Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2021, 2021:75

https://doi.org/10.28919/cmbn/6128

ISSN: 2052-2541

LOOKING FOR THE LINK BETWEEN THE CAUSES OF THE COVID-19 DISEASE USING THE MULTI-MODEL APPLICATION

PRASNURZAKI ANKI[1], ALHADI BUSTAMAM[1,2,*], RINALDI ANWAR BUYUNG[1]

[1]Department of Mathematics, Universitas Indonesia, Depok, Indonesia

[2]Data Science Centre, Universitas Indonesia, Depok, Indonesia

Abstract: Based on various reports of COVID-19 caused by a coronavirus, this disease causes numerous problems. There are many factors that can contribute to the rapid spread of COVID-19. In this study, multi-model application is used to identify the association among various variables that cause COVID-19. The use of programmes, artificial intelligence and input data that will be processed as a symptom of the COVID-19 disease, whose data processing uses a multi-model that will classify the variables causing COVID-19. The output data will be presented in a tabular form, so that the readers can easily understand the results. The best result obtained by this programme has an accuracy of 0.979767 using a decision tree with 0.3 of the total dataset as the test data. It is expected that the rate of spread of COVID-19 can be suppressed by detecting the disease using artificial intelligence. In the future, it is expected that a multidisciplinary collaboration between medical science and mathematics can be established with the help of this programme. This enables the authenticity of the research, and the model's performance be tested based on appropriate data in the field.

Keywords: COVID-19; multi model; variables; programme.

2010 AMS Subject Classification: 92C50, 93A30.

*Corresponding author

E-mail address: alhadi@sci.ui.ac.id

Received May 26, 2021

1. INTRODUCTION

COVID-19 has not only become a serious health problem to the society, but has also caused political and economic crisis in the affected countries. Thus, it is considered to be the greatest global public health threat today. The indicators of injustice and deprivation of social progress are also reflected in COVID-19 [1]. In the first reference, this paper discusses the COVID-19 pandemic that is occurring globally, the association between COVID-19 and the economy, as well as COVID-19 and the global environment, and the global strategy and control of COVID-19.

In this study, the first day used, namely January 22nd, 2020, was used because of the least number of detected cases, while April 4th, 2020, was used as the last day. In collecting Excel 2019 was used for the collection and integration of data over time the series in this discussion is used Excel 2019, while April 4th, 2020 was used as the last day. Then, to provide a consistent grouping of different countries against active, active cases in population and also developed in this area, its algorithms are used [2]. In the second reference, this paper analyses the grouping of countries based on the COVID-19 datasets, with time series periods. In addition, it explains the details related to the experimental designs, materials, and methods to be used in research.

In a brief news release, the General Director of the World Health Organization (WHO) stated that after 2 weeks, the number of cases in China has increased by 13-fold, whereas that in other countries that have already been affected by COVID-19 has tripled. Furthermore, he said, WHO is very concerned about the high level of spread let alone the alarming severity accompanied by inaction that is also alarming, so the action on efforts to combat the virus must be taken immediately by these countries [3]. In the third reference, this paper discusses what COVID-19, is the risk factors for COVID-19, the COVID-19 transmission from COVID-19, and how the ways to prevent COVID-19.

An online dashboard created by the Center for Systems Science and Engineering (CSSE) of Johns Hopkins University was used to present the official daily counts of COVID-19 cases and deaths. This online dashboard was made available to the public on January 22nd, giving an overview of the location and amount of confirmed cases, death and recovery of COVID-19 in all

affected countries. It was also used to galvanise the availability of researchers, public health, authorities and friendly tools used in the community, when users want to track outbreaks as they unfold. All data collected from this ashboard are presented in the Google Sheet, after which it is through the GitHub repository witha layer of dashboard features included in Esri Living Atlas [4].

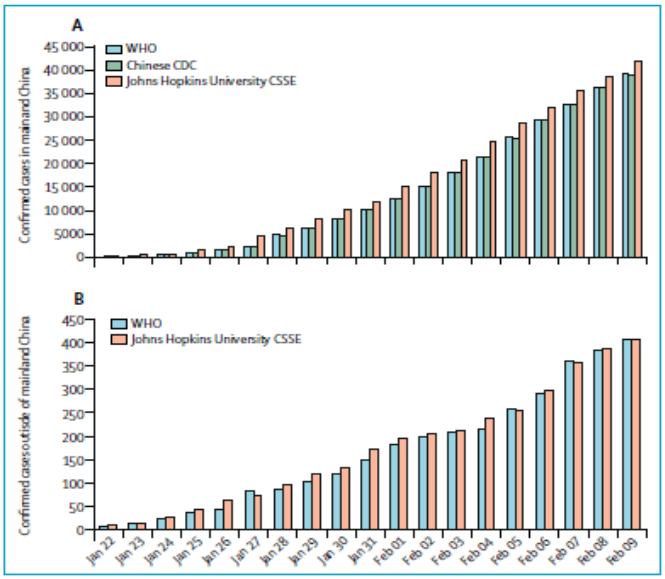

FIGURE 1. Comparison of the COVID-19 cases reported by different sources [4].

Figure 1, presents the report of the cumulative number of daily cases (starting from January 22, 2020), by of the CSSE of John Hopkins University and also the reports by the WHO and the Chinese Centers for Disease Control and Prevention (CDC China) [4]. In the fourth reference, this paper discusses interactive web dashboards used to directly track COVID-19 cases from multiple sources to verify which data is correct by comparing the data collection methods of each source.

The first case of COVID-19 occurred in China in December 2019, when individual contracted pneumonia with unknown causes from the seafood market in Wuhan. Then with the study, the corona beta virus found based on samples from patients with pneumonia symptoms. Where is the novel coronavirus will be isolated using epithelial cells located in the human airway named 2019-nCoV. Furthermore, in December 2019 and January 2020 are identified from patients in Wuhan hospital, China is the novel CoV (2019-nCoV). COVID-19 is characterised by the presence of fluid in the lungs in the infected patients' alveolar-lavage broncho fluid from the patients by sorting

from genome, culture and PCRpolymerase chain. This disease was originally called novel coronavirus-infected pneumonia (NCIP) [5]. In genaral coronaviruses infection in human can lead into severe acute respiratory illness [6].

2. Discussion of the Models and Methods

This session will discuss the characteristics of the dataset, model explanation, and evaluation method.

2.1. Characteristics of the Dataset

The dataset used in this study was obtained from India. It consists of the overall symptoms and COVID-19 test results, which are related to symptoms and the presence of COVID-19, which contains possible symptoms based on various WHO guidelines sourced from WHO Coronavirus Symptoms and AIIMS (All India Institute of Medical Science) [7]. Risk factors and the percentage level of possible exposure to COVID-19 are the dataset used in this study. Weight Criteria for Overseas Travel 8% Sore throat 10% Fever 10% Dry Cough 10% Headache 4% Fatigue 2% Gastrointestinal problems 1% Runny nose 5% Wear face mask 2% Sanitary products that are worth buying from supermarkets 2% Asthma 4% Chronic Lung Disease 6% Diabetes 1% Hypertension 1% Heart Disease 2% Contact with patients with COVID-19 8% Both you and your family attend Mass Gatherings 6% Visit public places such as malls, temples, etc. 4% Work in crowded places such as markets, hospitals 4% Shortness of breath 10%.

Many categories from various levels have stated from sources [8], which will then continue to monitor existing data and determine whether the symptoms are suggestive of COVID-19 through data processing using machine learning. Dry cough is the most common symptom of COVID-19, whereas headache is less common. Conversely, the most serious symptom is shortness of breath or difficulty breathing.

2.2. Model Explanation

2.2.1. SVM

The simplest that can be implemented in hardware because of its simple form of mathematical modelling it is called Support Vector Machine (SVM). In the reference of this study, we proposed the implementation of the SVM multi-class classifier in the asynchronous paradigm. SVM is a machine learning algorithm that is based on the theory of statistical learning and structural risk minimisation [9].

The m-dimension feature is used to map the first time of input vector $X = (x_1, x_2, \ldots, x_n)$ to the $k(x,x_i)$ kernel function; xiÞ, and then it is formed in a feature space that can be used for prediction by linear models. The description is the form of the SVM structure as depicted in Figure 2. The linear model is

$$y = wx + b \quad (1)$$

Where in the equation (1) $w = (w_1, w_2, \ldots, w_n)$, is the weight vector, b is the bias term. In Figure 2, the supporting vectors number is denoted by m. The support vector consists of subsets extracted from the training data using an algorithm. The model structure is only related to this support vector [9].

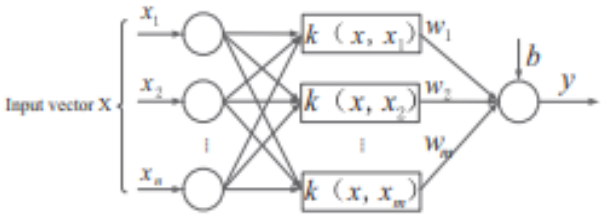

FIGURE 2. Structure of the SVM [9].

2.2.2. KNN

Based on [10], The K-Nearest Neighbor (KNN) algorithm can be used as an advanced method to find solutions to the classification problems. The scalability problem inherent in many existing

data mining methods can be solved using an algorithm. It can be handled by processing large amounts of training data, so that it can match the allocated memory. The KNN classifier applies the Euclidean distance or the value of the cosine similarity to look for differences in the tuple training and tuple test. It can be explained by the Euclidean distance between tuples x_i and x_t (t = 1,2,3 ... n) as follows:

$$d(y_i, y_t) = \sqrt{(y_{i1} - y_{it})^2 + (x_{i1} - x_{it})^2 + (x_{is} - y_{ts})^2} \quad (2)$$

where y_i, n and s are the constants of the tuples are, and they are numbers and their respective constraints. And can be expressed as follows:

$$Dist(y_1, y_2) = \sqrt{\sum_{i=1}^{n} (y_{1i} - y_{2i})^2} \quad (3)$$

The KNN algorithm is the basis of making this equation, where is at each point the closest neighbour to tuple test, it will be encapsulated in the space closest to tuple test.

2.2.3. Linear Regression

Linear regression models are often used to explore the relationship between sustainable outcome and independent variable [11]. To describe the relationship between scalar response Y and covariate X in the q dimension, the linear regression model can be used owing to its well-developed theory and ease of interpretation. This is a common linear regression model:

$$Y = \emptyset^T(X)\beta + \varepsilon \quad (4)$$

where $\emptyset$ is the known smooth function with the dimension p, β is the unknown parameter to be estimated and ε is the term error [12].

Each univariate analysis is carried out processing of the linear regression model, that is used with the aim of showing how much each independent variable will be able to predict by the dependent variable [13].

2.2.4. Logistic Regression

To isolate the effects and demonstrate which variables explain the variability that lies between accidents, the logistic regression model is used [14]. Based on [15], machine learning algorithms are employed to solve classification problems by dividing observations into different classes called logistic regression. Transferring data from any value to a value between 0 and 1, so it can be used

the sigmoid activation function, which functions to map different predictors on probability. It can be expressed as follows:

$$f(x) = \frac{1}{1+e^{-i(x)}} \quad (5)$$

For example, if there are two classifications, where no injury is 0 and the other is 1, then output can be rounded into two classes, namely, 0 and 1. Instead, of the mean squared error, which has been used for continuous responses, the cross-entropy cost function or log loss has been used for logistic regression. The cost functions for y = 1 and y = 0 can be expressed as follows:

$$cost(h_0(x), y) = -log\big(h_0(x)\big)\, If\ y = 1 \quad (6)$$

$$cost(h_0(x), y) = -log\big(1 - h_0(x)\big)\, If\ y = 0 \quad (7)$$

where x denotes the value of different predictors, y denotes the response and m denotes the number of observations.

2.2.5. Random Forest

Random forest is an algorithm aimed at constructing many parallel decision trees. The basic form of random forest learning is represented by these trees, and the characteristics are as follows:

• A different bootstrap sample from the data set is used to construct each tree, a mechanism called bagging. A bootstrap sample different from the dataset is used to construct each tree, mechanism referred to as bagging.

• The candidates for each separation are selected from the nodes on a number of randomsample variables (mtry). Furthermore, the best split point is selected from this set of random variables, a process called feature sampling. Before growing a forest, the "‘mtry"’ value must be set [16].

A considerable time is required when processing a large amount of data [17]. The random forest algorithm is employed to construct multiple tree models forming a random tree with the use of a single processor [18]. Thus, a parallel computation with a random forest design is proposed [19]. For future work, it is expected to improve the algorithm uses a sophisticated computing environment to process jobs in parallel [20].

2.2.6. Naive Bayes

Naive Bayes is a well-known discriminant reference in this type, so this marginal distribution allows for more flexibility, but this will result in completely any dependency between the X_j and X_l variables being ignored. Naive Bayes, which works with both variables for discrete and continuous data, expressed as follows:

$$P_f(D = k|X = x) = \prod_{j=1}^{d} P_{f_{(j)}}(D = k|X_j = x_j) \quad (8)$$

The marginal distribution of Xj is denoted by the $f_{(j)}$. Even though the property in (8) is not fulfilled, this estimate can work very well. Furthermore, parametric and non-parametric, the marginal distribution of (8) can be estimated, in both cases avoiding dimensional curses. In the case of identical population mean and covariance the naive Bayes can be used [21].

2.2.7. Decision Tree

The decision tree is other technique for predicting and mining data. It is used for the classification, grouping and prediction tasks, especially for solving classification problems [22]. To classify data in a discrete form, a tree-structured algorithm is employed, which is called a decision tree. The root nodes, leaves, internal nodes and branches are the contents of the standard tree. The series of vertices from the root to the leaf is represented by each branch, and an attribute is represented by each node. The main objective of the decision tree is to disclose the structural information contained there [23].

2.2.8. Gradient Boosting Regression

Random forest is a random collection of basic regression tree. If gradient boosting is like random forest, then it is a set of decision tree. In boosting, each new tree (Tree 2) matchesaa modified version of the original dataset (Tree 1). The idea is to improve the prediction of the first tree [24]. In this study, a gradient-enhancing regression method was employed to develop models that predict the thermal conductivity of soils [25]. Based on these two references, it can be deduced that the application of gradient boosting regression as a model in the programme to make predictions related to variables with COVID-19 can be tested in the programme. The name changes adjust from gradient boosting regression to gradient boosting regressor.

2.3. Evaluation Method

The following is a measurement reference to the evaluation method aimed at comparing the results of the model prediction with the original data based on [26]. In Table 1, a label will be assigned to each cell, as a collection of possible outcomes. For example, in the case of spam detection, true positive is a document that is recognised by the system as a spam (note in the Contingency Table) and is indeed a spam. False negative is a document that is recognized by the system as not a spam but is indeed a spam. True negative is a document that is recognised by the system as not a spam and is indeed not a spam. Accuracy can be expressed as follows:

$$accuracy = \frac{tp+tn}{tp+fp+tn+fn} \quad (9)$$

TABLE 1. Contingency Table [26].

		gold standard labels	
		system positive	*system negative*
system output labels	*system positive*	*true positive*	*false positive*
	system negative	*false negative*	*true negative*

In this research, we determined whether the prediction results of various models are in agreement with the original data in order to identify the most suitable model for classifying the cause of a person suffering from COVID-19 with the highest accuracy value.

3. Designing Programme Structures Based On The Predetermined Components Of The Programme

In this section, we describe the problem that needs to be solved, and the steps that need to be taken to create a programme.

3.1. Description of the Problem to Be Solved

Currently, COVID-19 is rapidly spreading worldwide and has dramatically impacted public health [27]. The COVID-19 pandemic has forced state, local and territorial public health agencies to take measures in order to protect and secure the health of the people [28]. The effects of

potentially unexpected on health results are due to overwhelming evidence that COVID19 and various levels that associated with non-communicable diseases [29]. Thus, monitoring the symptoms suggestive of COVID-19 is important to overcome the massive spread COVID-19.

3.2. The Steps to Create a Programme

The following are the steps that need to be taken to create a programme:

1. Create files based on different types of test data sizes.
2. Import command that will be used in the programme.
3. Input data to be used in the programme.
4. Check the characteristics of the data so ensure i ts compatibility with the research design that has been created.
5. Perform feature transformation on the data to be processed.
6. Re-check the contents of the data to be used.

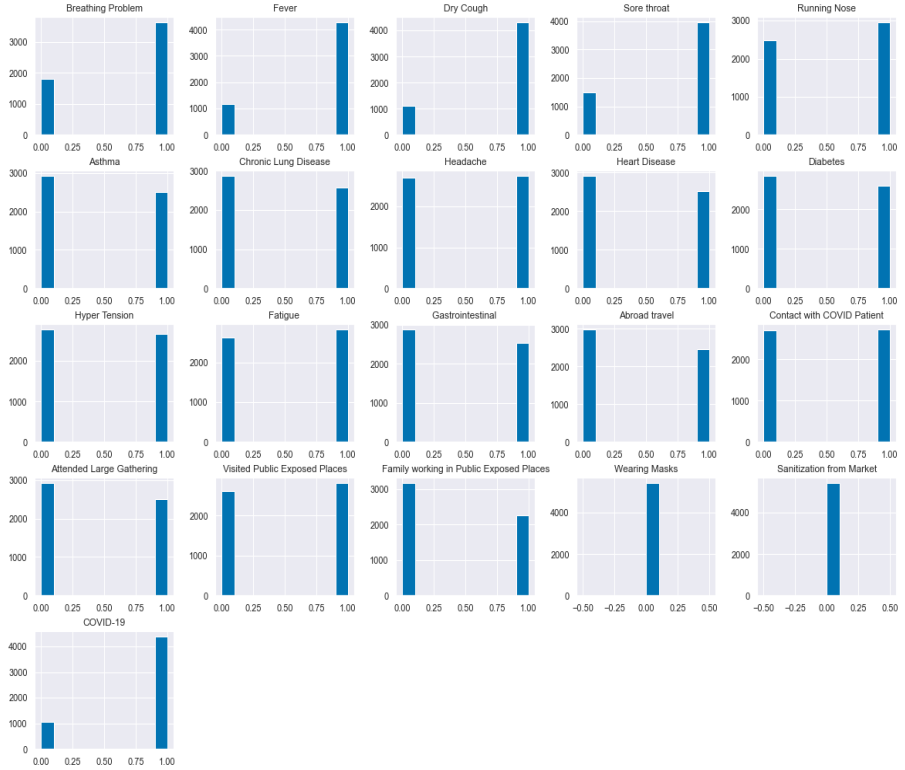

FIGURE 3. Results of the Checking of the Data Contents.

7. Eliminate monotonous data in certain features to maintain the diversity of the processed data.

8. Determine the correlation between the symptoms of a person with COVID-19 and thepossibility of a person contracting COVID-19.

9. Perform feature selection by removing features that have a relatively low correlation value in some features.

	Breathing Problem	Fever	Dry Cough	Sore throat	Hyper Tension	Abroad travel	Contact with COVID Patient	Attended Large Gathering	Visited Public Exposed Places	Family working in Public Exposed Places	COVID-19
Breathing Problem	1.000000	0.089903	0.159562	0.303768	0.045256	0.117795	0.214634	0.200304	0.066688	0.018295	0.443764
Fever	0.089903	1.000000	0.127580	0.322235	0.079001	0.128726	0.164704	0.070490	0.002252	0.012102	0.352891
Dry Cough	0.159562	0.127580	1.000000	0.213907	0.081989	0.331418	0.128330	0.117963	0.086176	0.163102	0.464292
Sore throat	0.303768	0.322235	0.213907	1.000000	0.042811	0.205986	0.189251	0.216438	0.079055	0.104378	0.502848
Hyper Tension	0.045256	0.079001	0.081989	0.042811	1.000000	-0.016382	0.027307	0.002911	0.019174	0.048152	0.102575
Abroad travel	0.117795	0.128726	0.331418	0.205986	-0.016382	1.000000	0.080210	0.113399	0.069609	0.143094	0.443875
Contact with COVID Patient	0.214634	0.164704	0.128330	0.189251	0.027307	0.080210	1.000000	0.234649	0.079800	0.006909	0.357122
Attended Large Gathering	0.200304	0.070490	0.117963	0.216438	0.002911	0.113399	0.234649	1.000000	0.083795	0.063776	0.390145
Visited Public Exposed Places	0.066688	0.002252	0.086176	0.079055	0.019174	0.069609	0.079800	0.083795	1.000000	0.028486	0.119755
Family working in Public Exposed Places	0.018295	0.012102	0.163102	0.104378	0.048152	0.143094	0.006909	0.063776	0.028486	1.000000	0.160208
COVID-19	0.443764	0.352891	0.464292	0.502848	0.102575	0.443875	0.357122	0.390145	0.119755	0.160208	1.000000

FIGURE 4. Correlation Matrix Based on the Selected Features Derived from the Feature Selection Results.

10. Determine the data that will be used as the dependent variable and the independent variable.

```
      Breathing Problem  Fever  Dry Cough  Sore throat  Hyper Tension  \
0                     1      1          1            1              1
1                     1      1          1            1              0
2                     1      1          1            1              0
3                     1      1          1            0              0
4                     1      1          1            1              1
...                 ...    ...        ...          ...            ...
5429                  1      1          0            1              0
5430                  1      1          1            0              1
5431                  1      1          1            0              1
5432                  1      1          1            0              0
5433                  1      1          1            0              1

      Abroad travel  Contact with COVID Patient  Attended Large Gathering  \
0                 0                           1                         0
1                 0                           0                         1
2                 1                           0                         0
3                 1                           0                         1
4                 0                           1                         0
...             ...                         ...                       ...
5429              0                           0                         0
5430              0                           0                         0
5431              0                           0                         0
5432              0                           0                         0
5433              0                           0                         0

      Visited Public Exposed Places  Family working in Public Exposed Places
0                                 1                                        1
1                                 1                                        0
2                                 0                                        0
3                                 1                                        0
4                                 1                                        0
...                             ...                                      ...
5429                              0                                        0
5430                              0                                        0
5431                              0                                        0
5432                              0                                        0
5433                              0                                        0

[5434 rows x 10 columns]
```

FIGURE 5. Independent Variables Used in the Programme.

```
0       1
1       1
2       1
3       1
4       1
       ..
5429    1
5430    1
5431    0
5432    0
5433    0
Name: COVID-19, Length: 5434, dtype: int32
```

FIGURE 6. Dependent Variable Used in the Programme.

Based on the data in Figures 5 and 6, a value of 1 means yes, and the value of 0 means no.

11. Conduct testing on eight predetermined models using the accuracy evaluation method.
12. Sort the test results of the eight models to determine which model yields the best test results.

4. RESULTS AND DISCUSSION

4.1. Describe the Variations in the Test Data Used in This Study

Based on the variation of the sample tested in [30], where the sample size tested was 10% and 20% or 0.1 and 0.2 of the total data. In this study, the data to be tested had several variations of 0.1; 0.2; and 0.3 of the total data used.

TABLE 2. Table of File Names and a Description of the Details in Each File.

Test Size	File No.	File Name
0.1	1	Testing Between Symptoms and the Presence of COVID- 19 Using a Multi-Model Test With Simulation 1
0.2	2	Testing Between Symptoms and the Presence of COVID- 19 Using a Multi-Model Test With Simulation 2
0.3	3	Testing Between Symptoms and the Presence of COVID- 19 Using a Multi-Model Test With Simulation 3

4.2. Results Which Obtained From Testing the Data on the Program

In this sub-section, we discuss the results of data testing using the programme.

TABLE 3. The Results of Programme Trials Conducted on Three Different Files.

File No. 1		File No. 2		File No. 3	
Model	Accuracy	Model	Accuracy	Model	Accuracy
Decision Tree	0.970588	Decision Tree	0.973321	Decision Tree	0.979767
SVM	0.961397	SVM	0.966881	SVM	0.972410
KNN	0.957721	KNN	0.966881	KNN	0.971183
Linear Regression	0.954044	Linear Regression	0.962282	Linear Regression	0.966891
Logistic Regression	0.954044	Logistic Regression	0.962282	Logistic Regression	0.966891
Random Forest	0.890994	Random Forest	0.897034	Random Forest	0.908178
Gradient Boosting Regressor	0.862832	Gradient Boosting Regressor	0.871956	Gradient Boosting Regressor	0.875629
Naive Bayes	0.753676	Naive Bayes	0.738730	Naive Bayes	0.735745

Based on the results of program trials from various models that have been tested, it can be concluded that in this study , the most effective model is the decision tree with File No. 3 with test size of 0.3, with an accuracy of 0.979767.

Compared with other classification techniques, based on [31], the decision tree files are faster and more accurate. In [32], concluded that a larger sample size leads to a more precise estimate. Based on these two statements, it can be concluded that the results obtained in this study, with a test data size of 0.3, which is greater than 0.1 and 0.2, and using the decision tree model that is better than the other models, are in accordance with the directed reference.

5. CONCLUSION

The conclusion obtained in this study is to examine various symptoms related to COVID-19, based on that the best model used which is based on a comparison of 8 models that have been tested, namely Decision Tree, SVM, KNN, Linear Regression, Logistic Regression, Random Forest, Gradient Boosting Regressor Naive Bayes that has been done in the existing programmes in this study is a decision tree, using a data set size of 0.3 of the total dataset 0.3 in the form of an accuracy of 0.979767.

ACKNOWLEDGEMENTS

This research is partially supported by PTUPT 2020 research grant by RISTEKDIKTI with contract number NKB-325/UN2.RST/HKP.05.00/2020. The authors are grateful for the support of members of the Bionformatics and Advanced Computing Laboratory (BACL) in the Mathematics and Data Science Center (DSC) Department at the Faculty of Mathematics and Natural Sciences, University of Indonesia. Our special thanks to Enago (www.enago.com) for the English review of this paper.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] I. Chakraborty, P. Maity, COVID-19 outbreak: Migration, effects on society, global environment and prevention, Sci. Total Environ. 728 (2020), 138882.

[2] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, et al., Clustering analysis of countries using the COVID-19 cases dataset, Data Brief. 31 (2020), 105787.

[3] D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic, Acta Biomed. 91(1) 2020, 157–160.

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

[5] N. Zhu, D. Zhang, W. Wang, et al., A Novel Coronavirus from Patients with Pneumonia in China, 2019, N. Engl. J. Med. 382 (2020), 727–733.

[6] A. Bustamam, E.D. Ulul, H.F.A. Hura, T. Siswantining, Implementation of hierarchical clustering using k-mer sparse matrix to analyze MERS–CoV genetic relationship, in: Depok, Jawa Barat, Indonesia, 2017: p. 030142.

[7] H. Hari, Symptoms and COVID presence, kaggle datasets, https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence. (August, 2020).

[8] World Health Organization, Coronavirus, https://www.who.int/health-topics/coronavirus#tab=tab_3. (April, 2020).

[9] J. Yue, K. Xu, W. Liu, SVM based measurement method and implementation of gas-liquid, Measurement. 145 (2020), 160-171.

[10] S. Nayak, M. Panda, G. Palai, Realization of optical ADDER circuit using photonic structure and KNN algorithm, Optik. 212 (2020), 164675.

[11] A.F. Schmidta, C. Finan, Linear regression and the normality assumption, J. Clinic. Epidemiol. 98 (2018), 146-151.

[12] X. Guo, L. Song, Y. Fang, Model checking for general linear regression with nonignorable missing response, Comput. Stat. Data Anal. 138 (2019), 1-12.

[13] A. S. Argawu, Linear Regression model for predictions of COVID-19 new cases and new deaths based on May/June data in Ethiopia, SSRN. (2020), 1-19. https://dx.doi.org/10.2139/ssrn.3680524.

[14] P. Gholizadeh, B. Esmaeili, Developing a multi-variate logistic regression model to analyze accident scenarios: case of electrical contractors, Int. J. Environ. Res. Public Health. 17(13) (2020), 4852.

[15] M. Rezapour, A. M. Molan, K. Ksaibati, Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models, Int. J. Transport. Sci. Technol. 9 (2020), 89-99.

[16] A. Callens, D. Morichon, S. Abadie, et al., Using random forest and gradient boosting trees to improve wave forecast at a specific location, Appl. Ocean Res. 104 (2020), 102339.

[17] A. Bustamam, K. Burrage, N.A. Hamilton, A GPU implementation of fast parallel Markov clustering in bioinformatics using EllPACK-R sparse data format, in: 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, IEEE, Jakarta, Indonesia, 2010: pp. 173–175.

[18] A. Bustamam, M.I.S. Musti, S. Hartomo, S. Aprilia, P.P. Tampubolon, D. Lestari, Performance of rotation forest ensemble classifier and feature extractor in predicting protein interactions using amino acid sequences, BMC Genomics. 20 (2019), 950.

[19] N. Azizah, L.S. Riza, Y. Wihardi, Implementation of random forest algorithm with parallel computing in R, J. Phys.: Conf. Ser. 1280(2) (2019), 022028.

[20] A. Bustamam, G. Ardaneswari, H. Tasman, D. Lestari, Performance evaluation of fast smith-waterman algorithm for sequence database searches using CUDA GPU-based parallel computing. J. Next Gen. Inform. Technol. 5(2) (2014), 38–46.

[21] H. Otneim, M. Jullum, D. Tjøstheim, Pairwise local Fisher and naive Bayes: Improving two standard discriminants, J. Econometrics. 216 (2020), 284-304.

[22] D. Hao, Q, Qiu, X. Zhou, et al., Application of decision tree in determining the importance of surface electrohysterography signal characteristics for recognizing uterine contractions, Biocybern. Biomed. Eng. 39 (2019), 806-813.

[23] J. Chen, Y. Lian, Y. Li, Real-time grain impurity sensing for rice combine harvesters using image processing and decision-tree algorithm, Computers Electron. Agric. 175 (2020), 105591.

[24] S. Kekre, Comparative study and proposed approach for multi-variate regression through gradient boosting, Int. Res. J. Eng. Technol. 7(5) (2020), 3640-3644.

[25] A.H. Yurttakal, Extreme gradient boosting regression model for soil thermal conductivity, Therm. Sci. 25(1) (2021), 1-7.

[26] D. Jurafsky, J.H. Martin, Speech and language processing, https://web.stanford.edu/~jurafsky/slp3/.

[27] G. A. Ramirez, M. Gerosa, L. Beretta, et al., COVID-19 in systemic lupus erythematosus: Data from a survey on 417 patients, Semin. Arthritis. Rheum. 50(5) (2020), 1150–1157.

[28] L. Kase, P. Demoly, B. Martin, et al., Allergy and coronavirus disease (COVID-19) international survey: Real-life data from the allergy community during the pandemic, World Allergy Organ. J. 14(2) (2021), 100515.

[29] M. R. Azarpazhooh, N. Morovatdar, A. Avan, et al., COVID-19 pandemic and burden of non-communicable diseases: An ecological study on data of 185 countries, J. Stroke Cerebrovasc. Dis. 29(9) (2020), 105089.

[30] B.K. Ayinde, T.J. Adejumo, G.S. Solomon, A study on sensitivity and robustness of one sample test statistics to outliers, Glob. J. Sci. Front. Res.: F Math. Dec. Sci. 16(6) (2016), 99-112.

[31] A.M. Ahmed, A. Rizaner, A.H. Ulusoy, A novel decision tree classification based on post-pruning with Bayes minimum risk, PLoS ONE. 13(4) (2018), e0194168.

[32] N. Asiamah, H.K. Mensah, E.F. Oteng-Abayie, Do larger samples really lead to more precise estimates? A simulation study, Amer. J. Educ. Res. 5(11) (2017), 9–17.