Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19

Artika Arista*
Information Systems, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
artika.arista@upnvj.ac.id

Submitted : Dec 15, 2021 | **Accepted** : Jan 3, 2022 | **Published** : Jan 3, 2022

Abstract: Many people today are unsure whether they have COVID-19. The frequent fever, dry cough, and sore throat are all signs and symptoms of COVID-19. If a person has signs or symptoms of coronavirus disease 2019 (COVID-19), he/she should see the doctor or go to a clinic as soon as possible. As a result, it's vital to learn and comprehend the fundamental differences. COVID-19 can cause a wide range of symptoms. The experiments were carried out using two Machine Learning Classification Algorithms, namely Decision Tree (DT) and Logistic Regression (LR). Both algorithms were written and analyzed using the Python program in Jupyter Notebook 6.4.5. From the results obtained in the experiments of covid symptoms dataset, on average, the DT model has obtained the best cross-validation average and the testing performance average compared to the LR machine learning models. For cross-validation results, the DT model has achieved an accuracy of 98.0%. For performance testing, the DT model has achieved an accuracy of 98.0%. The LR has obtained the second-best result on the average of cross-validation performance and the testing results. For cross-validation results, the LR model has achieved an accuracy of 96.0%. For performance testing, the LR model has achieved an accuracy of 97.0%. Consequently, the DT for the COVID-19 symptoms dataset is outperforming the LR for cross-validation and testing results.

Keywords: Covid-19; Machine Learning; Classification Algorithms; Decision Tree; Logistic Regression

INTRODUCTION

Technology has advanced significantly in recent years, particularly in the field of Machine Learning (ML), which is effective for minimizing human labor. In the field of artificial intelligence, machine learning (ML) combines statistics and computer science to create algorithms that become more efficient when given relevant data rather than precise instructions. ML is the study of computing methods that are automatically enhanced by experience, in addition to speech recognition, picture identification, and text localization. It is classified as a subset of artificial intelligence (Mohsin Abdulazeez et al., 2020; Charbuty & Abdulazeez, 2021).

ML algorithms generate a model population based on a sample, known as 'training data,' to predict or make decisions without being explicitly taught to do so (Charbuty & Abdulazeez, 2021). ML algorithms are used in a wide range of applications, such as email filtering and computer vision when creating classical algorithms to fulfill essential tasks is difficult or impracticable (Carleo et al., 2019). There are numerous applications for machine learning, the most popular of which is predictive data mining. Model development and model evaluation are two important methods of ML classification fulfillment (Hillel et al., 2021).

Any instance in every dataset utilized by ML algorithms is characterized using the same set of attributes. Continuous, categorical, or binary attributes could be used. Learning is called supervised when situations are recognized with recognized labels (correct outputs). Machine learning is the role of supervised learning in inferring a function from categorized training data. It also examines the results of the tests and generates a derived task that can be used to map new instances (Charbuty & Abdulazeez, 2021; Kunal Pahwa & Neha Agarwal, 2019; Sharma & Kumar, 2017).

*corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The goal of supervised learning methods is to figure out the relationship between input and target qualities. The model can then be used to forecast the value of the target attribute for new input data after it has been built. There are two types of supervised models: classification models and regression models, which are the ones we're interested in in this study. The input space (features) is mapped into one of the predetermined classes by classification models. Classifiers can be used to categorize things in an outside scene image as human, vehicle, tree, or structure, for example. Regression models, on the other hand, translate the input space into the real-vales domain. For example, a regression model can be created to forecast the price of a property based on its size, the number of rooms, and the size of the garden(Mashat et al., 2012).

Many people today wonder if they have COVID-19 or not. This question becomes a little more complicated for the millions of allergy sufferers across the country – allergies or COVID-19 (also known as the coronavirus), or even a cold or the flu? Many of the indications and symptoms of COVID-19, the common fever, dry cough, and sore throat. If a person experience signs or symptoms of coronavirus disease 2019 (COVID-19), he/she should contact the doctor or clinic as soon as possible for medical guidance (Narayan, 2021). So, how do we know whether we have got COVID-19 or not? As a result, it's critical to find out and understand the key distinctions. COVID-19 has a wide range of symptoms, depending on the variation we have. We use a supervised learning technique of building a decision tree model and logistic regression for covid symptoms dataset by hemanth hari (hemanth hari, 2020) to provide a comparison of two Machine Learning Classification Algorithms (Decision Tree and Logistic Regression) the findings are summarized, and the highest accuracy methods are achieved to tell the difference the covid and not.

LITERATURE REVIEW

Along with its widespread use in practice, a Decision Tree (DT) is also regarded as the most widely used form of a machine learning algorithm in classification tasks. The interpretability and accuracy of this algorithm in producing prediction models with an understandable structure that create relevant information on the corresponding area are key factors in its reputation(Banihashemi et al., 2017). Fig. 1 illustrate a structure of DT.

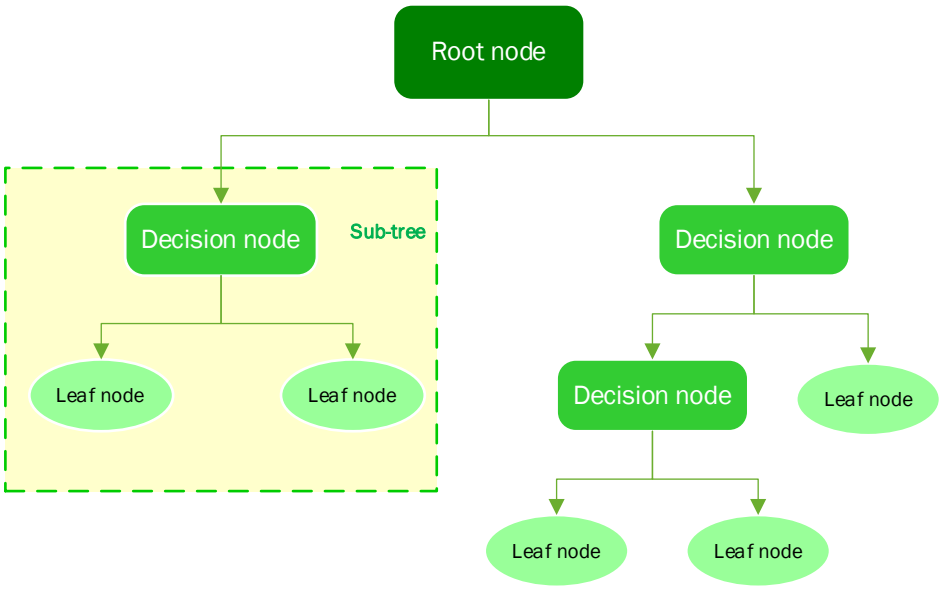

Fig. 1 Decision Tree (Charbuty & Abdulazeez, 2021)

Several recent works on the DT are discussed in this research. For diabetes mellitus prediction, Zou et al. (Zou et al., 2018) used decision tree (j48), Random Forest (RF), and neural network techniques. Physical research data for hospitals in Luzhou, China is included in the dataset. There are 14 different characteristics to consider. The training array extracts data from 68994 healthy humans and diabetic patients at random. To reduce dimensionality, they exploited the full significance of minimum Redundancy Maximum Relevance (mRMR) and Principal Component Analysis (PCA). In certain instances, the effects of RF, as opposed to the other classifiers, appeared to be larger. When compared to the other classifiers, the effects of RF appeared to be greater in some cases. In addition, the best result in the Luzhou data collection is 0.8084.

The DT method was utilized by Sathiyanarayanan et al. (Sathiyanarayanan et al., 2019) to detect breast cancer using the supervised learning mechanism. Here, breast cancer detection is carried out with a focus on

*corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

data, which is separated for the preparation and testing process. As a result, the results of the algorithms KNN and DT are compared. The results show that KNN achieves a 97 percent accuracy, while DT achieves a maximum accuracy of 99 percent. As a result, a supervised learning method known as a decision tree algorithm can predict the type of cancer.

Logistic Regression (LR) is another prominent classification approach. LR is a machine learning technique that can be applied to classification problems. Pierre Francais Verhulst defined the logistic function and its attributes in a paper published in Proceedings of the Belgian Royal Academy by specifying three parameters and the curve flowing through them. It is a relatively simple machine learning method that is widely utilized. A statistical approach for predicting binary classes is logistic regression. In this case, the dependent variable has a Bernoulli distribution. A sigmoid function, also known as a logistic function, is a 'S' shaped curve that takes values between 0 and 1. If the curve travels to positive infinity, 1 will be anticipated, and if it goes to negative infinity, 0 will be predicted (Majumder et al., 2021). A logistic function is illustrated in Figure 2.

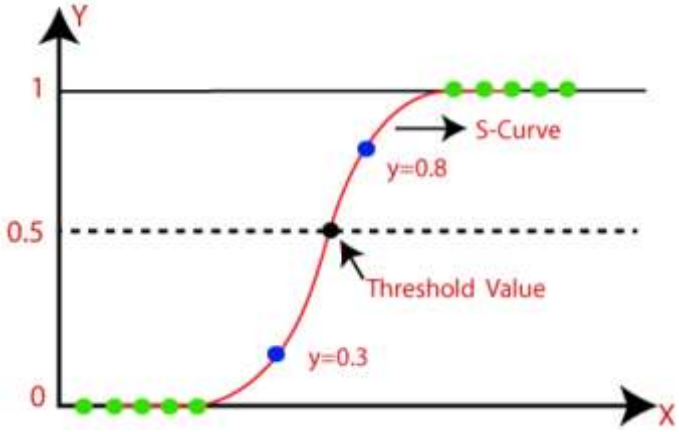

Fig. 2 Logistic Regression (Majumder et al., 2021)

Several recent works on the LR are discussed in this research. Authors of another study (Yan et al., 2020) collected blood samples from 404 patients in Wuhan, China, to identify disease-predictive biomarkers. The authors have presented a COVID-19 mortality prediction model based on artificial intelligence. For mortality prediction, they employed Artificial Neural Networks, RandomForest, Decision Tree, Logistic Regression, and K-Nearest Neighbor (KNN). The model had a 93 percent accuracy rate.

In a study conducted in the Netherlands (Hermans et al., 2020), they employed 319 patients as the dataset and used supervised learning as their approach, with logistic regression as the algorithm. They relied on the patient's chest CT scan scores and the RT–PCR test findings in this article. The results showed that chest CT, utilizing the CO-RADS rating system, is a specific beneficial tool that can lead to reliable COVID-19 diagnosis, even if RT–PCR tests are scarce during an epidemic. Combining a predictive machine-learning model with chest CT scans for COVID-19 patients may also increase diagnosis accuracy. Nonetheless, they advised that RT–PCR be kept as the principal method of diagnosis because up to 9% of patients with a positive RT–PCR test have cancer. Nonetheless, they suggested that RT–PCR be used as the primary testing method because up to 9% of patients with positive RT–PCR were missed by chest CT or the machine-learning model. Nonetheless, they suggested that RT–PCR be used as the primary testing method because up to 9% of patients with positive RT–PCR were missed by chest CT or the machine-learning model by (Kwekha-Rashid et al., 2021).

METHOD

A proposed workflow is depicted in the diagram below in Figure 3. Step 1. Data collection. There are a variety of web resources available, each with a large data set. The COVID-19 patient data were collected from Kaggle covid symptoms dataset by hemanth hari (hemanth hari, 2020).

Step 2. Data Preparation. The most crucial aspect of any machine learning-based application is data preprocessing. Data must be properly processed before being fed to the system to produce an accurate machine learning application. This data was preprocessed using Label Encoding of the Python machine learning framework to get target labels with value between 0 and 1.

Step 3. Data Splitting. A dataset from 80% of the training set and another dataset from 20% of the testing set are used in this step. The training set is put into ML models to determine what should be done with the data, while the test set is used to double-check the results.

*corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

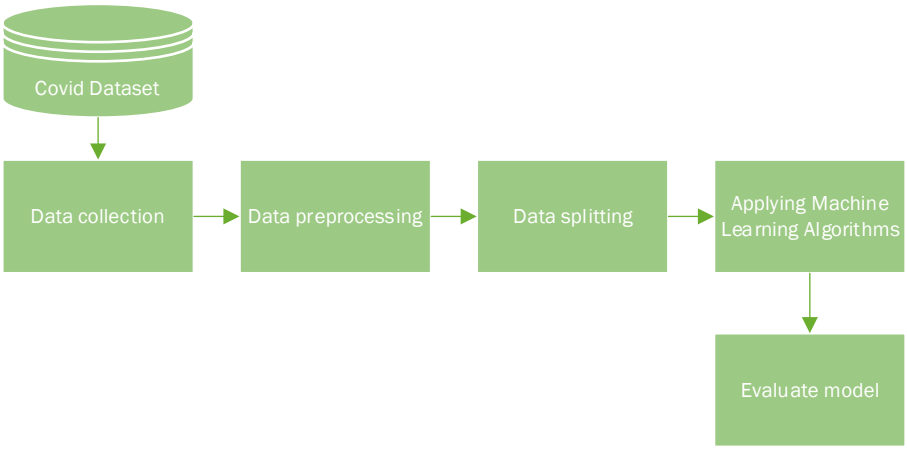

Fig. 3 Proposed workflow for classifying COVID-19 or not (Abdelminaam et al., 2021)

Step 4. Applying Machine Learning Algorithms. We have used two machine learning algorithms that can be utilized to solve a classification problem which is Decision Tree (DT) and Logistic Regression (LR) to classify covid symptoms or not.

The dataset (hemanth hari, 2020) highlights the possible risk factors associated with the disease such as Breathing Problems; Fever; Dry Cough; Sore throat; Running nose; Asthma; Chronic Lung Disease; Headache; Heart Disease; Diabetes; Hypertension; Fatigue; Gastrointestinal; Abroad travel; Contact with COVID Patient; Attended Large Gathering; Visited Public Exposed Places; Family working in Public Exposed Places; Wearing Masks; Sanitization from Market. The

Step 5. Evaluating models. Four statistics are used to determine the consistency of models. TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. Accuracy is shown in Equation (1) and Precision is shown in Equation (2). Recall is shown in Equation (3), and F1-Score is shown in Equation (4)(Abdelminaam et al., 2021; Molin et al., 2021).

Accuracy is a measure of totally correctly identified samples out of all the samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Precision is the ratio of true positives to everything flagged positive.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall gives us the true positive rate (TPR), which is the ratio of true positives to everything that was positive.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score is the harmonic mean of Precision and Recall.

$$F_1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (4)$$

RESULT

In this section, the results of applying two machine learning models DT and LR. The effectiveness of the classifier algorithms is studied here, (1) the performance of DT and LR algorithms. The experiments were carried out using a VivoBook Asus Laptop equipped Processor 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz and 8GB RAM. The algorithm was written and analyzed using the Python program in

*corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Jupyter Notebook 6.4.5. For criteria (1), using DT classifier, 0.98 or 98% accuracy is achieved whereas the LR classifier obtained 0.97 or 97% accuracy. This shows that DT performs better than LR for the selected dataset.

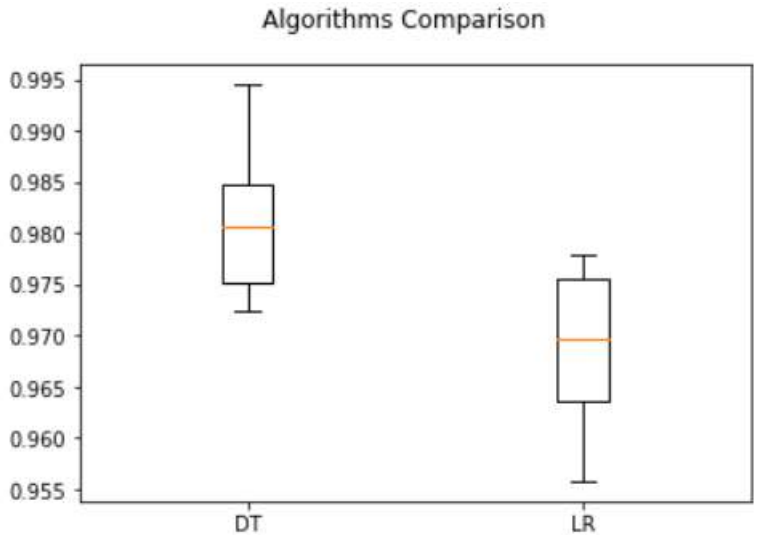

Fig. 4 Visualization using Boxplot for comparison

The model using the DT technique obtained the highest efficiency with an accuracy of 98.0%. The LR model achieved the best overall efficiency with an accuracy of 97.0%. Figure 4 is depicting visualization using Boxplot for comparison of accuracy between the DT and the LR model. Table 1. are depicting the overall performance comparison of the machine learning algorithms which shows the main classification metrics precision, recall, and f1-score on a per-class basis.

Comparison test result

Table 1. Overall performance comparison of the machine learning algorithms

Algorithms	Accuracy (%)	Class	Precision (%)	Recall (%)	F1-score
DT	98	Not covid	0.93	0.98	0.99
		Covid	0.99	0.98	0.99
LR	97	Not covid	0.95	0.94	0.95
		Covid	0.98	0.97	0.97

DISCUSSIONS

From the results obtained in the experiments of covid symptoms dataset. Figure 9 depicts the empirical results in the big picture for the cross-validation performances and the testing results, respectively. On average, the DT model has obtained the best cross-validation average and the testing performance average compared to the LR machine learning models. For cross-validation results, the DT model has achieved an accuracy of 98.0%. For performance testing, the DT model has achieved an accuracy of 98.0%.

The LR has obtained the second-best result on the average of cross-validation performance and the testing results. For cross-validation results, the LR model has achieved an accuracy of 96.0%. For performance testing, the LR model has achieved an accuracy of 97.0%. Consequently, the DT for the COVID-19 symptoms dataset is outperforming the LR for cross-validation and testing results.

According to the findings(Charbuty & Abdulazeez, 2021), multiple studies were conducted with various data sets, and the DT technique was utilized to improve performance. Several optimization techniques were used in the study (Nandhini & Marseline, 2020), to strengthen the decision tree on the UCI ML datasets stored; based on

*corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the assessment findings, it was determined that the DT approach had the highest accuracy of 99.93 %, compared to other techniques such as K-Nearest Neighbor (KNN), LR, Support Vector Machine (SVM), and Naïve Bayes (NB), which performed less well.

Two limitations are identified in relation to the analysis of the models in the studies in the review. The first limitation is this work only compares two Machine Learning Classification Algorithms, namely Decision Tree (DT) and Logistic Regression (LR). The second limitation is this work uses only 20 possible indications and symptoms of covid-19. Future study topics could include other Covid symptom datasets, large amounts of data, different types of diseases, and different classification techniques.

CONCLUSION

The diagnosis of covid-19 is regarded as a difficult research topic. Machine learning techniques for medical diagnostics have gained popularity in the medical community due to their capacity to forecast diseases early. Physicians will be able to determine the most appropriate treatment procedure with such early prediction power to covid. For analyzing covid prediction on several criteria such as accuracy, recall, precision, and F1-score, the decision tree and logistic regression techniques are employed in this research. The DT method outperforms the LR algorithm in the experiments.

ACKNOWLEDGMENT

This work is supported by the Institute for Research and Community Service (LPPM) Universitas Pembangunan Nasional "Veteran" Jakarta (UPNVJ), Faculty of Computer Science UPNVJ, Information Systems Study Program UPNVJ for providing funding support and assisting the implementation of this research.

REFERENCES

Abdelminaam, D. S., Ismail, F. H., Taha, M., Taha, A., Houssein, E. H., & Nabil, A. (2021). CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter. *IEEE Access*, *9*, 27840–27867. https://doi.org/10.1109/ACCESS.2021.3058066

Banihashemi, S., Ding, G., & Wang, J. (2017). Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption. *Energy Procedia*, *110*, 371–376. https://doi.org/10.1016/j.egypro.2017.03.155

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4). https://doi.org/10.1103/RevModPhys.91.045002

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

hemanth hari. (2020). *Symptoms and COVID Presence (May 2020 data) Kaggle*. https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence

Hermans, J. J. R., Groen, J., Zwets, E., Boxma-De Klerk, B. M., van Werkhoven, J. M., Ong, D. S. Y., Hanselaar, W. E. J. J., Waals-Prinzen, L., & Brown, V. (2020). Chest CT for triage during COVID-19 on the emergency department: myth or truth? *Emergency Radiology*, *27*(6), 641–651. https://doi.org/10.1007/s10140-020-01821-1

Hillel, T., Bierlaire, M., Elshafie, M. Z. E. B., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, *38*. https://doi.org/10.1016/j.jocm.2020.100221

Kunal Pahwa, & Neha Agarwal. (2019). Stock Market Analysis using Supervised Machine Learning. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019*.

Kwekha-Rashid, A. S., Abduljabbar, H. N., & Alhayani, B. (2021). Coronavirus disease (COVID-19) cases analysis using machine-learning applications. *Applied Nanoscience (Switzerland)*. https://doi.org/10.1007/s13204-021-01868-7

Majumder, A. B., Gupta, S., Singh, D., & Majumder, S. (2021). An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression. *Journal of Physics: Conference Series*, *1797*(1). https://doi.org/10.1088/1742-6596/1797/1/012011

Mashat, A. F., Fouad, M. M., Yu, P. S., & Gharib, T. F. (2012). A Decision Tree Classification Model for University Admission System. *IJACSA) International Journal of Advanced Computer Science and Applications*, *3*(10). www.ijacsa.thesai.org

*corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11243
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Mohsin Abdulazeez, A., Zeebaree, D., Abdulqader, D. M., & Zeebaree, D. Q. (2020). *Machine Learning Supervised Algorithms of Gene Selection: A Review*. *62*. https://www.researchgate.net/publication/341119469

Molin, S., Jee, K., O'Reilly for Higher Education (Firm), & Safari, an O. M. Company. (2021). *Hands-On Data Analysis with Pandas - Second Edition*.

Nandhini, S., & Marseline, D. J. (2020, February 1). Performance Evaluation of Machine Learning Algorithms for Email Spam Detection. *International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020*. https://doi.org/10.1109/ic-ETITE47903.2020.312

Narayan, S. (2021). *Allergies, Cold, Flu or COVID-19? How to Tell the Difference*. Emersonhospital.Org. https://www.emersonhospital.org/articles/allergies-or-covid-19

Sathiyanarayanan, M. P., Sai, S. M., & Vinayagar, M. (2019). Identification of Breast Cancer Using The Decision Tree Algorithm. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*.

Sharma, D., & Kumar, N. (2017). A Review on Machine Learning Algorithms, Tasks and Applications. In *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* (Vol. 6, Issue 10, pp. 2278–1323).

Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., … Yuan, Y. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection. *MedRxiv*. https://doi.org/10.1101/2020.02.27.20028027

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00515

*corresponding author