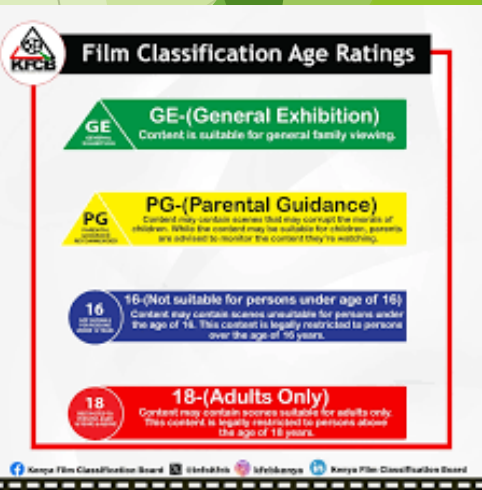
# Automating Classification of Audio-Visual Content and Rating for Regulation and Personal Use


DSC-PT-10

Authors:

- Gibson Ngetich
- Cindy Minyade
- Ayaya Vincent
- Maryan Daud
- Edwin Korir
- Margaret Njoroge

July 25th, 2025



# PROJECT OVERVIEW

- Develop a machine learning model to automate film classification based on genre, synopsis, platform, and other metadata.
- Enhance regulatory efficiency by providing intelligent rating suggestions aligned with Kenya's classification guidelines.
- Support parental control tools and age-based content filtering for safer content consumption.
- Promote discoverability and categorization of local content to increase visibility of Kenyan productions.

# Objectives

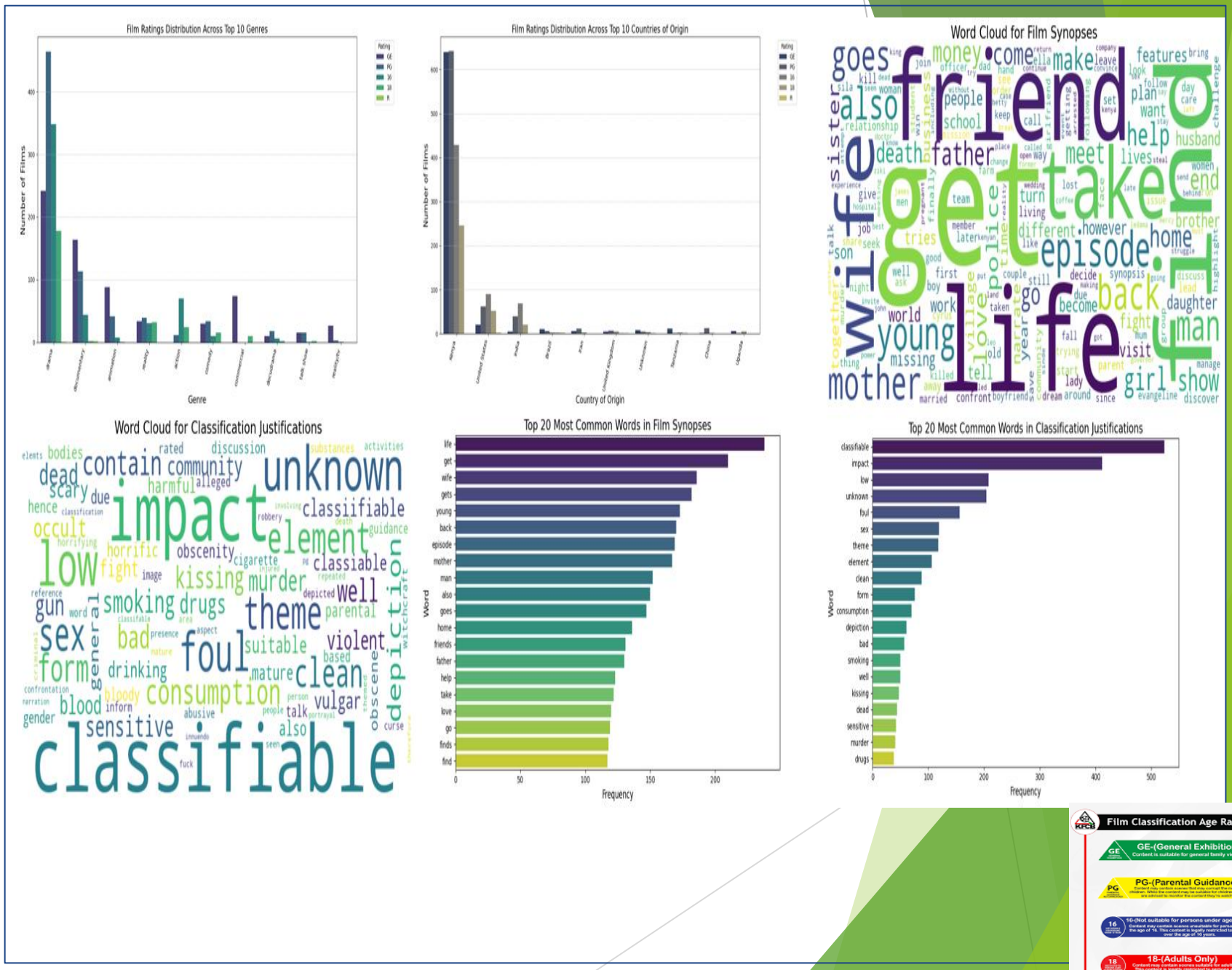
This project is aimed at assisting the regulator in automating the classification of audiovisual content in Kenya.

1. Automate film classification using machine learning to predict appropriate age ratings.
2. Improve regulatory efficiency by providing AI-based rating suggestions.
3. Support parental controls through age-based content filtering tools.
4. Promote and recommend local content through enhanced categorization and discoverability.

# DATA USED

**The data used for this project was collected by the Kenya Film Classification Board (KFCB) between July 2022 and June 2025.**
It was sourced from official classification records containing film metadata and regulatory decisions across different platforms (cinema, TV, online streaming).
For purposes of this study, the data was consolidated and structured as follows:
A unified dataset comprising 2,574 film records with 15 key attributes, including film title, genre, classification, synopsis, rating, and origin.
The data was cleaned and standardized to support analysis of content patterns, classification trends, and audience suitability.

# EDA Summary


Film Ratings Distribution Across Top 10 Genres
Number of Films
Genre
Film Ratings Distribution Across Top 10 Countries of Origin
Number of Films
Country of Origin
Word Cloud for Film Synopses
Word Cloud for Classification Justifications
Top 20 Most Common Words in Film Synopses
Word
Frequency
Top 20 Most Common Words in Classification Justifications
Word
Frequency


Film Classification Age Ratings
GE-(General Exhibition)
PG-(Parental Guidance)
16-(Not suitable for persons under age of 16)
18-(Adults Only)

# MODELLING

## Logistic Regression Model(Baseline)

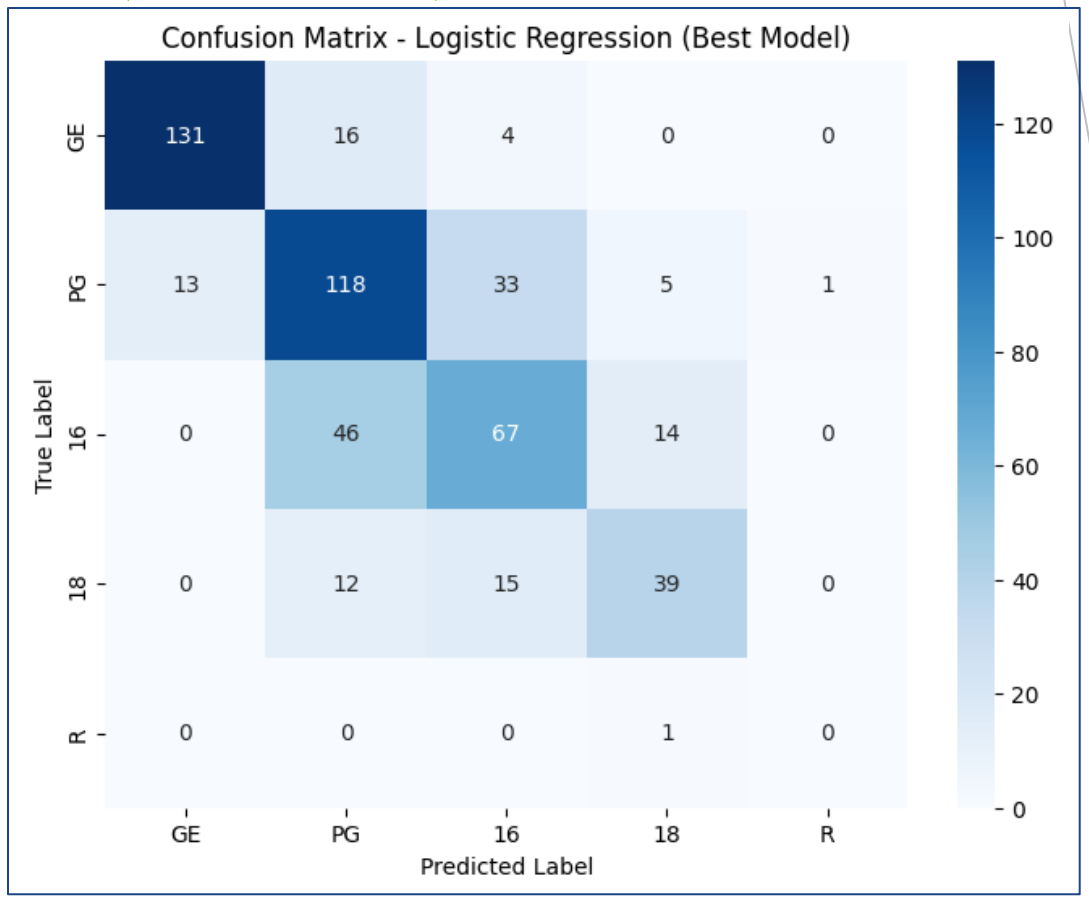
**1. Logistic Regression**

- **Best Parameters**: `C=1`, `solver='lbfgs'`

- **Accuracy**: 0.69

- **Best F1-Weighted Score (CV)**: 0.69

- **Notes**: Serves as a solid baseline with decent precision for 'GE' and 'PG' classes. Underperforms for rare classes like 'R'.




- The model performs well in predicting GE ratings, with minimal misclassifications. However, PG, 16, and 18 show significant overlap many PG films are misclassified as 16, and vice versa suggesting feature similarity in borderline content. The model struggles to correctly classify Restricted films due to the very limited training data in that class. Improving class balance and incorporating more discriminative features could enhance overall rating accuracy.

# MODELLING

## XGBoost Classifier Best Model

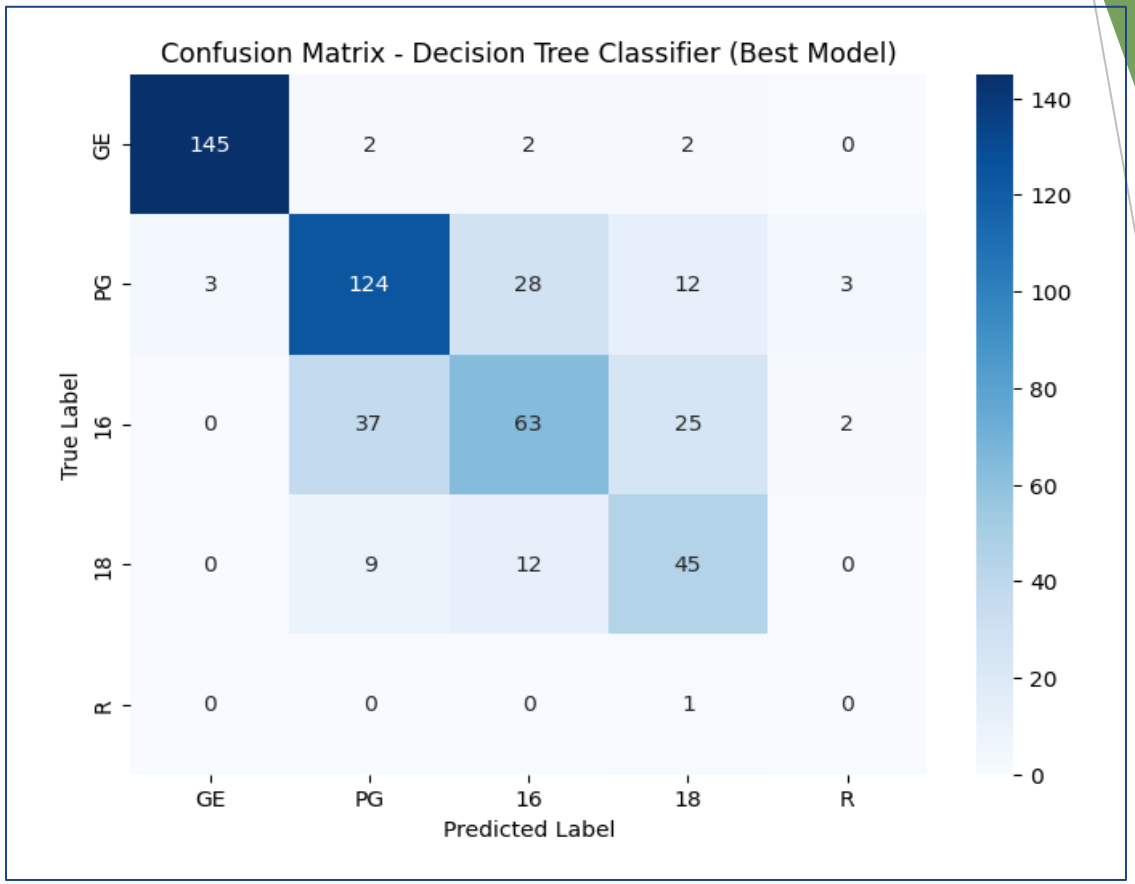
**XGBoost Classifier**

- Accuracy: **77.48%**

- F1-Weighted Score (Cross-Validation): **0.7472**

- Excellent balance between performance and interpretability

- Robust against overfitting and handles both categorical and numerical features effectively




**High True Positives** across major classes like "PG" and "GE", indicating strong classification performance on the most frequent categories.

- **Misclassifications** are more common between similar or adjacent rating classes (e.g., "16" misclassified as "18"), suggesting overlap in content characteristics.

- The **"R" and "18" classes**, which are less represented, show slightly lower recall—typical in imbalanced datasets.

# MODELS EVALUATION TABLE

| Model | Best Parameters | Accuracy | F1-Weighted | Notes |
|---|---|---|---|---|
| Logistic Regression | C=1, solver=lbfgs | 0.69 | 0.69 | Baseline; weak on rare class "R" |
| Decision Tree | max_depth=None | 0.70 | 0.71 | Captures "18"; risk of overfitting |
| Random Forest | n_estimators=200 | 0.76 | 0.76 | Balanced; strong overall |
| XGBoost | lr=0.1, n_estimators=200 | 0.77 | 0.76 | Top performer; efficient |
| LightGBM | lr=0.1, n_estimators=200 | 0.75 | 0.75 | Fast; comparable to XGBoost |
| Naive Bayes | alpha=1.0 | 0.75 | 0.75 | Great with text; good for "16" |

# CONCLUSION

- Successfully built a machine learning model to classify films based on age-appropriateness using KFCB guidelines.
- XG Boost classifier  performed best (Accuracy: **77.48%**, F1: **0.772**).
- Text features like synopses and justifications were key in improving prediction.
- EDA revealed rating patterns across genres, platforms, and countries.
- The solution supports regulators, parents, and content platforms in faster, scalable, and objective classification.

# RECOMMENDATIONS

- Use the ML model as a **pre-screening tool** for faster content review.
- **Integrate API** with KFCB or streaming platforms for real-time classification.
- Switch to **transformer models** (e.g., BERT) for better text analysis.
- Apply **SMOTE or class weighting** to handle rating imbalance.
- Add **human-in-the-loop** feedback to improve accuracy over time.
- Build a **parental control app** to help filter content by rating.
- Include **image/audio features** for richer content classification.
- Perform **regular audits** to detect and correct model bias.

# Next Steps

**Advanced NLP**: Use BERT/RoBERTa for better text understanding
**Multimodal**: Add image/audio features
**API & Dashboard**: Deploy for real-time use
**Bias Audit**: Fix class imbalance, monitor fairness
**Human Feedback**: Improve model via user input
**Scaling**: Localize for other countries/languages


## 🙏 Thank You

Thank you for your time and attention.
We appreciate your support and interest in our project.
— *The Team*