GMM 기반 oversampling 기법 적용 및 군집 별 이직 요인 분석

IBM HR Analysis Datasets를 활용하여

전남대학교

지리학과 문혁 (202466)
산업공학과 임창엽 (211782)

2026. 03. 30.

Contents

Motivation

1. 인적 자원 관리의 중요성과 이직 예측의 필요성

- **기업 경쟁력의 핵심 자산 보호:** 현대 경영 환경에서 직원의 이직은 단순히 개별 인력의 이탈을 넘어 조직 내 축척된 지식 자산의 유출과 생산성 저하를 초래하는 치명적인 위협 요인 (이수정 외, 2026)
- **경제적 손실 및 관리 비용 증대:** 핵심 인력의 이직은 신규 채용 및 교육 훈련에 따르는 막대한 재정적 비용과 시간적 기회비용을 발생시켜 기업의 운영 효율성을 저해함 (최준영, 2022)
- **선제적 대응 체계의 필요:** 따라서 데이터 기반의 객관적 분석을 통해 이직 가능성이 높은 직원을 사전에 식별하고 실효성 있는 이직 방지 대책 전략을 수립할 수 있는 정밀한 예측 모델이 필수적으로 요구됨 (이수정 외, 2026)

>> 이직하는 직원들의 이직 사유를 추정할 수 있는 정확한 모델이 필요함

2. 불균형 데이터 문제에 따른 예측 모델의 기술적 한계

- **이직 데이터의 희소성 문제:** 실제 현장에서 수집되는 HR Data는 잔류 직원에 비해 이직자 비중이 극히 적은 **클래스 불균형** 특성을 보임 (Juhi Padmaja P. et al., 2022)
- **다수 클래스 편향에 따른 성능 저하:** 불균형한 상태로 모델을 학습할 경우 머신러닝 알고리즘이 다수 클래스에 편향되어 실제 중요한 소수 클래스의 특징을 제대로 학습하지 못하는 결과가 발생함 (Krishna & Sidharth, 2022)
- **Accuracy의 함정과 식별력 상실:** 전체 정확도 Accuracy는 높게 나타나지만 정작 기업의 실질적인 피해를 주는 2종 오류(False Negative; FN)을 잡아내는 **Recall 수치**가 현저하게 떨어져 실전 예측 도구로서의 신뢰성이 낮아짐 (Krishna & Sidharth, 2022)

>> 불균형 데이터 세트를 활용하여 이직자를 정확하게 판별할 수 있는 모델이 필요함

Literature Review (1) : Setiawan et al. (2020)

Core

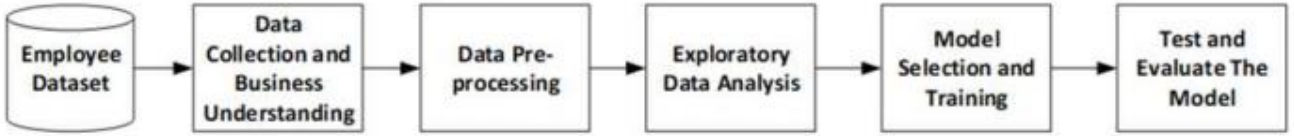


Figure 1. Attrition analysis steps.

- **로지스틱 회귀 귀반 이직 예측**
- **5단계 분석 프레임 워크:**

데이터 수집 및 이해 – 데이터 전처리 – 탐색적 데이터 분석 – 모델 선택 및 학습 – 모델 테스트 및 평가 5단계로 수행

- **분산 팽창 계수 활용:** 다중 공선성을 진단하기 위해 분산 팽창 계수를 활용하였음
- **예측 성능:** Accuracy 75%, Precision 35.7%, Recall 73%

Limitation

- **데이터 불균형 문제 :** 전형적인 불균형 세트이지만 별도의 **오버 샘플링 기법을 적용하지 않고** 로지스틱 회귀를 적용함.
- **선형 모델의 한계:** 복잡한 비선형 패턴을 잡지 못함

Literature Review (2) : Govindarajana et al. (2025)

Core

- **보팅 기반 하이브리드 모델 적용:** 단일 알고리즘의 약점 극복을 위해 RandomForest, SVM, Decision Tree 등 여러 모델을 결합한 **하드 보팅 기반 하이브리드 모델** 적용
- **변수 중요도 분석**을 통해 **변수를 선택**하고 모델의 복잡함을 해소함
- **SMOTE 오버샘플링**을 적용하여 데이터 불균형 문제 해소
- **변수 선택 최적화:** 상관관계 분석을 통해 영향력이 낮은 변수를 제거하여 복잡함 해소
- 단일 모델 대비 하이브리드 모델의 **Accuracy가 95%**를 기록하였음

Limitation

- **정확도 중심의 모델 성능 평가:** 모델의 Accuracy를 주요 성능 평가 지표로 삼음
  > Accuracy는 다수 클래스에 편향된 결과일 수 있음
- **모델의 복잡함:** 알고리즘을 결합한 형태이기 때문에 모델이 복잡함
- **샘플링 방식의 단순성:** SMOTE 선형 보간 방식은 데이터의 분포를 충분히 반영하지 못하여 합성된 데이터의 실제성이 낮을 수 있음

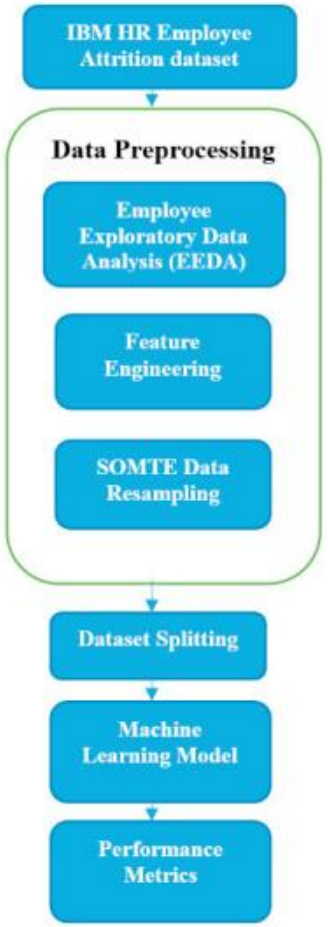


Fig. 1. Methodological analysis of the employee attrition prediction.

Literature Review (3) : Shobhanam Krishna & Sumati Sidharth (2022)

Core

- **단일 알고리즘 최적화:** 앙상블 기법인 Random Forest를 활용하여 고차원 HR 데이터의 특성을 학습함
- **불균형 데이터 처리:** SMOTE 오버샘플링 매커니즘을 적용하여 데이터 불균형 문제를 해결함
- **평가지표:** Accuracy, Precision, Recall, F1-Score, AUC-ROC 등 5가지 성능 지표를 사용하여 모델을

지표	Accuracy	Precision	Recall	F1-Score	AUC-ROC
SMOTE 적용 전	0.84	0.78	0.55	0.64	0.81
SMOTE 적용 후	0.87	0.81	0.58	0.68	0.83

Limitation

- **SMOTE 적용 후 Recall 개선 미미:** 데이터의 불균형 문제는 해결하였지만 Recall의 상승률이 크지 않음
- **모델 해석력 부족:** Random Forest의 변수 중요도를 확인할 수 있지만 SHAP, LIME 등 최신 XAI을 적용하지 않음
- **클래스 경계의 복잡성 반영 미흡:** SMOTE는 선형 보간 기반으로 새로운 소수 클래스 샘플을 생성하지만 다변량, 비선형 구조를 충분히 포착하지 못함

Contribution

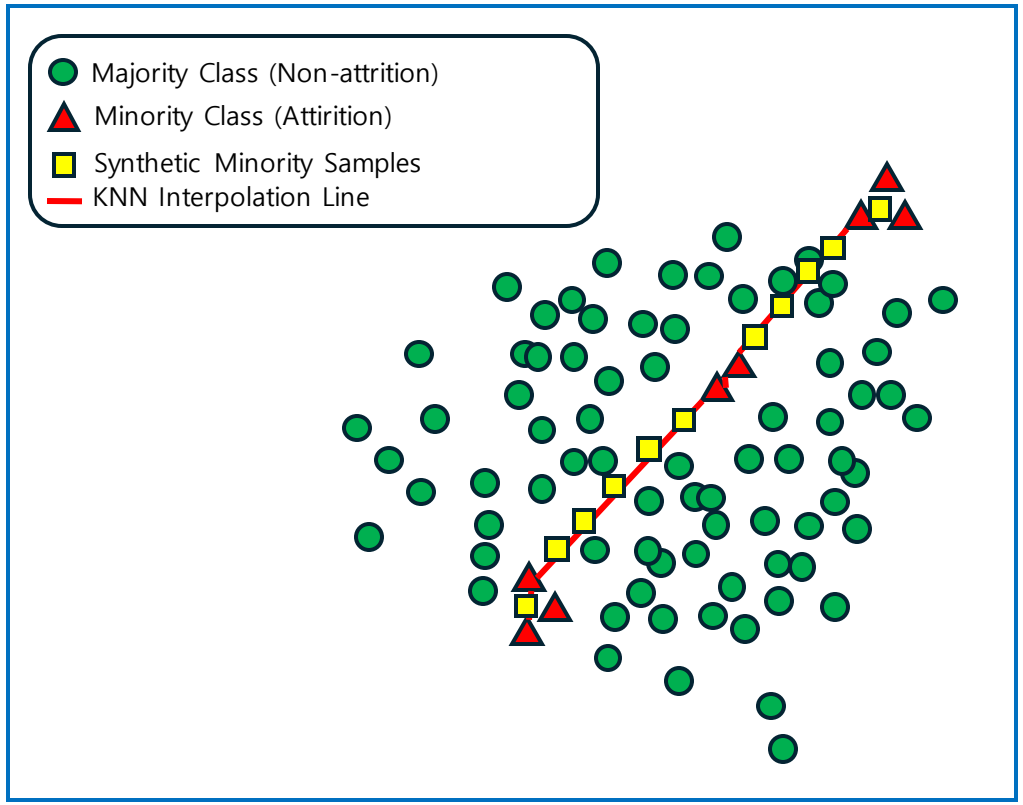
- 기존의 **SMOTE 기반 Over Sampling**은 소수 클래스 간의 직선적인 선형 보간에만 의존하여 실제 데이터가 가진 확률적 분포를 복원하는 데 한계가 있다.
- 또한 서로 다른 이직 사유를 가지는 데이터들이 하나의 선으로 연결되면서 데이터 고유의 패턴이 왜곡되거나 노이즈가 생성될 위험이 크고, 클래스 경계의 복잡성을 반영하기에 미흡하다.

따라서 본 연구는 **GMM(Gaussian Mixture Model) 기반 다중 군집 오버샘플링 방법**을 제안한다.

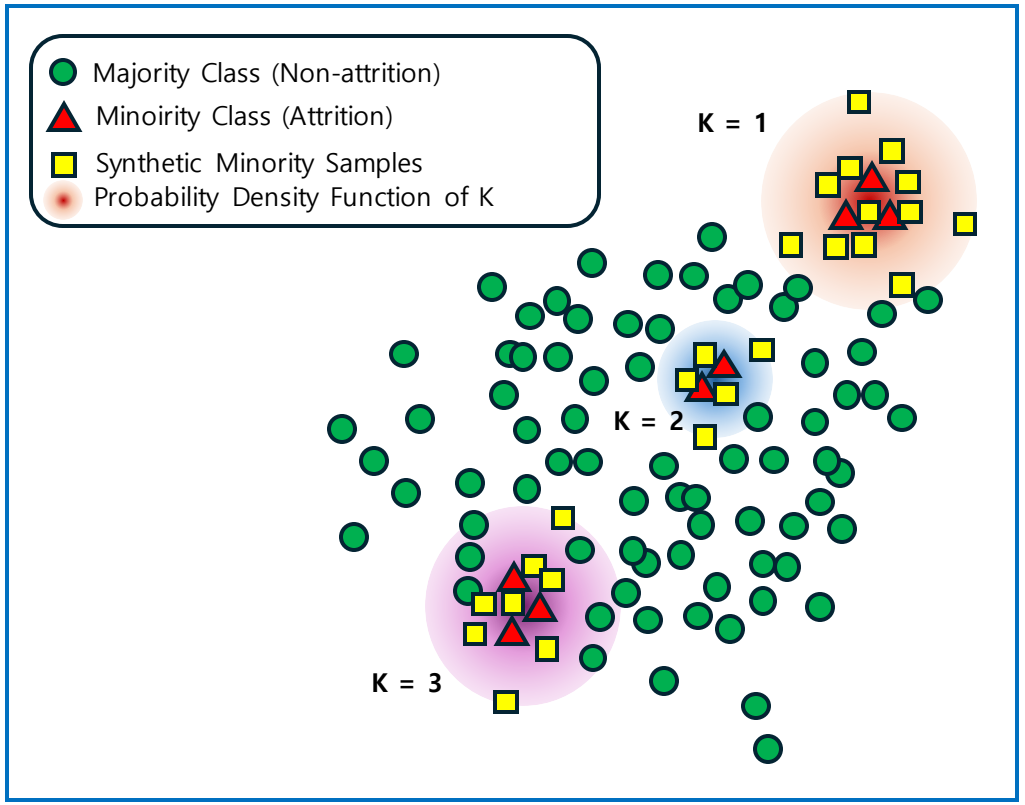
- GMM 기반 오버 샘플링은 BIC를 활용하여 데이터를 K개의 독립적인 정규분포 군집으로 세분화하고 각 군집의 정규분포를 결합한 혼합 모델을 구축하여 증강을 수행한다.
- BIC = -ln(L)/2 + ln(n)/k -> 최솟값 선택
- 단순 복제가 아닌 소수 클래스의 확률 밀도 함수(PDF)로부터 샘플을 추출하므로 실제 발생 가능한 잠재적 이직 패턴을 정교하게 생성할 수 있다.
- 생성된 K개의 군집별로 XAI 기반 이직에 영향을 미치는 특성을 분석하여 군집별 맞춤형 이직 방지 전략을 제안할 수 있다.

Contribution

SMOTE



GMM Based Oversampling



K는 군집의 수, 최적화 하는 과정이 필요함

GMM 기반 오버샘플링 기법 적용 및 군집 별 이직 요인 분석

Main Part

데이터 정제 및 변수 최적화	기초 통계 기반 전처리 및 VIF 기반 다중공선성 진단
학습 / 검증 데이터 분할	모델 학습 및 평가를 위한 데이터 세트 분할
군집 수(K) 설정	최적 군집 수 탐색
GMM 생성	군집별 정규분포를 활용하여 모델 샘플링 및 학습
모델 평가	검증 데이터로 모델 평가 후 피드백 (Recall 및 F1-Score 중심)
군집 별 이직 요인 분석	XAI 활용 군집 별 이직 영향 요인 분석

Main Part : Dataset

데이터 요약 표

항목	내용	비고
Data Source	IBM HR Analytics Dataset	Kaggle 공개 벤치마크 데이터
Total Samples	1,470명	
Total Features	35개 독립변수 34개 종속변수 1개	
Taret Variable	Attrition (Yes / No)	이직 여부
Class Distribution	No: 1,233(83.9%) Yes: 237(16.1%)	불균형 데이터
Data Types	범주형 정수형 연속형	혼합형 데이터

주요 변수 구성

변수 그룹	주요 변수
인적사항 정보	Age, Gender, Marital Status, Education
직무 성격 및 근무	JobRole, JobLevel, Department, BusinessTravel
업무 보상	MonthlyIncome, PrecentSalaryHike, StockOptionLevel
심리적 만족도	JobSatisfaction EnvironmentSatisfaction, WorkLifeBalance
경력 및 조직 충성도	YearsAtCompany, YearsSinceLastPromotion, TotalWorkingYears

Reference

- Setiawan, I., Suprihanto, S., Nugraha, A. C., & Hutahaean, J. (2020, April). HR analytics: Employee attrition analysis using logistic regression. In iop conference series: materials science and engineering (Vol. 830, No. 3, p. 032001). IOP Publishing.

- Vinoodhini, D. (2022, April). Effective classification of IBM HR analytics employee attrition using sampling techniques. In 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-6). IEEE.

- Krishna, S., & Sidharth, S. (2022). HR analytics: Employee attrition analysis using random forest. International Journal of Performability Engineering, 18(4), 275.

- Govindarajan, R., Kumar, N. K., & Reddy, S. (2025). Predicting employee attrition: a comparative analysis of machine learning models using the IBM human resource analytics dataset. Procedia Computer Science, 258, 4084-4093.

- 이수정, & 곽대혁. (2026). IBM HR 데이터를 활용한 직원 이직 예측 모델 구축 및 앙상블 기법 연구. 한국컴퓨터정보학회 학술발표논문집, 34(1), 19-22.

- 최준영. (2022). 인사관리효율화를 위한 기계학습 기반의 퇴직 예측: Retirement prediction for efficiency of HR management based on machine learning.

감사합니다

Thank you

전남대학교

지리학과 문혁 (202466)
산업공학과 임창엽 (211782)

2026. 03. 30.