

---

# Data Mining Predicts the Need for Immunization Vaccines Using the Naive Bayes Method

R Arri Widyanto <sup>1,\*</sup>, Meidar Hadi Avizenna <sup>2</sup>, Nugroho Agung Prabowo <sup>3</sup>, Kemal Alfata <sup>4</sup>, Agus Ismanto <sup>5</sup>

Department of Information Technology, Faculty of Engineering, Universitas Muhammadiyah Magelang, Magelang, Indonesia  
arri\_w@ummgl.ac.id <sup>1,\*</sup>; meidar@unimma.ac.id <sup>2</sup>; naprabowo@ummgl.ac.id <sup>3</sup>; kemalalfata11@gmail.com <sup>4</sup>;  
agusismanto509@gmail.com <sup>5</sup>  
\* corresponding author

(Received July 4, 2021 Revised August 10, 2021 Accepted August 28, 2021, Available online September 29, 2021)

---

## Abstract

In December 2019, SARS-CoV-2 caused the coronavirus disease (COVID-19) to spread to all countries, infecting thousands of people and causing death. COVID-19 causes mild illness in most cases, although it can make some people seriously ill. Therefore, vaccines are in various phases of clinical progress, and some of them have been approved for national use. The current state of affairs reveals that there is a critical need for a quick and timely solution to the need for a Covid-19 vaccine. Non-clinical methods such as data mining and machine learning techniques can help to do this. This study will focus on US COVID-19 Vaccination Advances using Machine learning classification algorithms and Using Geospatial analysis to visualize the results. The paper's findings indicate which algorithm is better for a given data set. Naive Bayes algorithm is used to run tests on real world data, and is used to analyze and draw conclusions. Period of Accuracy and performance, and it was found that Naive Bayes is very superior in terms of time and accuracy.

*Keywords:* Covid-19; Geospatial Analysis; Data Mining; Naive Bayes;

---

## 1. Introduction

Nowadays, Machine Learning (ML) algorithms are used in a wide computing area because of their effective performance. In December 2019, a new series of coronavirus known as SARS-Cov-2 was detected and caused a large number of deaths from irregular respiratory diseases. Recently, the virus known as COVID-19 began to spread from Wuhan, Hubei Province, China, then spread rapidly to all countries in the world. This virus can spread between humans and become a universal pandemic. COVID-19 has shown moderate symptoms and affects people of all ages. However, evidence shows that two groups of people infect this severe virus more than others, people who are elderly, over 60 years old and people with chronic diseases such as Cardiovascular Disease, Diabetes, Chronic Respiratory Diseases, and cancer.

The death rate continues to increase worldwide caused by COVID-19, therefore many countries are forced to take quick actions such as: wearing masks, social distancing, lockdowns and curfews. The rapid development of vaccines to prevent COVID-19 has become a global imperative. Sufficient vaccines against COVID-19 are expected to be produced by 2021 to vaccinate at least 70% of each country's population. Figures that could blunt the years of the pandemic sooner than expected if evenly distributed, according to a new analysis. Non-clinical methods such as data mining and machine learning techniques can help to do this. Data mining is used to derive rules from large amounts of data, whereas machine learning classification algorithms are data processing techniques that simplify the development of analytical models.

Today, machine learning is very important in healthcare. Machine Learning helps streamline administrative procedures in hospitals, map and manage infectious diseases, and personalize medical care. Machine Learning can also be used to provide a smart alternative for analyzing large amounts of data. By developing fast and efficient data-driven algorithms and models for real-time data processing, Machine Learning can generate accurate results and analysis. Therefore, we must increase physicians' access to additional detail so that they can make informed choices about medical diagnosis and treatment strategies while still recognizing the potential consequences and costs for all.

The value of machine learning in healthcare is its ability alone to generate large datasets beyond the scope of human capabilities, and then reliably transform the analysis of that data into clinical insights that assist clinicians in planning and delivering care, ultimately leading to better outcomes, lower treatment costs and increase patient satisfaction. For this problem, we collected a dataset on the success of the Covid-19 vaccine in America and analyzed the information using machine learning and data mining algorithms. It is a way to derive useful information from the data and suitable algorithms to make reliable predictive models from it. This research focuses on classifying machine learning algorithms and identifying the parts of the country most in need of vaccines with the greatest consistency and accuracy.

## 2. Literature Review

The healthcare system is one of the most important parts of developed countries. Nowadays the fear of COVID-19 has infected the whole world that It is associated with the rapid growth of essential data and the need to analyze the relationships and hierarchy between data leads to the need for machine learning and Data mining is in the health system that is very effective in prevention, diagnosis and treatment. COVID-19 Vaccination Progress Using Machine Learning Algorithms is the subject of these studies. This paper analyzed recent studies in this area, identified and determined the most efficient algorithm with the highest accuracy and precision.

IOP Series [17] performed an analysis in which a Deep Neural Network (DNN) was learned to distinguish between two types of datasets: serious illness. A simulation sample containing 70% of the data was used, with a research sample containing 30% of the data being used to verify the findings and quantify possible overtraining. The population data given by the Gauteng CityRegion Observatory was used to apply the DNN weights. The horizontal axis depicts the proportion of Gauteng's adult population, while the vertical axis depicts the decrease in serious disease. According to the DNN model, through vaccinating around 20% of the adult population, the incidence of serious COVID-19 disease may be decreased by over 80%. The smart algorithm's performance can be increased with further refinements such as multi-class labelling and broader datasets.

IOP Series [18] suggested a Fine-Tuned RF model also with the Adaboost algorithm as a boost. To predict the urgency of the circumstance and the probable result, death, the model uses the COVID-19 patient spatial, transport, health, and demographic details. The model has an F1 Score of 0.86 and a 94% accuracy rate. The data review reveals a strong association between patient gender and mortality, as well as the fact that the bulk of patients are between the ages of 20 and 70.

Vaccine Hesitancy (VH) has been proposed by Chen et al. [19] as a major obstacle to the next COVID-19 mass immunization program. Using area-level metrics readily accessible to policymakers, utilizing machine-learning algorithms to forecast populations at high risk of VH. Based on results from infant immunization programs for seven non mandatory vaccinations conducted in 6408 Italian municipalities in 2016, illustrate the strategy.

The region under the Receiver Operating Characteristics (ROC) curve is used to compare a collection of machine learning models. The Random Forest algorithm is shown to be the most effective in forecasting areas with a high chance of VH, improving the volatile baseline standard by 24% in terms of precision. The proportion of waste recycled, and the job rate was found to be the strongest predictors of high VH among the area-level indicators. Which will help policymakers target pro-vaccine advocacy programs at the local level.

To forecast COVID-19 vaccine needs, Watratan et al. [20] used the vaxign and the recently built-in Machine Learning based vaxign-ML reverse vaccinology methods. Vaxign-ML expected that the Spike proteins and nonstructural proteins (nsp3 and nsp8) will induce strong defensive antigenicity. The nsp3 gene was discovered to be more conserved in Sars-Cov-2, SARS-CoV-2, and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) than in 15 coronaviruses that infect humans and other species. The protein was also shown to include promiscuous Mhc-I and Mhc-Ii T-Cell epitopes, as well as linear B-cell epitopes, which were expected to be clustered on the protein surface. COVID-19 vaccine production may be competitive and secure if our expected future vaccine targets are realized.

Şengür [21] used the Naive Bayes Classification algorithm on Twitter data with the keyword 'COVID-19' indexed by the keyword 'vaccine', in Indonesian tweets in the second and third weeks of January 2021. Over that time, the study reported 39% optimistic sentiment, 56% pessimistic sentiment, and 1% favourable sentiment. Since the media did not think the vaccination was effective at the time, negative opinions were created. Further research is anticipated to use a variety of algorithms to produce more accurate outcomes in opinion analysis.

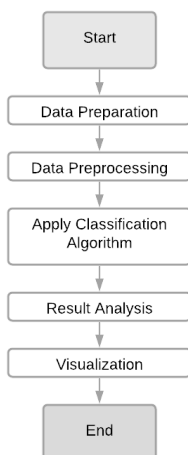
Roy et al. [22] performed research on Data Mining models for forecasting the recovery of patients infected with 2019-nCoV. The models estimate the time it will take for Covid-19 affected patients to recover, as well as when they will be discharged from isolation centers. They also determine which patients will not recover and may die as a result of the Covid-19 pandemic. The models assist health care providers in determining the recovery and stability of freshly contaminated individuals.

The models were created using the Korea Centers for Disease and Prevention (KCDC) dataset, which was used to extract instances of death and recovery reports from affected 2019- nCoV pandemic patients. Python programming and data mining techniques such as Support Vector Machine (SVM), Naive Bayes, Decision Tree, logistic regression, Random Forest (RF), and K-Nearest Neighbor (KNN) were used to create the models. The results indicate that the model developed using a Decision Tree data mining algorithm is more effective in predicting the probability of recovery of Covid-19 infected patients, with a 99.85% overall accuracy.

Forati and Ghose [23] developed two Deep Learning Classification models that use the hematochemical value from blood tests to distinguish between SARS-CoV-2 positive and negative patients, with accuracy and sensitivity varying from 82 to 86% and 92% to 95%, respectively. They've further developed a decision tree model to help doctors decide whether or not to interpret blood tests for suspected Covid-19 events (even off-line). Their findings showed that blood sample analysis and machine learning can be used to diagnose Covid-19 positive patients instead of Real-time Reverse Transcription Polymerase Chain Reaction (RRT-PCR), which is particularly useful in developing countries where RRT-PCR testing on respiratory tract specimens are scarce. They've created an open-source Web Based health reference and evaluation website.

### 3. Method

The study methodology basically consists of five main steps: Data preparation, data preprocessing, applying classification algorithms, result analysis and visualization as shown in figure 1.



**Figure. 1.** Research methodology

### 3.1. Dataset Preparation & processing

The “COVID-19 World Vaccination Progress Dataset’ is used from the Vaccine Reporting System Dataset in .CSV file format. The Vaccine dataset is an open-source archive that contains the most detailed and relevant information on COVID-19 World Vaccination Progress and Events accessible in kaggle.

There are 49 instances in the dataset we used for this study, with more than 14 attributes. Prior to implementing the technique, any processing of the dataset is usually done. Data transmission has to be improved in order to meet data consistency requirements. There are two main methods of data processing: data collection, convert nominal data to numeric data. Sample contents of the dataset that have been normalized and converted can be seen in Figure 2.

STATE	POPESTIMATE2019	lat	long	people_vaccinated	percent_vaccinated
Alabama	4903185	32.377716	-86.300568	2289141.0	0.466868
Alaska	731545	58.301598	-134.420212	386179.0	0.527895
Arizona	7278717	33.448143	-112.096962	4010968.0	0.551054
Arkansas	3017804	34.746613	-92.288986	1532747.0	0.507901
California	39512223	38.576668	-121.493629	26496378.0	0.670587
Colorado	5758736	39.739227	-104.984856	3568136.0	0.619604
Connecticut	3565287	41.764046	-72.682198	2561549.0	0.718469
Delaware	973764	39.157307	-75.519722	608891.0	0.625296
District of Columbia	705749	38.895110	-77.036370	465012.0	0.658891
Florida	21477737	30.438118	-84.281296	13174939.0	0.613423

**Figure. 2.** Dataset Sample

### 3.2. Naive Bayes Classification

The naïve Bayes model is a heavily simplified Bayesian probability model. The naïve Bayes classifier operates on a strong independence assumption. This means that the probability of one attribute does not affect the probability of the other. Given a series of n attributes, the naïve Bayes classifier makes 2nd independent assumptions. Nevertheless, the results of the naïve Bayes classifier are often correct. Naïve Bayes provides a method to predict the probability of different classes based on various attributes. This algorithm is often used in text classification and when grappling with multi-class problems. The work reported in examines the circumstances under which the naïve bayes classifier

performs well and why. It states that the error is a result of three factors: training data noise, bias, and variance. Training data noise can only be minimised by choosing good training data. The training data must be divided into various groups by the machine learning algorithm. Bias is the error due to groupings in the training data being very large. Variance is the error due to those groupings being too small.

#### 4. Results and Discussion

The previous section involves the study of each of the naïve Bayes introduced previously and implementing it on a set of COVID-19 Vaccine data. There are 14 attributes and 49 instances in the whole data collection. In this paper, this study did not use other algorithms because the other standard algorithm is not suitable for the classification of small datasets, despite its strong theoretical basis and high classification accuracy. This is because the training difficulty of the Classification algorithm is heavily dependent on the size of the dataset. For a large dataset, it is better to select the Decision Tree algorithm or Random tree algorithm if the dataset consists of more noisy data; the most suitable algorithm is the classification algorithm.

In table 1, Mean Absolute Error (MAE) or sometimes it is referred as L1 and Root Mean Square Error (RMSE) are used to measure the classification accuracy for the algorithm. MAE refers to the absolute error between predicted and observed values, while RMSE is a quadratic measure of the error between predicted and observed values. The main difference between two metrics is that RMSE provides a higher weight to large magnitude errors due to averaging the square of errors. As it is illustrated, Naïve Bayes the classifier has a large value of MAE (0.008). Naïve Bayes the classifier has a large percentage of relative absolute error (36.3292). While from Table 1 shows that the Naïve Bayes that has the lowest overall True Positive (0.697). The Precision values of the Nave Bayes classifier are also strong 0.856 respectively.

**Table 1.** Naïve Bayes Classification Result

MAE	RMSE	RAE	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
0.008	0.076	36.3292	0.697	0.002	0.856	0.697	0.689	0.705	0.958

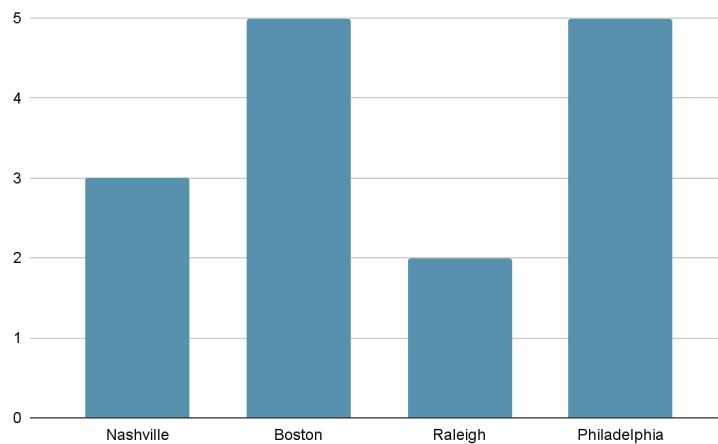
We observed that vaccine mapping approaches ranged from unadjusted to adjusted disease outcomes in space and time, with studies controlling for sex, age, occupation, and healthcare services. For example, during the early months of the epidemic, Samuel et al. [26] used a simple COVID-19 heat map approach to show the spatial patterns of the epidemic in China. Chamorro et al. [30] studied COVID-19 in disadvantaged areas of Liangshan Prefecture, China. Made et al. [34] applied geospatial visualization techniques to study the distribution of COVID-19 incidence rates in Bosnia and Herzegovina on a series of maps using ELIS (Epidemic Location Intelligence System). In another study by Chen et al. [32], spatial distribution of COVID-19 incidence was shown on a map, while the rest of the analysis was largely based on statistical analysis.

Spatial autocorrelation techniques were also used by many researchers. For example, hotspot analysis was used by Charpone et al. [37] to show the spatial intensity of infection in China. Similarly, another study carried out in China quantified the spatial distribution of COVID-19 by hotspot analysis. Likewise, Chamorro et al. [30] analyzed clusters of COVID-19 incidence in China by using global and local Moran's I in the ArcGIS environment. In the Hubei province of China, spatiotemporal analysis of COVID-19 daily cases and their incidence rate was done by Yang et al. [38], using hotspot and cluster analysis in ArcGIS. A simple ecological study by Roy et al. [22] explored spatial variation in COVID-19 incidence and mortality rates in Rio de Janeiro, Brazil, and globally. In this study, we will visualize geospatial analysis using folium. Folium is a Python library used to visualize geospatial data. It is an easy to use yet powerful library. Folium is a Python wrapper for Leaflet.js which is a leading open-source JavaScript library

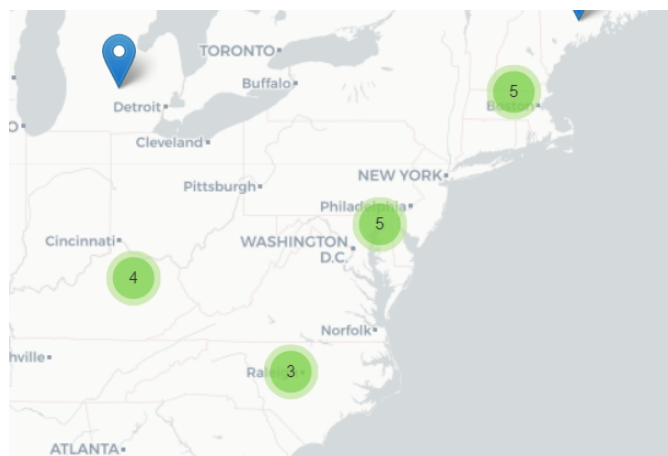
for plotting interactive maps. With the code as can be seen in figure 3 below. And the results can be seen in figures 4 and 5 below.

```
v_map = folium.Map(location=[42.32,-71.0589], tiles='cartodbpositron', zoom_start=4)
mc = MarkerCluster()
for idx, row in vaccinationAndPopulationByLocation.iterrows():
    if not math.isnan(row['long']) and not math.isnan(row['lat']):
        mc.add_child(Marker(location=[row['lat'], row['long']],
                             tooltip=str(round(row['percent_vaccinated']*100, 2))+"%"))
v_map.add_child(mc)
v_map
```

**Figure 3.** Folium code for Geospatial analysis



**Figure 4.** City with highest needs for vaccines



**Figure 5.** Geospatial analysis result

## 5. Conclusion

COVID-19 pandemic impacts millions of lives all around the world as a major public health concern. To win the fight against the COVID-19 pandemic, we'll need an efficient vaccine that can be distributed equally and broadly. Therefore, we retrieved datasets on covid-19 world vaccine progress and have processed this information using Naive bayes algorithms classification algorithms. Data mining algorithm performance assessment is important since it helps users to pick the suitable algorithm for their classification/prediction activities. The final results after applied the dataset and calculate the accuracy (The accuracy can be defined as the percentage of correctly classified instances divided by instances) for classification algorithm shows that the algorithm based on our data is the Naive bayes classifier with an accuracy of 99.9% and the total time taken to build the model is at 0.08 seconds. We also use geospatial analysis to visualize and predict city that need vaccines the most.

## References

- [1] Z. Reno, S. Elsi, H. Pratiwi, Y. Efendi, and R. Rusdina, "Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia," 2020, doi: 10.1088/1742-6596/1594/1/012007.
- [2] C. N. Poth, O. Bulut, A. M. Aquilina, and S. J. G. Otto, "Using Data Mining for Rapid Complex Case Study Descriptions : Example of Public Health Briefings During the Onset of the COVID-19 Pandemic," 2021, doi: 10.1177/15586898211013925.
- [3] C. Zhang, S. Xu, Z. Li, and S. Hu, "Understanding Concerns , Sentiments , and Disparities Among Population Groups During the COVID-19 Pandemic Via Twitter Data Mining : Large-scale Cross-sectional Study Corresponding Author :," vol. 23, no. 3, pp. 1–16, 2021, doi: 10.2196/26482.
- [4] C. Zhang, J. Jiang, H. Jin, and T. Chen, "The Impact of COVID-19 on Consumers ' Psychological Behavior Based on Data Mining for Online User Comments in the Catering Industry in China," 2021.
- [5] F. Stephany and L. Neuhäuser, "The CoRisk -Index: A data-mining approach to identify industry-speci fi c risk assessments related to COVID-19 in real-time," no. April, pp. 1–18, 2020.
- [6] R. E. C. Id, V. Purushothaman, J. Li, M. Cai, and K. M. Id, "Sub-national longitudinal and geospatial analysis of COVID-19 tweets," pp. 1–11, 2020, doi: 10.1371/journal.pone.0241330.
- [7] A. Fadli, A. Wisnu, W. Nugraha, M. S. Aliim, and A. Taryana, "Simple Correlation Between Weather and COVID-19 Pandemic Using Data Mining Algorithms Simple Correlation Between Weather Pandemic Using Data Mining Algorithms and COVID-19," 2020, doi: 10.1088/1757-899X/982/1/012015.
- [8] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," vol. 15, no. 1, pp. 55–64, 2021.
- [9] I. Franch-pardo, B. M. Napoletano, F. Rosete-verges, and L. Billa, "Science of the Total Environment Spatial analysis and GIS in the study of COVID-19 . A review," *Sci. Total Environ.*, vol. 739, p. 140033, 2020, doi: 10.1016/j.scitotenv.2020.140033.
- [10] Y. J. Juhn, P. Wheeler, C. Wi, E. Ryu, E. Ristagno, and C. Patten, "Role of Geographic Risk Factors in COVID-19 Epidemiology: Longitudinal Geospatial Analysis," *Mayo Clin. Proc. Innov. Qual. Outcomes*, 2021, doi: 10.1016/j.mayocpiqo.2021.06.011.
- [11] A. S. Albahri, R. A. Hamid, J. Alwan, A. A. Zaidan, B. B. Zaidan, and A. O. S. Albahri, "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus ( COVID-19 ): A Systematic Review," 2020.
- [12] L. J. M. Milon, I. Sani, S. Usman, and S. Islam, "Predictive Data Mining Models for Novel Coronavirus ( COVID - 19 ) Infected Patients ' Recovery," *SN Comput. Sci.*, vol. 1, no. 4, pp. 1–7, 2020, doi: 10.1007/s42979-020-00216-w.
- [13] W. The and W. Academy, "Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes Philippine Twitter Sentiments during Covid-19 Pandemic."

- [14] J. Yuan et al., "Pharmacotherapy Management for COVID - 19 and Cardiac Safety : A Data Mining Approach for Pharmacovigilance Evidence from the FDA Adverse Event Reporting System ( FAERS )," *Drugs - Real World Outcomes*, vol. 8, no. 2, pp. 131–140, 2021, doi: 10.1007/s40801-021-00229-8.
- [15] S. Kumar, "Monitoring Novel Corona Virus ( COVID - 19 ) Infections in India by Cluster Analysis," *Ann. Data Sci.*, vol. 7, no. 3, pp. 417–425, 2020, doi: 10.1007/s40745-020-00289-7.
- [16] D. Li, H. Chaudhary, and Z. Zhang, "Modeling Spatiotemporal Pattern of Depressive Symptoms Caused by COVID-19 Using Social Media Data Mining," 2020.
- [17] I. O. P. C. Series and M. Science, "Model Expert System for Diagnosis of Covid-19 Using Naïve Bayes Model Expert System for Diagnosis of Covid-19 Using Naïve Bayes Classifier," 2020, doi: 10.1088/1757-899X/1007/1/012067.
- [18] I. O. P. C. Series and M. Science, "Mapping the Spread of Covid-19 in Asia Using Data Mining X-Means Algorithms Mapping the Spread of Covid-19 in Asia Using Data Mining X-Means Algorithms," 2021, doi: 10.1088/1757-899X/1071/1/012018.
- [19] Q. Chen, A. Allot, and Z. Lu, "LitCovid : an open database of COVID-19 literature," vol. 49, no. November 2020, pp. 1534–1540, 2021, doi: 10.1093/nar/gkaa952.
- [20] A. F. Watratan, A. P. B, D. Moeis, S. Informasi, and S. P. Makassar, "JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY ( JACOST ) Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," vol. 1, no. 1, pp. 7–14, 2020.
- [21] D. Şengür, "Investigation of the relationships of the students ' academic level and gender with Covid -19 based anxiety and protective behaviors : A data mining approach Öğrencilerin akademik düzeyi ve cinsiyetinin Covid - 19 temelli kaygı ve koruyucu davranışlarla ilişkisinin incelenmesi : Bir veri madenciliği yaklaşımı," vol. 15, no. 2, pp. 93–99, 2020.
- [22] S. Roy, M. Saha, B. Dhar, S. Pandit, and R. Nasrin, "Geospatial analysis of COVID-19 lockdown effects on air quality in the South and Southeast Asian region," no. January, 2020.
- [23] A. M. Forati and R. Ghose, "Geospatial analysis of misinformation in COVID-19 related tweets," no. January, 2020.
- [24] K. Takakuwa, "Difficulties of Integrating Human Resources Management Globally by Japanese Corporations," *Int. J. Appl. Inf. Manag.*, vol. 1, no. 4, pp. 173–185, 2021, doi: 10.47738/ijaim.v1i4.19.
- [25] T. T. Kim Phuong, "Proposing a Theoretical Model to Determine Factors Affecting on Job Satisfaction, Job Performance and Employees Loyalty For Technology Information (IT) Workers," *Int. J. Appl. Inf. Manag.*, vol. 1, no. 4, pp. 201–209, 2021, doi: 10.47738/ijaim.v1i4.21.
- [26] N. A. Prabowo, "Social Network Analysis for User Interaction Analysis on Social Media Regarding E-Commerce Business," vol. 4, no. 2, pp. 95–102, 2021.
- [27] M. Fatima, K. J. O. Keefe, W. Wei, and S. Arshad, "Geospatial Analysis of COVID-19 : A Scoping Review," 2021.
- [28] N. M. Abdulkareem, "COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms," pp. 100–105, doi: 10.48161/Issn.2709-8206.
- [29] J. Samuel, G. G. N. Ali, M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification," pp. 1–22, doi: 10.3390/info11060314.
- [30] M. F. Beckman, F. B. Mougeot, and J. C. Mougeot, "Comorbidities and Susceptibility to COVID-19 : A Generalized Gene Set Data Mining Approach," 2021.
- [31] "Comment Geospatial digital monitoring of COVID-19 cases at high spatiotemporal resolution," vol. 2, no. 20, pp. 393–394, 2020, doi: 10.1016/S2589-7500(20)30139-4.
- [32] M. Azarafza, M. Azarafza, and H. Akgün, "Clustering method for spread pattern analysis of corona-virus ( COVID-19 ) infection in Iran," vol. 3, no. 1, pp. 1–6, 2021.
- [33] K. Chamorro, M. Fors, F. X. Mora, and M. Pulgar-s, "Biomarkers of severe COVID-19 pneumonia on admission using data-mining powered by common laboratory blood tests-datasets," no. January, 2020.
- [34] R. Ahasan, S. Alam, T. Chakraborty, and M. Hossain, "Applications of GIS and geospatial analyses in COVID-19 research : A systematic review [ version 1 ; peer review : awaiting peer review ]," pp. 1–14, 2021.



- 
- [35] T. Chen, J. Rong, L. Peng, J. Yang, G. Cong, and J. Fang, "Analysis of Social Effects on Employment Promotion Policies for College Graduates Based on Data Mining for Online Use Review in China during the COVID-19 Pandemic," 2021.
- [36] U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," vol. 5, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.
- [37] N. Made, A. Juli, D. Gede, H. Divayana, and G. Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," pp. 22–29, 2020, doi: 10.30864/jsi.v15i1.332.
- [38] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-elsoud, "Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy," no. January, 2020.
- [39] N. A. Mansour, A. I. Saleh, M. Badawy, and H. A. Ali, "Accurate detection of Covid - 19 patients based on Feature Correlated Naïve Bayes ( FCNB ) classification strategy," no. 0123456789. Springer Berlin Heidelberg, 2021.
- [40] C. Scarpone, S. T. Brinkmann, T. Große, D. Sonnenwald, M. Fuchs, and B. B. Walker, "A multimethod approach for county - scale geospatial analysis of emerging infectious diseases : a cross - sectional case study of COVID - 19 incidence in Germany," *Int. J. Health Geogr.*, pp. 1–17, 2020, doi: 10.1186/s12942-020-00225-1.
- [41] R. E. Cuomo, V. Purushothaman, J. Li, M. Cai, and T. K. Mackey, "A longitudinal and geospatial analysis of COVID-19 tweets during the early outbreak period in the United States," pp. 1–11, 2021.