



**UNDERGRADUATE THESIS**

**PREDICTING MOVIE POPULARITY IN  
INDONESIA BASED ON METADATA USING  
GRADIENT BOOSTING**

**APRINIA SALSABILA ROIQOH**  
NPM 22081010166

**THESIS ADVISORS**

Dr. Rizky Parluka, S.Kom., M.Kom.  
Dr. Firza Prima Aditiawan, S.Kom, M.T.I.

**MINISTRY OF HIGHER EDUCATION, SCIENCE, AND TECHNOLOGY  
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR  
FACULTY OF COMPUTER SCIENCE  
INFORMATICS STUDY PROGRAM  
SURABAYA  
2026**

## APPROVAL SHEET

### PREDICTING MOVIE POPULARITY IN INDONESIA BASED ON METADATA USING GRADIENT BOOSTING

By :  
APRINIA SALSABILA ROIQOH  
NPM. 22081010166

Has been defended before, and accepted by, the Board of Assessors of the Thesis Examination of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, on May 12, 2026:

Approved,

**Dr. Rizky Parlika, S.Kom, M.Kom**

NIP. 19840518 202121 1 003



(Advisor I)

**Dr. Firza Prima Aditiawan, S.Kom., MTI**

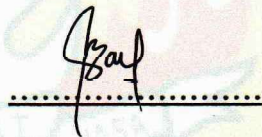
NIP. 19860523 202121 1 003



(Advisor II)

**Made Hanindia Prami Swari, S.Kom, M.Cs.**

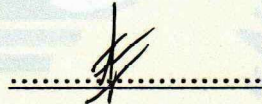
NIP. 19890205 201803 2 001



(Head Assessor)

**Budi Mukhamad Mulyo, S.Kom., M.T.**

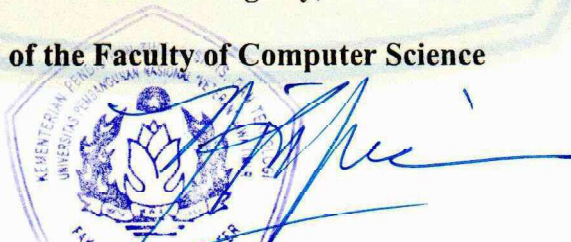
NIP. 19891118 202406 1 003



(Assessor I)

Acknowledge by,

Dean of the Faculty of Computer Science



**Prof. Dr. Ir. Novirina Hendrasarie, MT.**

NIP. 19681126 199403 2 001

## APPROVAL SHEET

### PREDICTING MOVIE POPULARITY IN INDONESIA BASED ON METADATA USING GRADIENT BOOSTING

By :  
APRINIA SALSABILA ROIQOH  
NPM. 22081010166

Approved to proceed to the Thesis Examination

Approved by,

Coordinator of Informatics Study Program  
Faculty of Computer Science



Dr. Intan Yuniar Purbasari, S.Kom. MSc.  
NIP. 19800602 202521 2 029

## STATEMENT OF ORIGINALITY

I am the undersigned:

Student Name : Aprinia Salsabila Roiqoh  
NPM : 22081010166  
Degree Program : Bachelor (S1)  
Study Program : Informatic  
Faculty : Faculty of Computer Science

Hereby declares that this undergraduate thesis contains no part of any other scientific work that has been submitted to obtain an academic degree at any higher education institution. Furthermore, it does not contain any work or opinions previously written or published by others, except for those which are explicitly cited in this thesis and listed completely in references.

And I declare that this scientific document is free from elements of plagiarism. If in the future indications of plagiarism are found in this Thesis, I am willing to accept sanctions in accordance with the applicable laws and regulations.

Thus, I made this statement without any coercion from anyone and to be used as it should.



Surabaya, May 25, 2026  
Declarant,



APRINIA SALSABILA ROIQOH  
NPM. 22081010166

## ABSTRACT

Student Name / NPM : Aprinia Salsabila Roiqoh / 22081010166  
Thesis Title : Predicting Movie Popularity In Indonesia Based  
On Metadata using Gradient Boosting  
Advisor : 1. Dr. Rizky Parluka, S.Kom., M.Kom.  
2. Dr. Firza Prima Aditiawan, S.Kom, M.T.I,

The film industry in Indonesia has experienced significant growth; however, the success of a film in attracting audiences remains difficult to predict accurately. This study aims to develop a model for predicting the number of moviegoers in Indonesia based on pre-release metadata using gradient boosting algorithms, namely XGBoost, LightGBM, and CatBoost. The dataset was collected from Cinepoint and TMDb, consisting of 3,464 initial records, which were reduced to 2,595 after the preprocessing stage. The preprocessing steps included data cleaning, selective handling of missing values, logarithmic transformation of the target variable, and feature engineering using a Bayesian smoothing approach. The models were trained using two data split scenarios (80:20 and 70:30), and hyperparameter optimization was performed using Random Search and Bayesian Optimization (Optuna). Model performance was evaluated using RMSE, MAE, MAPE, and  $R^2$  metrics. The results show that the best model was achieved by CatBoost with Random Search under the 80:20 data split scenario, yielding an  $R^2$  value of 0.8729, MAE of 0.5538, RMSE of 0.7698, and MAPE 5,02%. These results indicate that CatBoost provides the most accurate and stable prediction performance compared to XGBoost and LightGBM. Furthermore, hyperparameter tuning was proven to improve model performance in predicting movie audience numbers. Feature importance and SHAP analysis reveal that the main actors, directors, and genres are the most influential features in the prediction results. This indicates that pre-release metadata plays a significant role in determining movie popularity in Indonesia.

**Keywords:** Movie, Popularity, Metadata, Gradient Boosting, RMSE

## ACKNOWLEDGEMENTS

Praise be to Allah SWT for all His blessings, guidance, and grace bestowed upon the author, allowing this undergraduate thesis titled **“Predicting Movie Popularity in Indonesia Based on Metadata Using Gradient Boosting”** to be completed successfully.

The author would like to express the sincerest gratitude to Dr. Rizky Parluka, S.Kom., M.Kom. and Dr. Firza Prima Aditiawan, S.Kom., M.T.I., as the undergraduate thesis advisors, who have generously dedicated their time to provide guidance, invaluable advice, and motivation. The author has also received immense support from various parties, both morally, spiritually, and materially. Therefore, the author would like to extend appreciation to:

1. **Prof. Dr. Ir. Novirina Hendrasarie, M.T.**, as the Dean of the Faculty of Computer Science at Universitas Pembangunan Nasional “Veteran” Jawa Timur, for providing the facilities, opportunities, and a conducive learning environment that enabled me to pursue my studies optimally. Her leadership has been a great inspiration for students to continuously grow.
2. **Dr. Intan Yuniar Purbasari, S.Kom. MSc.**, as the Head of the Informatics Study Program, Faculty of Computer Science at Universitas Pembangunan Nasional “Veteran” Jawa Timur, for her guidance, academic support, and excellent management of the study program, ensuring that the academic process ran smoothly up to the final project presentation stage.
3. **First Advisor, Dr. Rizky Parluka, S.Kom., M.Kom.**, who with utmost patience and meticulousness provided guidance, constructive feedback, and direction throughout the undergraduate thesis writing process. His dedication has helped me understand many concepts and complete this research with a clearer focus.
4. **Second Advisor, Dr. Firza Prima Aditiawan, S.Kom., M.T.I.**, who generously spared his time, shared valuable insights, and provided tremendous guidance during the development of this undergraduate thesis. His motivation and suggestions were instrumental in refining my research.

5. **My beloved Parents**, who have always been the greatest source of strength, prayers, and motivation in my life Thank you for all the sacrifices, unwavering moral and material support, and endless love from the very beginning until the completion of this undergraduate thesis. Without your prayers and encouragement, I would not have been able to complete this academic journey successfully.
6. **My Sister**, who has always provided support, motivation, and advice during the writing of this undergraduate thesis. Thank you for the time and support given, which enabled me to complete this work on time.
7. **My closest friends**, who have constantly offered support, encouragement, and a place to share joys and sorrows throughout this undergraduate thesis journey. Thank you for the togetherness, cooperation, prayers, and laughter that made this journey feel lighter and more meaningful. Your presence has been an essential part of this process.
8. **Finally, a gentle note of gratitude to the author of this work, myself.** Thank you for holding on, for pushing through the exhaustion, and for refusing to give up despite the many hurdles in your way. You survived, and you finished. I hope the wisdom gathered during this long journey remains a guiding light for me, proving useful and meaningful for the rest of my life.

The author realizes that this undergraduate thesis is far from perfect and may contain shortcomings. Therefore, constructive criticism and suggestions from all parties are highly welcomed for the improvement of this work. Finally, despite all the limitations, the author hopes that this report will be beneficial to all readers in general and to the author in particular.

Surabaya, May 25<sup>th</sup> 2026

Author

## TABEL OF CONTENT

COVER PAGE.....	i
APPROVAL SHEET .....	iii
APPROVAL SHEET .....	v
STATEMENT OF ORIGINALITY .....	vii
ABSTRACT.....	ix
ACKNOWLEDGEMENTS .....	xi
TABEL OF CONTENT .....	xiii
LIST OF FIGURE.....	xvii
LIST OF TABLE .....	xix
CHAPTER I INTRODUCTION .....	1
1.1 Background .....	1
1.2 Research Question.....	4
1.3 Research Objectives .....	5
1.4 Research Significance .....	5
1.4.1 Academic Significance.....	5
1.4.2 Practical Significance.....	6
1.5 Research Limitations.....	6
CHAPTER II LITERATURE REVIEW.....	7
2.1 Literature Review.....	7
2.2 Theoretical Background .....	10
2.2.1 Movie Industry and Film Popularity .....	10
2.2.2 Movie Metadata .....	11
2.2.3 Bayesian Smoothing .....	12

2.2.4 XGBoost Regressor.....	13
2.2.5 CatBoost Regressor.....	15
2.2.6 LightGBM Regressor.....	17
2.2.7 Hyperparameter Tuning.....	19
2.2.8 Evaluasi Model.....	20
2.2.8.1 Mean Absolute Error (MAE).....	21
2.2.8.2 Root Mean Squared Error (RMSE).....	21
2.2.8.3 Mean Absolute Percentage Error (MAPE).....	22
2.2.8.4 Coefficient of Determination ( $R^2$ ).....	22
2.2.9 Feature Importance dan SHAP.....	23
2.2.10 Python.....	25
CHAPTER III RESEARCH METHODOLOGY.....	27
3.1 Research Workflow.....	27
3.2 Research Variable.....	28
3.3 Data Source and Collection.....	29
3.3.1 Data Source.....	29
3.3.2 Data Collection.....	30
3.4 Data Preprocessing.....	31
3.4.1 Data Cleaning.....	32
3.4.2 Handling Missing Value.....	33
3.4.3 Handling Outliers.....	34
3.4.4 Feature Engineering.....	36
3.5 Model Development and Training.....	37
3.5.1 Train-Test Split.....	38
3.5.2 Hyperparameter Tuning.....	40
3.5.3 Model Training.....	44

3.6 Model Evaluation .....	46
3.7 Feature Importance and SHAP Analysis .....	48
3.8 Visualization Implementation .....	50
CHAPTER IV RESULT AND DISCUSSION .....	53
4.1 Data Analysis .....	53
4.2 Data Preprocessing .....	60
4.3 Model Experiment Result .....	62
4.3.1 Model Performance Comparison .....	62
4.3.2 Hyperparameter Tuning Impact Analysis .....	70
4.3.3 Algorithm Comparison Analysis .....	71
4.4 Result Analysis .....	74
4.4.1 Feature Importance .....	74
4.4.2 SHAP Analysis .....	75
4.5 Streamlit Implementation and Prediction Experiment .....	78
4.5.1 Visualization Implementation .....	78
4.5.2 Prediction Experiment .....	84
CHAPTER V CONCLUSION AND SUGGESTIONS .....	87
5.1 Conclusion .....	87
5.2 Suggestions .....	88
REFERENCE .....	89
ATTACHMENT .....	93

## LIST OF FIGURE

Figure 3. 1 Workflow Diagram.....	27
Figure 3. 2 Movie Directory Interface on Cinepoint.....	30
Figure 3. 3 API Documentation Interface on TMDb.....	31
Figure 3. 4 Preprocessing Workflow Diagram .....	32
Figure 3. 5 Data Cleaning Result Visualization.....	33
Figure 3. 6 Handling Missing Value Result Visulization .....	34
Figure 3. 7 Data Distribution Before Transformation.....	35
Figure 3. 8 Data Distribution After Transformation .....	35
Figure 3. 9 Model Development Workflow Diagram.....	38
Figure 3. 10 5 K-Fold Visualization .....	39
Figure 4. 1 Admission Distribution Visualization .....	53
Figure 4. 2 Each Years Admission Visualization .....	54
Figure 4. 3 Genre Film Count Distribution.....	55
Figure 4. 4 Genre Admission Distribution.....	55
Figure 4. 5 Actor Film Count Distribution .....	56
Figure 4. 6 Writer Film Count Distribution .....	57
Figure 4. 7 Director Film Count Distribution .....	58
Figure 4. 8 Producer Film Count Distribution.....	59
Figure 4. 9 Feature Engineering Result Visualization .....	61
Figure 4. 10 Feature Importance Analysis .....	74
Figure 4. 11 Model SHAP Value Visualization.....	76
Figure 4. 12 SHAP Value Bar Visualization .....	77
Figure 4. 13 Dashboard Homepage.....	79
Figure 4. 14 EDA Page Interface .....	80
Figure 4. 15 Top 5 and Bottom 5 List Interface.....	81
Figure 4. 16 Prediction Page Interface .....	82
Figure 4. 17 Prediction Result Interface .....	83
Figure 4. 18 About Model Interface.....	84

## LIST OF TABLE

Table 2. 1 List of Prevoius Studies .....	7
Table 3. 1 List of Research Variables .....	29
Table 3. 2 Logarithmic Transformation Result.....	36
Table 3. 3 Data Split Result .....	38
Table 3. 4 5 K-Fold Split Result .....	39
Table 3. 5 List Of XGBoost Parameter .....	42
Table 3. 6 List Of CatBoost Parameter .....	42
Table 3. 7 List Of LightGBM Parameter .....	43
Table 3. 8 Model Training and Testing Scenarios .....	44
Table 3. 9 Feature and Train Data Ilustration .....	44
Table 3. 10 Gradient and Hessian Result .....	45
Table 3. 11 Manual Prediction Result.....	46
Table 4. 1 XGBoost Tasting Result Comparison.....	63
Table 4. 2 CatBoost Testing Result Comparison.....	65
Table 4. 3 LightGBM Testing Result Comparison .....	68
Table 4. 4 RMSE Value Comparison.....	70
Tabel 4. 5 Best Scenario Comparison.....	72
Table 4. 6 Lowest and Highest Popularity Sample .....	85
Tabel 4. 7 Sample Prediction Scenario .....	85