

CONTENTS

APPROVAL PAGE	ii
CONTENTS	iii
LIST OF TABLES	iv
LIST OF FIGURES	v
I. INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Scope	2
1.4 Research Objective	3
1.5 Research Advantage	3
II. LITERATURE REVIEW	4
III. BASIC THEORY	8
3.1 SMOTE	8
3.2 Random Undersampling	10
3.3 Random Oversampling	10
3.4 Support Vector Machine Classifier	10
3.5 Random Forest Classifier	12
3.6 Platt Scaling	13
3.6 K-Fold Cross-Validation	14
3.7 Hyperparameter Tuning	15
3.8 Confusion Matrix	16
3.9 Accuracy	16
3.10 Recall	17
3.11 Precision	17
3.12 F-Score	17
3.13 Receiving Operating Characteristics - Area Under Curve	17
IV. RESEARCH METHODOLOGY	19
4.1 Research Description	19
4.2 Tool and Material	20

4.3	Research Phase	20
4.3.1	Literature Study	20
4.3.2	System Design	20
4.3.3	Data Acquisition	22
4.3.4	Data Pre-processing	24
4.3.5	Sampling Method	25
4.3.6	Support Vector Machine (SVM) and Random Forest (RF)	26
4.3.7	Performance Evaluation	26
V.	IMPLEMENTATION	31
5.1	Import Necessary Libraries	31
5.2	Load Dataset	32
5.3	Dataset Cleaning	33
5.4	Dataset Description	34
5.5	Dataset Splitting	36
5.6	Random Sampling	37
5.7	SMOTE	39
5.8	Hyperparameter Tuning	41
5.9	Random Forest and Support Vector Machine Learner	42
5.10	Prediction Result	43
5.11	Probabilities Output	43
5.12	Evaluation	44
VI.	RESULTS AND DISCUSSIONS	46
6.1	Dataset Result	46
6.2	Data Preprocessing	49
6.3	Best Parameter Result	51
6.4	Evaluation Metrics and Final Comparison	53
V.	CONCLUSIONS AND RECOMMENDATIONS	58
	REFERENCES	60

LIST OF TABLES

Table 2.1 Research Comparison	6
Table 3.1 Confusion Matrix	16
Table 4.1 Example of BMKG Climate Data	23
Table 4.2 Example of BNPB Flood Disaster Data	23
Table 4.3 Example of Kuala Krai Climate Dataset	24
Table 4.4 Example of the New Combined Dataset	24
Table 6.1 Jakarta Best Hyperparameter Tuning	51
Table 6.2 Kerala Best Hyperparameter Tuning	52
Table 6.3 Jakarta Evaluation	54
Table 6.4 Kerala Evaluation	55

LIST OF FIGURES

Figure 3.1 Support Vector Machine Visualization	11
Figure 3.2 Pseudocode of Random Forest (Breiman, 2001)	13
Figure 3.3 Example of K-Fold Cross-Validation with k=10	15
Figure 3.4 Davis & Goadrich., 2006 ResearchAUC_PR Curve	18
Figure 4.1 Simplified System Design Phase	21
Figure 4.2 Simplified Random Under Sampling Design Phase	27
Figure 4.3 Simplified Random Over Sampling Design Phase	28
Figure 4.4 Simplified SMOTE Design Phase	29
Figure 5.1 Importing Necessary Libraries	32
Figure 5.2 Loading Dataset Code	32
Figure 5.3 Google Colab View of the Jakarta Dataset	33
Figure 5.4 Missing Values Code and Result	33
Figure 5.5 Function for Column Filling with Mean	34
Figure 5.6 Dataset Features	34
Figure 5.7 Jakarta Daily Rainfall Data	35
Figure 5.8 Code for Class Distribution	35
Figure 5.9 Code for Heatmap Visualization	35
Figure 5.10 Column Separation Function	36
Figure 5.11 Variable x Table	36
Figure 5.12 Variable y Table	37
Figure 5.13 Splitting Data for Training and Testing	37
Figure 5.14 Random Undersampling Code	38
Figure 5.15 Random Oversampling Code	38
Figure 5.16 Regular SMOTE Code	39
Figure 5.17 SVM SMOTE Code	39
Figure 5.18 Borderline SMOTE Code	40
Figure 5.19 K-Means SMOTE Code	40
Figure 5.20 SVM Hyperparameter Tuning.	41
Figure 5.21 RF Hyperparameter Tuning.	42

Figure 5.22 RF Learner Scikit-learn Code	42
Figure 5.23 SVM Base Learner Scikit-learn Code	43
Figure 5.24 Prediction Outcome Function for SVM	43
Figure 5.25 Prediction Outcome Function for RF	43
Figure 5.26 RF Probability Code	44
Figure 5.27 SVM Probability Code	44
Figure 5.28 Example Code for Evaluation Metrics	45
Figure 6.1 Class Distribution Bar Chart for Kerala Dataset	46
Figure 6.2 Class Distribution Bar Chart for Jakarta Dataset	47
Figure 6.3 Heatmap of Kerala Dataset	48
Figure 6.4 Heatmap of Jakarta Dataset	49
Figure 6.5 Jakarta Sampling Distribution Chart Comparison	50
Figure 6.6 Kerala Sampling Distribution Chart Comparison	50
Figure 6.7 Jakarta Best Model Confusion Matrix	57
Figure 6.8 Kerala Best Model Confusion Matrix	57

LIST OF EQUATIONS

3.1 Euclidean Distance Equation	8
3.2 Synthetic Data Equation	8
3.3 Hyperplane Equation	11
3.4 Soft Margin SVM Optimization Equation	12
3.5 Platt Scaling Equation	14
3.6 Accuracy Equation	16
3.7 Recall Equation	17
3.8 Precision Equation	17
3.9 F-Score Equation	17