🚢 Titanic Survival Prediction

머신러닝을 이용한 탑승객 생존 예측

Kaggle Competition
정확도: 0.77033 (상위 30%)

December 2025

📌 프로젝트 개요

✓ **목표**: 타이타닉 탑승객의 신상 정보로부터 생존 여부 예측
✓ **데이터 크기**: Train 891명, Test 418명
✓ **문제 유형**: 이진 분류 (Binary Classification)
✓ **기술 스택**: Python, Pandas, Scikit-learn, Matplotlib
✓ **최종 모델**: Support Vector Machine (SVM)

📊 데이터 탐색적 분석 (EDA)

구분	사망 (0)	생존 (1)
샘플 수	549명	342명
비율	61.6%	38.4%

주요 발견:

- 성별(Sex): 여성 74.2% vs 남성 18.9% (상관계수: -0.54)
- 객실 등급: 1등급 62.9% > 2등급 47.3% > 3등급 24.2%
- 운임(Fare): 높을수록 생존율 증가

🔧 피처 엔지니어링

파생 변수	처리 방법
AgeGroup	나이를 10세 단위로 구간화
Title	이름에서 호칭 추출 (Mr/Mrs/Miss)
Cabin	객실의 첫 글자(Deck) 추출
FamilySize	SibSp + Parch + 1
Fare	4분위수로 범주화

⭐ 결과: 원본 12개 피처 → 최종 9개 피처 (PassengerId, Ticket 제거)

⚙️ 모델 성능 비교

5-Fold Stratified Cross Validation 결과

모델	Train Accuracy	Test Accuracy	선택 이유
KNN	84.5%	79.6%	-
Decision Tree	92.0%	81.3%	과적합
Random Forest	90.3%	81.3%	Feature Importance 분석
SVM	**83.6%**	**82.9%**	최고 성능

🎯 최종 모델 선정: SVM

✓ **최고 검증 정확도**: 82.9%
✓ **과적합 최소화**: Train-Test 차이 0.7% (안정적)
✓ **일반화 성능**: 새로운 데이터에 강함
✓ **복잡한 결정 경계**: 비선형 데이터에 효과적

Kaggle 제출 정확도: 0.77033
(상위 30% 달성)

📈 Feature Importance 분석

Random Forest 모델 기반

순위	Feature	중요도	Insight
1️⃣	Sex (성별)	최고	가장 강한 예측 변수
2️⃣	Pclass (객실 등급)	높음	사회경제적 지위 반영
3️⃣	Fare (운임)	높음	Pclass와 높은 상관성
4️⃣	AgeGroup (나이대)	중간	아이와 노인 우대
5️⃣	FamilySize (가족 크기)	중간	집단 효과 분석

🔄 구현 프로세스

1

데이터 로드
& 결측치 처리

2

탐색적
데이터 분석

3

피처
엔지니어링

4

모델
비교 평가

5

최적 모델
선정

6

결과
제출

💡 핵심 인사이트

- **여성과 아이 우대 정책**: 성별이 생존율의 가장 큰 결정 요인
- **상층부 탈출 우위**: 1등급 승객의 생존율이 3배 이상 높음
- **경제 계층의 중요성**: Pclass와 Fare가 강한 예측 변수
- **피처 엔지니어링 효과**: Title, AgeGroup 파생 변수로 모델 성능 향상
- **과적합 회피**: 교차 검증을 통해 일반화 성능 높은 SVM 선정

🎓 결론 및 개선 방안

✅ 강점

- 체계적 ML 파이프라인
- 다중 모델 비교 분석
- 상세한 EDA & 가시화
- 논리적 피처 엔지니어링

🚀 개선 방향

- 하이퍼파라미터 튜닝 (GridSearchCV)
- 앙상블 모델 (VotingClassifier)
- Ticket 패턴 분석
- IsAlone 변수 추가

🎉 Thank You!

Titanic Survival Prediction
Machine Learning Competition

📊 정확도: 0.77033
📈 순위: 상위 30%

GitHub | Kaggle | LinkedIn

Speaker notes