

머신러닝을 활용한 공공도서관 이용자 세분화 및 연체 발생 예측에 관한 연구

A Study on Machine Learning-Based User Segmentation and Overdue Prediction for Public Libraries

정 희 진 (Heejin Jung)*

이 규 정 (Gyujeong Lee)**

목 차

- | | |
|----------|------------------|
| 1. 서 론 | 4. 성능 평가 및 결과 분석 |
| 2. 선행 연구 | 5. 결 론 |
| 3. 연구 방법 | |

초 록

본 연구는 머신러닝 기반 이용자 유형 세분화 및 유형별 연체 예측 모델 임계값 최적화 방식에 기반한 공공도서관의 예방적 연체 관리 전략의 적용 가능성을 고찰한다. 이를 위해 국내 C공공도서관 성인 이용자의 도서 대출·반납 데이터(2,148명, 75,377건)에 LightGBM을 적용하여 연체 예측 모델을 학습하고, K-Prototypes 군집 분석을 적용하여 전체 이용자를 연체율 차이(5~25%)가 뚜렷하게 나타나는 5개 유형으로 세분화한 후, 각 유형별 연체 예측 성능을 단일 최적 임계값과 유형별 최적 임계값을 각각 적용하여 비교 분석하였다. 분석 결과, 단일 최적 임계값 적용 시 유형별 재현율은 34.6~94.2%, F1-Score는 0.257~0.513으로 큰 편차를 보였으나, 유형별 최적 임계값 적용 시 모든 유형에서 재현율 75% 이상을 확보하면서 F1-Score도 단일 임계값 대비 대체로 향상됨을 확인하였다. 다만, 고령층 우수형의 경우 낮은 예측 성능을 보여 별도의 관리 방안이 필요함을 확인하였다. 이러한 결과는 일괄적인 연체 대응 방식의 대안으로 이용자 유형 세분화를 통한 차등화된 예측 및 관리 전략의 적용 가능성을 시사한다.

ABSTRACT

This study explores the feasibility of a preventive overdue management strategy for public libraries by segmenting user types and optimizing type-specific thresholds for an overdue prediction model based on machine learning. Based on the actual borrowing data from adult users of a C public library (2,148 users, 75,377 transactions), an overdue prediction model was trained with LightGBM. K-Prototypes clustering was then applied to segment users into five distinct types, which showed clearly different overdue rates ranging from 5% to 25%. For each segment, model performance was compared using a single optimal threshold versus segment-specific optimal thresholds. Results showed substantial variation in recall (34.6%-94.2%) and F1-Scores (0.257-0.513) across segments when a single optimal threshold was applied. In contrast, segment-specific thresholds kept recall above 75% for all segments and improved F1-Scores in most cases. One exception was the Senior Excellent segment, where the low overdue rate limited predictive performance, indicating the need for a supplementary management approach. These findings suggest that segment-specific prediction and management strategies can serve as an alternative to standard overdue response practices in public libraries.

키워드: 도서 연체, 이용자 세분화, 군집 분석, 연체 예측 모델, 공공도서관

Book Overdue, User Segmentation, Cluster Analysis, Overdue Prediction, Public Library

* 고려사이버대학교 융합정보대학원 석사과정(iris9181@koreacu.ac.kr / ISNI 0000 0005 3079 5876)
(제1저자)

** 고려사이버대학교 융합정보대학원 조교수(lgi413@koreacu.ac.kr / ISNI 0000 0005 3056 2217)
(교신저자)

논문접수일자: 2026년 7월 21일 최초심사일자: 2026년 8월 6일 게재확정일자: 2026년 8월 13일
한국문헌정보학회지, 60(3): 315-338, 2026. <http://dx.doi.org/10.4275/KSLIS.2026.60.3.315>

© Copyright © 2026 Korean Society for Library and Information Science

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. 서론

현대의 공공도서관은 지역사회의 핵심 역할을 하는 사회·경제·문화 인프라 시설로, 지역 주민들의 문화 및 여가 활동을 지원하는 복합문화공간으로서 그 역할과 기능이 확대되고 있다(고재민, 임호균, 2012; 배동식, 2022). 이와 같은 변화는 공공도서관 이용 증가에서도 확인된다. 2024년 공공도서관 기준 총대출 도서 수는 전년 대비 5.3% 증가한 1억 4,674만 권을 기록하였으며, 도서관당 평균 대출 도서 수도 전년 대비 3.3% 증가한 11만 3,227권에 달하는 것으로 나타났다(문화체육관광부, 2025).

한편 도서 대출이 증가함에 따라 연체 건수도 함께 증가하면서 연체 도서 회수 업무가 공공도서관 업무의 한 부분으로 자리 잡았다. 도서 연체의 증가는 다른 이용자의 자료 이용 기회를 제한할 뿐만 아니라, 연체자 관리를 위한 업무의 비중도 증가되어 도서관 사서들이 이용자 서비스에 할애해야 할 시간과 에너지가 부족해지는 부작용으로 이어진다. 특히 공공도서관은 불특정 다수의 이용자를 대상으로 서비스를 제공하기 때문에 대출 도서를 회수하는 데 많은 어려움을 겪는 것으로 보고되고 있다(박미환, 2009; 김종우, 2024).

이러한 연체 문제에 대응하기 위해 국내 공공도서관은 다양한 방지 대책을 운영하고 있다. 도서 연체 방지 대책은 반납예정일 전후로 문자메시지 등을 일괄 발송하는 예방적 대책과, 도서 연체 시 대출을 제한하거나 연체료를 부과하는 징벌적 대책으로 구분해볼 수 있다. 이중 예방적 대책은 반납예정일을 기준으로 모든 대출자에게 동일한 절차를 적용하고 있어, 고

위험 연체 이용자의 사전식별·개입과 같은 이용자 특성을 반영한 맞춤형 대응에는 한계가 있다. 따라서 연체 도서 증가에 효과적으로 대응하기 위해 도서관 이용자의 행동 패턴에 기반한 차등화된 예방적 연체 관리 방안의 검토가 요구된다.

하지만 도서관 이용자 행동 패턴 분석에 대한 기존 연구(이태호, 2019; 오민기 외, 2021; 남재우, 최원태, 2021; 박희진, 박성재, 2023; Zhang, 2023)는 기술통계, 현황 분석 수준에 그치거나 수요 예측, 맞춤형 도서 추천 등 서비스 개선에 중점을 두고 있다. 또한, 도서 연체에 대한 기존 연구는 현황 분석 수준의 결과를 제시하거나 징벌적 대책의 관점에서 국내외 도서관의 연체 제도에 대한 문헌적 분석에 중점을 두고 있다(백승욱, 2002; 박인호, 김일환, 2007; 윤희윤, 2020; 백선웅, 2024). 한편 이용자 행동 패턴과 연체 발생의 관계에 관한 기존 연구(Sung & Tolppanen, 2013; Liu, 2024)는 연체 발생 패턴이 이용자 집단별로 상이하게 나타남을 보였으나, 원인분석 수준에 머물러 있어 예방적 연체 방지 대책으로의 연계를 위한 추가 연구가 필요한 실정이다.

이에 본 연구는 도서 연체 방지를 위한 예방적 관리 전략에 중점을 두고, 공공도서관 이용자의 유형 세분화에 기반한 연체 발생 예측 모델의 적용 가능성에 대해 고찰한다. 구체적으로 국내 C공공도서관에서 수집된 실제 대출·반납 데이터를 기반으로 LightGBM 등 다양한 머신러닝 기법을 적용하여 대출 시점의 연체 발생 예측 모델을 학습하고, K-Prototypes 군집 분석을 통해 대출·연체 유형별로 이용자 집단을 세분화한 후, 유형별로 예측 모델 임계값

을 최적화한다. 이후 차등화된 임계값의 적용이 연체 발생 예측 성능에 미치는 영향을 검증하여 이용자 유형별 맞춤형 연체 관리 전략의 적용 가능성을 제시한다. 이를 통해 공공도서관의 행정 자원 효율화 및 한정된 장서의 순환을 극대화를 위한 시사점을 도출하고자 한다.

2. 선행 연구

2.1 도서관 이용자 행동 패턴 분석

도서관 이용자 행동 패턴 분석에 대한 기존 연구는 주로 도서 대출의 통계적 패턴 분석이나 대출 수요에 영향을 미치는 요인 분석, 그리고 행동 패턴에 기반한 맞춤형 서비스 제안을 중심으로 수행되어 온 것으로 확인된다.

먼저, 도서 대출의 통계적 패턴을 분석한 연구의 사례로 이태호(2019)는 대출 도서의 반납 행동 패턴을 확률분포와 큐잉 모델(Queuing Model)을 기반으로 분석하여, 연체 전후 이용자의 반납 행동 양식이 구분됨을 실증하였다. 비지도 학습 기반 기계학습 기법을 적용한 연구의 사례로 Larasati et al.(2019)은 대학도서관 이용자의 이용 동기 등을 포함한 설문조사 데이터를 기반으로 K-means와 X-means 군집 분석을 수행하여 이용자 유형을 분석한 바 있으며, Monika et al.(2025)은 실제 도서관 대출 거래 데이터를 기반으로 K-means를 포함한 다양한 군집 분석 기법을 적용하여 이용 이력을 중심으로 이용자를 유형화하였다.

대출 수요에 영향을 미치는 요인 분석 및 맞춤형 서비스에 관한 연구로는 공공도서관 도서

대출 이력에 K-Means 군집 분석과 다중회귀 모형을 병행 적용한 연구(오민기 외, 2021), 공공도서관 종단(longitudinal) 데이터를 기반으로 분석을 수행한 연구(박희진, 박성재, 2023)가 있으며, 도서 대출 수요 예측을 위해 인공지능 경망(Sun, 2021)이나 유전 알고리즘(Zhang, 2024)을 적용한 연구도 확인된다. 이 밖에도 남재우와 최원태(2021)는 공공도서관 이용자 설문조사 데이터의 다중회귀분석 결과를 바탕으로 맞춤형 우대서비스 제공방안을 제시하였으며, Zhang(2023)은 대출·반납 이력 데이터에 K-means 군집 분석과 Apriori 연관규칙 분석을 적용하여 이용자 프로파일을 구축하고 이를 바탕으로 개인화 추천 방안을 제안하였다.

2.2 도서 연체 개선방안

도서 연체 개선에 대한 기존 연구는 크게 연체 현황의 통계적 분석 및 데이터 마이닝 기반의 연체 요인 도출 결과를 바탕으로 연체 관리 방안을 제안한 연구와, 국내·외 도서관 대출 및 연체 관리 규정에 대한 문헌 분석 연구로 구분된다.

먼저, 연체 현황의 통계적 분석을 기반으로 한 연구의 사례로 박인호와 김일환(2007)은 대학도서관의 대출 및 연체 통계를 계열별·신분별로 분석하여 연체의 주요 원인을 식별하고, 이를 바탕으로 연체 관리 개선 방안을 제시하였다. 백승옥(2002)은 대학도서관의 대출·반납 데이터에 데이터 마이닝을 적용하여 연체 요인을 도출하고 고객관계관리(CRM) 모형을 설계하였으며, 대출 횟수와 연체율 데이터에 군집 분석을 적용하여 우수 고객을 선정하고 이들

을 대상으로 한 차별화된 서비스 전략을 제시하였다.

국내의 도서관의 대출 및 연체 관리규정의 문헌적 분석 연구의 사례로 윤희윤(2020)은 전국 공공도서관의 대출규정과 운영 현황을 대상으로 연체 제재 방식과 대출서비스 운영 특성을 비교·분석하여, 도서관 연체 관리 정책 및 제도의 개선방향을 제시하였다. 심효정과 송현경(2024)은 미국 주요 공공도서관과 서울시 공공도서관의 연체료 정책 및 운영 규정 자료를 수집하여 연체료 부과 방식과 연체 관리 정책을 비교·분석하고, 국내 공공도서관의 연체료 정책 개선방안을 제시하였다. 또한, 백선웅(2024)은 전국 173개 4년제 대학도서관의 대출규정과 운영 현황을 전수 조사하여 대출 자격, 대출 기간, 연체 제재 방식 등 대출서비스 운영 특성을 비교·분석하고, 이용자 중심의 대출 규정 개선방안을 제시하였다. 이 연구는 연체료 폐지 및 대출 기간 연장 이후 분실 자료가 감소한 사례를 근거로 연체료 규정의 전향적 개정을 제안하였으며, 연체 제재 방식보다 이메일·SNS를 통한 사전 안내 서비스가 실질적인 효과를 거두는 데 유리하다고 보았다.

2.3 이용자 행동 패턴과 연체 발생 관계

이용자 행동 패턴과 연체 발생의 관계에 관한 연구는 도서관 이용자의 집단별 특성에 따른 연체 발생 패턴의 차이를 통계적으로 분석한 연구와 이러한 차이가 발생하는 원인을 분석한 연구가 수행된 것으로 확인된다.

먼저, Liu(2024)는 연체 데이터에 경로 분석 및 구조 방정식 모형을 적용하여 누적 대출 횟

수, 자기조절 성향, 전공분야, 성별, 신분 등의 변수를 이용하여 연체 발생 빈도와 최대 연체 일수를 유의하게 예측 가능함을 실험적으로 검증하였다. Sung과 Tolppanen(2013)은 학부생, 대학원생, 교원 등으로 분류할 수 있는 이용자 집단의 특성에 따라 대출 도서의 반납 행동 패턴과 연체 경향에 뚜렷한 차이가 존재함을 정량적으로 확인하였다. Albergaria와 Lima(2022)는 사람의 시간적·심리적 행동성향과 연체 발생 간의 관계를 대출 도서의 반납 시점을 기준으로 분석하여 주말을 앞둔 금요일에 반납 지연이 급증하는 금요일 효과(Friday Effect)를 발견하였다. 한편, 타 분야의 유사 사례로 최봉진 외(2024)는 은행 대출 이용자의 거래 및 신용 이력 등 행동 패턴 데이터를 기반으로 대출 연체 예측 모델의 학습 및 적용 가능성을 제시하였다.

이상의 선행 연구를 종합하면, 이용자 행동 패턴 분석 연구는 실제 대출 이력 데이터를 활용한 정량적 분석을 시도하였다는 점에서 의의를 가지며, 도서 연체 개선방안에 관한 연구 역시 데이터 과학에 기반한 접근뿐 아니라 문헌 분석에 기반한 정책 연구도 활발히 수행되어, 도서 연체 문제 개선을 위한 다각적인 시도가 이루어져 왔음을 알 수 있다. 또한, 이용자 행동 패턴과 도서 연체의 관계에 관한 연구를 통해 이용자 유형과 같은 행동 패턴 변수가 연체 발생과 통계적으로 유의한 관계를 가짐을 확인할 수 있다.

그러나 이들 연구는 공통적으로 다음과 같은 한계를 지닌다. 머신러닝에 기반한 이용자 행동 패턴 분석 연구가 다수 시도되었으나 개인화 추천이나 수요 예측과 같은 서비스 개선 측면에 집중되었고, 연체 개선방안에 관한 연구

역시 기술통계 분석이나 제도·규정에 대한 문헌적 검토에 주로 머물렀다. 또한, 이용자 행동 패턴과 도서 연체의 관계에 관한 연구는 연체 발생 이후의 원인과 경향을 설명하는 데 초점을 두고 있어 개별 이용자 수준에서 대출 시점의 연체 발생 여부를 사전에 예측하는 단계까지는 나아가지 못하였다. 최근 AI 기술 발전에 따라 도서관 관리 전반에 데이터 분석 기반의 업무 프로세스를 적용하는 방향으로 연구 흐름이 전환되고 있는 시점에서(Nova et al., 2025; Islam et al., 2026) 도서관 운영상 지속적으로 제기되어 온 도서 연체 문제에 머신러닝 기반 기법을 적용하여 개선 방안을 탐색하는 것 그 자체로 의의를 지닌다고 할 수 있다.

따라서 본 연구는 이러한 선행 연구의 공백을 보완하고자, 국내 C공공도서관의 실제 대출·반납 데이터를 기반으로 대출 시점의 연체 발생 예측 모델을 학습하고, K-Prototypes 군집 분석을 통해 이용자 유형을 세분화한 후 유형별로 연체 예측 모델의 임계값을 최적화하여 예측 성능의 향상 여부를 검증한다. 이러한 결과를 바탕으로 효과적인 도서관 운영을 위한 이용자 유형별 맞춤형 연체 관리 전략을 제시하고자 한다.

3. 연구 방법

본 연구는 도서관 이용자의 대출 거래 데이터로 학습된 연체 발생 예측 모델과 군집 분석으로 세분화된 이용자 유형을 바탕으로, 각 유형별 모델 최적 임계값이 예측 성능에 미치는 영향을 실험적으로 확인하여 이용자 유형 세분

화에 기반한 맞춤형 연체 관리 전략의 적용 가능성을 검증하는 것을 목적으로 한다. 이를 위해 국내 C공공도서관 성인 이용자의 대출·반납 및 연체 데이터를 활용한다.

머신러닝 적용은 대출 시점의 연체 발생 여부를 예측하는 모델을 학습하는 단계와 이용자 유형을 대출 패턴에 따라 세분화하는 군집 분석 단계, 그리고 유형에 따라 예측 임계값을 조정하는 단계로 구분되며 각 단계는 목적에 맞게 서로 다른 분석 단위를 적용한다. 즉, 연체 발생 예측은 대출 시점에서 해당 대출 건의 연체 발생 여부를 판단하는 것을 목적으로 하므로, 개별 대출 건 단위로 수행되며, 이를 위해 대출 건별로 수집·생성된 데이터를 활용한다. 반면 군집 분석은 한 이용자가 일관된 하나의 이용자 유형에 속해야 한다는 목적에 부합하도록 이용자 단위로 수행하며, 이를 위해 개별 이용자의 누적된 행동 패턴을 요약한 데이터를 활용한다. 각 데이터는 앞서 언급한 국내 C공공도서관의 데이터를 원천 데이터로 하여 구성한다.

전체 연구 절차는 데이터 수집 및 전처리, 주요 변수에 대한 탐색적 분석, 군집 분석과 예측 모델 구축, 분석 결과 연계 해석의 네 단계로 설계하였다. 군집 분석 및 예측 모델 학습의 모든 과정은 Google Colab 환경에서 Python을 사용하여 수행되었다.

3.1 데이터 수집 및 전처리

데이터 수집은 국내 C공공도서관 이용자 중 2024년 1월부터 2026년 3월까지 대출 및 반납 실적이 있는 성인 이용자(2,907명)를 대상으로 하였다. 보호자 대리 대출이 빈번한 어린이 이

용자와 가입 회원 수 및 대출 이력의 모수가 적은 청소년 이용자의 대출·반납 실적은 데이터 왜곡 방지 및 통계적 유의성 확보를 위해 분석 대상에서 제외하였고, 수집 단계에서부터 이용자 성명 등 개인을 직접 식별할 수 있는 정보를 제거하는 등 비식별화 처리하였다.

분석에 활용된 데이터는 <표 1>과 같이 이용자의 기본 정보와 대출·반납 이력 그리고 연체 이력 등 3종의 데이터로 구성되며, 이용자 ID, 도서등록번호, 반납일(연체처리일)을 기준으로 결합하여 원천 데이터를 구성하였다. 결합 과정에서 데이터 불일치 또는 이상치를 보정·제거하였고, 분석 대상을 C공공도서관 이용자로 한정하기 위해, 전국 공공도서관 통합 대출 서비스인 책이음을 통한 대출실적이 있는 타 도서관 회원 데이터는 제외하였다. <표 2>는 전처리가 완료된 원천 데이터의 현황을 보인다. 본 연구에서 적용하는 연체율은 이용자 단위가 아닌 대출 건 단위의 연체율로, 데이터 수집 기간 동안 반납이 완료된 전체 대출 건 중 반납예정일을 초과하여 반납된 건의 비율을 의미한다.

따라서 한 이용자가 여러 권을 대출한 후 그 중 일부만 연체할 경우 연체가 발생한 개별 대출 건만이 연체 건으로 집계된다. 이와 달리 통상적으로 적용되는 연체율 산정방식은 특정 시점을 기준으로 그 시점에 대출 중인 도서 중 반납 기한을 초과한 미반납 도서의 비율을 측정하는 방식이다. 이 경우 이용자가 연체 도서를 반납하는 즉시 해당 건이 집계 대상에서 제외되므로, 본 연구에서 적용하는 연체율보다 낮은 수치를 보이는 경향이 있다. 다만 본 연구의 목적은 대출 건별 연체 예측 모델의 성능을 실험적으로 검증하는 데 있으므로, 예측 모델의 분석 단위인 대출 건과 일치하는 방식인 대출 건 단위의 연체율 산정방식을 적용한다.

<표 3>은 원천 데이터의 세부 항목 중 군집 분석과 예측 모델 구축에 적용할 변수 목록을 보인다. 전체 항목 중 서명과 자료실은 분석에 직접 활용되지 않은 항목으로 제외하였으며 분석 목적과의 연계성을 고려하여 새로운 파생변수로 정의하였다. 각 변수는 탐색적 데이터 분석을 바탕으로 유효성을 검증한 후 선별적으로

〈표 1〉 데이터 수집 현황

데이터 명	건 수	세부 항목
이용자 도서 대출·반납 이력	83,990건	이용자 ID, 신분, 도서등록번호, 서명, 자료실, 청구기호, 대출일, 반납예정일, 반납유형, 반납일
이용자 정보	2,907명	이용자 ID, 신분, 등록일, 가족회원 ID
이용자 연체 이력	15,146건	이용자 ID, 신분, 도서등록번호, 서명, 자료실, 청구기호, 연체일수, 연체처리일

〈표 2〉 데이터 전처리 결과

구 분	전처리 전	전처리 후	반납비율	
대출 및 반납 실적	83,990건	75,377건	정상반납	61,813건
			연체반납	13,564건(연체율 18.0%)
이용자 수	2,907명	2,148명	-	

〈표 3〉 원천 데이터셋의 세부 항목에 기반한 변수(파생변수 포함) 현황

구 분	변 수 명	변수 설명	변수 타입
공 통	senior	고령층 여부(0:일반, 1:노인)	int
	kdc_main	KDC 대분류(0~9)	int
	family	가족회원 여부(0:개인, 1:가족)	int
	overdue_day	연체일수	int
대출 건 단위 파생변수 (시점 중심)	tenure_days	가입 후 경과 일수	float
	new_member	1년 미만 회원 여부(0:일반, 1:신규)	int
	cumulative_overdue	누적 연체 횟수	int
	borrow_interval	대출 주기(일)	float
	avg_borrow_interval	평균 대출 주기(일)	float
	daily_borrow_count	일 대출 총 권수	int
	return_duration	대출 후 반납 소요일	int
	borrow_month	대출 월(1~12)	int
이용자 단위 파생변수 (총 이용기간 중심)	tenure_days_max	가입 경과일(최댓값)	float
	total_borrow_count	총 대출 권수	int
	overdue_count	누적 연체 권수	int
	avg_return_duration	평균 반납 소요일	float
	avg_overdue_day	평균 연체일수	float
	visit_count	방문 횟수	int
	avg_daily_borrow	1회 평균 대출 권수	float
	avg_visit_interval	평균 방문 주기(일)	float

군집 분석 및 예측 모델 학습에 활용한다.

〈표 4〉는 학습·검증·평가 데이터 구성현황을 보인다. 연체 발생 예측 모델의 학습과 평가 과정에서 데이터 누수와 과적합을 방지하고 시간적 순서를 반영하기 위해, 전체 데이터를 대출일을 기준으로 학습·검증·평가 데이터셋으로 분리하였다. 한편, 군집 분석은 이용자 특성에 따른 유형을 세분화하는 데 목적이 있으므로 전체 대출 이력을 이용자 단위로 집계한 데

이터셋을 대상으로 수행하였다. 이는 이용자 세분화에 따른 연체 예측 성능의 차이를 확인하기 위해 이용자 행동 특성에 기반하여 이용자 유형을 도출한 후 각 유형에 대해 예측 모델을 최적화·평가하는 본 연구의 분석 설계에 따른 것이다. 이때 도출된 군집 레이블은 예측 모델의 입력 변수로 사용하지 않고, 임계값 최적화와 성능 평가 과정에서 대출 데이터를 이용자 유형별로 구분하기 위한 기준으로 활용한다.

〈표 4〉 학습·검증·평가 데이터 분리 현황

구 분	기 간	대출 반납 이력
학습 데이터	2024. 1. ~ 2025. 9.	60,391건(연체율 17.9%)
검증 데이터	2025. 10. ~ 2025. 12.	6,668건(연체율 18.4%)
평가 데이터	2026. 1. ~ 2026. 3.	8,318건(연체율 18.0%)

3.2 군집 분석 및 예측 모델 구축

3.2.1 변수 선정

원천 데이터셋의 변수를 기반으로 군집 분석 및 예측 모델 학습에 활용할 변수를 선정하기 위해 대출 건 단위 데이터와 이용자 단위 데이터를 대상으로 각 변수의 특성과 연체 발생 간의 관계를 통계적으로 검토하였다.

군집 분석에 적용되는 변수는 탐색적 분석에서 통계적으로 유의한 것으로 확인된 변수와 이용자 단위로 분석할 수 있는 데이터를 고려하여 세 가지 기준에 따라 선정하였다. 첫째, 특정 변수가 군집 결과를 지나치게 지배하는지를 기준으로 검토하였다. 특정 변수가 군집을 지나치게 지배하여 다른 변수의 정보가 묻히는 경우 해당 변수의 투입을 재검토하였다. 둘째, 군집 해석 가능성과 변수 간 정보 중복 여부를 기준으로 검토하였다. 군집이 형성되더라도 각 군집의 특성이 명확하게 구분되지 않거나 도서관 현장에서 활용하기 어려운 경우, 변수 간 정보가 중복되어 추가적인 설명력을 갖지 못하는 경우 해당 변수의 투입을 재검토하였다. 셋째, 변수와 연체 발생 간의 실질적인 관련성을 고려하였다. 통계적으로 유의하더라도 상관관계수가 매우 낮아 연체와의 실질적인 관련성이 낮은 변수는 제외하였다.

예측 모델 학습에 적용되는 변수는 탐색적 분석에서 연체 발생과 통계적으로 유의한 관련성이 확인된 변수를 후보로 하여, 다음의 기준을 종합적으로 고려하여 선정하였다. 첫째, 변수가 대출 시점에 확인 가능한 정보인지를 고려하였다. 예측 모델은 대출이 이루어지는 시점에서 연체 발생 여부를 사전에 판단하는 것을 목적으

로 하므로, 반납 시점 또는 그 이후에 결정되는 변수는 모델 학습에 미래 정보가 유입되는 데이터 누수를 유발할 수 있어 후보에서 제외하였다. 둘째, 변수가 반영하는 정보의 층위가 중복되지 않고 다양한 차원에서 연체 위험을 설명할 수 있는지를 고려하였다. 이용자의 인구 통계적 속성, 가입 기간 및 과거 연체 이력, 대출 빈도 및 이용 패턴, 도서 특성이 균형 있게 포함되도록 하였으며, 동일한 정보를 중복적으로 담고 있어 추가적인 설명력이 낮은 변수는 제외하였다. 셋째, 변수가 도서관 현장에서 해석 가능하고 운영 시 활용될 수 있는지를 고려하였다. 통계적으로 유의하더라도 그 의미를 현장에서 해석하거나 정책적으로 적용하기 어려운 변수는 최종 변수에서 제외하였다.

3.2.2 군집 분석 방법 선정

K-means 군집 분석(Lloyd, 1982)은 데이터에서 패턴을 파악하기 위한 비지도학습 기법으로 널리 활용되며 도서관 이용자의 행동 패턴에 군집 분석을 적용한 기존 연구(오민기 외, 2021; Zhang, 2023; Monika et al., 2025)에서도 적용되었다. 그러나 본 연구의 경우 군집 분석에 적용하는 변수에 이진 변수와 연속형 변수가 혼재하기 때문에 K-means 군집 분석을 그대로 적용할 경우, 이진 변수의 값을 기준으로 군집이 단순하게 분리되어 결과가 왜곡될 가능성이 있다. 따라서 본 연구에서는 두 가지 변수 유형을 동시에 처리할 수 있는 K-Prototypes 군집 분석 기법(Huang, 1998; 김정화, 2018)을 적용한다. K-Prototypes 군집 분석은 이진 변수는 해밍 거리(Hamming Distance)로, 연속형 변수는 유클리드 거리(Euclidean Distance)로

각각 계산하여 적용하므로 K-means에서 발생할 수 있는 결과 왜곡을 방지할 수 있다는 장점이 있다. 또한, 변수 간 단위 차이로 인해 군집 결과가 특정 변수로 편향되는 현상을 방지하고 이진 변수와의 크기 균형을 맞추기 위해, 연속형 변수는 최댓값과 최솟값을 기준으로 0~1 범위로 정규화하는 최소-최대 정규화(Min-Max Normalization)를 적용하였다.

최적의 군집 수(K)는 K-Prototypes의 비용 함수(Cost) 감소량과 실루엣 점수(Silhouette Score)를 기준으로 결정하였다. 구체적으로 비용 함수가 급격히 감소하다가 완만해지는 엘보우 지점과 실루엣 점수가 높은 지점을 종합적으로 확인하였으며, 두 지표의 기준 외에도 각 군집 수에서의 군집별 특성을 함께 비교하여 최종 군집 수를 결정하였다.

3.2.3 예측 모델 학습 방법 선정

예측 모델은 대출 시점에서 연체 발생 여부를 예측하는 이진 분류 모델을 학습하는 것을 목적으로 한다. 학습 기법은 XGBoost(Chen & Guestrin, 2016; 최봉진 외, 2024), LightGBM(Ke et al., 2017), Logistic Regression을 각각 적용하여 성능을 비교하였다. XGBoost는 여러 개의 약한 학습기(결정 트리)를 순차적으로 학습시키면서 이전 모델이 틀린 부분을 다음 모델이 보완하는 부스팅 방식을 사용하며, 정규화가 내장되어 과적합을 효과적으로 방지한다. LightGBM 역시 부스팅 기반 알고리즘이지만, 리프 중심 트리 분할(Leaf-wise) 방식을 적용하여 XGBoost에 비해 학습 속도가 빠르고 메모리 사용량이 적다는 장점이 있다. Logistic Regression은 데이터가 특정 클래스에 속할 확

률을 시그모이드(sigmoid) 함수를 통해 0과 1 사이의 값으로 예측하는 선형 모델로, 구조가 직관적이고 연산 속도가 빠르며 해석 가능성이 높다는 특징이 있다.

한편 분석 대상 데이터는 정상반납(82%) 대비 연체반납(18%) 클래스의 비율이 낮은 클래스 불균형을 보인다. 클래스 불균형이 존재하는 상태에서 모델을 학습할 경우, 손실 함수가 다수 클래스인 정상반납 위주로 최소화되어 소수 클래스인 연체를 충분히 학습하지 못하는 문제가 발생할 수 있다. 이를 완화하기 위해 본 연구에서는 세 알고리즘의 Python 구현체에 공통으로 내장된 클래스 가중치 조정 기능을 활용하여, 손실 함수 계산 시 소수 클래스(연체)에 더 높은 가중치를 부여함으로써 모델이 연체 클래스의 패턴을 충분히 학습하도록 유도하였다.

또한, 알고리즘의 구조적 특성을 고려하여 모델별 입력 데이터셋을 차별화하였다. XGBoost와 LightGBM 같은 트리 기반 모델은 변수의 척도나 형태에 영향을 받지 않는 특성을 고려하여 별도의 전처리를 거치지 않은 데이터를 활용하였다. 반면, 선형 모델인 Logistic Regression은 모델의 정교한 학습과 해석의 왜곡을 방지하기 위해 연속형 변수에는 최소-최대 정규화를, 범주형 변수에는 원-핫 인코딩(One-Hot Encoding)을 적용한 데이터를 활용하였다.

예측 모델의 성능 평가는 재현율(Recall)을 주요 지표로 하고 정확도(Accuracy), 정밀도(Precision), F1-Score, AUC-ROC를 보조 지표로 함께 활용하였다. 이는 본 연구의 예측 모델이 실무자의 2차적 개입(예: 반납 독려 SMS 발송 등)을 돕기 위한 1차 필터링 역할을 수행

하기 때문이며 예측 결과에 따른 후속 조치가 징벌적 성격이 아닌 예방적 안내 서비스이므로, 도서 연체를 방지하여 발생하는 장서 회전을 저하(False Negative)의 손실이, 일부 정상 반납자에게 안내가 오발송되는(False Positive) 비용보다 크다고 판단했기 때문이다.

예측 모델의 분류를 위한 임계값은 예측 모델 학습 방법 선정 단계에서는 기본 임계값인 0.5를 적용하여 재현율을 비교한 이후 선정된 학습 기법에 대해 유형별 임계값 최적화를 적용한다. 임계값 최적화는 검증 데이터에서 수행하며, 선정된 임계값을 적용하여 평가 데이터에서 측정한 성능을 최종 성능으로 보고한다. 또한, 임계값을 낮출수록 재현율이 높아지는 반면 정밀도는 낮아지는 트레이드오프가 발생하므로, F1-Score를 함께 고려하여 유형별 최적 임계값을 탐색하였다.

4. 성능 평가 및 결과 분석

4.1 기술통계 및 탐색적 분석 결과

〈표 5〉는 대출 건 단위 데이터의 범주형 변수 기술통계를 보인다. 대출 및 반납 건수를 기준으로 senior(고령층) 이용자는 전체의 16.3%, family(가족회원)는 34.3%로 나타나 전체 대출 건수의 3분의 1 이상이 가족회원에 의해 발생하였으며, new_member(신규회원)는 0.3%에 불과하여 대부분이 기존 이용자로 구성되어 있음을 알 수 있다. 따라서 세 변수 중 극단적으로 편중된 new_member를 제외하고 senior, family만 군집 분석 및 예측 모델 학습에 활용한다. KDC 대분류별로는 문학(44.0%), 사회과학(14.1%), 기술과학·철학(8.3%) 순으로 나타나 도서 분야에 따른 대출 집중도의 차이가 뚜렷하게 나타났다.

〈표 5〉 대출 건 단위 데이터 주요 범주형 변수 기술통계

변 수	구 분	대출 및 반납 건수	비 율
senior	일반(0)	63,071건	83.7%
	노인(1)	12,306건	16.3%
family	개인(0)	49,537건	65.7%
	가족(1)	25,840건	34.3%
new_member	기존(0)	75,132건	99.7%
	신규(1)	245건	0.3%
kdc_main	문학(8)	33,137건	44.0%
	사회과학(3)	10,619건	14.1%
	기술과학(5)	6,265건	8.3%
	철학(1)	6,252건	8.3%
	역사(9)	4,737건	6.3%
	자연과학(4)	4,237건	5.6%
	언어(7)	3,722건	4.9%
	종교(2)	2,264건	3.0%
	예술(6)	2,226건	3.0%
	총류(0)	1,918건	2.5%

〈표 6〉은 대출 건 단위 데이터의 연속형 변수 기술통계를 보인다. overdue_day(연체일수)는 중앙값이 0으로 대부분의 이용자가 정상반납하고 있으나, 일부 장기 연체사례(534일)가 존재한다. cumulative_overdue(누적 연체 횟수, 평균 7.93, 표준편차 13.77)와 avg_borrow_interval(평균 대출 주기, 평균 8.16, 표준편차 17.00)이 평균보다 표준편차가 큰 형태로 이용 빈도의 편차가 크게 나타나는 것을 알 수 있다. return_duration(대출 후 반납 소요일)은 평균 12.53일로 대출 기간(14일) 내에 반납하는 패턴이 일반적임을 알 수 있다.

〈표 7〉은 이용자 단위 데이터의 연속형 변수 기술통계를 보인다. total_borrow_count(총 대출 권수)는 평균 35.09권, 중앙값 12권으로 소수의 다독 이용자가 평균을 크게 끌어올리는

우편향 분포를 보이며, overdue_count(누적 연체 권수) 역시 평균 6.31권, 중앙값 2권, 최댓값 120권으로 일부 이용자에게 연체가 집중되는 편중된 분포를 나타낸다. avg_visit_interval(평균 방문주기)도 사분위수 범위가 7.00~61.20일로 넓게 분포하여 이용 빈도 면에서 이용자 집단이 이질적임을 보여준다. 이러한 이용자 단위 데이터 변수의 기술적 통계는 각 변수가 단일 평균값으로 이용자 특성을 대표하기 어려움을 의미하며, 행동 패턴에 따라 유형 구분이 필요함을 시사한다.

4.2 군집 분석 결과

4.2.1 군집 분석 변수 선정

군집 분석의 변수는 4.1절에서 통계적으로

〈표 6〉 대출 건 단위 데이터 주요 연속형 변수 기술통계

변 수	평균	표준편차	최솟값	25%	중앙값	75%	최댓값
tenure_days	2625.12	2573.99	0.00	881.00	1648.00	3448.00	10788.00
overdue_day	1.30	5.54	0.00	0.00	0.00	0.00	534.00
cumulative_overdue	7.93	13.77	0.00	0.00	2.00	10.00	120.00
avg_borrow_interval	8.16	17.00	0.00	1.94	4.00	8.20	613.00
daily_borrow_count	5.16	3.18	1.00	3.00	5.00	7.00	21.00
return_duration	12.53	7.74	0.00	8.00	13.00	14.00	548.00

〈표 7〉 이용자 단위 데이터 주요 연속형 변수 기술통계

변 수	평균	표준편차	최솟값	25%	중앙값	75%	최댓값
tenure_days_max	1761.77	2098.91	0.0	567.00	1028.00	1994.00	10788.00
total_borrow_count	35.09	65.48	1.0	4.00	12.00	36.00	911.00
overdue_count	6.31	11.84	0.0	0.00	2.00	7.00	120.00
avg_return_duration	14.50	10.39	0.0	10.71	13.16	15.33	192.67
avg_overdue_day	2.76	9.00	0.0	0.00	0.30	2.08	178.67
visit_count	11.02	17.28	1.0	2.00	5.00	13.00	214.00
avg_daily_borrow	2.96	1.97	1.0	1.54	2.33	3.83	15.00
avg_visit_interval	51.64	80.92	0.0	7.00	24.46	61.20	758.00

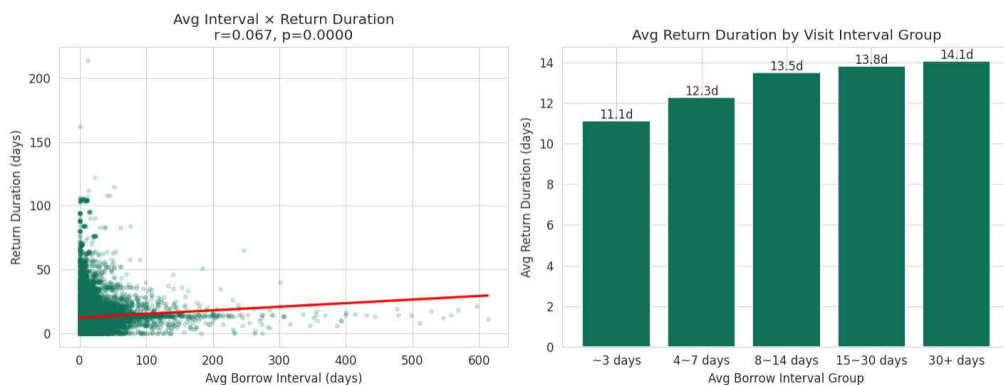
유의한 것으로 확인된 변수와 이용자 단위 데이터의 변수 총 10개를 대상으로 3.2.1절의 기준을 고려하여 <표 8>과 같이 5개의 변수를 선정하였다.

먼저, avg_overdue_day는 연체가 발생한 대출 건에만 값이 존재하고 정상반납 이용자는 0에 수렴하기 때문에 군집 결과에 편향을 제공하는 변수로 판단하여 제외하였다. total_borrow_count와 visit_count는 avg_daily_borrow와 정보가 중복되는 것으로 판단하여 제외하였다. avg_return_duration은 정상반납(0~14일)과 연체반납(15

일 이상) 정보가 모두 포함되어 산출되는 값으로, 정상과 연체 두 집단의 값이 혼재되므로 적용하지 않았다. avg_borrow_interval은 실제 반납 소요일과의 상관관계를 확인한 결과 <그림 1>과 같이 연체와의 상관계수($r=0.067$)가 낮게 측정되어 군집 분석 적용에서 제외하였다. 다만, 이는 단변량 분석의 한계일 수 있으므로, 변수 간의 다차원적 상호작용과 복합적 패턴을 자율적으로 학습하는 4.3절의 다변량 예측 모델에는 입력 변수로 유지하여 예측력을 보완할 수 있도록 설계하였다.

<표 8> 군집 분석 데이터셋 변수 선정 현황

변 수 명	변 수 설명	변 수 유형	선정여부	제외 사유
senior	고령층 여부(0:일반, 1:노인)	이진(0/1)	○	
family	가족회원 여부(0:개인, 1:가족)	이진(0/1)	○	
tenure_days_max	가입 경과일(최대값)	연속형	○	
total_borrow_count	총 대출 권수	연속형	×	중복 변수(avg_daily_borrow)
overdue_count	누적 연체 권수	연속형	○	
avg_return_duration	평균 반납 소요일	연속형	×	정상/연체 정보 중복
avg_overdue_day	평균 연체일수	연속형	×	분포 쏠림
visit_count	방문 횟수	연속형	×	중복 변수(avg_daily_borrow)
avg_daily_borrow	1회 평균 대출 권수	연속형	○	
avg_borrow_interval	평균 대출 주기(일)	연속형	×	연체 발생과 통계적 유의성 낮음



<그림 1> avg_borrow_interval과 실제 반납 소요일과의 상관관계 분석 결과

4.2.2 최적 군집 선정

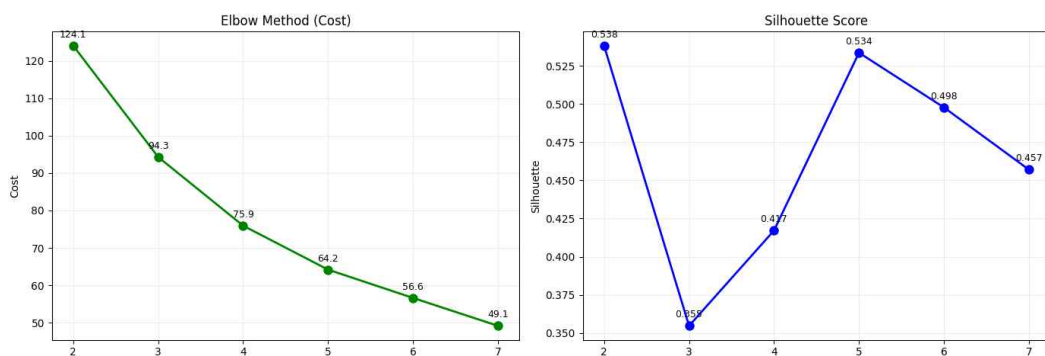
최적 군집 수를 선정하기 위해 통계적 지표 기반 후보 도출 후, 이용자 특성 및 연체 위험 변별력을 검토하였다. 〈그림 2〉는 군집 수(K)에 따른 손실함수(Cost) 감소량과 실루엣 점수를 보인다. 최적 군집 선정에 위한 탐색 범위는 $K=[2, 7]$ 로 설정하였다.

먼저, cost의 경우 K가 증가함에 따라 지속적인 감소 추세를 나타냈다. 탐색 범위 내에서 시각적으로 뚜렷한 엘보우 지점은 관찰되지 않았으나, 산술적으로 $K=4$ 이후로는 -11.7, -7.6, -7.5로 감소폭이 완만해지는 추세를 확인할 수 있었다. 실루엣 점수의 경우 $K=2$ 에서 가장 높은 값(0.538)을 보였고 $K=3$ 에서 급락(0.355)한 뒤, 다시 $K=5$ 에서 최댓값 수준으로 회복(0.534)하는 패턴을 보이고 이후 완만하게 감소하였다. 이러한 패턴을 바탕으로 cost 감소폭이 비교적 완만해지는 구간($K \geq 4$)과 실루엣 점수가 우수한 지점을 결합하여 $K=2, 5, 6$ 을 최적 군집 수의 비교 후보로 도출하였다.

비교 후보로 도출된 K에서 최적 군집을 선정하기 위해 연구의 목적, 군집별 이용자 특성

및 연체 위험 수준의 차이를 추가로 검토하였다. 먼저, $K=2$ 는 실루엣 점수(0.538)가 후보 중 가장 높았지만 두 번째로 높은 점수를 보인 $K=5$ 와의 차이가 크지 않았다. 또한, 본 연구가 이용자의 연체 위험 수준과 이용 특성을 함께 고려한 유형화를 목적으로 한다는 점에서, 두 집단으로 유형 간 특성 차이를 충분히 구분하기 어렵다고 판단하여 최적 군집 수 후보에서 우선 제외하였다. 나머지 후보인 $K=5$ 와 $K=6$ 의 비교를 위해 분할된 군집별 통계적 특성과 대출 건 단위 연체율을 〈표 9〉와 같이 산출하였다.

$K=5$ 의 경우 군집별 연체율이 5%~25% 범위에서 고위험(25%, 21%)-중위험(19%, 16%)-저위험(5%)의 3개 수준으로 뚜렷이 구분됨을 관찰할 수 있고, 특히 동일한 연체 위험 수준에서도 등록 기간(군집1 vs 군집3) 및 대출 권수(군집0 vs 군집4) 등에서 군집별로 차이가 나타남을 확인할 수 있다. $K=6$ 의 경우, $K=5$ 의 경우와 유사하게 군집별 연체율이 3개의 위험 수준으로 구분되었으나 군집4와 군집5는 연체율 차이가 크지 않아 유형 간 변별력이 높지 않다. 또한, 실루엣 점수도 0.498로 $K=5$ (0.534)보다



〈그림 2〉 군집 수에 따른 손실함수와 실루엣(Silhouette) 점수의 변화

〈표 9〉 K-prototypes 적응결과 군집별 기술통계

군집수	군집	이용자 수	연체율	누적 연체 권수	평균 연체일수	총 대출 권수	회당 대출 권수	등록 기간	가족 비율	고령층 비율
K=5	0	273	0.19	12.22	2.38	66.12	6.70	1086.16	0.10	0.04
	1	291	0.25	13.15	3.06	52.47	3.57	1914.37	0.99	0.02
	2	161	0.05	1.84	0.54	38.40	2.02	1820.71	0.02	1.00
	3	156	0.21	13.88	3.68	67.06	3.15	8035.45	0.82	0.10
	4	1,267	0.16	3.11	2.95	20.06	2.11	1092.36	0.00	0.00
K=6	0	257	0.19	11.60	2.33	62.60	6.77	929.80	0.10	0.04
	1	291	0.25	13.54	3.10	53.67	3.64	1884.18	1.00	0.02
	2	149	0.05	1.79	0.57	34.93	1.95	1556.64	0.02	1.00
	3	146	0.21	14.21	3.85	67.68	3.20	8214.54	0.83	0.09
	4	1,165	0.15	2.88	2.98	18.77	2.09	867.83	0.00	0.00
	5	140	0.14	6.77	2.29	48.01	2.58	3962.46	0.07	0.10

낮아 K=6으로의 추가 세분화 없이도 K=5만으로 연체 위험 수준에 따른 이용자 유형을 충분히 세분화할 수 있다고 판단하여 K=5를 군집 수로 선정하였다.

4.2.3 군집별 특성 분석

〈표 10〉은 각 이용자 군집별 특성을 분석한 내용이다. 각 군집은 특성에 따라 일반 다독형(General Heavy Reader), 가족회원 연체형(Family Overdue), 고령층 우수형(Senior Excellent), 장기가족회원 연체형(Long-term Family Overdue), 일반 표준형(General Standard)으로 구분하여 분석하였다.

4.3 예측 모델 학습 결과

4.3.1 예측 모델 학습 변수 선정

예측 모델 학습의 변수는 4.1절에서 통계적으로 유의한 것으로 확인된 변수와 대출 건 단위 데이터의 변수 총 11개를 대상으로 3.2.1절의 기준을 고려하여 〈표 11〉과 같이 9개의 변수

를 선정하였다. 제외된 `overdue_day`와 `return_duration`은 모두 대출 시점에서는 알 수 없는 정보이므로 제외하였다. 학습에 적용되는 데이터 세부 현황은 〈표 12〉와 같다.

4.3.2 학습 기법 선정

〈표 13〉은 기본 임계값을 적용했을 때 학습 기법별 예측 모델의 성능을 보인다. XGBoost와 LightGBM은 각각 재현율 91.2%, 92.3%를 보여 연체 발생 건의 90% 이상을 대출 시점에 예측할 수 있는 수준으로 나타났으며, AUC-ROC도 모델이 연체와 정상반납을 수용 가능한 수준으로 변별할 수 있음을 알 수 있다. 실험 결과에 따라 비교 대상 학습 기법 중 본 연구의 핵심 지표인 재현율에서 가장 우위를 보인 LightGBM에 대해서만 임계값 최적화를 적용하여 분석한다.

4.3.3 예측 모델 임계값 유형별 최적화

〈표 14〉는 LightGBM 모델에 다양한 기준의 임계값을 적용했을 때 이용자 유형별 연체

〈표 10〉 군집별 특성 분석

군집수	군집	이용자 수	연체율	군집 특성
K=5	0	273	19%	• 일반 다독형(General Heavy Reader) - 가족회원(10.0%), 고령층(4.0%) 비율이 낮은 일반 개인회원 중심 - 등록 기간 1,086일로 가장 짧은 최근 유입된 적극적 이용자층 - 회당 대출 권수가 6.70권으로 다섯 군집 중 가장 많은 대량 대출 집단 - 누적 연체 권수(12,22권)는 다소 높으나 이는 대출량 자체가 많기 때문이며, 연체율(19.0%)은 전체 군집 중 중간 수준으로 나타남
	1	291	25%	• 가족회원 연체형(Family Overdue) - 가족회원(99.0%), 고령층(2.0%) 비율로 일반 연령대 가족 중심 - 연체율(25%)이 군집 중 가장 높고, 누적 연체 권수(13,15권)는 두 번째로 높음. - 등록 기간(1,914일)이 길지 않음에도 연체가 빈번하여 이용 경력과 무관하게 선제적 관리가 필요한 고위험 군집
	2	161	5%	• 고령층 우수형(Senior Excellent) - 고령층 비율이 100.0%로 구성된 은퇴 세대 위주의 개인회원 중심 - 연체율(5.0%), 평균 연체일수(0.54일), 누적 연체 권수(1,84권) 모두 전체 군집 중 가장 낮아 반납 기한을 성실하게 준수하는 이용자층 - 회당 대출 권수가 2.02권으로 소량을 규칙적으로 대출하는 이용 패턴 - 등록 기간은 1,820일로 비교적 안정적인 이용 이력을 가진 그룹
	3	156	21%	• 장기 가족회원 연체형(Long-term Family Overdue) - 가족회원(82.0%)과 고령층(10.0%)이 공존하는 구성 - 등록 기간이 8,035일(약 22년)로 다섯 군집 중 압도적으로 길어 지역사회에 충성 이용자층으로 분류 - 총 대출 권수(67,06권), 누적 연체 권수(13,88권), 평균 연체일수(3.68일)가 모두 최고 수준으로 오랜 이용 이력과 함께 대출량 및 누적 연체 권수가 높은 특성을 보임
	4	1267	16%	• 일반 표준형(General Standard) - 가족회원과 고령층이 없는 순수 개인회원 중심 - 전체의 약 59%를 차지하는 가장 큰 군집 - 회당 대출 권수 2.11권의 소독(小讀) 패턴으로 대출량이 적어 연체 누적도 적은 평균적 이용자 그룹 - 등록 기간 1,092일로 비교적 최근 가입자 비중이 높음

〈표 11〉 예측 모델 학습 데이터셋 변수 선정 현황

변 수 명	변 수 설명	변 수 유형	선정여부	제외 사유
senior	고령층 여부(0:일반, 1:노인)	이진(0/1)	○	
family	가족회원 여부(0:개인, 1:가족)	이진(0/1)	○	
kdc_main	KDC 대분류(0~9)	범주형	○	
overdue_day	연체일수	연속형	×	대출 시점에서 확인 불가 정보
tenure_days	가입 후 경과 일수	연속형	○	
cumulative_overdue	누적 연체 횟수	연속형	○	
borrow_interval	대출 주기(일)	연속형	○	
avg_borrow_interval	평균 대출 주기(일)	연속형	○	
daily_borrow_count	일 대출 총 권수	연속형	○	
return_duration	대출 후 반납 소요일	연속형	×	대출 시점에서 확인 불가 정보
borrow_month	대출 월(1~12)	연속형	○	

〈표 12〉 이용자 유형별(K=5) 학습/검증/평가 데이터 통계

유 형	학습 데이터				검증 데이터	평가 데이터
	총 계	정상반납	연체	연체율		
가족회원 연체형	12,186	9,126	3,060	25%	1,332	1,752
장기 가족회원 연체형	8,678	6,919	1,759	21%	874	909
일반 표준형	20,069	16,946	3,123	16%	2,244	3,100
일반 다독형	14,350	11,685	2,665	19%	1,692	2,009
고령층 우수형	5,108	4,878	230	5%	526	548

〈표 13〉 예측 모델 학습 기법별 성능 비교(임계값: 0.50)

학습 기법	재현율	AUC-ROC	정밀도	F1-Score	정확도
XGBoost	0.912	0.791	0.274	0.421	0.549
LightGBM	0.923	0.793	0.276	0.424	0.550
LR	0.824	0.704	0.257	0.391	0.539

〈표 14〉 이용자 유형별(K=5) 최적 임계값 적용 시 성능 비교

유 형	기본 임계값(0.50)		단일 최적 임계값(0.61)		유형별 최적 임계값		
	재현율	F1-Score	재현율	F1-Score	임계값	재현율	F1-Score
가족회원 연체형	0.984	0.419	0.942	0.454	0.72	0.769	0.548
장기 가족회원 연체형	0.928	0.501	0.811	0.513	0.66	0.770	0.539
일반 표준형	0.876	0.375	0.771	0.398	0.61	0.771	0.398
일반 다독형	0.923	0.464	0.768	0.501	0.61	0.768	0.501
고령층 우수형	0.615	0.302	0.346	0.257	0.37	0.769	0.237
전체 데이터 학습	0.923	0.424	0.818	0.452	0.61	0.818	0.452

예측 성능을 보인다. 이때 각 유형별 최적 임계값은 검증 데이터에서 F1-Score가 가장 높은 임계값을 선택하며 재현율 기준은 본 연구의 목적에 부합하도록 75% 이상으로 설정하여 연체 탐지율을 유지하도록 하였다.

기본 임계값(0.5)을 적용했을 때 재현율은 이용자 유형별로 61.5~98.4%의 범위에서 편차를 나타냈다. 특히 고령층 우수형은 나머지 4개 유형이 87% 이상의 재현율을 보이는 데 비해서 상대적으로 낮은 재현율(61.5%)을 보였다. 이는 고령층 우수형의 연체율이 5.0%로 다섯 유형 중 가장 낮아 모델 학습 시 연체에 해당하는

데이터 샘플이 상대적으로 적어서 나타난 결과로 해석할 수 있다. 이와 같은 경향은 전체 데이터 기준으로 도출된 최적 임계값(0.61)을 일괄 적용할 경우에도 유사하게 확인되었다. 이 경우에는 재현율을 만족한 상태에서 기본 임계값을 적용했을 때 비해 F1-Score가 상승(0.452)하는 효과가 있었으나, 유형별 성능에서는 재현율의 편차가 기본 임계값에 비해 증가하였고, 특히, 고령층 우수형의 경우 34.6%까지 감소하여 상당수의 연체 건이 정상반납으로 오분류되는 결과로 이어진다. 이 결과는 단일 임계값 적용이 이용자 유형 간 예측 성능의 불균형을 초래할

수 있음을 시사한다.

반면 이용자 유형별 예측 모델이 75% 이상의 재현율을 확보하도록 임계값을 최적화한 결과(0.37~0.72), 고령층 우수형을 제외한 네 가지 유형의 F1-Score가 단일 최적 임계값을 적용한 경우에 비해 최대 0.094 향상됨을 확인할 수 있다. 다만, 가장 낮은 연체율을 보인 고령층 우수형의 경우 목표 재현율 달성을 위해 매우 낮은 수준의 임계값(0.37)의 적용이 요구되었고, F1-Score도 단일 임계값을 적용한 경우 대비 소폭 감소(0.02)함을 확인할 수 있다. 이는 단일 임계값 적용 시와 동일하게 고령층 우수형에 존재하는 클래스 불균형의 영향으로 해석할 수 있다.

〈표 15〉는 전체 데이터로 학습한 연체 예측 모델에 유형별 최적 임계값을 적용한 결과와, 유형별로 데이터를 분할하여 개별 학습한 모델에 각각 최적 임계값을 적용한 결과를 비교한 것이다. 재현율 75% 이상을 확보하는 조건에서 F1-Score를 비교한 결과, 5개 유형 중 4개 유형(가족회원 연체형, 장기 가족회원 연체형, 일반 다독형, 고령층 우수형)에서 전체 데이터 학습 모델이 유형별 개별 학습 모델보다 높은 F1-Score를 기록하였다. 특히 표본 수가 상대적으로 적은 고령층 우수형의 경우 유형별 개

별 학습 시 F1-Score가 0.237에서 0.177로 크게 하락하였으며, 임계값 역시 0.37에서 0.23으로 낮아져 모델의 클래스 분류 능력이 저하된 것으로 해석된다. 반면 일반 표준형은 유형별 학습 시 F1-Score가 0.398에서 0.413으로 소폭 개선되어, 표본 수가 크고 유형 내 동질성이 상대적으로 높은 경우에는 개별 학습의 이점이 제한적으로나마 나타날 수 있음을 보여준다.

4.4 논의

4.4.1 이용자 세분화 기반 연체 발생 예측 모델 성능 최적화 결과

일반적인 예측 모델은 전체 데이터로 학습한 모델에 단일 최적 임계값을 적용하고, 전체 평가 데이터 기준으로 성능 평가를 하여 모델의 적정성을 평가한다. 이러한 방식은 데이터 내 속성이 구별되는 집단(성별, 인종 등)이 존재하거나 클래스 불균형이 있는 경우 집단 간 예측 성능 편차의 주요한 요인이 된다. 본 연구에서 공공도서관 대출 기록으로 학습한 연체 예측 모델의 성능 평가에서 식별된 유형별 연체 예측 성능의 편차는 데이터 내에 존재하는 내재적 그룹 간 차이와 데이터의 규모로 인한 예측 성능 편향이 도서관 대출 이력 데이터에서

〈표 15〉 이용자 유형별(K=5)로 개별 학습된 예측 모델과의 성능 비교

유형	전체 데이터 학습 모델			유형별 데이터 개별 학습 모델		
	임계값	재현율	F1-Score	임계값	재현율	F1-Score
가족회원 연체형	0.72	0.769	0.548	0.48	0.949	0.488
장기 가족회원 연체형	0.66	0.770	0.539	0.58	0.752	0.521
일반 표준형	0.61	0.771	0.398	0.58	0.776	0.413
일반 다독형	0.61	0.768	0.501	0.58	0.788	0.466
고령층 우수형	0.37	0.769	0.237	0.23	0.769	0.177

도 존재함을 시사한다. 따라서 도서 연체 예측 모델을 실제 운영에 적용하기 위해서는 이용자의 유형화와 각 유형별 예측 모델의 성능 최적화 방안에 대한 고려가 필요하다.

이에 대한 방안으로 본 연구에서 제안한 이용자 유형별 분류 임계값 최적화 방식의 성능 평가 결과는 재현율 75% 이상을 만족하는 조건에서 고령층 우수형을 제외한 네 가지 유형의 F1-Score가 단일 최적 임계값을 적용한 경우에 비해 동등 이상의 성능을 보여 유형별 예측 임계값 최적화를 통한 차별화된 연체 관리 전략의 적용 가능성을 뒷받침한다. 다만, 고령층 우수형과 같이 클래스 불균형이 존재하는 집단의 경우 해당 집단의 특성을 반영한 운영 관점의 접근 방식의 고려가 필요하다. 또한, 유형별 데이터로 개별 학습한 모델이 전체 데이터를 이용해 학습한 모델에 비해서 성능이 향상되지 않음을 확인할 수 있었는데, 이는 학습 표본 감소에 따른 손실이 유형 특화 학습으로 얻는 이득보다 클 수 있음을 시사한다. 따라서 이러한 방식의 적용은 충분한 데이터가 축적된 환경이 전제되어야 할 것이다.

4.4.2 이용자 유형별 맞춤형 연체 관리 전략

4.4.1의 분석 결과를 바탕으로, 연체 예방에 중점을 둔 이용자 유형별 관리 전략의 도입을 다음과 같이 고려할 수 있다. 단, 본 연구의 데이터 분석 결과로부터 적용 가능성 차원의 운영적 제안으로, 향후 현장 검증이 요구된다.

첫 번째로 예측 모델의 유형별 임계값 차등 운영전략이다. 본 연구의 실험을 통해 단일 임계값은 데이터 수가 많은 유형에 적합하도록 최적화되어 소수 유형의 예측 성능이 저하될 가능

성이 있음을 확인하였다. 따라서 하나의 예측 모델을 공통으로 사용하되, 이용자 유형에 따라 서로 다른 임계값을 적용하는 방안을 고려할 수 있다. 이는 개별 학습에서 나타난 표본 감소의 손실을 피하면서 유형 간 탐지 성능의 편차를 통제할 수 있고, 유형별 개별 모델 운영유지에 따른 관리 부담도 경감할 수 있다는 장점이 있다. 다만, 대출 기록이 누적됨에 따라 주기적으로 유형 분류 및 임계값 최적화를 수행해야 하는 점을 고려해야 한다.

두 번째 전략은 반납 알림 체계의 세분화와 자동화 방안이다. 모든 이용자에게 동일한 내용과 주기로 발송되는 현행 알림 체계를 유형별 시나리오로 세분화하고 연체 위험이 높게 예측된 이용자에게는 집중 관리하는 방안을 적용할 수 있다. 유형별 시나리오의 예시로는 회당 평균 6.7권을 대출하는 일반 다독형은 안내 문자에 대출 중인 도서의 제목과 권수를 함께 기재하여 반납 대상을 명확히 인지하도록 돕는 방식을 고려해 볼 수 있고, 가족회원 연체형과 장기 가족회원 연체형에 대해서는 대출 목록 공유에 대한 동의를 전제로 가족 단위의 대출 현황과 반납예정일을 통합 안내하는 방안을 적용해 볼 수 있다. 또한, 고령층 우수형은 실험결과에서 목표 재현율 확보를 위한 임계값이 0.37로 나타나 예측 모델 적용이 제한적인 것으로 확인되었으므로 머신러닝 기반의 차등적 전략 적용보다는 전체 인원을 대상으로 대출 영수증을 제공하거나 대출반납일력표를 통해 안내하는 운영 차원의 보완적 방식을 적용하는 방안을 고려할 수 있다.

세 번째는 대출 규정의 인지를 위한 주기적 환기 전략이다. 평균 등록 기간이 8,035일에 이

르면서도 연체율이 21%로 높은 장기 가족회원 연체형은 이용 경험의 축적이 반납 준수로 이어지지 않은 사례로 해석될 수 있다. 이러한 반복적인 연체를 방지하기 위해 회원 가입 시 대출 규정과 연체 제재 사항을 이용 안내 리플릿과 모바일 안내문을 통해 명확히 전달하고 연 1회 수준의 규정 환기 안내를 검토할 수 있다.

마지막으로 자율적 행동 교정을 위한 시각화 피드백과 인센티브 전략의 적용을 고려할 수 있다. 개인별 대출·반납 성향과 연체 위험도를 시각화한 리포트를 모바일 앱 또는 웹을 통해 제공하면 이용자가 자신의 행태를 인지하고 스스로 개선하도록 유도할 수 있으며, 제재 중심의 접근을 보완하여 성실 이용자에게 대출권수 확대 등의 혜택을 부여하는 방안도 병행할 수 있다.

5. 결 론

본 연구는 국내 C공공도서관에서 실제로 수집한 도서 대출·반납 데이터를 활용하여 이용자 세분화와 연체 발생 예측을 연계한 분석 방법을 적용한 연구로서 다음과 같은 학술적·실무적 의의를 갖는다.

첫째, 기존 도서관 연구가 이용 현황의 정량적 분석이나 연체 발생 이후의 원인 규명에 머물렀던 것과 달리, 본 연구는 실제 대출·반납 이력을 활용하여 이용자 특성을 정량적으로 분석하여 유형화하고 예측 모델을 구축하였다는 점에서 차별성을 갖는다.

둘째, 비지도 학습 기반 군집 분석에 기반한 유형별 예측 최적 임계값을 적용함으로써 전체

성능 지표만으로는 파악하기 어려운 이용자 유형별 예측 성능 편차 문제를 군집 분석과의 연계를 통해 드러냈다. 이 점은 유사한 구조의 데이터를 다루는 후속 연구에서 참조할 수 있는 분석 방법론이 된다는 점에서 의미가 있다.

셋째, 이용자를 단일 집단으로 간주하는 획일적 관리 방식에서 벗어나 유형별로 차별화된 연체 예방 전략을 제안하였다. 이는 도서관 현장에서 한정된 행정 자원을 고위험 이용자 집단에 집중하는 방식으로 운영 효율성을 높이는 한편, 연체를 사전에 방지하여 장서 순환율을 개선하는 데에도 기여할 것으로 기대된다.

본 연구는 다음과 같은 한계점을 가진다.

첫째, 본 연구는 C공공도서관의 단일 기관 데이터를 활용한 연구로 다양한 환경에서 모델의 성능을 확인하는 일반화 측면의 검증은 제한된다. 도서관의 규모, 지역적 특성, 이용자 구성에 따라 연체 행동 패턴이 상이하게 나타날 수 있으므로 향후 다양한 규모와 지역의 도서관 데이터를 추가하여 모델의 범용성을 검증하는 연구가 필요하다.

둘째, 평가 데이터가 특정 기간(2026년 1월~3월)에 한정되고, 유형별 표본 규모의 편차가 존재한다. 따라서 연체 예측 모델의 일반화 성능 검증 및 유형별 최적 임계값의 유효성을 확인하기 위해 다양한 시기 데이터의 충분한 확보가 요구된다.

셋째, 예측 모델에 활용된 변수가 도서관 대출·반납 이력 데이터 내의 내재적 변수에 한정되었다. 이용자의 직업, 거주지와 도서관 간 거리, 대중교통 이용 여건 등 연체 행동에 영향을 줄 수 있는 외재적 요인과 연령, 성별 등 인구통계학적 정보는 분석에서 제외되었다. 다만,

이는 개인정보보호 및 데이터 비식별화 조치로 인한 것으로 향후 개인정보보호 대책을 강구한 후 이용자 특성 변수를 추가로 적용하면 예측 모델의 성능을 향상시킬 수 있을 것으로 기대된다.

넷째, 본 연구에서 적용한 군집 분석은 이용자의 유형별 세분화를 목적으로 전체 관측기간의 대출 이력을 기반으로 수행하였다. 그러나 실제 도서관 운영환경에서는 누적 연체 권수와 같이 시간에 따라 갱신되는 정보에 예측 시점 이후의 기록이 포함될 가능성이 있다. 이러한 정보의 포함이 본 연구의 이용자 세분화 결과에 미친 영향을 확인하기 위해 평가기간을 제외한 데이터로 군집을 재산정하여 비교한 결과, 두 군집화 결과의 일치도를 나타내는 조정 랜드 지수(ARI)는 0.958, 동일 군집 유지율은 98.4%로 나타나 평가기간 정보의 포함 여부에 따른

이용자 유형 구조가 본 연구의 실험환경에서는 거의 동일하게 유지되었음을 확인하였다. 다만 실제 운영환경에 적용할 경우에는 예측 시점까지 확보된 정보만을 이용하여 이용자 유형을 결정하고, 지속적으로 누적되는 신규 데이터를 반영하기 위해 이용자 유형과 유형별 임계값을 주기적으로 갱신할 필요가 있다. 따라서 향후 연구로 지속적으로 축적되는 데이터 흐름을 고려한 자동화된 운영 파이프라인을 구축하고 그 효과를 검증할 필요가 있다.

다섯째, 본 연구가 제시한 유형별 관리 방안은 분석 결과와 선행 연구에 근거하여 도출한 제안으로, 실제 개입을 통한 효과가 측정되지 않았다. 향후 소규모 파일럿 적용을 통해 사전·사후 연체율을 비교하는 검증 절차가 요구된다.

참 고 문 헌

- 고재민, 임호균 (2012). 공공도서관의 복합화 경향에 관한 연구. 디자인융복합연구, 11(5), 45-56.
- 김정화 (2018). k-prototypes 군집분석에 관한 고찰. 석사학위논문, 고려대학교.
- 김종우 (2024.12.04.). [인&아웃] 도서 연체자 10만명 '특별사면'. 연합뉴스, <https://www.yna.co.kr/view/AKR20241203090500546>
- 남재우, 최원태 (2021). 공공도서관 이용 증진을 위한 우수 이용자 우대서비스 제공에 관한 연구. 한국 문헌정보학회지, 55(1), 305-321. <http://doi.org/10.4275/KSLIS.2021.55.1.305>
- 문화체육관광부 (2025). 2025년 전국도서관통계 공공도서관 통계조사 결과보고서.
- 박미환 (2009). 공공도서관 연체 자료의 효율적 회수 방안: 서울시립 Y도서관 사례를 중심으로. 문헌정보학연구지, 35, 71-85.
- 박인호, 김일환 (2007). 대출도서 이용경향과 연체요인분석을 통한 연체율 개선방안 연구. 사대도협회지, 8, 121-134.

- 박희진, 박성재 (2023). 잠재성장모형과 성장혼합모형을 이용한 도서관 대출권수의 중단적 분석. *한국 문헌정보학회지*, 57(1), 45-68. <http://doi.org/10.4275/KSLIS.2023.57.1.045>
- 배동식 (2022). 공공도서관의 사회적 역할 변화에 따른 공간구성에 관한 연구: 대구광역시 공공도서관을 중심으로. *한국공간디자인학회논문집*, 17(3), 305-314. <http://dx.doi.org/10.35216/kisd.2022.17.3.305>
- 백선웅 (2024). 국내 대학도서관 대출규정 분석과 개선방안. 석사학위논문, 대구대학교.
- 백승옥 (2002). 대학도서관의 CRM 구축에 관한 연구: C 대학도서관의 대출서비스를 중심으로. 석사학위논문, 전남대학교.
- 심효정, 송현경 (2024). 미국과 국내 공공도서관 연체료 정책 분석을 통한 개선 방안 연구. *한국문헌정보학회지*, 58(3), 145-168. <http://doi.org/10.4275/KSLIS.2024.58.3.145>
- 오민기, 김경래, 정원웅, 김건욱 (2021). 문화적 특성을 고려한 공공도서관 도서 대출수요 분석: 대구광역시 시립도서관을 사례로. *디지털융복합연구*, 19(3), 55-64. <http://doi.org/10.14400/JDC.2021.19.3.055>
- 윤희윤 (2020). 전국 공공도서관 대출규정 조사 연구. *한국도서관·정보학회지*, 51(1), 349-372. <http://doi.org/10.16981/kliss.51.1.202003.349>
- 이태호 (2019). 도서관에서 도서 대여/반납에 대한 인간의 행동과 인간-책 네트워크의 구조. 석사학위논문, 인하대학교.
- 최봉진, 이성우, 김연국 (2024). 설명 가능한 인공지능을 활용한 은행대출연체 예측 모델 연구. *전산회계연구*, 22(2), 44-61. <http://doi.org/10.32956/kaoca.2024.22.2.43>
- Albergaria, M. & Lima, G. (2022). A note on inattention: Evidence from library services in São Paulo. *Revista Brasileira de Economia*, 76(1), 1-9. <https://doi.org/10.5935/0034-7140.20220001>
- Chen, T. & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283-304.
- Islam, N., Islam, K., Wider, W., Butdiswan, S., Ahmad, S., Chakrabarty, R., Matin, A., & Ahmad, N. (2026). Big data, chatbots, and data mining: transforming academic libraries-a bibliometric analysis. *New Review of Academic Librarianship*, 32(1), 24-51. <https://doi.org/10.1080/13614533.2025.2544531>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing*

- Systems, 30, 3149-3157.
- Larasati, A., Hajji, A. M., Handayani, A. N., Azzahra, N., Farhan, M., & Rahmawati, P. (2019). Profiling academic library patrons using k-means and x-means clustering. *International Journal of Technology*, 10(8), 1567-1575. <https://doi.org/10.14716/ijtech.v10i8.3440>
- Liu, J. (2024). The challenge of procrastination: gaining insights from overdue books in university libraries. *Med Clin Res*, 9(8), 01-19. <https://doi.org/10.21203/rs.3.rs-4426009/v1>
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129-137. <https://doi.org/10.1109/TIT.1982.1056489>
- Monika, W., Nasution, A. H., Syam, F. A., & Wijesundara, C. (2025). Clustering of Library's Patron Behavior Using Machine Learning. *Digital Zone: Journal Teknologi Informasi dan Komunikasi*, 16(1), 1-11. <https://doi.org/10.31849/digitalzone.v16i1.19680>
- Nova, N. A., Morales, H., Pájaro, J., & González, A. (2025). Advancing library operations with AI: data-driven insights for academic resource management. *Information Research an International Electronic Journal*, 30(CoLIS), 105-120. <https://doi.org/10.47989/ir30CoLIS52261>
- Sun, J. (2021). Prediction and estimation of book borrowing in the library: machine learning. *Informatica*, 45(1), 163-168. <https://doi.org/10.31449/inf.v45i1.3431>
- Sung, J. & Tolppanen, B. (2013). Do library fines work?: analysis of the effectiveness of fines on patron's return behavior at two mid-sized academic libraries. *The Journal of Academic Librarianship*, 39(6), 506-511. <https://doi.org/10.1016/j.acalib.2013.08.011>
- Zhang, C. (2023). Cloud data resources and library subject information services. *Applied Mathematics and Nonlinear Sciences*, 9(1), 1-17. <https://doi.org/10.2478/amns-2024-1007>
- Zhang, H. (2024). Machine learning algorithms for predicting and estimating book borrowing in university libraries. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 28(5), 1204-1209. <https://doi.org/10.20965/jaciii.2024.p1204>

• 국문 참고자료의 영어 표기

(English translation / romanization of references originally written in Korean)

- Bae, Dong-Sig (2022). A study on the spatial composition of public libraries according to the change in the social role: focused on the public library in Daegu. *Journal of the Korea Institute of Spatial Design*, 17(3), 305-314. <http://dx.doi.org/10.35216/kisd.2022.17.3.305>
- Baek, Seoung-Ok (2002). A Study on the Construction of CRM in University Library: Focused

- on the Circulation Service of C University Library. Master's thesis, Chonnam National University, Korea.
- Baek, Sun-Ung (2024). A Study on the Book Circulation Rules of University in Korea. Master's thesis, Daegu University, Korea.
- Choi, Bong-Jin, Lee, Seong-Woo, & Kim, Yeon-Kook (2024). Developing consumer bank loan delinquency prediction model using XAI. *Journal of Korean Association of Computer and Accounting*, 22(2), 44-61. <http://doi.org/10.32956/kaoca.2024.22.2.43>
- Kim, Jong-Woo (2024, December 4). [In & Out] 'Special amnesty' for 100,000 book delinquents. *Yonhap News*. <https://www.yna.co.kr/view/AKR20241203090500546>
- Kim, Jung-Hwa (2018). A Study on Clustering Analysis Based on K-prototypes. Master's thesis, Korea University, Korea.
- Ko, Jae-Min & Lim, Ho-Kyun (2012). A study on the complex trend of domestic public libraries. *Design Convergence Study*, 11(5), 45-56.
- Lee, Tae-Ho (2019). Human Behavior for Book Borrowing/Returning in Library and Structure of Human-Book Networks. Master's thesis, Inha University, Korea.
- Ministry of Culture, Sports and Tourism (2025). 2025 Public Library Statistical Survey Result Report (Based on 2024 Data) of the National Library Statistics.
- Nam, Jae-Woo & Choi, Won-Tae (2021). A study on the provision of preferential services for excellent users to promote the use of public libraries. *Journal of the Korean Society for Library and Information Science*, 55(1), 305-321. <http://doi.org/10.4275/KSLIS.2021.55.1.305>
- Oh, Min-Ki, Kim, Kyung-Rae, Jeong, Won-Oong, & Kim, Keun-Wook (2021). Analysis of borrows demand for books in public libraries considering cultural characteristics. *Journal of Digital Convergence*, 19(3), 55-64. <http://doi.org/10.14400/JDC.2021.19.3.055>
- Park, Hee-Jin & Park, Sung-Jae (2023). A longitudinal analysis of the number of checked-out books using latent growth model and growth mixture modeling. *Journal of the Korean Society for Library and Information Science*, 57(1), 45-68. <http://doi.org/10.4275/KSLIS.2023.57.1.045>
- Park, In-ho & Kim, Il-Hwan (2007). A study on improving loan overdue rates through analyzing book loan trends and overdue factors. *Journal of the Korea Private University Library Association*, 8, 121-134.
- Park, Mi-hwan (2009). A study on effective methods for retrieving overdue materials in public libraries: focusing on the case of Seoul metropolitan Y library. *Journal of Library and*

Information Science Research, 35, 71-85.

Sim, Hyo-Jung & Song, Hyun-Kyung (2024). A study on the improvement measures through analysis of late fee policy in public libraries in the United States and Korea. Journal of the Korean Society for Library and Information Science, 58(3) 145-168.
<http://doi.org/10.4275/KSLIS.2024.58.3.145>

Yoon, Hee-Yoon (2020). A study on the book circulation rules of public library in Korea. Journal of Korean Library and Information Science Society, 51(1), 349-372.
<http://doi.org/10.16981/kliss.51.1.202003.349>