Development of Regression Models for COVID-19 Trends in Malaysia

SOFIANITA MUTALIB[1,3*], SITI NURJEHA MOHD PUNGUT[4], AIDA WATI ZAINAN ABIDIN[2], SHAMIMI A HALIM[1], ISKANDAR SHAH MOHD ZAWAWI[2]

[1]School of Computing Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor,
MALAYSIA

[2]School of Mathematical Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor,
MALAYSIA

[3]Research Initiative Group Intelligent Systems,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor,
MALAYSIA

[4]Xplode Media Private Limited,
Lot No. A-07-2 Paragon Point, Seksyen 9 Pusat Bandar Baru Bangi Bangi, 43650, Selangor,
MALAYSIA

**Corresponding Author*

Abstract: - COVID-19 has emerged as the biggest threat to the world's population, since December 2019. There have been fatalities, financial losses, and widespread fear as a result of this extraordinary occurrence, especially in Malaysia. Using available COVID-19 data from the Ministry of Health (MOH) Malaysia website, from 25/1/2020 to 17/6/2022, this study generated regression models that describe the trends of COVID-19 cases in Malaysia, taking into account the unpredictable nature of COVID-19 cases. Three techniques are used in Weka software: 60:40 / 70:30 split ratio, 10 and 20-fold cross-validation, Support Vector Regression (SVR), Multi Linear Regression (MLR), and Random Forest (RF). Based on new instances among adults, the study's findings indicate that RF has the strongest coefficient correlation and the lowest Root Mean Square Error of 22.7611 when it comes to predicting new COVID-19 deaths in Malaysia. Further investigation into prospective characteristics like vaccination status and types, as well as other external factors like locations, could be added to this study in the future.

Key-Words: - COVID-19, Regression Models, Random Forest, Support Vector Regression, Linear Regression, Supervised.

Received: August 27, 2022. Revised: September 28, 2023. Accepted: October 7, 2023. Published: November 3, 2023.

1 Introduction

Since the first case was reported in Wuhan in 2019, the Coronavirus Disease 2019 (COVID-19) has been affecting the world for nearly two years, resulting in numerous cases and unnecessary deaths. Beginning at the start of December 2019, this disease spread rapidly throughout Wuhan City, Hubei Province, China, [1]. The World Health Organization declared SARS-COVID-19 to be a global pandemic on March 11, 2020. Right after the first case outside of China was identified, Malaysia began its strict screening by prohibiting the entry of foreigners and closing its border. After that, Malaysia experienced new waves of the disease until April 2020, with the first wave being completely and effectively managed and the second wave beginning in early March 2021 with cautious measures taken.

The number of COVID-19 cases in Malaysia continues to rise daily, according to data from Malaysian states through January 7, 2022. The state with the highest number of confirmed cases is Selangor, with 792043, followed by Sarawak and Johor with 252461 and 247439 cases, respectively. Wilayah Persekutuan Labuan, Wilayah Persekutuan

Putrajaya, and Perlis have a low number of COVID-19 cases, with 10897, 9569, and 7321 confirmed cases, respectively. The monitoring also applies to the total number of deaths broken down by the states. The highest fatality rates are still seen in Johor and Selangor, with 3876 and 9988 deaths, respectively. Sarawak had only 1618 total deaths, which is a modest amount considering it was the second state with the most verified cases.

Based on publicly available statistics, this epidemic began when many people at a Wuhan fish market contracted the virus. In addition to seafood, this market sells a broad variety of unusual animals, including birds, snakes, marmots, and bats. Because of their diet, they are high-risk hosts of several viruses and bacteria. Because of the sharp increase in reported cases, the Malaysian government has begun implementing preventative measures to halt the virus's spread. To halt the global spread of SARS-CoV-2, traditional measures must be put into place and adhered to, [2]. The Malaysian government has recommended several preventive measures, including frequent hand washing, avoiding handshakes with strangers, wearing masks and gloves, maintaining a minimum of one metre of social space when walking outside, and avoiding crowds, to stop the virus from spreading. It has been demonstrated that the adoption of non-pharmaceutical social distancing or lockdown measures has significantly reduced the pandemic's scope, [3].

Health systems can better organize their resources, monitor outbreak management, and get ready for pandemics by forecasting the COVID-19 virus. It is useful to use mathematical and statistical models to forecast how infectious diseases may develop. Researchers have been conducting experiments and analyzing the COVID-19 prediction. Several models have been used to forecast the number of COVID-19 cases. The Autoregressive Integrated Moving Average (ARIMA) model, [4], is an example of a statistical model that was used to forecast COVID-19 cases and estimate COVID-19 in Italy, Spain, and France, [5]. In addition, a polynomial model for daily COVID-19 case forecasting was proposed, [6]. Susceptible-Infected-Removed (SIR) model example to estimate and analyze the COVID-19 spread in Kuwait with fixed variables, [7], and estimation of the final size of the COVID-19 epidemic in Pakistan based on the reported cases, [8]. A few additional studies contemplate utilizing the traditional Susceptible-Exposed-Infectious-Removed (SEIR) model, [9], [10], [11].

Our study aims to test several regression algorithms for prediction related to COVID-19 new deaths in Malaysia. The regression models can be useful in analyzing the non-linear relations in historical cases, as aforementioned. Therefore, three different regression models are applied for this prediction. The contribution of this paper is the different experiments in predicting target (new death) based on selected scenarios: general, adolescent, children, and adult. Therefore, the developed models can be used for monitoring the spread of the pandemic.

The rest of this paper is structured as follows: Section 2 provides the related works, and Section 3 presents the literature about predictive models. Section 4 introduces the methods and dataset used in this study. The results and findings are given in Section 5. Section 6 concludes the study.

2 Related Works

When the COVID-19 disease was identified in China, it has been a lot of effort was made to test many types of treatments and solutions across multiple populations to reduce its impact and spread. Since the first case of the current pandemic, COVID-19 was identified more than two years ago, the immunization program for this year in Malaysia did not begin until early February, with a primary focus on immunization and infection control. Herd immunity reduces the likelihood of ineffective contact between a susceptible individual and an infected host, thereby offering indirect protection to susceptible members of a sufficiently immune group. In its most fundamental form, herd immunity occurs when a population reaches the herd immunity threshold, or when a certain proportion of people are immune to a virus, [12].

Numerous viral pandemics, including H1N1, have utilized the SIR model, demonstrating that modelers can approximate disease behavior by predicting a small number of parameters. The SIR model is a mathematical method that may have difficulty being applied to the dynamics of an epidemic. Additionally, the model may suffer if the procedure is overly simplified. The revised data for the COVID-19 pandemic are complex and daily; therefore, they must be discarded in day-series order. The SIR model also assumes that total immunity can be acquired through infection, thereby including the epidemiological concept of natural herd immunity. However, the authors noted earlier that the SIR model could not account for such dynamics in this pandemic, so employing artificial intelligence or machine learning techniques would strengthen the reliability of COVID-19 data, [13], [14]. Awareness of the SIR model in COVID-19 can help researchers reduce other infectious diseases as well as other problems, such as computer viruses and natural disasters.

The COVID-19 pandemic has had some influence and impact on other studies as well, such as mental

health and economic sectors. Studies used statistical and also machine learning techniques for sentiment analysis, for MySejahtera apps, which are mandatory for all Malaysians to report their movements, [15]. In analyzing the feeds and tweets in social media, Latent Dirichlet Allocation (LDA) is implemented to view the three main topics and issues related to mental health, [16]. Meanwhile, Naive Bayes were tested based on VADER and TextBlob scores to classify the records into positive and negative sentiments. The next section explains the machine learning algorithms used in developing predictive regression models, which were mainly adapted in our study for COVID-19 in Malaysia.

3 Predictive Models

Predictive models are used to predict future events, including in economics and also in public health areas. Predictive models are also dynamics that are regularly updated or verified to take into consideration changes to the underlying data. Predictive models rely on assumptions that are grounded in historical and contemporary events. In the case of COVID-19, the development of a predictive model is demonstrated by several research papers.

There is a study that used IoT-based technologies and machine learning to rapidly identify the spread of coronavirus cases, monitor the clinical outcomes of survivors, and collect and analyze pertinent data to establish the presence of the virus, [17]. The collected data culminated in an 80-symptom list, and this study utilized eight algorithms to compare the precision of providing information or tracking COVID-19 cases for each patient. The results show that the machine learning algorithms achieved a good accuracy rate that exceeded 90%, demonstrating that this method of tracking and monitoring is effective.

Another research paper attempts to forecast when infected patients will recover or not (released or deceased) using machine learning algorithms, [18]. The variables in the dataset include gender, age, infection case, and number of days. The Decision Tree (DT) model was found to be the most accurate with a 99.85% accuracy rate, followed by Random Forest (RF) with 99.60%, Support Vector Machine (SVM) with 98.85%, K-nearest neighbor (KNN) with 98.06%, Naive Bayes (NB) with 97.52%, and Linear Regression (LR) with 97.49%. The developed models would be of great assistance in the healthcare industry's battle against COVID-19. Table 1 shows more research papers and the algorithms applied in the COVID-19 study.

Table 1. Machine learning algorithms used in building predictive models.

Reference	Description	Techniques and results
[19]	A prediction model based on demographic and clinical features, with target: positive or negative.	Logistic Regression, DT, SVM, NB Best result: SVM (accuracy 93.34%)
[17]	A framework with IoT devices to monitor and track survivors' clinical measures (80 symptoms)	Neural Network, Decision Table, SVM, NB, k-NN, Dense Neural Network (DNN) – more than 90% LSTM & OneR – less than 90%
[20]	Target: Number of active cases, death and recovery.	LASSO, RF, DT Regressor, LR, SVM, Polynomial Regression The performance of the algorithms varies.
[21]	20 attributes that are possible factors related to acquiring the virus, to predict whether it is positive or not.	J48 DT, RF, SVM, k-NN, NB, MLP, LR, ANN SVM is the best model, RF is the second-best model.

3.1 Multiple Linear Regression

Multiple regression, or multiple linear regression (MLR), is a statistical technique that combines multiple explanatory variables to predict the outcome of a response variable. Using multivariate linear regression, the mathematical relationship between random variables is established. A study focuses on estimating solar radiation using multi-linear regression techniques, artificial neural networks, and empirical equations such as the Hargreaves equation, [22]. Meanwhile, another study compared the predictability of the monthly streamflow using MLR, ANN, ANFIS, and KNN, [23]. The results demonstrated that all three nonlinear models, ANN, ANFIS, and KNN, performed admirably, with the ANFIS model outperforming the others due to its utilization of both fuzzy inference systems and neural networks. MLR models have a different advantage in model development, as they are designed to replicate a linear relationship between inputs and outputs, though, in their study, they failed to accurately forecast monthly flows.

3.2 Random Forest

The output of the RF algorithm for classification

problems is the class selected by the majority of trees. According to, [24], this study uses the RF algorithm to assess and issue warnings regarding the security risk associated with extensive group activities by conducting experiments with a 10-fold cross-validation method. In the study, [24], a comparison was done between RF and KNN, NB, and the CART algorithms. The results revealed that RF achieved the highest accuracy of 86% while KNN achieved 81% accuracy, NB achieved 74% and the CART algorithm achieved the lowest accuracy out of the four algorithms, indicating that RF is a reliable method for predicting ability, [25].

Another study demonstrates that RF models are superior to decision trees, as the accuracy of random forest models for both the train and test sets was slightly higher than that of decision tree models, [26]. The purpose of this study is to evaluate and contrast, on a regional scale, the performance of two cutting-edge machine learning models, DT and RF model, about seven modeled major rainfall-triggered landslides on the Japanese island of Izu-Oshima. This investigation's samples were chosen based on the presence or absence of landslip data, and a classification tree was constructed. At each branch node, a random subset of the potential cause factors is selected.

3.3 Support Vector Regression

Support Vector Regression (SVR), a supervised learning method, is used to forecast discrete values. SVR and SVM are based on a similar concept. The primary objective of SVR is to locate the optimal line. For SVR, the model with the hyperplane with the most points is the optimal fit line. A research paper discovered that it is possible to combine SVR with any other technique. The purpose of this study is to compare the best model for predicting the average monthly temperature in Iran and to demonstrate the effectiveness of combining SVR with the Firefly optimization algorithm, [27].

In another study, SVR and genetic algorithm (GA) are combined to predict the water temperature in numerous reservoirs, [28]. This study employed the root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and Nash-Sutcliffe efficiency coefficient (NSE) to compare the accuracy of these strategies. The results demonstrate that the GA-SVR model outperforms the SVR model on all metrics, while the M-GASVR model is superior, and the ANN model is the least effective of the four models.

4 Methodology

4.1 Data Acquisition

In this phase, our study manipulates the available data and the references from previous research and uses them as the starting point for the modeling of the COVID-19 trend. The daily data was obtained from the website of our Ministry of Health (MOH), [29]. Following the readings, previous studies identified the patterns that might influence the instances of COVID-19 fatalities and used them as variables to assess the accuracy of each model. Collecting COVID-19 datasets was one of the activities performed during this phase. These datasets contain actual information regarding Malaysia's daily cases, daily deaths, and population. These datasets capture the records from 25/1/2020 – 17/6/2022. To view the daily trends, the plotted graphs are provided in Figure 1 for the daily new cases of COVID-19, Figure 2 for the daily new deaths, Figure 3 for the number of vaccination recipients, and Figure 4 is the average number of daily new cases based on the category.

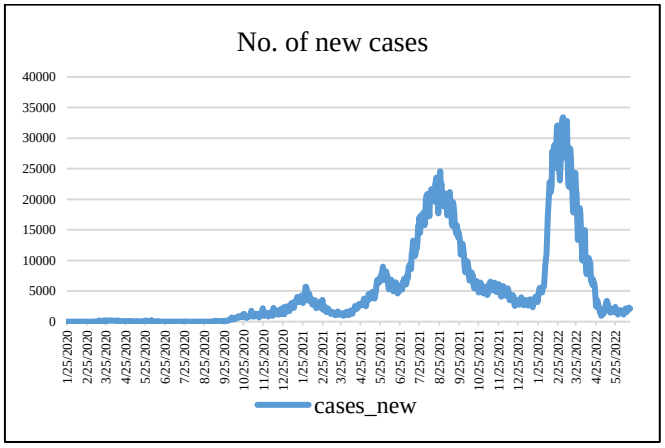

Fig. 1: The daily new cases of COVID-19

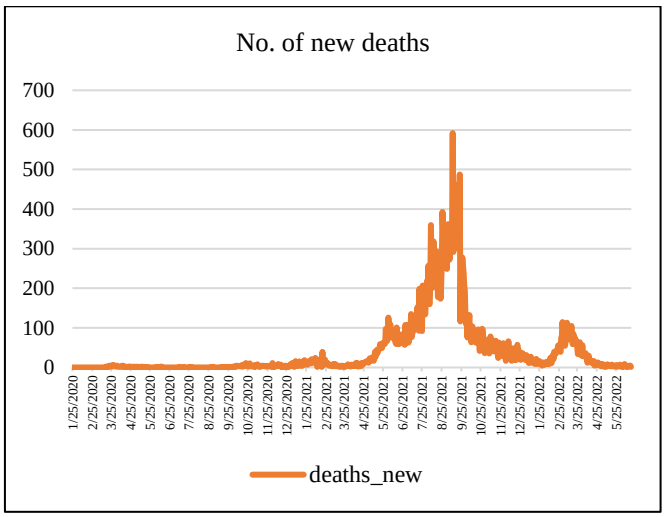

Fig. 2: The number of daily new deaths

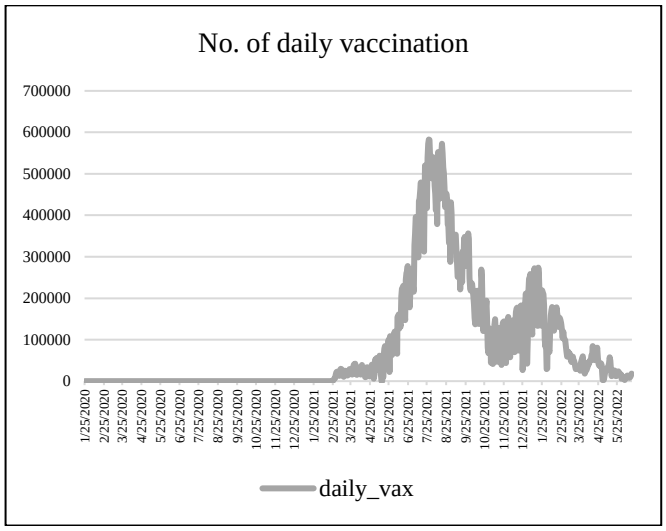

Fig. 3: The number of vaccination recipients

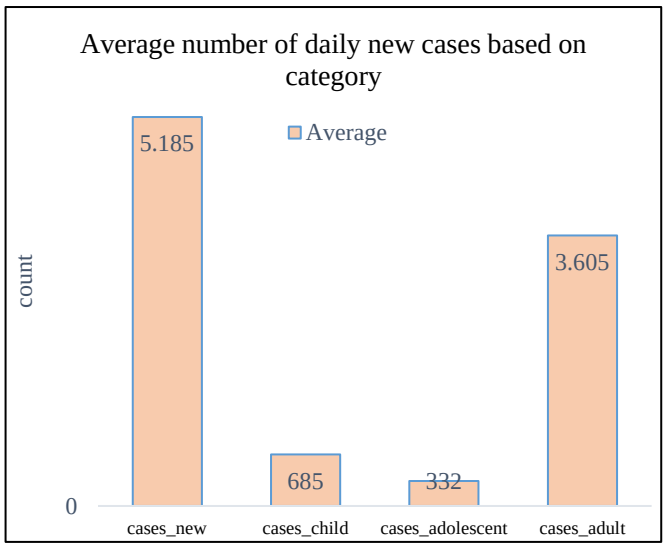

Fig. 4: The average of daily new cases based on category

4.2 Data Modeling

The attributes of daily cases, immunization, and the group of children, adolescents, and adults were selected and converted into a format compatible with Weka. This study's objective is to determine which trends could have an impact on the number of deaths in Malaysia, so deaths new serves as y and other trends serve as x in a series of experiments based on Linear Regression (LR).

$$y = Ax + B. \tag{1}$$

To measure the relationship between the independent variable (x) and dependent variable (y), constant and can be estimated by fitting the experimental data on the variables and through the method of least squares.

Next, the relationship of the input and output variables was also tested using LR, with configuration in Figure 5, RF in Figure 6, and SVR, as in Figure 7.

The training and testing process was done using the hold-out method with a ratio of 60:40 and 70:30, and also the cross-validation (CV) method, with partitions of 10 and 20-fold.

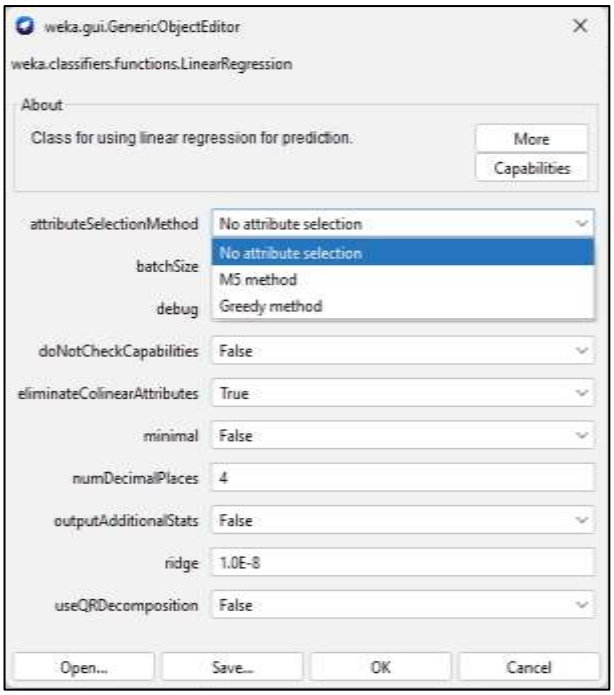

Fig. 5: Configuration for MLR model

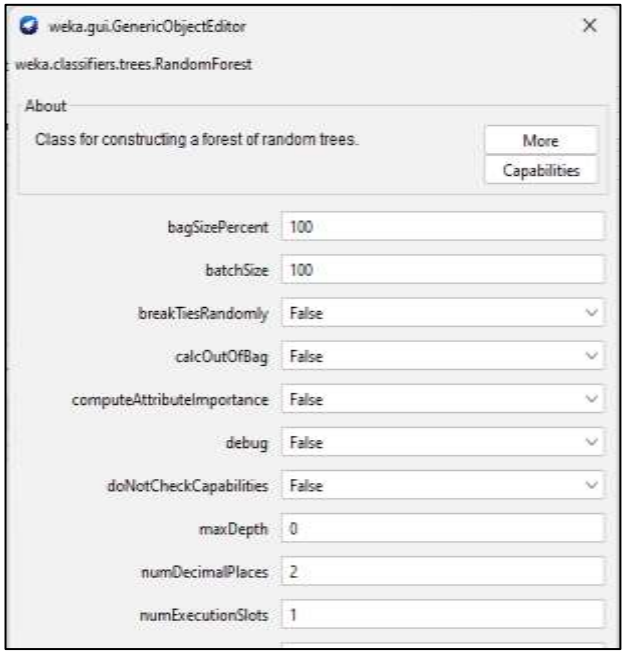

Fig. 6: Configuration for RF model

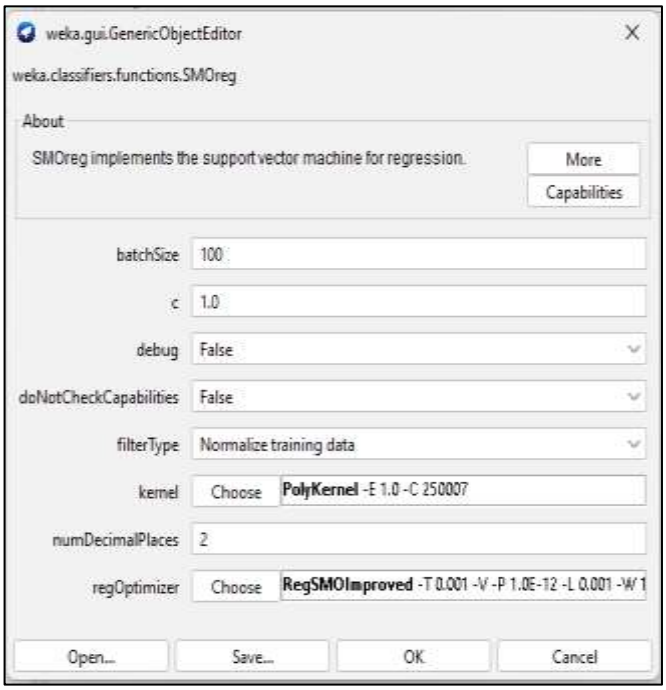

Fig. 7: Configuration for SVR model

4.3 Model Evaluation

During the evaluation phase, the performance of the algorithms was compared based on how each algorithm satisfied the established criteria. In addition, the researcher must ensure that the procedures are followed precisely to construct the most accurate model. This study concentrates on RMSE and Karl Person's correlation coefficient to evaluate the performance of each regression model.

The RMSE is the square root of the mean squared error (MSE). For a sample of n observations $y(y_i, i = 1,2,...,n)$ and n corresponding model predictions $\hat{y}$, the RMSE is given as follows, [30]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}, \qquad (2)$$

Correlation is a technique that measures the nature, degree, and extent of association existing between two continuous variables. Karl Pearson's correlation coefficient is a measure of the degree of relationship between two variables x and y, which is expressed as follows, [31]:

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}, \qquad (3)$$

where is the number of observations. The values are between -1 and 1; a positive value is a positive relationship, while a negative value indicates a negative association of variables. A value is an indicator of a negligible relationship between variables. The correlation values are 'moderate' and 'strong', with values of 0.5 – 0.7, and > 0.7, respectively. The correlation value ($r < 0.5$) implies a weak correlation.

5 Results and Findings

There are five experiments reported for the prediction model of new deaths, based on input attributes, as shown in Table 2 using regression methods. The presented graphs are based on 10-fold cross-validation, 20-fold cross-validation, and a split percentage of 60:40 and 70:30 with different algorithms. The results of each model are then compared to get the smallest RSME value among them.

Table 2. The five experiments of regression models.

Experiment	Input variable, x	Outcome (target), y
1	x = daily new cases	y = number of new deaths
2	x_1 = number of immunizations x_2 = daily new cases	y = number of new deaths
3	x_1 = number of immunizations to children x_2 = daily new cases among children	y = number of new deaths
4	x_1 = number of immunization receivers x_2 = new patient among adolescent	y = number of new deaths
5	x_1 = number of immunization receivers x_2 = new patient among adult	y = number of new deaths

5.1 Experiment 1: To Predict a New Death number

In experiment 1, our study applied SLR, RF, and SVR modeling techniques to model the COVID-19 trends, and the independent variable selected for *x*, which is *cases_new* to predict the *deaths_new* that holds the value of *y* in LR. Figure 8 shows RF has the lowest RMSE of the three models, with a split percentage of 54.9854 at 10-fold cross-validation and 54.8599 at 20-fold cross-validation.

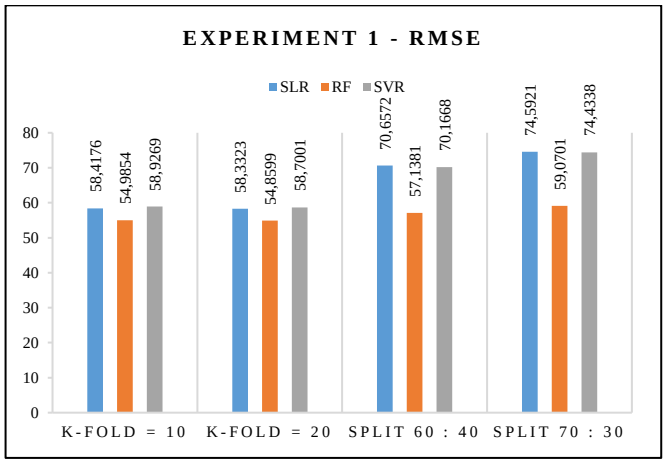

Fig. 8: The comparison of RSME for the developed models in Experiment 1.

5.2 Experiment 2: To Predict the New Death Number

In this experiment, another variable, which is the number of immunizations, was combined with the daily number of new cases applied in the model to forecast the occurrence of new deaths. With 20-fold cross-validation, RF produces the lowest RMSE for the second experiment with 27.1664, as shown in Figure 9.

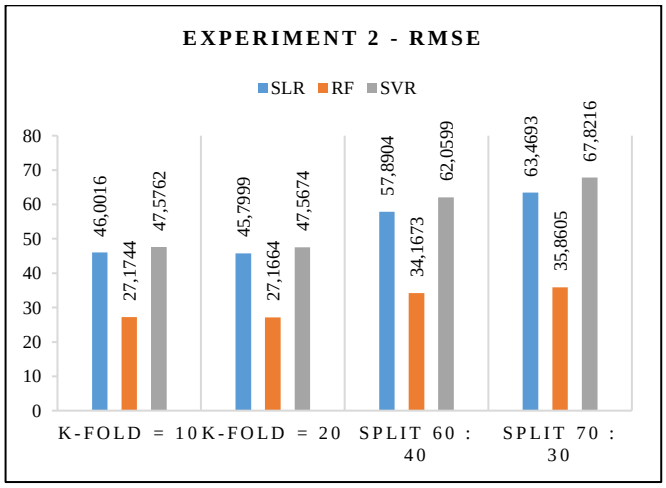

Fig. 9: The comparison of RMSE for the developed models in Experiment 2.

5.3 Experiment 3: To Examine the Influence of Immunization and the New Cases among Children on the New Deaths

Experiment 3 examined the number of immunizations given to children and the number of new COVID-19 cases among children in terms of influencing death. From Figure 10, RF has the lowest RSME with 27.4661 and 25.7022 at 10-fold and 20-fold cross-validation.

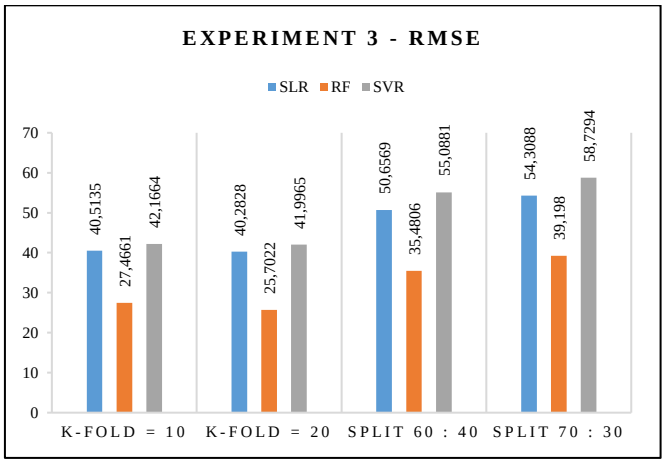

Fig. 10: The comparison of RMSE for the developed models in Experiment 3.

5.4 Experiment 4: To Examine the Influence of Immunization and the New Cases among Adolescents to New Deaths

The fourth experiment concentrated on the adolescent new COVID-19 patients and the quantity of immunizations they had received to predict the deaths. Based on Figure 11, RF was the most performed regression model in this experiment, based on the RSME scoring at 32.7991 and 32.1197 at a 10-fold cross-validation and 20-fold cross-validation.

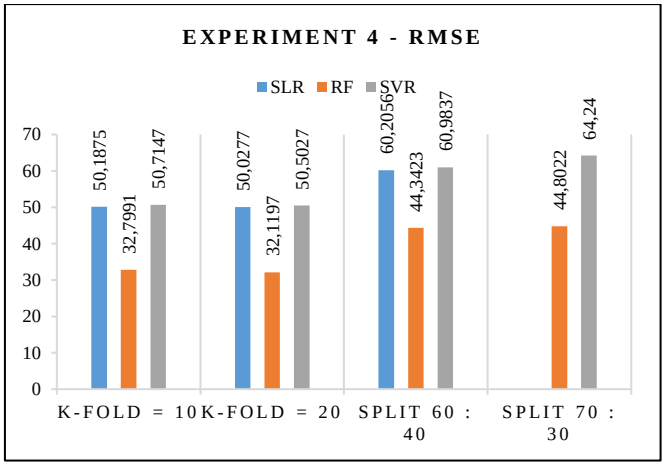

Fig. 11: The comparison of RMSE for the developed models in experiment 4.

5.5 Experiment 5: To Examine the Influence of Immunization and the New Cases among Adults in the New Deaths

In this experiment, the number of immunizations and the new cases among adults are the input. Based on Figure 12, RF showed the lowest score of RSME with 22.8123 and 22.7611 at cross-validation 10-fold and 20-fold.

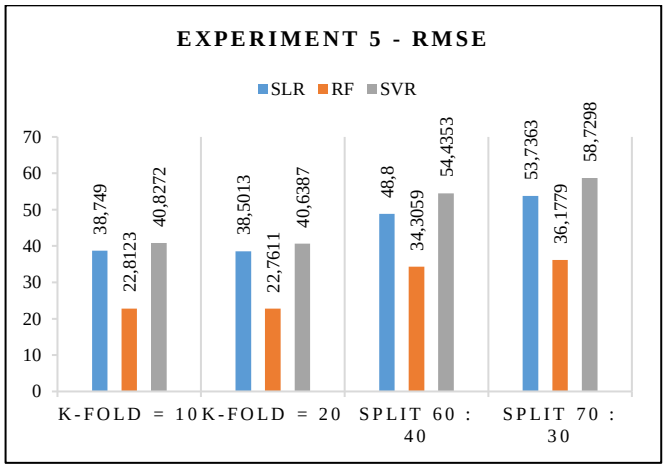

Fig. 12: The comparison of RMSE for the developed models in Experiment 5.

In comparing the RF results among the five experiments, the lowest RSME is gained when predicting death based on immunization and new cases among adults. Adolescent and child new cases influence a little bit lower if we compare the RSME score. Meanwhile, Figure 13 shows the average correlation coefficient among the regression models in five experiments for four setups of training and testing. In Figure 9, SLR/MLR and SVR have a lower median line than RF. The distribution of the correlation coefficient for all experiments is shifted upwards relative to SLR/MLR and SVR. The upper quartile for SLR/MLR and SVR overlaps with the lower quartile for RF. The correlation coefficient variability of each algorithm is almost similar to each other, with the median being negatively skewed due to being closer to the upper value. It is meant that the majority of the correlation values are 'moderate' or 'strong', concerning the values of 0.5 - 0.7 or more than 0.7, respectively. Compared to the studies in, [17], [18], and, [19], they used supervised methods based on classification to the level of risk or either positive or negative class labels, so their results are produced in an accuracy measure.

The main limitation of this analysis was that it takes the daily data without considering the exact location within Malaysia. Some areas are contributing to a higher number of cases and a faster rate of spread. Despite all the limitations, the biggest strength of this study was several experiments done, mainly in the general type of age, among adults, children, and adolescents. Despite that, the training and testing methods were implemented in several settings of cross-validation and split methods to confirm the performance of regression models.

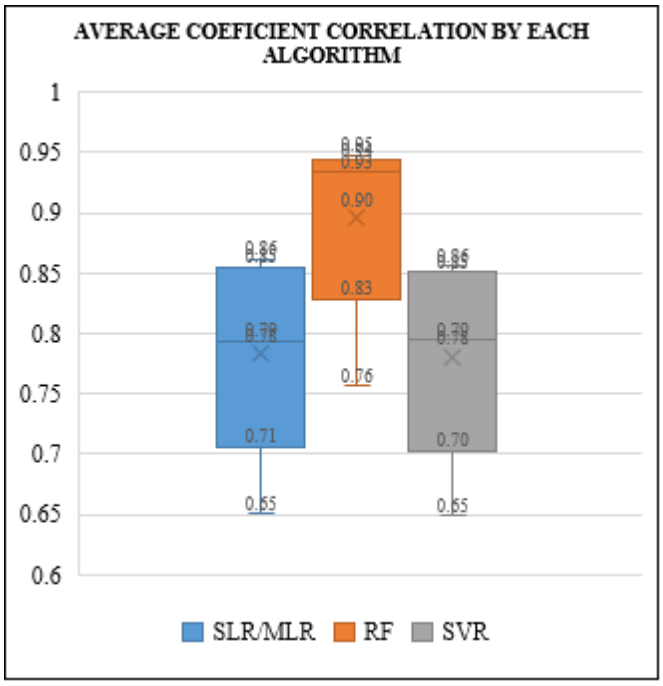

Fig. 13: The box plot of the average correlation coefficient for the developed models.

6 Conclusion

The COVID-19 attributes that are affecting Malaysian death cases have been explored. The goal of our study is to perform experiments in developing a prediction model that will be more significant in the future. Next, the best prediction techniques for modeling COVID-19 trends are identified. After doing several experiments with each modeling technique, the RF regression model achieved the lowest RMSE in all experiments, which is the lowest RMSE of 22.8123 at experiment 5. The performance of RF is also maintained in predicting the occurrence of fatalities. There are opportunities to enhance research in the creation of regression models in the future. Studies can be conducted to identify potential future trends that may impact pandemics or other pandemic fatalities. A new experiment incorporating the variable of vaccination type can also be conducted during the current pandemic. The dataset will soon have over 1,000 data points per attribute, which may increase the experiment's accuracy. At the current time, a modeling experiment was done with a dataset that has not exceeded 1000 data points, so the study's complexity has not been fully realized. Moreover, the developed model in this study can be expanded to build an application with the following advantages:

- for individuals, to identify the possibility of COVID-19 based on symptoms appearing.
- for organizations, predicting the possibility of the risk becoming higher or lower based on the importance of locations.
- for the government, to monitor the risk of spread.

Research incorporating further information or symptoms from hospital records, virus-acquired individuals, COVID-19 survivors, and patients undergoing examination, or treatment may also be taken into consideration. To give more details about the required actions and potential therapies to take into consideration, a model that can determine the likely severity of COVID-19 can also be constructed.

Acknowledgments:
The authors would like to express their gratitude to the Research Management Center, Universiti Teknologi MARA, Malaysia for the research fund 600-RMC/GPM ST 5/3 (021/2021) and College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support. The authors also thank Muhammad Danish Hazim Bin Rahmad and Muhammad Umarul Aiman Bin Mohamad Zulhilmi for their assistance.

References:

[1] A. U. M. Shah, S. N. A. Safri, R. Thevadas, N. K. Noordin, A. A. Rahman, Z. Sekawi, A. Ideris, and M. T. H. Sultan, "COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government," Int. J. Infect. Dis., vol. 97, pp. 108-116, 2020. https://doi.org/10.1016/j.ijid.2020.05.093

[2] A. Elengoe, "COVID-19 Outbreak in Malaysia," Osong Public Health Res. Perspect., vol. 11, no. 3, pp. 93-100, 2020. https://doi.org/10.24171/j.phrp.2020.11.3.08

[3] S. Moore, E. M. Hill, L. Dyson, M. J. Tildesley, and M. J. Keeling, "Modelling optimal vaccination strategy for SARS-CoV-2 in the UK," PLoS Comput. Biol., vol. 17, no. 5, pp. e1008849, 2021. https://doi.org/10.1371/journal.pcbi.1008849

[4] S. Singh, B. M. Sundram, K. Rajendran, K. B. Law, T. Aris, H. Ibrahim, S. C. Dass, and B. S. Gill, "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models," The Journal of Infection in Developing Countries, vol. 14, no. 09, pp. 971–976, 2020. https://doi.org/10.3855/jidc.13116

[5] S. Singh, B. M. Sundram, K. Rajendran, K. B. Law, T. Aris, H. Ibrahim, S. C. Dass, and B. S. Gill, "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models," The Journal of Infection in Developing Countries, vol. 14, no. 09, pp. 971–976, 2020. https://doi.org/10.3855/jidc.13116

[6] M. Ekum and A. Ogunsanya, "Application of hierarchical polynomial regression models to predict transmission of COVID-19 at global level," Int J Clin Biostat Biom, vol. 6, no. 1, pp. 27, 2020.

[7] M. N. Alenezi, F. S. Al-Anzi, and H. Alabdulrazzaq, "Building a sensible SIR estimation model for COVID-19 outspread in Kuwait," Alexandria Engineering Journal, vol. 60, no. 3, pp. 3161–3175, 2021. https://doi.org/10.1016/j.aej.2021.01.025

[8] F. Syed and S. Sibgatullah, "Estimation of the Final Size of the COVID-19 Epidemic in Pakistan," https://doi.org/10.1101/2020.04.01.20050369

[9] F. Nyabadza, F. Chirove, W. Chukwu, and M. V. Visaya, "Modelling the potential impact of social distancing on the COVID-19 epidemic in South Africa," https://doi.org/10.1101/2020.04.21.20074492

[10] H. B. Taboe, K. V. Salako, J. M. Tison, C. N. Ngonghala, and R. G. Kakaï, "Predicting COVID-19 spread in the face of control measures in West Africa," Mathematical Biosciences, vol. 328, p. 108431, 2020. https://doi.org/10.1016/j.mbs.2020.108431

[11] C. Wang, L. Liu, X. Hao, H. Guo, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, A. Pan, S. Wei, and T. Wu, "Evolving Epidemiology and Impact of Non-pharmaceutical Interventions on the Outbreak of Coronavirus Disease 2019 in Wuhan, China," https://doi.org/10.1101/2020.03.03.20030593

[12] H. E. Randolph and L. B. Barreiro, "Herd Immunity: Understanding COVID-19," Immunity, vol. 52, no. 5, pp. 737-741, 2020. https://doi.org/10.1016/j.immuni.2020.04.012

[13] K. B. Law, M. P. K, H. Mohd Ibrahim, and N. H. Abdullah, "Modelling infectious diseases with herd immunity in a randomly mixed population," Sci. Rep., vol. 11, no. 1, pp. 20574, 2021. https://doi.org/10.1038/s41598-021-00013-2

[14] K. M. A. Kabir, K. Kuga, and J. Tanimoto, "Analysis of SIR epidemic model with information spreading of awareness," Chaos, Solitons & Fractals, vol. 119, pp. 118-125, 2019. https://doi.org/10.1016/j.chaos.2018.12.017

[15] P. A. R. Azmi, A. W. Z. Abidin, S. Mutalib, I. S. M. Zawawi and S. A. Halim, "Sentiment Analysis on MySejahtera Application during COVID-19 Pandemic," 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), IPOH, Malaysia, 2022, pp. 215-220, DOI: 10.1109/AiDAS56890.2022.9918748.

[16] N. Khalid, S. Abdul-Rahman, W. Wibowo, N. S. Abdullah, and S. Mutalib, “Leveraging

social media data using latent dirichlet allocation and naïve bayes for mental health sentiment analytics on Covid-19 pandemic," International Journal of Advances in Intelligent Informatics, 9(3), 457-471, 2023, https://doi.org/10.26555/ijain.v9i3.1367

[17] A Aljumah, "Assessment of Machine Learning Techniques in IoT-Based Architecture for the Monitoring and Prediction of COVID-19," Electronics. 2021; 10(15):1834. https://doi.org/10.3390/electronics10151834

[18] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," SN Comput. Sci., vol. 1, no. 4, pp. 206, 2020. https://doi.org/10.1007/s42979-020-00216-w

[19] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset," SN Comput Sci, 2(1), 11, 2021, https://doi.org/10.1007/s42979-020-00394-7

[20] V. Bhadana, A. S. Jalal and P. Pathak, "A Comparative Study of Machine Learning Models for COVID-19 prediction in India," 2020 IEEE 4th Conference on Information Communication Technology (CICT), 1–7, 2020.

[21] C. N. Villavicencio, J. J. E. Macrohon, X. A. Inbaraj, J-H Jeng and J-G Hsieh, "Covid-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA," Algorithms 2021, 14, 201. https://doi.org/10.3390/a14070201

[22] V. Z. Antonopoulos, D. M. Papamichail, V. G. Aschonitis, and A. V. Antonopoulos, "Solar radiation estimation methods using ANN and empirical models," Comput. Electron. Agric., vol. 160, pp. 160-167, 2019. https://doi.org/10.1016/j.compag.2019.03.022

[23] A. Khazaee Poul, M. Shourian, and H. Ebrahimi, "A Comparative Study of MLR, KNN, ANN and ANFIS Models with Wavelet Transform in Monthly Stream Flow Prediction," Water Resour. Manag., vol. 33, no. 8, pp. 2907-2923, 2019. https://doi.org/10.1007/s11269-019-02273-0

[24] Y. Chen, W. Zheng, W. Li and Y. Huang, "Large group activity security risk assessment and risk early warning based on random forest algorithm." Pattern Recognition Letters, 144, pp1-5, 2021, https://doi.org/10.1016/j.patrec.2021.01.008

[25] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, "A Time-Dependent SIR Model for COVID-19 With Undetectable Infected Persons," IEEE Trans. Netw. Sci. Eng., vol. 7, no. 4, pp. 3279-3294, 2020. https://doi.org/10.1109/TNSE.2020.3024723

[26] J. Dou, A. P. Yunus, D. Tien Bui, A. Merghadi, M. Sahana, Z. Zhu, C. W. Chen, K. Khosravi, Y. Yang, and B. T. Pham, "Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan," Sci. Total Environ., vol. 662, pp. 332-346, 2019. https://doi.org/10.1016/j.scitotenv.2019.01.221

[27] P. Aghelpour, B. Mohammadi, and S. M. Biazar, "Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA," Theoretical and Applied Climatology, vol. 138, no. 3-4, pp. 1471-1480, 2019. https://doi.org/10.1007/s00704-019-02905-w

[28] Q. Quan, Z. Hao, H. Xifeng, and L. Jingchun, "Research on water temperature prediction based on improved support vector regression," Neural Comput. Appl., vol. 34, no. 11, pp. 8501-8510, 2020. https://doi.org/10.1007/s00521-020-04836-4

[29] Ministry of Health Malaysia, "Official data - COVID-19," 2022. [Online], https://github.com/MoH-Malaysia/covid19-public (Accessed Date: October 31, 2023)

[30] T. O. Hodson, T. O., "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development," 15(14), 5481–5487, 2022. https://doi.org/10.5194/gmd-15-5481-2022

[31] A. A. Suleiman, U. A. Abdullahi, A. Suleiman, S. A. Suleiman, and H. U. Abubakar, "Correlation and Regression Model for Physicochemical Quality of Groundwater in the Jaen District of Kano State, Nigeria," Journal of Statistical Modeling and Analytics, Vol. 4, Issue 1, 2022. https://doi.org/10.22452/josma.vol4no1.2

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

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

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

The authors would like to express their gratitude to the Research Management Center, Universiti Teknologi MARA, Malaysia for the research fund 600-RMC/GPM ST 5/3 (021/2021)

Conflict of Interest

The authors have no conflicts of interest to declare.