

**Predictive Modeling of Regional Economic Growth in East Java using Aggregated
Village Infrastructure Data: a Comparative Machine Learning Approach**

By
Annisaa Bunga Fajar February
22451034

MASTER'S DEGREE
in

Data Science Business Informatics
Faculty of Engineering and Information Technology



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia
Revision after thesis defense on 5 February 2026

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.



Annisaa Bunga Fajar February

3/3/2026

Student

Date

Approved by:



Dr. Ir. Eka Budiarto, S.T., M.Sc

3/3/2026

Date

Thesis Advisor

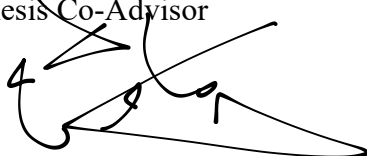


Dr. Maulahikmah Galinium, S.Kom, M.Sc

3/3/2026

Date

Thesis Co-Advisor



Dr. Maulahikmah Galinium, S.Kom, M.Sc

3/3/2026

Date

Dean

ABSTRACT

Predictive Modeling of Regional Economic Growth in East Java using Aggregated Village Infrastructure Data: a Comparative Machine Learning Approach

By

Annisaa Bunga Fajar February

Dr. Ir. Eka Budiarto, S.T., M.Sc., Advisor

Dr. Ir. Maulahikmah Galinium, S.Kom., M.Sc., Co-Advisor

SWISS GERMAN UNIVERSITY

This research develops a machine learning predictive model for East Java's regional economic growth (GRDP per capita) using population-normalized village infrastructure data (PODES) to address the "Core-Periphery" paradox and mitigate spatial size bias. Comparing four algorithms, Gradient Boosting emerged as the optimal architecture (MAPE: 33.493%), drastically outperforming Linear Regression (MAPE: 142.410%) and proving the highly non-linear, threshold-dependent nature of regional spatial economies. Feature importance analysis reveals an agglomeration paradox: capital-intensive facilities (e.g., Higher Education, Specialized Health Centers) act as primary accelerators of high GRDP, whereas inflated per-capita ratios of baseline utilities (e.g., government lighting) merely proxy sprawling, low-output rural peripheries. Additionally, the absence of vocational training (NO LKP) acts as a severe spatial penalty trapping regions in economic stagnation. Ultimately, regional welfare is propelled by specialized human capital rather than basic physical facilities, prompting recommendations for the government to decentralize "Growth Poles" and reorient investments toward vocational agility.

Keywords: GRDP per Capita, Machine Learning, Gradient Boosting, PODES Data Aggregation, Data-Driven Policy, Agglomeration Paradox.

© Copyright 2026

by Annisaa Bunga Fajar February

All rights reserved



DEDICATION

To my parents, for their endless prayers and sacrifices.

*And to the people of East Java—may this research contribute, in a small way, to bridging the
gaps and building a more equitable future.*



ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to God Almighty for the blessings, strength, and patience granted to me, which enabled the completion of this thesis titled "Predictive Modeling of Regional Economic Growth in East Java using Aggregated Village Infrastructure Data: A Comparative Machine Learning Approach." This thesis would not have been possible without the support and guidance of numerous individuals. I would like to extend my sincere appreciation to:

1. Dr. Ir. Eka Budiarto, S.T., M.Sc. my thesis supervisor. Thank you for your invaluable mentorship, insightful feedback, and patience throughout the research process. Your expertise in Data Science and your critical perspective on model optimization were instrumental in shaping the direction of this study.
2. The entire faculty and staff of the Master of Data Science Business Informatics program at Swiss German University. Thank you for the knowledge shared during the coursework, which provided the foundational skills in Machine Learning and Business Intelligence necessary to execute this research.
3. Badan Pusat Statistik (BPS) of East Java Province, for providing the Village Potential (PODES) data and other regional economic indicators. The granularity of this data was the backbone of this study's analysis.
4. My beloved parents, Dr. Arif Badrudin, M.Mgt.Stud. and Mustikasari Damajanti, A.md, and my family. Thank you for your endless prayers, unconditional love, and sacrifices. You are my primary motivation to pursue higher education and strive for excellence.
5. My fellow graduate students and friends in the Data Science. Thank you for the intellectual discussions, the shared struggles during late-night coding sessions, and for being a constant source of encouragement.
6. To my friends who accompanied me on occasional golf rounds and coffee breaks, thank you for helping me maintain my sanity and balance during this intense academic journey.

Finally, I realize that this thesis is far from perfect. However, I hope that the findings regarding the comparison of Random Forest, Gradient Boosting, and other algorithms in this study can contribute to the development of regional economic policies in Indonesia.

Jakarta, 18 February 2026

Annisaa Bunga Fajar February

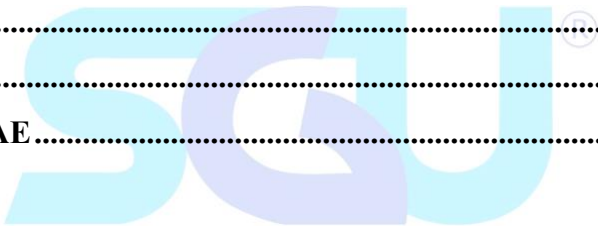


LIST OF CONTENTS

COVER	1
STATEMENT BY THE AUTHOR	2
ABSTRACT	3
© Copyright 2026	4
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
LIST OF CONTENTS	8
LIST OF FIGURES	11
LIST OF TABLES	12
CHAPTER 1 – INTRODUCTION	13
1.1 Background	13
1.2 Problem Statement	14
1.3 Research Objectives	14
1.4 Research Questions	15
1.5 Hypothesis	15
1.6 Significance of the Study	16
1.6.1 Theoretical Significance	16
1.6.2 Practical Significance	16
1.7 Scope and Limitations	16
1.7.1 Scope	16
1.7.2 Limitations	17
1.8 Structure of the Thesis.....	17
CHAPTER 2 - LITERATURE REVIEW.....	18
2.1 Regional Economic Growth Theory	18
2.1.1 GRDP per Capita as a Welfare Indicator	18
2.1.2 Endogenous Growth Theory	18
2.2 Infrastructure as a Determinant of Rural-Urban Economy	19
2.3 Overview of the Study Area: East Java Province	19
2.4 The Machine Learning Paradigm in Regional Economics.....	20
2.4.1 Limitations of Traditional Econometrics	20

2.4.2	Advantages of Supervised Machine Learning	21
2.5	Rationale for the Selection of Machine Learning Algorithms for Predictive Modeling	21
2.6	Model Evaluation and Interpretability	23
2.6.1	Permutation Feature Importance	23
2.7	Conceptual Framework	24
2.8	Research Gap.....	27
2.8.1	Identified Research Gaps	27
CHAPTER 3 – RESEARCH METHODS.....		35
3.1	Research Design.....	35
3.2	Data Sources.....	35
3.3	Data Preprocessing (Aggregation Strategy).....	36
3.4	Research Framework.....	37
CHAPTER 4 – RESULTS AND DISCUSSIONS.....		40
4.1	Overview of Analytical Workflow and Dataset.....	40
4.2	Description of the Aggregated Dataset	40
4.3	Data Aggregation and Descriptive Analysis	43
4.3.1	Descriptive Statistics of Infrastructure Variables (PODES)	43
4.4	Predictive Analysis of Key Spatial.....	46
4.4.1	Predictive Analysis of Healthcare Infrastructure Density	46
4.4.2	Model Evaluation: The Digital Connectivity (Signal) Scenario	49
4.4.3	Model Performance Analysis: Electricity Infrastructure Density	50
4.4.4	Predictive Performance Analysis: Educational Infrastructure and Regional Economic Growth	52
4.5	Empirical Evaluation of Machine Learning Models Across Predictive Scenarios ..	54
4.6	Transitioning from Predictive Accuracy to Policy Formulation	56
4.6.1	Analysis of Absolute Education Infrastructure	59
4.6.2	Spatial-Demographic Transformation of Telecommunication Capacity	62
4.6.3	Spatial Distribution of Healthcare Infrastructure and Regional Output.....	63
4.6.4	Spatial Dynamics of Transportation Infrastructure and Mobility Networks.....	65
4.6.5	Spatial Dynamics of Electrification and Public Utilities per 10,000 Population .	66
4.6.6	The Multi-Dimensional Determinants of Regional GRDP	68

4.7	General Insight About Important Features	71
4.8	The Architecture of Regional Human Capital: Synergizing Education, Vocational Skills, and Infrastructure	73
4.9	Discussion	75
4.9.2	Policy Implication: Reorienting Human Capital Investment	76
4.9.3	Addressing Regional Inequality	77
4.10	Recommendation	77
4.11	Validation Expert	78
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS		81
5.1	Conclusions	81
5.1.1	The Effectiveness of Supervised Machine Learning Models	81
5.1.2	The Primary Determinants of Regional Wealth	82
5.2	Strategic Policy Recommendations	82
5.3	Limitations and Recommendations for Future Research	83
GLOSSARY		85
REFERENCES		86
CURRICULUM VITAE		89



LIST OF FIGURES

Figures	Page
Figure 1 Conceptual Framework: From Village Data Aggregation To Policy Insight ..	25
Figure 2 Operational Framework.....	38



LIST OF TABLES

Table	Page
Table 1. 1 Research Gaps	27
Table 3. 1 Feature Extraction Strategy	36
Table 4. 1 Dataset Summary	40
Table 4. 2 Name Of Bakorwil In East Java.....	44
Table 4. 3 Model Performance Comparison For Healthcare Density (Mape)	47
Table 4. 4 Model Performance Comparison For Signal Infrastructure (Mape).....	49
Table 4. 5 Model Performance (Mape) Using Electricity Density Features	50
Table 4. 6 Performance Measurement Cross Validation With 20 Folds.....	53
Table 4. 7 Analysis Of Absolute Telecommunication Infrastructure Data On Regional Economic Growth.....	57
Table 4. 8 Feature Importance Of Scenario 1	59
Table 4. 9 Spatial-Demographic Transformation Of Telecommunication Capacity	62
Table 4. 10 Increase Mape Of Healthcare And Regional Output.....	64
Table 4. 11 Transportation Infrastructure And Mobility Networks.....	65
Table 4. 12 Electricity Dan Public Utilities	66
Table 4. 13 Multi Dimensional Of Regonal Grdp	68
Table 4. 14 Important Features That Appeared Most Often In The Scenarios	71