


## Article

# A Model for Rapid Selection and COVID-19 Prediction with Dynamic and Imbalanced Data

Jeonghun Kim <sup>1</sup> and Ohbyung Kwon <sup>2,\*</sup> <sup>1</sup> Department of Management, Kyung Hee University, Seoul 02447, Korea; adsky0719@khu.ac.kr<sup>2</sup> School of Management, Kyung Hee University, Seoul 02447, Korea

\* Correspondence: obkwon@khu.ac.kr; Tel.: +82-2961-2148

**Abstract:** The COVID-19 pandemic is threatening our quality of life and economic sustainability. The rapid spread of COVID-19 around the world requires each country or region to establish appropriate anti-proliferation policies in a timely manner. It is important, in making COVID-19-related health policy decisions, to predict the number of confirmed COVID-19 patients as accurately and quickly as possible. Predictions are already being made using several traditional models such as the susceptible, infected, and recovered (SIR) and susceptible, exposed, infected, and resistant (SEIR) frameworks, but these predictions may not be accurate due to the simplicity of the models, so a prediction model with more diverse input features is needed. However, it is difficult to propose a universal predictive model globally because there are differences in data availability by country and region. Moreover, the training data for predicting confirmed patients is typically an imbalanced dataset consisting mostly of normal data; this imbalance negatively affects the accuracy of prediction. Hence, the purposes of this study are to extract rules for selecting appropriate prediction algorithms and data imbalance resolution methods according to the characteristics of the datasets available for each country or region, and to predict the number of COVID-19 patients based on these algorithms. To this end, a decision tree-type rule was extracted to identify 13 data characteristics and a discrimination algorithm was selected based on those characteristics. With this system, we predicted the COVID-19 situation in four regions: Africa, China, Korea, and the United States. The proposed method has higher prediction accuracy than the random selection method, the ensemble method, or the greedy method of discriminant analysis, and prediction takes very little time.



**Citation:** Kim, J.; Kwon, O. A Model for Rapid Selection and COVID-19 Prediction with Dynamic and Imbalanced Data. *Sustainability* **2021**, *13*, 3099. <https://doi.org/10.3390/su13063099>

Academic Editors: Son Nghiem and Demetris Lamnisis

Received: 9 January 2021

Accepted: 8 March 2021

Published: 11 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** COVID-19 pandemic; classification algorithms; data availability; big data analytics; decision tree; data imbalance

## 1. Introduction

As COVID-19 has recently become a global pandemic, each country is striving to respond appropriately. To this end, accurately predicting trends in the number of confirmed cases of COVID-19 is important in determining quarantine policies.

The factors that affect COVID-19 infection are related to direct contact with the infectious agent, but COVID-19 has a latent time during which it is difficult to identify infection externally. Therefore, in addition to rapid discovery and isolation of symptomatic infectious agents, it is important to predict trends in confirmed cases and to imagine how many infected people will be in a space where infected and healthy people without symptoms are mixed. Therefore, infection models are used, such as the susceptible, infected, and recovered (SIR) model [1]; the susceptible, exposed, infected, and resistant (SEIR) model; the Gaussian mixture model [2]; and regression analysis [3]. These models use epidemiological data such as the number of previously infected people and the total population. Infection trends are also predicted using several parameters such as the latency period and probability of healing. However, these simple models do not reflect various socio-static and economic factors that may profoundly affect the course of the virus. This is why predictions

based on economic and social data in addition to epidemiological data in the analysis of COVID-19 trends are necessary.

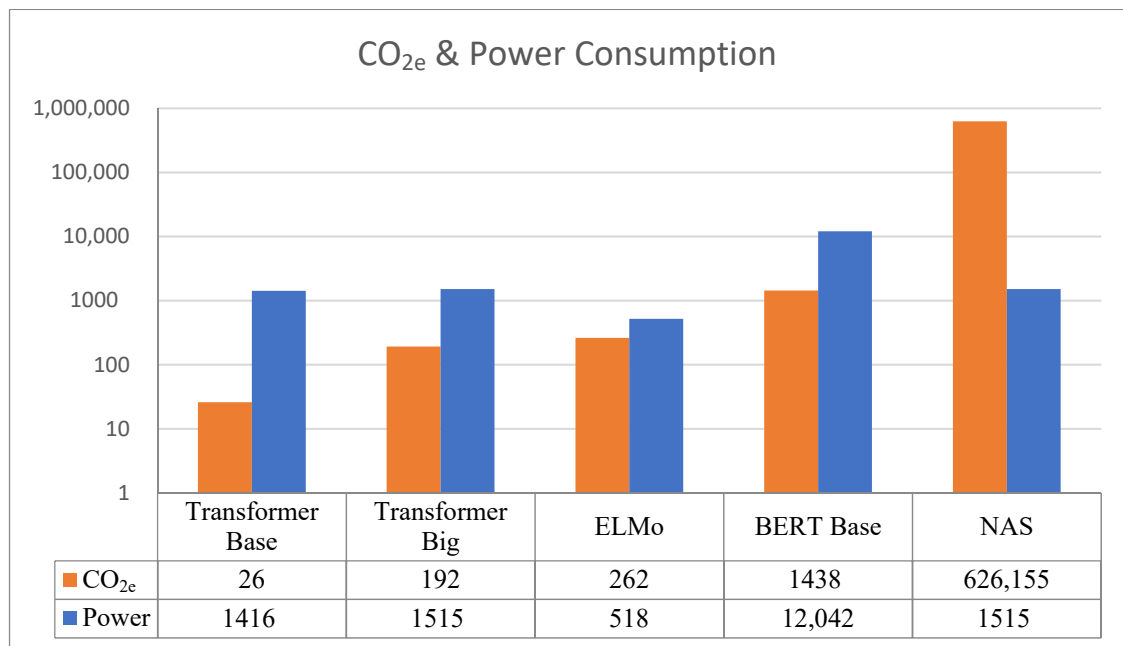
Other confounding factors make prediction difficult. In particular, COVID-19-related datasets are often highly imbalanced, which negatively affects the accuracy of predictions. To predict well using highly imbalanced datasets, statisticians often use sampling methods, cost-sensitive learning, and feature selection. However, there are no guidelines on which method and which prediction model to use depending on the nature of the dataset. Second, the economic and social variables that seem significant in predicting infectious diseases such as COVID-19 have not been confirmed. This leads to a proliferation of predictive models and makes it difficult to determine the optimal model to be applied in any given country and situation. Third, the availability of reliable economic and social data differs by country. For example, the degree of exposure to fake information related to quarantine causes confusion and increases the risk of COVID-19 infection. However, reliable exposure data are not available in some countries. Confusingly, the features included in COVID-19 prediction models are inevitably different for each country. Finally, COVID-19 infection is still progressing every day. Therefore, the urgent decision of which sampling method and prediction algorithm to use according to the characteristics of each country's dataset must be made quickly.

Thus, as part of efforts to prevent the spread of COVID-19 infection, a country-specific infection-level prediction model would solve the methodological problem of prompt and optimal prediction with an imbalanced data set consisting of features that are difficult to define in advance. Unfortunately, no meta-model that enables decisions about which characteristics of a certain dataset should be included in which sampling method with which prediction algorithm has yet been introduced. As interest in the application of artificial intelligence (AI), including machine learning, is increasing in the business domain, selecting the optimal discrimination algorithm is gaining attention, but excessive waste of resources (e.g., hardware, money, time, manpower) has occurred, which makes a prompt response difficult [4]. A guideline for selecting a sampling and discrimination algorithm based on the meta-characteristics of the dataset would make it possible to alleviate problems related to COVID-19 by quickly finding a model that can provide good prediction performance.

Therefore, the purpose of this study is to present a model that quickly predicts confirmed cases by recognizing appropriate sampling methods and algorithms according to the data characteristics of each country and region, focusing on class-imbalanced data according to the COVID-19 situation of each country. To do this, we first selected various datasets consisting of numerical and nominal values, including a socioeconomic dataset related to COVID-19, to explore the relationship between data characteristics, sampling method, and algorithm prediction performance, after which we used the results to make a decision. Using the extracted selection guidelines, we then examined how performance improves by selecting and predicting a sampling method and classification algorithm using an available dataset for each region in which COVID-19 has been detected in four countries. Using the proposed model, companies and researchers can perform eco-friendly machine learning, reduce repetitive experiments, and waste fewer resources.

## 2. Sustainable Machine Learning

Recently, many information systems have introduced machine learning for personalization and AI. However, developing an AI system through cloud computing and tensor processing consumes a huge amount of power, resulting in an indirect carbon footprint greater than the average carbon dioxide generated by a single car in a year [5]. Figure 1 is a graph representing the amount of carbon generated and the amount of power consumption required to make an AI algorithm. Efforts are therefore needed to reduce the amount of carbon generated while developing AI algorithms.



**Figure 1.** CO<sub>2</sub>e and power consumption (modified from Strubell et al., 2019). ELMo, Embeddings from Language Model; BERT, Bidirectional Encoder Representations from Transformers; NAS, Neural Architecture Search.

In terms of sustainability, machine learning can be classified into greening machine learning and greened machine learning. First, greening machine learning refers to the use of machine learning methods to maintain the planet's sustainability by doing things such as reducing carbon dioxide emissions. Using deep learning improves the accuracy of carbon dioxide emission predictions to facilitate decision-making and reduce carbon dioxide [6,7]. For example, studies have demonstrated the efficacy of machine learning for reducing the amount of repetitive work in experiments on reducing carbon dioxide emissions [8].

Greened machine learning, on the other hand, involves devising a method to reduce power consumption or costs. However, recently developed deep learning algorithms have large numbers of parameters and require considerable computation time to calculate [9]. In particular, running repeated experiments to find a suitable predictive model is costly and causes huge CO<sub>2</sub> emissions. These experiments must be repeated frequently because classification algorithms used in machine learning are not generalized to all data. Accordingly, some researchers have argued for the necessity of a generalization method that would take into account the characteristics of the dataset [10], but so far no solution has been found.

Although various dataset characteristics are being studied, there is a lack of guidelines for the selection of sampling methods or classification algorithms according to the characteristics of the dataset. Efforts have been made to elucidate the causal relationship between classification performance and data characteristics in some studies, but no guidelines for selecting a classification algorithm have been provided [11,12]. In addition, even in studies using sampling methods, there are cases where a simple classification performance comparison is limited [13]. Therefore, in this study, we propose a method of recommending a classification algorithm and a sampling method that considers the characteristics of the dataset and that also reduces the unnecessary resource consumption that is common in repeated experiments.

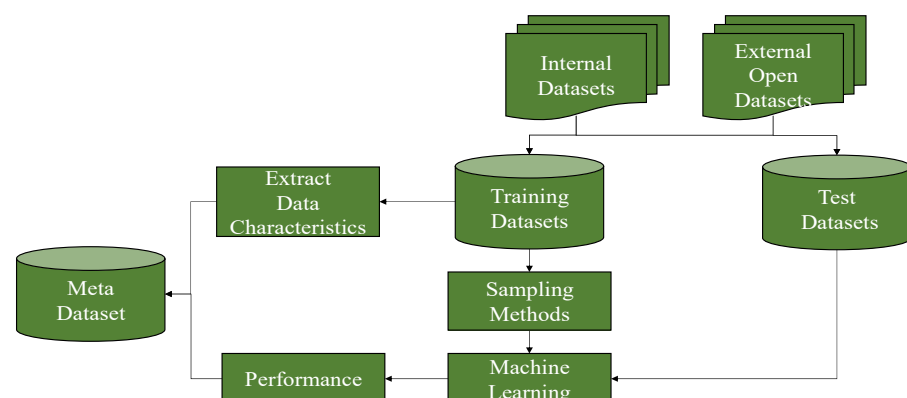
Figure 1 shows the results of calculations regarding the amount of power and carbon dioxide generated for training and hyperparameter tuning of a natural language processing (NLP) model. Examples of machine translation models include the Transformer model, the Embeddings from Language Model (ELMo), the Bidirectional Encoder Representations from Transformers (BERT) model, and the Neural Architecture Search (NAS) model.

The difference between Big and Base arises from the number of parameters: 65 million parameters in the base model and 213 million parameters in the big model. The amounts of carbon emissions that are generated in the process of learning in the BERT and NAS models (BERT = 1438 CO<sub>2</sub>e, NAS = 626,155 CO<sub>2</sub>e) are greater than the carbon emission of one passenger who travels from New York to San Francisco, which is 1984 CO<sub>2</sub>e. Many companies and data scientists are generating a large volume of carbon by repeating experiments with AI algorithms for high performance. If there is a guideline for the production of AI algorithms, it may be motivated by the need to develop AI algorithms in a more eco-friendly manner, while reducing the amount of carbon generated in repeated experiments. In particular, regulations on carbon emissions have emerged as international concerns. Currently, carbon emissions from transportation and factories are actively regulated, but with the advent of the fourth industry, AI developers based on machine learning will also be subject to such regulations due to carbon emissions from graphics processing units (GPUs) and tensor processing units (TPUs). Therefore, AI-based companies will have to make efforts to reduce carbon emissions. Companies may gain benefits from their eco-friendly images that are associated with the efforts to reduce carbon emissions, especially for AI-based companies.

### 3. Methods

#### 3.1. Metadata Collection for Finding Classification Algorithm Selection Rules

The metadata used in this study were collected through the process shown in Figure 2. The data were provided by the UCI Machine Learning Repository (see Table 1), which provides a variety of benchmark data and is frequently used by data mining researchers to verify research results. A dataset with an imbalanced class (Balance Scale, Car Evaluation, Dermatology, Wine, Contraceptive Method Choice, Glass, New-Thyroid, Hayes Roth) was selected. The selected dataset was divided using 10-fold cross-validation and then the characteristics of the dataset were extracted. For this study, several datasets with different degrees of imbalance were generated using the sampling method. The data were increased by 10% by applying the weight to the oversampling method; finally, oversampling the number of classes to the same level was attempted. For undersampling, a method of reducing the total number of classes by 10% was used. This procedure was not applied to data that did not need weight. Finally, a 10% to 100% sampling method was applied to one dataset, and 10 datasets (excluding the original dataset) were used. The total amount of metadata generated was 28,160 records.



**Figure 2.** Metadata collection process.

**Table 1.** Training datasets. HHI, Herfindahl–Hirschman index.

Data	Number of Classes	Number of Features	Number of Instances	Imbalance Ratio	HHI
Wine	3	13	178	0.676	0.342
new_thyroid	3	6	215	0.200	0.533
hayes_roth	3	5	132	0.588	0.350
Cmc	3	10	1473	0.529	0.354
Dermatology	6	35	366	0.179	0.201
Glass	6	10	214	0.118	0.263
balance_scale	3	5	625	0.170	0.431
page_block	5	11	5473	0.006	0.810

### 3.2. Dataset Characteristics Used in the Metadata

The following characteristics of datasets can affect the selection of the classification algorithm.

#### 3.2.1. Number of Instances

Number of instances is a representation of the size of the dataset. In general, with more instances, we can have a more accurate classification. However, the larger the dataset, the more noise occurs and the more learning time is required in the machine learning process. Too little noise makes correct classification difficult due to lack of information. According to Brazdil et al., 1994 [14], some classification algorithms can be determined by the number of instances.

#### 3.2.2. Number of Numeric Variables

Numeric variables provide more abundant information than nominal variables. However, compared to nominal variables, numeric variables may have an ambiguous classification boundary, which may cause difficulties in classification. Therefore, fewer numeric variables can negatively affect classification accuracy [15].

#### 3.2.3. Number of Nominal Variables

The number of nominal variables is a factor that positively influences classification accuracy [15]. Moreover, nominal variables have clear boundaries compared to numeric variables, so they can be advantageous in classification problems. However, this factor can be unfavorable for probability problems using methods such as logistic regression and naïve Bayes.

#### 3.2.4. Number of Missing Values

Missing values often appear in COVID-19 databases in various fields. Missing values cause loss of information and are related to the quality of the data. If there are many missing values, information for classification will be insufficient, and many errors will result [16].

#### 3.2.5. Herfindahl–Hirschman Index (HHI)

The HHI is an indicator of market concentration, which is the sum of the squared market share of all operators in the market [17,18]. When comparing the market share to the frequency of each class, the higher the class balance, the lower the HHI value. In addition, higher HHI values are expected to affect discrimination performance negatively because, in general, the greater the data imbalance problem, the lower the accuracy of the decision [19,20].

#### 3.2.6. Number of Variables

Reducing the number of input variables can reduce data complexity, solve multicollinearity problems between variables, and improve prediction, classification, and cluster-

ing [21,22]. However, if too few variables are considered, the complexity of the predictive model will improve, but the accuracy will be lower than when considering a large number of variables [23].

### 3.2.7. Number of Classes

The number of classes represents the dimensions of a class [24,25]. Since the expected probability decreases according to the number of classes, it will be closely related to the accuracy of the class. In general, the greater the number of classes, the lower the discrimination accuracy. Krawczyk, 2016 [26], argued for the need to find an appropriate number of classes to resolve class imbalances.

### 3.2.8. Entropy

Entropy is a variable that can quantitatively measure the uncertainty of information in a given dataset because it is related to the amount and purity of information [27,28]. The rarer the information, the better, and the more common the data, the more consistent it is. Entropy can be an indicator of whether the data are consistently pure. We measured this variable using the method proposed by Shannon [27].

### 3.2.9. Silhouette Score

The silhouette score is an index used to verify whether a cluster is correctly formed during clustering analysis. The closer the silhouette score is to 1, the more appropriate the number of clusters is. In addition, the more that clusters are grouped, the more homogeneity there is within the cluster; however, heterogeneity between clusters is better. In this study, k-means clustering was performed, where k was determined as the number of classes in the data, and then silhouette coefficients were measured.

### 3.2.10. Data Nonlinearity

In general, data nonlinearity negatively affects discrimination performance [29]. Barella et al., 2018 [12], found a correlation between data complexity and discriminant performance, and revealed a negative relationship between the nonlinearity of a linear classifier and the classification performance (G-mean). In this study, a method of measuring the nonlinearity of linear classifiers was used. This method considers both the linearity of the data and the outliers, first classifying errors through a non-linear classifier and then comparing the results classified through the linear kernel function of the support vector machine (SVM) [30].

### 3.2.11. Hub Score

The hub score is an index that measures the cohesiveness of the data using the concept of network connectivity. The score is determined by configuring a network using the given data and measuring the number of nodes connected to the network; the more connected nodes, the higher the score [30]. The hub score has values ranging from 0 to 1. The more nodes there are connected, the closer the value will be to 1. On the other hand, if the hub score is close to 0, there is a high probability of error due to overlapping with other classes [31].

### 3.2.12. Feature Overlap

The maximum Fisher's discriminant ratio measure represents the degree of overlap between variables as a ratio. The higher the data overlap, the more ambiguous the division of decision boundaries. Therefore, as the overlap of variables increases, the use of under-sampling methods such as Tomek link, edited nearest neighbors (ENN), and condensed nearest neighbors (CNN) may be more advantageous.

### 3.2.13. Neighborhood

Classes that are labeled incorrectly or that are randomly labeled are placed on different class boundaries. In some cases, it is difficult to maximize the distance between classes in a linear separation problem, and of course, this may negatively affect the discrimination performance [32].

### 3.2.14. Dimensionality

Dimensionality involves the reduction of the dimensions of the variables of the data used through principal component analysis (PCA), and the expression of the difference between the reduced variables and the original data variables as a ratio. If the dimensions are reduced more than the original data, it means that the variables in the original data are not efficient [33]. Variables that are not efficient create the risk of generating noise. In addition, if there are many unnecessary variables, machine learning may take longer.

### 3.3. Sampling Methods Used in the Metadata

Various classification studies are being conducted using algorithms such as the support vector machine (SVM), naïve Bayes classifier, k-nearest neighbor (k-NN), and decision tree. However, according to the ‘no free lunch theory’ of Wolpert and Macready, 1997 [34], an algorithm optimized for a specific problem does not exhibit the same performance in other problems. Actual data has various problems such as data imbalance, data errors, and high dimensions [34]. In particular, class imbalance has a significant negative impact on classification performance. Class imbalance means that the properties of the target variable to be classified are unbalanced. This phenomenon occurs in various fields such as fraud detection, medical diagnosis, network intrusion detection, and modern manufacturing plants. Various studies have been conducted using various methods to resolve class imbalances [35].

First, there is a cost-sensitive learning method that mitigates the bias of multiple classes by modifying the existing algorithm. Cost-sensitive learning is a method of reducing classification errors by using a cost matrix for misclassified data, unlike the data extraction method, which operates according to the distribution of classes [36]. In cases of cost-sensitive learning, it shows good performance because it is applied according to the characteristics of actual data.

Second, data preprocessing is a method of sampling training data to fit the classification algorithm. For example, undersampling balances the class distribution by removing the majority class. The problem with the undersampling method is the loss of information. On the other hand, the oversampling method balances a minority class by replicating it to fit a large number of classes. However, the oversampling method may generate noise in the data, and learning may be prolonged due to the increase in data. The data preprocessing method may be less accurate than the cost-sensitive learning method, but it does not provide a cost metric, so if you have no expertise or experience with the data, it takes a lot of time and money to find an appropriate cost metric [37]. Therefore, it would be more efficient to use the preprocessing method when dealing with a large amount of data or in cases where domain knowledge is lacking.

Looking at existing class imbalance resolution studies, we see that the performance of the sampling method differs according to the data characteristics and that performance varies according to the strategy of the classification algorithm. Therefore, in this study, the oversampling method and the undersampling method were targeted. Random oversampling (ROS), the synthetic minority oversampling technique (SMOTE), and the adaptive synthetic sampling approach for imbalanced learning (ADASYN) were selected as oversampling methods, and random undersampling (RUS), ENN, the Tomek link method, CNN, and the neighborhood cleaning rule (NCL) were selected as undersampling methods. These sampling methods are described in Table 2.

**Table 2.** Sampling Methods.

Method	Description
Random Oversampling (ROS)	This is a method of iteratively recovering and extracting data by randomly selecting data until a few classes are equal to the data size of many classes. The random overextraction method has the advantage of being very convenient to use, with almost no loss of information. However, if the sampling rate is unreasonably increased, an overfitting problem may occur because data of a minority class are repeatedly reconstructed and extracted.
Synthetic Minority Oversampling Technique (SMOTE)	This is a method of selecting random data of a minority class and artificially generating new data between k-nearest neighbors [38]. Unlike ROS, which restores and extracts fractional class data, SMOTE has been proposed to avoid the overfitting problem by generating new data.
Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN)	This is a method of generating data with consideration of the density distribution of a minority class based on SMOTE [39]. The method is similar to the distribution of the original data because it considers the density distribution of the data.
Random Undersampling (RUS)	This is a sampling method in which the majority class is randomly deleted and its proportion adjusted. RUS has the advantage of being easy to use with large-scale data, which can reduce the cost by reducing the amount of data. However, there is a high possibility of losing important information because the data are arbitrarily reduced.
Condensed Nearest Neighbors (CNN)	CNN is a method of removing data until there are no data concentrated in a majority class, leaving only representative data in the data distribution. The CNN method leaves data with clear boundaries of different classes [40]. Data are stored one-by-one and a suitable dataset is constructed by removing duplicate data.
Edited Nearest Neighbors (ENN)	Unlike CNN, if the value included in set X is misclassified, it can be excluded from X [41].
Tomek link	Based on the CNN sampling method, Tomek link is a method of removing internal data near the decision boundary. The method has the effect of removing ambiguous data overlapping with other classes [42]. Therefore, it is regarded as an efficient sampling method for removing abnormal data.
Neighborhood Cleaning Rule (NCL)	NCL is a method that combines condensed nearest neighbors (CNN) and edited nearest neighbors (ENN). It has the effect of clarifying class boundaries by removing data from multiple classes rather than the nearest data, avoiding fractional data [43].

### 3.4. Classification Algorithms Used in the Metadata

In this study, random forest, SVM, naïve Bayes classifier, k-NN, and logistic regression were considered.

First, random forest is an ensemble technique based on the majority voting method that generates several decision tree models and tends to show superior performance compared to other decision tree models.

Second, SVM is a method used to classify data from n-dimensional data using an n-1-dimensional hyperplane. The dividing line between the two classes allows selection of the hyperplane with the largest width of the two classes. This method is called linear classification. In addition, SVM is also capable of nonlinear classification using kernel functions such as polynomial kernel, sigmoid kernel, or radial base function (RBF) kernel [44].

Third, the naïve Bayes classifier is a probabilistic approach that uses Bayes' theorem during supervised learning. Although the naïve Bayes model is relatively simple and the calculation process is not complicated, it is known to exhibit excellent performance [45].

Next, unlike the naïve Bayes classifier, SVM, and the decision tree model, the k-NN method is a lazy learning method that does not use training data and moves only when empirical data are given. This is a method of selecting k neighbors among the nearest neighbors and classifying them into the most common class.

Lastly, logistic regression is a probability model that predicts the likelihood of an event using a linear combination of independent variables.

### 3.5. Classification Performance Measurement

The F-score was used to measure classification performance. Overall accuracy can be useful for measuring algorithm performance in class-balanced data. However, in real



datasets, data are often skewed to one side, so overall accuracy alone cannot be a sure indicator of the true performance of the algorithm. Therefore, in this study, the F-score was used as a performance measurement method considering the class-imbalanced situation. The formula for calculating the F-score is presented in Equation (1).

$$\text{F-score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (1)$$

### 3.6. Extraction of Classification Algorithm Recommendation Rules According to Data Characteristics

First, only the classification algorithm and the class imbalance resolution method that showed the best classification performance for each fold were extracted from the collected metadata to identify the class imbalance resolution method and classification algorithm recommendation rule. We selected classification algorithms and class imbalance resolutions that showed the greatest improvements in F-scores and if F-scores were identical, we compared them in the order of G-mean, overall accuracy, and elapsed time. Elapsed time was calculated from algorithms and class imbalance resolutions that consumed the least amount of time. In total, 70 metadata were finally used. The standard of extraction was the difference in classification performance, which was derived by applying the class imbalance resolution method from the classification performance of original data. The most improved ones were extracted.

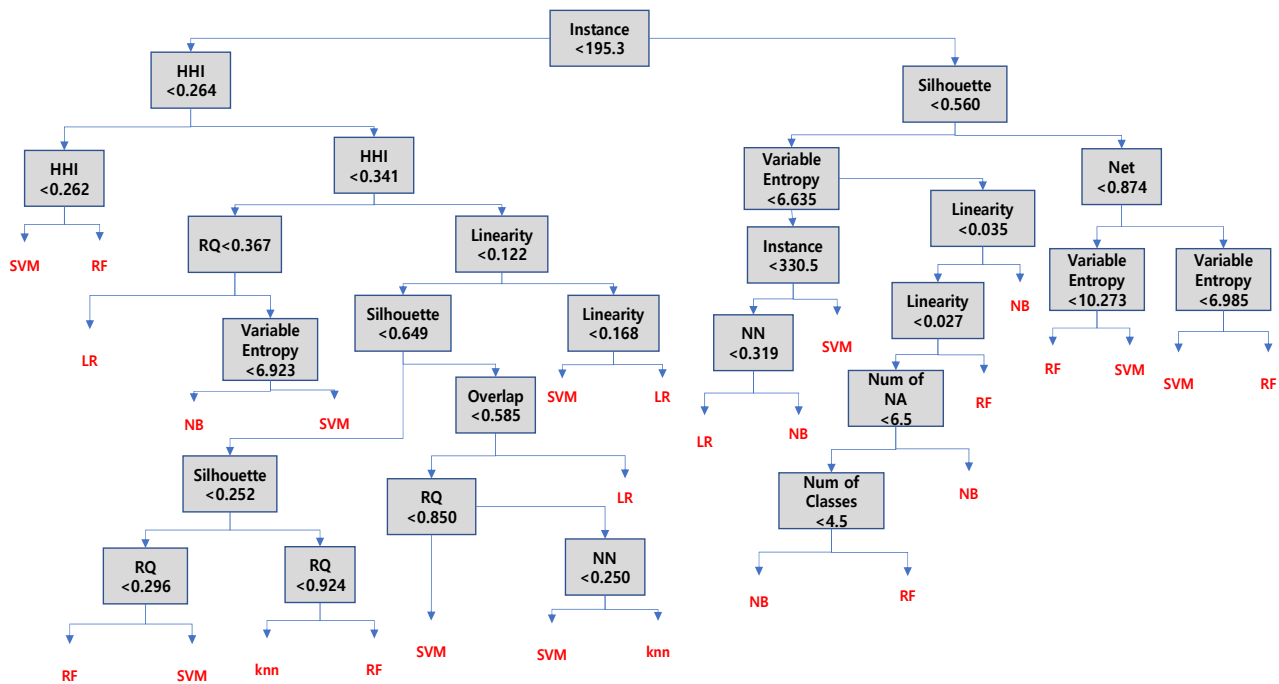
Table 3 shows the results of applying the random forest method to the data characteristics selected in this study to identify the most important factors in selecting a classification algorithm. Mean decrease accuracy refers to the expected decrease in accuracy when no factor is selected, and the mean decrease Gini refers to the decrease in data purity when no factor is selected.

**Table 3.** Importance of variables in the choice of classification algorithm.

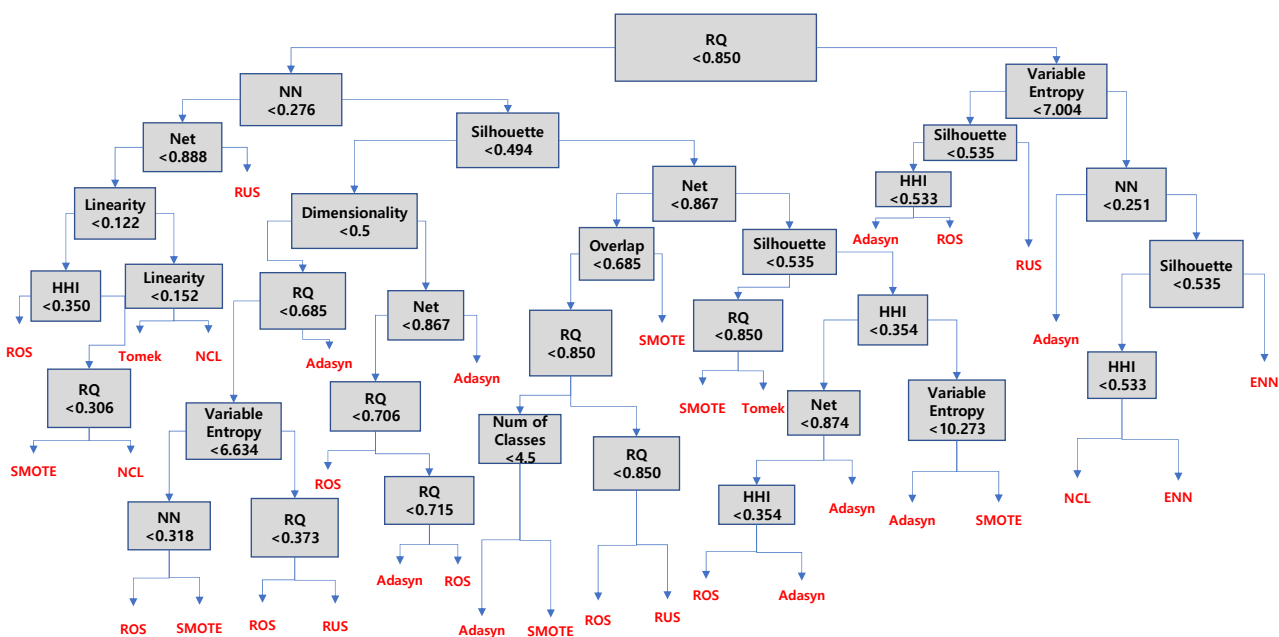
	Mean Decrease Accuracy	Mean Decrease Gini
Variable Entropy	41.594	33.339
Network	39.014	33.990
Coefficient of Determination	38.380	27.771
Mean Silhouette Score	36.215	26.814
Linearity	34.618	25.051
Overlap	33.996	25.400
Neighborhood	33.315	25.976
Number of Instances	26.958	18.619
Entropy of Classes	20.992	15.608
HHI	19.424	13.738
Dimensionality	15.510	8.321
Number of Missing Values	13.542	3.068
Number of Numeric Variables	10.892	3.782
Number of Features	10.744	3.804
Number of Classes	6.075	0.964
Number of Nominal Features	0.000	0.000

First, looking at Table 3, the mean decrease accuracy was 41.594, and the entropy of features was the most important factor in selecting a classification algorithm. The next most important factors were the network and coefficient of determination. Figures 3 and 4 show the decision trees that represent classification algorithm selection rules and sampling method selection rules, respectively. The decision trees were constructed based on the indicators (HHI, mean silhouette score, variable entropy, coefficient of determination) that characterized the datasets. The decision tree indicated that the data characteristics considered in this study were used in almost all rules. In addition, the higher the number of instances, the more divided the nodes, which resulted in the most effective classification

algorithm in a class imbalance state. Thus, the data characteristics considered in this study played an important role in selecting suitable classification algorithms.



**Figure 3.** Decision tree for algorithm selection rules. Abbreviations: Coefficient of determination = RQ, mean silhouette score = Silhouette, linearity = Linearity, overlap = Overlap, neighborhood = NN, dimensionality = Dimensionality, network = Net, variable entropy = Variable Entropy, entropy of classes = Entropy of Classes, HHI = HHI, number of missing values = Num of NA, number of numeric variables = Num of Num, number of features = Num of Features, number of classes = Number of Classes.



**Figure 4.** Decision tree for sampling method selection rules. Page: 11. Abbreviations: coefficient of determination = RQ, mean silhouette Score = Silhouette, linearity = Linearity, overlap = Overlap, neighborhood = NN, dimensionality = Dimensionality, network = Net, variable entropy = Variable Entropy, entropy of classes = Entropy of Classes, HHI = HHI, number of missing values = Num of NA, number of numeric variables = Num of Num, number of features = Num of Features, number of classes = Number of Classes.

Table 4 shows the importance of various variables in resolving class imbalances. The main factors involved in selecting the method for resolving these imbalances were variable entropy, network, and the coefficient of determination. Looking at the decision tree for resolving the imbalance problem, we see that most data can be classified based on the coefficient of determination, network, and entropy of variables. The reason for this result is that most class imbalance resolution methods delete or create data based on distance. The coefficient of determination means the extent to which the independent variable explains the dependent variable. In other words, it represents the efficiency of the variable. Network and variable entropy represent noise data.

**Table 4.** Importance of variables in the choice of sampling method.

	Mean Decrease Accuracy	Mean Decrease Gini
Overlap	41.423	34.153
Coefficient of Determination	39.421	29.778
Network	38.493	31.584
Feature Entropy	38.077	32.219
Mean Silhouette Score	37.022	31.640
Neighborhood	36.737	29.920
Linearity	33.677	24.640
Number of Instances	25.214	15.447
Entropy of Classes	23.919	14.360
HHI	23.122	13.595
Number of Missing Values	18.096	4.990
Dimensionality	15.644	6.733
Number of Features	8.798	1.925
Number of Numeric Variables	8.464	1.759
Number of Classes	6.160	0.532
Number of Nominal Features	0	0

In particular, the ADASYN, SMOTE, and Tomek methods use the strategy of calculating and removing or generating distances between data. Therefore, if the efficiency of the variables constituting the dataset is low, better results can be obtained if the class imbalance is resolved after preprocessing through methods such as dimension reduction and feature selection. In addition, as the rule was created using network, variable entropy, dimensionality, and overlap, it seems that it was determined by the structure and distribution of the data. If the variables constituting the dataset do not have a marginal hyperplane, an error may occur when creating or removing data.

## 4. Validation

### 4.1. COVID-19 Datasets

In order to check the performance of the optimal COVID-19 prediction model (decision tree type) proposed in this study in terms of the data characteristics and data imbalances, the following datasets were collected.

For this study, the United States, China, Korea, and the African continent were targeted, and the variables secured in the open dataset included humidity, temperature, population, and economic variables (GDP, store sales, etc.). The US dataset was collected by the United States Census, Bureau of Economic Analysis. In the United States Census, the number of men, women, and total population by state, and the number of students enrolled in school for more than three years were collected, and quarterly GDP was extracted from the Bureau of Economic Analysis. For the Korean dataset, population, population movement data, and sales datasets were collected from the Korean Statistical Information Service (Kosis), and temperature, humidity, and wind speed were collected from the Korea Meteorological Agency. To build the China dataset, economic, education, energy, environment, and population data were collected by China Knowledge Resource Integrated (CKNI), and weather data were collected from <http://data.sheshiyuanyi.com/> (accessed on 8 January

2021) in China. In the case of China and Korea, data were collected by city, and in the case of the United States, they were collected by state. Africa data were collected by country, and the variables used were life expectancy, GDP, population, and Gini coefficients from World Bank Open Data. The variables included in each country's dataset depended on the circumstances in each country. For a fair comparison, data analysis experts were given 24 h to build a dataset. Each dataset was collected in state units, except in Africa, where they were collected in national units. The number of samples was increased to 10 times that of the original dataset, which was too small. The values of the variables obtained were between the mean and the standard deviation.

As for the dependent variable of each dataset, differences in the number of confirmed cases or deaths between August and September 2020 were calculated. Classes included increase, flat, and decrease. The composition of the datasets is shown in Table 5.

**Table 5.** Test datasets.

Data	Number of Classes	Number of Variables	Number of Instances	Imbalance Ratio	HHI
Africa	3	5	583	0.379	0.402
China	3	18	341	0.190	0.513
South Korea	2	29	187	0.307	0.640
United States	3	10	561	0.275	0.423

Table 6 shows the characteristics of the COVID-19 datasets in the four regions. Looking at Table 6, we see that the data characteristics differ because the data that could be collected differed for each country. Even if the same data were available, the collection period, the type of data, and the completeness of the data all differed.

**Table 6.** Characteristics of COVID-19 datasets.

Number of Classes	Africa	China	Korea	United States
	3	3	2	3
HHI	0.403	0.513	0.64	0.423
Number of Instances	524.7	306.9	168.3	504.9
Number of Variables	5	18	29	10
Number of Nominal Variables	4	17	28	9
Number of Nominal Features	0	0	0	0
Number of Missing Values	6	10	16	0
Coefficient of Determination	0.013	0.199	0.684	0.233
Entropy of Classes	1.444	1.22	0.787	1.394
Entropy of Variables	9.461	8.114	7.733	8.927
Dimensionality	0.25	0.353	0.107	0.444
Linearity	0.242	0.133	0.000	0.150
Network	0.882	0.71	0.749	0.899
Overlap	0.949	0.952	0.929	0.932
Neighborhood	0.478	0.489	0.47	0.472
Mean Silhouette Score	0.677	0.361	0.638	0.479

#### 4.2. Performance Comparison

To verify the superiority of the method proposed in this study, a performance comparison was conducted. For the experiment, the following four methods were compared.

Method 1 (Random): After randomly selecting one of the classification algorithms, the algorithm predicts and discriminates the number of confirmed cases of COVID-19 in the area.

Method 2 (Ensemble): Using five classification algorithms such as k-NN, logistic regression (LR), naïve Bayes (NB), random forest (RF), and SVM, the majority vote of the

results obtained is used to predict the number of confirmed cases of COVID-19 in the region.

Method 3 (Greedy): After predicting the number of confirmed cases of COVID-19 in the region using each of the five discrimination algorithms, k-NN, LR, NB, RF, and SVM, the best performance is determined.

Method 4 (Proposed): After analyzing the characteristics of the dataset, the number of confirmed cases of COVID-19 in the area is predicted by selecting a classification algorithm according to the selection method proposed in this study.

The results are reported in Table 7. The proposed method (Proposed) was the best based on the F-score standard, and its standard error was also the lowest, indicating that its performance was very stable. In addition, even in terms of elapsed time, it can be seen that the proposed method produced results more quickly than the other methods (on average 0.740 s faster). In fact, it was more than twice as fast as Method 2 (Random), which ran quite fast. In this study, relatively few samples were used for the experiment, but if the number of confirmed cases of COVID-19 is predicted using a large dataset in the future, the effect of saved execution time will be even greater. Thus, the proposed method is excellent in terms of discrimination performance and time required for discrimination.

**Table 7.** Performance comparison among classification methods.

Method	F-Score	F-Score (s.d.)	Elapsed Time (s)	Elapsed Time (s.d.) (s)
Random	0.618	0.187	1.858	0.870
Ensemble	0.555	0.260	5.018	1.936
Greedy	0.763	0.263	9.288	4.348
Proposed	0.765	0.126	0.740	0.916

The method of selecting a method for resolving class imbalances was also compared as follows.

Method 1 (Random): After randomly selecting one of the methods for resolving class imbalances, the algorithm predicts the number of confirmed cases of COVID-19 in the region.

Method 2 (Ensemble): Based on the principle of majority vote and using the results obtained by utilizing the eight class imbalance resolution methods (ADASYN, CNN, ENN, NCL, ROS, RUS, SMOTE, Tomek), we predict and discriminate the number of confirmed cases of COVID-19 in the region.

Method 3 (Greedy): Each of the eight class imbalance resolution methods (ADASYN, CNN, ENN, NCL, ROS, RUS, SMOTE, Tomek) is used to predict the number of confirmed cases of COVID-19 in the region, and the best performance is determined.

Method 4 (Proposed): After analyzing the characteristics of the dataset, we predict the number of confirmed cases of COVID-19 in the area by selecting a class imbalance resolution method according to the method proposed in this study.

Table 8 shows the results of the performance comparison between the class imbalance resolution method and other selection methods. The proposed method (Proposed) was almost no different from the Greedy method in terms of F-score, and was a safer method in terms of performance changes ( $0.173 < 0.418$ ). Furthermore, when using the proposed method, the elapsed time was the least required. Thus, the superiority of the method of selecting the method for resolving the class imbalances proposed in this paper was demonstrated.

#### 4.3. Recommendations

Table 9 lists classification algorithms derived from the dataset characteristics utilized in this study. In Africa, random forest was recommended, naïve Bayes was recommended for China, k-NN and logistic regression were recommended for Korea, and naïve Bayes and random forest were recommended for the United States. Table 10 shows the recommendation results for class imbalance resolution methods.

ADASYN and SMOTE were recommended for Africa; ROS for China; SMOTE for Korea; and ADASYN, SMOTE, and ROS for the USA. The method recommended in this study derives from the rules trained using the metadata presented in Table 2. The example of classification algorithm selections is as follows (see Figure 3 and Table 6). The number of instances from China is larger than 195.3, so it goes in the right direction. Next, the mean silhouette score goes in the left direction because it is smaller than 0.560. Naive Bayes is selected because the entropy of the variable is greater than 6.635 and the linearity is greater than 0.035.

**Table 8.** Results of performance comparison between class imbalance resolution method selection methods.

Method	F-Score	F-Score (s.d.)	Elapsed Time (s)	Elapsed Time (s.d.) (s)
Ensemble	0.679	0.172	5.627	2.704
Random	0.568	0.241	1.654	0.750
Greedy	0.693	0.418	13.236	5.998
Proposed	0.688	0.173	0.567	0.371

**Table 9.** Recommended classification algorithms.

Data	Performance	k-NN	Logistic Regression (LR)	Naïve Bayes (NB)	Random Forest (RF)	SVM	Ensemble	Recommended Algorithm
Africa	F-score	0.49	0.17	0.67	0.56	0.34	0.27	Random Forest
	Elapsed	0.01	0.11	0.00	0.08	0.03	0.21	
	Elapsed Sum	0.37	3.46	0.21	4.64	1.83	6.58	
	Total Time	4.30	5.99	0.78	5.10	2.36	9.23	
China	F-score	0.39	0.53	0.71	0.53	0.55	0.45	Naive Bayes
	Elapsed	0.01	0.04	0.01	0.13	0.04	0.14	
	Elapsed Sum	0.46	1.16	0.37	7.18	2.35	4.71	
	Total Time	1.94	3.22	0.96	7.79	2.96	6.80	
South Korea	F-score	0.72	0.81	0.94	0.95	0.88	0.88	k-NN, Logistic Regression
	Elapsed	0.01	0.01	0.01	0.04	0.02	0.08	
	Elapsed Sum	0.38	0.45	0.59	2.11	0.87	2.40	
	Total Time	1.23	1.14	0.97	2.31	1.19	3.06	
United States	F-score	0.54	0.56	0.73	0.67	0.62	0.62	Naive Bayes, Random Forest
	Elapsed	0.01	0.05	0.00	0.11	0.04	0.19	
	Elapsed Sum	0.38	1.80	0.27	6.11	2.16	6.38	
	Total Time	3.29	6.05	0.86	6.59	2.74	10.38	

**Table 10.** Recommended class-imbalanced resolution methods.

Data	Performance	ADASYN	CNN	ENN	NCL	ROS	RUS	SMOTE	Tomek	Recommended Sampling Method	Suggested Method Elapsed
Africa	F-score	0.15	0.25	0.57	0.04	0.55	0.34	0.53	0.34	ADASYN, SMOTE	
	Sampling Time	0.05	0.11	0.02	0.09	0.04	0.10	0.04	0.21		
	Sum of Sampling Time	0.38	0.77	0.18	0.65	2.83	7.32	3.29	1.67		
	Total Time	0.61	8.62	0.31	1.25	3.14	7.42	4.18	2.23		
China	F-score	0.28	0.45	0.49	0.06	0.62	0.51	0.62	0.41	ROS	0.02
	Sampling Time	0.08	0.09	0.03	0.06	0.06	0.06	0.06	0.07		
	Sum of Sampling Time	0.60	0.72	0.23	0.49	4.82	4.62	4.25	0.50		
	Total Time	0.79	5.15	0.35	1.10	5.16	4.90	5.41	0.81		
South Korea	F-score	0.91	0.81	0.92	0.89	0.92	0.79	0.92	0.85	SMOTE	
	Sampling Time	0.03	0.04	0.02	0.05	0.02	0.03	0.02	0.04		
	Sum of Sampling Time	0.22	0.25	0.17	0.36	1.54	2.30	1.64	0.32		
	Total Time	0.29	1.71	0.27	0.45	1.77	2.55	2.14	0.72		
United States	F-score	0.68	0.54	0.67	0.60	0.68	0.58	0.66	0.56	ADASYN, SMOTE, ROS	
	Sampling Time	0.06	0.15	0.03	0.09	0.05	0.08	0.05	0.10		
	Sum of Sampling Time	0.45	1.22	0.23	0.74	4.00	5.87	3.86	0.73		
	Total Time	0.69	10.80	0.39	1.71	4.22	5.98	4.89	1.23		

## 5. Discussion

### 5.1. Contributions

In this study, we have proposed a method to recommend an appropriate classification algorithm and sampling method according to the characteristics of the dataset. This study makes the following theoretical contributions. First, in this study, 13 types of data characteristics were used, and the main factors determining classification algorithms and sampling methods were identified. With consideration of various data characteristics, a selection rule was first proposed.

Even if the characteristics of domains or datasets are different, our method can quickly identify an appropriate classification algorithm or sampling method. Domain knowledge is very important in data analysis. For example, in the microarray field, there are many variables, but relatively few instances, but in the social science field, there are relatively few variables and many instances. In order to overcome such differences arising from the characteristics of the domain, our proposed method generates a rule that determines the optimal classification algorithm and sampling method by converting the data into metadata. This method can reduce resource consumption caused by repeated measurements in the field of data science.

This study has the following practical implications. First, the results of this study can be used to develop a classification algorithm recommendation system according to the characteristics of the dataset. In particular, in the current COVID-19 global pandemic, in situations where it is necessary to predict trends quickly using different datasets by country or region, our method allows for a review of all possible discrimination algorithms or data imbalances in order to find the best alternative. Details can be missed when there is no systematic way of determining the optimal algorithm or sampling method. However, applying the method proposed in this study will facilitate the rapid recommendation of an excellent prediction method regardless of regional characteristics and data availability, thus contributing to preventing the spread of COVID-19.

Second, there are differences in discrimination performance due to parameter setting in the area of machine learning. Because all the characteristics of the data are different, knowing the optimal parameters required according to the data is necessary. Many data scientists perform repeated experiments to determine the optimal parameters and algorithms to solve new problems; this effort is time-consuming and expensive (Garcia et al., 2018). Many of the parameters of machine learning are related to the characteristics of the data. If the method proposed in this study is used, repetition of experiments will be reduced and the optimal algorithm may be selected in a relatively short time.

### 5.2. Limitations

This study has several limitations. First, not all existing class imbalance solutions were considered. For brevity, in this study, only oversampling and undersampling were considered, and feature selection and cost-sensitive learning were not considered. In future research, we plan to propose a selection algorithm that considers all of these factors.

Second, this study does not consider the data characteristics of unstructured datasets. Although most input features used in predicting confirmed cases of COVID-19 are numerical data, prediction will present few problems using the current method; however, future studies will need to include data characteristics of unstructured data.

Third, the results of this study may be generalized by repeated experimentation with the data of other countries. With any dataset, the proposed method may be applied because metadata can be constructed, as in this study. Furthermore, in this study, deep learning algorithms of artificial neural networks were not included in the algorithm recommendations. However, if the learning data include multimedia such as images, artificial neural networks (ANNs) may be included. This is an area for further study.

## 6. Conclusions

Recently, interest has been increasing in using AI methods to respond to global crises in various fields. However, since many iterations must be performed, the process of finding an appropriate algorithm can be a stumbling block in a crisis situation. Therefore, in this study, we proposed a method of identifying decision rules that determine classification algorithms and sampling methods using existing datasets. The method proposed in this study will be useful for machine learning and data mining researchers, practitioners, and machine learning-based system developers. In particular, it will contribute to the improvement of the quality of intelligent information systems using classification algorithms and minimize wasteful, resource-heavy repeated experiments by data scientists. We believe that the proposed method will also contribute to understanding the current status of COVID-19 in all countries of the world and enable policymakers to respond quickly.

Our method increases environmental sustainability by consuming as little power as possible in the process of finding an optimal machine learning algorithm in a trial-and-error manner. Until now, repeated experiments have been necessary to find the optimal algorithm, and in this process, resources are unnecessarily consumed or wasted and carbon emissions are increased. In addition, the number of parameters used in the algorithms in the development of AI continues to increase exponentially, and the consumption of resources also increases accordingly. As the rapidity and diversity of the spread of COVID-19 poses a threat to national defense systems, especially those in developing countries, accurate prediction models are vital to overcoming this crisis. Another problem is that the COVID-19 prediction models in developed countries cannot be reused as-is in developing countries due to data differences; a separate prediction model must be built. This time-consuming, expensive process threatens the lives of citizens of underdeveloped countries. We believe that companies or researchers can utilize eco-friendly machine learning to reduce repetitive experiments and wasted resources using the method proposed in this paper.

**Author Contributions:** Conceptualization, O.K.; methodology, O.K.; software, J.K.; validation, O.K.; formal analysis, J.K.; investigation, J.K.; resources, J.K.; data curation, J.K.; writing—original draft preparation, O.K.; writing—review and editing, O.K.; visualization, J.K.; supervision, O.K.; project administration, O.K.; funding acquisition, O.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by a grant from Kyung Hee University in 2019 (KHU-20191209).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is available at <https://archive.ics.uci.edu/ml/index.php>, accessed on 9 January 2021.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Zhong, L.; Mu, L.; Li, J.; Wang, J.; Yin, Z.; Liu, D. Early prediction of the 2019 novel coronavirus outbreak in the mainland china based on simple mathematical model. *IEEE Access* **2020**, *8*, 51761–51769. [[CrossRef](#)] [[PubMed](#)]
- Zhang, X.; Ma, R.; Wang, L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. *Chaos Solitons Fractals* **2020**, *135*, 109829. [[CrossRef](#)]
- Ghosal, S.; Sengupta, S.; Majumder, M.; Sinha, B. Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**, *14*, 311–315. [[CrossRef](#)]
- Garcia, L.P.; Lorena, A.C.; de Souto, M.C.; Ho, T.K. Classifier recommendation using data complexity measures. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 874–879.
- Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. *arXiv* **2019**, arXiv:1906.02243.
- Zhang, Z.; Schott, J.A.; Liu, M.; Chen, H.; Lu, X.; Sumpter, B.G.; Fu, J.; Dai, S. Prediction of carbon dioxide adsorption via deep learning. *Angew. Chem.* **2019**, *131*, 265–269. [[CrossRef](#)]
- Mardani, A.; Liao, H.; Nilashi, M.; Alrasheedi, M.; Cavallaro, F. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *J. Clean. Prod.* **2020**, *275*, 122942. [[CrossRef](#)]



8. Siebert, M.; Krennrich, G.; Seibicke, M.; Siegle, A.F.; Trapp, O. Identifying high-performance catalytic conditions for carbon dioxide reduction to dimethoxymethane by multivariate modelling. *Chem. Sci.* **2019**, *10*, 10466–10474. [[CrossRef](#)]
9. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green ai. *arXiv* **2019**, arXiv:1907.10597.
10. Sun, S.; Shi, H.; Wu, Y. A survey of multi-source domain adaptation. *Inf. Fusion* **2015**, *24*, 84–92. [[CrossRef](#)]
11. Cano, J.R. Analysis of data complexity measures for classification. *Expert Syst. Appl.* **2013**, *40*, 4820–4831. [[CrossRef](#)]
12. Barella, V.H.; Garcia, L.P.; de Souto, M.P.; Lorena, A.C.; de Carvalho, A. Data complexity measures for imbalanced classification tasks. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
13. Zhu, B.; Baesens, B.; vanden Broucke, S.K. An empirical comparison of techniques for the class imbalance problem in churn prediction. *Inf. Sci.* **2017**, *408*, 84–99. [[CrossRef](#)]
14. Brazdil, P.; Gama, J.; Henery, B. Characterizing the applicability of classification algorithms using meta-level learning. In Proceedings of the European Conference on Machine Learning, Catania, Italy, 6–8 April 1994; Springer: Berlin/Heidelberg, Germany, 1994; pp. 83–102.
15. Dogan, N.; Tanrikulu, Z. A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Inf. Technol. Manag.* **2013**, *14*, 105–124. [[CrossRef](#)]
16. Sim, J.; Lee, J.S.; Kwon, O. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Math. Probl. Eng.* **2015**, *2015*, 538613. [[CrossRef](#)]
17. Matsumoto, A.; Merlone, U.; Szidarovszky, F. Some notes on applying the Herfindahl–Hirschman Index. *Appl. Econ. Lett.* **2012**, *19*, 181–184. [[CrossRef](#)]
18. Lu, C.; Qiao, J.; Chang, J. Herfindahl–Hirschman Index based performance analysis on the convergence development. *Clust. Comput.* **2017**, *20*, 121–129. [[CrossRef](#)]
19. Wu, G.; Chang, E.Y. Aligning boundary in kernel space for learning imbalanced dataset. In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04), Brighton, UK, 1–4 November 2004; pp. 265–272.
20. Andrić, K.; Kalpić, D.; Bohaček, Z. An insight into the effects of class imbalance and sampling on classification accuracy in credit risk assessment. *Comput. Sci. Inf. Syst.* **2019**, *16*, 155–178. [[CrossRef](#)]
21. Nemhauser, G.; Wolsey, L. The scope of integer and combinatorial optimization. In *Integer and Combinatorial Optimization*; John Wiley & Sons: New York, NY, USA, 1999; pp. 1–26.
22. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
23. Morán-Fernández, L.; Bolón-Canedo, V.; Alonso-Betanzos, A. Can classification performance be predicted by complexity measures? A study using microarray data. *Knowl. Inf. Syst.* **2017**, *51*, 1067–1090. [[CrossRef](#)]
24. Rok, B.; Lusa, L. Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 64.
25. Prabakaran, S.; Sahu, R.; Verma, S. Classification of multi class dataset using wavelet power spectrum. *Data Min. Knowl. Discov.* **2007**, *15*, 297–319. [[CrossRef](#)]
26. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [[CrossRef](#)]
27. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
28. Brissaud, J.B. The meanings of entropy. *Entropy* **2005**, *7*, 68–96. [[CrossRef](#)]
29. SáEz, J.A.; Luengo, J.; Herrera, F. Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification. *Pattern Recognit.* **2013**, *46*, 355–364. [[CrossRef](#)]
30. Garcia, L.P.; de Carvalho, A.C.; Lorena, A.C. Effect of label noise in the complexity of classification problems. *Neurocomputing* **2015**, *160*, 108–119. [[CrossRef](#)]
31. Lorena, A.C.; Maciel, A.I.; de Miranda, P.B.; Costa, I.G.; Prudêncio, R.B. Data complexity meta-features for regression problems. *Mach. Learn.* **2018**, *107*, 209–246. [[CrossRef](#)]
32. Leyva, E.; González, A.; Perez, R. A set of complexity measures designed for applying meta-learning to instance selection. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 354–367. [[CrossRef](#)]
33. Lorena, A.C.; Costa, I.G.; Spolaôr, N.; De Souto, M.C. Analysis of complexity indices for classification problems: Cancer gene expression data. *Neurocomputing* **2012**, *75*, 33–42. [[CrossRef](#)]
34. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [[CrossRef](#)]
35. L'heureux, A.; Grolinger, K.; Elyamany, H.F.; Capretz, M.A. Machine learning with big data: Challenges and approaches. *IEEE Access* **2017**, *5*, 7776–7797. [[CrossRef](#)]
36. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
37. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [[CrossRef](#)]
38. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
39. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE world congress on computational intelligence. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks, Chemnitz, Germany, 1–8 June 2008; pp. 1322–1328.
40. Hart, P. The condensed nearest neighbor rule (Corresp.). *IEEE Trans. Inf. Theory* **1968**, *14*, 515–516. [[CrossRef](#)]

41. Wilson, D.L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.* **1972**, *3*, 408–421. [[CrossRef](#)]
42. Tomek, I Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.* **1976**, *6*, 769–772.
43. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Proceedings of the Conference on Artificial Intelligence in Medicine in Europe, Hong Kong, China, 1–8 June 2008; Springer: Berlin/Heidelberg, Germany, 2001; pp. 63–66.
44. Hussain, M.; Wajid, S.K.; Elzaart, A.; Berbar, M. A comparison of SVM kernel functions for breast cancer detection. Imaging and Visualization. In Proceedings of the 2011 Eighth International Conference Computer Graphics, Washington, DC, USA, 17–19 August 2011; pp. 145–150.
45. Wu, X.; Kumar, V.; Ross, J.Q.; Ghosh, J.; Yang, Q.; Motoda, H.; Geoffrey, J.M.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [[CrossRef](#)]