COVID-19 SEVERITY CLASSIFICATION AND ANALYSIS USING MACHINE LEARNING

Marco Antonio[1]
Marcos Vinicius[2]
Diego Lisboa[3]

[1]Federal University of Pará, PPGEE, Operational Research Laboratory, Belém, Pará
[2]Universitary Center of State of Pará, Belém, Pará
[3] Federal University of Pará, PPGEE, Operational Research Laboratory, Belém, Pará

ABSTRACT

In the last years, with the alarming growth of Covid-19 cases, a highly contagious viral disease, new forms of diagnosis and control for this sickness have become necessary to the spread decreases until the population is effectively vaccinated. In this context, Artificial Intelligence (AI) and its subfields appear as possible alternatives to help and provides a response to combat the virus. Some Machine Learning (ML) methods are shown as an answer to control this disease, these methods can perform an analysis based on a set of symptoms presented by the patient and consequently indicating the diagnosis, as well as streamline the treatment process. To achieve this goal in this paper, three models that use ML methods to predict COVID-19 severity on different degrees are proposed, unlike other works whose purpose was to diagnose only the presence or absence of Covid-19, this paper aims to improve the classification of the patient's disease state. The results in each of these models are evaluated through the metrics established in this work. Furthermore, there are distinct suggestions to improve the analysis and make predictions with greater accuracy.

Keywords: Covid-19, Artificial Intelligence, Machine Learning, Improve, Analysis, Severity.

INTRODUCTION

The Covid-19 is a highly contagious viral disease reported for the first time in China as an abnormal case of pneumonia. Since then, the number of related cases of this disease has exponentially grown, causing several fatalities in multiple countries around the world, being considered by the World Health Organization (WHO) as a case of international public emergency. Since then, the governments have imposed social and border restrictions as well as strengthened hygiene habits to reduce the virus spread (Ye et al., 2021; Chowdhury et al., 2020).

Within this context, new ways of identifying and controlling possible cases of Covid-19 are needed. One of these ways is through the use of Artificial Intelligence (AI), which is a technology that can ease track the spread of the virus, identify high-risk patients and predict mortality through analysis of data. AI has subfields of Machine Learning (ML) and Deep Learning (DL), and ML is a subfield of AI where existing data is used to predict and respond to future data (Chowdhury et al., 2020; Vaishya et al., 2020).

There are several ML algorithms, these algorithms based on the feature selection approach can choose the most relevant data in unbalanced and large-scale datasets. Some of these algorithms stand out in performance in comparison to other approaches based on feature filters, both in computational cost and accuracy. Its uses in various health fields to predict problems, such as cancer and heart diseases, show significant efficiency in most cases. The aforementioned indicates that if ML can be effectively used for the diagnosis of other health problems, its application can be effective as well for the diagnosis of Covid-19 (Boyarshinov, 2005; Liu and Zhou, 2017; Felice and Polimeni, 2020).

In this line of reasoning, this paper proposes to use three ML techniques to diagnose and classify the degree of severity of Covid-19, comparing the performance obtained in each of them according to the metrics defined in this work, verifying which fits better for the diagnosis of the disease. The aim is to contribute to the scientific community and its diverse segments in the control of the pandemic. The paper is organized as follows: Section 2 presents the related works, section 3 comprises the methodology proposed in this work, section 4 presents the results obtained and puts an end to section 5 shows the conclusion and future work.

CORRELATED WORKS

Multiple works use ML as a diagnostic and control tool, like Hua Ye et al. (2021), which develops a framework that uses Harris Hawk's optimizer (HHO), which trains a Fuzzy K-nearest Neighbor (FKNN) model so the resulting model (HHO-FKNN) can be used for COVID-19 severity diagnosis. An improvement is obtained in comparison to the FKNN, however, due to factor limitations during the beginning of the pandemic, multiple features which are disease-related aren't considered.

Sung-Bae and Hong-Hee (2003) used four techniques, MLP, KNN, SVM, and Structure Adaptive Self-Organizing Map (SASOM) applied to three benchmark datasets for DNA analysis. The performance evaluation is based on the accuracy obtained in each of the techniques for all datasets. The comparison is made by combining about 42 characteristics and classifiers, showing MLP and KNN as better compared to other techniques, however, the problem in question does not take into account values with low confidence rates in one of the datasets in order to improve the accuracy, which leads to a better result, but less plausible.

Authors Rustam et al. (2020) use different ML techniques, which are, Linear Regression (LR), LASSO Regression (LL), Support Vector Machine (SVM), and Exponential Smoothing (ES) to predict Covid-19 propagation foci. The technique that obtained the best performance in relation to the others was ES due to the nature and size of the dataset used, but because of this same factor and other aspects of the SVM technique, it was underutilized, therefore it had the worst performance.

Liu et al. (2017) show a Decision Tree using Gradient Boosting XGBoost (DTGB) is applied to classify the patient's symptoms severity which has a previous psychiatric evaluation. It's possible to visualize that the conventional DT technique regarding DTGB has a superior performance due to the XGBoost improvement dealing with missing values and tree complexity. However, the confusion matrix shows a high quantity of instances misclassified since DTGB doesn't handle well with a multiclass problem.

The authors Ambesange et al. (2020) use Logistic Regression (LR) to predict multiple heart diseases with adjustment techniques and hyperparameter sets to diagnose multiple heart diseases. Optimizations made with tuning and hyperparameter techniques are effective, however, the same LR algorithm is used for all cases, showing that there is a gap in the performance evaluation that can be filled by applying other techniques to the problem in question. Considering what was presented before, this work will try to perform COVID-19 severity classification using different ML algorithms that fit a multiclass problem together with adjustment and optimization techniques, so it's expected no subtilization of any techniques as well as acceptable performance.

MATERIALS AND METHODS

This section shows each step of the COVID-19 severity classification used in this work. The tools for implementation and development were: Anaconda 1.9.12 with Python 3.8.3 using Jupyter Notebook Integrated Development Environment.

PRE-PROCESSING

The dataset used is based on guidelines provided by WHO. It is composed of 16 attributes, which are: fever, tiredness, dry cough, difficulty breathing, sore throat, absence of symptoms, nasal congestion, stuffy nose, diarrhea, no experience, age, gender, contact with infected, and country. Data pre-processing was done in several steps. The initial data cleaning phase was done removing attributes with little relevance. Then it was necessary to identify the correlations between features. After that, 15 attributes remained to be selected, being 14 input parameters and 1 output parameter.

The dataset used has a large amount of data but, the frequency at which determined data appear about others is unbalanced. Thus, after data selection, it was necessary to balance these data, so that the techniques subsequently applied do not have their performance impaired. For this, SMOTE (synthetic minority over-sampling technique) was used. It is an algorithm capable of reducing the influence of unbalanced classes through the generation of artificial minority instances using k-nearest neighbors and random variables (Lee et al., 2018).

Then the dataset was distributed in a Gaussian way using the Power Transformer pre-processing technique to avoid non-constant variations. After that, the data was split into the 25% test and the remaining training. So state randomization together with the stratification of its data was applied for each of the techniques. Cross-validation is being used to estimate hyperparameters through the method of data division in two disjoint parts (training and testing), to obtain a better prediction and validation (Mu et al., 2018).

MACHINE LEARNING ALGORITHMS

The Decision Tree (DT) algorithm used in this article for classification is based on Classification and Regression Tree (CART). This operation creates decision trees using past data with assigned classes for all observations. CART builds resolutions by dividing information into two parts with as much homogeneity as allowed. These factors allow the algorithm to achieve good accuracy with a low computational cost (Liu and Zhou, 2017).

The Random Forest (RF) algorithm establishes multiple DTs over the database, from which an estimative is obtained from each one of these. Being a classification problem involving symptoms of COVID-19, the tree with the highest number of votes among the estimates is selected as the best. Furthermore, Random Forest (RF) can reduce overfitting, one of the biggest problems of ML, and it also provides an advantage in its accuracy by not ignoring atypical observations (Yaşar et al., 2021).

The last algorithm used, Logistic Regression (LR), can indicate a correlation between two or more data elements. With these linear relationships between the data and in the case of COVID-19, the technique can be suitable for the dataset used, since there are more linear relationships between the data than non-linear ones (Kim et al., 2020).

FEATURE ENGINEERING

Feature engineering is a procedure that involves identifying correlations between dependent or independent attributes. Through this process, it was possible to optimize the results obtained using the algorithms mentioned above (Mars et al., 2018).

HYPERPARAMETER OPTIMIZATION

Hyperparameter optimization is an essential part of building an effective ML model by selecting the best parameters for the dataset. Thus, Grid Search optimization of hyperparameters was applied for each of the three techniques to improve the results obtained in each one (Ambesange et al., 2020).

MODEL EVALUATION

Model evaluations according to the metrics (accuracy, precision, recall, ROC) shown in table 1 were performed. In addition, pre-processing, resource engineering, optimization, and assembly techniques parameters are described. Several executions were realized to guarantee the veracity of the results obtained in each model.

Table 1. Applied Models

Algorithm	Pre-Processing	Feature Engineering	Hyperparameter Optimization	Ensemble Technique
Decision Tree (CART)	SMOTE; Power Transformer; method= yeo-johnson	Correlation Matrix	GridSearchCV (random_state =42, max_iterations=300, max depth = 20, criterion= ‘gini’, ‘entropy’)	Cross-validation = 10
Random Forest	SMOTE; Power Transformer; method= yeo-johnson	Correlation Matrix	GridSearchCV (random_state =42, max_iterations=200, max depth = 30,n_estimators=100’,criterion= ‘gini’, ‘entropy’)	Cross-validation= 10
Logistic Regression	SMOTE; Power Transformer; method= yeo-johnson	Correlation Matrix	GridSearchCV (random_state =42, max_iterations=200, penalty= ‘l1’,’l2,’elastic-net’’ class_weigth= ‘balanced’, ‘dict’’)	Cross-validation = 10

RESULTS

After applying all processes in Table 1, the lowest accuracy, precision, and F1-Score obtained were in the Logistic Regression technique, reaching an average of 52.29%. Decision Tree achieved better results compared to the previous technique, achieving an average of 67.53%. Its prominence was mainly in the execution time, reaching an average time about the others. Random Forest was the technique that demanded the highest execution time and computational cost; however, it obtained an average of 82.5%, showing better results than the other two techniques.

Figures 1,2,3 show the learning curve, evaluating the accuracy of each aforementioned algorithms. RF obtained the best performance about the other two techniques in terms of accuracy.

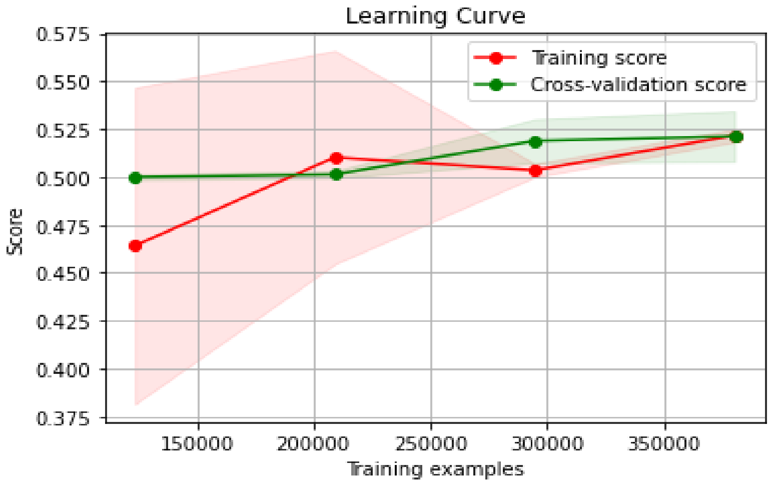

Figure 1. LR Learning Curve

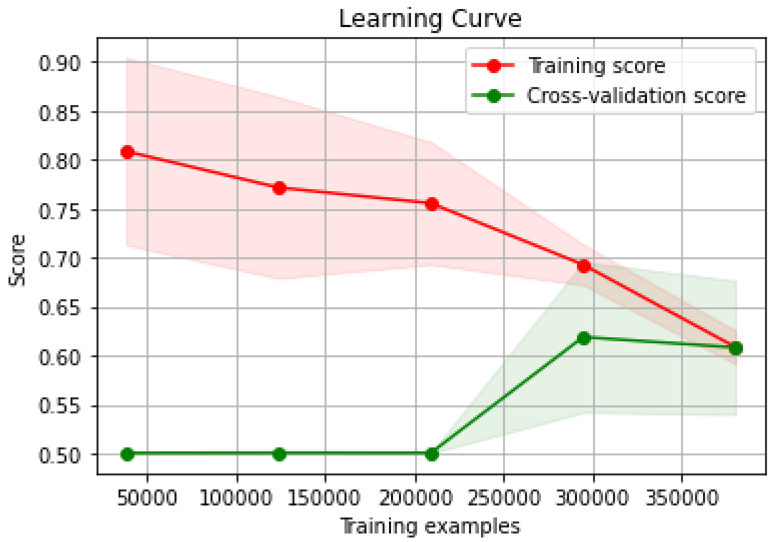

Figure 2. DT Learning Curve

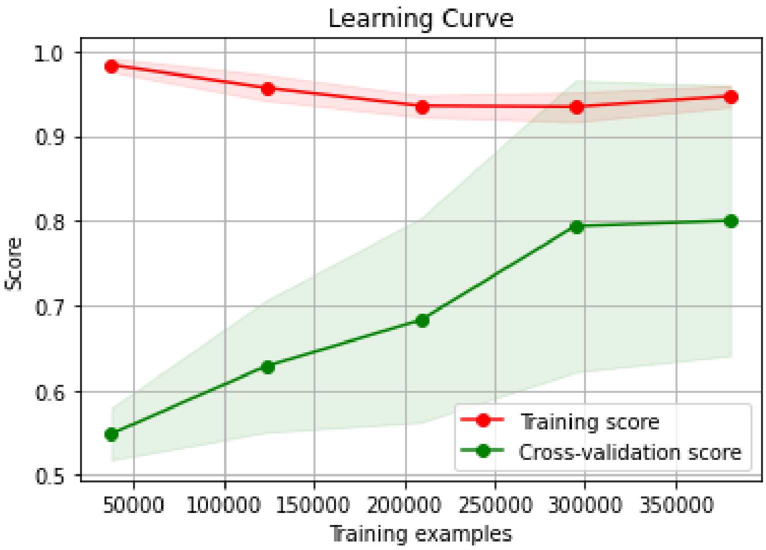

Figure 3. RF Learning Curve

Figures 3,4 and 5 show the precision and recall curve for each algorithm. Class 0 indicates the user as not ill or asymptomatic, while class 1 indicates one of three possible grades for COVID-19. As you can see, the algorithm that obtained the smallest curve variation about the others was LR. DT showed great variation in its curves, resembling lines at intervals: 0.6 to 1.0 (accuracy) and 0.0 to 0.4 (recall) for class 0. For class 1 at intervals: 0, 8 to 1.0 (accuracy) and 0.0 to 0.2 (recall). RF, like DT, has large variations in its curves. Although it performs better than the other two techniques, there is a large variation in the precision range from 0.6 to 1.0.

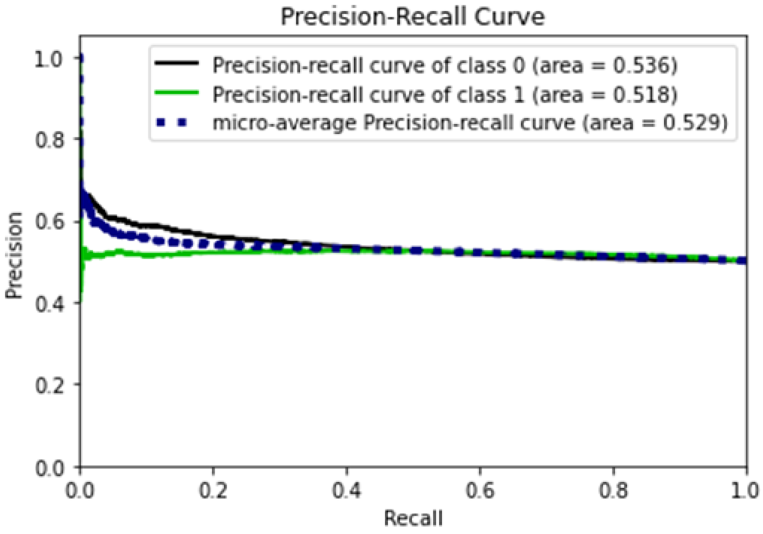

Figure 4. LR Precision-Recall Curve

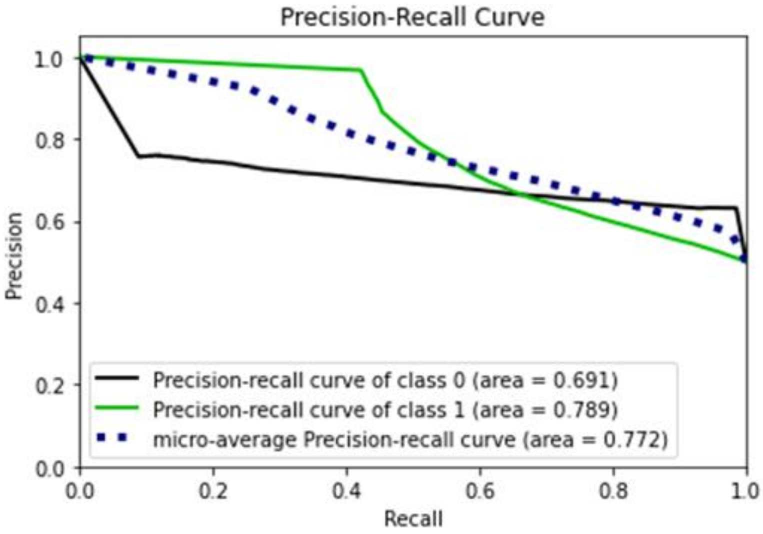

Figure 5. DT Precision-Recall Curve

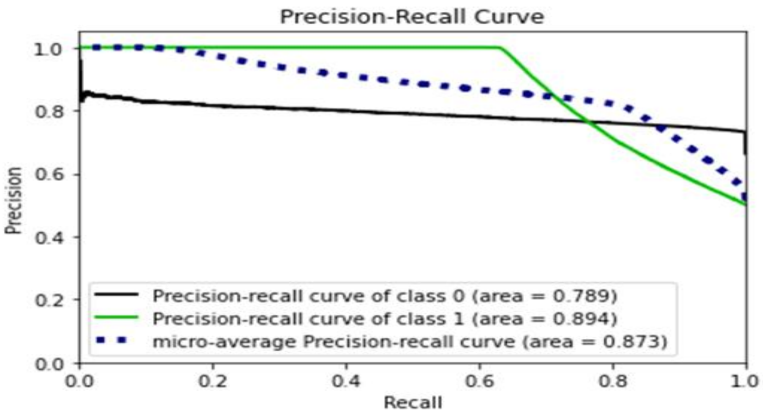

Figure 6. RF Precision-Recall Curve

Figures 7, 8, and 9 show the ROC curve for each algorithm. LR maintains a curve without major variations while DT and RF show a behavior similar to the previous Figures (5,6). It's possible to observe that DT and RF curves resemble lines and RF what it obtained the best true positive rate statistic and false positive rate (TPR/FPR) relative to others.

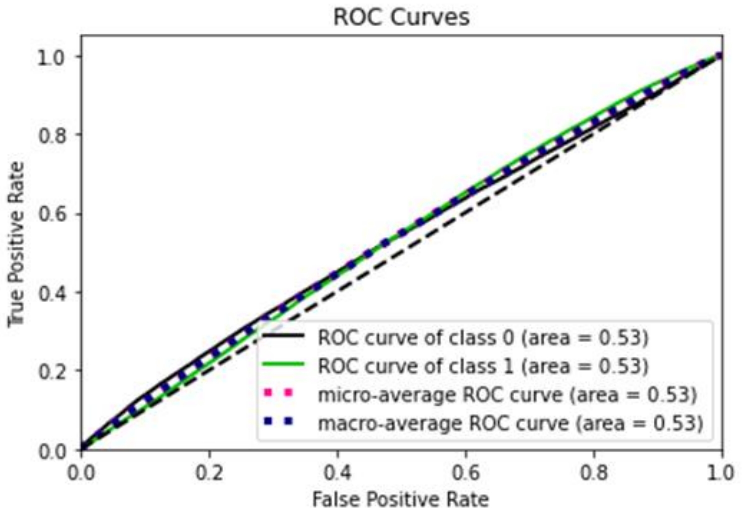

Figure 7. LR ROC Curve

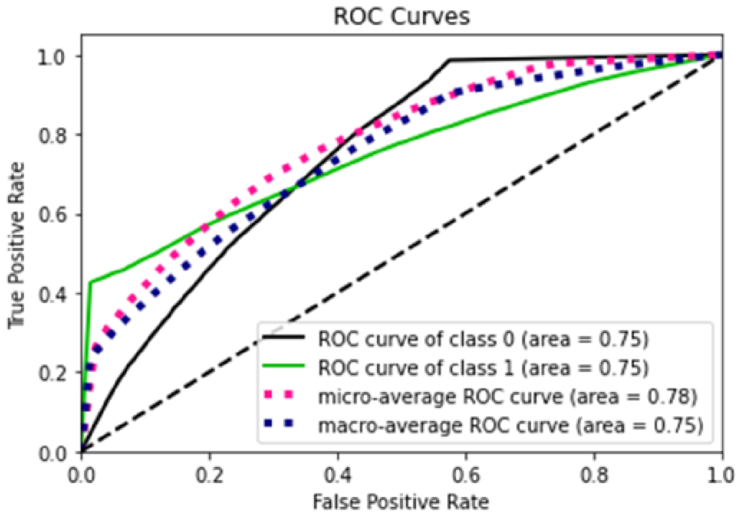

Figure 8. DT ROC Curve

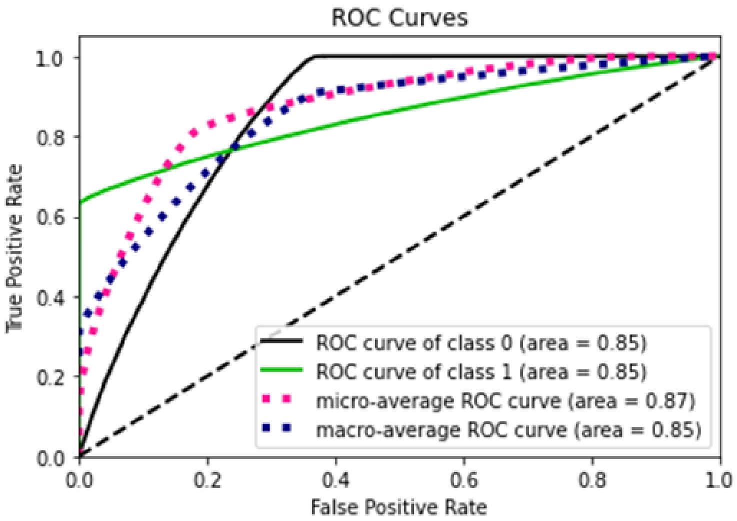

Figure 9. DT ROC Curve

CONCLUSION

Covid-19 proved to be a big challenge to be overcome in the global health scenario. In the last two years, several studies that use ML techniques to help control and diagnose this disease have shown promising results for the classification. In this sense, this work proposed and evaluated three models to classify the severity of COVID-19 in different degrees to help control and diagnose the disease. According to the results, it was possible to see that the model with the best percentages in classifying the severity of COVID-19 was RF, with an overall average of about 82.5%, which can be taken into account during a previous diagnosis of this disease, although this model in terms of computational time and cost is the one that demanded more compared to the others.

Furthermore, as future works, there are opportunities for new model applications to carry out a more complete assessment for the classification of this disease. Other forms of pre-processing and balancing also can be used to improve the results obtained. Another alternative for getting better analysis is to use Deep Learning with Big Data to perform a more comprehensive analysis in terms of the amount of data.

REFERENCES

Ambesange, S. et al. Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques. Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability, WS4 2020, p. 827–832, 2020.

Boyarshinov, V. Machine Learning Machine Learning. [s.l: s.n.]. v. 2005.

Chowdhury, M. E. H. et al. Can AI help in screening viral and COVID-19 pneumonia? arXiv, v. 8, p. 132665–132676, 2020.

De Felice, F.; Polimeni, A. Coronavirus disease (COVID-19): A machine learning bibliometric analysis. In Vivo, 2020.

Kim, M. K.; Kim, Y. S.; Srebric, J. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustainable Cities and Society, v. 62, n. September 2019, p. 102385, 2020.

Lee, H. et al. Synthetic minority over-sampling technique based on fuzzy c-means clustering for imbalanced data. 2017 International Conference on Fuzzy Theory and Its Applications, iFUZZY 2017, v. 2017- Novem, p. 1–6, 2018.

Liu, H.; Zhou, M. Decision tree rule-based feature selection for large-scale imbalanced data. 2017 26th Wireless and Optical Communication Conference, WOCC 2017, p. 1–6, 2017.

Liu, Y. et al. Symptom severity classification with gradient tree boosting. Journal of Biomedical Informatics, v. 75, p. S105–S111, 2017.

Mars, P.; Chen, J. R.; Nambiar, R. Learning Algorithms. Learning Algorithms, n. Icosec, p. 65–71, 2018.

Mu, B.; Chen, T.; Ljung, L. Asymptotic Properties of Hyperparameter Estimators by Using Cross-Validations for Regularized System Identification. Proceedings of the IEEE Conference on Decision and Control, v. 2018-Decem, n. Cdc, p. 644–649, 2019.

Rustam, F. et al. COVID-19 Future Forecasting Using Supervised Machine Learning Models. IEEE Access, v. 8, p. 101489–101499, 2020.

Sung-Bae C., Hong-Hee W. (2003). Machine Learning in DNA Microarray Analysis for Cancer Classification. Asia-Pacific Bioinformatics Conference, 19. https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV19Cho.pdf

Yaşar, Ş.; Çolak, C.; Yoloğlu, S. Artificial Intelligence-Based Prediction of Covid-19 Severity on the Results of Protein Profiling. Computer Methods and Programs in Biomedicine, v. 202, 2021.

Ye, H. et al. Diagnosing Coronavirus Disease 2019 (COVID-19): Efficient Harris Hawks-inspired Fuzzy K-nearest Neighbor Prediction Methods. IEEE Access, v. 9, p. 17787–17802, 2021.