최종 분석 결과 정리

영화 장르 분류 모델을 구축하고 다양한 최적화 기법을 적용해 보았지만, Accuracy가 크게 향상되지 않았습니다. 이는 데이터셋의 특성, 모델의 한계, 또는 최적화 기법의 부적합성 등 다양한 원인이 복합적으로 작용했을 가능성이 있습니다.

**주요 문제점:**

1. **데이터셋의 복잡성:** 영화 장르는 명확하게 구분되지 않는 경우가 많고, 여러 장르가 혼합되어 있는 경우가 많습니다. 이러한 데이터셋의 복잡성은 모델 학습을 어렵게 만듭니다.
2. **텍스트 데이터의 한계:** 영화 줄거리만으로 장르를 완벽하게 예측하기는 어려울 수 있습니다. 감독, 배우, 포스터 이미지 등 추가적인 정보가 필요할 수 있습니다.
3. **모델의 한계:** 랜덤 포레스트는 다중 레이블 분류에 적합하지 않을 수 있습니다. 다른 모델 (예: 딥러닝 모델)을 사용하는 것이 더 효과적일 수 있습니다.
4. **최적화 기법의 한계:** GridSearchCV를 사용하여 하이퍼파라미터를 튜닝했지만, 최적의 하이퍼파라미터를 찾지 못했을 수 있습니다. 더 넓은 범위의 하이퍼파라미터를 탐색하거나 다른 최적화 기법을 적용해야 할 수 있습니다.

**개선 방향:**

1. **데이터 전처리:** 텍스트 데이터 전처리 과정을 개선하여 노이즈를 줄이고, 의미 있는 특징을 추출해야 합니다. 예를 들어, 어간 추출, 표제어 추출, 불용어 제거, 품사 태깅 등을 적용할 수 있습니다.
2. **특징 엔지니어링:** 영화 제목, 감독, 배우, 포스터 이미지 등 추가적인 특징을 추출하여 모델에 입력해야 합니다. 외부 데이터베이스에서 영화 정보를 가져와 특징으로 사용할 수도 있습니다.
3. **모델 변경:** 랜덤 포레스트 대신 다중 레이블 분류에 적합한 다른 모델 (예: ClassifierChain, MultiOutputClassifier, 딥러닝 모델)을 사용해야 합니다.
4. **최적화 기법:** 더 넓은 범위의 하이퍼파라미터를 탐색하거나 다른 최적화 기법 (예: RandomizedSearchCV, Bayesian Optimization)을 적용해야 합니다.
5. **데이터 증강:** 데이터 증강 기법을 사용하여 데이터셋의 크기를 늘려야 합니다. 텍스트 데이터의 경우, 동의어 치환, 문장 순서 변경 등의 방법을 사용할 수 있습니다.
6. **클래스 불균형 해소:** 샘플 수가 적은 장르에 대한 데이터를 추가하거나, 오버샘플링 기법을 적용하여 클래스 불균형 문제를 해결해야 합니다.

**결론:**

영화 장르 분류는 복잡한 문제이며, 높은 정확도를 달성하기 위해서는 다양한 노력이 필요합니다. 데이터 전처리, 특징 엔지니어링, 모델 선택, 최적화 기법 등을 개선하고, 데이터 증강 및 클래스 불균형 해소를 통해 모델 성능을 향상시킬 수 있습니다.

하지만, 완벽한 정확도를 달성하기는 어려울 수 있으며, 실제 환경에서 사용하기 위해서는 모델의 성능뿐만 아니라 속도, 안정성, 유지보수 등 다양한 요소를 고려해야 합니다.