

CONTENTS

CHAPTER I.....	1
1.1. Background.....	1
1.2 Problem Statement.....	4
1.3 Research Scope.....	5
1.4 Research Objective.....	6
1.5 Benefit of Study.....	6
CHAPTER II.....	9
Table 2.1 Literature Study.....	12
CHAPTER III.....	14
3.1 Air Pollution Forecasting.....	14
3.2 Machine Learning.....	16
3.2.1 Xtreme Gradient Boosting (XGBoost).....	17
3.3 Neural Network.....	26
3.3.1 Recurrent Neural Network (RNN).....	30
3.3.2 Long Short-Term Memory (LSTM).....	32
3.3.3 LSTM Memory Cell.....	33
3.4 Hyperparameter Tuning.....	38
3.5 Evaluation Metric.....	38
CHAPTER IV.....	40
4.1 Research Description.....	40
4.2 Research Procedure.....	41
4.2.1 Data Acquisition.....	43
4.2.2 Data Preprocessing.....	43
4.2.3 Feature Engineering.....	44
4.2.4 Modeling and Training.....	46
4.2.5 Model Evaluation.....	48
CHAPTER V.....	49
5.1 Research Tools.....	49
5.2 Data Acquisition Implementation.....	49
5.3 Data Preprocessing Implementation.....	51
5.3.1 Standardized Data Formatting.....	51
5.3.2. One-Hot Encoding.....	52
5.3.3 Data Integration.....	53

5.3.4 Outlier Removal.....	54
5.3.5 Feature Scaling.....	55
5.3.6 Missing Value Handling.....	55
5.4 Feature Engineering Implementation.....	56
5.4.1 Date-Time Feature Extraction Implementation.....	56
5.4.2 Lag and Rolling Features Implementation.....	57
5.5 Modeling and Training Implementation.....	57
5.5.1 Dataset Splitting.....	58
5.5.2 XGBoost Hyperparameter Tuning Implementation.....	58
5.5.3 XGBoost Training Implementation.....	60
5.5.4 XGBoost Evaluation Implementation.....	61
5.5.5 LSTM Hyperparameter Tuning Implementation.....	62
5.5.6 LSTM Training Implementation.....	65
5.5.7 LSTM Evaluation Implementation.....	66
CHAPTER VI.....	67
6.1 Training Dataset.....	67
6.2 Hyperparameter Tuning XGBoost.....	68
6.3 Hyperparameter Tuning LSTM.....	69
6.4 Training and Evaluation.....	70
CHAPTER VII.....	77
7.1 Conclusions.....	77
7.2 Recommendations.....	77
REFERENCES.....	79

LIST OF FIGURE

Figure 3.1 Diagram of Artificial Intelligence and Data Science Fields Pandey, R. K., et al. (2020).....	14
Figure 3.2 Different Types of Machine Learning Approaches Experfy. (n.d.).....	15
Figure 3.3 General Architecture of XGBoost Algorithm with input and output parameters Wang, W., et al. (2020).....	17
Figure 3.4 Tree Pruning in XGBoost Devopedia. (n.d.).....	20
Figure 3.5 Representation of How a Decision Tree in XGBoost Handles Missing Data Using Default Directions Xu (n.d.).....	20
Figure 3.6 Architecture of Neurons in Neural Networks Rowe (2019).....	25
Figure 3.7 Different Types of Activation Functions Yanikoglu (n.d.).....	27
Figure 3.8 Leaky ReLU Can Solve The Dying ReLU Issue Through Allowing Non-zero Gradient Li, Z., et al. (2022).....	28
Figure 3.9 Recurrent Neural Network (RNN), Processing Sequential Data With Memory Across Time Steps via Shared Weights Khalifa, Y., et al. (2020).....	19
Figure 3.10 Illustration of Back-propagation Through Time (BPTT) in Unrolled RNN structure Gupta, P. (2019).....	30
Figure 3.11 Architecture of An LSTM Memory Cell with input and output parameters	34
Figure 4.1 Research Procedure Flowchart.....	41
Figure 4.2 Visualization of Bayesian Optimization which probabilistically searches for the optimal hyperparameters Gomedé (2024).....	47
Figure 4.3 Process Overview for modeling and training.....	48
Figure 5.1 Satu Data Indonesia platform.....	50
Figure 5.2 Stored dataset inside Google Drive Cloud Storage.....	51
Figure 5.3 Identified and fix irregular column naming.....	52
Figure 5.4 Identified irregular date format.....	52
Figure 5.5 One-hot-encoding preprocessing for categorical feature.....	53
Figure 5.6 Integrate the data from different years of dataset into a single CSV file...	53
Figure 5.7 Outliers removal for target variable pm10.....	55
Figure 5.8 Feature scaling implemented on the trained feature.....	55
Figure 5.9 Removing missing values rows.....	56
Figure 5.10 Extractinng date-time features.....	57
Figure 5.11 Lag and rolling features application for PM10.....	57
Figure 5.12 Dataset splitting.....	58
Figure 5.13 Bayesian hyperparameter tuning for XGBoost.....	60

Figure 5.14 XGboost training using hyperamater tuned parameters.....	61
Figure 5.15 Plotting the actual test values against the forecasted values.....	62
Figure 5.16 LSTM hyperparameter tuning using Bayesian Optimization.....	64
Figure 5.17 LSTM training using hyperamater tuned parameters.....	65
Figure 5.18 LSTM training result visualization.....	67
Figure 5.19 LSTM validation loss graph visualization.....	66
Figure 6.2 RMSE result of XGBoost.....	71
Figure 6.3 Plotted predicted PM10 value from XGBoost compared to actual PM10 value.....	71
Figure 6.4Plotted LSTM validation loss.....	72
Figure 6.5 Final RMSE result after training.....	72
Figure 6.6 Plotted actual vs predicted values of PM10 from the LSTM model.....	73
Figure 6.7 Negative slope for PM10 level in the whole dataset.....	75
Figure 6.8 Negative slope for PM10 level in the test dataset.....	75
Figure 6.9 Negative slope for PM10 forecasted level using XGBoost.....	76
Figure 6.10 Negative slope for PM10 forecasted level using LSTM.....	76

LIST OF TABLE

Table 2.1 Literature Study.....	12
Table 4.1 Dataset Sample for the Year 2022.....	43
Table 4.2 The Final Features of The Dataset.....	45
Table 4.3 Preprocessed dataset sample.....	46
Table 6.1 Training Dataset Sample.....	67
Table 6.2 Best parameters for XGBoost hyperparameter tuning.....	69
Table 6.3 Best parameters for LSTM hyperparameter tuning.....	71