

Classifier Selection for the Prediction of Dominant Transmission Mode of Coronavirus Within Localities: Predicting COVID-19 Transmission Mode

Donald Douglas Atsa'am, University of the Free State, South Africa

Ruth Wario, University of the Free State, South Africa

 <https://orcid.org/0000-0003-3733-3485>

ABSTRACT

The coronavirus disease-2019 (COVID-19) pandemic is an ongoing concern that requires research in all disciplines to tame its spread. Nine classification algorithms were selected for evaluating the most appropriate in predicting the prevalent COVID-19 transmission mode in a geographic area. These include multinomial logistic regression, k-nearest neighbour, support vector machines, linear discriminant analysis, naïve Bayes, C5.0, bagged classification and regression trees, random forest, and stochastic gradient boosting. Five COVID-19 datasets were employed for classification. Predictive accuracy was determined using 10-fold cross validation with three repeats. The Friedman's test was conducted, and the outcome showed the performance of each algorithm is significantly different. The stochastic gradient boosting yielded the highest predictive accuracy, 81%. This finding should be valuable to health informaticians, health analysts, and others regarding which machine learning tool to adopt in the efforts to detect dominant transmission mode of the virus within localities.

KEYWORDS

Classification, COVID-19, Data Mining, Mining Methods and Algorithms, Transmission Mode

INTRODUCTION

In December, 2019, the first coronavirus disease 2019 (COVID-19) case was detected in the city of Wuhan, China (Cortegiani, Ingoglia, Ippolito, Giarratano, & Einav, 2020; Khan & Atangana, 2020; Shereen, Khan, Kazmi, Bashir, & Siddique, 2020). The COVID-19 is a highly contagious infection caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) (Shereen et al., 2020; Rothan & Byrareddy, 2020), which symptoms include fever, coughing and problems with breathing. According to the report by the World Health Organization (WHO), as of 29th June, 2020, a total of 10,021,401 COVID-19 cases were confirmed across the world (WHO, 2020). Out of the total

DOI: 10.4018/IJEHMC.20211101.oa1

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

confirmed cases, a total of 499,913 deaths were recorded from the viral infection from inception of the pandemic to 29th June, 2020.

Literature evidence suggests that artificial intelligence (AI) is a valuable tool in the efforts to contain the pandemic. Vaishya, Javid, Khan, and Haleem (2020) identified several ways in which AI can be applied in the fight against COVID-19. According to Vaishya et al (2020), AI has capabilities to detect early infection of the virus in an individual, monitoring of COVID-19 patient's condition, and contact tracing of exposed persons. Mei et al. (2020) deployed AI to develop a rapid system for diagnosing COVID-19 patients. The diagnostic system uses AI algorithms to integrate chest computed tomography (CT) findings with exposure history, laboratory testing and clinical symptoms to diagnose COVID-19 positive persons. Among other merits of the AI-enabled diagnostic system such as enhanced accuracy, the system is reported to produce test results faster than the regular virus-specific reverse transcriptase polymerase chain reaction test, which takes two days to complete (Mei et al, 2020). In another study, Muhammad, Islam, Usman, and Ayon (2020) investigated the best classification algorithm for use in constructing a model for predicting the possibility of recovery from COVID-19 infection by patients. The experiments were conducted using COVID-19 data in South Korea and the predictive accuracies of support vector machines, decision tree, logistic regression, naïve Bayes, random forest, and k-nearest neighbor models were noted. The study concluded that the decision tree, which achieved 99.85% predictive accuracy, is the best classification algorithm for use in the prediction of whether or not a patient is going to recover from the infection (Muhammad et al., 2020). The research by Tuli Shreshth, Tuli Shikhar, Tuli Rakesh, and Gill (2020) integrated machine learning with cloud computing to develop a model that can predict the growth and trend of COVID-19 across countries of the world. The study used the iterative weighting in fitting a generalized inverse Weibull distribution, which served the purpose of developing a predictive model. The prediction framework, which can be deployed on a cloud computing platform, is capable to perform real-time prediction of the growth pattern and trend of the epidemic across the world (Tuli Shreshth et al., 2020).

One of the effective ways of controlling the spread of coronavirus is to determine the prevalent transmission mode from person-to-person within a geographic area. Knowledge of the dominant transmission mode specific to an area is vital so that experts can advise on strategies to be put in place to tame the spread. Against this backdrop, this study was aimed at determining the most appropriate classification algorithm to be adopted for predicting the prevalent transmission mode of coronavirus in a given area. Several classification algorithms are in existence, but not all of these can be deployed in every problem domain. The question of which machine learning algorithm will produce optimum results has to be determined experimentally through spot-checking. In the end of this study, the classification algorithm that offers the highest performance in predicting the dominant COVID-19 transmission mode within an area is advised.

CLASSIFICATION ALGORITHMS RELEVANT TO THIS STUDY

Class prediction is a technique of data mining concerned with development of models, decision trees and IF-THEN rules to predict class labels of observations (Alola & Atsa'am, 2020; Atsa'am, 2020; Bodur & Atsa'am, 2019; Kantardzic, 2011). Let $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$ be a dataset containing m observations, where $x = (x_1, \dots, x_n)$ are independent variables of dimension $x \in R^n$; and $y \in C$ where C is the outcome variable. Classification or class prediction is the mapping $t : R^n \rightarrow C$ where t is a predictive model (Bodur & Atsa'am, 2019; Genuer, Pogy, & Tuleau-Malot, 2010). In situations where C consists of only two categories, the prediction problem is termed binary class prediction; and where C has more than two categories, it is termed multinomial class prediction (Kleinbaum & Klein, 2010). The classification algorithms relevant to this study are described in this section.

Multinomial Logistic Regression

This variant of the logistic regression models the relationship between a dependent, multi-level, nominal variable and independent variable(s) (Anderson & Rutkowski, 2008). When fitting a multinomial logistic regression model on a dataset with n independent variables and one categorical outcome variable with k categories ($j = 1, 2, \dots, k$), one of the categories is usually set aside as the reference or baseline category (Anderson & Rutkowski, 2008; El-Habil, 2012). It should be pointed out that all logits are computed relative to the baseline category. Let the k^{th} category denote the baseline category, and let P_j denote the multinomial probability that an observation belongs to the j^{th} category. Then, the multinomial logistic regression is given as in Equation (1) (El-Habil, 2012).

$$\ln \left(\frac{P_j(x_i)}{P_k(x_i)} \right) = b_{0j} + b_{1j}X_1 + b_{2j}X_2 + \dots + b_{nj}X_n \quad (1)$$

where, $j = 1, 2, \dots, k-1$; b_{0j} is the intercept of the j^{th} category, $b_{1j}, b_{2j}, \dots, b_{nj}$ are parameter estimates for each j^{th} category, X_1, X_2, \dots, X_n are data values for each variable in the dataset. The Equation (1) effectively models the natural logarithm of the probability of an observation belonging to a particular category versus the probability of belonging to the baseline category. All the probabilities in Equation (1) sum to one, and the equation can be written in terms of probabilities as shown in the Equation (2) (El-Habil, 2012).

$$P_j(x_i) = \frac{\exp(b_{0j} + b_{1j}X_1 + b_{2j}X_2 + \dots + b_{nj}X_n)}{1 + \sum_{j=1}^{k-1} \exp(b_{0j} + b_{1j}X_1 + b_{2j}X_2 + \dots + b_{nj}X_n)} \quad (2)$$

When predicting the class of an observation, all probabilities are computed and the j^{th} category with the maximum probability is predicted as the class of the observation.

K-Nearest Neighbour (KNN)

In order to predict the class of a vector, the KNN searches the entire training set for K instances that are most similar to the vector being classified, and then classifies it to the category with the highest probability (Harrison, 2018; Kim, Choi, Moon, & Mun, 2011). Consider a training set with N samples and n categories, c_1, c_2, \dots, c_n . Let d_i represent a neighbor in the training set, and X be a vector to be classified. Then, $y(d_i, c_k)$ shows whether d_i is of class c_k , and $S(X, d_i)$ is a function that evaluates the similarity between X and d_i . The probability that the vector X belongs to class c_k , denoted by $P(X, c_k)$, is given in Equation (3) (Kim et al., 2011).

$$P(X, c_k) = \sum_{d_i \in KNN} S(X, d_i) \cdot y(d_i, c_k) \quad (3)$$

Depending on the type of data, the similarity, $S(X, d_i)$, can be evaluated using distance measures such as Euclidean, Manhattan, and Minkowski.

Support Vector Machines (SVM)

The SVM is a classification algorithm that uses a hyperplane to separate input data points according to their class (Qi, Silvestrov, & Nazir, 2017). The task of the learning algorithm is to find coefficients

that yield the best separation of the classes by the hyperplane. The objective is that the hyperplane should be at the maximum margin from the various classes. This makes the classifier stable and unaffected by noise within the data (Harrington, 2015). One of the implementations that extend the SVM from just a binary classifier to a multi-class classifier is the one-against-all scheme (Harrington, 2015). In this approach, an SVM model is built with one class while all data objects within other classes are grouped into a single opposing class. In the prediction problem, the model that produces the largest output represents the predicted class of the input vector.

Linear Discriminant Analysis (LDA)

With this classification algorithm, the data is assumed to follow multivariate normal distribution. To predict the class, i , of an observation, X ; the vector mean, μ_i , and prior probability, $P(i)$, for each class, and the common covariance matrix, C , are computed from the data. Then, the formula in Equation (4) is evaluated to predict the class to which the observation belongs (Teknomo, 2019).

$$f_i(X) = \mu_i C^{-1} X^T - \frac{1}{2} \mu_i C^{-1} \mu_i^T + \ln(P(i)) \quad (4)$$

From the Equation (4), the conditional probability of X belonging to class i , denoted by $f_i(X)$, is computed and the vector is assigned to class i if $f_i(X) > f_j(X) \forall i \neq j$.

Naïve Bayes

The naive Bayes classification is built on the Bayes theorem. The theorem assumes that the set of x_1, x_2, \dots, x_n variables are independent, given the class, y . With this assumption, the classifier is given as in Equation (5) (Bishop, 2006; Duda, Hart, & Stork, 2003).

$$P(X | y) = P(x_1 | y) \times P(x_2 | y) \times \dots \times P(x_n | y) \quad (5)$$

In the classification problem, the probability of an observation belonging to each class, y , is computed as shown in Equation (6); and the class that maximizes this probability is predicted as the class of the observation (Bishop, 2006; Duda et al., 2003).

$$y = \arg \max_y [P(y) * \prod_{i=1}^n P(x_i | y)] \quad (6)$$

C5.0 Algorithm

The C5.0 is one of the classification algorithms that implement the decision tree. At the initial step, the algorithm uses entropy to evaluate the mix of class values within the dataset (Han & Kamber, 2006; Yobero, 2018). Entropy is computed as; $Entropy(S) = \sum_{i=1}^C -\pi_i \log_2(\pi_i)$ where S is a given data segment, C is the number of class levels, and π_i is the proportion of values within the class level i (Yobero, 2018). The algorithm then uses the purity obtained from entropy to compute information gain for each feature. Information gain for a feature, F , is computed as $InformationGain(F) = Entropy(S_1) - Entropy(S_2)$, which is the difference between entropy before the split (S_1) and the entropy in the partitions after the split (S_2). The $Entropy(S_2)$ is computed by evaluating the total entropy in all the partitions. This is achieved by weighing the entropy in each

partition by the proportion of observations within that partition, which can be obtained as shown in Equation (7) (Han & Kamber, 2006; Yobero, 2018).

$$Entropy(S_2) = \sum_{i=1}^n w_i Entropy(\pi_i) \quad (7)$$

A feature that produced the highest information gain is selected for splitting. The C5.0 algorithm has been designed to automatically decide when to prune the decision tree and provide generalization for unobserved data.

Bagged Classification and Regression Tree (Bagged CART)

Bagging is a bootstrapping approach that constructs several decision trees by iteratively selecting random subsets of the original training data (Lawrence, Bunn, Powell, & Zambon, 2004). The Bagged CART is a specific implementation of bagging where multiple samples are taken from the training data and decision trees are fitted on each sample. During classification of a new observation, each decision tree performs a prediction and the averages of the predictions are obtained to determine the class of the observation (Lawrence et al., 2004; Le, 2020). It is on record that in most cases, bagging algorithms produce higher classification accuracies than single decision tree algorithms (Optiz & Maclin, 1999).

Random Forest

Random forest is an ensemble of several independent decision tree predictors $\{h(X, v_k), k = 1, \dots, n\}$, where X is an input vector and $\{v_k\}$ are independent random vectors within the same distribution across all trees in the forest (Breiman, 2001; Genuer et al., 2010). In the task of predicting the class of an input vector X , each tree in the forest casts a single vote and the class with highest number of votes is predicted as the class to which X belongs (Breiman, 2001).

Stochastic Gradient Boosting (SGB)

Boosting is an ensemble method that constructs a strong classification model from a number of weak models (Friedman, 2001; Friedman, 2002; Le, 2020). The stochastic gradient boosting, also called generalized boosting modeling, is an implementation of boosting that builds a tree from the training data, then creates another tree that corrects the errors observed from the previous model. At every iteration, the algorithm randomly fetches a subsample from the training dataset and uses the same to construct a tree classifier. The algorithm keeps adding models until predictive accuracy is maximized on the training set or optimum number of models has been added (Friedman, 2002; Lawrence et al., 2004). Unlike single decision tree classifiers, the SGB is not prone to overfitting.

METHODOLOGY

Experimental Datasets

Following the outbreak of COVID-19, the WHO set up a global surveillance system to gather essential information to monitor the pandemic. On a daily basis, national authorities of member-states send in count of confirmed cases and deaths, in addition to other set of information vital to monitoring the progression or otherwise of the pandemic. The WHO collects these data from the national authorities and then makes the same available on the internet in form of COVID-19 situation reports. For the purpose of this study, five datasets on COVID-19 situation reports on five different days as reported

by WHO were randomly selected and employed for experiments. The datasets, described below, are freely available on the WHO website (WHO, 2020).

- **COVID-19_May4:** This dataset consists of COVID-19 cases across 214 countries, territories and areas as of 4th May, 2020. The WHO presents the data as coronavirus situation report – 105.
- **COVID-19_May20:** This dataset consists of COVID-19 cases across 215 countries, territories and areas as of 20th May, 2020. The WHO presents the data as coronavirus situation report – 121.
- **COVID-19_May31:** This dataset consists of COVID-19 cases across 214 countries, territories and areas as of 31st May, 2020. The WHO presents the data as coronavirus situation report – 132.
- **COVID-19_June1:** This dataset consists of COVID-19 cases across 214 countries, territories and areas as of 1st June, 2020. The WHO presents the data as coronavirus situation report – 133.
- **COVID-19_June2:** This dataset consists of COVID-19 cases across 213 countries, territories and areas as of 2nd June, 2020. The WHO presents the data as coronavirus situation report – 134.

Each datasets is composed of seven variables as follows:

Reporting Country/Territory/Area: This holds the name of the country, territory or area whose COVID-19 situation data is being reported.

Total confirmed cases: This reports the total number of confirmed cases in a particular country, territory or area from inception of the pandemic. According to WHO, a confirmed case is when laboratory test result shows that someone is infected with the coronavirus even when they do not show the symptoms.

Total confirmed new cases: This variable records the total number of confirmed new cases on the reporting date in a country, territory or area.

Total deaths: This variable keeps track of the total number of deaths caused by the coronavirus disease in a particular country, territory or area from inception of the pandemic.

Total new deaths: This records the total number of new deaths from COVID-19 on the reporting date in a particular country, territory or area.

Days since last reported case: This holds the total number of days between the last confirmed COVID-19 case in a particular country, territory or area and the reporting date when another case is confirmed.

Transmission classification: This consists of four categories that indicate the prevalent COVID-19 transmission mode from person-to-person in a particular country, territory or area. The categories include; sporadic cases, clusters of cases, community transmission, and pending. The sporadic cases transmission category indicates where one or more cases are imported or detected locally. The clusters of cases transmission category is where several cases take place which are clustered in time, geographic location and or by common exposures. The community transmission category includes larger outbreaks of local transmission resulting from multiple unrelated clusters in several areas of the country, territory or area. The pending category, also referred to as unknown, indicates when the transmission mode has not been reported to WHO. Where multiple modes of transmission have been reported, the mode with the highest cases is adopted by WHO.

The Table 1 shows the distribution of records by transmission category in the various COVID-19 experimental datasets.

Analytical Approach

The analytical approach followed in this study to determine the best machine learning algorithm for use in predicting the prevalent COVID-19 transmission mode is presented in this section.

Table 1. number of records in datasets per transmission class

| Dataset | #Sporadic | #Clusters | #Community | #Pending | Total |
|----------------|-----------|-----------|------------|----------|-------|
| COVID19_May4 | 50 | 91 | 52 | 21 | 214 |
| COVID19_May20 | 38 | 87 | 74 | 16 | 215 |
| COVID19_May31 | 36 | 83 | 83 | 12 | 214 |
| COVID-19_June1 | 36 | 82 | 84 | 12 | 214 |
| COVID-19_June2 | 35 | 82 | 84 | 12 | 213 |

Data Preparation: The first step towards preparing the datasets for experiments was attribute selection. The variable that holds the name of the geographic entity where COVID-19 cases are reported has no relevance in classification modeling. Therefore, the variable was eliminated from all the experimental datasets.

Selection of Classifiers for Spot-Checking: Nine classification algorithms were selected for spot-checking on the five COVID-19 datasets. These include; the multinomial logistic regression (Log), k-nearest neighbour (KNN), support vector machines (SVM), linear discriminant analysis (LDA), naïve Bayes (NB), C5.0 (C5O), bagged classification and regression trees (CART), random forest (RF), and stochastic gradient boosting (SGB). The choice of these algorithms was deliberate such that the major classification methods including linear, non-linear, trees and rules, and ensemble of trees, have been represented.

Building and Evaluating the Models: As noted, the task is to determine which machine learning algorithm provides the highest accuracy in predicting the prevalent transmission mode of coronavirus within a geographic area. This is a multi-class prediction problem as there are four categories of transmission mode (sporadic cases, clusters of cases, community transmission, and pending). For each of the five COVID-19 datasets, the following procedures were followed:

1. The prepared dataset was loaded into the R programming language environment
2. The “train” function in the Caret package (Kuhn et al., 2018) was invoked to fit a classification model using each of the nine machine learning algorithms.
3. The repeated k-fold cross validation method (Atsa’am, 2020) was used as the technique for testing the predictive accuracy of the classifiers. Specifically, the 10-fold cross validation with three repeats was used.
4. The predictive accuracy produced by each classifier was noted.

The results of the various classification accuracies produced by each classifier on the five datasets are shown in Table 2.

Table 2. Predictive Accuracies of Classification Algorithms on Five Covid-19 Datasets

| Dataset | Accuracy | | | | | | | | |
|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | Log | KNN | SVM | LDA | NB | C5O | CART | RF | SGB |
| COVID-19_May4 | 0.73 | 0.60 | 0.50 | 0.55 | 0.54 | 0.63 | 0.62 | 0.63 | 0.84 |
| COVID-19_May20 | 0.70 | 0.56 | 0.40 | 0.52 | 0.34 | 0.54 | 0.54 | 0.59 | 0.78 |
| COVID-19_May31 | 0.71 | 0.53 | 0.51 | 0.49 | 0.49 | 0.60 | 0.55 | 0.55 | 0.81 |
| COVID-19_June1 | 0.73 | 0.52 | 0.49 | 0.48 | 0.48 | 0.56 | 0.54 | 0.56 | 0.80 |
| COVID-19_June2 | 0.72 | 0.53 | 0.51 | 0.49 | 0.46 | 0.58 | 0.56 | 0.55 | 0.82 |
| Average | 0.72 | 0.55 | 0.48 | 0.51 | 0.46 | 0.58 | 0.56 | 0.58 | 0.81 |

The repeated cross validation is a robust technique of testing predictive accuracy of models when there is class imbalance in the datasets, such as the datasets used in this study. As could be observed in Table 1, the records distributions in the datasets per COVID-19 transmission mode category are not the same across categories. To determine classification accuracy with 10-fold cross validation with three repeats, it means that in each case, the dataset was randomly split into 10 groups and the classifier was trained on nine groups, while one group was set aside for testing the predictive accuracy of the model. At the end of the procedure, each observation must have been part of the training set nine times, and part of the test set at least once. The procedure was repeated three times before the average value was obtained to represent the predictive accuracy of a classifier.

It should be pointed out that some classification algorithms such as LDA, SVM, and KNN perform better on normalized data. Consequently, the preprocess function in the Caret package was included in the arguments of these classifiers to transform the data to uniform scales before modeling.

RESULTS AND DISCUSSION

The results in Table 2 show that the stochastic gradient boosting algorithm performed better than other algorithms, with average accuracy of 81%. Furthermore, the Naïve Bayes produced the least accuracy compared to other algorithms, with average accuracy of 46%.

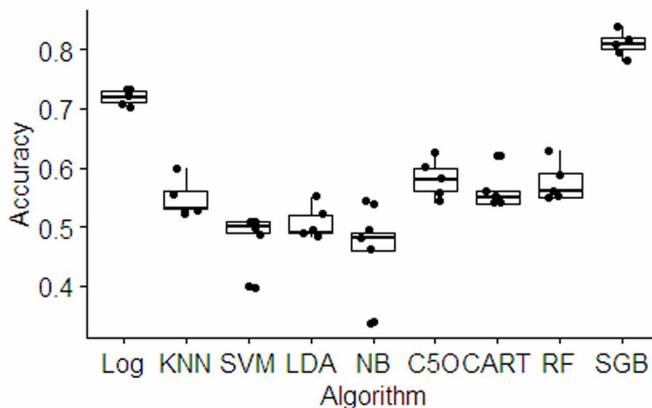
The performances of the nine algorithms on five different COVID-19 datasets are represented by the boxplot in Figure 1.

In Figure 1, the predictive accuracies of each algorithm on the five datasets are represented in a box. These show how accurate each algorithm can predict the prevalent transmission mode of coronavirus within an area.

The Friedman test was conducted on the predictive accuracies in Table 2 to determine if there is a statistically significant difference in the performances of the nine algorithms. The following result was obtained; $Friedman\ chi\text{-squared} = 35.1, df = 8, p\text{-value} = 0.00$, thus confirming that the performances of the algorithms are different generally. The Friedman test is a nonparametric statistic that assesses whether the differences in the distributions of paired groups are statistically significant (Riffenburgh, 2012).

To confirm further whether the predictive accuracy performances of the nine algorithms on the five COVID-19 datasets are different, the effect size of the Friedman test was evaluated on Table 2, using the Kendall's W (Tomczak M. & Tomczak E., 2014). Irrespective of whether or not the differences in the distributions among groups are statistically significant, the goal of effect size

Figure 1. Boxplot of algorithms performance on COVID-19 datasets



evaluation is to test if the differences between the groups' averages are meaningfully large (Lakens, 2013; Tomczak M. & Tomczak E., 2014). The Kendall's W effect size yielded 0.88, indicating a large effect size among the average predictive accuracies of the classification algorithms. It should be pointed out that Kendall adopts the Cohen's guidelines for interpreting effect sizes as follows; 0.1 to < 0.3 (small effect), 0.3 to < 0.5 (medium effect), 0.50 and above (large effect) (Tomczak M. & Tomczak E., 2014).

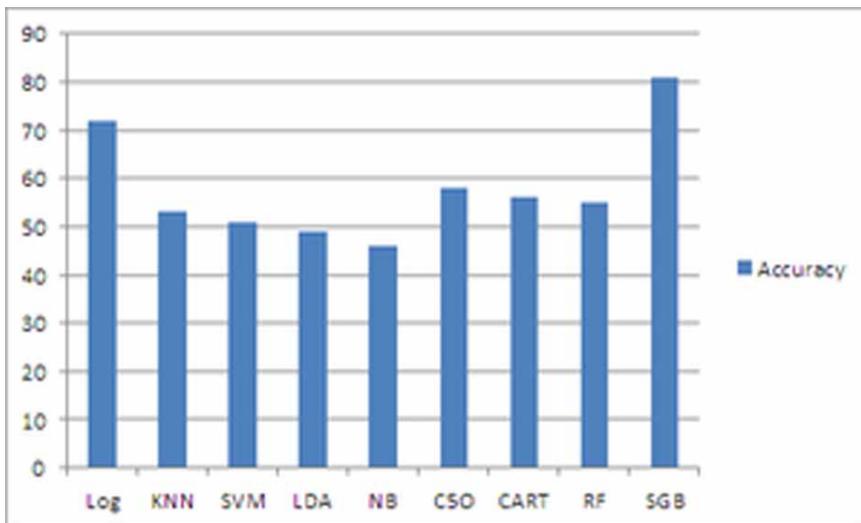
The outcome of the Friedman test indicated that the predictive accuracies of the nine classification algorithms are different on a general note. However, a follow-up pairwise test was conducted to determine which pairs of algorithms specifically produced different predictive accuracies. The post-hoc Conover's pairwise comparisons test (Conover, 1999) was conducted and the results in Table 3 were produced.

The results in Table 3 show that the differences between the predictive accuracies produced by LDA and SVM, NB and LDA are not statistically significant. The overall performance of each algorithm on the five COVID-19 datasets in terms of the average predictive accuracy is shown in Figure 2.

Table 3. Pairwise Comparisons of Predictive Accuracies using Conover's tests

| | Log | KNN | SVM | LDA | NB | C5O | CART | RF |
|------|-------|-------|--------------|--------------|-------|-------|-------|-------|
| KNN | 0.000 | - | - | - | - | - | - | - |
| SVM | 0.000 | 0.000 | - | - | - | - | - | - |
| LDA | 0.000 | 0.000 | 0.672 | - | - | - | - | - |
| NB | 0.000 | 0.000 | 0.019 | 0.058 | - | - | - | - |
| C5O | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | - | - | - |
| CART | 0.003 | 0.011 | 0.000 | 0.000 | 0.000 | 0.002 | - | - |
| RF | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.003 | - |
| SGB | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Figure 2. Average predictive accuracies per classifier



The mean accuracies show that the stochastic gradient boosting algorithm offers better prediction of the prevalent transmission mode of the coronavirus within an area.

CONCLUSION

It cannot be decided in advance which classifier to use for a prediction problem until experiments are conducted through spot-checks. In the present study, nine classification algorithms that support multi-class prediction were selected for evaluating which is the most appropriate for predicting the prevalent coronavirus transmission mode in a geographic area. The algorithms deployed for spot-checking include; the multinomial logistic regression, k-nearest neighbour, support vector machines, linear discriminant analysis, naïve Bayes, C5.0, bagged classification and regression trees, random forest, and stochastic gradient boosting. These algorithms cut across linear models, non-linear models, trees and rules, and ensemble of trees. Five COVID-19 datasets were employed and classifiers were fitted using these algorithms. The average predictive accuracies indicated that the stochastic gradient boosting algorithm is the best model for predicting the transmission mode of coronavirus (81% accuracy). Accuracy was determined using the 10-fold cross validation with three repeats. Though the datasets used for experiments were obtained from the WHO website, where data is reported at the national level of countries, the outcome is equally applicable to data obtained at the level of villages, towns, cities, states or regions. The outcome of this study is a valuable contribution in the efforts of using data mining and machine learning in the quest to tame the COVID-19 pandemic. The research finding has resolved the question of which machine learning algorithm to deploy in the task of predicting the prevalent coronavirus transmission mode within a specific geographic area. This knowledge should be valuable to health informaticians, epidemiologists, health analysts and other stakeholders. There are numerous classification algorithms in existence, but this study deployed only nine. Further research should deploy more classifiers for spot-checking. In future research direction, the possibility of developing a mobile or computer application that implements the stochastic gradient boosting algorithm for predicting the dominant transmission mode of the coronavirus from person-to-person should be explored.

REFERENCES

- Alola, U. V., & Atsa'am, D. D. (2020). Measuring employees' psychological capital using data mining approach. *Journal of Public Affairs*, 20(2), e2050. doi:10.1002/pa.2050
- Anderson, C. J., & Rutkowski, L. (2008). Multinomial logistic regression. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 390–409). SAGE Publishing. doi:10.4135/9781412995627.d31
- Atsa'am, D. D. (2020). Feature selection algorithm using relative odds for data mining classification. In A. Haldorai & A. Ramu (Eds.), *Big data analytics for sustainable computing* (pp. 81–106). IGI Global. doi:10.4018/978-1-5225-9750-6.ch005
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
- Bodur, E. K., & Atsa'am, D. D. (2019). Filter variable selection algorithm using risk ratios for dimensionality reduction of healthcare data for classification. *Processes (Basel, Switzerland)*, 7(4), 222. doi:10.3390/pr7040222
- Breiman, L. (2001). Random forests: Machine learning. *Scientific Research*, 45(1), 5–32. doi:10.1023/A:1010933404324
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). Wiley.
- Cortegiani, A., Ingoglia, G., Ippolito, M., Giarratano, A., & Einav, S. (2020). A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19. *Journal of Critical Care*, 57, 279–283. Advance online publication. doi:10.1016/j.jcrc.2020.03.005 PMID:32173110
- Duda, R. O., Hart, P. E., & Stork, D. G. (2003). *Pattern recognition* (2nd ed.). Wiley.
- El-Habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan Journal of Statistics and Operations Research*, 8(2), 271–291. doi:10.18187/pjsor.v8i2.234
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. doi:10.1214/aos/1013203451
- Friedman, J. (2002). Stochastic gradient boosting: Nonlinear methods and data mining. *Computational Statistics & Data Analysis*, 38(4), 367–378. doi:10.1016/S0167-9473(01)00065-2
- Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236. doi:10.1016/j.patrec.2010.03.014
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Elsevier.
- Harrington, P. B. (2015). Support vector machine classification trees. *Analytical Chemistry*, 87(21), 11065–11071. doi:10.1021/acs.analchem.5b03113 PMID:26461495
- Harrison, O. (2018). *Machine learning basics with the k-nearest neighbors algorithm. Towards data science*. <https://link.medium.com/lbtqwRhYE7>
- Kantardzic, M. (2011). *Data mining concepts, models, methods, and algorithms* (2nd ed.). Wiley & Sons. doi:10.1002/9781118029145
- Khan, M. A., & Atangana, A. (2020). Modeling the dynamics of novel coronavirus (2019-nCov) with fractional derivative. *Alexandria Engineering Journal*, 59(4), 2379–2389. Advance online publication. doi:10.1016/j.aej.2020.02.033
- Kim, K. S., Choi, H. H., Moon, C. S., & Mun, C. W. (2011). Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current Applied Physics*, 11(3), 740–745. doi:10.1016/j.cap.2010.11.051
- Kleinbaum, D. G., & Klein, M. (2010). Ordinal logistic regression. In *Logistic regression: Statistics for biology and health* (pp. 463–488). Springer. doi:10.1007/978-1-4419-1742-3_13
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., & Kenkel, B. (2018). Caret: Classification and regression training. R package version 6.0-77. CRAN. <https://CRAN.R-project.org/package=caret>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. doi:10.3389/fpsyg.2013.00863 PMID:24324449

Lawrence, R., Bunn, A., Powell, S., & Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90(3), 331–336. doi:10.1016/j.rse.2004.01.007

Le, J. (2020). *A tour of the 10 top algorithms for machine learning newbies*. Built In. <https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies>

Mei, X., Lee, H., Diao, K., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P. M., Chung, M., Bernheim, A., Mani, V., Calcagno, C., Li, K., Li, S., Shan, H., Lv, J., Zhao, T., Xia, J., & Yang, Y. et al. (2020). Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nature Medicine*, 26(8), 1224–1228. Advance online publication. doi:10.1038/s41591-020-0931-3 PMID:32427924

Muhammad, L. J., Islam, M. M., Usman, S. S., & Ayon, S. I. (2020). Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. *SN Computer Science*, 1(4), 206. doi:10.1007/s42979-020-00216-w PMID:33063049

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, 169–198. doi:10.1613/jair.614

Qi, X., Silvestrov, S., & Nazir, T. (2017). Data classification with support vector machine and generalized support vector machine. *Proceedings of the AIP Conference*, 1798. doi:10.1063/1.4972718

Riffenburgh, R. H. (2012). Tests on ranked data. In R. H. Riffenburgh (Ed.), *Statistics in medicine* (3rd ed., pp. 221–248). Elsevier. doi:10.1016/B978-0-12-384864-2.00011-1

Rothan, H. A., & Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *Journal of Autoimmunity*, 109, 102433. doi:10.1016/j.jaut.2020.102433 PMID:32113704

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*, 24, 91–98. doi:10.1016/j.jare.2020.03.005 PMID:32257431

Teknomo, K. (2019). *Derivation of linear discriminant analysis formula*. Revoledu Design. <https://people.revoledu.com/kardi/tutorial/LDA/LDA%20Formula.htm>

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1, 19–25.

Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, 11, 100222. doi:10.1016/j.iot.2020.100222

Vaishya, R., Javid, M., Khan, I. H., & Haleem, A. (2020). Artificial intelligence (AI) application for COVID-19 pandemic. *Diabetes & Metabolic Syndrome*, 14(4), 337–339. doi:10.1016/j.dsx.2020.04.012 PMID:32305024

World Health Organization. (2020). *Data from: Coronavirus disease (COVID-19) situation reports* [Dataset]. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>

Yobero, C. (2018). *Determining creditworthiness for loan applications using C5.0 decision trees*. RPubs. <https://rpubs.com/cyobero/C50>

Donald Douglas Atsa'am is currently a Postdoctoral Research Fellow at the University of the Free State, South Africa. He has been a Computer Science lecturer with the University of Agriculture, Makurdi, Nigeria, since 2012. Donald holds a Ph.D. in Applied Mathematics and Computer Science from the Eastern Mediterranean University, Famagusta, North Cyprus. His research interests are in Data Mining and Knowledge Discovery, Computational Intelligence, and Machine Learning. Donald is a Certified Information Systems Auditor with several years of experience in Systems Audit and Control.

Ruth Wario is a Senior Lecturer at the Department of Computer Science and Informatics at the University of the Free State, South Africa. She is currently teaching and supervising both undergraduate and postgraduate students. She teaches programming, software engineering, human computer interaction, database management and information management system. Her area of specialization is ICT4D and Human Computer Interaction.