# Evaluation of using Parametric and Non-parametric Machine Learning Algorithms for Covid-19 Forecasting

Ghada E. Atteia[1], Hanan A. Mengash[2], Nagwan Abdel Samee[3]*
Information Technology Department, College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11461 Saudi Arabia[1, 3]
Information Systems Department, College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11461 Saudi Arabia[2]
Computer Engineering Department, Misr University for Science and Technology, Giza, 12511 Egypt[3]

***Abstract*—Machine learning prediction algorithms are considered powerful tools that could provide accurate insights about the spread and mortality of the novel Covid-19 disease. In this paper, a comparative study is introduced to evaluate the use of several parametric and non-parametric machine learning methods to model the total number of Covid-19 cases (TC) and total deaths (TD). A number of input features from the available Covid-19 time sequence are investigated to select the most significant model predictors. The impact of using the number of PCR tests as a model predictor is uniquely investigated in this study. The parametric regression including the Linear, Log, Polynomial, Generative Additive Regression, and Spline Regression and the non-parametric K-Nearest Neighborhood (KNN), Support Vector machine (SVM) and the Decision Tree (DT) have been utilized for building the models. The findings show that, for the used dataset, the linear regression is more accurate than the non-parametric models in predicting TC & TD. It is also found that including the total number of tests in the mortality model significantly increases its prediction accuracy.**

***Keywords*—Covid-19; parametric regression; non-parametric regression; linear regression; log regression; polynomial regression; generative additive regression; spline regression; k-nearest neighborhood; KNN; support vector machine; SVM; decision trees; DT***

## I. Introduction

Once the coronavirus pandemic, Covid-19, broke out at the late of December 2019, in Wuhan, China, the virus has been spread all over the world by the Spring of 2020. The coronavirus pandemic has so far followed a wave pattern, with increases in new cases followed by reductions [1]. SARS-CoV-2, the coronavirus that causes Covid-19, has mutated since the beginning of the pandemic, resulting in variations of the disease symptoms [2]. The delta variation is one of these mutations. The delta coronavirus is one of the most contagious coronavirus strains to date [3]. Presently, some countries are suffering from the fourth wave of the pandemic with the severest mutated version of the virus, delta variant. The current total number of confirmed cases of Covid-19 approaches 245 million persons worldwide with nearly five million total deaths [4]. The unpredictable rapid spread of the pandemic all over the world has caused unprecedented global lockdowns and overwhelmed the healthcare systems. As no medicine has been approved yet for this virus, the World Health Organization (WHO) has guaranteed the availability of Covid-19 clinical data for the majority of countries and encouraged the research community to provide support in this pandemic to "fight panic with information" [5][6]. This would certainly aid in directing governments toward proper crisis management and effective resource utilization to contain the pandemic.

Many recent studies have tackled the problem of forecasting the spread and mortality of the new coronavirus disease using various machine learning prediction methods. Based on the survey done in [7], most studies focused only on addressing the relationship between the numbers of confirmed and recovered cases and deaths to build models for predicting the spread of the coronavirus disease. However, there are other features that would significantly affect the prediction accuracy of these models.

In this paper, we propose a comparative study to evaluate the use of several parametric & non-parametric machine learning regression methods to model the two main folds of Covid-19 spread: the total number of confirmed cases and the total number of deaths. Within the study framework, we seek for the most significant input features of the models and investigate the impact of the number of tests on the prediction performance. The proposed framework has two phases: The Data Analytics & Modeling Phase and the Future Prediction Phase. In the first phase, Covid-19 time sequence dataset is preprocessed, and several significant predictors are selected according to a correlation criterion. These predictors are then used to build several regression models using several parametric & non-parametric methods using the training subset of the data. The model that shows the best prediction performance in terms of the least RMSE value will be considered for making the future predictions in the following phase. In the Future Prediction Phase, the values of the total deaths & the number of the total cases are to be predicted at future dates. In order to do so, the selected predictors should be estimated at the required future dates as well. Therefore, in this phase, each predictor is modeled individually against time (the day count referenced to an origin date) using a set of parametric & non-parametric methods. The best model is then used to estimate the value of the corresponding predictor at the required future date and predictor value is then substituted in

*Corresponding Author

the total cases model as well as the total death model. The proposed framework has been applied on the Covid-19 dataset of Saudi Arabia over 116 days from April 25 till August 8, 2020 for training & testing the prediction models and these models have been used for estimating the future values of the total number of cases and total number of deaths.

## II. Literature Review

Several factors have influenced whether new Covid-19 cases are increasing or decreasing in specific locations during the pandemic. Some of these factors include the efficiency of vaccination, adhering the precautionary measures, the virus mutations, and the PCR tests. For instance, there was a huge surge in the number of Covid-19 confirmed cases during the winter of 2021 in the United States as a result of people not adhering to the COVID-19 precautions and regulations. Additionally, in many countries, vaccinating the citizens has aided in bringing new infection levels down until the spring season of 2021.

The number of PCR tests is one of the most important features that could significantly contribute to the prediction accuracy of the spread/ mortality models as it is explicitly affecting the number of confirmed cases. Nonetheless, no studies, to the best of our knowledge, have included the number of tests as an input feature to the Covid-related prediction models, nor have they examined its impact on the prediction accuracy of those models. For instance, the study of Yuanyuan et al. The work done in [8] utilized a linear regression analysis to create a model between the number of Wuhan roaming people and the cumulative number of Covid-19 cases in Henan province, China. Another study by Sansa et al. [9] conducted a correlation analysis and built a simple linear regression model between the numbers of confirmed cases and recovered cases in China over one month period. In another study[10], the epidemic peak in Saudi Arabia was predicted using the (Susceptible-Infected-Recovered) model [11], and the Logistic Growth model[12]. In that study, four variables were considered in the prediction models which are the number of daily confirmed, accumulated confirmed, recovered and deaths cases. Other studies utilized a number of non-parametric machine learning approaches to forecast the worldwide spread & death rate of Covid-19 and other pandemic-related variables as in [13][14][15] . The Naïve method, averaging, and Holt linear/winters method have been used in[14] to predict the value of the number of deaths in the next day based on the value of the present day. Another work in [16] has presented the application of linear and logistic regression for the prediction of the risk periods and survival of Covid-19 in different ages. However, the Decision Tree (DT)[17], K-Nearest Neighborhood (KNN) [18], and Support Vector Machine (SVM) [19] have been employed for the classification of patients (risk/mild) and hence the significant features have been extracted to distinguish between the classes of patients. In addition, DT, SVM, Random Forest, KNN, Naive Bayes, and logistic regression were employed in [20] to predict the number of days needed to recover from Covid-19 and the age of patient that may result in risky outcomes of the disease.

## III. Materials

In this work, a data set of COVID-19 records for Saudi Arabia [3] is used for building and evaluating the regression models. This dataset is published in the upstream repository at Johns Hopkins University Center for Systems Science and Engineering website [17]. The Covid-19 data set records the number of new confirmed cases, new deaths and recovered cases daily along with the corresponding accumulated total numbers. Other auxiliary entries like the median patient age, population, diabetes prevalence and others are also included in the data [2]. These auxiliary entries have constant values across the days. The number of new tests and total tests were recorded as well starting May 13th, 2020 for the Saudi Arabia data [2]. In this work, the entries with variable values are only used to model the number of the total confirmed cases and the total deaths using regression while the auxiliary entries were ignored as they do not contribute significantly to the models. There were four missing entries for the total tests and their values were estimated using the average of its two adjacent values. Day counts have been created to be used in reference to the required date. Day counts start from April 25th, 2020; i.e. Day 1 corresponds to April 25th, Day 2 to April 26th and so on. The available records are divided randomly into a training data set and a testing data set with a ratio of 8:2. The training data is used to estimate the regression coefficients of the prediction models while the testing set is used to evaluate the prediction accuracy of the proposed models. In order to unify the range of the input observations, the min-max normalization [18] is used to normalize the input features before building the models. All the codes of this work are created using the R programming language. For convenience, the following notations are used for the variables throughout the paper. TC, TR, ND, TD, TT, and DC denotes the number of the Total Confirmed Cases, the number of the Recovered Cases, the number of the New Deaths, the number of the Total Deaths, the number of the Total Tests, the Day Count.

## IV. Methods

Regression is a supervised machine learning technique that is used for the prediction of a continuous quantitative outcome. For this purpose, the relationship between a dependent (response) variable and one or more independent variables (predictors) in a labeled dataset is estimated during the regression analysis process. Regression can be implemented using parametric and non-parametric algorithms. If a dataset is collected about a response variable Y, and predictor variables $(x_1, x_2, x_3, .., x_m)$, the relationship between Y and X can be modeled as in Eq. (1) [21].

$$Y = f(X, C) + C_o \quad (1)$$

Where, $C$ is a vector of m parameters, $C_o$ is an error term that shows the deviation of the actual values from the model predictions and f(.) is some function that maps the relationship between Y and X. The selection to use the parametric, semi-parametric or nonparametric method to implement the regression model depends mainly on the prior knowledge about the form of the function f(.). If f(.) is known a priory, parametric methods is to be used; otherwise, non-parametric methods should be used. Semi-parametric methods can be used if f(.) is known partially [21]. The function f(.) could be

linear or non-linear function in the model parameters and accordingly the model becomes a linear or non-linear parametric model respectively. Parametric models require the estimation of the model parameters $C$ and $C_o$. It is noteworthy mentioning that parametric models perform the best when the relational function is known and correct. In contrast, using the wrong function would result in larger bias when compared to the other competitive models [21] and would make inaccurate predictions. The most common parametric regression is the linear regression in which a linear model is composed of linear combination of the input predictors. Non-parametric regression methods do not require pre-knowing the form of f(.) and consequently, they provide more flexibility in analyzing the relationship between the variables [21]. Many machine learning algorithms that are used for classification can be used as non-parametric regressors with some structural amendments when the response variable is continuous rather than discrete. The K-Nearest Neighborhood (KNN), Support Vector Machine (SVM) and Decision Tree (DT) algorithms are examples of such non-parametric regression methods.

### *A. Parametric Machine Learning Regression*

To get sense of the relation between the dependent variable and each of the predictors, a set of scatter plots are provided in Fig. 1 for the total number of deaths and in Fig. 2 for the total number of confirmed cases. The scatter plots show that the relationship between the response variables and all predictors, individually, are increasing and could be linearly modeled using the multivariate parametric linear regression.

TD Linear Regression Models

As the TD is highly correlated with the TC, TR & TT, the proposed prediction model of the TD in Experiment 1 is given in Eq. (2) while that of Experiment 2 after excluding TT, is given in Eq. (3) :

$$TD = C_0 + C_1 * TC + C_2 * TR + C_3 * TT \quad (2)$$

$$TD = C_0 + C_1 * TC + C_2 * TR \quad (3)$$

Where $C_0, C_1, C_2, C_3$ are the regression coefficients of the model which represent the association of the model predictors to the dependent variable.

TC Linear Regression Models

The proposed prediction model of (TC, TT&TR) is given as in Eq. (4) and that of the (TC,TR) is given in Eq. (5):

$$TC = B_0 + B_1 * TR + B_2 * TT \quad (4)$$

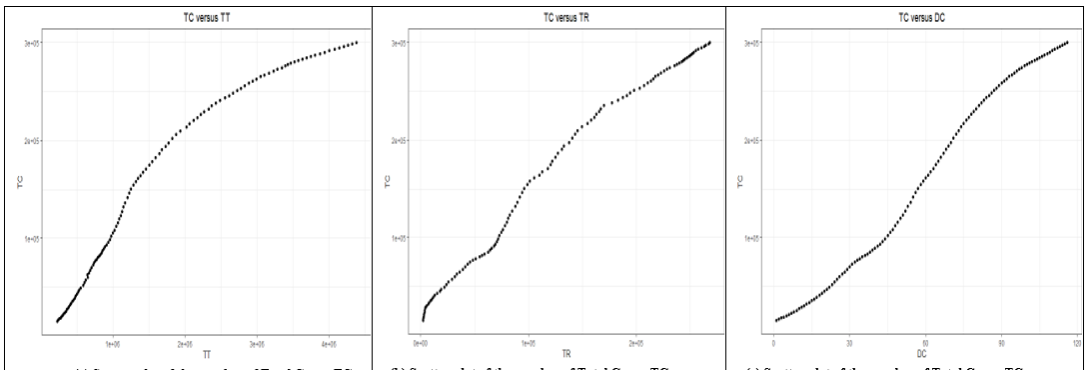
$$TC = B_0 + B_1 * TR \quad (5)$$


(a) Scatter plot of the number of Total Deaths, TD, versus the number of Total Cases, TC.

(b) Scatter plot of the number of Total Deaths, TD, versus the number of Total Tests, TT.

(c) Scatter plot of the number of Total Deaths, TD, versus the number of Total Recovered, TR

Fig. 1. Scatter Plot of the Total Deaths (TD) Versus Total Cases (TC), Total Tests (TT), and Total Recovered (TR).




(a) Scatter plot of the number of Total Cases, TC versus the number of Total Tests, TT.

(b) Scatter plot of the number of Total Cases, TC versus the number of Total Recovered, TR.

(c) Scatter plot of the number of Total Cases, TC versus Day Count, DC.

Fig. 2. Scatter Plot of the TC Versus Total Tests (TT), Total Recovered (TR), and Day Count (DC).

Where $B_0, B_1, B_2$ are the regression coefficients of the model. The model coefficients for all of the linear models built in this study are estimated using the Least Squares Estimation algorithm.

### B. Non-Parametric Machine Learning Regression

In this part, the TC and the TD are modeled using a number of supervised learning non-parametric algorithms. Non-parametric algorithms do not make an assumption about the relationship between the response and predictors or the underlying distribution of the data and the model structure is configured from the data itself. In this study, the KNN, SVR and the DT algorithms are used for manipulating the regression problem.

KNN is a non-parametric supervised machine learning algorithm that is used for classification and regression. KNN approximates the association between the input features and the response variable using feature similarity[22]. In classification, KNN finds the majority votes of a number of neighbors (called k) of an input instance to select the appropriate class. However, in regression, the response variable is estimated by averaging the observations in the nearest neighborhood of the input instance based on a similarity measure. The similarity measure employed herein is the Euclidian Distance [23]. In order to select the optimal value of k, we run the KNN algorithm on the training dataset with k values starts from 3 up to 8 and calculate the RMSE at each k value then select the value that minimizes the root mean-squared error. k values of 1 & 2 are excluded as they cause unstable predictions. Also, k values greater than 8 are excluded as it has been observed that the RMSE values keep increasing as k increases.

Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for classification and regression tasks. In a classification problem, SVM tries to find a hyperplane in the input feature space to distinctly classify the input data points[24]. Finding the hyperplane is an optimization problem to select the plane that achieves the maximum margin between the data points of two classes using the aid of kernel functions[25]. For a regression problem, SVM is known as SVR (Support Vector Regressor) and the problem then is to find a function that approximates input features to real numbers instead of discrete classes. This function itself defines the hyperplane in the regression problem and is used for the prediction of the response variable. This is again an optimization problem that aims to find the best hyperplane that passes through the maximum number of points within a given decision boundary at distance "ε" from the hyperplane. Let's consider that the hyperplane is a straight line as in Eq. (6) [24]:

$$y = wx + b \tag{6}$$

Where $w, b$ are the parameters of the line. Then the decision boundary can be defined as in Eq. (7), and Eq. (8):

$$wx + b = +\varepsilon \tag{7}$$

$$wx + b = -\varepsilon \tag{8}$$

So, any hyperplane that satisfies our SVR should satisfy Eq. (9) [24]:

$$-\varepsilon < y - wx + b < +\varepsilon \tag{9}$$

In this part of study, as no assumptions are made about the multivariate input or their relationships to the response variable, therefore, multiple kernel functions are used to adapt to the patterns in the data. The linear, polynomial, Gaussian radial basis and the sigmoid kernel functions [25] have been employed to non-linearly map the data from the original space into a higher dimensional space.

Decision Tree (DT) is a well-established supervised machine learning algorithm that can be used for classification and regression [26]. A decision tree makes decisions by splitting nodes into sub-nodes using the "if, then" condition multiple times until reaching the terminal homogeneous nodes. In this work, the Recursive partitioning has been employed to build the regression models of the response variables. The models are built against the predictors that show very high correlation with the response as depicted in Table I. As we are tackling a regression problem, we used the ANOVA splitting rule as the partitioning method of the tree. ANOVA rule is based on the Reduction of Variance concept to split the nodes. For each split, ANOVA calculates the variance of each node and then the variance of the split and then selects the split with the lowest variance. This process is repeated until all nodes with zero variance are reached and marked as the terminal nodes. At this end, no further splits are needed[26]. The ANOVA splitting rule is used as the partitioning methods of the tree. To pre-prune the Decision Tree, three hyperparameters are tuned and optimized. That is, the Complexity Parameter (CP), the Maximum Depth (MD) and the Minimum Split (MS). Complexity Parameter is used to save computing time by pruning off splits that does not improve the fit's R-squared value by the value of (CP). The Maximum Depth indicates how deep the tree can be. The Minimum Split of the parent node which is the minimum number of observations in the parent node that can be split further[27]. To optimize the values of these hyperparameters, the R function "Rpart.tune" is used.

### C. The Study Framework

In this study, two models are to be built for the prediction of two response variables separately: the total number of confirmed cases (TC) and the total number of deaths (TD). Several parametric and non-parametric machine learning regression methods are utilized to build the models. The models will be evaluated based on some performance metrics and the best performing model will be considered for the future predictions of the response variables. The framework, shown in Fig. 3, is composed of two phases:

Phase 1: Data Analytics and Modelling

As a first step in this phase, data is explored to determine the significant predictors (the independent variables) to be used in building the models. A correlation analysis between all the input variables in the data has been conducted and the Pearson Correlation Coefficients (PCC)[28] are depicted in the correlation matrix in Table I. Only highly correlated variables (PCC>0.9) with the response variable are considered significant and used as predictors of the corresponding model. In Table I, highly correlated variables with the total confirmed

cases are highlighted in light grey while those highly correlated with the total deaths, are highlighted in dark grey.

After selecting the significant predictors, several parametric & non-parametric regression methods are used to model the total number of confirmed cases and the total number of deaths. At last, the model that shows the best prediction performance is selected for the future prediction in phase 2 of the framework.

The prediction model of the total number of deaths are built using the predictors that show high correlation with it which are the total number of tests, the total number of recovered cases and the total number of confirmed cases as shown in Table I. However, it was noted that the effect of the total number of tests on the Covid-19 prediction models is not investigated widely in the literature. Most probably this is because recording the TT on a daily basis was started late in most countries. Therefore, it has been decided in this study to figure out the impact of the total number of tests on the prediction accuracy of the proposed regression models. This is achieved by conducting two experiments for modeling the TD. In Experiment 1, all predictors that are highly correlated with the TD (which are TT, TR and TC) are used to build the model using the multivariate regression paradigm. On the other, the TT is excluded in Experiment 2 and the model is constructed using only TR and TC.

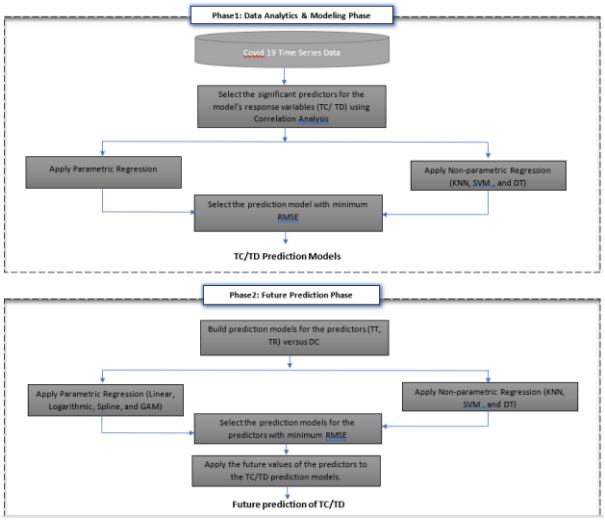



Fig. 3. The Study Framework for Predicting the Total Number of Cases & Total Number of Deaths of Covid-19 Outbreak.

TABLE I. CORRELATION MATRIX OF ALL VARIABLES INCLUDED IN THE STUDY

| | DC | TR | TC | TT | TD | ND | NC |
|---|---|---|---|---|---|---|---|
| DC | 1 | 0.99 | 0.994 | 0.969 | 0.982 | 0.658 | 0.032 |
| TR | 0.992 | 1 | 0.992 | 0.988 | 0.995 | 0.592 | 0.072 |
| TC | 0.994 | 0.992 | 1 | 0.965 | 0.989 | 0.658 | 0.034 |
| TT | 0.969 | 0.988 | 0.965 | 1 | 0.988 | 0.498 | 0.186 |
| TD | 0.981 | 0.994 | 0.989 | 0.988 | 1 | 0.580 | 0.083 |
| ND | 0.657 | 0.592 | 0.658 | 0.497 | 0.580 | 1 | 0.590 |
| NC | 0.033 | 0.072 | 0.034 | 0.186 | 0.083 | 0.59 | 1 |

The prediction of the total number of confirmed cases is one main fold in tracing the spread of a pandemic. Therefore, an accurate model should be developed for the prediction of the total number of confirmed cases. In this study, two approaches are used to build and select the suitable TC model. In the first approach, a univariate prediction model is built for the TC using the day count as will be described later in this section. In the second approach, the multivariate regression is used to model the TC against the most significant predictors according to the high correlation criterion following the two experiments as in the TD model. In Experiment 1, according to the correlation criterion and as depicted in Table I, the TR and the TT achieve the highest correlation with the TC with PCC > 0.9 and hence are used as the model predictors in this approach. Although, the TD shows high correlation with the TC, the former has been excluded while building the TC model. This has been decided to avoid any inaccuracy due to duplication as the TD model is considered the primary model and has already taken the TR and the TT in the prediction of TD. In Experiment 2, the TT is excluded from the model and the TR is the only predictor of the model.

After the TD & TC models from the two approaches are built by a set of parametric and non-parametric regressors, some performance metrics are then applied to evaluate the performance of the prediction models on the testing data set. The model that achieves the highest performance measures on the testing dataset are selected to be used for the prediction of the TC.

Phase 2: Future Prediction

As it is one of our objectives in this study to track the spread of Covid-19, values of the total number of confirmed cases and the total number of deaths are to be calculated at future dates. Given that the prediction models require the future values of their correspondent predictors, the values of these predictors are unknown apriori and need to be estimated beforehand at the required dates. Therefore, in this phase, each of the selected predictors is modeled individually against the day count. After that, the predictors' future values are substituted in the TC/TD forecasting models to find their corresponding future predictions. A number of parametric & non-parametric regressors are used to model the univariate predictors against the day count and the model with the least RMSE value is considered.

## V. RESULT AND DISCUSSION

In this section, the results related to the TC model are presented first followed by the results of the TD model. Within this arrangement, we present the models built using the parametric linear regression then those built using the non-parametric methods. To evaluate the performance of the regression models developed in this study, a number of well-known performance metrics are utilized. The Min-Max accuracy, MAPE, the Root Mean Squared Error (RMSE), the R-Squared, Error rate of the RMSE referenced to the mean of the actual values and the correlation accuracy are used to evaluate the accuracy of predictions on the testing data[29][30][31]. The model that achieves the highest significance and prediction accuracy will be used for making the future prediction of the total cases and deaths.

### A. The Total Number of Confirmed Cases Prediction Model (TC Model)

Within the proposed framework for TC prediction, two approaches are used to model the total number of confirmed cases. In one approach, a univariate model that relates the TC with the DC is constructed. However, in the other approach, the highly correlated predictors with the TC (which are the TT & TR) are used to build the model. Under this approach, two experiments are conducted to investigate the effect of the TT on the TC prediction model. In Experiment 1, a model that relates the TC to both the TT & TR is built while in Experiment 2, the TT is excluded, and a univariate regression model is constructed using the TC & TR training data. Several regression models are built using the parametric linear regression and the KNN, SVR & DT non-parametric methods. The performance of each of the proposed models is assessed using the measures described in the Methods Section. The model that best fit the training data and that provides the highest prediction accuracy on the testing data is selected to be used in estimating the future value of the TC predictor required in the TD model.

#### 1) Parametric Linear Regression

In this part, the relation between the predictors (TR, TT, DC) and the dependent variable (TC) is assumed to be linear. We have used two approaches in modeling TC. In the First Approach, TC is modeled versus predictors with high correlation with the response variable. And in the second approach TC is modeled only versus DC. In the first experiment under the first approach, we model TC versus TR & TT. To check the statistical significance of the estimated model coefficients, the standard error, p-value and the t-value are calculated after building the model using the training dataset as shown in Table II. The low values of these metrics reveal that the estimated coefficients are significant.

The accuracy of the TC model on the testing data has been evaluated using the Min-Max accuracy, the Mean Absolute Percentage Error (MAPE) and the R-Squared metrics. An average value between the maximum and minimum predictions has been retrieved as 94 % with a MAPE value of 0.063 which show a good accuracy of the prediction model over the testing data. The RMSE value of 6826 implies that there is an average alteration between the actual and the predicted values in the testing subset with an error rate of 5.27%. The value of the 0.99 for the R-squared reveals the high correlation between the actual and predicted values. This is consistent with the correlation accuracy of 0. 9973 computed after predicting the TC for the test data. This implies that the actual and the predicted values have analogous directional movement in which the actuals values increase as the predicted values increase and vice-versa.

TABLE II. SUMMARY OF THE STATISTICAL SIGNIFICANCE OF THE ESTIMATED COEFFICIENTS OF THE (TC- TT& TR) PREDICTION MODEL

| | **Estimated Coefficient** | **STD Error** | **t-value** | **p-value** |
|---|---|---|---|---|
| $B_0$ | 25520 | 1560 | 16.36 | < 2e-16 |
| $B_1$ | -194135 | 18922 | -10.26 | < 2e-16 |
| $B_2$ | 464305 | 17319 | 26.81 | < 2e-16 |

TABLE III. SUMMARY OF THE STATISTICAL SIGNIFICANCE OF THE ESTIMATED COEFFICIENTS OF THE (TC- TR) PREDICTION MODEL

| | Estimated coefficient | STD Error | t.value | p-value |
|---|---|---|---|---|
| B0 | 1.084e+00 | 1.426e-02 | 75.98 | <2e-16 |
| B1 | 2.884e+04 | 2.091e+03 | 13.79 | <2e-16 |

In the second experiment under this approach, the first approach, we model TC versus TR only. Like what has been done in Experiment 1, the statistical significance of the (TC, TR) model (given in Equation 4) are calculated and shown in Table III. The retrieved results of Min-Max accuracy, MAPE, RMSE, and R-squared are 91%, 0.1, 12812, and 0.98 respectively which are worse than the values for the (TC, TT&TR) model. The values of the performance measures depict that excluding the TT from the model reduces its statistical significance and reduces the prediction accuracy as well.

In the Second Approach, TC versus DC Model, the training dataset of the day count and the total number of cases (DC, TC) is used to fit a model for the TC. Five models have been built using the Linear, Logarithmic, Spline, Polynomial and the Generative Additive Regression. Scatter plots of these models are shown in Fig. 4. The R-squared values of these models vary from 0.8 to nearly 1. The Logarithmic regression provides the worst fit with the lowest R-squared value of (0.79) followed by the Linear regression model. The Spline regression and the Polynomial regression provide comparable R-squared values while the Generative Additive Model (GAM) provides the best fit in terms of the highest R-squared value. Therefore, the GAM model is considered here for further statistical significance analysis.

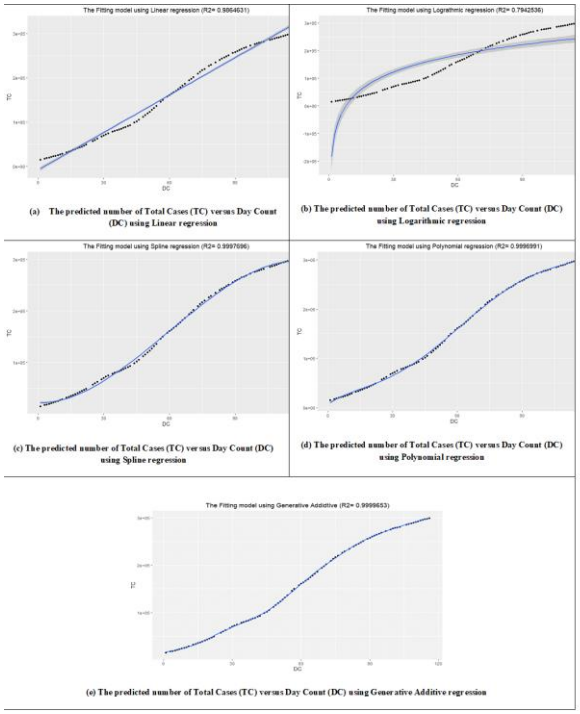



Fig. 4. The Predicted Total Cases (TC) Versus Day Count (DC) using: a) Linear Regression b) Logarithmic Regression c) Spline Regression d) Polynomial Regression e) Generative Additive Regression.

In an assessment of the prediction accuracy of the GAM model on the training data, the Adjusted and Multiple R-squared and the F-statistics are computed. The values of all R-squared measures are 1 which indicate that the variability in the TC is captured perfectly by the prediction model. This is supported by the very large value of the F-statistic (124906) and the very low p-value which reflect the high significance of the model. Therefore, this model was used to predict the TC values for the testing data and the performance metrics were computed to evaluate the prediction accuracy of the model. A Min-Max accuracy of 98.9% and a MAPE value of 0.011 were obtained for the model. The RMSE value of 1018 implies that there is a low average alteration between the actual and the predicted values in the testing subset with an error rate of 0.63%. The value of the 0.9999 for the R-squared reveals the high correlation between the actual and predicted values. This is consistent with the correlation accuracy of 0.9999 computed after predicting the TC for the test data.

*2) Non-parametric Machine Learning Regression*

In this part, no assumptions about the relation between the predictors (TR, TT, DC) and the dependent variable (TC) are made and the TC model is estimated from the data using the KNN, SVM and the DT regression methods. The performance measures calculated for all non-parametric methods are depicted in a table for each model and the model with the lowest RMSE is highlighted in light grey to facilitate the visual interpretation of the results. At the end, a comparison is conducted between the parametric and non-parametric models based on the RMSE measure to select the model that will be used for future predictions. Also, we have used two approaches in modeling TC as done in the Parametric regression.

In the First Approach, TC is modeled versus predictors with high correlation with the response variable. In the first experiment under this approach, we model TC versus TR & TT non-parametrically. Table IV shows the summary of the accuracy metrics for the models built by the KNN, SVM and the Decision Tree Regression. For the KNN, it is obvious that as the K increases, the larger the RMSE values are. Among all K values, the lowest RMSE & MAPE are achieved when the number of neighbor points equals 3. This k value also corresponds to the highest R-squared & Min-Max accuracy. For the SVM regression, the optimization tuning function "tune.svm" in the R language is used to deliver the best Gamma & cost parameters values for the Polynomial, Sigmoid & the Radial bases kernels for the SVM model. Values of the retrieved parameters are given in the caption of the table. It is noticed that the Radial kernel offers the least RMSE among the other kernels, yet still performing worse than the KNN. The Decision Tree Regressor has the worst performance over all non-parametric methods while the opposite is true for the KNN.

TABLE IV. SUMMARY OF THE ACCURACY OF THE (TC- TT & TR) PREDICTION MODEL ON THE TESTING DATASET USING THE KNN, SVM (GAMMA = 0.001, COST = 10 FOR POLYNOMIAL, RADIAL, SIGMOID KERNEL FUNCTION), AND DECISION TREE (BEST PARAMETERS: MAX DEPTH=3, CP=0.002, AND MINI SPLIT=10)

| | **Learning Parameters** | **RMSE** | **R2** | **Min-Max Accuracy** | **MAPE** |
|---|---|---|---|---|---|
| KNN | k=3 | 1907.4 | 0.999 | 0.976 | 0.026 |
| | k=4 | 2356.3 | 0.999 | 0.974 | 0.029 |
| | k=5 | 2085.1 | 0.999 | 0.970 | 0.034 |
| | k=6 | 2501 | 0.999 | 0.970 | 0.036 |
| | k=7 | 2969.6 | 0.999 | 0.964 | 0.043 |
| | k=8 | 3099 | 0.999 | 0.961 | 0.048 |
| SVM | Linear Kernel | 9349 | 0.991 | 0.907 | 0.118 |
| | Polynomial Kernel | 36644 | 0.85 | 0.77 | 0.29 |
| | Radial Kernel | 6326.7 | 0.996 | 0.927 | 0.099 |
| | Sigmoid kernel | 12713 | 0.649 | 0.38 | 0.778 |
| DT | Anova Partitioning Method | 11388.9 | 0.98 | 0.894 | 0.142 |

In the Second Approach, TC is modeled versus DC. Table V shows that the KNN with k=3 achieves the lowest error and the highest accuracy over all KNNs. Also, it has been found that the Radial kernel SVM is the best performer over all SVRs followed by the linear kernel. Decision tree performs comparably with the linear SVM and better than the Sigmoid SUM. However, again, the KNN with k = 3 is the best regressor over the other non-parametric algorithms and is highlighted in grey in Table V.

TABLE V. SUMMARY OF THE ACCURACY OF THE (TC-DC) PREDICTION MODEL USING KNN, SVM (GAMMA = 0.1, COST = 10 FOR POLYNOMIAL, RADIAL, SIGMOID KERNEL FUNCTION), AND DECISION TREE (BEST PARAMETERS: MAX DEPTH=3, CP=0.002, AND MINI SPLIT=10)

| | **Learning Parameters** | **RMSE** | **R2** | **Min-Max Accuracy** | **MAPE** |
|---|---|---|---|---|---|
| KNN | k=3 | 2232 | 0.99 | 0.97 | 0.032 |
| | k=4 | 2619 | 0.99 | 0.96 | 0.03 |
| | k=5 | 2938 | 0.99 | 0.96 | 0.04 |
| | k=6 | 2806 | 0.99 | 0.96 | 0.04 |
| | k=7 | 3639 | 0.99 | 0.95 | 0.05 |
| | k=8 | 4004 | 0.99 | 0.94 | 0.06 |
| SVM | Linear Kernel | 13143 | 0.98 | 0.87 | 0.18 |
| | Polynomial Kernel | 36607 | 0.85 | 0.76 | 0.31 |
| | Radial Kernel | 7913 | 0.99 | 0.94 | 0.06 |
| | Sigmoid kernel | 18408 | 0.96 | 0.86 | 0.16 |
| DT | Anova Partitioning Method | 12087 | 0.98 | 0.89 | 0.12 |

### B. The Total Number of Deaths Prediction Model (TD Model)

In order to build the TD model, two experiments were conducted as aforementioned in Sec 3 in which the impact of the total number of tests on the prediction accuracy of the TD model is investigated. Several models are built using the parametric linear regression and the KNN, SVR & Decision Tree Non-parametric methods. The performance of each of the proposed models is assessed and the best fit will be used to estimate the total number of deaths.

#### 1) Parametric Linear Regression

As a first Experiment, the TT, TR and the TC are used to model the TD using linear regression given in Equation 1. These predictors show very high correlation with the TD as illustrated in the scatter plots of Fig. 1. Table VI shows that the TC & TT coefficients have highest significance followed by the TR.

The accuracy of the TD model on the testing data has been evaluated. A Min-Max accuracy of 86% with a MAPE value of 0.13 is obtained for this model. The RMSE value of about 72 implies that there is very low average alteration between the actual and the predicted values in the testing data with an error rate of 4.25 %. A value of 0.995 for the R-squared and a correlation accuracy of 0. 998 show that the actual and predicted values are highly correlated.

In the second Experiment, the TT is excluded, and the TR and the TC are used to model the TD using linear regression given in Equation 2. Table VII demonstrates the model significance over the training data. This table shows that the model coefficients have higher STD error, p-value & t-value than those obtained in Table VI for Experiment 1 using the TT as a model predictor. The accuracy of the TD model on the testing data has been computed. It has been found that the retrieved results of the Min-Max accuracy, MAPE, RMSE, and R-squared are 82%, 0.19, 97, 0.992 correspondingly which are worse than the values for the (TC, TT&TR) model.

TABLE VI. SUMMARY OF THE STATISTICAL SIGNIFICANCE OF THE ESTIMATED COEFFICIENTS OF THE (TD- TC& TR& TT) PREDICTION MODEL

| | Estimated coefficient | STD Error | t-value | p-value |
|---|---|---|---|---|
| C0 | -75.43 | 13.99 | -12.378 | 5.32e-07 |
| C1 | 2873.00 | 217.65 | -5.212 | < 2e-16 |
| C2 | -2218.01 | 377.07 | 12.933 | 6.47e-08 |
| C3 | 2927.68 | 8.251e-04 | 203.60 | < 2e-16 |

TABLE VII. SUMMARY OF THE STATISTICAL SIGNIFICANCE OF THE ESTIMATED COEFFICIENTS OF THE (TD- TC& TR) PREDICTION MODEL

| | Estimated coefficient | STD Error | t-value | p-value |
|---|---|---|---|---|
| C0 | -30.99 | 24.45 | -1.268 | 0.208 |
| C1 | 2741.32 | 273.18 | 10.035 | <2e-16 |
| C2 | 592.64 | 267.18 | 2.218 | 0.029 |

#### 2) Non-parametric Regression

In the first Experiment, TD is modeled versus (TT-TR-TC). And as depicted in Table VIII, we can notice that the RMSE values for all KNN regressors used to build the (TD- TC& TR& TT) model is less than all other non-parametric models. Specifically, the least RSME is achieved by the KNN regressor with k =3 which is highlighted in grey in Table VIII. In contrast, it has been noticed that the Decision Tree has the worst performance metrics. For the SVMs, the radial kernel outperforms the linear & the sigmoid kernels.

TABLE VIII. SUMMARY OF THE ACCURACY OF THE (TD- TC& TR& TT) PREDICTION MODEL USING KNN, SVM (GAMMA = 0.01, COST = 10 FOR POLYNOMIAL, RADIAL, SIGMOID KERNEL FUNCTION), AND DECISION TREE (BEST PARAMETERS: MAX DEPTH=3, CP=0.015, AND MINI SPLIT=40)

| | **Learning Parameters** | **RMSE** | **R2** | **Min-Max Accuracy** | **MAPE** |
|---|---|---|---|---|---|
| KNN | k=3 | 25.44 | 0.99 | 0.97 | 0.02 |
| | k=4 | 27.25 | 0.99 | 0.97 | 0.02 |
| | k=5 | 29.89 | 0.99 | 0.97 | 0.02 |
| | k=6 | 36.65 | 0.99 | 0.97 | 0.03 |
| | k=7 | 40.34 | 0.99 | 0.96 | 0.03 |
| | k=8 | 43.94 | 0.99 | 0.96 | 0.03 |
| SVM | Linear Kernel | 85.25 | 0.99 | 0.81 | 0.19 |
| | Polynomial Kernel | 1131.2 | 0.83 | 0.45 | 1.69 |
| | Radial Kernel | 70.44 | 0.99 | 0.84 | 0.15 |
| | Sigmoid kernel | 91.22 | 0.99 | 0.80 | 0.19 |
| DT | Anova Partitioning Method | 232.82 | 0.95 | 0.78 | 0.32 |

TABLE IX. SUMMARY OF THE ACCURACY OF THE (TD- TR& TC) PREDICTION MODEL USING KNN, SVM (GAMMA = 0.01, COST = 10 FOR POLYNOMIAL, RADIAL, SIGMOID KERNEL FUNCTION), AND DECISION TREE (BEST PARAMETERS: MAX DEPTH=3, CP=0.015, AND MINI SPLIT=40)

| | **Learning Parameters** | **RMSE** | **R2** | **Min-Max Accuracy** | **MAPE** |
|---|---|---|---|---|---|
| KNN | k=3 | 46.89 | 0.99 | 0.96 | 0.03 |
| | k=4 | 48.42 | 0.99 | 0.96 | 0.03 |
| | k=5 | 48.04 | 0.99 | 0.96 | 0.03 |
| | k=6 | 45.89 | 0.99 | 0.97 | 0.027 |
| | k=7 | 55.58 | 0.99 | 0.96 | 0.03 |
| | k=8 | 53.15 | 0.99 | 0.97 | 0.03 |
| SVM | Linear Kernel | 117.7 | 0.99 | 0.78 | 0.23 |
| | Polynomial Kernel | 1188 | 0.87 | 0.44 | 1.75 |
| | Radial Kernel | 73.64 | 0.99 | 0.83 | 0.16 |
| | Sigmoid kernel | 125.2 | 0.98 | 0.75 | 0.25 |
| DT | Anova Partitioning Method | 269 | 0.94 | 0.83 | 0.22 |

In the second Experiment 2, TD is modeled versus (TR-TC). Table IX shows that the (TD- TR& TC) model also behaves like the (TD- TC& TR& TT) model in terms of the RMSE values but with larger values. It has been noticed that all KNN regressors has less RMSE values than all other non-parametric models. However, unlike the (TD- TC& TR& TT) model, the least RSME & MAPE and the highest accuracy & R-squared values are achieved by the KNN with k =6 (highlighted in grey in Table IX). Moreover, it has been found that the Decision Tree has the worst performance metrics. For the SVMs, the radial kernel performs better than the linear & the sigmoid kernels.

### C. Selecting the basic Models

In order to select the basic models that will be considered for the future prediction of the total number of confirmed cases & the total number of deaths, we compared the performance metrics for all the models created to the TC & TD variables using the parametric & non-parametric regression methods. The RMSE is selected to be used as the reference for the comparison as the R-squared values are convergent between most models, the Min-Max accuracy behaves consistently with it and the MAPE behaves consistently with the RMSE. The Bar graphs of Figures 5 & 6 are bar charts that show the lowest RMSE values for the parametric & non-parametric regression models built for the TC & the TD models respectively in this study. For the TC models, the RMSE values of only the KNN with k=3 and the Gaussian radial kernel SVM along with the Decision Tree are depicted in Fig. 5. However, for the TD models, the records of the KNN with k=6, radial kernel SVM & the Decision Tree are shown in Fig. 6.

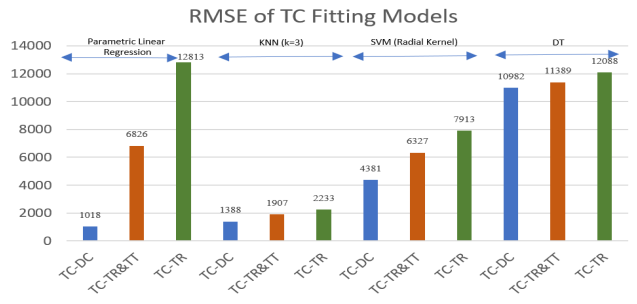



Fig. 5. Bar Chart for the Minimum RMSE retrieved for the TC Fitting Models.

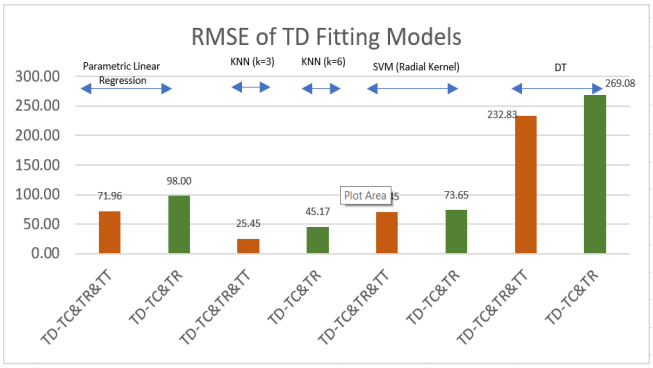



Fig. 6. Bar Chart for the Minimum RMSE retrieved for the TD Fitting Models.

For the TC, it is obvious that the (TC-DC) models have the best performance over all other models when estimated by both the parametric & non-parametric methods. Conversely, the (TC-TR) models are the worst consistently over all methods. Also, it has been observed that adding the TT as a predictor to the (TC-TR) model apparently improves the performance of the model but yet the (TC-TD) model outperforms the (TC-TR&TT) model. In order to select the best (TC-DC) model, we select the modeling method that provides the least RMSE. It has been found that the parametric linear regression model outperforms the KNN, SVM & DT non-parametric regressors. Therefore, it has been decided in this study to consider the linear regression model of the (TC-DC) model as the basic model for tracking the TC growth and for estimating the future values of the TC predictor in the TD model.

For the TD, we can see that adding the TT to the TC& TR reduces the RMSE for all parametric & non-parametric models. Although the reduction in RMSE is slight for almost all regression methods, for the KNN (k=6), the presence of TT in the model reduces the RMSE by nearly 50%. However, we can see that TT has negligible effect for the SVM (Radial) Regressor. It is also noticed that the non-parametric KNN (k=6) performs the best over the other non-parametric models and the parametric linear model followed by the SVM regressor. It is clear also that the linear regression & the SVM performs comparably for the (TD-TC&TR&TT) Nevertheless, it is decided in this study to consider the (TD-TC& TR&TT) build by the Radial Kernel SVM to be used for predicting the future values of the TD instead of the KNN. By finding the future prediction for the unseen data at multiple future dates, we found that all TD predictions have the same values. This could be explained in the light of knowing the nature of the KNN algorithm in associating the unseen data to its neighbors. That is, all upcoming future values appear in the neighborhood of the last training example (Day 116) in the training dataset which always uses this neighborhood to find the future perdition which will give surely the same value for the predictions for all days after Day 116.

### D. Prediction of the Predictor's Future Values

The future predictions of the TD are estimated using the (TD-TC&TR& TT) model. However, the future values of the predictors TC, TR and TT are yet to be predicted against the Day Count. The (TC-DC) model has been previously built and its linear regression model will be used for predicting the future TC value. However, in this part, we model each of the predictors (TT and TR) with respect to the DC using parametric & non-parametric regression methods. Five parametric models have been built using the Linear, Logarithmic, Spline, Polynomial and the Generative Additive Regression [32][33][34]. However, the non-parametric models have been built using the KNN, SVM & DT regression. Afterward, we select the model that has the least RMSE value for the future prediction of the corresponding predictor. Fig. 7 & 8 show the parametric models of the predictors while Fig. 9 & 10 show the non-parametric models. The values of the RMSE corresponding to each model are depicted in Table X. It is clear from this table that the GAM models have the least RMSE over all other models therefore, they have been selected to find the future values of the predictors.

TABLE X. THE VALUES OF THE RMSE & R-SQUARED VALUES FOR THE (TT/TR VERSUS DC) MODELS BUILT USING SEVERAL PARAMETRIC AND NON-PARAMETRIC REGRESSION METHODS. LEAST RMSE VALUES ARE HIGHLIGHTED IN GREY

| Predictor | Method | RMSE | R2 |
|---|---|---|---|
| TT | linear | 275906 | 0.94 |
| | log | 720572.8 | 0.59 |
| | Splines | 15480.06 | 0.99 |
| | polynomial | 31395.33 | 0.99 |
| | GAM | 12382.7 | 0.99 |
| | KNN | 15145.8 | 0.99 |
| | DT | 183297.9 | 0.98 |
| | SVM | 90788.64 | 0.99 |
| TR | linear | 11444.48 | 0.98 |
| | log | 48101.85 | 0.66 |
| | Splines | 2037.958 | 0.99 |
| | polynomial | 2585.802 | 0.99 |
| | GAM | 1247.912 | 0.99 |
| | KNN | 1295.82 | 0.99 |
| | DT | 11798.02 | 0.98 |
| | SVM | 5981.838 | 0.99 |

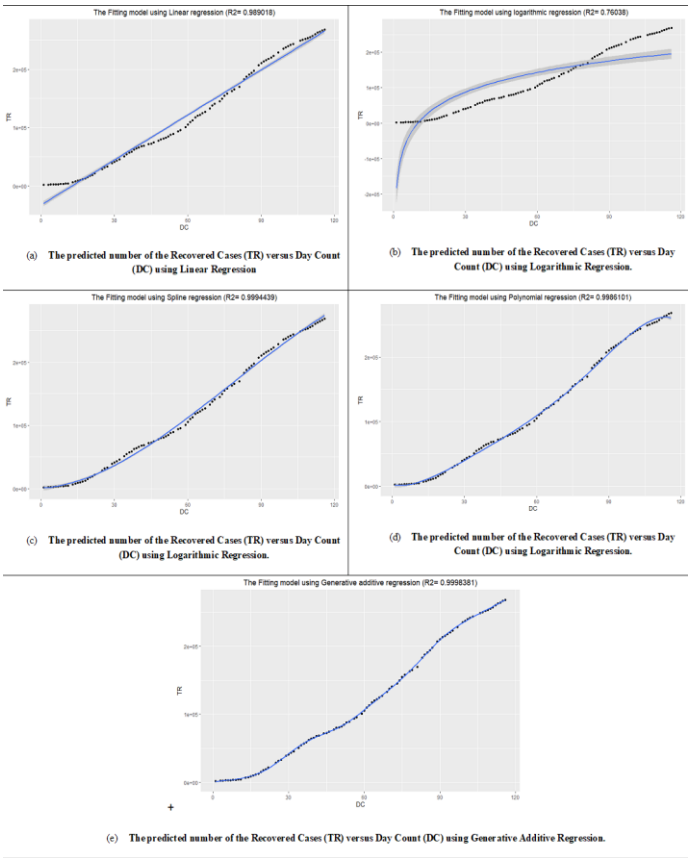



Fig. 7. The Predicted Number of the Recovered Cases (TR) Versus Day Count (DC) using: a) Linear Regression b) Logarithmic Regression c) Spline Regression d) Polynomial Regression e) Generative Additive Regression.

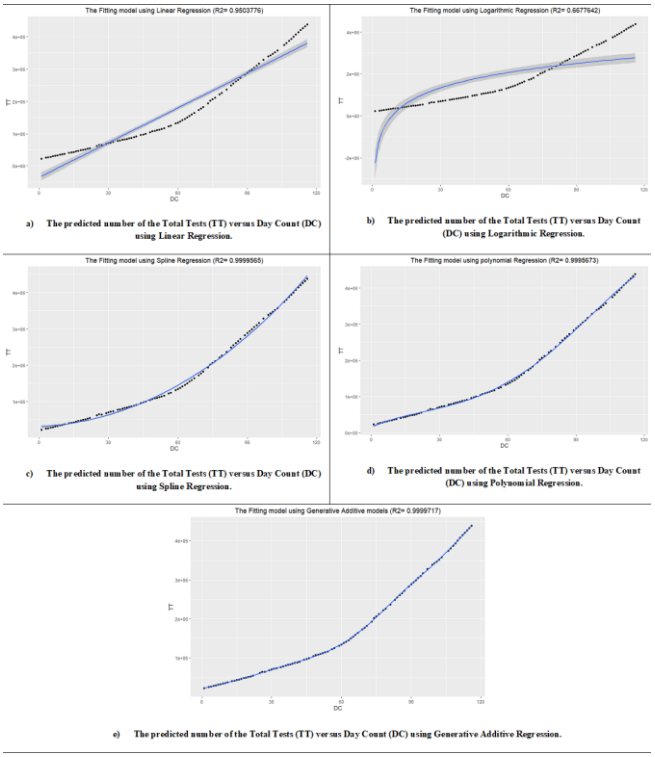



Fig. 8. The Predicted Number of the Total Tests (TT) Versus Day Count (DC) using: a) Linear Regression b) Logarithmic Regression c) Spline Regression d) Polynomial Regression e) Generative Additive Regression.

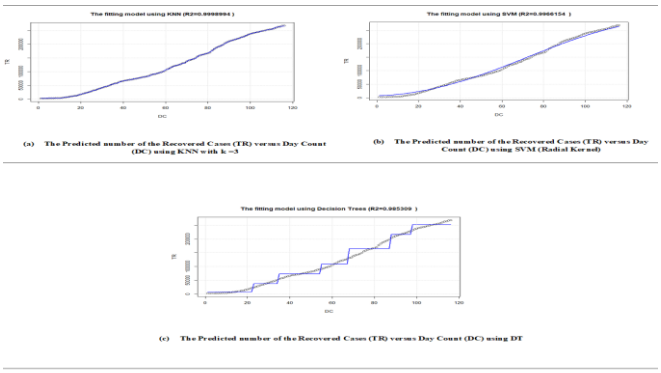



Fig. 9. The Predicted Number of the Recovered Cases (TR) Versus Day Count (DC) using Non-Parametric Regression: a) KNN with k =3 b) SVM (Radial Kernel) c) DT.

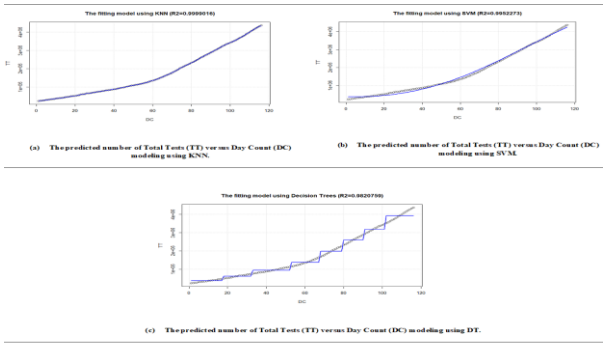



Fig. 10. The Number of Predicted Total Tests (TT) Versus Day Count (DC) using Non-Parametric Regression: a) KNN, K =3 b) SVM c) DT.

## VI. Conclusion

The main objective of this work is to investigate the power of the parametric and non-parametric machine learning methods in the accurate prediction of the spread and mortality of Covid-19 pandemic. Different features in the used Covid-19 dataset have been examined. Very high correlation between the models' response variable and the input predictors is used as the feature selection criterion. The significance of using the number of PCR tests as a model predictor has been investigated. Within the framework of this study, the data is preprocessed, and the most significant predictors are selected to build a number of regression models for the TC & TD separately. The parametric linear regression and the non-parametric KNN, SVM and DT are used for individually modeling the response variables against the selected predictors. The models that show the best prediction performance are considered the basic models to be used for the future prediction of the response variables. The predictors are modeled individually against a time variable using a variety set of parametric & non-parametric methods. The best model is then used to estimate the value of the corresponding predictor at the required future date. The findings show that, for the given dataset, the linear regression performs better than the non-parametric models for predicting TC & TD. It is also found that including of the total number of tests in the mortality model significantly increases its prediction accuracy.

## Acknowledgment and Funding

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

## References

[1] G. Cacciapaglia, C. Cot, & F. Sannino, "Multiwave pandemic dynamics explained: how to tame the next wave of infectious diseases," Sci. Rep., vol. 11, 2021.

[2] W. T Harvey et al. "SARS-CoV-2 variants, spike mutations and immune escape," Nat. Rev. Microbiol., vol. 19, pp. 409–424, 2021.

[3] A. Sheikh, J. McMenamin, B. Taylor, & C. Robertson, "SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness,"Lancet., vol. 397, pp. 2461–2462, 2021.

[4] Worldometer, "Coronavirus disease (COVID-19) outbreak." World Health Organization, Europe.

[5] E. Dong, H. Du & L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," Lancet. Infect. Dis., vol. 20, pp. 533–534, 2020.

[6] M. Wolkewitz & L. Puljak, "Methodological challenges of analysing COVID-19 data during the pandemic," BMC Med. Res. Methodol., vol. 20, pp.1–4, 2020.

[7] G. Shinde et al. "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art," SN Comput. Sci., vol. 14, issue 1, pp. 1–15, 2020.

[8] ak and control in Henan province caused by the output population from Wuhan," medRxiv, 2020. doi:10.1101/2020.05.03.20089193.

[9] N. Sansa, "The Correlation between COVID-19 Confirmed and Recovered Cases in China: Simple Regression Linear Model Evidence," SSRN Electron. J., 2020. doi:10.2139/SSRN.3556549.

[10] D. Alboaneen, B. Pranggono, D. Alshammari, N. Alqahtani, & R. Alyaffer, "Predicting the Epidemiological Outbreak of the Coronavirus Disease 2019 (COVID-19) in Saudi Arabia," Int. J. Environ. Res. Public Health., vol 17, pp. 1–10, 2020.

[11] W. Kermack, & A. McKendrick, "Contributions to the mathematical theory of epidemics," Bull. Math. Biol., vol. 53, pp. 33–55, 1991.

[12] G. Chowell, L. Simonsen, C. Viboud, & Y. Kuang, "Is West Africa Approaching a Catastrophic Phase or is the 2014 Ebola Epidemic Slowing Down? Different Models Yield Different Answers for Liberia," PLoS Curr. 6, 2014.

[13] V. Chaurasia, & S. Pal, "Application of machine learning time series analysis for prediction COVID-19 pandemic," Res. Biomed. Eng., pp. 1–13, 2020.

[14] B. S. Frey, & H. Weck, "Estimating the Shadow Economy: A" Naïve" Approach," Oxf. Econ. Pap., vol. 35, pp. 23–44, 1983.

[15] A. K. Dubey, S. Narang, A. Kumar, S. Sasubilli, & V. García-Díaz, "Performance estimation of machine learning algorithms in the factor analysis of COVID-19 dataset," Comput. Mater. Contin., vol. 66, pp. 1921–1936, 2020.

[16] Kaliappan et al. "Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate," Front. Public Heal., vol. 1319, 2021.

[17] B. Yahaya, L. Muhammad, N. Abdulganiyyu, F. Ishaq, & Y. Atomsa, "An Improved C4.5 Algorithm using Hospital Rule for Large Dataset," Indian J. Sci. Technol., vol. 11, pp. 1–5, 2017.

[18] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Miscellaneous clustering methods," Cluster analysis, pp. 215–255, 2011.

[19] M. Islam, H. Iqbal, M. Haque, & M. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," Proc. 5th IEEE Reg. 10 Humanit. Technol. Conf. , pp. 226–229, 2018.

[20] L. Muhammad, M. Islam, S. Usman, & S. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," SN Comput. Sci., vol. 1, 2020.

[21] H. Mahmoud, Parametric versus Semi and Nonparametric Regression Models. Int. J. Stat. Probab., vol. 10, 2019.

[22] Z. Yao, & W. Ruzzo, "A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data," BMC Bioinforma., vol. 71, issue 7, pp. 1–11, 2006.

[23] N. Ali, D. Neagu, & P. "Trundle. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," SN Appl. Sci., vol. 112, issue 1, pp. 1–15, 2019.

[24] N. Parveen, S. Zaidi & M. Danish, "Support vector regression model for predicting the sorption capacity of lead (II)," Perspect. Sci., vol. 8, pp. 629–631, 2016.

[25] T. Hofmann, B. Schölkopf, & A. "Smola, Kernel methods in machine learning,", vol. 36, pp. 1171–1220, 2008.

[26] S. Uddin, A. Khan, M. Hossain, & M. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Med. Informatics Decis. Mak., vol. 191, issue 19, pp. 1–16, 2019.

[27] L. Breslow, & A. Leonard, W. David, "Simplifying decision trees: A survey," Knowl. Eng. Rev. , vol. 12, pp. 1–40, 1997.

[28] S. Boslaugh, "The Pearson Correlation Coefficient," in Statistics in a Nutshell, 2nd Edition, O'Reilly Media, Inc., pp. 80–92, 2012.

[29] J. Fan, "Nonparametric Models," in Nonlinear Time Series, New York: Springer, pp. 313–403, 2008.

[30] J. Fan, "Nonparametric Density Estimation," in Nonlinear Time Series, Springer New York, pp. 193–214, 2008. doi: 10.1007/978-0-387-69395-8_5.

[31] J. Fan, Nonlinear Time Series - Nonparametric and Parametric Methods. Springer New York.

[32] A. Gonçalves, E. Orton, J. Boon, & M. Salman, "Linear, logarithmic, and polynomial models of M-mode echocardiographic measurements in dogs," Am. J. Vet. Res, vol. 63, pp. 994–999, 2002.

[33] B. Wang, W. Shi, & Z. Miao, Comparative "Analysis for Robust Penalized Spline Smoothing Methods," Math. Probl. Eng., 2014.

[34] K. Ravindra, P. Rattan, S. Mor, & A. Aggarwal, "Generalized additive models: Building evidence of air pollution, climate change and human health," Environ. Int., vol. 132, 2019.