

UNIVERSITI TEKNOLOGI MARA

**IMPROVING AIR QUALITY PREDICTION USING
BAYESIAN OPTIMIZATION FOR FEATURE
SELECTION, HYPERPARAMETER TUNING AND
ENSEMBLE LEARNING**

MUHAMAD HAFIZZUDDIN BIN HASLIN

**BACHELOR OF COMPUTER SCIENCE (Honours)
FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES**

JULY 2026

SUPERVISOR APPROVAL

IMPROVING AIR QUALITY PREDICTION USING BAYESIAN OPTIMIZATION FOR FEATURE SELECTION AND HYPERPARAMETER TUNING

By

MUHAMAD HAFIZZUDDIN BIN HASLIN

2024693394

This thesis was prepared under the supervision of the project supervisor, Azlina Narawi. It was submitted to the Faculty of Computer and Mathematical Sciences and was accepted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science (Hons.).

Approved by

.....

Azlina Narawi

Project Supervisor

JULY 11, 2026

ACKNOWLEDGEMENT

Alhamdulillah, praises and thanks to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given. First of all, my sincere gratitude goes to my supervisor, Azlina Narawi for her guidance and patience in writing this thesis and her constant support along the way. This project has been formed through her willingness to read through my work and provide me with feedback and guidance when I got lost. I am very lucky to have her to supervise me, I appreciate her time and hard work. My sincere appreciation also goes to Dr. Liew Chin Ying, my class advisor, for her attentiveness in reading through my thesis and for the thoughtful comments and corrections she provided. Her comments on the form and clarity of my writing have benefitted this document in many ways. I would also like to thank my parents for their prayers, support and encouragement which have helped me during some of the most challenging times of this journey. They are the reason that I have been able to accomplish anything. Finally, I would like to thank my very close friends and course mates who helped me brainstorm, motivated me and had a great time with me during this last year. Thank you all.

ABSTRACT

Air pollution from fine particulate matter (PM_{2.5}) is a continuing environmental and public health issue in Malaysia, where seasonal haze and local emissions cause air quality to fluctuate considerably. Most machine learning models for predicting PM_{2.5} rely on the feature selection and hyperparameter tuning stage as two separate steps, instead of optimising two simultaneously, limiting the accuracy and generalization of PM_{2.5} forecasting. This study develops an enhanced prediction model applying Bayesian Optimization (BO) to feature selection and hyperparameter tuning simultaneously, within a stacking ensemble framework. Daily air quality records for Kuching, Sarawak (3,026 entries, 2016-2024) were pre-processed into 2,557 cleaned observations. The input features used to predict PM_{2.5} were PM₁₀, nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃). BO was used to determine the most influential feature subset and optimize four base learners: Random Forest, XGBoost, CatBoost and Support Vector Regression. Evaluation used RMSE, MAE and R² under a chronological train-test split. All the base learning models performed better after optimization, but SVR was most successful with the highest enhancement. Its coefficient of determination (R²) increased from 0.7790 to 0.9051, while the Mean Absolute Error (MAE) improved from 5.03 to 3.56. The stacking ensemble had a test R² of 0.9276, RMSE of 2.4534, and MAE of 1.5983. The robustness of the stack was also tested for five different train-test split ratios and the mean R² was 0.9164 ± 0.0273 . The predicted pollutant concentrations were also converted into AQI categories using the EPA breakpoint level and the classification of the AQI was 90.04% for 512 test samples. The final forecasting model was also integrated into a prototype web-based dashboard to offer a practical tool to help with decision making. The main contribution of this study is a systematic and reproducible framework in which feature selection and hyperparameter tuning are optimised jointly through Bayesian Optimization, helping environmental authorities and the public to predict and respond to poor air quality.

TABLE OF CONTENTS

CONTENT	PAGE
ACKNOWLEDGEMENT	i
ABSTRACT	ii
TABLE OF CONTENTS.....	iii
LIST OF FIGURES	viii
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi

CHAPTER 1 INTRODUCTION

1.0 Background of the Study.....	1
1.2 Problem Statement	2
1.3 Project Aim and Objective	3
1.4 Scope	4
1.5 Significance of Study	5
1.6 Conclusion	6

CHAPTER 2 LITERATURE REVIEW

2.0 Introduction	7
2.1 Air Pollution.....	8
2.1.1 Definition of Air Pollution.....	8
2.1.2 Causes of Air Pollution.....	9
2.1.3 Effects of Air Pollution.....	10
2.1.4 Comparison of Related Work in AQI Prediction.....	11
2.2 Machine Learning in Air Quality Forecasting	14
2.2.1 Role of Machine Learning in Environmental Modelling.....	14
2.2.2 Random Forest.....	15
2.2.3 Support Vector Regression (SVR).....	16
2.2.4 XGBoost (Extreme Gradient Boosting).....	17
2.2.5 CatBoost.....	18