IEESE International Journal of Science and Technology (IJSTE), Vol. 13 No. 1, March 2024, 7-13
ISSN : 2252-5297

Age Rating Classification for Film Content: A Robust Approach with XGBoost

Akhmad Badawi[1]

[1] Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, **Indonesia**
[1]Email: akhmadbadawi04@gmail.com

Abstract. Movies serve as a primary source of entertainment globally, offering a vast array of content. However, not all content suits every age group, making age ratings essential to ensure viewers access age-appropriate material. These ratings are crucial as they affect the reach and profitability of production houses. This study examines the content types that significantly impact film age ratings using the XGBoost algorithm. The content types analyzed include nudity, violence, profanity, alcohol, and frightening elements. XGBoost was selected for its efficiency and ability to assess feature importance, offering insights into how each content type influences age rating classification.
The study utilized a dataset from IMDb, containing 6,178 films. The data was split into 75% for training and 25% for testing to develop and validate the model. The analysis indicates that frightening content is the most critical factor in determining age ratings, followed by nudity, violence, profanity, and alcohol. These results suggest that understanding the impact of various content types can help film producers create content that aligns with the desired age ratings, effectively targeting appropriate audiences.
Key Words: *Film Age Ratings, Content Impact Analysis, XGBoost Classification*

1. Introduction

Film is one of the forms of art that has become an integral part of modern society. The film industry generates billions of dollars in its production because films serve as a source of entertainment in any country [1]. As an audiovisual medium, film not only acts as a primary means of entertainment but can also be a tool to convey messages, inspire, and influence viewers. Due to its significant impact, it is crucial for the film industry to ensure that the content displayed in a film is appropriate for the maturity and understanding levels of the intended audience.

In various countries, there are film rating systems. These ratings aim to provide guidance to viewers about the suitability of a film for different age groups [2]. The process of determining a film's rating can refer to several factors, one of which is the content within the film. However, there are no detailed explanations specifying the types of content that can influence the age rating of a film.

In a study by Latif and Afzal [1], a comparison of machine learning methods for classifying the success prediction of a film was presented. These methods include Forest, Decision Tree, K-Nearest Neighbours (KNN), NLP, XGBoost Classifier, and Deep Neural Network. In that paper, XGBoost was identified as the classification method with the highest accuracy rate of 90%. Additionally, this method can also determine the significance of the impact provided by a feature using feature importances.

Referring to that study, this research aims to explore the use of the XGBoost Classifier algorithm in a similar context. The use of the XGBoost Classifier in this study will attempt to identify which content has the most significant impact among the five types of content used as features in the classification of film age ratings. The five contents analysed include Nudity, Violence, Profanity, Alcohol, and Frightening. The subsequent structure of this research consists of section 2, which contains the research methodology. The third section presents the testing results, followed by a discussion in section 3. Finally, the conclusion of this research is presented.

2. Proposed Methodology

The research methodology is the phase in a study that is organized systematically to tackle and resolve research issues.. The detailed stages in Figure 1 of the approach in this research include: 1) Data Collection; 2) Data Preprocessing; 3) Data Splitting; 4) Model Selection; 5) Model Training; 6) Model Evaluation [3].

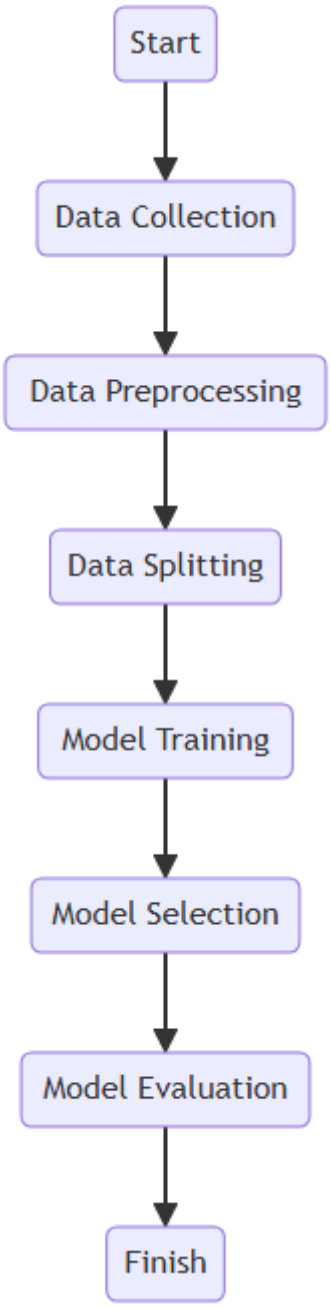

Figure 1. Research Design

2.1. Data Collection

This research utilizes a dataset obtained from IMDb, consisting of 6,178 films. The dataset was sourced from the well-known dataset site, Kaggle.com. Each film in the dataset is accompanied by content-related information such as Nudity, Violence, Profanity, Alcohol, and Frightening. This dataset also includes age rating labels commonly used in the film industry.

2.2. Data Preprocessing

The data preprocessing stage refers to the process of transforming raw data into a more suitable format for analysis, modeling, or other applications [4]. This stage involves data cleaning, feature encoding, and category encoding to remove noise, handle missing values, and enhance the quality and relevance of the data for its intended use.

2.2.1. Data Cleaning

Data cleaning begins with the removal of unused columns such as title, release year, rating, votes, genre, duration, type, and number of episodes. Cleaning continues by deleting rows with empty data or invalid values, such as "No Rate," which means unrated.

2.2.2 Feature Selection

Feature selection involves choosing a subset of the most relevant features and discarding less informative ones to reduce model complexity. In the context of this research, the most relevant features from the dataset are Nudity, Violence, Profanity, Alcohol, and Frightening.

2.2.2 Feature and Label Encoding

Encoding is the process of converting raw data into a format that can be processed by a model. The format change involves converting categorical data into numerical format. The purpose of encoding is to improve model performance during computation [5].

The feature encoding process is done manually. The manual method used is mapping, which converts categorical values {None, Mild, Moderate, Severe} into numerical values {0, 1, 2, 3}. This step is performed on the five selected features.

Table 1. Features before preprocessing

Nudity	Violence	Profanity	Alcohol	Frightening
None	Moderate	None	Mild	Moderate
Mild	None	Severe	Mild	None
Mild	Moderate	Moderate	None	Mild
No Rate	No Rate	No Rate	No Rate	No Rate
None	Moderate	Mild	Mild	Moderate

Table 2. Features afterpreprocessing

Nudity	Violence	Profanity	Alcohol	Frightening
0	2	0	1	2
1	0	3	1	0
1	2	2	0	1
0	2	1	1	2

The encoding process for the broadcast rating category is carried out using the LabelEncoder method. LabelEncoder is a method for converting categorical labels into numerical labels from the Scikit-learn library in Python. The purpose of LabelEncoder is to assist models that do not support categorical labels by converting them into unique numerical values [6].

Table 3. Label Encoding

Before	After
Banned	0
Approved	1
E	2
G	3
GP	4
M	5
M/PG	6
NC-17	7
PG	8
PG-13	9
Passed	10
R	11
TV-14	12
TV-G	13
TV-MA	14
TV-PG	15
TV-Y	16
TV-Y7	17
TV-Y7-FV	18
Unrated	19
X	20

2.2.3 Model Selection

The data in this study is divided into 75% for training and 25% for testing. The choice of a 75:25 ratio allows the model to learn more general patterns and classify data more effectively. This ratio can help improve model performance and reduce prediction errors [7]. This ratio has been proven to reduce prediction errors and increase model accuracy as shown in studies by Alkaf, Baskara, & Ainiyyah [8] and Karo [9].

The model used in this research is based on the XGBoost algorithm. XGBoost, which stands for eXtreme Gradient Boosting, is an enhancement algorithm based on gradient boosting decision tree algorithms [10]. The algorithm in this model starts by building a simple decision tree model and then calculating its prediction error. This error is used to build the next decision tree model, which aims to better predict the residual errors of the previous model. This process is repeated multiple times, with each new model attempting to reduce the errors of the previous models. This model is chosen for its flexibility with various types and sizes of data, as well as its reliability in handling overfitting. The application of XGBoost in classification can be represented by the following equation 1.

$$\text{Obj}(\theta) = \sum_{i=1}^{n} = l(y_i, \hat{y}_i) + \sum_{k=1}^{K} = \Omega(f_k) \quad (1)$$

Where:
• l is the loss function that measures the difference between the prediction ŷ and the actual value yi.
• Ω is the regularization term that controls the model's complexity to prevent overfitting.
• θ is the model parameter.
• K is the number of trees in the ensemble.

The computational process of the XGBoost algorithm in this research uses the Scikit-Learn library from Python. The implemented model algorithm is the XGBoost Classifier. The use of this library facilitates computation where the model can be easily adapted to the prepared training data. Scikit-Learn also offers flexibility in hyperparameter tuning, where parameters can be easily adjusted and combined with tuning algorithms.

2.2.4 Model Training

The created model is then subjected to tuning for its hyperparameters. The objective of hyperparameter tuning is to find the optimal combination of hyperparameters that yields the best model performance [11]. The tuning process is carried out using the GridSearch method. GridSearch is a simple and easy method for hyperparameter tuning where it involves dividing the search space into a grid containing possible combinations of hyperparameters and evaluating each combination. Although computationally expensive, it is effective for small to medium-sized search spaces [12]. From the several combinations provided, this method will select the most effective combination in training the model.

Table 4. Hyperparameter Combination

Hyperparameter	**Frightening**
learning_rate	0.01, 0.1, 0.2
n_estimators	100, 300, 500
max_depth	3, 6, 9

2.2.5 Model Evaluation

The trained model is then evaluated using the prepared testing data. The performance evaluation results of the model are then captured in a Confusion Matrix. The Confusion Matrix is highly useful in tasks involving multi-class classification where each instance can be labeled with one class. The matrix is a square table with the number of classes as both rows and columns. Rows represent the actual classes of the samples (instances) that are input to the classifier, while columns represent the classes predicted by the classifier [13].

The Confusion Matrix summarizes information from the classification of each class including precision, recall, and support. Precision is a measure of the classifier's accuracy in identifying instances belonging to a particular class [13]. It is computed by dividing the number of true positives by the sum of true positives and false positives ($TP/\ (TP+FP)$). Recall is a measure of the classifier's accuracy in identifying all instances belonging to a particular class [14]. It is computed by dividing the number of true positives by the sum of true positives and false negatives ($TP/(TP+FN)$). F1-score is the harmonic mean of precision and recall, providing a measure of balance between the two [15]. The F1 score is calculated as the average of precision and recall ($2((Precision * Recall)/\ (Precision + Recall))$).

The model evaluation process is also conducted to gather information about the influence of features in the classification process. This stage is carried out using the feature importances function. Feature importances are a

concept used to determine the measure of how much each feature influences the model's output, and this measure is used to identify the most relevant features in the dataset [16].

3. Experimental Result and Discussion

After going through the series of processes in the previous stage, several pieces of information were obtained after evaluating the model.

Table 5. Hyperparameter Combination

Precision	Recall	F1-Score
0.00	0.00	0.00
1.00	0.12	0.22
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.18	0.07	0.11
0.48	0.52	0.50
0.42	0.69	0.52
0.00	0.00	0.00
0.61	0.78	0.68
0.29	0.11	0.16
0.73	0.47	0.57
0.53	0.21	0.31
0.33	0.08	0.13
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00
0.50	0.22	0.31
0.00	0.00	0.00

The report from the confusion matrix indicates that the model is able to predict several classes quite accurately, as seen from the obtained f1-score values. However, some other classes are still biased, which is due to the dataset used not being sufficiently representative to represent certain classes. In another aspect, the evaluation of the model also received a report from feature importances, providing information on the significance of the features used in the classification process.

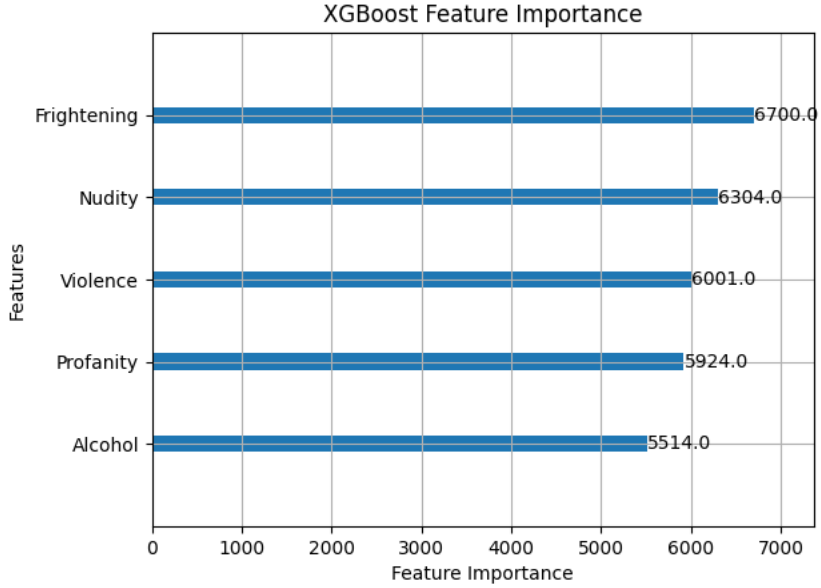

Figure 2. Feature Importance

In the evaluation of feature importances conducted, the values measuring the influence of the features used in the classification model were obtained. Of the five features used in the classification model, it is shown that "Frightening" has the highest value, followed by "Nudity," "Violence," "Profanity," and "Alcohol" respectively.

The higher the value of a feature, the more significant its influence is in determining the rating class of a film screening.

4. Conclusion

The research results indicate that the XGBoost algorithm can be used in classifying the age ratings of films based on the content features of the film. The model evaluation results from the obtained matrix show that the accuracy level can still be improved by referring to the f1-score values, which are still relatively low on average, with only one class receiving a score of 0.6. Accuracy improvement can be achieved by using a more representative dataset for each age rating class. The applied XGBoost algorithm also provides information on how significant a particular type of content is in influencing the age rating label of a film through feature importances. The information obtained from feature importances indicates that the five analyzed types of content all have an impact on the process of determining the age rating of films. The level of influence of these contents, in order, is Frightening with an importance value of 6700, Nudity with 6304, Violence with 6001, Profanity with 5924 points, and Alcohol with a value of 5514. These results can serve as a reference for film production houses in adjusting the content of films if they intend to target specific age groups as their audience.

References

[1] LATIF, M.H., & AFZAL, H., 2016. Prediction of movies popularity using machine learning techniques. International Journal of Computer Science and Network Security, 16(8), pp. 127-131 http://paper.ijcsns.org/07_book/201608/20160820.pdf
[2] DEWANTO, I. S., & MULYADI, D. V., 2020. Efektivitas Flat Design dalam Motion Graphic "Pentingnya Rating Usia Film Bagi Anak". MIND (Multimedia Artificial Intelligent Networking Database) Journal, 5(2), pp. 149-159.
[3] Ahmed, D. M., Hassan, M. M., & Mstafa, R. J. (2022). A review on deep sequential models for forecasting time series data. Applied Computational Intelligence and Soft Computing, 2022(1), 6596397.
[4] ALBAHRA, S., GORBETT, T., ROBERTSON, S., D'ALEO, G., KUMAR, S. V. S., OCKUNZZI, S., ... & RASHIDI, H. H. (2023, March). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. In Seminars in Diagnostic Pathology (Vol. 40, No. 2, pp. 71-87). WB Saunders.
[5] PARGENT, F., PFISTERER, F., THOMAS, J., & BISCHL, B., 2022. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, 37(5), pp. 2671-2692.
[6] ERNAWATI, N. W., KUMARA, I. N. S., & SETIAWAN, W. , 2023. PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE DAN NAÏVE BAYES PADA ANALISIS SENTIMEN KENDARAAN LISTRIK. Jurnal SPEKTRUM, 10(3), pp. 106-114.
[7] OKTAFIANI, R., HERMAWAN, A., & AVIANTO, D. ,2023. Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning. Jurnal Sains dan Informatika, 9(1), pp. 19-28.
[8] ALKAFF, M., BASKARA, A., & AINIYYAH, A., 2023. Penerapan Metode XGBoost Untuk Memprediksi Jumlah Kejadian Kecelakaan Lalu Lintas di Kota Banjarmasin. Generation Journal, 7(1), pp. 61-69.
[9] KARO, I. M. K. (2020). Implementasi metode XGBoost dan feature important untuk klasifikasi pada kebakaran hutan dan lahan. Journal of Software Engineering, Information and Communication Technology (SEICT), 1(1), pp. 11-18.
[10] DARAPANENI, N., BELLARMINE, C., PADURI, A. R., ENTOORI, S., KUMAR, A., VYBHAV, S. V., and MONDAL, K., 2020. Movie success prediction using ml. In 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 0869-0874). IEEE.
[11] KOSKELA, A., & KULKARNI, T. D., 2024. Practical differentially private hyperparameter tuning with subsampling. Advances in Neural Information Processing Systems, 36.
[12] LIN, C. Y., 2023. Multiresponse surface methodology for hyperparameter tuning to optimize multiple performance measures of statistical and machine learning algorithms. Quality and Reliability Engineering International, 39(7), pp. 2995-3013.
[13] AMIN, F., & MAHMOUD, M., 2022. Confusion matrix in binary classification problems: a step-by-step tutorial. Journal of Engineering Research, 6(5), pp. T1-T12.
[14] HEYDARIAN, M., DOYLE, T. E., & SAMAVI, R., 2022. MLCM: Multi-label confusion matrix. IEEE Access, 10, pp. 19083-19095.
[15] FOURURE, D., JAVAID, M. U., POSOCCO, N., & TIHON, S. (2021, September). Anomaly detection: how to artificially increase your f1-score with a biased evaluation protocol. In Joint European Conference on

Machine Learning and Knowledge Discovery in Databases (pp. 3-18). Cham: Springer International Publishing.

[16] LEE, Y. DAN SEO, J. 2023. Suggestion of statistical validation on feature importance of machine learning, Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2023, pp. 1-4. Available at: https://doi.org/10.1109/EMBC40787.2023.10340208