

# 📄 Day 4 완료 보고서: NSMC 머신러닝 모델 비교 분석

## 🎯 Day 4 목표 및 달성도

### 설정 목표

- ✅ 1차 목표: 75% 이상 정확도 달성
- ✅ 2차 목표: 3개 알고리즘 체계적 비교
- ✅ 3차 목표: 각 모델의 특성 이해 및 문서화
- ✅ 효율성 평가: 성능 vs 학습시간 트레이드오프 분석

### 최종 달성 결과

목표	달성도	결과
75% 정확도	✅ 초과 달성	최고 79.25% 달성
모델 비교	✅ 완료	3개 모델 완전 분석
특성 이해	✅ 완료	각 모델의 강점/약점 파악
효율성 분석	✅ 완료	상세한 성능/시간 분석

## 📊 모델 성능 종합 비교

### 🏆 최종 성능 순위

순위	모델	정확도	학습시간	효율성 점수*	상태
🥇 1위	나이브 베이즈	79.25%	0.014초	5,661	✅ 채택
🥈 2위	로지스틱 회귀	79.17%	0.604초	131	✅ 대안
🥉 3위	SVM	78.77%	2,558초	0.03	❌ 제외

\*효율성 점수 = 정확도 / 학습시간 × 1000

## 📈 상세 성능 지표

### 나이브 베이즈

- 교차 검증: 79.35% ± 0.27%
- 검증 정확도: 79.25%
- F1-Score: 0.78 (긍정), 0.80 (부정)
- 학습 시간: 0.014초
- 특징: 초고속 학습, 높은 안정성

로지스틱 회귀

- 교차 검증: 79.13% ± 0.28%
- 검증 정확도: 79.17%
- F1-Score: 0.78 (긍정), 0.80 (부정)
- 학습 시간: 0.604초
- 수렴: 6번 반복으로 정상 수렴
- 특징: 완벽한 해석성, 균형잡힌 성능

SVM (제외됨)

- 교차 검증: 78.77% ± 0.27%
- 학습 시간: 2,558초 (42분)
- 서포트 벡터: 52.7% (과도한 복잡성)
- 제외 사유: 학습시간 과다, 성능 향상 없음

🔍 핵심 발견사항

1. 데이터 특성에 따른 모델 적합성

NSMC 데이터 특성:

- 평균 리뷰 길이: 35.2자 (매우 짧음)
- 평균 단어 수: 7.6개 (단순함)
- 감정 구분 단어: 매우 명확 ("최고" vs "최악")

모델 적합성 분석:

짧고 단순한 텍스트 + 명확한 감정 단어  
→ 복잡한 모델보다 단순한 모델이 더 효과적

결과: 나이브 베이즈 > 로지스틱 회귀 > SVM

2. "No Free Lunch Theorem" 실증

핵심 교훈: 복잡한 모델이 항상 더 좋은 것은 아니다

모델	복잡도	성능	시간	결론
SVM	높음	78.77%	42분	과도한 복잡성
로지스틱 회귀	중간	79.17%	0.6초	균형점
나이브 베이즈	낮음	79.25%	0.014초	최적 효율

### 3. 실무 관점에서의 모델 선택

#### 성능 차이 분석:

- 나이브 베이즈 vs 로지스틱 회귀: 0.08%p 차이 (실질적으로 동일)
- 학습 시간 차이: 43배 (0.014초 vs 0.604초)






#### 실무 적용 가이드:

실시간 처리 중요 → 나이브 베이즈  
해석성 중요 → 로지스틱 회귀  
복잡한 패턴 → (이 프로젝트에서는 해당없음)



### 각 모델별 상세 분석

#### 나이브 베이즈: 최고 효율성

##### 강점:

-  최고 성능 (79.25%)
-  초고속 학습 (0.014초)
-  완벽한 해석성 (단어별 확률)
-  메모리 효율적
-  텍스트 분류 특화

##### 약점:




-  단어 독립성 가정의 한계
-  복잡한 패턴 포착 어려움



##### 핵심 발견:

- "ㅎㅎ": 84.2% 긍정 지표
- "???": 70.3% 부정 지표
- 한국어 이모티콘의 강력한 감정 신호 확인



#### 로지스틱 회귀: 최고 해석성

##### 강점:

-  완벽한 가중치 해석
-  나이브 베이즈와 동등한 성능 (79.17%)
-  빠른 수렴 (6번 반복)

-  확률 출력으로 예측 신뢰도 확인
-  맥락 이해 능력 우수

#### 약점:

-  나이브 베이즈 대비 43배 느린 학습
-  선형 모델의 한계

#### 핵심 발견:

최강 긍정 단어: "여운이" (가중치: 5.26)  
최강 부정 단어: "최악의" (가중치: -7.00)  
중요 패턴: 작품성 평가 단어들의 높은 가중치

### SVM: 비효율적 복잡성

#### 성능 분석:

- 정확도: 78.77% (가장 낮음)
- 학습 시간: 42분 (비현실적)
- 서포트 벡터: 52.7% (과도함)

#### 제외 사유:

1. **성능:** 나이브 베이즈보다 0.48%p 낮음
2. **효율성:** 21,672배 더 오랜 학습 시간
3. **복잡성:** 전체 데이터의 절반이 서포트 벡터
4. **적합성:** NSMC 데이터 특성에 부적합

**교훈:** 고차원 희소 데이터라고 해서 무조건 SVM이 좋은 것은 아님

---

### 프로젝트 인사이트

#### 1. 한국어 감정분석 특성

##### 강력한 감정 지표 단어들:

긍정: "여운이", "최고", "수작", "명작"  
부정: "최악의", "아깝다", "쓰레기", "지루하다"

#### 한국어 특수 표현:

- "아깝다": 시간/돈 낭비에 대한 한국어 특유의 표현

- "수작": 작품의 완성도를 나타내는 긍정 표현
- "여운이": 영화의 깊이를 나타내는 최고 지표

## 2. 텍스트 분류 모델 선택 기준

데이터 복잡도에 따른 모델 선택:

단순한 텍스트 (NSMC) → 나이브 베이즈  
중간 복잡도 텍스트 → 로지스틱 회귀  
복잡한 패턴 텍스트 → 딥러닝/Transformer

실무 고려사항:

1. 성능 요구사항: 75% vs 90% vs 95%
2. 응답 시간: 실시간 vs 배치
3. 해석 필요성: 블랙박스 vs 화이트박스
4. 리소스 제약: CPU/메모리/시간

## 다음 단계 계획

### Phase 3: 딥러닝 모델 (6-7주차)

목표: 85%+ 정확도 달성

예정 모델:

- LSTM (순환 신경망)
- CNN (합성곱 신경망)
- Word2Vec/FastText 임베딩

기대 효과:

- 단어 조합 패턴 포착
- 시퀀스 정보 활용
- 복잡한 맥락 이해

### Phase 4: Transformer 모델 (8주차)

목표: 90%+ 정확도 달성

예정 모델:

- KoBERT (한국어 특화)
- KoELECTRA

## 기대 효과:

- 최신 기법 활용
  - 문맥 이해 극대화
  - 최고 성능 달성
- 

## ✅ Day 4 성과 요약

### 🎯 목표 달성도

- 1차 목표 (75%): ✅ 79.25%로 초과 달성
- 모델 이해: ✅ 3개 모델 완전 분석
- 실무 적용성: ✅ 상황별 모델 선택 가이드 완성

### 📚 핵심 학습 성과

1. "No Free Lunch" 실증: 복잡한 모델이 항상 좋지 않음
2. 데이터 기반 선택: NSMC 특성에 맞는 모델 발견
3. 효율성 분석: 성능 vs 시간 트레이드오프 이해
4. 해석성 가치: 로지스틱 회귀의 가중치 분석 활용

### 🔧 기술적 성취

- ✅ TF-IDF + 3개 알고리즘 완전 구현
  - ✅ 교차 검증을 통한 안정적 성능 측정
  - ✅ 체계적인 모델 비교 프레임워크 구축
  - ✅ 실무 지향적 분석 및 해석
- 

## 📊 최종 권장사항

### 🎯 프로덕션 적용 시

1순위 추천: 나이브 베이즈

사용 케이스:

- 실시간 리뷰 모니터링
- 대용량 배치 처리
- 프로토타입 개발
- 리소스 제약 환경

장점: 최고 성능 + 초고속 + 간단함

## 2순위 추천: 로지스틱 회귀

사용 케이스:

- 비즈니스 인사이트 도출
- 모델 설명이 필요한 상황
- 단어별 영향력 분석
- 규제 업종에서의 설명 가능한 AI

장점: 완벽한 해석성 + 동등한 성능

## 비추천: SVM

이유:

- 42분 학습 시간 (비현실적)
- 낮은 성능 (78.77%)
- 과도한 복잡성
- NSMC 데이터 특성에 부적합

## Day 4 완료 선언

Day 4는 계획했던 모든 목표를 성공적으로 달성했습니다!

- ✅ 75% 목표 → 79.25% 달성 (4.25%p 초과)
- ✅ 3개 모델 완전 분석 (나이브 베이즈, 로지스틱 회귀, SVM)
- ✅ 실무 지향적 인사이트 도출
- ✅ 체계적 문서화 완성

다음 단계: 딥러닝 모델 도전으로 85%+ 정확도 달성 시도

보고서 작성일: 2025년 1월

작성자: NSMC 감정분석 프로젝트팀

버전: v1.0