healthcare

MDPI

Article

Cluster Analysis of US COVID-19 Infected States for Vaccine Distribution

Dong-Her Shih [1,*], Pai-Ling Shih [2], Ting-Wei Wu [1], Cheng-Jung Li [1] and Ming-Hung Shih [3]

[1] Department of Information Management, National Yunlin University of Science and Technology, Douliu 64002, Taiwan; wutingw@yuntech.edu.tw (T.-W.W.); abc867123@gmail.com (C.-J.L.)
[2] Department of Information Management, National Chung Cheng University, Chiayi 621301, Taiwan; d08530003@ccu.edu.tw
[3] Department of Electrical and Computer Engineering, Iowa State University, 2520 Osborn Drive, Ames, IA 50011, USA; mshih@iastate.edu
* Correspondence: shihdh@yuntech.edu.tw

Abstract: Since December 2019, COVID-19 has been raging worldwide. To prevent the spread of COVID-19 infection, many countries have proposed epidemic prevention policies and quickly administered vaccines, However, under facing a shortage of vaccines, the United States did not put forward effective epidemic prevention policies in time to prevent the infection from expanding, resulting in the epidemic in the United States becoming more and more serious. Through "The COVID Tracking Project", this study collects medical indicators for each state in the United States from 2020 to 2021, and through feature selection, each state is clustered according to the epidemic's severity. Furthermore, through the confusion matrix of the classifier to verify the accuracy of the cluster analysis, the study results show that the Cascade K-means cluster analysis has the highest accuracy. This study also labeled the three clusters of the cluster analysis results as high, medium, and low infection levels. Policymakers could more objectively decide which states should prioritize vaccine allocation in a vaccine shortage to prevent the epidemic from continuing to expand. It is hoped that if there is a similar epidemic in the future, relevant policymakers can use the analysis procedure of this study to determine the allocation of relevant medical resources for epidemic prevention according to the severity of infection in each state to prevent the spread of infection.

Keywords: COVID-19; clustering analysis; classification validation; vaccine distribution; machine learning

Citation: Shih, D.-H.; Shih, P.-L.; Wu, T.-W.; Li, C.-J.; Shih, M.-H. Cluster Analysis of US COVID-19 Infected States for Vaccine Distribution. *Healthcare* **2022**, *10*, 1235. https://doi.org/10.3390/healthcare10071235

Academic Editors: Chao Wang and Chunyan Ding

Received: 6 June 2022
Accepted: 1 July 2022
Published: 2 July 2022

1. Introduction

Since the emergence of the new coronavirus COVID-19 in December 2019, it has spread through the world at a rapid rate, causing a catastrophe in the field of human public health. This virus not only affects world transportation but also causes irreparable economic and human losses. The United States, which dominates the global economic system, has failed to make timely corresponding policies for epidemic prevention [1]. Although the number of infections continues to rise, the United States seems to have failed to take effective vaccine distribution and isolation measures [2]. According to [3] who investigated vaccine allocation, it is important for a region to prioritize who receives vaccines. According to [4], reasonable vaccine distribution protocols are needed in the case of regions with unstable resources. The authors used more than 100 countries to conducted data analysis of more than 100 countries and found that problems with vaccine distribution are a main cause of the spread of influenza. In [5], authors collected the global coronavirus collective infection data, used the cluster technique to analyze, and found that virus transmission is related to family and community infection. The authors in [6] found data on influenza, used feature dimension reduction to extract important information from the data, and finally used classification algorithms to evaluate the outcome.

Clustering is the disassembly of a dataset from group to group, and the comparison between clusters shows whether the difference within the cluster is small and the difference between the groups is significant; the difference is measured by the distance between observations. Academic studies in all areas often use clustering to help analyze data and reach conclusions. For example, in [7], the authors detected gas leaks by monitoring mass spectrometer data and cluster analysis. Because of disputes between tourism development and natural landscape protection, various stakeholders were included in a cluster analysis, and the analysis results were divided into four groups: conservative to radical [8]. According to financial strategies of different companies [9], non-financial companies are often analyzed by clusters. Some strategies are suitable for high-tech economic industries, while others are suitable for basic industries.

Classification refers to establishing a data classification model based on known data and their category attributes, which can help predict which label the target data will be assigned. According to [10], sonar datasets are selected through the short-time Fourier transform, and then the sonar targets are classified by few-shot learning in the small sample learning method, which improves classification accuracy. In [11], a support vector machine was used to classify different hand movements according to experimental subjects' real-time and non-real-time EMG data, and the results showed that the human muscles set were as repetitive as fingerprints or retinas. Dritsas and Trigka [12] using different machine learning techniques to predict stroke, found that ensemble machine learning was the best approach.

In the face of the shortage of vaccines during the COVID-19 pandemic, the United States did not put forward effective epidemic prevention policies in time to prevent the infection from further expanding, resulting in an increasingly serious epidemic in the United States. This study mainly uses the COVID-19 infection case indicators collected by the COVID Tracking Project in various states of the United States. Through cluster analysis, we can distinguish the severity of infection (low, medium, and high) in each state and further allocate vaccines to states with high infection rates to carry out priority epidemic prevention measures. In this study, the data mining software WEKA is used to conduct cluster analysis and classification. After the experiment, the cluster is named after confirming the classification accuracy of the confusion matrix with classification verification. This study hopes that understanding the indicators of infection cases in each state can help allocate medical resources in the future and provide a reference for decision makers in vaccine distribution.

The structure of this study is as follows. Section 2 is a preliminary description of COVID-19, medical indicators, and vaccine distribution, combined with machine learning processes, which include feature selection, clustering method, and classification verification. Section 3 describes the dataset and the experimental process. Section 4 is the result of grouping analysis and classification verification. Section 5 is the discussion, and Section 6 is the conclusion.

2. Preliminary

2.1. COVID-19

Severe Special Infectious Pneumonia (SARS-CoV-2) aka Novel Coronavirus (COVID-19) has seriously affected people's lives worldwide, and many scholars have conducted related studies on COVID-19. A survey [13] of seriously ill patients collected many characteristics of COVID-19 symptoms, found it to be a dangerous virus, and conjectured that early pulmonary fibrosis was a substantial basis. As the epidemic became more serious, some scholars began to explore factors closely related to the death of patients. Zhou et al. [14] used univariate and multivariate logistic regression to explore factors associated with in-hospital mortality. Zheng et al. [15] analyzed the clinical characteristics of severe and non-severely ill COVID-19 patients in 13 articles with a total of 3027 patients to identify risk factors for developing severe disease or death in COVID-19 patients to predict disease

progression effectively, respond to treatment early, and allocate medical resources in a better way.

2.2. Medical Indicators and Vaccine Allocation

According to [16], a survey of Taiwanese medical institutions found that the quality of medical care and its organization and management are highly correlated. The Taiwan Quality Indicators Project (TQIP), which collected hospital datasets from 1998 to 2004, used this dataset and a survey in the United States to conduct data envelopment analysis and found that improvements in medical quality services could reduce costs [17]. In [18], the authors analyzed common disease characteristics to improve the quality of medical care in the country and found that data analysis could help the hospital to make better decisions.

According to [19], research on vaccine distribution has studied how to stop disease transmission effectively and found that when the number of transmissions is reduced, the number of deaths can be effectively reduced. The vaccine allocation study in [20] found that 5 to 15% of people worldwide die each year from epidemics. To counter this threat, the United States mass produces a variety of vaccines. However, these vaccines have no effect on newly emergent diseases. COVID-19 required the development and production of a completely new vaccine. Then the problem of vaccine distribution had to be considered.

2.3. Feature Selection Techniques

In the clustering process, if there are too many variables, the problem will often become more complicated; therefore, feature selection or feature extraction becomes very important. It can reduce the dimension of variables and make the problem simpler. The following are some commonly used feature selection methods.

2.3.1. Principal Component Analysis (PCA)

According to the discussion of the principal component analysis in [21], this technique is mainly aimed at finding the vector after the data projection, hoping to maximize the data variation, find the C (covariance) value, and finally get the covariance. Algorithm 1 is shown below:

Algorithm 1 PCA based Feature Selection.

Inputs: X = {x_1, x_2, … … , x_D} // D-dimension training dataset
Outputs: Y = {y_1, y_2, … … , y_d} // lower dimensionality d-dimensional feature set where d <= D
1 Do PCA on X for dimensionality reduction
2 Compute mean of input dataset (x)′
3 Calculate the covariance matrix Cov (x)
4 Find spectral decomposition of Cov (x) and the corresponding Eigen vectors and values E_1, E_2, … E_D to get the principal components P = ($x_{1'}$, $x_{2'}$, … … , $x_{n'}$) which is a subset of X.

The purpose of using PCA to extract features is to generate a new collection of dimensionally reduced features compared to the original dataset. This will convert a *D*-dimensional dataset to a new lower d-dimensional dataset where *D* <= d, as shown in Equation (1). Let *X* be the original dataset and *xi* be the individual variables in the dataset:

$$Consider\ a\ D - dimensional\ dataset\ X = (x_1,\ x_2,\ x_3,\ \ldots\ldots,\ x_N) \quad (1)$$

In this study, PCA was used to reduce the dimension of the data, and the specific steps were as follows. The first step was to calculate the mean value of *X* with Equation (2), *N* = number of observations.

$$(x)' = \frac{1}{N}\sum_{i=1}^{N}.(xi) \quad (2)$$

This will help standardize the data and calculate the covariance. Standardization places variables and values of data within a given range to achieve unbiased results.

The next step is computing the covariance matrix. The covariance matrix is used to identify the correlation and dependencies among the features as shown in Equation (3).

$$\mathrm{Cov(x)} = 1/N \sum_{i=1}^{N} .(xi - xi')(xi - xi')T \quad (3)$$

The last phase is spectral decomposition of the covariance matrix using eigenvectors $£_1$, $£_2, \ldots \ldots, £_D$ and eigenvalues $\lambda_1, \lambda_2, \ldots \ldots, \lambda_D$. This gives Y as shown in Equation (4). Let Y be the lower dimensional set and yi be the variables.

$$Y = (y_1,\ y_2,\ y_3,\ \ldots\ldots,\ y_P) \quad (4)$$

such that Y is the lower d-dimensional dataset and has the principal components, as shown in Equation (5).

$$(\mathrm{Y} = (£T1(x - xi'), £T2(x - xi'), £T3(x - xi'),\ \ldots\ldots,\ £Td(x - xi'))T) \quad (5)$$

such that for the original dataset X, the new dimensional representation is Y which has principal components.

2.3.2. Information Gain

Information gain uses a feature sorting method to rank the variables in the dataset. It mainly uses an entropy principle to measure a set of randomly generated variables [22]. The information gain-based feature selection is shown in Algorithm 2:

Algorithm 2 Information gain-based feature selection

Inputs: Dataset D
Outputs: Selected Features FS
Start
  Initialize threshold for gain **gt**
  Initialize feature -gain map G
Get attributes from D into A provided c
for each attribute **a** in **A**
  Find gain **g**
  IF **(g > gt)** THEN
    Add attribute **a** and **g** to **G**
  End IF
End For
For each element in **G**
  IF feature is found useful THEN
    Update **FS** with the feature
  END IF

As can be seen in Algorithm 2, the information gain-based approach finds the information gain pertaining to the importance of features in the dataset. Equations (6) and (7) are used to compute the entropy of x and y.

$$\mathrm{En(x)} = -\sum p(x) \log p(x) \quad (6)$$

$$\mathrm{En(y)} = -\sum p(y) \log p(y) \quad (7)$$

Once entropy is computed, the difference is computed to know the gain value. In fact, the gain from x on y is the reduction in entropy values and is computed using Equation (8).

$$\mathrm{IG(y,x)} = En(\mathrm{y}) - En(y/x) \quad (8)$$

2.3.3. Gain Ratio

Karegowda et al. [23] introduced the extension technique of information gain in which after the information gain result appears, it will branch it separately, and then find the best score from it. The information gain metric is used to select test attributes at each node of the decision tree. Let s be a set consisting of s data samples with m distinct classes. The expected information needed to classify a given sample is given by Equation (9).

$$\mathrm{I(s)} = -\sum_{i=1}^{m} P_i \log_2(P_i) \tag{9}$$

P_i is the probability that an arbitrary sample belongs to class $\mathrm{C_i}$ and is estimated by $\mathrm{s_i/s}$. Attribute A has v distinct values. Let $\mathrm{s_{ij}}$ be the number of samples of class $\mathrm{C_i}$ in a subset $\mathrm{S_j}$. $\mathrm{S_j}$ contains those samples in S that have value $\mathrm{a_j}$ of A. The entropy information based on the partitioning into subsets by A, is shown in Equation (10).

$$\mathrm{E(A)} = -\sum_{i=1}^{m} I(\mathrm{S}) \frac{\mathrm{S}_{1i} + \mathrm{S}_{2i} + \cdots \mathrm{S}_{mi}}{s} \tag{10}$$

The encoding information that would be gained by branching on A is shown in Equation (11)

$$\mathrm{Gain(A)} = \mathrm{I(S)} - \mathrm{E(A)} \tag{11}$$

The gain ratio which applies normalization to information gain using a value defined is shown in Equation (12)

$$SplitInfo_A(S) = -\sum_{i=1}^{v} (|S_i|/|S|) \log 2(|S_i|/|S|) \tag{12}$$

The above value represents the information generated by splitting the training dataset S into v partitions corresponding to v outcomes of a test on the attribute A. Finally, the gain ratio is defined as shown in Equation (13).

$$\text{Gain Ratio (A)} = \frac{Gain(A)}{SplitInfo_A(S)} \tag{13}$$

2.4. Cluster Analysis

Cluster analysis aims to divide samples into different groups to maximize homogeneity within each group and maximize heterogeneity between groups. This concept is similar to "intra-group homogeneity and inter-group heterogeneity" in market segmentation. There are two commonly used clustering methods introduced as follows.

2.4.1. K-Means

According to [24], K-means clustering consists of two independent stages: the first stage is to select k centers, where the k value is pre-fixed, and the next stage is to bring each data object to the nearest center. Supposing that the target object is x, x_i indicates the average of cluster C_i. The criterion function is defined as shown in Equation (14).

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2 \tag{14}$$

E is the sum of squares of errors of all objects in the dataset. The distance of the criterion function is the Euclidean distance, which is used to determine the closest distance of each data object to the cluster center. The distance between one vector $x = (x_1, x_2, \ldots\ x_\mathrm{n})$ and another vector $y = (y_1, y_2, \ldots\ y_\mathrm{n})$, is the Euclidean distance $d(x_i, y_j)$. It can be calculated as shown in Equation (15).

$$d(x_i, y_j) = \left[\sum_{i=1}^{n} (x_i - y_j)^2\right]^{1/2} \tag{15}$$

2.4.2. Cascade K-Means

Another clustering is based on [25], in which there is a Calinski–Harabasz metric that evaluates the number of clusters, calculates the distance, and then uses the metric to decide whether to continue the cluster and find the optimal number of clusters. The matrices $WG^{\{k\}}$ are square symmetric matrices of size p × p. Let WG denote their sum for all the clusters as shown in Equation (16).

$$WG = \sum_{k=0}^{K} WG^{\{k\}} \tag{16}$$

The matrices $WG^{\{k\}}$ represent a positive semi-definite quadratic form Q_k, and their eigenvalues and their determinant are greater than or equal to 0. The within-cluster dispersion, noted as $WGSS^{\{k\}}$ or $WGSS_k$, is the trace of the scatter matrix $WG^{\{k\}}$, as shown in Equation (17).

$$WGSS^{\{k\}} = Tr\left(WG^{\{k\}}\right) = \sum_{i \in I_k} \left|\left|M_i^{\{k\}} - G^{\{k\}}\right|\right|^2 \tag{17}$$

The within-cluster dispersion is the sum of the squared distances between the observations $M_i^{\{k\}}$ and the barycenter $G^{\{k\}}$ of the cluster. Finally, the pooled within-cluster sum of squares WGSS is the sum of the within-cluster dispersions for all the clusters as shown in Equation (18).

$$WGSS = \sum_{k=0}^{K} WGSS^{\{k\}} \tag{18}$$

The between-group dispersion measures the dispersion of the clusters between each other. This sum is the weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k, as shown in Equation (19).

$$BGSS = \sum_{k=1}^{K} n_K \left|\left|G^{\{k\}} - G\right|\right|^2 \tag{19}$$

Using the notations of Equations (18) and (19), the Calinski–Harabasz metric is shown as in Equation (20).

$$\mathbf{CH} = \frac{\mathrm{BGSS}/(\mathrm{K}-1)}{\mathrm{WGSS}/(\mathrm{N}-\mathrm{K})} = \frac{\mathrm{N}-\mathrm{K}}{\mathrm{K}-1}\frac{\mathrm{BGSS}}{\mathrm{WGSS}} \tag{20}$$

2.5. Classification

Classification is a critical method of data mining. The classification concept is to learn a classification function or construct a classification model (usually called a classifier) based on existing data. The function or model can map data records in the database to one of the given categories, and apply it to data prediction. The classifier is a general term for classifying samples in data mining, including decision tree, logistic regression, naive Bayes, neural networks, and other algorithms. The following is an introduction to the two classifiers used in this study.

2.5.1. Random Forest

According to [26], the generation technology of random trees evolved from decision trees. In addition, the pattern generation method of a tree can also be used without selecting the data target first. This technique is an ensemble learning algorithm that uses bagging plus random feature sampling. The random forest training algorithm applies the general bagging technique to tree learning. Given a training set $X = x_1, \ldots, x_n$ and a target $Y = y_1, \ldots, y_n$, the bagging method is repeated (B times) to sample from the training set with replacement and then train a tree model on these samples:

For $b = 1, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.

After training, the prediction for the unknown sample x can be achieved by averaging the predictions of all individual regression trees on x as in Equation (21):

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{21}$$

2.5.2. Neural Network

According to [27], the neural network is one of the classic technologies inspired by the human brain. human brain can process different information content because it has different neurons. The includes hundreds of millions of neurons that can connect and share. When a neuron can continuously connect to other neurons, it can trigger the brain to control the human body to complete some behaviors. This behavior is essentially a process of learning and absorbing knowledge, while new neural connections stimulate the brain to learn new actions. The process mentioned above is similar to the one used in neural networks, where neurons can be defined as one or more nodes used. The structure of neurons is an input layer, a hidden layer, and an output layer which can be shown in Equation (22), where $\sigma()$ is called the activation or transfer function, N is the number of input neurons, V_{ij} is the weights, x_j is inputs to the input neurons, and T_i^{hid} is the threshold terms of the hidden neurons.

$$H_i = \sigma\left(\sum_{j=1}^{N} V_{ij}x_j + T_i^{hid}\right) \tag{22}$$

3. Material and Methods

3.1. Dataset

The dataset of this study was collected from various states in the United States (The COVID Tracking Project). The dataset variables are divided into various levels. The variables marked as grade A or above indicate that the information provided by this state is relatively sufficient and complete and vice versa. There are 44 variables in the dataset, roughly divided into cases, PCR tests, antibody tests, antigen tests, hospitalizations, death outcomes, and the state metadata, as shown in Table 1.

Table 1. Dataset.

Item	Variables	Attribute	Item	Variables	Attribute
1	date	String	24	inlcucumulative	Numerical
2	state	String	25	inlcuCurrently	Numerical
3	dataQualityGrade	String	26	onVentilatorCumulative	Numerical
4	positive	Numerical	27	onVentilatorCurrently	Numerical
5	positive Increase	Numerical	28	death	Numerical
6	probable Cases	Numerical	29	death Increase	Numerical
7	positiveScore	Numerical	30	death Probable	Numerical
8	positiveCasesViral	Numerical	31	death Confirmed	Numerical
9	positiveTestsViral	Numerical	32	recovered	Numerical
10	positiveTestsPeopleAntibody	Numerical	33	totaltestResults	Numerical
11	positiveTestsAntibody	Numerical	34	totalTestResultsIncrease	Numerical
12	positiveTestsPeopleAntigen	Numerical	35	totalTestsViral	Numerical
13	positiveTestsAntigen	Numerical	36	totalTestsViralIncrease	Numerical
14	negative	Numerical	37	totalTestsPeopleViral	Numerical
15	negativeTestsViral	Numerical	38	totalTestsPeopleViralIncrease	Numerical
16	negativeTestsPeopleAntibody	Numerical	39	totalTestEncountersViral	Numerical
17	negativeTestsAntibody	Numerical	40	totalTestEncountersViralIncrease	Numerical

Table 1. *Cont.*

Item	Variables	Attribute	Item	Variables	Attribute
18	negativeIncrease	Numerical	41	totalTestsAntigen	Numerical
19	Pending	Numerical	42	totalTestsPeopleAntigen	Numerical
20	hospitalized	Numerical	43	totalTestsAntibody	Numerical
21	hospitalized Increase	Numerical	44	totalTestsPeopleAntibody	Numerical
22	hospitalized Cumulative	Numerical			
23	hospitalized Currently	Numerical			

3.2. Conceptual Framework

The clustering analysis and classification scenario of this study is shown in Figure 1, which are data preprocessing, cluster experiment, and classification verification. Feature selection was performed before the dataset was clustered, and essential feature variables were identified. The results of the two clustering methods were statistically compared for these essential variables. Finally, after clustering analysis, the classification method was used to verify the confusion matrix.

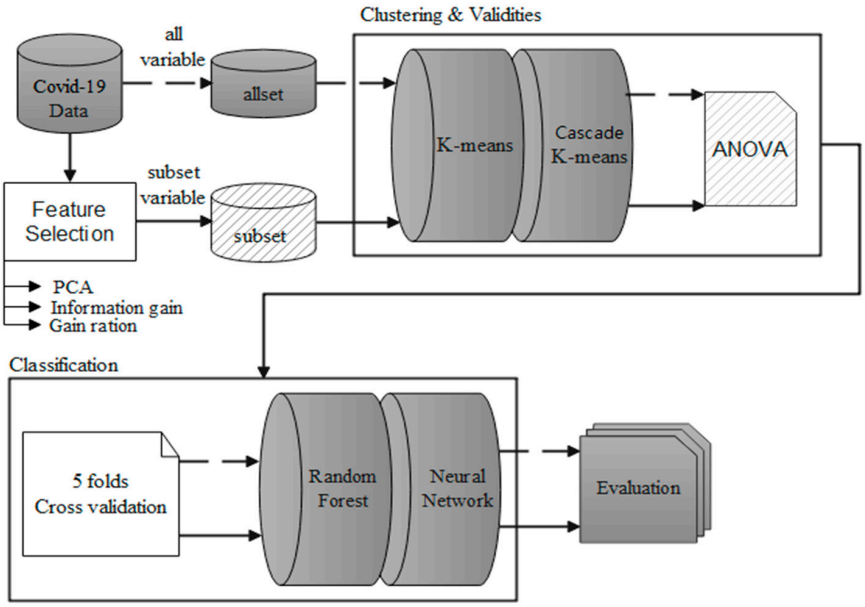

Figure 1. Clustering analysis and classifications scenario.

3.3. A Brief Review of Clustering Techniques

Clustering is to group all data and classify similar data into the same group. A piece of data only belongs to a particular group, and each group is called a cluster. Defining the so-called similarity is usually judged by the distance between data points. The closer the distance is, the more similar it is presumed to be. The denser the neighbors, the more similar they are presumed to be. The clustering techniques are pretty diverse. In the 1970s, most of the published studies were performed with hierarchical-based algorithms. In addition to generating a tree graph, these algorithms can also present the division of relatively important and target clusters [28]. However, when the merge or split decision is implemented in the pure hierarchical clustering method, the quality of the clustering will be affected, and it will not undo previous operations. Moreover, an object cannot move to another cluster [29]. The density-based algorithm plays a vital role in nonlinear shapes and structures derived from density. The concepts of the density-based algorithm are density accessibility and density connectivity. However, most of the indicators used to evaluate or

compare cluster analysis results are not suitable for evaluating the results of density-based clustering analysis [30].

In the past few years, there has been much work on graph-based clustering, and theorists have extensively studied the properties of clustering and the quality measures of various clustering algorithms using elegant mathematical structures established in graph theory [31]. For example, the Markov Cluster Algorithm is a fast and scalable graph (also known as a network) unsupervised clustering algorithm based on the simulation of (random) flow in a graph [32]. A comparison table of different clustering techniques is shown in Table 2. The pros and cons of different clustering techniques are also included. This study employs a partitioning algorithm, a non-hierarchical approach, to evaluate clustering results by constructing various partitions. Criteria are globally optimal or efficient heuristics. K-means is the most commonly used method [33], which needs to define the number of clusters in advance to meet the requirements of specific clusters. The dataset in this study does not belong to a complex real network or shape, so the cluster analysis in this study is performed by well-integrated K-means and modified Cascade K-means.

Table 2. Comparison of different clustering techniques.

Category	Hierarchical	Density-Based	Graph-Based	Partitioning
Based on	Linkage methods	Density accessibility Density connectivity	Graph theory	Mean Centroid Mediod-Centriod
Type of Data	Numerical	Numerical	Mix data	Numerical
Pros	Easy to implement Good for small datasets	Found clusters of arbitrary shapes and sizes	Perform well with complex shapes of data	Easy to implement Robust and easier to understand
Cons	Fails on larger sets Hard to find the correct number of clusters	Doe not work well in high dimensionality data.	Can be costly to compute	Unable to handle noisy data and outliers

3.4. Data Preprocessing

3.4.1. Feature Selection

According to [34], when the data are more complex, the readability is lower; therefore the reduction of the data dimension becomes a matter of course, and the reduction of the data dimension is also a method usually used for feature selection. Uğuz [35] used the method of data dimension reduction to extract features in different mixed models using two-stage testing and finally obtained better results. Therefore, this study used PCA, IG, and GR to select essential variables and then sorted out co-occurring variables. The method of feature selection and the detailed WEKA feature selection process is shown in Figure 2. The selected common variables are used for the next step of cluster analysis.

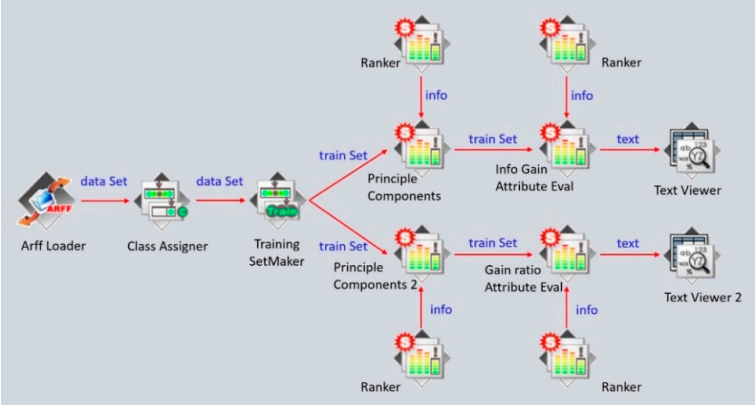

Figure 2. WEKA Feature selection process.

3.4.2. Clustering Analysis and Classification

This study mainly used the dataset from 2020 to 2021 in the COVID Tracking Project to conduct cluster analysis. In the experiment, fearing that the single clustering method was too subjective, we intentionally compared the clustering results of the two clustering methods after data preprocessing. Both K-means and Cascade K-means were used in the cluster experiments in this study. The classification verification after the cluster analysis was completed. Previous research found that the random forest method (RF) and the neural network (NN) have good performance [36]. We therefore used these two methods to compare the confusion matrix in the classification results and verify the effect of cluster analysis. The detailed WEKA cluster experiment and classification verification process are shown in Figure 3.

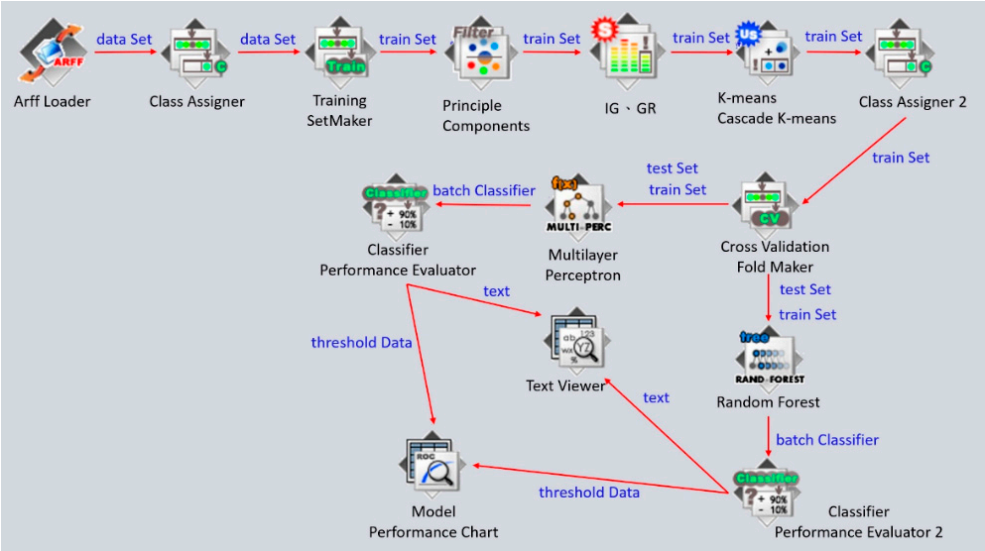

Figure 3. Clustering and classification validation with WEKA.

4. Results

4.1. Feature Selection and Clustering Analysis Result

The descriptive statistics such as the minimum, maximum, average, and standard deviation of the data fields marked as grade A in The COVID tracking project dataset are shown in Table 3.

Table 3. Dataset descriptive statistics.

Data Field	Minimum	Maximum	Mean	Standard Deviation
Death	3.563	20,146.993	3012.726	4063.720
deathConfirmed	0.000	11,873.819	1612.264	2564.015
deathIncrease	0.229	139.083	23.802	28.186
deathProbable	0.000	1557.983	116.001	241.290
Hospitalized	0.000	74,908.536	6795.447	12,375.868
hospitalizedCumulative	0.000	74,908.536	6795.447	12,375.868
hospitalizedCurrently	0.000	5706.697	1016.505	1178.133
hospitalizedIncrease	−0.868	574.361	52.537	93.318
inIcuCumulative	0.000	4167.848	374.140	935.623
inIcuCurrently	0.000	1594.500	198.782	310.440
negative	3193.583	5,676,868.213	1,217,574.982	1,393,361.630

Table 3. *Cont.*

Data Field	Minimum	Maximum	Mean	Standard Deviation
negativeIncrease	225.347	66,790.590	10,430.280	13,289.475
negativeTestsAntibody	0.000	274,785.535	9939.138	42,390.124
negativeTestsPeopleAntibody	0.000	307,446.436	9400.504	44,786.195
negativeTestsViral	0.000	5,069,123.190	376,163.876	913,586.361
onVentilatorCumulative	0.000	1210.757	44.743	189.114
onVentilatorCurrently	0.000	383.805	77.040	115.523
positive	290.354	689,808.865	126,025.819	137,216.541
positiveCasesViral	0.000	644,108.814	99,837.740	122,787.805
positiveIncrease	16.833	6633.789	1390.662	1405.567
positiveScore	0.000	0.000	0.000	0.000
positiveTestsAntibody	0.000	32,553.559	2128.500	6739.150
positiveTestsAntigen	0.000	28,745.608	1748.258	5172.510
positiveTestsPeopleAntibody	0.000	30,843.331	969.265	4476.315
positiveTestsPeopleAntigen	0.000	19,372.812	912.666	3418.733
positiveTestsViral	0.000	746,688.084	69,991.507	149,301.079
recovered	0.000	548,376.917	56,375.980	91,104.014
totalTestEncountersViral	0.000	6,252,107.282	436,938.962	1,266,757.976
totalTestEncountersViralIncrease	0.000	70,634.798	4496.100	13,115.623
totalTestResults	3643.785	6,252,107.282	1,575,543.098	1,650,858.474
totalTestResultsIncrease	250.514	70,634.798	14,555.709	15,895.585
totalTestsAntibody	0.000	336,182.488	33,358.436	81,366.549
totalTestsAntigen	0.000	329,705.523	23,679.767	57,553.140
totalTestsPeopleAntibody	0.000	338,396.716	13,807.533	51,122.147
totalTestsPeopleAntigen	0.000	121,896.261	6203.051	21,180.165
totalTestsPeopleViral	0.000	3,939,157.669	424,758.638	718,630.307
totalTestsPeopleViralIncrease	−251.653	31,530.365	3374.921	5665.766
totalTestsViral	0.000	5,972,478.403	1,244,215.985	1,591,549.099
totalTestsViralIncrease	0.000	52,143.061	11,304.525	14,366.747

We first performed principal component analysis (PCA) on the dataset and extracted the first 20% of the variables from the PCA results. The detailed results are shown in Table 4.

Table 4. PCA feature selection results.

	Features										
Group	Pc1	Pc2	Pc3	Pc4	Pc5	Pc6	Pc7	Pc8	Pc9	Pc10	Pc11
Variation	15.89	6.05	4.69	2.64	1.90	1.38	1.11	0.92	0.64	0.54	0.49
Variation Percentage	0.42	0.16	0.12	0.07	0.05	0.04	0.03	0.02	0.02	0.01	0.01
Cumulative contribution ratio	0.42	0.58	0.70	0.77	0.82	0.85	0.88	0.91	0.93	0.94	0.96

Next, we continued to extract the feature variables from the IG and GR methods. Feature variable selection of the IG and GR methods also takes the top 20% of the feature variables and the top 20% of the PCA feature variables for sorting and comparison. The results are shown in Table 4. The feature variables that repeatedly appear in Table 4 are selected in this study. We list these important feature variables that repeatedly appear in Table 5. Table 6 is the cluster feature variables selected for this study, and there are 10 variables in total.

Table 5. Feature selection sorting.

Rank	PCA	IG	GR	Average Rank
1	A30	A2	A39	2.7302
2	A19	A19	A11	2.7302
3	A21	A30	A13	2.7302
4	A31	A31	A16	2.7302
5	A8	A13	A18	2.7302
6	A15	A21	A19	2.7302
7	A24	A12	A38	2.7302
8	A34	A4	A12	2.7302
9	A14	A8	A9	2.7302
10	A36	A27	A22	2.6814
11	A9	A38	A8	2.4945
12	A33	A39	A3	2.4945
13	A6	A20	A4	2.4945
14	A7	A7	A5	2.3984
15	A23	A6	A6	2.3984
16	A3	A11	A7	2.3269
17	A5	A9	A21	2.3269
18	A39	A36	A20	2.2745
19	A29	A18	A2	1.9183
20	A18	A3	A31	1.9183

Table 6. Clustering variable.

Code	Variable	Definition
A3	deathConfirmed	Number of confirmed deaths
A6	Hospitalized	Number of hospitalizations
A7	hospitalizedCumulative	Cumulative hospitalizations
A8	hospitalizedCurrently	Number of people currently hospitalized
A9	hospitalizedIncrease	New hospitalizations
A18	onVentilatorCurrently	Number of respirators currently in use
A19	positive	Number of confirmed cases
A21	positiveIncrease	The number of new diagnoses
A31	totalTestResults	Total number of tests
A39	totalTestsViralIncrease	Number of new PCR tests

According to the critical cluster characteristics in Table 6, the number of confirmed cases, the number of deaths, the number of hospitalizations, the number of PCR tests,

and other variables are related. It can be seen from this information that the results of future cluster analysis will have a positive relationship with the severity of the confirmed outbreak, and age, hospitalization, and death are closely related [37].

This study next carried out cluster analysis. The variables after feature selection (FS) in Table 5 were analyzed by K-means and Cascade K-means methods of WEKA. The results of the cluster analysis of US states are shown in Table 7.

Table 7. Clustering results.

Method	**FS + K-Means Clustering**			**FS + Cascade K-Means Clustering**		
Group	Cluster1	Cluster2	Cluster3	Cluster3	Cluster1	Cluster2
Count	5 states	5 states	41 states	7 states	22 states	22 states
cluster member	LA MO NC PA TN	IL MA MI OH TX	AK, AL, AR, AZ, CO, CT, DC, DE, FL, GA, GU, HI, IA, ID, IN, KS, KY, MD, ME, MN, MS, MT, ND, NE, NH, NJ, NM, NV, NY, OK, OR, PR, RI, SC, SD, UT, VA, VT, WA, WI, WY	IL LA MI MO NC PA TX	AL, AR, AZ, CO, FL, GA, IN, KY, MA, MD, MN, MS, NJ, NM, NY, OH, OK, SC, TN, UT, VA, WI	AK, CT, DC, DE, GU, HI, IA, ID, KS, ME, MT, ND, NE, NH, NV, OR, PR, RI, SD, VT, WA, WY

It can be observed from Table 6 that the clustering results of these two types of cluster analysis are not the same.

To determine which cluster analysis results were better, we first performed an ANOVA analysis of variance between clusters. Table 7 shows the results of the analysis of variance (ANOVA) between clusters after using K-means for cluster analysis. Table 8 shows that 7 of the 10 selected variables are significant. Moreover, Table 9 shows the results of ANOVA analysis between clusters after using Cascade K-means for cluster analysis. Table 9 shows that all 10 variables are significant. Therefore, the preliminary analysis results show that there are significant differences between clusters using Cascade K-means cluster analysis results, and the effect of clustering is better.

Table 8. ANOVA analysis results with K-means clustering.

Source	**Sum of Squares**	**df**	***p*-Value**
deathConfirmed	96,309,899.004	2	0.000 ***
hospitalized	163,810,675.372	2	0.595
hospitalizedCumulative	163,810,675.372	2	0.595
hospitalizedCurrently	20,454,751.196	2	0.000 ***
hospitalizedIncrease	10,452.301	2	0.558
onVentilatorCurrently	171,650.041	2	0.001 ***
positive	297,933,527,997.266	2	0.000 ***
positiveIncrease	33,214,641.966	2	0.000 ***
totalTestResults	56,194,090,472,628.98	2	0.000 ***
totalTestsViral	73,640,463,254,532.75	2	0.000 ***

*** $p < 0.001$.

4.2. Classification Validation

Although the cluster analysis results verify that the clustering effect of Cascade K-means is better, we still need more evidence to prove its clustering effect. Therefore, we used two classifiers to train and test their classification effects in this section. The accuracy

of the confusion matrix after classification by the classifiers can represent the clustering effect after cluster analysis. The higher the accuracy is, the better the clustering effect.

Table 9. ANOVA analysis results with Cascade K-means clustering.

Source	Sum of Squares	df	*p*-Value
deathConfirmed	92,467,765.941	2	0.000 ***
hospitalized	2,542,735,930.524	2	0.000 ***
hospitalizedCumulative	2,542,735,930.524	2	0.000 ***
hospitalizedCurrently	28,731,039.605	2	0.000 ***
hospitalizedIncrease	146,195.267	2	0.000 ***
onVentilatorCurrently	182,801.937	2	0.000 ***
positive	428,233,021,802.748	2	0.000 ***
positiveIncrease	47,433,754.861	2	0.000 ***
totalTestResults	62,516,522,292,791.39	2	0.000 ***
totalTestsViral	53,777,136,225,256.51	2	0.000 ***

*** $p < 0.001$.

4.2.1. Validation of Random Forest

This section verifies the clustering results of K-means through the confusion matrix of the random forest classifier. As shown in Table 7, the clustering results of K-means are divided into three groups, the first group has 5 states, the second group has 5 states, and the third group has 41 states. This study uses WEKA's random forest (RF) classifier for training and testing. The confusion matrix of the test results is shown in Table 10. In the classification of the first group, a total of five states were misclassified to the third group, and four states of the second group were misclassified to the third group. The accuracy of random forest classification using the clustering results of K-means was 82.35%.

Table 10. Random forest classification validation for K-means clustering.

Confusion Matrix			Clustering Class		
			Cluster1	Cluster2	Cluster3
			a	b	c
Prediction Class	Cluster1	a	0/5	0	5
	Cluster2	b	0	1/5	4
	Cluster3	c	0	0	41/41

In addition, as can be seen from Table 7, the cluster analysis results of Cascade K-means are also divided into three groups. The first group has 22 states, the second group has 22 states, and the third group has 7 states. This study still uses WEKA's random forest (RF) classifier for training and testing. The confusion matrix of the test results is shown in Table 11. In the classification of the first group, only one state was misclassified to the third group, and the rest were classified correctly, with a classification accuracy rate of 98.03%. This study again shows that the cluster analysis results of Cascade K-means are better than the cluster analysis results of K-means.

4.2.2. Validation of Neural Network

In case the results of random forest are too subjective, in this section we will use another neural network classifier to verify the results of cluster analysis again. Table 12 is the confusion matrix verified by the K-means cluster analysis results through neural networks classification. As can be seen from Table 12, five states were misclassified to the third group in the classification of the first group. In addition, the second group had

three states misclassified, and the third group had two states misclassified from the second group, and the classification accuracy for K-means clustering was 80.39%.

Table 11. Random forest classification validation for Cascade K-means clustering.

Confusion Matrix			Clustering Class		
			Cluster1	Cluster2	Cluster3
			a	b	c
Prediction Class	Cluster1	a	21/22	1	0
	Cluster2	b	0	22/22	0
	Cluster3	c	0	0	7/7

Table 12. Neural network classification validation for K-means clustering.

Confusion Matrix			Clustering Class		
			Cluster1	Cluster2	Cluster3
			a	b	c
Prediction Class	Cluster1	a	0/5	0	5
	Cluster2	b	2	2/5	1
	Cluster3	c	0	2	39/41

Table 13 is the confusion matrix verified by the Cascade K-means cluster analysis results through neural networks' classification. It can be seen from Table 13 that under the classification and verification of the neural network method, there were three states in the first group and the second group each with three misclassifications, and the classification accuracy rate is 88.23%. Table 14 shows the comparison of classification verification of the two clustering methods, and it can be seen that Cascade K-means has a better clustering result in terms of accuracy, precision, and recall. The study results again show that the cluster analysis results of Cascade K-means are still better than the cluster analysis results of K-means.

Table 13. Neural network classification validation for Cascade K-means clustering.

Confusion Matrix			Clustering Class		
			Cluster1	Cluster2	Cluster3
			a	b	c
Prediction Class	Cluster1	a	19/22	3	0
	Cluster2	b	0	22/22	0
	Cluster3	c	3	0	4/7

Table 14. Comprehensive comparison of two clustering methods.

Clustering	K-Means		Cascade K-Means	
Validation	RF	NN	RF	NN
Accuracy	0.8235	0.8039	0.9803	0.8823
Precision	1	0.911	1	0.938
Recall	0.824	0.872	0.98	0.938

The above study results show that the cluster analysis results of Cascade K-means are the best, so the following discussions are based on the cluster analysis results of Cascade K-means for more in-depth discussions.

5. Discussion

After classification and verification, this study uses the cluster analysis results of Cascade K-means with the highest accuracy as the cluster basis. Figure 4 shows the characteristics of the 10 feature variables of the 3 clusters. The proportion of variables in the first cluster is all below 20%, The second cluster of variables is mainly characterized by (total TestResults) and (deaths Confirmed). The third cluster of variables is mainly characterized by hospitalization-related from (hospitalizedIncrease), (hospitalizedCurrently), and (hospitalizedCumulative). Next, we will label the three clusters of the cluster analysis results. In this study, the average of the 10 feature variables of the 3 clusters was used to label the high, medium, and low severity of the epidemic infection in each state in the United States. Table 15 is the cluster labeling result of this study.

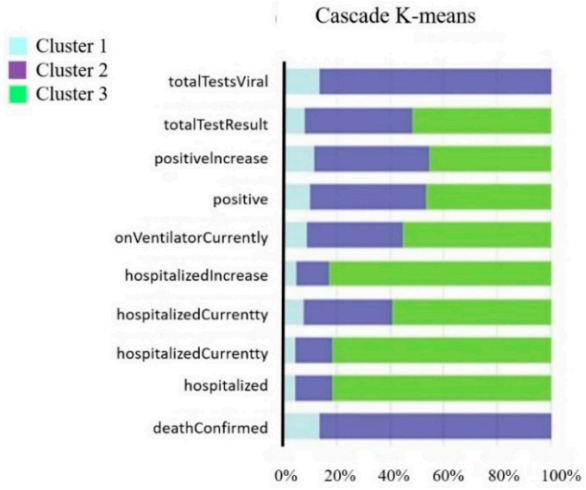

Figure 4. Cluster characteristics.

Table 15. Cluster labeling.

Severity	States
Low	AK, CT, DC, DE, GU, HI, IA, ID, KS, ME, MT, ND, NE, NH, NV, OR, PR, RI, SD, VT, WA, WY
Medium	AL, AR, AZ, CO, FL, GA, IN, KY, MA, MD, MN, MS, NJ, NM, NY, OH, OK, SC, TN, UT, VA, WI
High	IL, LA, MI, MO, NC, PA, TX

Nevertheless, Figure 5 is a color map of the severity of the outbreak in each US state based on the findings of this study. California in the lower-left corner was omitted because the data collected in California at the beginning of the outbreak were not sufficient and were not listed as A-level. From the color map, it can be seen that the southeastern United States was more serious. At the same time, the United States also had a large outbreak of influenza in 2021, as shown in Figure 6, and the southeastern United States was also the main infection area. That as an interesting finding. In the future, decision makers can refer to these two graphs and deduce an epidemic prevention strategy after comparing the severity to improve future epidemic prevention efficacy.

COVID-related vaccine supplies increased throughout 2021, but it will take months to get enough vaccines for everyone who needs them; existing vaccines must be distributed to priority groups until there is a sufficient supply for all [37]. In this study, the clustering variables include the number of confirmed cases, the number of deaths, PCR tests used, and the number of respirators. In addition to determining possible priorities through the clustering results, other external information can also be used to coordinate vaccine allocation. for example, to older age groups, front-line workers, or vaccine distribution models [38,39]. Wingert et al. [40] used machine learning methods to establish

a cluster analysis of regional severity, which may provide an alternative perspective for future vaccine allocation by using multivariate analysis of the severity of risk factors to allocate vaccine.

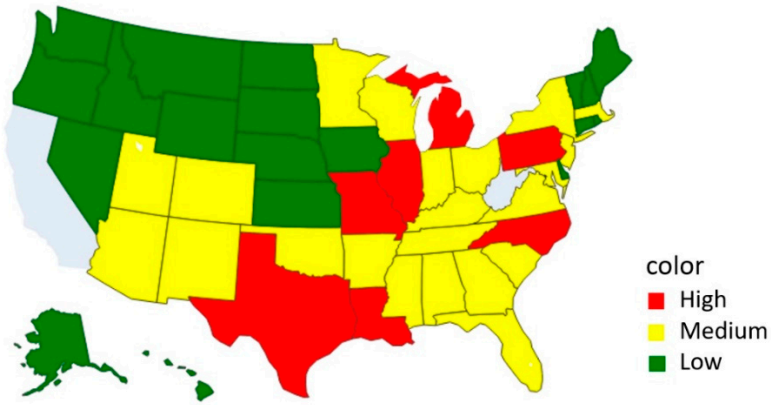

Figure 5. Clustering results from Cascade K-means.

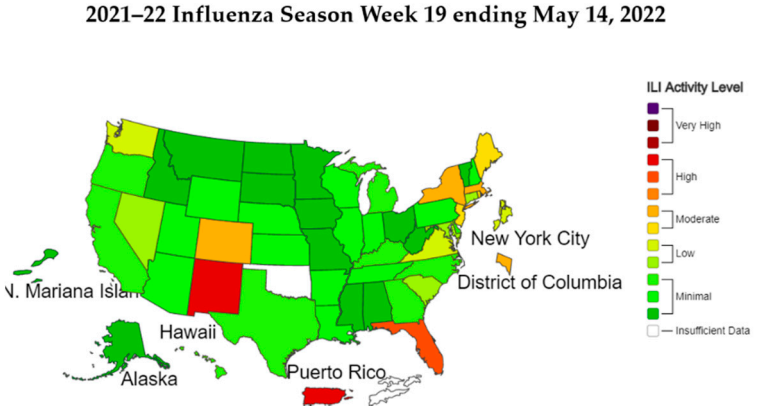

Figure 6. Influenza activity according to the CDC.

6. Conclusions

COVID-19 has spread and ravaged worldwide since December 2019, but in the case of vaccine shortage, how to distribute vaccines is a critical issue. This study uses machine learning technique combined with the medical indicators of the COVID Tracking Project to perform a cluster analysis of various states in the United States to distinguish the severity of COVID-19 infection. The dataset of this study was collected from 2020 to 2021. After features selection, and clustering methods, and then through the classification method to verify the data, the verification results show that the clustering results of Cascade K-means are better, and the classification accuracy is the highest. This study also marked the three clusters of the analysis results of US states as high, medium, and low infection so that policymakers can better objectively decide which states should prioritize vaccine allocation to prevent the epidemic from continuing to expand in the event of a vaccine shortage. It is hoped that, if there is a similar disease pandemic in the future, relevant policymakers can use the procedure of this study to allocate relevant medical resources according to the severity of infection in each state to prevent the spread of infection and the loss of more lives. The clustering results from this study can also be combined with other external information to shape a more careful vaccine distribution plan, helping back decision makers in the event of a potential future outbreak.

Author Contributions: Conceptualization, P.-L.S. and M.-H.S.; data curation, C.-J.L.; formal analysis, T.-W.W. and C.-J.L.; funding acquisition, D.-H.S. and P.-L.S.; investigation, P.-L.S.; methodology, D.-H.S.; project administration, D.-H.S.; resources, T.-W.W. and M.-H.S.; supervision, D.-H.S.; vali-

dation, M.-H.S.; visualization, P.-L.S.; writing—original draft, C.-J.L.; writing—review and editing, D.-H.S., T.-W.W. and M.-H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Taiwan Ministry of Science and Technology (grants MOST 110-2410-H-224-010). The funder has no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. CDC COVID-19 Response Team; Bialek, S.; Boundy, E.; Bowen, V.; Chow, N.; Cohn, A.; Dowling, N.; Ellington, S.; Gierke, R.; Hall, A.; et al. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, 12 February–16 March 2020. *Morb. Mortal. Wkly. Rep.* **2020**, *69*, 343–346.
2. Jit, M.; Jombart, T.; Nightingale, E.S.; Endo, A.; Abbott, S.; Edmunds, W.J. Estimating number of cases and spread of coronavirus disease (COVID-19) using critical care admissions, United Kingdom, February to March 2020. *Eurosurveillance* **2020**, *25*, 2000632. [CrossRef]
3. Chen, S.I.; Wu, C.Y.; Wu, Y.H.; Hsieh, M.W. Optimizing influenza vaccine policies for controlling 2009-like pandemics and regular outbreaks. *PeerJ* **2019**, *7*, e6340. [CrossRef]
4. Kurbucz, M.T. A joint dataset of official COVID-19 reports and the governance, trade and competitiveness indicators of World Bank group platforms. *Data Brief* **2020**, *31*, 105881. [CrossRef]
5. Liu, T.; Gong, D.; Xiao, J.; Hu, J.; He, G.; Rong, Z.; Ma, W. Cluster infections play important roles in the rapid evolution of COVID-19 transmission: A systematic review. *Int. J. Infect. Dis.* **2020**, *99*, 374–380. [CrossRef] [PubMed]
6. Álvarez, J.D.; Matias-Guiu, J.A.; Cabrera-Martín, M.N.; Risco-Martín, J.L.; Ayala, J.L. An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. *BMC Bioinform.* **2019**, *20*, 1–12. [CrossRef]
7. Hasegawa, M.; Sakurai, D.; Higashijima, A.; Niiya, I.; Matsushima, K.; Hanada, K.; Kuroda, K. Towards automated gas leak detection through cluster analysis of mass spectrometer data. *Fusion Eng. Des.* **2022**, *180*, 113199. [CrossRef]
8. Trelohan, M.; François-Lecompte, A.; Gentric, M. Tourism development or nature protection? Lessons from a cluster analysis based on users of a French nature-based destination. *J. Outdoor Recreat. Tour.* **2022**, *39*, 100496. [CrossRef]
9. Dzuba, S.; Krylov, D. Cluster analysis of financial strategies of companies. *Mathematics* **2021**, *9*, 3192. [CrossRef]
10. Ghavidel, M.; Azhdari, S.M.H.; Khishe, M.; Kazemirad, M. Sonar data classification by using few-shot learning and concept extraction. *Appl. Acoust.* **2022**, *195*, 108856. [CrossRef]
11. Tepe, C.; Demir, M.C. *Real-Time Classification of EMG Myo Armband Data Using Support Vector Machine*; IRBM: Pomezia, Italy, 2022.
12. Dritsas, E.; Trigka, M. Stroke risk prediction with machine learning techniques. *Sensors* **2022**, *22*, 4670. [CrossRef]
13. Huang, W.; Wu, Q.; Chen, Z.; Xiong, Z.; Wang, K.; Tian, J.; Zhang, S. The potential indicators for pulmonary fibrosis in survivors of severe COVID-19. *J. Infect.* **2021**, *82*, e5–e7. [CrossRef]
14. Zhou, F.; Yu, T.; Du, R.; Fan, G.; Liu, Y.; Liu, Z.; Cao, B. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **2020**, *395*, 1054–1062. [CrossRef]
15. Zheng, Z.; Peng, F.; Xu, B.; Zhao, J.; Liu, H.; Peng, J.; Tang, W. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *J. Infect.* **2020**, *81*, e16–e25. [PubMed]
16. Pan, T.; Fang, K. An effective information support system for medical management: Indicator based intelligence system. *Int. J. Comput. Appl.* **2010**, *32*, 119–124.
17. Chang, S.J.; Hsiao, H.C.; Huang, L.H.; Chang, H. Taiwan quality indicator project and hospital productivity growth. *Omega* **2011**, *39*, 14–22. [CrossRef]
18. Mainz, J.; Krog, B.R.; Bjørnshave, B.; Bartels, P. Nationwide continuous quality improvement using clinical indicators: The Danish National Indicator Project. *Int. J. Qual. Health Care* **2004**, *16* (Suppl. 1), i45–i50. [CrossRef]
19. Medlock, J.; Galvani, A.P. Optimizing influenza vaccine distribution. *Science* **2009**, *325*, 1705–1708. [CrossRef]
20. Enayati, S.; Özaltın, O.Y. Optimal influenza vaccine distribution with equity. *Eur. J. Oper. Res.* **2020**, *283*, 714–725. [CrossRef]
21. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
22. Ramesh, G.; Madhavi, K.; Reddy, P.D.K.; Somasekar, J.; Tan, J. Improving the accuracy of heart attack risk prediction based on information gain feature selection technique. *Mater. Today Proc.* **2022**, in press. [CrossRef]
23. Karegowda, A.G.; Manjunath, A.S.; Jayaram, M.A. Comparative study of attribute selection using gain ratio and correlation based feature selection. *Int. J. Inf. Technol. Knowl. Manag.* **2010**, *2*, 271–277.

24. Na, S.; Xumin, L.; Yong, G. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In Proceedings of the 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China, 2–4 April 2010; pp. 63–67.
25. Desgraupes, B. Clustering indices. *Univ. Paris Ouest-Lab Modal'X* **2013**, *1*, 34.
26. Shi, T.; Horvath, S. Unsupervised learning with random forest predictors. *J. Comput. Graph. Stat.* **2006**, *15*, 118–138. [CrossRef]
27. Wang, S.C. Artificial neural network. In *Interdisciplinary Computing in Java Programming*; Springer: Boston, MA, USA, 2003; pp. 81–100.
28. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2.1* **2012**, *2*, 86–97. [CrossRef]
29. Rani, Y.; Rohil, H. A study of hierarchical clustering algorithm. *Int. J. Inf. Comput. Technol.* **2013**, *3*, 1115–1122.
30. Campello, R.J.; Kröger, P.; Sander, J.; Zimek, A. Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1343. [CrossRef]
31. Chen, Z.; Ji, H. Graph-based clustering for computational linguistics: A survey. In Proceedings of the TextGraphs-5-2010 Workshop on Graph-Based Methods for Natural Language Processing, Uppsala, Sweden, 16 July 2010; pp. 1–9.
32. Somasekar, H.; Naveen, K. Text Categorization and graphical representation using Improved Markov Clustering. *Int. J.* **2018**, *11*, 107–116. [CrossRef]
33. Kameshwaran, K.; Malarvizhi, K. Survey on clustering techniques in data mining. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 2272–2276.
34. Mitra, P.; Murthy, C.A.; Pal, S.K. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 301–312. [CrossRef]
35. Uğuz, H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowl-Based Syst.* **2011**, *24*, 1024–1032. [CrossRef]
36. Zekić-Sušac, M.; Has, A.; Knežević, M. Predicting energy cost of public buildings by artificial neural networks, CART, and random forest. *Neurocomputing* **2021**, *439*, 223–233. [CrossRef]
37. Reilev, M.; Kristensen, K.B.; Pottegård, A.; Lund, L.C.; Hallas, J.; Ernst, M.T.; Thomsen, R.W. Characteristics and predictors of hospitalization and death in the first 11 122 cases with a positive RT-PCR test for SARS-CoV-2 in Denmark: A nationwide cohort. *Int. J. Epidemiol.* **2020**, *49*, 1468–1481. [CrossRef]
38. Swift, M.D.; Sampathkumar, P.; Breeher, L.E.; Ting, H.H.; Virk, A. Mayo Clinic's multidisciplinary approach to Covid-19 vaccine allocation and distribution. *NEJM Catal. Innov. Care Deliv.* **2021**, *2*, 1–9.
39. Bertsimas, D.; Ivanhoe, J.; Jacquillat, A.; Li, M.; Previero, A.; Lami, O.S.; Bouardi, H.T. Optimizing vaccine allocation to combat the COVID-19 pandemic. *medRxiv* **2020**. [CrossRef]
40. Wingert, A.; Pillay, J.; Gates, M.; Guitard, S.; Rahman, S.; Beck, A.; Hartling, L. Risk factors for severity of COVID-19: A rapid review to inform vaccine prioritisation in Canada. *BMJ Open* **2021**, *11*, e044684. [CrossRef]