기말프로젝트 중간보고서

Naive Bayes, KNN, 나무모델

최현석, 황성진

2025-05-30

Contents

0.1 변수 구성 관련 설명 1
0.2 세 가지 모델 적합 비교 결과 1
0.3 3. 결론 3

0.1 변수 구성 관련 설명

모델 비교의 객관성을 위해 가능한 한 동일한 성능 지표를 활용하였으나,
각 모델의 특성과 데이터 요구 조건에 따라 일부 변수 구성이 달라졌음을 밝힌다.

Table 1: 모델별 변수 구성

모델	사용 변수
KNN	”Years.of.Service”, ”Position”, ”GS”, ”PPG”, ”X3P”, ”FT”, ”TS.”, ”X2PA”, ”AST”, ”BLK”, ”TOV”, ”USG.”, ”DBPM”, ”VORP”
Naive Bayes	”PPG”, ”VORP”, ”DRB”, ”TOV”, ”Position”, ”Years.of.Service”, ”AST”, ”MP”
결정나무	모든 변수 사용

0.2 세 가지 모델 적합 비교 결과

0.2.1 전체 성능 요약

Table 2: 모델별 전체 성능 요약

모델	교차검증 정확도	표준편차	교차검증 Kappa	검증셋 정확도	검증셋 Kappa
KNN (gaussian)	0.6561	0.0604	0.4446	0.6818	0.5014
Naive Bayes	0.7246	0.0444	0.5709	0.6860	0.4535
결정나무	0.6609	0.0490	0.4554	0.6250	0.4108

해석 및 요약

- **Naive Bayes**는 교차검증 정확도(0.7246)와 Kappa 값(0.5709) 모두 가장 높아, **모델의 일관성과 안정성이 우수**한 것으로 평가된다.
- **KNN (Gaussian)** 모델은 검증 셋 기준으로 **정확도(0.6818)** 및 **Kappa 값(0.5014)** 이 가장 높아, **실제 분류 상황에서의 성능이 뛰어날 가능성**이 있다.
- **결정나무**는 구조가 단순하여 해석은 용이하지만, 전체적으로 정확도와 일관성이 낮아, **성능 면에서는 비교적 취약**한 것으로 분석된다. 이는 샘플 수 대비 변수가 많을 경우 **과적합 없이 단순하게 모델이 구성되었기 때문**일 가능성이 있다.

→ 따라서 모델의 **안정성 및 전반적 성능을 고려할 때는 Naive Bayes가 가장 적절한 선택**이며, KNN은 **실제 예측 상황에서의 실용성 측면에서 보완적 활용 가능성**이 있다.

0.2.2 세부 성능 요약

Table 3: 세부 성능 요약

모델	Class	민감도	특이도	정밀도	균형 정확도	AUC
Naive Bayes	Low	0.5810	0.8910	0.7500	0.7360	0.8152
Naive Bayes	Mid	0.7436	0.6383	0.6300	0.6910	0.7070
Naive Bayes	High	0.7500	0.9430	0.7500	0.8460	0.9268
KNN	Low	0.7200	0.8413	0.6429	0.7806	0.8900
KNN	Mid	0.6512	0.7333	0.7000	0.6922	0.6980
KNN	High	0.7500	0.9265	0.7500	0.8382	0.8590
결정나무	Low	0.6000	0.8095	0.5556	0.7048	0.7968
결정나무	Mid	0.6050	0.6889	0.6500	0.6468	0.6811
결정나무	High	0.7000	0.8970	0.6667	0.7985	0.8103

- **Naive Bayes**에서는 **High 클래스**가 가장 뛰어난 성능을 보임. 전체적으로 AUC가 세 클래스 모두 0.7 이상으로 고르게 분포되어 있어 안정적인 구분 성능을 보임.
  Mid 클래스에서는 정밀도(0.630)가 다소 낮아, 해당 구간에서 **오분류 가능성**이 존재함.
- **KNN (Gaussian)**에서는 **Low, High 클래스**에서 민감도·정밀도 모두 높음.
  AUC 또한 모든 클래스에서 0.7~0.89 수준으로 양호.
  다만 Mid 클래스에서 AUC 0.698로 다소 낮으며, 민감도와 정밀도도 다른 클래스에 비해 상대적으로 낮음
  → **중간 구간 예측이 다소 약함**.
- **결정나무**에서는 **High 클래스**에서 AUC 0.8103으로 괜찮은 성능을 보이나,
  Low, Mid 클래스에서 민감도·정밀도 모두 낮음 (예: Low 정밀도 0.5556).
  AUC가 클래스 간 편차가 크고 Mid에서 0.6811로 가장 낮음 → **불안정한 분류 성능을 시사**.

결론
→ **High 클래스**는 모델이 상대적으로 **일관적으로 잘 구분**되는 반면, **Mid 클래스**는 **오분류 가능성이 높음**.
또한 **Low 클래스**는 **모델별 강점이 달라, 모델 선택에 따라 성능 편차가 존재**함.

0.2.3 모델별 ROC 커브

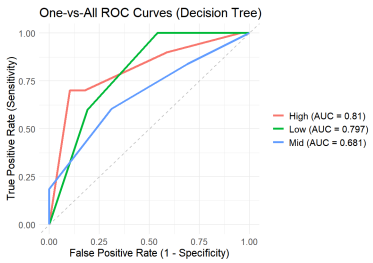

Figure 1: *
Decision Tree

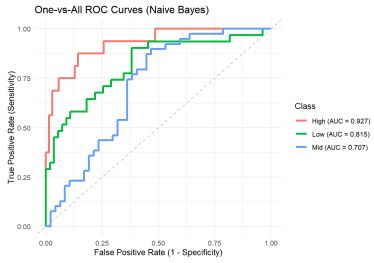

Figure 2: *
Naive Bayes

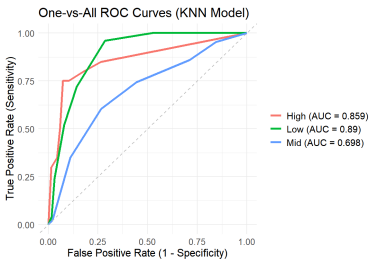

Figure 3: *
KNN

Figure 4: 모델별 ROC 커브

0.3 3. 결론

본 보고서는 **NBA 선수 데이터**를 기반으로 연봉 구간 분류를 위해 **KNN**, **Naive Bayes**, **결정나무** 모델을 사용하였다. 성능 비교 결과

- **Naive Bayes** 모델이 교차검증과 검증 셋 모두에서 가장 안정적이고 우수한 성능을 보였다.
- **KNN** 모델은 **Low** 클래스에서 상대적으로 좋은 성능을 보였으나 **Mid** 클래스에서의 분류 정확도가 낮았다.
- **결정나무** 모델은 변수 **선택 없이** 단순 구조로 적용되었지만, 정확도와 안정성 모두에서 다른 모델보다 낮은 성능을 보였다.

추가로 **Mid 클래스에 대한 낮은 성능**은 클래스 불균형 문제의 영향일 가능성이 있으며, 이를 위한 추가 조치가 필요할 수 있다.

결론적으로, 전체적인 성능과 안정성 측면에서 **나이브 베이즈 모델**이 가장 적절한 분류 모델로 평가되며, 상황에 따라 **KNN**을 보완적으로 활용할 가능성도 있다.