ARABIC MOVIE RECOMMENDATION SYSTEM

Team Members:

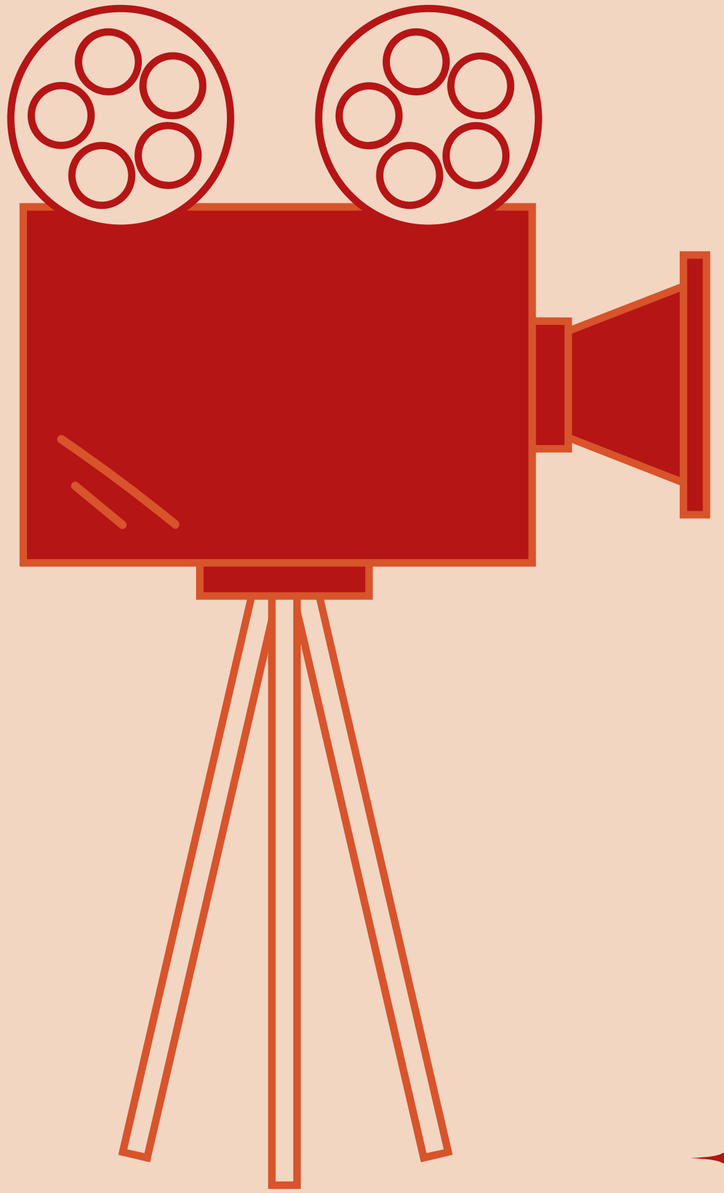
Lana Altaweel
Hanin Alturki
Ghadi Alayed
Reema Almutairi
Reema Alojairy

PROBLEM STATEMENT

Arabic cinema especially Egyptian movies has a long and rich history, but modern digital tools for discovering and recommending Arabic films are still limited.
Many viewers struggle to find movies that match their preferences because most recommendation systems are designed for Hollywood or global cinema, not Arabic content.

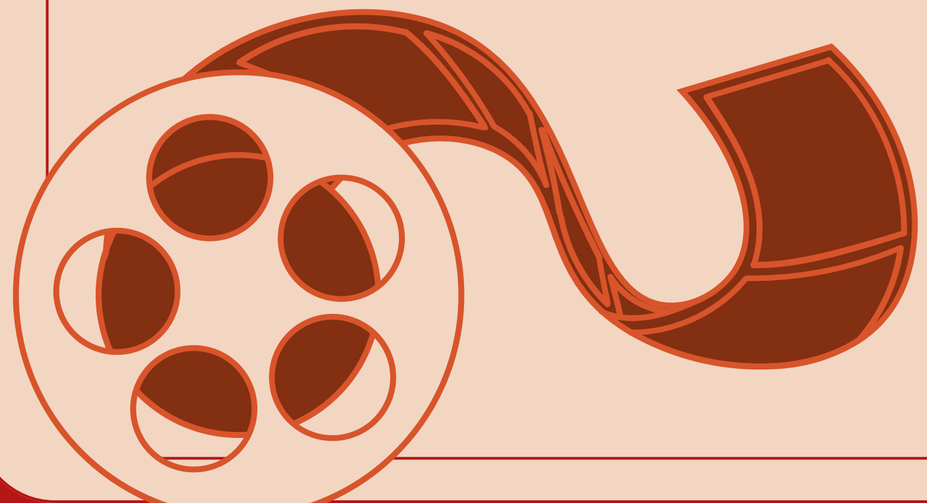
GOAL

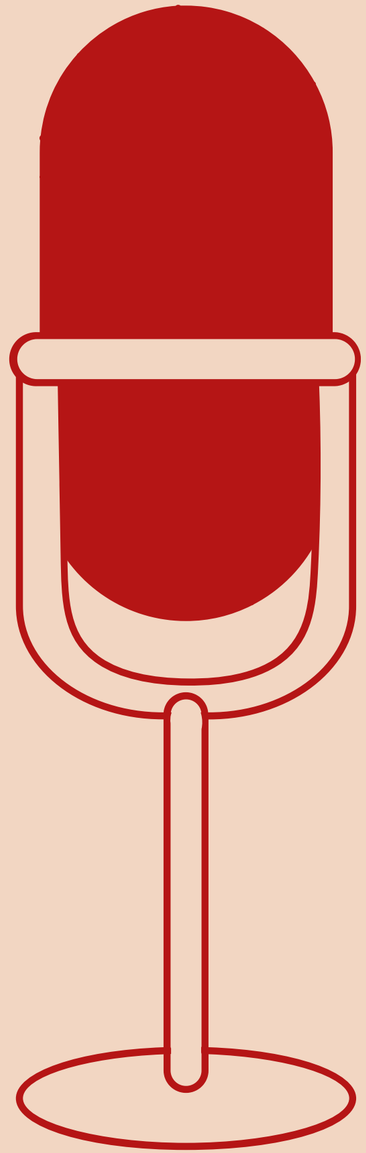
Is to build an Arabic Movie Recommendation System based specifically on Egyptian movies. The system aims to help users discover films they are likely to enjoy.



IMPACT

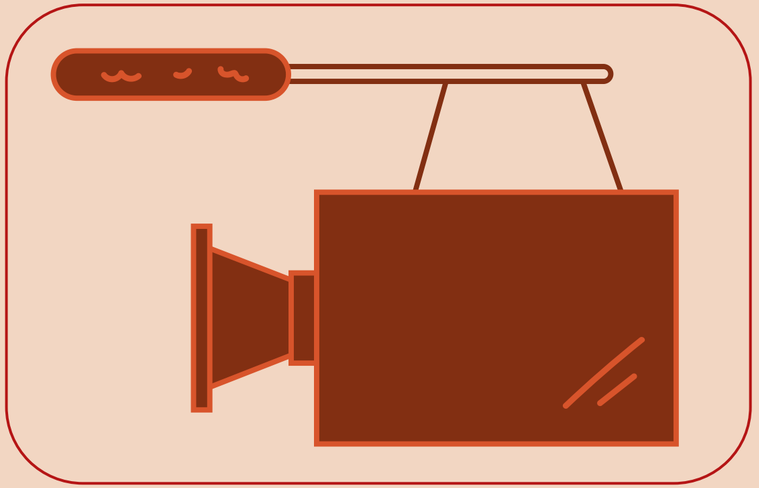
- Personalized recommendations
- Cultural Impact





DATASET

The dataset used in this project is CIMA.xlsx, which contains 2,368 Egyptian movies and 15 columns collected from an online repository.



Dataset Source: Kaggle, “Data Analysis of Egyptian Arabic Movies”

Main Columns in the Dataset

- Movie Title
- Genre
- Release Year
- Duration
- summary
- Cast

Data Types

- Textual
- Categorical
- Numerical

OUR APPROACH

Data Cleaning

- Handle missing values and duplicates
- Normalize Arabic text

EDA

- Genre distribution
- Text length analysis
- Identifying imbalance

Supervised Learning

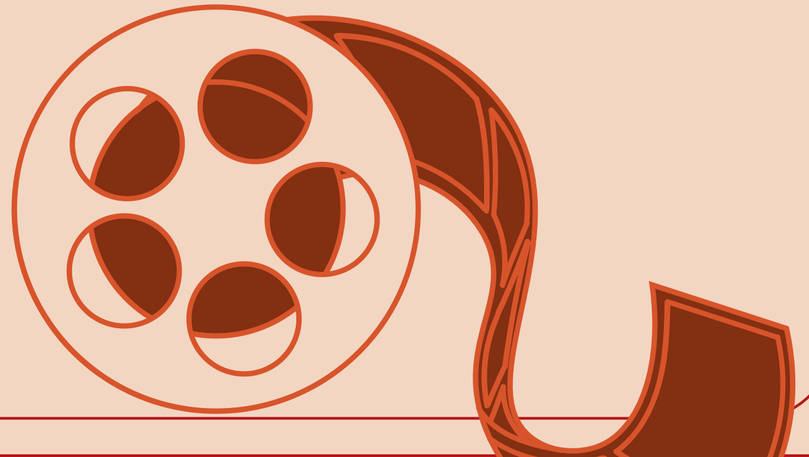
- Convert text to numerical features (TF-IDF)
- Train and evaluate classification models
- Model evaluation

Unsupervised Learning

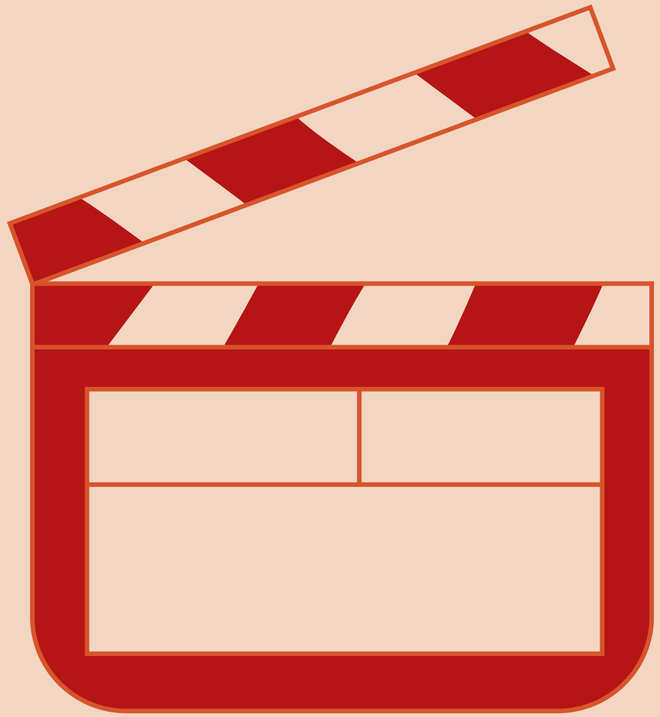
- Cluster movies to explore hidden patterns
- Evaluate cluster quality and interpret results

Generative AI Integration

- Use AI to recommend a movie.
- Design and test prompt templates.



TOOLS USED



Python

Pandas & NumPy: numerical operations processing

Scikit-learn: ML models, clustering

NLTK / Arabic NLP tools

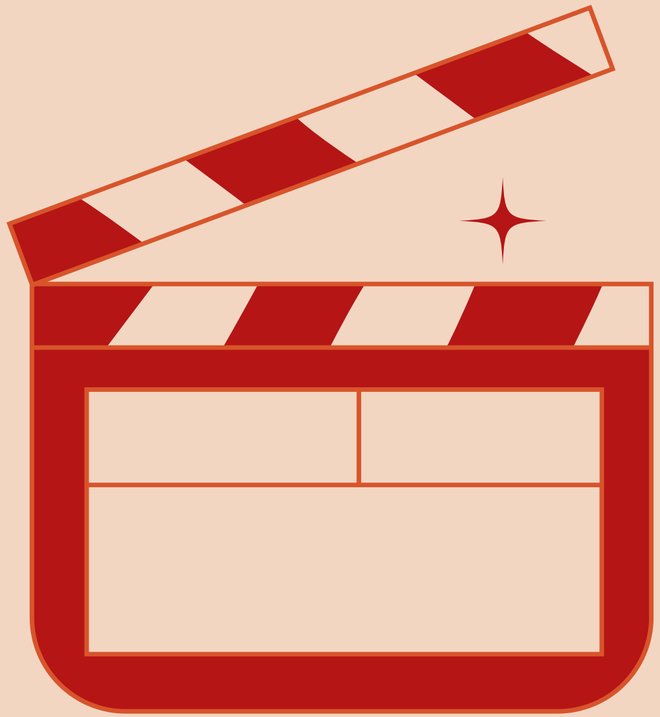
Matplotlib , Seaborn & Plotly: Data visualization

Groq

Google colab

Github

FINDING AND ANALYSIS



PHASE 1&2

Our goal we wanted to understand the structure of Egyptian movie data and test whether AI can use film features (text + metadata) to support meaningful movie prediction.

What we explored

- Analyzed 2,368 Egyptian Arabic films.
- Columns included: title, genre, duration, summary, actors, directors, writers, and production crew.
- Examined distributions, missing values, inconsistencies, and overall data quality.
- Cleaned Arabic text, normalized spelling, and handled missing entries.
- Created TF-IDF text features from movie summaries.
- Added numeric features (duration, year).
- Trained supervised learning models to measure how well structured features can support genre prediction.

Key Insights:

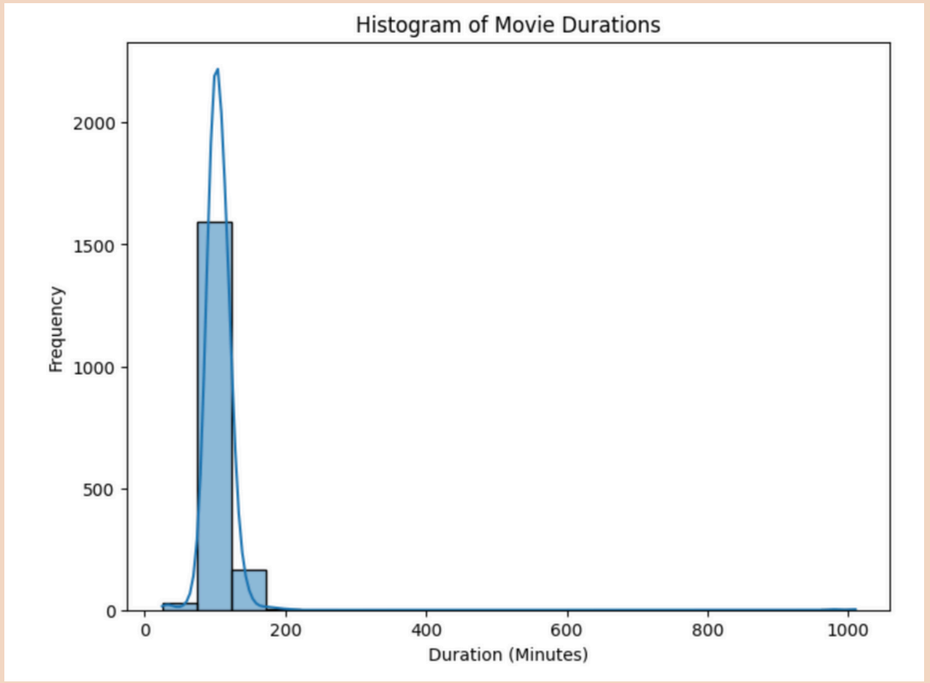
- Most movies fall in the 80–120 minute range.
- The dataset is heavily imbalanced, with Drama dominating the catalog.
- Arabic text required extensive preprocessing (normalization + cleaning).
- Movie metadata alone is not strong enough to make high-quality recommendations on its own.
- Supervised learning helped extract useful patterns, but true personalization needed a more advanced method(leading to Phase 4’s generative AI recommender).

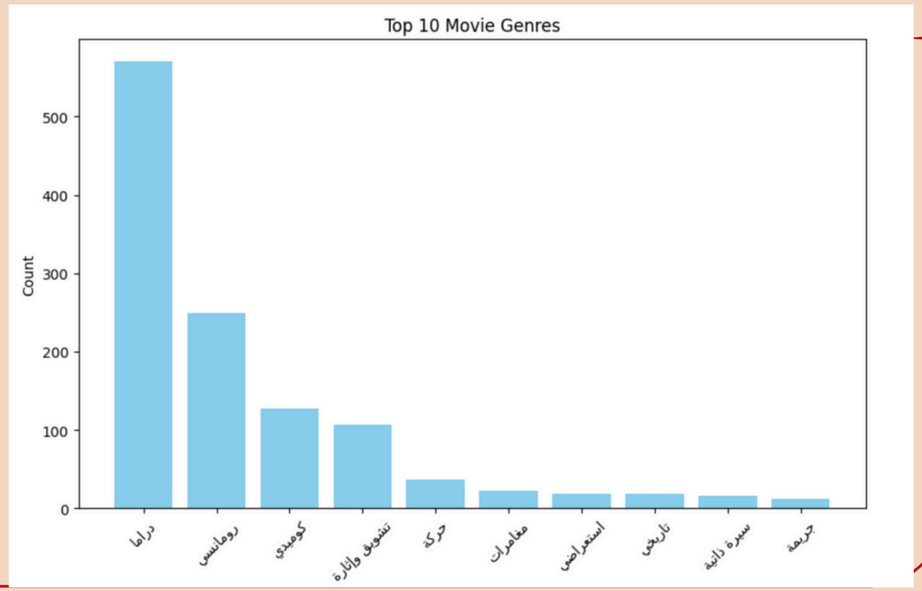
PHASE 1 DATA EXPLORATION RESULTS LEARNING RESULTS

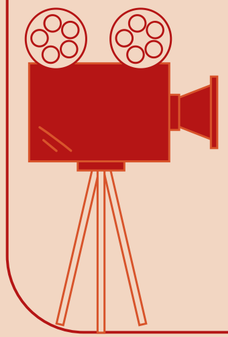
- Cleaned and Prepared the Data

1. Fixed inconsistent formats (e.g., duration, year).
2. Handled missing values in several columns.
3. Removed duplicates and corrected obvious data-entry errors.
4. Ensured numeric fields were valid and usable.

- Most films fall between 80–120 minutes → typical movie length.
- The dataset is heavily imbalanced → Drama dominates all other genres.
- Some movies lacked complete metadata → required imputation.
- Dataset is rich, but not evenly distributed across genres.





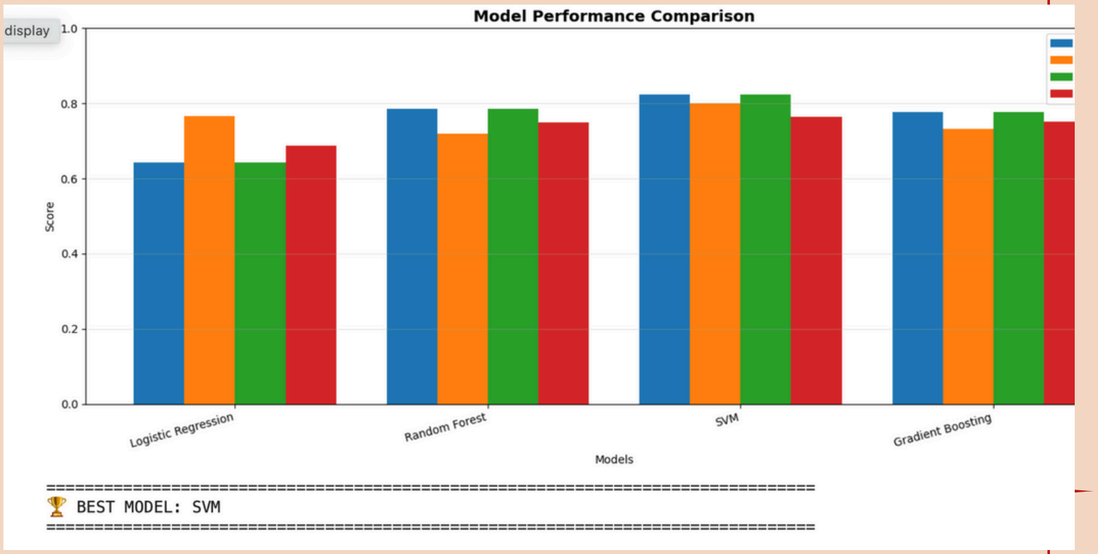


PHASE 2 SUPRIVISED LEARNING

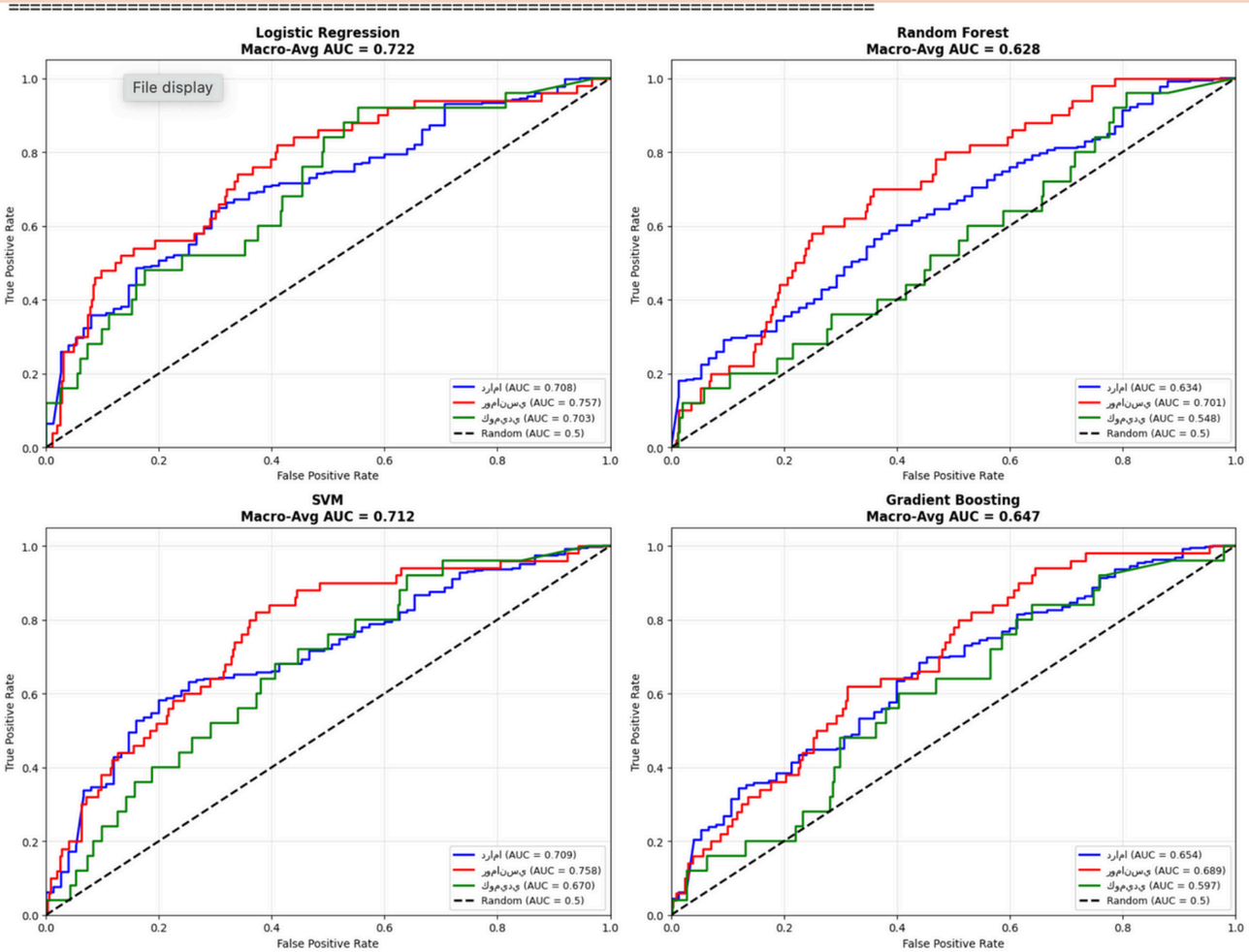
- Focused the task on the top 3 genres: دراما، رومانسي، كوميدي to reduce extreme imbalance.
- Converted movie summaries (ملخص) into numeric features using TF-IDF (unigrams + bigrams, 500 features).
- Applied SMOTE on the training set to oversample Romance and Comedy and create a balanced training set.Trained four models with imbalance-aware settings:

1. Logistic Regression
2. Random Forest Classifier
3. Support Vector Machine (SVM)
4. Gradient Boosting Classifier

- Best model: SVM with ~82% accuracy and ~76% weighted F1-score on the test set.
- Algorithm choice is good (SVM works best for our TF-IDF text features).
- Data imbalance is the bottleneck → the model over-predicts Drama because that is what it sees most.



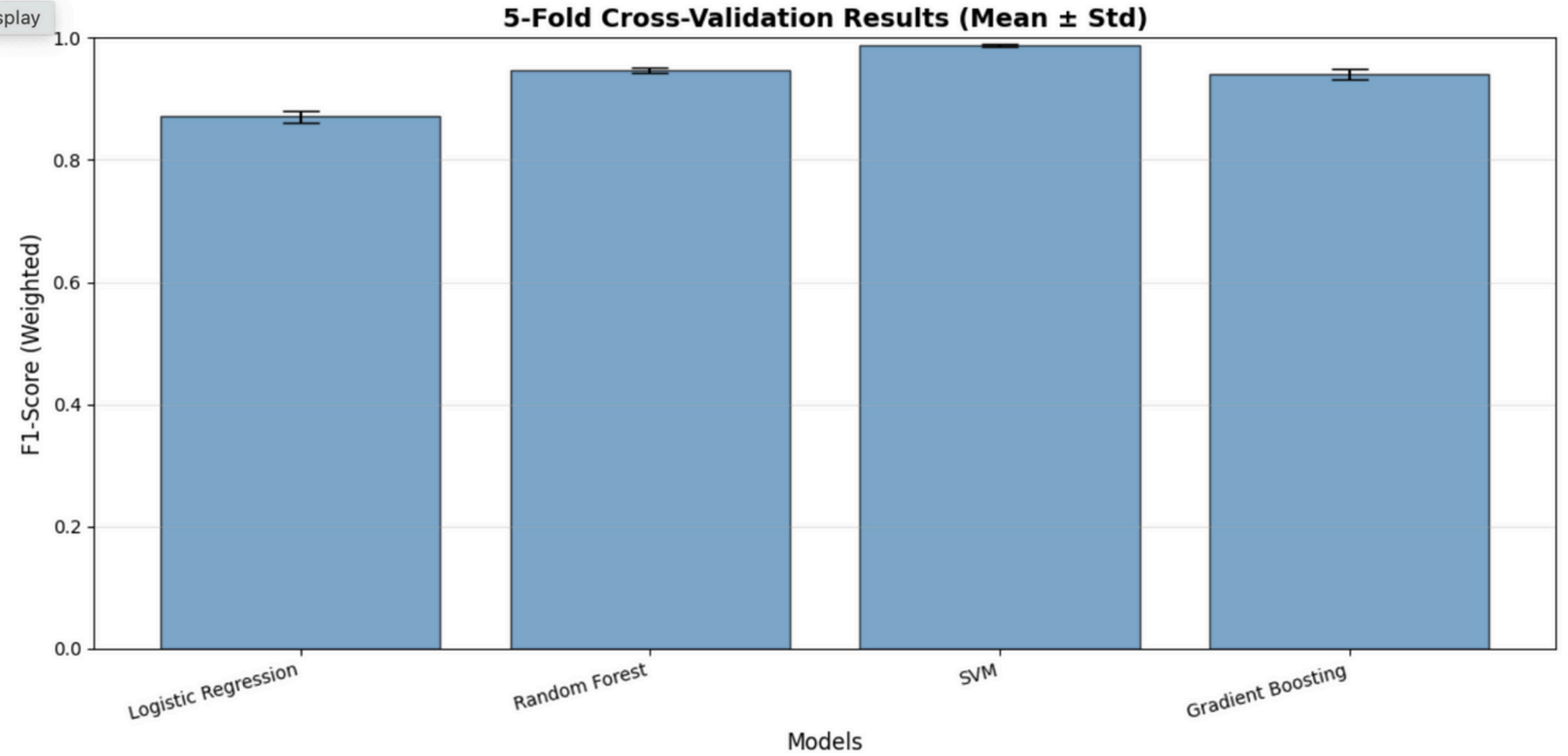
PHASE 2 SUPRIVISED LEARNING



✅ ROC-AUC curves show model performance across all genres
- Higher AUC = Better discrimination ability

PHASE 2 SUPRIVISED LEARNING

File display

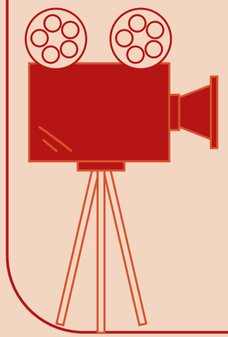


✅ Cross-validation provides robust performance estimates
- Mean: Average performance across 5 folds
- Std: Performance variability (lower = more stable)
- Error bars show consistency across different data splits

PHASE 2 SUPRIVISED LEARNING

File display

Model	Accuracy	F1-Score
SVM	82.34%	76.47%
Random Forest	78.52%	74.97%
Gradient Boosting	77.80%	75.21%
Logistic Regression	64.20%	68.66%



PHASE 3&4

Our goal was to understand whether Arabic film data has a hidden structure and whether we can use AI to give meaningful movie advice.

We explored two different approaches:

(1) Clustering to see if movies naturally group together

(2) Generative AI to provide personalized recommendations

Each phase revealed important insights that helped us understand what works best for Arabic film data

Key Insights:

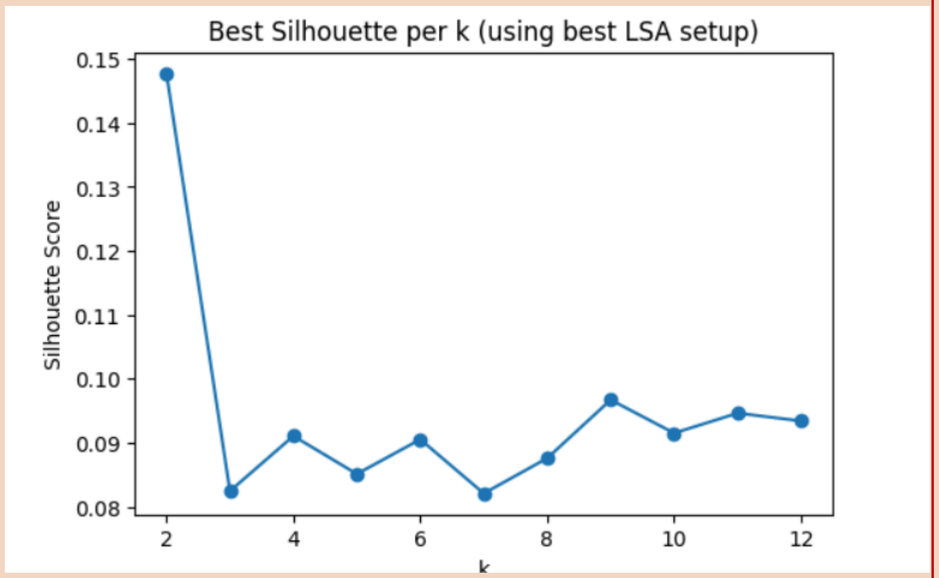
- Clustering struggled because films share similar vocabulary.
- Generative AI understood user intent and produced helpful recommendations.
- The system shifted from simple filtering → to real personalized advice.

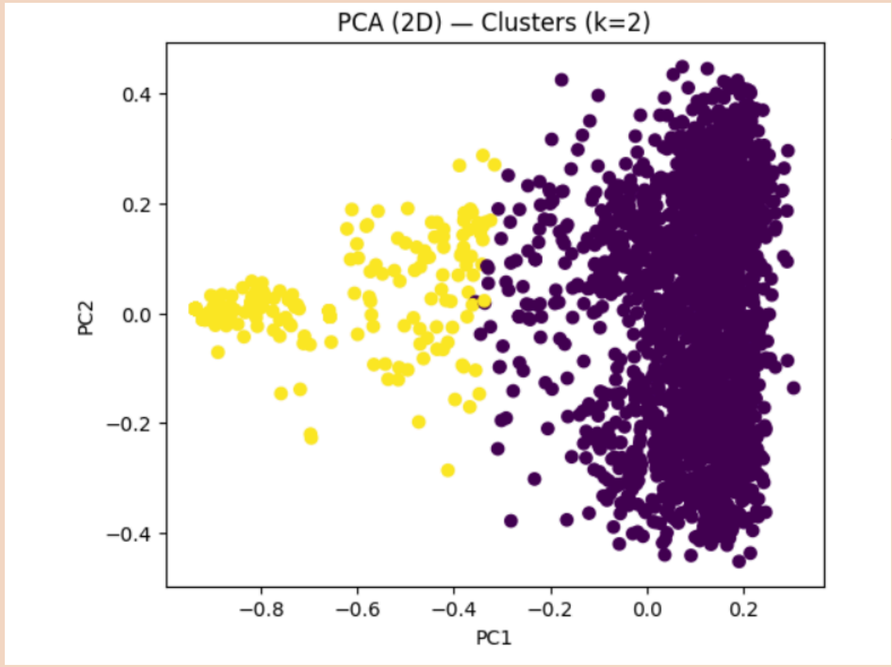
PHASE 3 UNSUPERVISED LEARNING RESULTS

- Cleaned Arabic text
- Combined: ملخص + تمثيل + تأليف + إخراج.
- Applied K-Means with k = 2 → 12.

Actual Results :

- Best configuration: LSA = 50 + k = 2
- Silhouette score = 0.148 → very weak separation
- Clusters contain mixed genres → not meaningful
- BCubed Precision/Recall low → clustering does NOT reflect genre



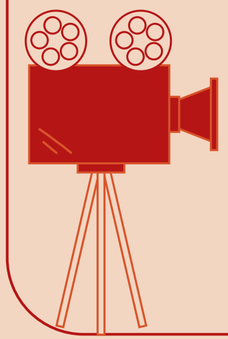


PHASE 4 PERSONALIZED MOVIE RECOMMENDATION WITH AI

- Integrated Llama 3.1 model via Groq API.
- The user writes a description (mood, duration, family-friendly, preferred genre).
- We first filter movies (genre, duration, year).
- Then we send 8 candidate movies to the AI model.

Two Templates Tested:

- Template A: One recommendation + short reason
- Template B: Top 3 recommendations + deeper reasoning



```
========== TEMPLATE A (نصيحة مباشرة) ==========
الفيلم المقترح: شمشون ولبلب
التصنيف: كوميدي
سبب الترشيح: هذا الفيلم مناسب للمستخدم لأنها يبحث عن فيلم كوميدي خفيف يمكنه مشاهدة مع العائلة، ومدة الفيلم أقل من ساعتين، وهو أيضًا فيلم مصري يعتبر من الأفضل في هذا النوع.
مناسب للعائلة؟: نعم


========== TEMPLATE B (ترتيب + شرح مفصل) ==========
1) اسم الفيلم 1>: شمشون ولبلب - السبب: هذا الفيلم مناسب للمستخدم بسبب كوميديته الخفيفة والرومانسية الموجودة فيه. كما أنه يحتوي على شخصيات محببة ومفارقات كوميدية ساخرة. يعتبر> هذا الفيلم مناسباً للمزاج الخفيف للمستخدم ويمكن مشاهدته مع العائلة.

2) اسم الفيلم 2>: حلوة وشقية - السبب: هذا الفيلم مناسب للمستخدم بسبب كوميديته الخفيفة والرومانسية الموجودة فيه. كما أنه يحتوي على شخصيات محببة ومفارقات كوميدية ساخرة. يعتبر> هذا الفيلم مناسباً للمزاج الخفيف للمستخدم ويمكن مشاهدته مع العائلة.

3) اسم الفيلم 3>: خلي بالك من جيرانك - السبب: هذا الفيلم مناسب للمستخدم بسبب كوميديته الخفيفة والرومانسية الموجودة فيه. كما أنه يحتوي على شخصيات محببة ومفارقات كوميدية ساخ> يعتبر هذا الفيلم مناسباً للمزاج الخفيف للمستخدم ويمكن مشاهدته مع العائلة.

نصيحة عامة: يبدو أن المستخدم يناسبه أفلام كوميدية خفيفة ورومانسية، لذلك يمكنه مشاهدة أفلام مثل الأفلام الموجودة في القائمة أو أفلام أخرى من نفس النوع.
```

CONCLUSION & RECOMMENDATIONS

CONCLUSION & RECOMMENDATIONS

Conclusion

- Dataset is rich but highly imbalanced, with Drama dominating all other genres.
- Supervised learning showed that SVM performed best (~82% accuracy), but minority genres (Romance, Comedy) remain hard to classify due to limited data.
- Unsupervised clustering (K-Means + LSA) failed to separate genres, confirming that summaries share similar vocabulary.
- Generative AI produced the strongest recommendations, offering personalized, context-aware movie suggestions beyond traditional ML techniques.

Recommendations

- Expand the dataset, especially Romance and Comedy, to reduce imbalance and improve model fairness.
- Add more metadata (keywords, ratings, posters, user behavior) to strengthen prediction accuracy.
- Use advanced text embeddings (AraBERT, LLaMA embeddings) instead of TF-IDF for better semantic understanding.
- Build a hybrid recommender (content-based + collaborative + generative AI) for Netflix-style personalization.
- Develop a simple UI to turn the system into an interactive, user-ready product.

THANK YOU

Movies Are More
Than Entertainment