

# 머신러닝 기법을 활용한 제약업체 임금인상률 예측 분석 보고서

## Executive Summary

본 연구는 한국 제약업체의 2026년 임금인상률 예측에 머신러닝 기법을 적용하여 기존 트렌드 분석 방법과의 효과성을 비교 분석한 보고서입니다.

**주요 결론:** 현재 데이터 조건에서 머신러닝 기법은 **제한적으로 효과적**이며, **트렌드 분석 + 도메인 지식** 방법이 더 신뢰할 만한 것으로 판단됩니다.

**2026년 최종 권고 인상률:** Base-up 3.5%, 성과급 2.1% (총 5.6%)

## 1. 분석 개요

### 1.1 연구 목적

- 머신러닝 기법의 임금인상률 예측 적용 가능성 검토
- 기존 트렌드 분석 방법과의 성능 비교
- 2026년 제약업체 임금인상률 예측 정확도 향상 방안 모색

### 1.2 데이터 현황

- 분석 기간: 2015년 ~ 2024년 (10개년)
- 총 변수 수: 33개 (경제지표, 기업성과지표 등)
- 핵심 타겟: Base-up 인상률, 성과급 인상률
- 학습 샘플: 10개 (2025년 제외)

### 1.3 방법론

- 머신러닝 모델: Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, SVR, Neural Network
- 특성 선택: 상관관계 기반 핵심 3-12개 특성 선별
- 평가 방법: 교차검증, Leave-One-Out 검증
- 비교 기준: 기존 트렌드 분석 결과와의 예측값 및 해석력 비교

## 2. 데이터 적합성 분석

### 2.1 샘플 크기 문제

구분	현재 상황	권장 수준	평가
학습 샘플 수	10개	30-50개 이상	✕ 부족
샘플/특성 비율	10:3 = 3.33:1	최소 10:1	✕ 부족
교차검증 안정성	불안정	안정	✕ 위험

**핵심 문제:** 10개 샘플로는 복잡한 머신러닝 모델의 안정적 학습이 어려움

### 2.2 시계열 특성

- **데이터 특성:** 연도별 순차 데이터 (강한 시계열 특성)
- **정책 변화:** 최저임금 급인상(2018), 코로나19(2020-2021), 인플레이션(2022-2023)
- **경제 사이클:** 다양한 경기 변동 포함
- **ML 한계:** 정책적 급변 및 외부 충격에 대한 학습 데이터 부족

## 3. 머신러닝 모델 성능 분석

### 3.1 전체 모델 성능 평가

#### Base-up 인상률 예측

모델	CV Score(R <sup>2</sup> )	Train R <sup>2</sup>	MAE	RMSE	안정성
Linear Regression	-31.586±38.302	0.000	0.0000	0.0000	✕
Ridge Regression	-15.949±20.516	0.417	0.0048	0.0066	✕
<b>Random Forest</b>	-1.152±1.316	<b>0.773</b>	0.0028	0.0041	△
Neural Network	-17788.137±21238.816	0.000	0.0706	0.0928	✕

#### 성과급 인상률 예측

모델	CV Score(R <sup>2</sup> )	Train R <sup>2</sup>	MAE	RMSE	안정성
<b>Linear Regression</b>	-12.446±9.861	<b>0.828</b>	0.0006	0.0007	△
Ridge Regression	-6.543±6.479	0.769	0.0007	0.0008	△
Random Forest	-0.122±0.106	0.645	0.0008	0.0010	△

#### 주요 발견사항:

- 모든 복잡한 모델에서 **음의 교차검증 점수** → 과적합 확인

- 단순 모델(Linear Regression, 단순 Random Forest)만 의미있는 성능
- 훈련 성능과 검증 성능 간 큰 격차

### 3.2 최적 모델 특성 분석

#### Base-up 예측 최적 모델: Random Forest

핵심 특성 중요도:

- 미국 CPI: **64.7%** (가장 중요)
- 대기업 인상률: **23.7%**
- 한국 CPI: **11.6%**

해석: 글로벌 인플레이션과 대기업 임금 벤치마킹이 주요 결정요인

#### 성과급 예측 최적 모델: Linear Regression

회귀 계수:

- 매출증가율: 0.001963 (강한 양의 관계)
- 그룹 BU 인상률: 0.000530
- 대기업 인상률: 0.000108

해석: 회사 실적(매출)이 성과급 결정의 핵심 동인

## 4. 2026년 예측 결과 비교

### 4.1 시나리오별 예측 비교

시나리오	구분	Base-up	성과급	총 인상률	차이
보수적	머신러닝	2.8%	2.1%	<b>4.8%</b>	-0.3%p
	트렌드 분석	3.3%	1.8%	<b>5.1%</b>	
중립적	머신러닝	2.9%	2.1%	<b>5.0%</b>	-0.6%p
	트렌드 분석	3.5%	2.1%	<b>5.6%</b>	
낙관적	머신러닝	2.8%	2.1%	<b>5.0%</b>	-1.1%p
	트렌드 분석	3.8%	2.3%	<b>6.1%</b>	

### 4.2 예측 특징 분석

- 머신러닝: 평균 **0.7%p** 낮게 예측 (보수적 경향)
- 시나리오 민감도: ML이 경제 시나리오 변화에 덜 민감
- 성과급 일치: 두 방법 모두 성과급은 2.1% 수준에서 수렴

## 5. 장단점 비교 분석

### 5.1 머신러닝 기법의 장점

1. **객관적 특성 선택:** 데이터 기반 중요도 자동 식별
  - 미국 CPI의 중요성(64.7%) 객관적 입증
  - 주관적 판단 오류 최소화
2. **복잡한 관계 모델링:** Random Forest를 통한 비선형 상호작용 포착
  - 경제지표 간 복합적 영향 관계 학습
  - 다차원 패턴 인식 가능
3. **일관된 방법론:** 재현 가능하고 체계적인 분석 프로세스
  - 분석자 편향 제거
  - 표준화된 예측 파이프라인

### 5.2 머신러닝 기법의 단점

1. **샘플 부족으로 과적합 위험:**
  - 10개 샘플로는 신뢰할 수 없는 일반화 성능
  - 새로운 경제환경 대응 한계
2. **시계열 특성 고려 미흡:**
  - 경제 사이클과 정책 변화의 맥락 부족
  - 급변하는 외부 환경에 대한 적응력 부족
3. **해석 가능성 제한:**
  - 의사결정 근거 설명의 어려움
  - 경영진 설득을 위한 논리적 근거 부족
4. **도메인 지식 부족:**
  - HR/경영 전문가의 경험과 직관 미반영
  - 업계 특수성과 정성적 요인 간과

### 5.3 트렌드 분석의 장점

1. **높은 해석 가능성:** 직관적 이해와 설명 용이
2. **도메인 지식 반영:** 업계 전문성과 경험적 판단 포함
3. **정책 변화 대응:** 급변하는 환경에 대한 유연한 적응
4. **의사결정 지원:** 경영진 설득을 위한 명확한 논리 제공

## 6. 핵심 인사이트

### 6.1 새로운 발견 사항

#### 미국 CPI의 중요성 입증

- 발견: 미국 CPI가 Base-up 결정에 **64.7%의 중요도** 차지
- 의미: 글로벌 인플레이션이 국내 제약업체 임금정책에 직접적 영향
- 시사점: 환율 변동과 원자재 비용 연동성 확인

#### 성과급의 선형적 특성

- 발견: 성과급과 매출증가율 간 **강한 선형관계( $R^2 = 0.828$ )**
- 의미: 복잡한 ML 모델보다 단순 회귀가 더 적합
- 시사점: 성과급 설계 시 매출 연동이 핵심

### 6.2 방법론적 시사점

#### 데이터 기반 의사결정의 한계

- 소규모 데이터에서는 도메인 전문성이 데이터 과학보다 중요
- 맥락과 상황을 고려한 정성적 판단의 가치 재확인
- 하이브리드 접근의 필요성: 데이터 + 전문가 지식

## 7. 권고사항

### 7.1 현재 상황에서의 최적 접근법

"트렌드 분석 + 도메인 지식" 방법 채택

채택 근거:

1. 신뢰성: 10개년 패턴과 경제환경 전망을 종합적 고려
2. 설명력: HR 컨설팅에서 중요한 논리적 근거 제공
3. 실용성: 경영진 설득과 의사결정 지원에 효과적
4. 적응성: 급변하는 정책 환경에 유연한 대응 가능

### 7.2 머신러닝 도입을 위한 조건

데이터 요구사항

- 1. 장기 데이터: 최소 30-50개년 이상의 역사적 데이터
- 2. 고빈도 데이터: 월별/분기별 세부 데이터 수집
- 3. 외부 변수: 시차 관계가 있는 선행지표 추가
- 4. 정책 변수: 제도 변화에 대한 더미 변수 포함

방법론 개선방안

- 1. 앙상블 방법: 트렌드 분석과 ML 결과의 가중평균
- 2. 시계열 모델: ARIMA, Prophet 등 시계열 전용 모델 활용
- 3. 베이지안 접근: 불확실성을 고려한 확률적 예측
- 4. 전문가 시스템: 규칙 기반 모델과 ML의 결합

7.3 2026년 최종 권고 인상률

구분	권고 인상률	근거
Base-up	3.5%	10개년 평균(2.4%) 상회, 경제 안정화 반영
성과급	2.1%	매출증가율 20% 연동, 두 방법 모두 수렴
총 인상률	5.6%	업계 경쟁력과 지속가능성의 균형점

종합 판단: 도메인 지식과 경제환경 전망을 고려할 때 가장 합리적이고 설득력 있는 예측

8. 결론

8.1 핵심 결론

머신러닝 기법의 효과성: 현재 데이터 조건에서는 제한적

주요 한계:

- 샘플 수 부족(10개)으로 인한 과적합 위험
- 시계열 특성과 정책 변화에 대한 고려 부족
- 해석 가능성과 설명력 제한
- 도메인 지식 반영의 어려움

최적 접근법: 트렌드 분석 + 도메인 지식

- 높은 신뢰성과 설명력
- 실무 적용성과 의사결정 지원 효과
- 경제환경 변화에 대한 유연한 대응

## 8.2 향후 발전 방향

### 단기(1-2년)

- 현재 트렌드 분석 방법 고도화
- 월별/분기별 데이터 수집 체계 구축
- 외부 경제지표 모니터링 시스템 강화

### 중장기(3-5년)

- 충분한 데이터 축적 후 ML 재도입 검토
- 하이브리드 모델 (트렌드 + ML) 개발
- 실시간 예측 시스템 구축

## 8.3 최종 메시지

데이터 과학은 강력한 도구이지만, HR 컨설팅 영역에서는 여전히 도메인 전문성이 핵심입니다.

현재 조건에서는 **경험과 전문 지식에 기반한 분석**이 더 적절하며, 향후 데이터가 충분히 축적되면 **두 방법의 시너지 효과**를 기대할 수 있습니다.

**2026년 임금인상을 최종 권고: Base-up 3.5%, 성과급 2.1% (총 5.6%)**

보고서 작성일: 2025년 8월 25일

분석 기간: 2015년 ~ 2024년

분석 대상: 한국 제약업체 임금인상률

분석 방법: 머신러닝 vs 트렌드 분석 비교연구