| **빅데이터 분석 정의서** |
| --- |

| **주제 :** | **자격증 합격을 위한 스마트 스터디 플랫폼** |
| --- | --- |

| **2025.06.26** |
| --- |

**Ⅰ. 개요**

**1. 아이디어 주제:**  
생성형 AI와 텍스트 마이닝을 활용한 자격증 대비 스마트 스터디 플랫폼 개발

**2. 개발 목표:**  
자격증 수험생이 주제별로 효율적으로 학습하고, AI 해설 및 실전 CBT 환경을 통해 학습 효율을 높이고 합격률을 향상시키는 것을 목표로 함

**3. 개발 내용:**

기출문제 기반 문제은행 구축 및 주제 자동 분류  
Gemini API를 활용한 AI 해설 챗봇 기능  
실제 시험과 유사한 CBT 문제풀이 UI 구현  
사용자별 학습 성취도 분석 및 시각화  
학습 알림 서비스, 시험 합격 시 환급 기능 포함

**Ⅱ. 기능별 빅데이터 분석 명세서**

| **기능명** | 문제은행 구축 및 주제 자동 분류 및 신빙성 확보 |
| --- | --- |
| 1. **데이터 준비** | |
| **데이터 정의** | 정보처리기사 기출문제 (문제문, 선택지, 정답, 해설), 과목, 회차, 주제 분류 정보 |
| **데이터 획득 방법** | - 크롤링: https://www.cbtbank.kr/category/정보처리기사  - OCR 처리: 이미지로 된 문제의 텍스트 추출  - 예상 수집량: 약 800문제 |
| **수집 데이터** | - 문제: “다음 중 정규화에 해당하지 않는 것은?”  - 선택지: ① 제1정규형 … ④ 제4정규형  - 정답: ④  - 해설: “제4정규형은 고급 정규화로 …”  - 과목: 데이터베이스 / 회차: 2023년 1회 |
| **2. 전처리** | |
| **전처리 과정** | 1. HTML 태그 및 공백 제거  2. 이미지 문제 OCR로 텍스트화  3. 정답/해설 분리  4. 주제 키워드 추출  5. 학습용 JSON/CSV 포맷으로 구조화 |
| **전처리 데이터** | json<br>{<br> "question": "다음 중 정규화에 해당하지 않는 것은?",<br> "choices": ["제1정규형", "제2정규형", "제3정규형", "제4정규형"],<br> "answer": "제4정규형",<br> "subject": "데이터베이스",<br> "topic": "정규화"<br>} |
| **3. 데이터 분석** | |
| **데이터 분석 목표** | 수집한 데이터를 통해 분석할 목표  ex) 얼굴 이미지에서 눈 위치 찾기 |
| **데이터 분석 시나리오** | 데이터 분석에 사용할 알고리즘  ex)  1) Dlib 라이브러리를 사용해서 이미지에서 눈의 좌표 찾기  2) 찾은 좌표값을 OpenCV라이브러리를 사용해서 이미지에 표현 |
| **데이터 분석 결과** | 데이터 분석 결과를 표시 |
| **4. 모델 생성 및 학습** | |
| **모델링 목표** | 인공지능 모델을 사용해서 어떤 것을 예측하게 할것인지 |
| **학습 모델** | 모델링 구현에 사용할 모델 나열 |
| **학습 방법** | 모델구현을 위한 학습방법 나열 |
| **모델 예측 결과** | 모델 예측 결과를 표시 |
| **5. 검증** | |
| **분석 검증 방안** | 구현한 모델에 대한 검증 방안 나열 |
| **분석 평가 결과** | 학습 후 모델에 대한 평가 결과 |