과정이 제대로 이루어지지 않으면 부적절한 모델이 도출되거나, 결과 해석이 왜곡될 위험이 높아진다.

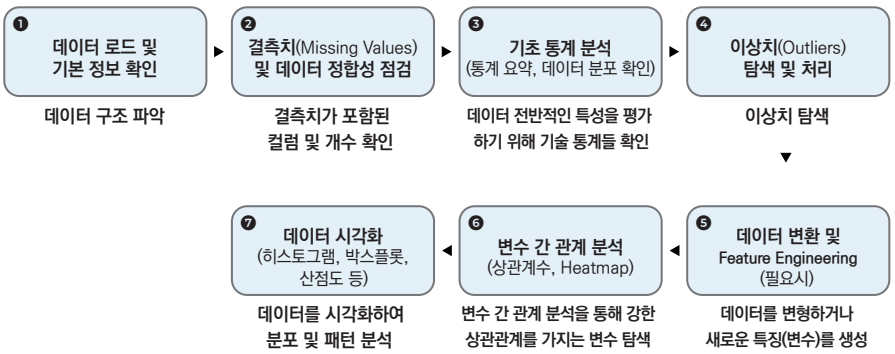
- **데이터의 품질을 사전 점검:** 결측치, 이상치, 중복 데이터 등을 확인하여 데이터 정제의 필요성을 파악한다.
- **변수 간 관계 파악:** 예를 들어 회귀분석 수행 시, 변수 간 상관관계를 분석하여, 모델링에서 중요한 변수를 선별하고 **다중공선성(Multicollinearity)*** 문제를 사전에 식별한다.
  * **다중공선성:** 회귀분석에서 독립변수들끼리 강한 상관관계가 있어 계수 추정이 불안정해지는 현상임. 이로 인해 해석이 어렵고 예측 정확도가 떨어질 수 있음
- **분석 방향 설정:** 데이터의 분포를 확인하고, 목표 변수(Target Variable, Y값)의 경향성을 분석하여 적절한 알고리즘과 파라미터를 결정하는 데 도움을 준다.
- **비즈니스 도메인과 연결:** 분석 결과가 도메인 전문가의 기대와 일치하는지 검토하고, 필요하면 분석 방향을 수정한다.

EDA 수행 과정

EDA는 데이터를 다양한 각도에서 탐색하고 시각화하여 패턴을 발견하며 데이터의 품질을 평가하는 과정으로서, 모델링 및 분석을 위한 필수적인 단계이며, 데이터를 정제하고 이해하는 데 중요한 역할을 한다.

- **데이터 로드 및 기본 정보 확인:** 데이터의 전체적인 구조(행, 열 개수 등), 변수의 종류와 데이터 타입 등 기본적인 정보 파악
- **데이터 품질 점검:** 결측치 존재 여부와 분포 파악, 데이터 타입 오류나 일관성 등 품질 문제 점검
- **기초 통계 분석(통계 요약, 데이터 분포 확인):** 기술 통계량(평균, 중앙값, 표준편차 등)을 파악하여 데이터의 분포, 경향성, 산포도 등을 확인
- **이상치(Outlier) 탐색 및 처리:** 일반적인 데이터 분포에서 크게 벗어나는 값을 탐색
- **데이터 시각화(히스토그램, 박스플롯 등):** 데이터를 시각화하여 분포 및 패턴 분석
- **변수 간 관계 분석(상관계수, Heatmap):** 변수 간 관계 분석을 통해 강한 상관관계를 가지는 변수 탐색
- **데이터 변환 및 Feature Engineering(필요시):** 모델 성능 향상을 위해 변수 재조합, 스케일링, 인코딩 등 데이터를 변형하거나 새로운 특징(변수)을 생성 (로그 변환, 원-핫 인코딩, 파생변수 생성, 정규화/표준화 등)

[그림 12] EDA 수행 과정



EDA 결과에서 도출되어야 할 사항

분석 담당자가 통계 및 차트 분석을 통해 데이터 분포와 경향성을 파악하여 모델링의 투입 변수 후보를 예상할 수 있어야 한다. 구체적으로 아래와 같은 사항들에 대해 EDA를 수행하여 결정을 내릴 수 있어야 한다.

데이터 품질 점검 결과

- 결측치 처리 방안(삭제, 대체 등)
- 이상치 처리 방안(조정, 변환, 제거 등)

데이터 변환 필요 여부 결정

- **로그 변환(Log Transformation)***, **표준화(Standardization)***, **정규화(Normalization)*** 등 필요 여부 판단

  ***로그 변환:** 데이터의 크기 차이가 너무 클 때 숫자를 줄여서 분포를 고르게 만드는 방법

  ***표준화:** 평균은 0, 표준편차는 1로 만들어서 데이터의 단위를 맞추는 방법

  ***정규화:** 데이터 값을 0과 1 사이의 범위로 바꿔서 비율로 비교할 수 있게 하는 방법

중요 변수 선별

- 모델링에 활용할 주요 변수(Feature Selection) 선정
- 상관 분석을 통한 다중공선성 변수 제거 여부 판단

분석 모델링 방향성 설정

- 목표 변수(Y값)와 독립 변수(X값)의 관계성 확인
- 모델링에서 사용할 적절한 알고리즘 탐색 (선형 회귀 vs. 비선형 모델 등)

  ***목표 변수:** 예측하고자 하는 대상 (ex. 주택 가격)

  ***독립 변수:** 목표변수를 예측하기 위해 필요한 Feature 값 (ex. 주택의 면적, 방 개수 등)

추가 데이터 확보 필요성 검토

- 내부 데이터로 충분한지, 외부 데이터 연계가 필요한지 평가

EDA 수행 시 고려해야 할 사항

검토 필요사항

Key Question : EDA 수행 시 고려사항
비즈니스 목표와의 일치 여부 • **데이터 분석 방향이 기관의 정책 목표와 부합하는지 점검해야 함** ex. 금융 부문의 대출 부실 예측 모델을 구축한다면, 연체율과 관련된 주요 변수를 확인
데이터의 대표성 확인 • **특정 기간 또는 특정 집단에 편향(Bias)이 있는지 확인해야 함** ex. 특정 연령대에만 집중된 데이터라면 일반화가 어려울 수 있음
데이터 변환 필요성 고려 • **데이터가 정규 분포를 따르지 않을 경우, 변환이 필요한지 판단** ex. 로그 변환 등
EDA 결과에 대해 도메인 전문가 자문 등을 통해 합리적인지 자문 수행 • **분석 결과가 실제 정책 결정에서 활용 가능한지 전문가 검토 필요** ex. 특정 변수가 중요할 것이라고 판단했지만, 전문가가 볼 때 연관성이 낮다면 모델링 방향을 조정해야 함

데이터 전처리와 EDA의 차이점

데이터 분석은 **'데이터 전처리 ▸ EDA ▸ 데이터 최적화'**의 흐름으로 진행하는 것이 효과적이다.

데이터 전처리(Preprocessing)는 데이터 품질 확보에 초점을 두지만(결측치 처리, 정제, 변환 등) EDA는 분석을 위한 데이터 탐색이 목적이며, 데이터를 어떻게 활용할지 방향을 설정하는 것으로 차이가 있다.

데이터 전처리(Preprocessing) – **전반적인 품질 관점**

- 원천 데이터의 결측치, 중복 데이터, 이상치 등 데이터 품질 점검
- 데이터의 형식 변환(숫자형, 범주형 등) 및 데이터 정제(클리닝)
- 변수 간 데이터 정합성 체크 (ex. 날짜 형식 통일, 단위 변환 등)
- 분석에 사용할 데이터셋을 정리하고 기본적인 품질을 보장하는 단계

EDA(탐색적 데이터 분석) – **분석 방향 및 데이터 현황 파악**

- 주요 변수의 평균, 분포, 상관관계, 퍼센트 값 등을 확인하여 데이터 특성을 파악
- 변수 간 관계(상관행렬, 다중공선성 등)를 분석하여 어떤 변수를 활용할지 결정
- 이상치가 분석 결과에 미치는 영향을 고려하여 필요한 데이터 변환(로그 변환, 정규화 등) 여부 판단
- 분석 기획과 연계하여 모델링 방향성 설정 (ex. 선형 vs. 비선형 관계 탐색)

데이터 최적화 – 분석모델 적용을 위한 추가 변환

- EDA를 통해 분석에 적절한 데이터 변환 방법 결정 (ex. 스케일링, 로그 변환 등)
- 분석모델이 올바르게 작동하도록 필요한 데이터 조정 및 전처리 수행
- 분석용 데이터셋 최종 정리

Box Tip

데이터 전처리와 EDA의 핵심 차이점 요약

데이터 전처리(Preprocessing)는 데이터 품질 확보에 초점을 둠(결측치 처리, 정제, 변환 등) EDA는 분석을 위한 데이터 탐색이 목적이며, 데이터를 어떻게 활용할지 방향을 설정함.

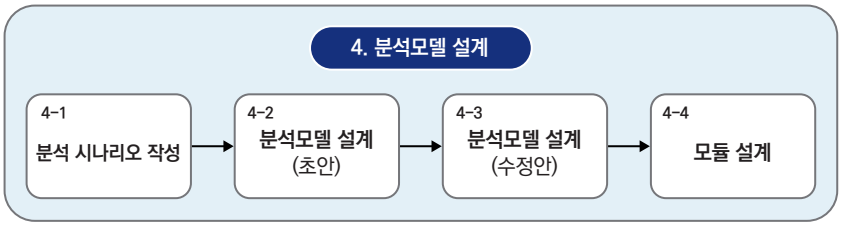
즉, EDA 결과를 바탕으로 추가적인 전처리(변환, 스케일링 등)를 수행하여 모델에 최적화된 데이터셋을 구축하는 것이 최종 목표.

따라서, **데이터 전처리 ▸ EDA ▸ 데이터 최적화**의 흐름으로 데이터 분석과제를 진행하는 것이 적절한 접근 방식이 됨

2.4 분석모델 설계

분석모델 설계는 데이터 분석의 목적을 달성하기 위해 분석 과정과 절차, 구조를 구체적으로 계획하고 체계화하는 작업을 의미한다. 분석모델 설계는 데이터 분석의 목표와 절차를 구체화하고 실행 가능성을 높이기 위해 반드시 필요한 단계로, 먼저 분석 시나리오 작성을 통해 분석의 목적, 범위, 문제 정의 등을 명확히 하고, 이를 토대로 분석모델 설계 초안을 작성하여 초기 방향성과 기본 구조를 도출한다. 이후 검토와 피드백을 거쳐 분석모델 설계수정안을 마련하고, 최종적으로 모듈 설계를 통해 기능별 모듈과 시스템 구조를 구체화한다.

[그림 13] 분석 모델링 단계의 주요 활동



2.4.1 분석 시나리오 작성

분석 시나리오는 데이터 분석 프로젝트에서 분석의 전체 흐름과 방향성을 사전에 계획하는 단계로, 분석 목표를 달성하기 위한 구체적인 절차와 방법을 기술한 문서다. 이 시나리오는 어떤 데이터를 대상으로 어떤 방법론으로 분석할지, 그리고 그 결과를 어떻게 활용할지를 한눈에 파악할 수 있도록 설계도의 역할을 한다. 분석 대상, 분석 범위, 데이터 정의, 모델 후보군, 평가 방식 등 분석 전 과정의 주요 사항을 구조적으로 정리하며, 모델 개발 단계에서의 시행착오를 줄이는 데 중요한 기능을 한다.

분석 시나리오는 분석 과정에서 방향성을 상실하거나 중간에 불필요한 반복을 방지하기 위해 필요하다. 프로젝트 초기 단계에서 목표와 절차가 명확하지 않으면 분석 결과가 요구사항과 엇나가거나 실무 적용이 어렵게 되는 경우가 발생한다. 시나리오는 이해관계자 간의 기대치를 조율하고 데이터 수집, 모델링, 검증까지의 흐름을 일관되게 이끌어준다. 특히 복잡하거나 다부서 협업이 필요한 프로젝트일수록 사전 시나리오 작성은 필수적이며, 향후 문제 발생 시 초기 기획과의 비교 기준으로도 활용된다.

분석 시나리오의 핵심 내용은 우선 분석 목표를 명확히 정의하는 것으로 시작한다. 다음으로, 분석 대상과 범위를 설정하여 분석이 적용될 데이터셋과 시간적·공간적 범위를 구체화한다. 분석에 필요한 데이터의 종류와 출처, 데이터 전처리 방안도 포함된다.
주요 분석 변수(독립변수, 종속변수)와 예비 변수(Feature)들을 정의하고, 분석 목표에 따라 군집, 분류, 회귀 등 적절한 분석 방법론과 모델 후보군을 선정한다. 마지막으로 결과 검증 및 평가 계획을 세우고, 분석 결과를 어떻게 업무에 적용할지까지 시나리오로 정리한다.

예를 들어 **'고객 이탈 예측'**을 목표로 하는 분석 시나리오는 다음과 같이 구성할 수 있다. 분석 대상은 최근 2년간 가입 고객 중 6개월 이상 거래 내역이 있는 고객으로 설정한다. 분석 범위는 전국 단위이며 월 단위 데이터를 활용한다. 내부 데이터(고객 기본 정보, 결제 이력, 상담 기록)와 외부 데이터(경기지수 등)를 결합한다. 종속변수는 이탈 여부이며, 독립변수로는 결제 빈도, 월평균 금액, 고객 등급 등을 정의한다. 분석 모델은 로지스틱 회귀와 랜덤포레스트로 선정하며, AUC를 주요 평가 지표로 삼는다. 분석 결과는 마케팅팀에 고위험 고객군으로 전달하여 맞춤형 전략을 수립하는 데 활용한다.

분석 시나리오 작성 시 가장 주의할 점은 분석 목표와 데이터의 정합성을 철저히 검토하는 것이다. 요구사항에 맞지 않거나 확보가 어려운 데이터를 전제로 시나리오를 작성하면 실행 단계에서 큰 차질이 발생할 수 있다. 또한 분석 범위를 지나치게 넓게 설정하면 데이터가 과도하게 방대해지고, 반대로 너무 좁게 잡으면 일반화 가능성이 떨어진다. 분석 방법론과 모델 선정 시에는 단순히 최신 기법만을 고려하기보다는 목표에 적합한 방법을 선택해야 한다. 검증 계획도 미리 구체화하지 않으면 분석 결과의 신뢰성을 담보하기 어렵다.
끝으로, 분석 결과가 어떻게 현업과 연결될지 실질적 적용 방안을 반드시 명확히 해야 한다.

2.4.2 모델 설계(초안, 수정안)

분석 모델링은 정책 결정, 서비스 개선, 업무 효율화 등을 위해 데이터를 기반으로 인사이트를 도출하고 예측하거나 분류하는 통계적·기계학습 기반 모델을 구축하는 과정이다.
공공부문에서는 정책 목적, 행정 수요, 데이터 특성을 고려하여 분석모형을 설계하고 최적화해야 한다.

분석설계안은 이러한 분석 모델링을 체계적으로 수행하기 위한 기획 문서로, 분석 목적에 부합하는 모델링 방향성과 방법론을 사전에 정리하고 이해관계자 간 공감대를 형성하는 데 중요한 역할을 한다. 설계안은 모델링 진행 중 중간 점검 기준이 되며, 결과 해석의 기준점도 제공한다.

분석설계안은 데이터 수집 및 정제 완료 이후, 본격적인 분석에 착수하기 직전 단계에서 마련해야 한다. 이 시점은 데이터의 품질과 범위, 결측 여부, 전처리 결과 등이 확인된 시점이며, 이를 바탕으로 현실적이고 실현 가능한 모델링 전략을 수립할 수 있다.

분석설계안은 분석 목적 및 기대효과, 분석 대상 및 범위, 사용 데이터 목록 및 특성, 분석변수 정의 및 전처리 전략, 예측·분류·군집 등 분석 방법론 선택 사유, 모델링 절차, 성능 평가지표, 한계점 및 고려사항과 같은 내용을 포함한다.

설계안은 관계자(발주기관, 현업 담당자, 데이터 분석가 등)와의 협의를 통해 초안이 검토되며, 파일럿 분석 결과 또는 모델 시범 구축 결과에 따라 현실성과 정합성을 점검하여 수정한다.

수정된 분석설계안은 내부 검토와 외부 자문(필요 시)을 거쳐 분석 모델링 단계 진입 전까지 최종 확정되어야 한다. 확정안은 분석 결과 해석과 정책적 활용 방안 도출의 기준 문서로 기능하며, 최종 결과 보고서에도 부록 형태로 포함될 수 있다.

검토 필요사항

Key Question : 분석모델 설계

[분석모델 설계(초안)]

필요한 데이터 항목이 정해졌는가?

- 분석모델을 설계할 때, 모델이 제대로 작동하고 신뢰할 수 있는 결과를 도출하기 위해 필요한 데이터 항목이 충분히 정의되었는지 확인해야 함

[필요한 데이터 항목이 정해졌는지 확인하는 주요 과정]

- **분석 목적 및 모델링 목표 정의:** 모델이 해결하려는 비즈니스 문제를 정의함. 목표에 따라 필요한 데이터 항목이 다를 수 있음
- 분석모델에서 예측하거나 분류하려는 변수인 종속 변수(타깃 변수, Y) 확인
- **독립변수 선정:** 통계적 상관관계, 도메인 지식 등 활용
- **데이터 품질 점검:** 결측치, 이상치, 데이터 정합성 확인
- 새로운 변수(Feature) 생성 가능성을 확인하기 위하여 Feature Engineering(특성 엔지니어링) 검토 (ex. 날짜 ▸ 요일, 월별 패턴 생성 가능)
- 모델 성능 평가를 통해 변수의 적절성 검토

데이터 항목별 표준화 방법을 정하였는가?

- 데이터를 분석하고 모델링할 때, 각 항목(변수)의 특성에 따라 적절한 표준화(Scaling) 방법을 선택하는 것이 중요
  ex. 변수의 유형(연속형, 범주형 등) 확인, 데이터의 분포 확인(정규분포 여부 판단), 변수 간 크기 차이(Scale Difference) 확인, 변수 유형에 따른 적절한 표준화 방법 선택, 표준화 적용 후 데이터의 변화 검토

데이터 수집 가능 항목에 따라, 단계별로 모델이 설계되었는가?

- 분석모델을 효과적으로 설계하려면, 현재 수집 가능한 데이터 항목에 맞춰 단계별로 모델을 구성하고, 데이터가 추가될 때마다 모델이 점진적으로 발전할 수 있도록 설계해야 함

[단계별 모델 설계 검증 주요 과정]

- 모델의 최종목표가 무엇인지 파악해야 함
- 분석 목표를 명확하게 정의하고, 단계별 모델의 역할을 설정해야 함
- **기본 모델**(Baseline Model) ▸ **중간 모델** ▸ **최종 모델**로 발전 가능성이 있는지 점검
  ex. 고객 이탈 예측 모델의 경우, **기본 모델**[기본적인 고객 정보를 기반으로 이탈 예측(**필요 데이터 항목:** 성별, 연령, 가입일)] ▸ **중간 모델**[고객의 행동 데이터를 추가하여 예측 성능 개선(**필요 데이터 항목:** 구매 횟수, 최근 방문일, 평균 구매 금액)] ▸ **최종 모델**[실시간 데이터 및 외부 데이터를 반영해 정밀한 예측(**필요 데이터 항목:** 고객 서비스 문의 이력, 실시간 웹사이트 방문 로그)]

- **단계별 모델에서 필요한 데이터가 확보되었는지 검토:** 데이터가 부족한 경우 대체 변수 사용 가능성 검토 등
- 단계별 모델 성능 비교 및 개선 가능성 분석
- 데이터 추가 시 모델 발전 계획 확인

분석 검증 통계기법을 정하였는가?

- 검증 방법이 타당한지 판단
- 분석 목적에 맞는 검증 기법을 정하였는지, 사용한 검증 기법이 적절한 통계적 가정을 충족하는지, 데이터의 특성(정규성, 독립성, 분산성 등)에 따라 검증 방법이 적절한지, 검증 방법이 데이터 크기 및 샘플링 방법과 일치하는지, 분석 결과를 해석할 때 통계적 유의성이 충분한지 등을 확인

[분석모델 설계(수정안)]

현업, 빅데이터 전문가, 실무자들로 구성된 자문위원단과 분석모델 설계를 검토하였는가?

- 자문위원단은 분석모델 검토 시 각자의 역할이 다름
- **현업 전문가(도메인 전문가):** 비즈니스 목표와 모델의 연관성 검토
  빅데이터 전문가(데이터 사이언티스트): 모델의 수학적/통계적 타당성 검토
  실무자(데이터 엔지니어, 분석가): 데이터 수집 및 기술 구현 가능성 검토
- **데이터 가용성과 활용 가능성 검토:** 데이터 품질, 추가 데이터 수집이 필요한지 등
- 모델 검증 및 성능 평가 기준 검토
- 모델의 실무 적용 가능성 및 피드백 반영

분석모델 설계(수정안)가 타당한가?

- 모델의 수정안이 기존 모델 대비 어떤 점이 개선되었는지를 확인함
- 모델에 사용된 데이터와 특징이 적절한지 검토(변수 추가/제거가 모델 성능 향상에 기여했는가 등)
- **기존 모델과 수정된 모델의 성능을 비교하여 개선이 되었는지 평가:** 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score 등
- 모델의 해석 가능성과 비즈니스 적용성 검토: 모델이 해석 가능하고, 실무에서 쉽게 활용될 수 있는지 확인
- **실무 적용 가능성과 유지보수 용이성 평가:** 모델이 안정적으로 운영될 수 있는지, 데이터 업데이트 주기와 재학습 전략이 수립되었는지 등 점검

분석모델 설계가 타당한지 자문위원회 회의를 거쳐 논의하고, 내용을 추가 반영하여 분석 모델 설계를 완성한다.

2.4.3 주요 산출물

 # 분석모델 설계서

목적변수를 정의하여 학습용 데이터를 활용해, 모델을 학습하고 평가하는 일련의 AI 모델에 대한 설계와 이를 활용한 예측 모델과 케이스에 대해 정의한 문서

- 모델 정의
- 목적변수 정의
- 학습용 데이터 정의
- 모델 학습 및 평가 방법
- 활용 모델
- 모델 개발 케이스

2.4.4 모듈 설계

분석 모델링 전에 수행하는 모듈 설계는 데이터 수집, 정제, 전처리 등 모델 개발에 필요한 작업을 체계적이고 재사용 가능한 단위로 나누어 구조화하는 과정이다. 이 단계는 모델링 이전의 데이터 준비 과정을 효율화하기 위해 수행된다.

주요 내용은 데이터 수집 모듈, 데이터 정제 및 표준화 모듈, 결측치 처리 모듈, 탐색적 데이터 분석(EDA) 모듈 등으로 구성된다. 각 모듈은 입력과 출력이 명확하고, 후속 모델링 단계에서 바로 활용할 수 있도록 설계한다. 또한, 전처리된 결과의 안전하고 효율적인 관리를 위한 데이터 저장 체계를 갖추어야 한다.

모듈 설계를 통해 데이터 준비 과정을 자동화하고 일관성을 확보할 수 있다. 특히 분석 대상이 변경되거나 데이터가 갱신될 때 빠르게 재적용할 수 있어 유지보수성이 높아진다. 팀 단위 개발에서도 역할 분담이 용이해지며, 전체 분석 품질을 일정 수준 이상으로 유지할 수 있다.

예를 들어 고객 데이터를 활용하는 프로젝트에서는 내부 DB에서 고객 데이터를 추출하는 수집 모듈, 결측값과 이상치를 처리하는 정제 모듈, 고객별 요약 지표를 만드는 피처 생성 모듈, 기본 통계와 시각화를 수행하는 EDA 모듈로 나눌 수 있다.
각 모듈은 함수나 스크립트 단위로 구현되며, 중간 산출물을 CSV나 데이터베이스에 저장해 모델링 단계에서 바로 사용할 수 있게 한다.

모듈 설계에서 주의할 점은 모듈 간 의존성을 최소화하고, 입력 및 출력 형식을 표준화해야 한다는 것이다. 초기에는 모델 결과를 연계하는 시스템적 요소보다 데이터 품질과 처리의 신뢰성을 우선시해야 하며, 오류 발생 시 빠르게 대응할 수 있는 로깅과 검증 절차도 설계에 포함해야 한다. 지나치게 복잡한 모듈은 유지보수가 어렵기 때문에 단순하고 명확한 구조를 유지하는 것이 중요하다.

검토 필요사항

Key Question : 모듈 설계

[모듈 기능 정의 및 설계]

모듈에 대한 기능을 정의하였는가?

- 기존에 존재하는 언어 사전이 분석에 활용이 가능한지 확인
- 누락된 알고리즘이 없는지 점검
- 중복되거나 누락된 모듈 기능이 없는지 검토

모듈 설계를 실시하였는가?

- H/W와 S/W등의 제반 상황을 고려하여 모듈 구축 가능성을 검토
- 분석모델 프로세스와 모듈 설계안이 일치하는지 검토

검토 필요사항

Key Question : 표준화 모듈 설계

표준화 모듈 설계를 하였는가?

- 표준화할 데이터 속성(변수) 정의 및 분석
- 데이터 표준화 규칙 설정(단위 변환, 포맷 통일, 인코딩 방식 등)
- 데이터 변환 및 정제 모듈 설계
- 표준화된 데이터 검증 및 오류 처리 로직 설계
- 데이터 표준화 모듈 자동화 및 유지보수 설계

검토 필요사항

Key Question : 데이터 정제 모듈 설계

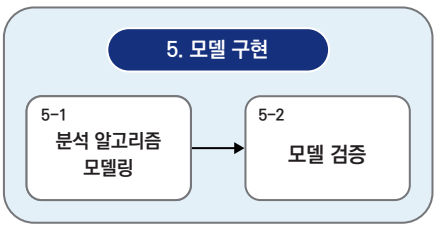
데이터 정제 모듈 설계를 하였는가?

- 데이터 정제(Data Cleansing) 모듈은 데이터 품질을 향상시키기 위해 이상치, 중복, 결측치, 오류 데이터를 탐지하고 수정하는 과정을 자동화하게 됨
- 데이터 정제 요구사항 분석 및 주요 정제 항목 정의
- **정제 로직 설계:** 결측치 처리, 중복 제거, 이상치 탐지 및 수정 등
- **데이터 정제 모듈 개발:** Python 기반 코드 구현
- 데이터 정제 자동화 및 검증 프로세스 설계
- **성능 최적화 및 유지보수 전략 수립:** 대량 데이터를 효율적으로 처리할 수 있도록 최적화

2.5 모델 구현

모델 구현은 분석의 목표를 실제로 이루는 단계로, 미리 설계한 분석모델을 실제 데이터에 적용해서 가장 좋은 결과를 얻도록 하는 과정이다.

[그림 14] 모델구현 단계의 주요활동



2.5.1 분석 알고리즘 모델링

분석 모델링이란?

분석 모델링(Analytical Modeling)은 데이터를 기반으로 패턴을 발견하고, 인사이트를 도출하며, 이를 바탕으로 예측하거나 최적의 의사결정을 내리기 위한 수학적/통계적 모델을 설계하고 구현하는 과정을 의미한다.

분석 모델링 단계에서는 이전 단계에서 마련된 분석설계안을 바탕으로 실제 데이터를 활용하여 분석 알고리즘을 적용하고, 예측·분류·군집 등의 모델을 구현한다.
이 과정은 단순히 알고리즘을 선택하는 것을 넘어서, 업무 목적과 데이터 특성에 맞는 분석 전략을 세우고 모델을 반복적으로 실험·검증하여 최적화하는 작업이 포함된다.

분석 모델링에는 통계적 기법뿐 아니라 기계학습 기반의 알고리즘도 활용되며, 각각의 모델은 데이터 구조, 분석 목적, 예측 정확도, 해석 가능성 등에 따라 장단점이 다르기 때문에 적절한 판단이 필요하다.
이 과정에는 통계적 이해뿐 아니라 실제 모델링 경험과 시행착오를 바탕으로 한 엔지니어의 노하우가 요구되며, 경험이 풍부한 분석 전문가나 데이터 과학자의 참여가 핵심적이다.

특히 공공데이터는 결측치 처리, 행정 데이터 간 연계 문제 등 데이터 품질 이슈가 많아, 모델링 이전에 철저한 데이터 점검과 전처리 전략 수립이 필요하다. 모델 구현은 단일 모델로 끝나는 것이 아니라, 여러 모델을 비교 평가하여 가장 효과적인 모델을 선정하고, 필요시 앙상블 등 복합 전략도 적용하게 된다.

공공기관 실무자 입장에서는 분석 모델링의 기술적 세부사항까지 숙지할 필요는 없지만, 어떤 모델이 왜 선택되었고, 그 모델이 실제 정책 또는 서비스에 어떤 기여를 할 수 있는지를 이해하는 것이 중요하다. 이를 통해 분석 결과의 해석 및 활용에 대한 실무 판단력이 높아지고, 기관의 데이터 기반 의사결정 역량도 강화될 수 있다.

검토 필요사항

Key Question : 분석 알고리즘 모델링

[분석 알고리즘 모델링]

분석모델에 대하여 조사하였는가?

- **분석모델에서 사용한 분석기법이 알고리즘화가 가능한지 점검:** 분석모델에서 사용한 분석기법이 알고리즘화(Automation & Systemization) 가능한지 점검하는 과정으로서, 해당 기법을 코드로 구현하고, 자동화 및 시스템에 적용할 수 있는지 평가
- **이를 검증하기 위해 분석기법의 수식화 가능 여부 점검:** 분석기법을 수식으로 표현하고, 코드로 변환할 수 있는지 점검
- **데이터 흐름의 자동화 가능성 점검:** 데이터 입력 ▸ 출력 ▸ 프로세스를 자동화할 수 있는지
- **연산 효율성 점검:** 데이터 크기가 커질수록 연산 시간이 지나치게 증가하는 모델은 알고리즘화가 어려우므로, 분석기법이 대량의 데이터를 처리할 수 있도록 연산 효율성이 높은가를 검토
- 실시간 또는 배치 처리를 통해 운영 가능한지, 모델 배포 및 유지보수 용이성이 확보되었는지 등 점검

분석 모델링과 AI/ML 활용방안

분석 모델링에서는 주로 머신러닝(ML) 알고리즘을 사용한다.
ML 알고리즘은 AI 시스템이 작업을 수행하기 위해 사용하는 일련의 규칙 또는 프로세스로, 주로 새로운 데이터 인사이트와 패턴을 발견하거나 주어진 입력 변수 세트에서 출력 값을 예측하기 위해 사용된다.

- ML은 AI의 여러 분야 중 하나이다. 컴퓨터 시스템은 ML 알고리즘을 사용하여 대량의 기록 데이터를 처리하고 데이터 패턴을 식별한다. ML은 AI이지만 모든 AI 활동을 ML이라고 할 수는 없다.
- 인공지능(AI)은 기계를 더 인간처럼 만드는 데 사용되는 다양한 전략과 기술을 포괄한다. AI에는 Alexa와 같은 스마트 비서부터 로봇 청소기 및 자율 주행 자동차에 이르기까지 모든 것이 포함된다.

분석 모델링의 핵심 가치

EDA와 분석 모델링은 서로 보완적인 역할을 한다. 즉, EDA는 분석 모델링을 위한 출발점이고, 분석 모델링은 데이터에서 발견한 패턴을 실질적인 의사결정에 활용하는 과정으로 볼 수 있다. EDA의 패턴 분석 결과는 분석 모델링으로 확장하여 사용할 수 있다.

#예시

❶ EDA에서 안전에 대한 민원 데이터를 3가지 그룹으로 나눌 수 있는 패턴을 발견하였다면 분석 모델링에서 K-Means 클러스터링을 적용하여 그룹화한다.

❷ EDA에서 특정 고객이 구매율이 높은 패턴을 발견하였다면 분석 모델링에서 머신러닝을 활용한 고객 분류모델을 구축한다.

따라서 분석 모델링은 데이터를 단순히 분석하는 것을 넘어서, 실질적인 문제 해결과 업무 최적화 및 자동화 하는 것이 핵심 과정이다.

모델 선정

분석 모델링을 위해 데이터의 속성과 분석 목적에 맞게 종속 및 독립변수, ML 알고리즘을 먼저 선정해야 한다.

종속변수 유무에 따라 알고리즘을 결정할 경우, 종속변수가 존재하면, 즉 예측하거나 분류할 대상이 명확하다면 지도학습의 분류/회귀 알고리즘을 사용하고 존재하지 않는다면 비지도학습의 군집/차원축소/연관성 알고리즘을 사용한다.

데이터 유형에 따라 분석모델을 결정할 경우, 종속변수가 범주형이면 분류 알고리즘, 연속형이면 회귀 알고리즘을 주로 사용한다.

검토 필요사항

Key Question : ML 알고리즘 적용을 위한 최적의 분석모델을 선정하였는가?

분석주제

- 데이터로 나타낼 수 있는 구체적인 분석 주제, 즉 가설을 설정

모델유형

- 학습유형 또는 데이터 유형에 따른 최적의 분석모델 선정

Y변수(종속변수) 정의

- 데이터를 통해 얻고자 하는 목표를 구체적으로 설정

X변수(독립변수) 정의

- 목표실행을 위해 필요한 설명 가능한 요인을 구체적으로 설정

데이터범위 정의

데이터분리

- 전체 데이터를 학습용(Train), 검증용(Validation), 테스트용(Test) 데이터로 분리
- **일반적인 분할 비율:**
  . Train : Validation : Test = 64 : 16 : 20 (Train 80% 내에서 Validation 20% 즉, 전체 데이터에서 Train용으로 먼저 80%를 확보하고, 그중 20%를 Validation으로 배정)
- **데이터 분리 목적:**
  .Train: 모델 학습용 데이터
  . Validation: 하이퍼파라미터 튜닝 및 검증용 데이터
  . Test: 최종 모델의 일반화 성능을 평가

분석 알고리즘 모델을 설계하였는가?

- **분석모델을 기반으로 알고리즘을 설계하였는지를 확인:** 분석모델을 기반으로 알고리즘을 설계하는 과정은 이론적으로 수립한 분석모델을 실제 코드 및 시스템으로 구현할 수 있도록 구조화하는 것
- 분석모델이 알고리즘으로 변환될 수 있도록 정의되었는지, 알고리즘 설계에 필요한 데이터 입력 ▸ 처리 ▸ 출력 프로세스가 정의되었는지, 알고리즘이 분석모델의 수학적 가정을 충족하는지, 알고리즘이 실행 가능하고, 성능 평가를 통해 검증되었는지, 알고리즘이 실무에서 운영 가능하도록 최적화되었는지 등을 평가

분석 결과가 최선인지에 대한 근거를 마련하는 것이 중요

데이터 분석모델을 개발할 때, 단순히 모델링과 구현에 집중하는 것이 아니라, 선정한 알고리즘이 최선의 선택인지에 대한 근거를 명확하게 제시하는 과정이 필수적이다.

특히, 알고리즘 선정의 타당성을 확보하기 위해서는 도메인 지식, 기존 연구 검토, 다양한 알고리즘 비교 실험 등의 단계가 논리적으로 진행되어야 한다.

먼저, 도메인에 대한 기존 연구를 조사하여 알고리즘의 적합성을 검토하는 과정이 필요하다.

- 특정 주제에 대한 연구 논문을 검색하여 해당 문제를 해결하기 위해 주로 사용된 알고리즘을 확인하고, 각 연구에서 사용한 변수들이 설명력이 충분한지를 검토해야 한다.
- 이 과정에서 단순히 연구 결과를 참조하는 것이 아니라, 연구에서 사용한 데이터 특성과 분석 목적이 본 과제와 얼마나 유사한지를 판단하여 알고리즘의 적합성을 검증해야 한다.
- 이를 통해 이론적으로 알고리즘이 해당 문제를 해결하는 데 타당한 근거를 갖는지 확인하는 것이 중요하다.

다음으로, 다양한 알고리즘을 테스트하고 비교 분석하는 과정이 필요하다.

- 특정 알고리즘이 많이 사용되었다는 이유만으로 채택하는 것이 아니라, 여러 알고리즘을 동일한 데이터에 적용하여 비교하는 실험적 접근이 필수적이다.
- 만약 기존 연구에서 사용되지 않은 알고리즘을 선택했다면, 기존 방법과 비교하여 그 알고리즘이 왜 더 적합한지에 대한 설명이 필요하다. 예를 들어, 다른 연구에서는 결정 트리(Decision Tree)나 서포트 벡터 머신(SVM)을 사용했지만, 본 연구에서는 신경망(DNN)을 활용했다면 신경망이 해당 데이터에서 더 높은 설명력을 가지는 이유를 명확히 제시해야 한다.

마지막으로, 가장 설명력이 높은 모델을 실험 근거를 바탕으로 선정하는 과정이 필요하다.

- 여러 모델을 비교한 후, 정확도, 설명력, 실행 속도 등의 측면에서 최적의 모델을 선택하고, 그 근거를 정리해야 한다.
- 이를 위해, 다양한 모델의 성능을 비교하는 표를 활용하면 보다 명확한 근거를 제시할 수 있다. 예를 들어, 특정 데이터에서 신경망 모델(DNN)이 99%의 정확도를 보이고, 기존 연구에서 사용된 알고리즘보다 설명력이 높다면, 해당 모델이 최적의 선택이라는 것을 논리적으로 입증할 수 있다.

결과적으로, 데이터 분석모델을 개발할 때는 기존 연구를 기반으로 알고리즘의 적합성을 검토하고, 다양한 알고리즘을 비교 실험한 후, 가장 설명력이 높은 모델을 근거를 바탕으로 선정하는 과정이 필요하다. 이러한 과정이 체계적으로 수행될 때, 모델의 신뢰성과 활용 가능성이 더욱 높아진다.

2.5.2 모델 검증

모델 검증은 데이터 분석을 위한 과정의 일환으로 개발한 모델이 실제 분석 업무에 사용 가능한지를 평가한다. 별도로 분리해 놓은 검증 데이터셋을 통해 모델을 검증하고, 객관적인 평가지표를 통해 업무에 사용 가능한지를 평가하고, 이것이 기존 운영체계와 연계 및 통합 가능한지를 점검하여 지속적으로 모델을 개선해야 한다.

모델의 적합성을 검증하기 위해서는 생성한 모델에 오류가 없는지, 목표하고자 하는 예측값(Y값) 및 검증모델이 누구에게나 설명 가능한 합리성을 갖춘 것인지 검토가 필요한다.

모델 검증의 유효성

모델 검증의 유효성을 확보하기 위해서는 별도의 데이터셋을 통해 모델을 검증해야 한다. 이때 합리적인 평가지표를 선정해 선정 사유와 검증 결과 제시 및 향후 모델의 성능 평가를 위한 주요 평가지표를 제시해야 한다.

검토 필요사항

Key Question : 모델 검증 및 효과평가 방법이 제시되어 있는가?
모델 검증 • 분리한 검증 데이터셋을 이용해 예측값(Y값) 산출 및 샘플 제시 • 기대하는 목표서비스 구현을 위해 Y값으로 설명 가능한지 검토
모델 수정 • 필요시 X값, Y값의 추가·삭제·조작적 정의 변경 • 변경 전후의 차이점 비교, 결과 기록
전문가 검토 • 최적의 모델을 잘 선정했는지, 예측 및 검증보완 방법 등에 대한 전문가 의견 필요

모델 성능 평가 객관성

학습데이터에서 높은 정확도를 보이지만 테스트 데이터나 새로운 데이터에서는 부정확한 모델이 만들어지는 경우(과대적합), 모델이 단순하여 데이터 내부의 패턴이나 규칙을 잘 학습하지 못하는 경우(과소적합) 등이 종종 존재하기 때문에 학습, 평가, 검증 데이터에 대해 모두 정확하게 예측하는 모델, 즉 일반화된 모델이 잘 선정되었는지 성능을 평가하는 것은 중요하다.

모델의 성능을 평가하기 위해서는 과대·과소평가 된 것은 아닌지, 생성한 모델이 이전 모형보다 더 우수한 성과를 보이는지, 고려된 모형들 중 어느 것이 가장 우수한가를 검토하여야 하며, 이는 **[모델 평가지표]**를 통해 평가할 수 있다.
모델 평가지표란 모델의 예측 정확도, 오류율, 설명력 등을 수치로 나타내어 모델의 성능을 객관적으로 판단할 수 있게 해주는 지표를 말한다.

모델 성능 평가의 객관성을 확보하기 위해서는 모델 및 검증에 사용하지 않은 별도의 데이터를 활용해 모델 성능 평가를 점검하고, 모델을 이용한 시스템을 구현하고자 할 때는 사후 컨설팅을 받아야 한다.

검토 필요사항

Key Question : 모델 성능 평가를 객관적으로 2차 검증하였는가?

모델 성능평가 기준 선정

- 모델의 성능 평가 기준은 모델의 일반화 가능성, 효율성, 예측 및 분류 정확성으로, 이를 평가하기 위한 모델별 평가지표 확인

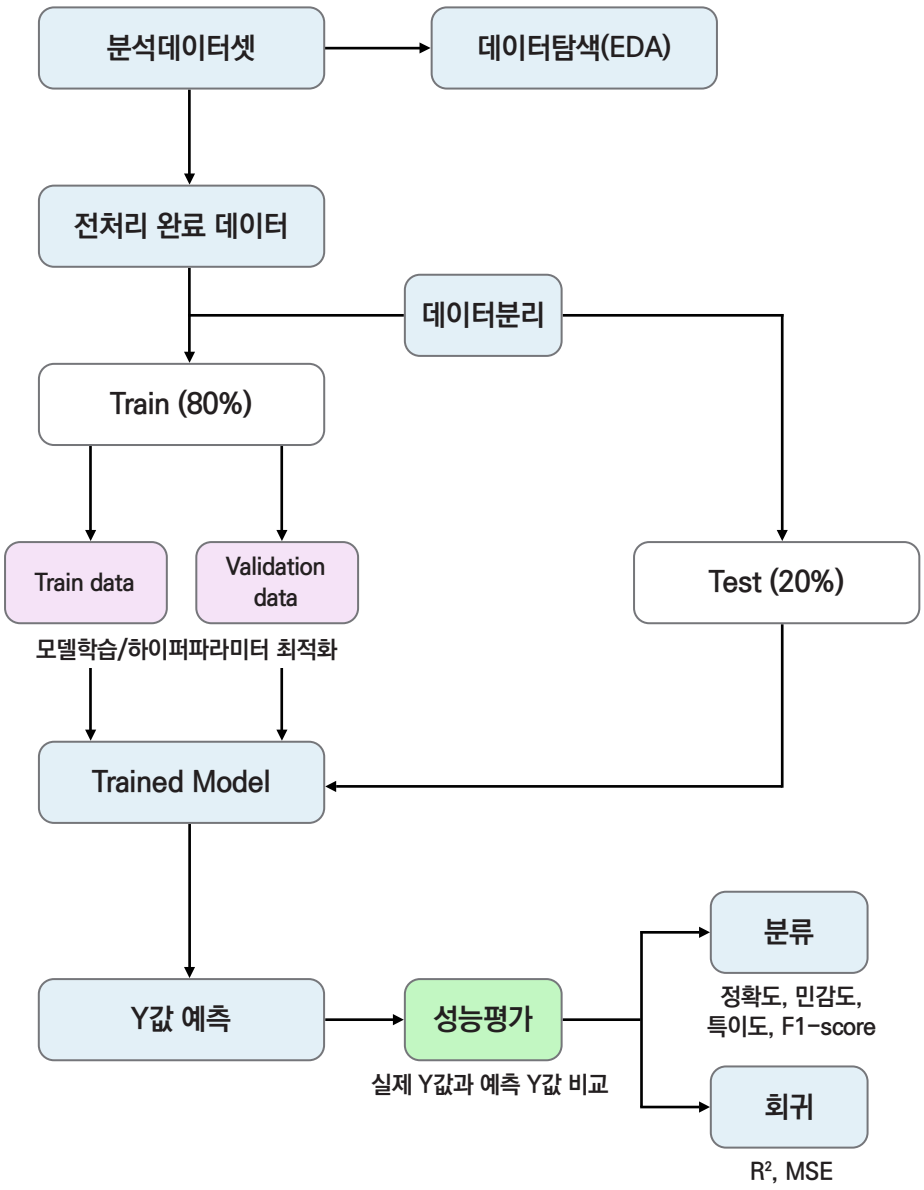
분석 결과 검토

- 평가 기준에 따른 결과값, 실질적인 활용 가능성에 대한 검토
- 알고리즘별로 파라미터를 변경하며 수행-변경 전후의 차이점 비교, 결과, 기록

모델 개선

- 제한된 학습 데이터셋에 지나치게 특화되어 새로운 데이터에 대한 오차가 커지는 과대적합 문제 해결(일반화 부족)
  - ▸ **과대적합 방지방법:** 데이터 증강(데이터 변형을 통한 데이터의 양 증가), 모델복잡도 감소, 가중치 규제 적용
- 지나치게 단순한 모델이거나 데이터에 내재된 구조를 학습하지 못해서 생기는 과소적합 문제 해결과 데이터 보안을 위해 데이터 관리 명세서 작성

분석절차


분석데이터셋
데이터탐색(EDA)
전처리 완료 데이터
데이터분리
Train (80%)
Train data
Validation data
Test (20%)
모델학습/하이퍼파라미터 최적화
Trained Model
Y값 예측
성능평가
실제 Y값과 예측 Y값 비교
분류
정확도, 민감도, 특이도, F1-score
회귀
R^2, MSE

모델 성능 평가지표

모델종류	평가지표	비고
분류모델	**혼동행렬(Confusion Metrix):** **모델이 분류한 예측범주와 실제 분류범주를 교차표로 정리한 행렬** • 정확도: 실제 분류를 정확하게 예측한 비율 • 정밀도: 모델이 '참'으로 예측한 결과 중 실제로 '참'인 비율 • 민감도(재현율): 실제 '참'을 '참'으로 예측한 비율 • 특이도: 실제 '거짓'을 '거짓'으로 예측한 비율 • 거짓긍정률: 거짓긍정률 = 1 – 특이도 • F1score: 정밀도와 민감도(재현율)의 조화평균 • 오차비율: 오차비율 = 1 – 정확도 • 카파통계량: 두 관찰자가 측정한 범주값의 일치도를 측정	F1-score = $2 \times \frac{\text{재현율} \times \text{정밀도}}{\text{재현율} + \text{정밀도}}$
	ROC Curve • 가로축을 거짓긍정률(1-특이도), 세로축을 민감도로 한 곡선. • AUC(Area Under ROC): ROC곡선 아래 면적	.ROC 곡선이 왼쪽 꼭대기에 가까울수록 분류성능 우수 .AUC는 0.5 이상이어야 좋음, 1에 가까울수록 모델 정확도 높음
	이익도표(Gain Chart) • 목표범주에 속한 개체들이 임의로 나눈 등급별로 얼마나 분포하고 있는지 나타내는 값 • 이익 도표 = 이익 곡선 = 리프트 곡선	
	IOU(Intersection Over Union) • 객체 검증의 정확도를 평가하는 지표로 일반적으로 객체탐지에서 개별 객체에 대한 검출이 성공하였는지로 결정.	
	mAP(mean Average Precision) • 객체 검출 또는 이미지 분류에서 성능을 평가하는 대표적 지표. 여러 클래스 또는 여러 임계값에 대한 평균 정밀도를 계산하여 모델의 전반적인 성능을 측정 • P-R Curve(Precision-Recall Curve)의 아래 면적(AUC)을 계산하여 각 클래스의 AP를 계산 후 평균하여 계산	0~1 사이의 값
회귀모델	**실제값,예측값,평균값 : 실제값,예측값,평균값 기본적으로 검토**	
	오차제곱합(SSE): 예측값과 실제값의 차이(오차) 제곱합 • 전체제곱합(SST): 실제값과 평균값의 차이 제곱합 • 회귀제곱합(SSR): 예측값과 평균값의 차이 제곱합 • 기타: 평균오차(AE), 평균절대오차(MAE), 평균 제곱근오차(RMSE), 평균 절대 백분율 오차(MAPE), 평균백분율 오차(MPE) 등	
	결정계수(R^2) • R^2 = SSR/SST으로 회귀모형이 실제값을 얼마나 잘 나타내는지에 대한 비율 • 모형의 변수 개수가 증가할 때 그 변수가 유의하지 않더라도 결정계수는 증가하기 때문에 독립변수 개수가 많은 모형의 경우 부적합	0~1 사이의 값

모델종류	평가지표	비고
회귀모델	**조정된 결정계수(R^2adj)** • 결정계수 R의 단점을 보완한 지표. 독립변수의 개수를 고려하여 결정계수 값 조정 • 유의하지 않은 독립변수 추가하면 패널티부과, 모형이 유용한 독립변수를 추가하면 계수 증가	0~1 사이의 값 조정된 결정계수는 결정계수보다 항상 작음
	Mallow's Cp • 적절하지 않은 독립변수 추가에 대한 패널티를 부과한 통계량 • 값이 작을수록, 실젯값을 잘 설명하는 모형임	
언어모델	**PPL(Perplexity)** • PPL은 당혹감, 혼란의 의미로 모델이 정답을 결정할 때 얼마나 헷갈렸는가를 나타내는 지표로 문장이 완성될 때까지 선택된 토큰들의 누적된 확률을 기반으로 계산한 값 • 값이 낮을수록 모델이 덜 헷갈린 상태로 확신을 가지고 답을 냈다는 의미	
	BLEU(Bilingual Evaluation Understudy) • 문장의 길이와 단어의 중복을 고려하여 정답문장과 예측문장 사이의 겹치는 정도를 계산하는 지표로 사용	1에 가까울수록 정확도가 높음
	SSA(Sensibleness and Specificity Average) • 구글에서 발표한 자율 발화 모델에 대한 평가지표로 사람이 직접 자율 발화 모델과 대화를 하며 점수를 평가 • Sensibleness는 사람의 말에 대한 봇의 답변이 합리적인지를, Specificity는 해당 답변이 단답형이 아닌 구체적인 답변인지를 나타냄	
	ROUGE(Recall-Oriented Understudy for Gisting Evaluation) • 주로 문서요약(TextSummarization)의 품질을 평가하는데 사용되는 지표로 단어나 구의 재현율 측정.	1에 가까울수록 정확도가 높음
	휴먼평가지표 • AI답변의 정확성 등을 평가하기 위해 별도의 평가지표를 만들어 사람이 직접 평가하는 방식	
	기타 • (자체보유 성과지표 솔루션) AI업체 등에서 자체 보유하고 있는 성과지표가 있는 경우 사용 • (국내외 AI평가지표) 국내외 논문, 각종 AI리더보드 대회(H6) 등에서 사용하는 객관적이고 공신력 있는지표 제시하여 사용 • 분류모델에서 사용하는 여러가지 지표	

2.6 시각화

2.6.1 분석 결과 시각화

데이터 시각화는 분석을 통해 얻은 복잡한 결과와 인사이트를 그래프, 차트, 대시보드 등 시각적인 형태로 변환하여 사용자가 쉽고 빠르게 이해할 수 있도록 만드는 과정으로서, 복잡한 데이터 패턴을 직관적으로 파악할 수 있도록 하는 것이 중요하다.

- 분석 결과를 이해하기 쉽게 표현함으로써 분석 전문가가 아닌 정책 결정자나 실무 담당자도 합리적인 의사결정을 할 수 있게 한다. 시각화를 볼 대상자가 누구인지, 전달하고자 하는 핵심 메시지가 무엇인지 명확하게 파악하는 것이 중요하다.

또한, 최근 들어 데이터 분석에 있어 시각화는 단순 분석과정의 도구 역할에서 더 발전하여 성공적 분석 결과 도출을 위한 **'전략'** 측면으로 그 비중이 커지고 있다.

따라서, 시각화 단계에서는 분석 결과가 분석가 이외의 사람들이 이해하기 쉽도록 그림 및 도표를 활용하여 표현한다.

수집한 데이터 및 분석 결과를 담당자 및 관련 기관들과 공유하기 위한 목적으로 필요할 경우, Open-API 형태로 제공하도록 개발한다.

분석 결과 시각화 단계에서는 분석 결과를 다양한 관점에서 고찰하고 패턴을 쉽게 발견할 수 있도록 시각화를 설계하고 개발한다.

시각화 설계

검토 필요사항

Key Question : 시각화 설계
시각화 설계 목차로 메뉴 구조도를 작성하였는가?
목차마다 화면이 설계되어 있는가?
화면 설계에 구체적인 설명이 되어있는가? • 화면 구성이 이해하기 쉽고 한 눈에 들어오는지 확인
화면을 스토리 보드에 맞게 구성하였는가? • 시각화에 분석 주제의 전반적인 내용을 이해할 수 있는 흐름으로 통계나 차트가 담겨있는지 검토
분석 결과가 잘 반영되었는가? • 분석 결과 의미가 잘 표현되었는지 검토 • 분석 대상의 기본정보를 제공하여 분석의 이해를 돕고 있는지 검토 • 다양한 관점의 시각화 방식이 구현 가능한 설계인지 검토

검토 필요사항

Key Question : 시각화 개발
분석 결과별로 시각화를 개발하였는가? • 분석 결과를 통해 데이터를 바로 확인할 수 있도록 시각화하였는지 검토 • 모든 분석 결과의 의미가 시각화로 쉽게 이해가 되는지 검토
시각화 화면 설계 시 실무자의 의견이 반영되었는가?
데이터 분석 결과를 추출, 시각화 작동되는 단계까지 작동에 오류가 없는가?

2.6.2 주요 산출물

데이터 분석과제의 관리자 관점에서, 본 단계에서 검토해야 할 주요 산출물은 아래와 같다.

 # 시각화 설계서

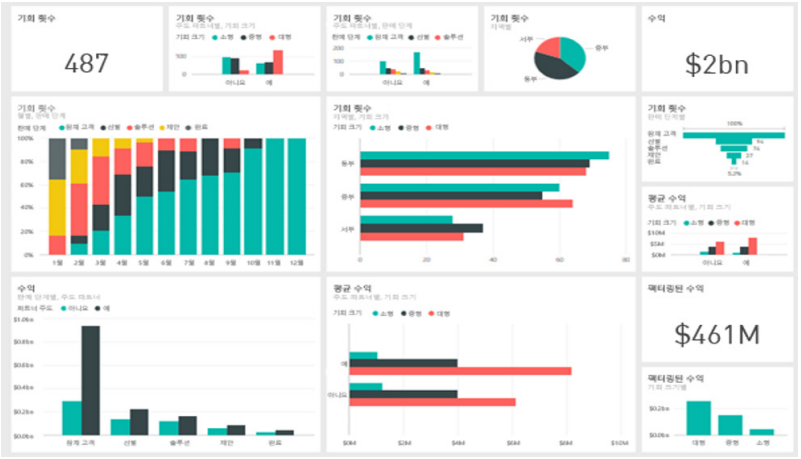
이해관계자(데이터 활용자)의 요구사항을 수렴하여 시각화를 설계한 문서

- 전체 시각화 구성
- 페이지 구성

[그림 15] 시각화 설계서

1. 전체 시각화 구성

구현되는 시각화 서비스는 아래와 같이 구성



 # 시각화 결과서

시각화 설계서를 반영해 시각화를 구축한 결과를 보여주는 보고서로, 개정 이력을 통해 요구사항의 변천을 확인할 수 있도록 관리하는 문서

- 개정 이력
- 시각화 구성 요소
- 시각화 결과(개정 이력 등)

2.7 모듈 개발

2.7.1 모듈 개발

모듈 개발 단계에서는 분석 알고리즘을 독립적인 소프트웨어 구성 요소(모듈)로 설계하고, 이를 실제 시스템에서 실행할 수 있도록 구현하는 과정을 수행한다.

모듈이란 분석 알고리즘이나 모델을 하나의 독립적인 소프트웨어 기능 단위로 구현한 것을 의미한다. 쉽게 말해, 특정 분석 기능(ex. 예측, 분류, 이상탐지 등)을 하나의 블록처럼 구현하여 시스템에 연결할 수 있도록 만든 단위다. 알고리즘은 수학적 방법, 모델은 분석 결과를 도출하는 틀이라면, 모듈은 이를 실제 시스템에서 동작할 수 있게 만든 실행 가능한 구성 요소라고 할 수 있다. 모듈은 데이터 분석 시, 시스템 탑재 및 상시적 활용을 위해 필요하다.

모듈 개발 단계에서는 이전 단계에서 선택된 분석 알고리즘이나 모델을 기반으로 하여, 이를 실제 공공시스템이나 운영환경에 적용 가능한 형태로 구현한다.
예를 들어 Python, R, Java, 또는 공공기관에서 사용하는 분석도구 등에 맞춰 코드를 작성하고, 해당 기능이 제대로 작동하는지 테스트한다.

공공부문에서는 모듈화를 통해 분석 결과를 반복 활용하거나, 다른 시스템과 연계하여 정책·행정 업무에 지속적으로 적용 가능한 구조를 만들 수 있다. 또한, 모듈 단위로 검증하고 개선할 수 있기 때문에 분석 결과의 신뢰성과 유지 관리의 효율성이 높아진다.

다만, 단발성 또는 실험적 성격의 1회성 분석 과제의 경우에는, 모델 구현과정에서 시각화나 리포트 형태의 결과 도출에 초점을 두며, 시스템 적용을 위한 모듈 설계 및 개발은 제외하는 경우도 있다. 이때는 분석 결과를 일회성 참고자료로 활용하거나, 향후 필요 시 모듈화 대상으로 전환할 수 있도록 문서화 및 코드 정리가 병행되어야 한다.

따라서, 분석 알고리즘을 적용할 모듈을 명확하게 정의하고, 이를 체계적으로 설계하며 동작 검증을 수행함으로써, 분석 알고리즘이 실무 시스템에서 정상적으로 작동할 수 있도록 보장해야 하며 분석 결과의 신뢰성과 일관성을 유지할 수 있어야 한다.

검토 필요사항

Key Question : 모듈 개발

[모듈 개발]

모듈 개발 결과물이 모델 설계와 일치하는가?

- 모듈 개발 결과물이 모델 설계와 일치하는지 확인하는 것은 설계 단계에서 정의한 모델이 실제 개발된 모듈에서 정상적으로 구현되었는지 검증하는 과정임
- 이 과정에서 모델 설계서, 데이터 흐름, 알고리즘, 성능 기준 등이 일관되게 적용되었는지 점검해야 함

  ex. 설계 단계에서 정의한 모델과 실제 개발된 코드가 일치하는지 검토(모델의 주요 기능(데이터 전처리, 학습, 예측 등)이 설계 문서와 동일하게 구현되었는지 확인, 설계 문서에서 정의한 모듈(데이터 처리 모듈, 모델 실행 모듈)이 실제 코드에서 구현되었는지 검토하는 등)

구축한 모듈이 정상적으로 동작하는가?

- 모듈이 예상한 대로 입력을 처리하고, 정확한 출력을 생성하며, 성능이 요구사항을 충족하는지 검증함
- 모듈이 정상적으로 데이터를 처리하고, 예상된 출력을 생성하는지 단위 테스트(Unit Test)를 통해 검증
- **모듈이 데이터를 올바르게 처리하고, 예상된 단계를 거쳐 결과를 도출하는지 점검:** 입력 데이터가 원하는 형식으로 변환되고, 중간 과정에서 변형되지 않는지 검증
- 모듈이 대량 데이터를 처리할 때 성능 저하 없이 정상적으로 동작하는지 검토
- 잘못된 입력, 네트워크 오류, 데이터 유실 등의 예외 상황에서 모듈이 정상적으로 대응할 수 있는지 점검
- 모듈이 운영 환경에서 실행될 때 오류 및 주요 이벤트를 기록하는 로깅 시스템이 구축되었는지 검토

데이터 저장

데이터 저장 단계에서는 데이터 구조를 표준화하고, 모든 데이터(테이블)를 융합(매칭)할 주요 키(Key) 값을 찾는다.

데이터를 융합하여 데이터의 정보 및 패턴을 파악하기 위해 기초 통계 분석을 하며, 기관 자체 인프라를 활용하여 데이터 관리가 자동화되도록 한다.

데이터 표준화 및 융합

검토 필요사항

Key Question : 데이터 표준화 및 융합

수집된 데이터의 속성(칼럼명, 유형(문자형, 숫자형 등), 크기 및 길이, 필수 여부 등) 정의가 되어있는가?

- 데이터베이스(DB) 또는 데이터 파일의 테이블 구조(스키마)가 명확히 정의되어 있는지 확인
- **칼럼명(Column Names) 일관성 점검:** 칼럼명이 일관성 있는 네이밍 규칙을 따르는지
- **데이터 유형(Data Type) 검증:** 데이터 유형이 사전에 정의된 필드 타입과 일치하는지
- **데이터 크기 및 길이 제한 검사:** 칼럼별 문자열 최대 길이 또는 숫자 범위 제한이 적절한지 점검
- 필수 필드에 NULL 값이 존재하는지 점검 및 결측치 처리(삭제, 대체 등) 수행 여부 확인 등

분석에 필요한 데이터 목록과 유형이 정의되고 변환 기준이 정의되었는가?

- **분석 목적에 맞는 데이터 목록 정의 여부 확인:** 데이터 분석을 수행하기 위해 어떤 데이터가 필요한지 먼저 정의
- **데이터 유형(Type) 및 스키마 정의 확인:** 각 데이터 필드(칼럼)의 유형(Type)이 명확하게 정의되었는지 확인
- 데이터 변환 기준(Preprocessing Rules) 정의 여부 검토
  ex. 데이터 항목(age) ▸ 원본 데이터(25) ▸ 변환 기준(범주화) ▸ 변환 예제(20대)
- 결측치 및 이상치 처리 기준 확인
- 데이터 문서화 및 정의서(데이터 사전) 존재 여부 점검

여러 유관 기관의 데이터 융합 시, 수집된 데이터는 구조를 통일하고 병합하는 표준화 과정을 거쳤는가?

- **데이터 필드(칼럼) 구조 통일 여부 확인:** 데이터가 서로 다른 시스템에서 수집된 경우, 칼럼명이 다르거나 일부 필드가 누락될 수 있기에 이를 확인하고 일관된 구조로 정리해야 함
  ex. 데이터셋 A user_id vs 데이터셋 B id
- 데이터 유형이 일관되지 않으면 병합(Join, Merge) 시 오류 발생하므로 데이터 유형을 통일
  ex. 숫자형, 날짜형, 문자열 필드 표준화 등
- **코드값 및 범주형 데이터 표준화 여부 점검:** 데이터가 여러 소스에서 들어올 경우 범주형 값(Categorical Values)이 다를 수 있음
  ex. M/F, Male/Female, 1/2 등
- **데이터 병합 및 정렬(Join&Merge) 검증:** 데이터 병합을 위해 키(Key)가 일관되게 설정되었는지 확인
- 중복 데이터 및 일관성 점검: 중복된 레코드 제거, 동일한 사용자의 데이터가 일관되게 유지되었는지 점검

데이터 관리 및 수집 자동화

검토 필요사항

Key Question : 데이터 관리 및 수집 자동화

데이터 추출 – 변환 – 탑재(ETL, Extract-Transform-Load) 하기 위한 변환 시스템은 정상 작동하나?

- ETL 프로세스는 데이터 수집, 정제, 변환 및 저장을 자동화하는 핵심 시스템이다. ETL 시스템이 정상적으로 작동하는지 확인하려면 각 단계(추출, 변환, 적재)의 성공 여부를 점검하고, 오류 발생 시 원인을 분석하여 복구할 수 있는 시스템 모니터링 체계를 구축해야 함

ETL 시스템 검증 주요 단계

- **데이터 추출(Extract) 단계 확인:** 원본 데이터에서 정상적으로 데이터를 가져오는가?
- **데이터 변환(Transform) 단계 점검:** 변환 로직이 올바르게 적용되는가?
- **데이터 적재(Load) 단계 검토:** 변환된 데이터가 목적지에 정상적으로 저장되는가?
  ex. 적재 후 데이터 손실 또는 중복 발생되는지
- **오류 로그 및 예외 처리 검토:** 장애 발생 시 원인을 파악할 수 있는가?
  ex. 오류 시 로그가 기록되고 알람이 전송되는가?
- 성능 모니터링 및 SLA(Service Level Agreement) 준수 여부 점검
  ex. ETL 실행시간이 SLA 기준을 충족하는지

데이터 보유기관으로부터 데이터를 주기적으로 수집할 수 있도록 연계가 이루어졌는가?

- 데이터 연계 방식(API, DB 연결, FTP 등) 점검
  ex. API 호출이 정상적으로 이루어지는지, API 응답시간이 적절한지(SLA 기준) 등 확인하여 정상적으로 데이터를 가져올 수 있는지 확인
- 보유기관의 데이터베이스에 주기적으로 접근하여 데이터를 가져오는 경우, DB 연결 상태 점검
  ex. 데이터베이스에서 최신 데이터를 조회할 수 있는지 점검
- FTP 또는 SFTP 방식으로 보유기관에서 파일을 주기적으로 제공하는 경우, 파일이 정상적으로 업로드 및 다운로드 되는지 확인
- 연계된 데이터 수집 일정(주기적 자동화 프로세스) 확인
- 데이터 수집 중 발생하는 오류를 기록하고 대응할 수 있는 시스템이 구축되었는지 확인
  ex. 오류 발생 로그, 장애 발생 시 자동 복구 등
- 수집된 데이터의 최신성 및 무결성 검토
- **데이터 보유기관과의 협약 및 SLA 준수 여부 확인:** 협약된 제공 주기, SLA 요구 정확도와 무결성 유지 등

시스템 연계

시스템 연계 단계에서는 기관 자체 인프라를 활용하여 데이터 정제 및 융합이 자동화될 수 있도록 데이터 연계 개발을 수행한다.

검토 필요사항

Key Question : 분석 프로젝트 시스템 연계 개발

분석 프로젝트를 위한 시스템 연계 데이터를 정하였는가?

- 시스템 연계를 성공적으로 수행하려면 연계 대상 시스템과 데이터 포맷, 연계 방식(API, ETL, 데이터베이스 동기화 등), 보안 및 인증 방식 등을 체계적으로 정의해야 함
- **연계할 시스템 및 목적 정의:** 어떤 시스템과 데이터를 연계할 것인지 확인하고, 연계 목적(데이터 수집, 실시간 분석, 자동화 보고서 생성 등)을 정의
- 연계 대상 데이터 항목 선정 및 데이터 구조 정의
- 데이터 연계 방식(API, ETL, 데이터 동기화 등) 결정
- 데이터 연계 주기 및 처리 방식 결정(실시간 vs 배치 처리)
- 데이터 보안 및 인증 방식 설정
- 연계 데이터 품질 검토 및 오류 처리 방식 정의

분석 프로젝트를 위한 시스템 연계 방식을 설계하였는가?

- **데이터 연계 목적 및 대상 시스템 정의:** 데이터를 연계하는 목적과 연계할 시스템을 정의
- **연계할 데이터 유형 및 구조 결정:** 정형 데이터, 비정형 데이터, 실시간 데이터 등
- 데이터 연계 방식(API, ETL, 데이터 동기화 등) 설계
- **데이터 흐름 및 변환(Transform) 로직 정의:** 데이터 클린징, 데이터 표준화, ETL 처리 방식(변환 후 저장할지, 연계 후 저장할지 등)
- 데이터 연계 주기(실시간 vs 배치) 및 성능 요구 사항 설정
- 보안 및 인증 방식(API Key, OAuth, TLS 등) 적용
- 오류 처리 및 예외 상황 대응 방식 수립

시스템 연계하여 활용할 데이터 저장방식을 정하였는가?

- 저장할 데이터의 유형(정형, 비정형, 실시간 데이터) 결정
- 데이터 저장 기술 선택(RDB, NoSQL, 데이터 레이크 등)
- **데이터 접근 속도 및 성능 요구 사항 분석:** 데이터 저장방식이 성능 요구사항을 충족하는지 고려
- 데이터 보안 및 접근 권한 설정
- 데이터 저장소의 백업 및 유지보수 전략 수립
- 확장성(Scalability) 및 비용 효율성 검토

데이터 보안 관리에 대하여 고려하였는가?

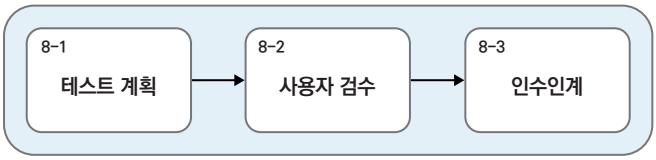
- 데이터 접근 제어가 가능한지 점검
- 사용자 인증 절차 기능이 있는지 점검
- 분석 프로세스 단계별로 나오는 결과들이 외부의 불법적인 침입으로 유출되지 않도록 조치가 취하여졌는지 확인

2.8 테스트 및 안정화

분석모델의 결과가 신뢰성을 가지는지 확인하고 실제 환경에서 안정적으로 작동하도록 조치하는 단계이다.

테스트 및 안정화 단계에서는 사업 전반에 대한 결과물(시각화, 분석 프로세스 자동화 등)을 통합 테스트하여 검증하고, 사용자 테스트 후 검수를 거쳐 분석 결과를 사용할 실무자를 위한 교육을 실시하고 인수인계를 수행한다.

[그림 16] 테스트 및 안정화 단계의 주요 활동



2.8.1 테스트 계획

테스트 계획 단계에서는 프로젝트 추진과 관련된 테스트 활동 중 통합테스트 계획을 세워 통합테스트 요구사항과 목표, 테스트 범위, 절차, 요구되는 자원, 일정 등에 대한 계획을 수립한다.

검토 필요사항

Key Question : 테스트 계획
테스트 대상 및 범위를 선정하였는가?
테스트 대상 및 범위가 사업 결과물(자동화 시스템 및 시각화 등)을 판단하는 데 적합한가?
테스트 유형별로 점검항목을 구체화하였는가? • 기능 테스트, 성능 테스트, 보안 테스트 등 세부 유형별 점검항목
테스트 항목을 단계적으로 정하였는가?
테스트 예상 결과를 도출하였는가?
테스트 수행 주체 및 참여자를 선정하였는가?

2.8.2 사용자 검수

사용자 검수 단계에서는 테스트 계획에 따라 통합테스트를 실시하여 프로젝트를 점검하고 검수한다.

검토 필요사항

Key Question : 사용자 검수

[통합테스트]

테스트 계획에 따라 진행하였는가?

- 테스트 결과가 예상 결과와 동일한가, 다르다면 그 원인을 찾고 해결하였는가?
- 분석결과 검증을 위해 관련 전문가와 동행한 현장 실사를 진행하였나?

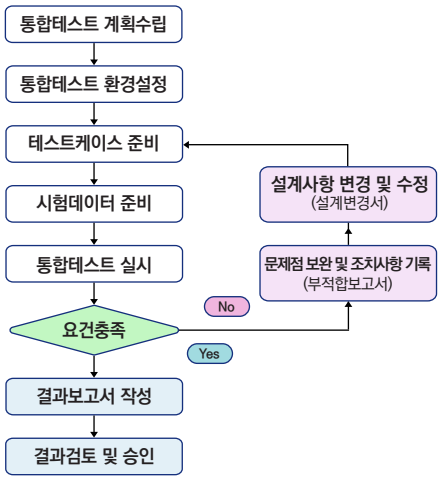
  ex. CCTV 사각지대의 경우, 분석 결과 사각지대로 파악된 지점을직접 찾아가 실사하고 분석결과와 실제가 일치하는지 검증 필요

[사용자 승인 테스트]

사용자가 통합테스트에 참석하였는가?

혹은 사용자를 위한 시연을 실시하였는가?

[그림 17] 통합 테스트 절차 예시

시험 절차	수행 내역
	**통합테스트 계획서 작성** • 목적 및 범위, 방법, 시험 케이스 분류, 시험 데이터 확보, 시험일정 계획, 시험 추진조직 **사전 준비사항** • 시험 항목, 시험 데이터, 화면 및 항목 설명서, 통합테스트 계획서 확인 **시험 케이스 도출** • 통합시나리오 및 테스트케이스 작성 **시험 항목 설정** **시험 데이터 확보** • 실 운영 데이터를 전환하여 활용 **시험 실시** • 시험 항목별 시험데이터 준비하여 통합테스트 실시, 시험 점검표에 처리결과 기록 • 통합 시험 시 오류사항은 문제점을 보완하여 재시험 수행

2.8.3 사용자 교육

인수인계 단계에서는 실제 업무에서 분석 결과를 활용하게 될 실무자를 위한 교육을 실시한다.

검토 필요사항

Key Question : 인수인계
사용자를 위한 시스템(분석 자동화 및 시각화 화면) 사용자 지침서를 작성하였는가? • 분석 프로세스는 단계별 설명이 충분한지 검토 • 알고리즘 단계별 주석의 설명이 이해하기 쉬운지 검토 • 시각화 화면의 구성에 대한 설명이 모두 되어있는지 검토 • 시스템 작동 단계별 매뉴얼이 구성되어 있는지 검토
사용자 및 운영자를 위한 교육을 계획하고 실시하였는가? • 교육 대상자가 적합한지 검토 • 인수인계를 위해 교육 내용이 충분한지 검토 • 교육 대상자의 참석률이 높았는지 확인

3. 분석모델 활용 및 고도화

활용 및 분석모델 고도화 단계에서는 분석 결과 업무적용에서 효과검증, 홍보, 유지관리, 분석모델 고도화에 이르기까지의 과정을 다룬다.

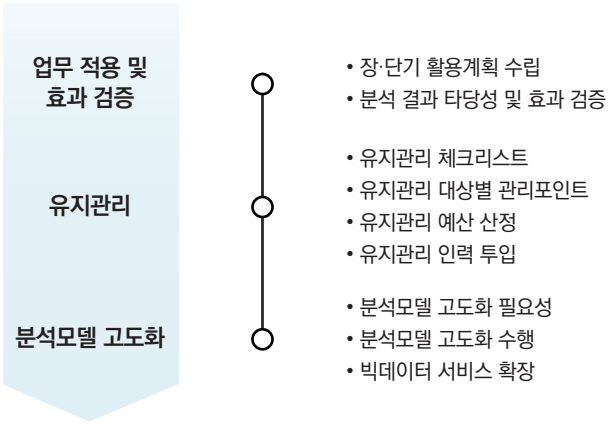
기 추진한 분석과제 결과를 업무에 적용하기 위한 준비, 실제 업무에 적용, 그리고 적용 후에 나타나는 효과에 대한 검증까지 자세히 안내한다.

이때 분석 결과의 실효성을 높이기 위해, 분석 예측값과 실제 데이터를 비교하는 검증 과정이 필요하다. 예를 들어, 의약품 수급 불안 예측 모델의 경우, 예측된 품절 위험 시기와 실제 품절 발생 시점을 비교함으로써 모델의 정확도를 점검하고 결과를 보완할 수 있다. 이러한 비교 검증은 분석 결과의 신뢰성과 활용 가능성을 제고하는 핵심 과정이다.

분석 서비스의 안정적인 운영과 현행화된 데이터의 수집·저장·관리, 분석 결과의 정책개발 및 행정업무 지원 등의 지속적인 활용을 위해 유지관리가 필요하다.

또한, 분석모델에 대한 주기적이고 지속적인 고도화에 대한 필요성이 제기됨에 따라 기관에서는 분석 결과에 대한 활용 및 확산, 분석모델에 대한 지속적인 고도화를 통해 데이터 분석과제 추진의 목적을 달성할 수 있다.

[그림 18] 분석모델 활용 및 고도화 추진 절차



3.1 분석 결과 활용

분석 결과의 효과를 검증하고 이를 실질적으로 업무에 반영하여 성과를 극대화한다.

분석 계획을 수립할 때 세웠던 분석의 목표성과를 달성했는지 정기적으로 점검하고 모델이 유의미한지 검증한다.

데이터를 성공적으로 활용하려면 다음과 같은 5가지 과정을 거쳐야 한다.

❶ 가치 있는 데이터 보유
❷ 내·외부 데이터 통합 및 서비스
❸ 통합 데이터에서 인사이트 발견
❹ 발견한 인사이트를 현업에 적용
❺ 현업 적용에 필요한 변화 관리와 지속적 모니터링

그런데 현실에서는 많은 데이터 기반 과제가 분석만 하거나 인사이트 보고서로 마무리하는 경우가 많다. 이는 분석 관련 프로젝트를 진행하다가 중간에 멈춘 것이나 다름없다.

- 분석을 거쳐 인사이트를 발견했다면, 이에 따른 활용계획 수립 ▸ 실행 ▸ 평가가 뒤따라야 한다

Box Tip

분석결과 활용 체크포인트

- **활용자의 아이디어가 수렴되었나?**
  데이터 분석 및 활용에 대한 계획을 수립할 때 받았던 전체 조직 구성원들의 아이디어와 의견이 수렴되었는지
- **초기 활용 목표가 달성되었나?**
  분석 결과를 활용하고, 적용하고, 활성화하려는 계획이 달성되었는지
  - 분석 결과의 단기 활용계획과 중장기 활용 계획 달성 여부
  - 데이터 활용과 관련해 단계적이고 구체적인 계획 달성 여부
- **분석 종료 후 결과 활용을 위한 부서 협조 및 예산확보는?**
  - 수립한 데이터 분석 및 활용 방안이 기술적, 경제적으로 충분히 실현 가능했는지 점검

3.2 분석모델 고도화 및 유지관리

기존 모델을 개선하여 분석 정확도를 높이고, 최신 데이터와 기술 변화에 적응할 수 있도록 고도화 작업을 수행한다.

3.2.1 분석모델 고도화

분석모델 고도화 필요성

주변 환경의 변화를 빅데이터 분석모델에 지속적으로 반영하기 위해 분석모델의 고도화 방안이 필요하다.

이를 위해 실시간으로 변화하는 데이터를 통해 도출되는 분석 결과와 모델의 성능을 지속적으로 모니터링하고 분석하여 고도화가 필요한 시점을 판단해야 한다.

또한, 분석 대상 업무와 환경은 지속적으로 변화하고 이에 따라 관련된 데이터도 변화하거나 다양화 되어 분석모델이 변화된 업무와 데이터를 계속해서 수용할 수 있어야 한다.

데이터 분석모델 고도화를 위해 **기관 내·외부 데이터 축적 ▸ 업무 활용 ▸ 데이터 관련 기술 성숙 ▸ 데이터 축적의 선순환 구조**가 작동해야 한다.

분석모델 고도화 수행

초기 분석모델보다 더 좋은 결과물을 도출하기 위해서는, 기존 분석과정 및 결과를 재검토하고 보완하여 모델을 고도화하는 것이 필요하다.

- 고도화 범위에 반영하기 위하여, 데이터 수집, 전처리, 분석 방법론, 분석 결과까지 과제의 전반에 대하여 보완하거나 신규 추가할 사항을 정리하여 개선방안을 도출한다.
- 분석모델의 고도화를 위해 주체(기관)별로 장·단기 제도, 운영 방안을 수립한다.

각 방안을 단계적으로 구체화하고 분석과제의 보완사항을 반영한다.

- 단계별 적용 업무 작성 및 각 방안의 근거를 마련한다.

데이터 서비스 확장

분석 결과를 지속적으로 활용하기 위해 데이터 서비스 확장이 필요하며, 다음과 같은 점검 사항을 확인하여 계획을 수립한다.

- 분석모델 고도화는 모델 자체의 **'질적'** 개선임에 반하여 데이터서비스 확장은 활용 기관의 범위 확대 등 모델의 **'양적'** 확산으로 개념상 구분할 수 있다.

분석 서비스 추가

- 추가되는 서비스의 유형을 고려: 데이터 가시화 방법 변경, 분석 프로세스 변경, 데이터 수집/가공/분석기법 추가 적용 등
- 비용 대비 효과에 대한 고려

분석 범위의 확장

- 분석 서비스 추가나 분석 품질의 향상이 요구되는 상황에서 분석 범위를 확장하여 달성할 수 있는지 고려
- 수집, 분석 대상 데이터의 범위 및 데이터의 종류를 확대, 기존 분석 결과와 결합함으로써 분석 범위 확장
- 분석 대상과 분석 항목을 확장하고 다양한 분석모델 적용 확대

데이터의 이질성 고려

- 기 데이터와 추가 데이터의 구조적인 이질성과 데이터 항목 간의 의미적인 이질성, 데이터값의 이질성을 식별해야 하며 분석을 위한 통합 모델 수립
- 통합 모델에 따라 데이터 가공 및 필터링 방법의 변경이 필요한지 검토
- 통합 모델은 메타데이터 관리체계에 편입하고 관리

3.2.2 유지관리

분석 서비스의 안정적인 운영과 현행화된 데이터의 수집·저장·관리, 분석 결과의 정책개발 및 행정업무 지원 등의 활용, 성과 공유 등을 위해 유지관리가 필요하다.

기관 담당자는 유지관리를 위해 사업 추진 이후 제공되는 서비스에 대한 내용 파악이 필요하며, 파악된 제공 서비스 내용을 기반으로 유지관리 점검항목을 활용하여 세부 유지관리 사항을 구체화하여 점검해야 한다.

[표 13] 유지관리 점검항목

구분	점검항목	확인
운영 관리단계	운영 관리를 위한 전담인력을 선별하여 구성하였는가?	
	유지관리 관련 기관의 책임과 역할을 명확히 정의하였는가?	
	운영 인력의 업무역량 분석 및 교육훈련 방안을 마련하였는가?	
	데이터 품질, 분석, 보안 및 서비스 운영관리를 위해 협력하는가?	
	지속적인 운영관리를 위한 협력체계를 확보하였는가?	
확장단계	분석 서비스의 추가가 필요한가?	
	분석 범위와 항목 확대가 필요한가?	
	초기 데이터와 추가 데이터의 이질성을 고려하는가?	
	분석 품질 향상을 위해 데이터 및 서비스의 상호운용성을 고려하는가?	
	다양한 데이터의 수집, 분석 및 재활용성을 위한 변화를 관리하는가?	

데이터 분석과제가 한 번에 끝나는 원타임(One-time) 프로젝트가 아님을 감안하여 매년 **'유지·관리 및 업그레이드'**, **'새로운 업무 반영'**이라는 유지관리 목표를 수립하고, 유지관리 대상별로 구분하여 그에 따른 유지관리 수행 내용을 작성해야 한다.

[표 14] 유지관리 대상별 관리 포인트

대상	관리 포인트
정책/제도	정부 정책/제도 개발 및 적용
업무	신규 업무 반영, 기존 업무 업그레이드
관련 시스템	빅데이터 포털 및 Open API 현행화
인력	업무역량, 책임과역할(R&R), 교육훈련 등
서비스	분석과제 발굴, 활용방안 마련, 성과관리 등
분석모델	분석 알고리즘 주기, 변수, 소스(데이터 원천) 등
데이터	현 시점의 현행화 데이터 확인

유지관리에 필요한 활동을 도출하고 각 활동별 투입공수를 산정하여 유지관리 예산을 수립한다.

유지관리 수행 시 상주·비상주 여부, 특정 유지관리 항목은 전문인력이 필요하다는 내용 등의 제약사항 관리가 필요하다.

유지관리가 제대로 수행되지 않아 분석 결과가 최신 데이터와 업무 내용을 반영하지 못한다면 현업 부서에서는 빅데이터 분석 결과를 사용할 수 없다.

서비스의 최신성·지속가능성

데이터는 지속적으로 수집·생성되고, AI 기술·모델은 지속적으로 발전한다. 따라서 아무리 성공적으로 구현된 서비스라고 할지라도 지속적으로 업데이트를 하지 않으면 무용지물이 된다.

이에 서비스가 지속 운영될 수 있도록 향후 수집되는 데이터에 대한 수집·관리방안을 마련하고, 결정된 최적의 모델일지라도 변화된 상황에 따라 지속적으로 업데이트를 하는 것이 매우 중요하다.

검토 필요사항

Key Question : 서비스가 지속 운영될 수 있도록 업데이트 방안과 활용방안이 마련되어 있는가?

데이터 업데이트

- 업데이트가 필요할 때 특별한 제한 없이 AI 모델을 재학습하게 하는 방안 마련
- 향후 수집되는 데이터에 대한 수집 주체, 수집 방법, 예상되는 규모 등 관리 프로세스 방안 마련

기술 및 모델 업데이트

- 향후 기술 및 모델의 업데이트 필요성 및 가능성에 대한 검토 분석 결과 검토

분석 결과 검토

- AI 유지보수 전문인력 배정 및 인수인계서 관리

유지보수 관리

- 데이터 운영·활용을 위한 데이터 설명서 등을 포함한 사용자 활용 가이드라인 마련

첨부1: 데이터 분석과제 단계별 주요 점검 항목

단계	점검 및 고려사항 일람표 – 가이드 본문에 질문별 가이드 답변 수록 –	
1.1 추진체계 구성	이해관계자 범위	• 데이터 분석 이해관계자 점검 및 고려사항
	업무의 체계성	• 과제 추진을 위한 역할분담·지원체계 방안이 마련되어 있는가
1.2 데이터 분석과제 발굴	주제 도출 방법	• 데이터 분석 주제는 어떻게 도출하는가
	데이터 분석 주제 (1)	• 데이터 분석 담당자들이 과제 발굴 시 실제 고려하는 점검항목
	데이터 분석 주제 (2)	• 데이터 분석 주제 점검 점검항목
	분석과제 선정	• 분석과제 선정 점검항목
1.3 분석 추진계획 수립	분석 계획 요소	• 성공적인 데이터 분석 계획 수립을 위한 고려사항은 무엇인가
	사업의 타당성	• 구현하고자 하는 목표서비스에 대해 명확히 설명할 수 있는가
	기술의 효용성	• 적용하고자 하는 AI 기술이 목표서비스 구현을 위한 최선의 방법인가
	데이터 확보방안 검토	• 데이터 분석을 위한 다양한 데이터 원천을 파악하고 있는가 • 분석 대상 외부 데이터 확보를 위한 기관 간 협의를 했는가 • 외부 데이터는 내부 데이터와 통합해 분석 가능한가 • 수집 예정 데이터에 개인정보가 포함되었는가
	분석 비용 및 기간 산정	• 분석비용 및 기간산정이 최적화될 수 있도록 프로젝트의 범위가 명확히 정의되었는가 • 예산수립 및 기간산정의 적정성 여부를 검토하였는가

1.4 분석 결과 활용계획 수립	분석 결과 활용계획 수립 점검사항	• 데이터 분석 및 활용에 대한 방안을 수립 시, 전체 이해관계자들의 아이디어와 의견이 반영되었는가 • 분석 결과를 활용하고, 적용하고, 활성화하려는 계획을 세웠는가 • 데이터 분석 종료 후 결과 활용을 위한 관련부서 협조와 예산확보가 가능한가 • 분석모델을 지속적으로 유지하고 업그레이드 할 수 있는가
	성과관리 계획	• 과제목적에 적합한 성과목표를 세웠는가 • 성과목표에 대해 대상, 기간, 측정 항목들이 명확하고 구체적으로 작성되었는가 • 성과 측정시기별 계획안이 마련되었는가 • 성과관리 지표가 분석과제를 평가하기에 타당한가 • 성과지표의 단기/중기/장기 활용계획을 수립하였는가 • 과제 수행 결과 활용 계획을 세웠는가 • 분석모형의 일반화, 확산 및 운영방안 수립 등의 활용방안을 제시하였는가
2.1 요건 정의	일정계획 수립	• 일정계획을 수립할 때 고려해야 할 사항
	분석과제 요구사항	• 분석과제의 요구사항을 잘 정의하였는가
	인터뷰 수행	• 인터뷰를 통해 구체화가 필요한 내용 예시
	선행사업 분석 및 자료조사	• 선행사례 분석 시 분석 항목 예시
	업무 및 관련 시스템 현황 분석	• 기술과 서비스 인프라가 호환 가능한가 • 호환 가능성 검토 사항 예시
	수집 환경 분석	• 내부망 또는 외부망에서 데이터 수집이 가능한지 점검했는가 • 외부망 활용 시 데이터 수집 루트를 정하였는가 • 데이터 수집을 위한 개발 환경 계획이 잘 되었는가 • 데이터 수집을 위한 환경 분석이 완료되었는가

2.2 데이터 수집	**필요 데이터 정의**	• 데이터 수집 대상 목록을 작성하였는가
	데이터 확보방안 수립 및 점검	• 조사한 데이터는 사용가능한 데이터인가 • 조사한 데이터는 수집 가능한가 • 데이터 수집 항목 중 민간 기업에서 수집되는 경우가 있다면, 데이터 수집비용은 적절한가 • 수집대상 데이터 보유기관 및 담당자를 조사하였는가 • 보유기관이 데이터를 제공하는 형식을 조사하였는가 • 보유기관으로부터 데이터 메타정보(칼럼, 건수, 크기, 길이)를 입수하였는가
	개인정보에 대한 점검	• 수집 예정 데이터에 개인정보가 포함되었는가 • 수집 예정 데이터에 개인정보가 포함되어 있다면, 유출될 문제에 대해 방안을 세웠는가
	기술의 안전성	• AI 적용시 예상되는 위험요소를 분석하였는가 (법적·윤리적 이슈, 비용-효과 이슈, 기술적 호환성, 전문 인력 호환성 등)
	데이터의 보안성 및 투명성 확보	• 데이터 보안등급 설정 및 관리 절차를 투명하게 하였는가
	수집 환경 구성	• 타 기관 데이터의 수집 환경을 구성하였는가 • 수집할 데이터의 크기 및 최대 저장기간 등을 고려해 저장 공간을 설계하였는가 • 데이터 수집 기술이 웹 크롤링인 경우, 수집 환경 구성에 아래와 같은 사항을 검토하였는가 • 데이터 수집을 위한 사전테스트를 진행하였는가
	데이터 수집	• 데이터를 수집하기 위한 보유기관과의 협의 일정을 세웠는가 • 보유기관에 요청할 데이터 리스트는 구체적으로 작성하였는가 • 기관 내외부 데이터 수집을 완료하였는가
	데이터 수집 및 연계방안 점검사항	• 내부 데이터의 구조와 연계 가능성을 확인하였는가 • 외부 데이터를 활용할 경우, 내부 데이터와의 연계 가능성을 미리 검토하였는가

2.3 데이터 전처리	데이터 정제	• 데이터 기초분석을 하였는가 • 오류 데이터 제거, 캐릭터셋 통일, 길이체크, 날짜 포맷 통일 등 데이터 검증 및 정제 과정을 거쳤는가 • 코드 데이터가 정의된 코드값에 맞게 변환되었는가 • 데이터에 포함된 개인정보는 모두 비식별화를 하였는가 • 사용하는 모든 데이터가 개인정보보호법에 따른 제약사항을 준수하고 있는가
	데이터 품질 확보	• 데이터 품질관리 방안을 검토하였는가
	데이터의 보안성 및 투명성 확보	• 데이터 보안등급 설정 및 관리 절차를 투명하게 하였는가
	데이터 변환	• 코드 데이터가 코드값에 맞게 변환되었나?
	EDA 수행 시 고려사항	• 비즈니스 목표와 일치하는가 • 데이터의 대표성을 확인했는가 • 데이터 변환 필요성을 고려했는가 • EDA 결과에 대한 도메인 전문가 등을 통해 합리적인 자문을 수행했는가
2.4 분석모델 설계	분석모델 설계	• 필요한 데이터 항목이 정해졌는가 • 데이터 항목별 표준화 방법을 정하였는가 • 데이터 수집 가능 항목에 따라 단계별로 모델이 설계되었는가 • 분석 검증 통계기법을 정하였는가 • 현업, 데이터전문가, 실무자들로 구성된 자문위원단과 분석 모델 설계를 검토하였는가 • 분석모델 설계(수정안)가 타당한가
	모듈 설계	• 모듈에 대한 기능을 정의하였는가 • 모듈설계를 실시하였는가 • 모듈 개발 결과물이 모델 설계와 일치하는가 • 구축한 모듈이 정상적으로 작동하는가
	표준화 모듈 설계	• 표준화 모듈 설계를 하였는가
	데이터 정제 모듈 설계	• 데이터 정제 모듈 설계를 하였는가

2.5 모델 구현	분석 알고리즘 모델링	• 분석모델에 대해 조사하였는가
	모델 선정	• AI 알고리즘 적용을 위한 최적의 분석모델을 선정하였는가
	모델 검증의 유효성	• 모델 검증 및 효과평가 방법이 제시되어 있는가
	모델 성능 평가 객관성	• 모델 성능 평가를 객관적으로 2차 검증을 하였는가
2.6 시각화	시각화 설계	• 시각화 설계 목차로 메뉴 구조도를 작성하였는가 • 화면 설계에 구체적이 설명이 되어 있는가 • 화면을 스토리보드에 맞게 구성하였는가 • 분석 결과가 잘 반영되었는가
	시각화 개발	• 분석 결과별로 시각화를 개발하였는가 • 시각화 화면 설계 시 실무자의 의견이 반영되었는가 • 데이터 분석 결과를 추출, 시각화가 작동되는 단계까지 작동에 오류가 없는가
2.7 모듈 개발	모듈 개발	• 모듈 개발 결과물이 모델 설계와 일치하는가 • 구축한 모듈이 정상적으로 작동하는가
	데이터 표준화 및 융합	• 수집된 데이터의 속성(칼럼명, 유형(문자형, 숫자형 등), 크기 및 길이, 필수 여부 등) 정의가 되어있는가 • 분석에 필요한 데이터 목록과 유형이 정의되고 변환 기준이 정의되었는가 • 여러 유관 기관의 데이터 융합 시, 수집된 데이터는 구조를 통일하고 병합하는 표준화 과정을 거쳤는가
	데이터 관리 및 수집 자동화	• 데이터 추출 – 변환 – 탑재 (ETL, Extract-Transform-Load)하기 위한 변환시스템은 정상 작동하는가 • 데이터 보유기관으로부터 데이터를 주기적으로 수집이 가능하도록 연계가 이루어졌는가
	분석 프로젝트 시스템 연계 개발	• 분석 프로젝트를 위한 시스템 연계 데이터를 정하였는가 • 분석 프로젝트를 위한 시스템 연계방식을 설계하였는가 • 시스템 연계하여 활용할 데이터 저장방식을 정하였는가 • 데이터 보안관리에 대해 고려하였는가

2.8 테스트 및 안정화	테스트 계획	• 테스트 대상 및 범위를 선정하였는가 • 테스트 대상 및 범위가 사업결과물(자동화 시스템 및 시각화 등)을 판단하는데 적합한가? • 테스트 유형별로 점검항목을 구체화하였는가 • 테스트 항목을 단계적으로 정하였는가 • 테스트 예상 결과를 도출하였는가 • 테스트 수행 주체 및 참여자를 선정하였는가
	사용자 검수	• 테스트 계획에 따라 진행하였는가 • 사용자가 통합테스트에 참석하였는가 혹은 사용자를 위한 시연을 실시하였는가
	인수인계	• 사용자를 위한 시스템(분석 자동화 및 시각화 화면) 사용자 지침서를 작성하였는가 • 사용자 및 운영자를 위한 교육을 계획하고 실시하였는가
3. 분석모델 활용 및 고도화	분석 결과 활용 체크포인트	• 활용자의 아이디어가 수렴되었는가 • 초기 활용목표가 달성되었는가 • 분석 종료 후 결과 활용을 위한 분석 협조 및 예산 확보가 가능한가
	유지관리 체크리스트	• 운영관리단계 점검항목 • 확장단계 점검항목
	서비스의 최신성 및 지속가능성	• 서비스가 지속 운영될 수 있도록 업데이트 방안과 활용 방안이 마련되어 있는가

첨부2: 데이터 분석 주요 용어

데이터 분석과제

- 해결하고자 하는 문제를 명확히 규정하고, 이를 위해 어떤 데이터를 활용해 어떤 분석을 수행할지를 구체적으로 설정하는 과정

데이터 분석모델

- 설정된 분석 목적에 따라 데이터를 수학적 또는 통계적 구조로 표현하여, 예측·분류·설명 등 특정 목표를 달성하기 위한 알고리즘의 기반을 설계하는 과정

할루시네이션(Hallucination)

- AI 모델이 실제로 존재하지 않는 정보를 생성하거나 왜곡된 데이터를 출력하는 현상

매칭키(Matching Key)

- 데이터 분석이나 데이터베이스에서 서로 다른 데이터 세트를 연결하거나 비교하기 위해 사용하는 고유한 식별자. 주로 두 개 이상의 데이터 소스에서 관련된 정보를 결합하기 위해 사용

데이터 전처리

- 데이터를 정제·가공하여 데이터의 품질을 높이고 결과의 신뢰성을 높이는 과정

데이터 정제(Cleansing)

- 데이터에서 결측값, 중복값, 이상값 등을 제거하거나 수정하는 과정

데이터 변환(Transformation)

- 데이터를 분석에 적합한 형태로 변환하는 과정

데이터 통합(Integration)

- 여러 출처에서 수집된 데이터를 하나의 일관된 데이터셋으로 통합하는 과정

데이터 축소(Reduction)

- 분석의 효율성을 높이기 위해 데이터를 요약하거나 축소하는 과정

첨부2: 데이터 분석 주요 용어

데이터 결합 전문기관

- 서로 다른 개인정보처리자가 보유한 가명정보를 안전하게 결합하기 위해 지정된 기관

데이터 가명처리 지원센터

- 개인정보를 안전하게 활용하기 위해 가명처리 및 결합을 지원하는 전문기관

탐색적 데이터 분석(EDA: Exploratory Data Analysis)

- 데이터의 분포, 패턴, 이상치, 결측치, 상관관계 등을 확인하여 데이터를 이해하고 특성을 파악하기 위해 수행하는 과정. 데이터 분석에서 EDA는 단순한 사전 점검이 아니라 데이터의 본질을 이해하고 분석 방향을 설정하는 핵심 과정

분석모델 설계

- 데이터 분석의 목적을 달성하기 위해 분석과정과 절차, 구조를 구체적으로 계획하고 체계화하는 작업

분석 시나리오

- 데이터 분석 프로젝트에서 분석의 전체 흐름과 방향성을 사전에 계획하는 단계로, 분석 목표를 달성하기 위한 구체적인 절차와 방법을 기술한 문서

분석 모델링

- 정책 결정, 서비스 개선, 업무 효율화 등을 위해 데이터를 기반으로 인사이트를 도출하고 예측하거나 분류하는 통계적·기계학습 기반 모델을 구축하는 과정

분석 설계안

- 분석 모델링을 체계적으로 수행하기 위한 기획 문서

모듈

- 분석 알고리즘이나 모델을 하나의 독립적인 소프트웨어 기능 단위로 구현한 것

첨부2: 데이터 분석 주요 용어

모듈

- 분석 알고리즘이나 모델을 하나의 독립적인 소프트웨어 기능 단위로 구현한 것

모듈 설계

- 데이터 수집, 정제, 전처리 등 모델 개발에 필요한 작업을 체계적이고 재사용 가능한 단위로 나누어 구조화하는 과정

모델 구현

- 분석 모델링과 모델 검증을 실제로 수행하여 분석 목표를 실현하는 단계로, 설계된 분석모델을 데이터에 적용하고 최적의 결과를 도출하는 과정

ML(Machine Learning) 알고리즘

- AI 시스템이 작업을 수행하기 위해 사용하는 일련의 규칙 또는 프로세스로, 주로 새로운 데이터 인사이트와 패턴을 발견하거나 주어진 입력 변수 세트에서 출력 값을 예측하기 위해 사용

모델 검증

- 데이터 분석을 위한 과정의 일환으로 개발한 모델이 실제 분석 업무에 사용 가능한지를 평가

데이터 시각화

- 분석을 통해 얻은 복잡한 결과와 인사이트를 그래프, 차트, 대시보드 등 시각적인 형태로 변환하여 사용자가 쉽고 빠르게 이해할 수 있도록 만드는 과정