Count Data Models and Machine Learning Methods with Applications to New COVID-19 Cases using Air Pollution Data in Bangkok, Thailand

CHATCHAYA NANGSUE[1], CHAINARONG TOCHARUS[2], NUANPAN LAWSON[3,*]
[1]Faculty of Medicine, Chiang Mai University,
50200,
THAILAND

[2]Department of Anatomy, Faculty of Medicine,
Chiang Mai University,
50200,
THAILAND

[3]Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
Bangkok, 10800,
THAILAND

**Corresponding Author*

Abstract: - The novel coronavirus causing the COVID-19 pandemic and morbidity and mortality around the world has left unprecedented repercussions on all aspects of life. Estimation of the incidence of coronavirus patients will reimburse planning and preparation for new waves. The link between COVID-19 and air pollution data can allow us to estimate COVID cases. In this study, count data models; Poisson regression and negative binomial regressions models and machine learning methods; support vector regression and random forest are used to investigate the link between new COVID-19 cases and the air pollution data in Bangkok, Thailand. The root means square error and means absolute percentage error are used as the criterions to compare the efficiency from each method. The results from the real data showed that the support vector regression outperforms random forest and also Poisson regression and negative binomial regression models.

Key-Words: - Count data models, Poisson regression model, Negative binomial model, Machine learning, Support vector regression model, Random forest, COVID-19.

Received: November 29, 2024. Revised: May 21, 2025. Accepted: June 13, 2025. Published: July 8, 2025.

1 Introduction

The COVID-19 pandemic has profoundly afflicted global health, economies, and lives. Since the start of the outbreak, efforts to predict the incidence and spread of the coronavirus have been imperative for formulating public health interventions and mitigating repercussions on the population such as morbidity and mortality. Global vaccination programs were enforced but some groups remain susceptible including immunocompromised individuals, [1]. In many countries around the world, there has been a rise in pollution such as fine particulate matter that significantly contribute to increased risk and severity of respiratory and cardiovascular conditions that may be non-communicable or infectious. Exposure to long term pollution is also linked to COVID-19 mortality. The most recent situation is that the air pollution all over Thailand is threatening health, so extreme policies have been reinforced yet again, such as closing schools for the meantime and providing free sky train services to deter pollution from car exhausts. Many studies have found significant positive correlations between air pollution and COVID-19 incidence including a systematic review of studies in North America and Europe using particulate matter with diameter $\leq$2.5 or 10 μm (PM2.5 or PM10), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2) and carbon monoxide (CO) to predict cases, [2].

A study using a linear regression model found that nitrogen dioxide levels were linked to COVID-related deaths [3]. Methods such as estimators [4]

and [5] have been applied to predict COVID-19 cases focusing on real world circumstances of missing data, but machine learning models for forecasting COVID-19 incidence from pollution are limited, although they can benefit greatly with its accuracy and can save time and costs while utilizing large sets of data. A study utilized machine learning to predict COVID-19 cases from previous infection, fine particulate matter (PM2.5), temperature, and humidity data in Nigeria, but may not have found the most suitable algorithm to apply to PM2.5 data as it did not yield superior accuracy in comparison to the other data, [6]. Air pollution caused by exhausts also leads to oxidative stress, inflammation, and EGFR gene mutation resulting in lung cancer, this mechanism has herbal drugs currently being researched along with its effects in treating COVID, [7]. Count data models such as Poisson regression and negative binomial regression models are usually used to models count data just like new COVID-19 cases. Some works have developed estimators for the estimation of the interested population mean to yield the highest accuracy, [8], [9].

In this study, we aim to investigate the association between COVID-19 incidence and the air pollution data e.g., PM2.5, PM10, CO, NO2, SO2 using two count data models; Poisson regression and negative binomial regression models and two machine learning methods; support vector regression and random forest. The root means square error (RMSE) and mean absolute percentage error (MAPE) are considered as the criterions in comparison for each method.

2 Research Methodology

2.1 Count Data Models

The count data models; Poisson regression and negative binomial regression models and machine learning methods are considered in this study. The details are as follows.

Poisson regression model is a model for count data which is a special case of the generalized linear model where mean and variance of the Poisson random variable must be equal. The density function of Poisson distribution is shown in equation (1), [10].

$$f(y_i \mid x_i) = \frac{e^{-\mu}\mu^{y}}{y!},\ y = 0,1,2,..., \quad (1)$$

where μ is the rate parameter.

Mean and variance of the response variable are shown in equation (2).

$$E(y_i \mid x_i) = V(y_i \mid x_i) = \mu. \quad (2)$$

The negative binomial regression model is also used when the response variable is a count variable that has negative binomial regression distribution. It is less strict than the Poisson regression model which can be used when the mean and variance are not equal which is called overdispersion when the variance is larger than the mean, [11]. The density function of negative binomial is shown in equation (3).

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\,y!}\frac{\mu^{y}\theta^{\theta}}{(\mu+\theta)^{y+\theta}}, \quad (3)$$

where θ is the shape parameter, μ is the mean and $\Gamma(\cdot)$ is the gamma function.

Mean and variance of the response variable are respectively.

$$E(y_i \mid x_i) = \mu, \quad (4)$$

$$V(y_i \mid x_i) = \mu_i(1+1/\theta)\mu_i. \quad (5)$$

2.2 Machine Learning

One of the artificial intelligence (AI) is machine learning or it is called (ML). Machine learning is used as the brain of AI to create intelligence which refer to the models that are created by computers where humans are only responsible for programming by learning from data and let machine handle the rest by itself. It is designed by a data scientist which requires a mechanism that is a program or called an algorithm that has many different types, and it needs to be compared with others to find the most suitable algorithm for actual use. Example tasks of machine learning are data analysis, predicting data, images categorization (see e.g. [12], [13], [14], [15], [16], [17], [18]). In this study, we considered two machine learning methods; support vector regression model and random forest.

Support vector regression (SVR) is a special case of support vector machines (SVM) that is applied to regression issues. The algorithm of SV is a nonlinear generalization of the Generalized Portrait algorithm that is established in the sixties in Russia [19], [20]. As a result, in the past three decades, SV algorithm is grounded framework in statistical learning theory or VC theory which can be used for classification in machine learning very well for unseen data and has been developed by [21], [22], [23], [24], [25], [26], [27], [28]. The

kernels, sparse solution and Vapnik-Chervonenkis (VC) theory are used to categorize the SVR in margin controller and the number of support vectors, [29]. Support vector regression challenges by minimizing the generalization error bound instead of observed error to get the generalized performance. SVR used the basic concept of the computation of high dimensions of the feather space which map the input data with a nonlinear function in a linear regression function, [30].

Random forest is one of the machine learning models that were suggested by [31] and were inspired by the work of [32] that aggregates the output from multiple decision trees in order to create a single result for a predictor variable. Random forest can handle both a categorical response variable which is referred to as classification and a continuous response which is referred to as regression so therefore the predictor variables can be either categorical or continuous data [33], [34], [35], [36], [37], [38]. Both support vector regression and random forest can be useful in the prediction of the predictor variable.

3 Research Results

3.1 Count Data Model Results

The response variable in this study is daily new COVID-19 cases from the top three ranks in Bangkok, Thailand; Bankok Noi, Din Dang and Huai Khwang from Ministry of Public Health, Thailand between the third and fourth stages of the outbreak between 1st of April 2021 and 30th of April 2022 [39] and the independent variables are the air pollution data from pollution control department [40] consist of PM2.5, PM10, NO, CO, SO and O3. Stepwise regression method is used to select the best model from Poisson regression and negative binomial regression models. Figure 1, Figure 2 and Figure 3 represent the number of new COVID-19 cases from Bangkok Noi, Din Dang and Huai Khwang respectively.

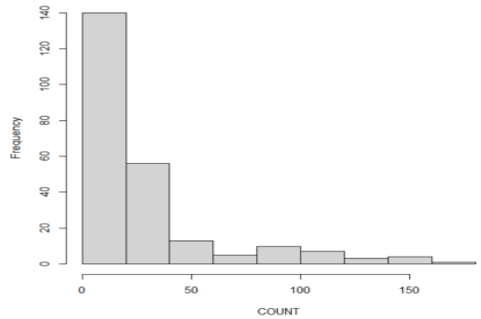

Fig. 1: The number of new COVID-19 cases from Bangkok Noi

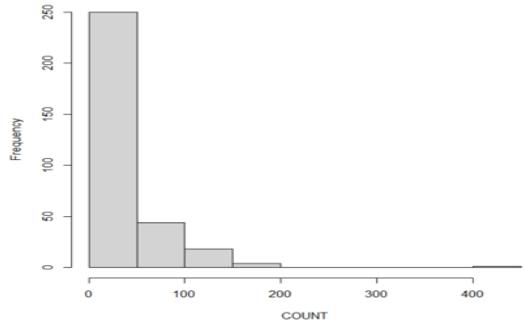

Fig. 2: The number of new COVID-19 cases from Din Dang

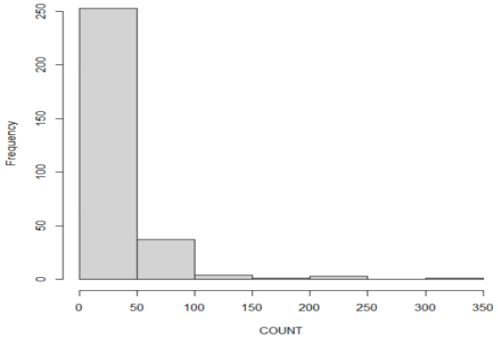

Fig. 3: The number of new COVID-19 cases from Huai Khwang

The descriptive statistics for new COVID-19 cases and the air pollution variables are represented in Table 1. The estimated values of parameters for count data models are represented in Table 2.

From Table 1, we can see that the variance of new COVID-19 cases is larger than the mean for all three districts; Bangkok Noi, Din Dang and Huai Khwang.

Table 1. Descriptive statistics for new COVID-19 cases and the air pollution variables

District	Variable	Mean	Standard deviation	Minimum	Median	Maximum
Bangkok Noi	New COVID-19 cases	28	33.664	1	17	169
	PM2.5	20.969	10.033	9	17.375	67.708
	PM10	37.808	15.944	16.792	32.333	104.542
	NO	9.772	8	0.261	7.957	43.409
	CO2	0.406	0.243	0.076	0.3350	1.071
	SO	2.332	1.436	0.043	2.826	4.870
	O3	19.281	7.994	14.435	17.478	40.565
Din Dang	New COVID-19 cases	32.854	42.241	1	17	405
	PM2.5	29.023	11.422	13.087	25.542	89.708
	PM10	5.949e+01	1.924e+01	2.775e+01	5.387e+01	1.326e+02
	NO	29.317	14.138	7.769	25.261	80.091
	CO2	1.239	0.329	0.557	1.178	2.934
	O3	14.077	6.991	5.556	12.227	54.913
Huai Khwang	New COVID-19 cases	28.682	36.532	1	17	349
	PM2.5	17.248	9.072	5.958	14.125	61.417
	PM10	30.384	12.275	12.208	27.125	74.875
	O3	19.333	8.67	3.348	18.13	42.87

Table 2. The estimated values of parameters for count data models

District	Variable	Mean	Standard deviation	Minimum	Median	Maximum
Bangkok Noi	New COVID-19 cases	28	33.664	1	17	169
	PM2.5	20.969	10.033	9	17.375	67.708
	PM10	37.808	15.944	16.792	32.333	104.542
	NO	9.772	8	0.261	7.957	43.409
	CO2	0.406	0.243	0.076	0.3350	1.071
	SO	2.332	1.436	0.043	2.826	4.870
	O3	19.281	7.994	14.435	17.478	40.565
Din Dang	New COVID-19 cases	32.854	42.241	1	17	405
	PM2.5	29.023	11.422	13.087	25.542	89.708
	PM10	5.949e+01	1.924e+01	2.775e+01	5.387e+01	1.326e+02
	NO	29.317	14.138	7.769	25.261	80.091
	CO2	1.239	0.329	0.557	1.178	2.934
	O3	14.077	6.991	5.556	12.227	54.913
Huai Khwang	New COVID-19 cases	28.682	36.532	1	17	349
	PM2.5	17.248	9.072	5.958	14.125	61.417
	PM10	30.384	12.275	12.208	27.125	74.875
	O3	19.333	8.67	3.348	18.13	42.87

Table 2 shows the results found from the Poisson regression model and negative binomial model.

Stepwise regression model is used to select the best model for each model. We can see that the negative binomial regression model is suitable to use when compared to Poisson regression model considering it has a smaller AIC and BIC. Moreover, it is also supported by Table 1 that the variance of new COVID-19 cases is larger than the mean which is more suitable to use with negative binomial regression.

3.2 Comparison of the Models

According to the results in section 3.1 we now compare the negative binomial model with the support vector regression model and random forest using RMSE and MAPE. The results are shown in Table 3.

Table 3. Root mean square errors and mean absolute percentage errors for count data models and machine learning methods.

District	Model	RMSE	MAPE
Bangkok Noi	Negative binomial regression	31.387	300.852
	Support vector regression	30.774	123.999
	Random Forest	29.477	251.39
Din Dang	Negative binomial regression	36.235	344.287
	Support vector regression	34.607	176.719
	Random Forest	39.255	350.240
Huai Khwang	Negative binomial regression	33.992	276.221
	Support vector regression	34.524	162.416
	Random Forest	37.631	258.577

From Table 3 we can see that the support vector regression performed the best in this situation as it produced the smallest MAPE and also if we consider RMSE we can see that for Din Dang district, support vector regression gave the smallest RMSE and also similar RMSE for both support vector regression and random forest for two districts; Bangkok Noi and Huai Khwang. As a result, for the new COVID-19 data that we considered, the support vector regression is recommended to use to estimate new COVID-19 cases for all three districts. As we can see from Figure 4, Figure 5 and Figure 6 that the predicted values for new COVID-19 cases from support vector regression method are closer to the actual new COVID-19 cases for all three districts.

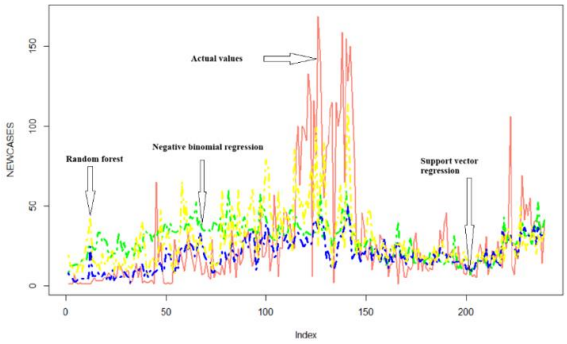

Fig. 4: The actual values of new COVID-19 cases compared to predicted values from negative binomial regression model, support vector regression and random forest methods for Bangkok Noi

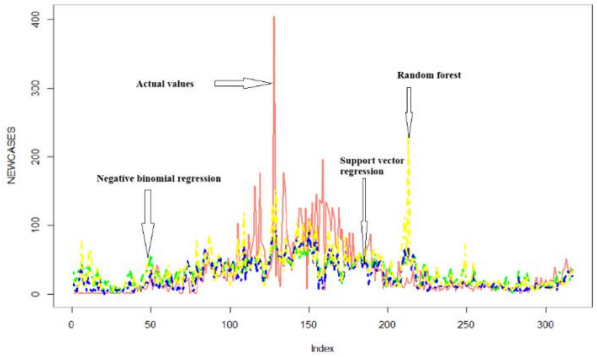

Fig. 5: The actual values of new COVID-19 cases compared to predicted values from the negative binomial regression model, support vector regression and random forest methods for Din Dang

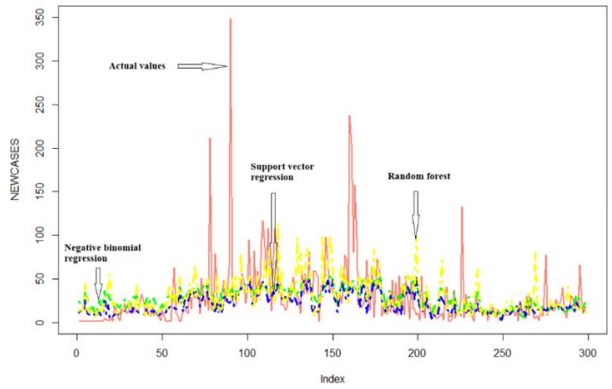

Fig. 6: The actual values of new COVID-19 cases compared to predicted values from the negative binomial regression model, support vector regression and random forest methods for Huai Khwang

4 Conclusions

The count data models both the Poisson regression and negative binomial regression models are used to compare with the results based on machine learning; support vector regression model and random forests. We can see that the machine learning using the support vector regression model performed the best in this situation for estimating new COVID-19 cases using air pollution data from three districts in Thailand.

This work applies knowledge of the profound association between air pollution and COVID-19 cases, causing more morbidity and mortality through respiratory damage leading to vulnerability to the virus. The prevailing measures for deterring the progression of the pandemic have focused on preventing transmission and ameliorating healthcare services although this is seldom not enough as research has shed light on association of the virus to air pollution, a long-term dilemma. Predicting COVID-19 incidence is critical for gaining insights for future public health plans, while also highlighting the importance of reducing pollution which contributes to the increased prevalence and disease severity. This study emphasizes the need for health policies that eradicate both public health issues and environmental risks that threaten health, especially when they are associated together in which the mitigation of one would lead to the improvement of the other. For future works, other count data models and machine learning can be considered to predict COVID-19 incidence or another real data set which can be useful in policy planning for global issues.

Acknowledgement:
Many thanks to all the unknown referees for the valuable comments to help improving the manuscript.

References:

[1] Nangsue, C., Srisurapanont, K. and Sudjaritruk, T., A Comparison of the Immunogenicity and Safety of an Additional Heterologous versus Homologous COVID-19 Vaccination among Non-Seroconverted Immunocompromised Patients after a Two-Dose Primary Series of mRNA Vaccination: A Systematic Review and Meta-Analysis. *Vaccines*, Vol. 12, No. 05, 2024, pp. 468. https://doi.org/10.3390/vaccines12050468.

[2] Hernandez Carballo, I., Bakola, M., and Stuckler, D., The impact of air pollution on COVID-19 incidence, severity, and mortality: A systematic review of studies in Europe and North America. *Environmental Research*, Vol. *215*(Pt 1), 2022, pp.114-155. https://doi.org/10.1016/j.envres.2022.114155.

[3] Sangon, A., Air Pollution and Mortality from COVID-19 in Thailand, *PSRU Journal of Science and Technology*, Vol. 7, No. 1, 2022, pp. 55–70, [Online]. https://ph01.tci-thaijo.org/index.php/Scipsru/article/view/246871 (Accessed Date: October 15, 2024).

[4] Lawson, N., Thongsak, N., Transformed Regression Type Estimators in the Presence of Missing Observations: Case Studies on COVID-19 Incidence in Chiang Mai, Thailand, *WSEAS Transactions on Biology and Biomedicine*, Vol. 21, 2024, pp. 2224-2902. https://doi.org/10.37394/23208.2024.21.13.

[5] Ponkaew, C. and Lawson, N., New Product Estimators for Population Mean Under Unequal Probability Sampling with Missing Data: A Case Study on the Number of New COVID-19 Patients, *Thailand Statistician,* Vol. 22, No. 3, 2024, pp. 634-656.

[6] Ogunjo, S. T., Fuwape, I. A. and Rabiu, A. B., Predicting COVID-19 Cases From Atmospheric Parameters Using Machine Learning Approach. *GeoHealth*, Vol. 6, No. 4, 2022, pp. 1-8. https://doi.org/10.1029/2021GH000509.

[7] Nangsue, C., Srathong, P., Choowongkomon, K., Promnikorn, K., Tocharus, C., Anticancer Activity of the Thai Herbal Formula “Kerra” Versus Afatinib and Doxorubicin in Vitro,

WSEAS Transactions on Biology and Biomedicine, Vol. 21, 2024, pp. 155-161. https://doi.org/10.37394/23208.2024.21.16.

[8] Koc, H., Ratio-Type Estimators for Improving Mean Estimation Using Poisson Regression Method, *Communications in Statistical-Theory and Methods*, Vol. 50, No. 20, 2021, pp. 4685–4691. https://doi.org/10.1080/03610926.2020.1777307.

[9] Koc, H., Tanis, C. and Zaman, T., Poisson Regression-Ratio Estimators of the Population Mean Under Double Sampling, With Application to Covid-19, *Mathematical Population Studies*, Vo. 29, No. 4, 2022, pp. 226–240. https://doi.org/10.1080/08898480.2022.2051988.

[10] Agresti, A., *An Introduction to Categorical Data Analysis*, New York: Wiley, 2007.

[11] Zwilling, Michael L., Negative Binomial Regression, *The Mathematica journal*, Vol. 15, 2013. dx.doi.org/10.3888/tmj.15-6.

[12] Basuchoudhary, A., Bang, J. T. and Sen, T., *Machine-Learning Techniques in Economics: New Tools for Predicting Economic Growth*, New York: Springer, 2017.

[13] Witten, I. H. and Frank, E., *Weka: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann publisher, 2005.

[14] Aware, If AI is artificial intelligence, then what is machine learning?, [Online]. https://www.aware.co.th/th/machine-learning (Accessed Date: October 15, 2024).

[15] Coursera, What Is Machine Learning? Definition, Types, and Examples, [Online]. https://www.coursera.org/articles/what-is-machine-learning (Accessed Date: May 15, 2025).

[16] Alam, A, What is Machine Learning?, 2023. https://doi.org/10.5281/zenodo.8231581

[17] Ławrynowicz, A. and Tresp, V., *Introducing Machine Learning, Perspectives on Ontology Learning*, Publisher: AKA Heidelberg /IOS Press, 2014.

[18] Shalev-Shwartz, S. and Ben-David, S., *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.

[19] Vapnik, V. and Lerner, A., Pattern Recognition Using Generalized Portrait Method, *Automationand Remote Control*, Vol. 24, 1963.

[20] Vapnik, V. and Chervonenkis, A., A Note on One Class of Perceptrons, *Automation and Remote Control*, Vol. 25, 1964.

[21] Vapnik, V. and Chervonenkis, A., Theory of Pattern Recognition (in Russian), *Nauka*, 1974.

[22] Vapnik, V., *Estimation of dependences based on empirical data*, Springer Verlag, 1982.

[23] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.

[24] Alex, J. Smola and Scholkope, B., A Tutorial on Support Vector Regression, *Statistics and Computing*, Vol. 14, 2004, pp. 199–222.

[25] Montesinos López, O.A., Montesinos López, A.and Crossa, J., *Support Vector Machines and Support Vector Regression. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-89010-0_9.

[26] Pal, D. S. and Patranabis, D. C., Support Vector Regression, *Neural Information Processing – Letters and Reviews*, Vol. 11, No. 10, 2007, pp. 203 – 224.

[27] James, G., Witten, D., Hastie, T. and Tibshirani, R., *An Introduction to Statistical Learning*, New York: Springer, 2009.

[28] Yeh, C. Y., Huang, C. W. and Lee, S. J., *A Multiple-Kernel Support Vector Regression Approach for Stock Market Price Forecasting*, Expert System with Application, Vol. 38, 2011, pp. 2177-2186.

[29] Awad, M., Khanna, R., *Support Vector Regression, In: Efficient Learning Machines*, Apress, Berkeley, CA, 2015. https://doi.org/10.1007/978-1-4302-5990-9_4.

[30] Basak, D., Pal, S. and Patranabis, D.C., Support Vector Regression, *Statistics and Computing*, Vol. 11, No. 10, 2008, pp. 203-224.

[31] Breiman, L., Random Forests, *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32.

[32] Amit, Y., Geman, D., Shape Quantization and Recognition with Randomized Trees, *Neural Computation*, Vol. 9, No. 7, 1997, pp. 1545-1588.

[33] Biau, G., Analysis of a Random Forests Model, *Journal of Machine Learning Research*, Vol. 13, 2012, pp. 1063-1095.

[34] Schonlau, M. and Zou, R. Y., The Random Forest Algorithm for Statistical Learning, *The Stata Journal*, Vol. 20, No. 1, 2020, pp. 3–29. https://doi.org/ 10.1177/1536867X20909688.

[35] Au, T. C., Random Forests, Decision Trees, and Categorical Predictors: The "Absent

Levels" Problem, *Journal of Machine Learning Research*, Vol. 19, 2018, pp. 1-30.
[36] Kulkarni, V. Y. and Sinha, P. K., Random Forest Classifiers: A Survey and Future Research Directions, *International Journal of Advanced Computing*, Vol. 36, No. 1, 2013, pp. 1144-1153.
[37] Bernard S, Heutte L, Adam S, Dynamic Random forests, *Pattern Recognition Letters*, Vol. 33, 2012, pp. 1580-1586.
[38] Rodriguze J, Kuncheva L, Rotation Forest: A New Classifier Ensemble Method, *IEEE Transaction on Pattern Analysis and Machine intelligence*, Vol 28, No. 10, 2006, pp.1619-1630.
[39] Department of Disease Control, COVID-19 Data, Thailand, 2020, [Online]. https://covid19.ddc.moph.go.th (Accessed Date: October 22, 2024).
[40] Pollution Control Department, Daily PM2.5 concentration; 2022, [Online]. http://air4thai.pcd.go.th/webV2/history/ (Accessed Date: October 22, 2024).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Chatchaya Nangsue was responsible for the statistics and writing the manuscript.
- Chainarong Tocharus was responsible for the research planning and editing of the manuscript.
- Nuanpan Lawson was responsible for the research planning and execution. She carried out the simulation studies, was responsible for the statistics, writing, review and editing of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research was funded by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok Contract no. KMUTNB-FF-68-B-23.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.