UNDERGRADUATE THESIS

CLASSIFICATION OF PRIORITY RECIPIENTS FOR RTLH ASSISTANCE USING LIGHTGBM WITH BAYESIAN OPTIMIZATION IN JOMBANG REGENCY

FERDI FIRDAUS EGA PRATAMA
NPM 22081010247

THESIS ADVISORS
Eva Yulia Puspaningrum, S.Kom., M.Kom.
Muhammad Muharrom Al Haromainy, S.Kom., M.Kom.

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

APPROVAL SHEET

CLASSIFICATION OF PRIORITY RECIPIENTS FOR RTLH ASSISTANCE USING LIGHTGBM WITH BAYESIAN OPTIMIZATION IN JOMBANG REGENCY

By:
FERDI FIRDAUS EGA PRATAMA
NPM 22081010247

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 20, 2026:

Approved,

Eva Yulia Puspaningrum, S.Kom., M.Kom. NIP. 19890705 202121 2 002	....	**(Advisor I)**
Muhammad Muharrom Al Haromainy, S.Kom., M.Kom. NIP. 19950601 202203 1 006	....	**(Advisor II)**
Dr. Rizky Parlika, S.Kom., M.Kom. NIP. 19840518 202121 1 003	....	**(Head Assessor)**
Achmad Junaidi, S.Kom., M.Kom. NIP. 19781110 202521 1 048	....	**(Assessor I)**

Acknowledge by,

Dean of the Faculty of Computer Science

Prof. Dr. Ir. Novirina Hendrasarie, MT.
NIP. 19681126 199403 2 001

APPROVAL SHEET

CLASSIFICATION OF PRIORITY RECIPIENTS FOR RTLH ASSISTANCE USING LIGHTGBM WITH BAYESIAN OPTIMIZATION IN JOMBANG REGENCY

By:
FERDI FIRDAUS EGA PRATAMA
NPM 22081010247

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	:	Ferdi Firdaus Ega Pratama
NPM	:	22081010247
Degree Program	:	Bachelor (S1)
Study Program	:	Informatics
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 28, 2026
Declarant,

METERAI TEMPEL
A1EDCAOX000568872

FERDI FIRDAUS EGA PRATAMA
NPM. 22081010247

ABSTRACT

Student Name / NPM	:	FERDI FIRDAUS EGA PRATAMA / 22081010247
Thesis Title	:	CLASSIFICATION OF PRIORITY RECIPIENTS FOR RTLH ASSISTANCE USING LIGHTGBM WITH BAYESIAN OPTIMIZATION IN JOMBANG REGENCY
Advisor	:	1. Eva Yulia Puspaningrum, S.Kom., M.Kom. 2. Muhammad Muharrom Al Haromainy, S.Kom., M.Kom.

ABSTRACT

The Uninhabitable Housing Assistance (RTLH) assistance program in Jombang Regency still faces challenges in determining aid recipient priorities because the selection process is conducted manually and involves many assessment variables. This study aims to develop a classification model for RTLH assistance recipient priorities using the Light Gradient Boosting Machine (LightGBM) algorithm with hyperparameter optimization through Bayesian Optimization. The dataset used in this study was obtained from the Department of Housing and Settlement Areas of Jombang Regency, consisting of 9,173 data records. After data cleaning, the dataset was reduced to 7,280 records with 36 variables representing housing physical conditions and household socio-economic characteristics. The preprocessing stage included missing value removal, categorical data transformation using Label Encoding, and numerical data normalization using Min-Max Scaler. The evaluation results showed that the baseline model achieved an accuracy of 93.44%, precision of 94.89%, recall of 93.95%, and F1-Score of 94.42%. After optimization using Bayesian Optimization, the model performance improved with an accuracy of 94.71%, precision of 95.26%, recall of 95.81%, and F1-Score of 95.53%. The optimized model was implemented into a Streamlit-based web application to support the initial selection process of RTLH assistance recipients more quickly and objectively. Overall, the combination of LightGBM and Bayesian Optimization improved the classification performance in determining RTLH assistance recipient priorities in Jombang Regency.

Keywords **:** RTLH Classification, LightGBM, Bayesian Optimization, Machine Learning, Decision Support System

ACKNOWLEDGEMENTS

All praise and gratitude be to Allah SWT, *Alhamdulillahirabbil'alamin*, for His abundant mercy, blessings, and grace, which have enabled the author to successfully complete this undergraduate thesis entitled, *"***CLASSIFICATION OF PRIORITY RECIPIENTS FOR RTLH ASSISTANCE USING LIGHTGBM WITH BAYESIAN OPTIMIZATION IN JOMBANG REGENCY***"*.

During the process of completing this undergraduate thesis, the author realized that many parties had provided assistance, support, guidance, and motivation, which greatly contributed to the successful completion of this research. Therefore, the author would like to express sincere gratitude to:

1. Allah SWT, who has bestowed health, strength, ease, and guidance upon the author, enabling the completion of this undergraduate thesis.
2. The author's beloved parents and family, who have continuously provided prayers, support, encouragement, motivation, and endless sacrifices throughout the author's educational journey.
3. The author's younger sibling, who has always provided support, encouragement, and entertainment throughout the process of completing this undergraduate thesis.
4. Maya Yunita Sutrisno, S.H., who has provided support, care, motivation, and encouragement to the author throughout the academic journey and the process of completing this undergraduate thesis.
5. Prof. Dr. Ir. Novirina Hendrasarie, M.T., as the Dean of the Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.
6. Dr. Intan Yuniar Purbasari, S.Kom., M.Sc., as the Head of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.
7. All lecturers of the Informatics Study Program, Faculty of Computer Science, UPN "Veteran" Jawa Timur, who have provided knowledge,

experience, and valuable learning throughout the author's academic journey.

8. Mrs. Eva Yulia Puspaningrum, S.Kom., M.Kom., as the first thesis advisor, who has dedicated time, effort, and thought in providing guidance, direction, suggestions, and motivation to the author throughout the process of completing this undergraduate thesis.
9. Mr. Muhammad Muharrom Al Haromainy, S.Kom., M.Kom., as the second thesis advisor, who has provided valuable suggestions, guidance, and support in the completion of this research.
10. Dr. Rizky Parlika, S.Kom., M.Kom., and Achmad Junaidi, S.Kom., M.Kom., as the thesis examiners, who have provided constructive criticism, suggestions, and feedback for the improvement of this undergraduate thesis.
11. The author's fellow students of the Informatics Study Program class of 2022, who have continuously provided support, assistance, and encouragement throughout the academic journey and the process of completing this undergraduate thesis.

The author realizes that this undergraduate thesis is still not perfect. Therefore, constructive criticism and suggestions are sincerely appreciated for future improvement. Hopefully, this undergraduate thesis can be useful and provide benefits for readers and future researchers.

Surabaya, May 28, 2026

Author

TABLE OF CONTENTS

APPROVAL SHEET .... **iii**

APPROVAL SHEET .... **v**

STATEMENT OF ORIGINALITY .... **vii**

ABSTRACT .... **ix**

ACKNOWLEDGEMENTS .... **xi**

TABLE OF CONTENTS .... **xiii**

LIST OF FIGURES .... **xvii**

LIST OF TABLES .... **xix**

LIST OF CODE PROGRAM .... **xxi**

NOMENCLATURE .... **xxiii**

CHAPTER I INTRODUCTION **1**

1.1 Background 1

1.2 Problem Statement 3

1.3 Research Objectives 3

1.4 Research Benefits 4

1.5 Scope and Limitations 4

CHAPTER II LITERATURE REVIEW **7**

2.1 Previous Research 7

2.2 Uninhabitable Housing Assistance 9

2.3 Machine Learning 11

2.4 Classification 12

2.5 Data Transformation 13

2.6 Data Normalization 14

2.7 Gradient Boosting Framework 15

2.8 Light Gradient Boosting Machine (LightGBM) 16

2.9 Hyperparameter Tuning 18

2.10 Bayesian Optimization 19

2.11 Classification Model Evaluation 21

CHAPTER III RESEARCH METHODOLOGY 23
3.1 Literature Review.. 24
3.2 Data Collection 24
3.3 Data Preprocessing.. 25
3.3.1 Data Cleaning.. 25
3.3.2 Data Transformation 26
3.3.3 Data Normalization 29
3.3.4 Data Selection 30
3.4 RTLH Assistance Priority Labels 32
3.4.1 Source and Validity of Labels... 33
3.4.2 Legal Basis for Label Determination 33
3.4.3 Categorical Value Transformation.. 34
3.5 LightGBM Model 36
3.6 Bayesian Optimization.. 39
3.7 Model Training 42
3.8 Model Evaluation... 43
3.9 Hyperparameter Optimization Using Bayesian Optimization 43
3.10 Testing Scenarios... 44
3.11 Model Deployment 45
3.12 Results Analysis... 45
3.13 System Validation Using Cohen's Kappa....................................... 46
CHAPTER IV RESULTS AND DISCUSSION 49
4.1 Data Analysis 49
4.2 Data Preprocessing.. 50
4.2.1 Data Cleaning.. 51
4.2.2 Data Transformation 54
4.2.3 Data Normalization 61
4.2.4 Data Selection 63
4.3 LightGBM Model 65
4.3.1 Dataset Splitting.. 65
4.3.2 LightGBM Classification... 67
4.4 Bayesian Optimization.. 72

4.5 Model Evaluation ... 79
4.6 Model Evaluation Results ... 82
4.6.1 Comparison of Confusion Matrices ... 82
4.6.2 Comparison of Evaluation Metrics ... 84
4.6.3 ROC-AUC Comparison ... 85
4.6.4 Comparison of Feature Importance ... 87
4.7 Results Analysis ... 88
4.8 Model Deployment ... 92
4.9 Validation Using Field Data ... 95
4.10 System Validation Using Cohen's Kappa ... 97
4.10.1 Validation Questionnaire Instrument ... 97
4.10.2 Assessment Scale ... 98
4.10.3 Validator Assessment Results ... 98
4.10.4 Cohen's Kappa Calculation Results ... 99
CHAPTER V CONCLUSION ... 101
5.1 Conclusions ... 101
5.2 Suggestions ... 102
BIBLIOGRAPHY ... 103
ATTACHMENT ... 109

LIST OF FIGURES

Figure 3. 1 Research Methodology 23
Figure 3. 2 RTLH Dataset 25
Figure 3. 3 Columns Containing Missing Values 26
Figure 3. 4 LightGBM Workflow 36
Figure 3. 5 Bayesian Optimization Workflow 39
Figure 4. 1 Dataset Display 50
Figure 4. 2 Missing Value Checking 52
Figure 4. 3 Missing Value Removal Results 54
Figure 4. 4 Successfully Encoded Columns: Education, Gender, Occupation, and Income 55
Figure 4. 5 Successfully Encoded Columns: Housing Status, Housing Assets, Land Status, Land Assets, Electricity Access, and Assistance Status 56
Figure 4. 6 Successfully Encoded Columns: Area Classification, Room Function, Foundation, Columns, Frame Structure, and Ceiling 57
Figure 4. 7 Successfully Encoded Columns: Beams, Sloof, Windows, Ventilation, Floor Material, and Floor Condition 58
Figure 4. 8 Successfully Encoded Columns: Wall Material, Wall Condition, Roof Material, Roof Condition, Water Source, and Distance to Waste Disposal Site 59
Figure 4. 9 Successfully Encoded Columns: Bathroom Facilities, Toilet Seat, and Waste Disposal System Type 60
Figure 4. 10 Data Normalization Results 63
Figure 4. 11 Dataset after Preprocessing 65
Figure 4. 12 Confusion Matrices of the Baseline Model and the Bayesian Optimization Model 82
Figure 4. 13 Comparison of Evaluation Metrics 84
Figure 4. 14 ROC-AUC Comparison 86
Figure 4. 15 Comparison of Feature Importance 87
Figure 4. 16 Model Deployment Interface 94
Figure 4. 17 Input Data Testing Results 95

Figure 4. 18 Field Validation Data 96
Figure 4. 19 Validation Accuracy Using Field Data from the RTLH Department of Jombang Regency 96
Figure 4. 20 Case Processing Summary Cohen's Kappa 99
Figure 4. 21 Crosstabulation Results of Validator Assessments 100
Figure 4. 22 Cohen's Kappa Measurement Results 100

LIST OF TABLES

Table 2. 1 Confusion Matrix 21
Table 3. 1 Example of a Categorical Dataset before Transformation.................. 27
Table 3. 2 Example of a Categorical Dataset after Data Transformation 28
Table 3. 3 Example of Data Used for Calculation 37
Table 3. 4 Example of Calculation Results.. 38
Table 3. 5 Initial Bayesian Optimization Experiments 41
Table 3. 6 Bayesian Optimization Iteration Results... 41
Table 3. 7 Bayesian Optimization Search Space 44
Table 3. 8 Comparison of Evaluation Metrics across Testing Scenarios............. 44
Table 3. 9 Interpretation of Cohen's Kappa Values... 47
Table 4. 1 Bayesian Optimization Results 78
Table 4. 2 Comparison of LightGBM Hyperparameters Before and After Bayesian Optimization.. 80
Table 4. 3 Validation Questionnaire Instrument 97
Table 4. 4 Validation Assessment Scale 98
Table 4. 5 Validator Assessment Results... 98

LIST OF CODE PROGRAM

Code Program 4. 1 Dataset Loading 49
Code Program 4. 2 Displaying Data 50
Code Program 4. 3 Missing Value Checking 51
Code Program 4. 4 Missing Value Removal 53
Code Program 4. 5 Data Transformation 54
Code Program 4. 6 Data Normalization 61
Code Program 4. 7 Data Selection 63
Code Program 4. 8 Displaying Data 65
Code Program 4. 9 Dataset Splitting 66
Code Program 4. 10 Hyperparameter Initialization and Configuration 68
Code Program 4. 11 Model Initialization and Initial Training 69
Code Program 4. 12 Library Import and Environment Preparation 72
Code Program 4. 13 Defining the Objective Function (lgbm_cv) 73
Code Program 4. 14 Hyperparameter (pbounds) 75
Code Program 4. 15 BayesianOptimization 76
Code Program 4. 16 Baseline Model Evaluation 80
Code Program 4. 17 Confusion Matrix Baseline 80
Code Program 4. 18 Evaluation of the Model with Bayesian Optimization 81
Code Program 4. 19 Confusion Matrix Bayesian Optimization 81
Code Program 4. 20 Model .pkl File 92
Code Program 4. 21 Streamlit 92
Code Program 4. 22 Input Data Preprocessing and System Prediction 93

NOMENCLATURE

x	:	Original data value
x'	:	Normalized value
x_{min}	:	Minimum value of a variable
x_{max}	:	Maximum value of a variable
$F_0(x)$	:	Initial boosting model
$F_m(x)$	:	Boosting model at the m-th iteration
$f_m(x)$	:	Weak learner or decision tree at the m-th iteration
r_{im}	:	Residual or pseudo-residual
y_i	:	Actual value of the i-th observation
$\widehat{y_i}$	:	Predicted value of the i-th observation
$y_i - \widehat{y_i}$	:	Difference between actual and predicted values
η	:	Learning rate
λ	:	L2 regularization parameter
γ	:	Node-splitting penalty
G_L	:	Sum of gradients in the left branch
G_R	:	Sum of gradients in the right branch
H_L	:	Sum of Hessians in the left branch
H_R	:	Sum of Hessians in the right branch
$\mu(x)$	:	Gaussian Process mean function
$k(x, x')$	:	Kernel or covariance function
$f(x)$	:	Objective function
$f(x^+)$	:	Best objective function value
$EI(x)$	:	Expected Improvement
$\mathcal{N}\left(\mu(x), k(x, x')\right)$	:	Gaussian Process distribution
TP	:	True Positive
TN	:	True Negative
FP	:	False Positive

FN	:	False Negative
$TP + TN$	:	Number of correct predictions
$TP + FP$	:	Total positive predictions
$TP + FN$	:	Total actual positive instances
$Accuracy$	:	Model accuracy
$Precision$	:	Correctness of positive predictions
$Recall$	:	Ability to identify positive instances
$F1\text{-}Score$	:	Harmonic mean of precision and recall
$\frac{TP + TN}{TP + TN + FP + FN}$	:	Accuracy formula
$\frac{TP}{TP + FP}$	:	Precision formula
$\frac{TP}{TP + FN}$	:	Recall formula
$\frac{2 \times Precision \times Recall}{Precision + Recall}$	:	F1-Score formula
$\frac{x - x_{min}}{x_{max} - x_{min}}$	:	Min-Max normalization formula
$\sum_{m=1}^{M} f_m(x)$	:	Sum of boosting models
$L(y_i, F(x\_i))$	:	Loss function
$\arg m\,ax$	:	Optimal maximum value