Covid19-Mexican-Patients’ Dataset (Covid19MPD) Classification and Prediction Using Feature Importance

Khaled Mohamad Almustafa

Department of Information Systems, College of Computer and Information Systems, Prince Sultan University, Riyadh, Kingdom of Saudi Arabia

Correspondence
Khaled Mohamad Almustafa, Department of Information Systems, College of Computer and Information Systems, Prince Sultan University, Riyadh, Kingdom of Saudi Arabia.
Email: kalmustafa@psu.edu.sa

Funding information
Prince Sultan University

Summary
Coronavirus disease, Covid19, pandemic has a great effect on human health worldwide since it was first detected in late 2019. A clear understanding of the structure of the available Covid19 datasets might give the healthcare provider a better understanding of identifying some of the cases at an early stage. In this article, we will be looking into a Covid19 Mexican Patients’ Dataset (Covid19MPD), and we will apply a number of machine learning algorithms on the dataset to select the best possible classification algorithm for the death and survived cases in Mexico, then we will study the performance enhancement of the specified classifiers in terms of feature selection in order to be able to predict severe, and or death, cases from the available dataset. Results show that J48 classifier gives the best classification accuracy with 94.41% and \( \text{RMSE} = 0.2028 \) and \( \text{ROC} = 0.919 \), compared to other classifiers, and when using feature selection method, J48 classifier can predict a surviving Covid19MPD case within 94.88% accuracy, and by using only 10 out of the total 19 features.

KEYWORDS
classification, Covid19, feature importance, feature selection, machine learning, prediction

1 | INTRODUCTION

Coronavirus disease, Covid19, pandemic has a great effect on human health worldwide since it was first detected in late 2019 in Wuhan, China. Until now, many counties got affected by severe social and economic crises due to this disease. As of the end of July 2020, more than 15 million cases of Covid19 have been recorded worldwide, and more than 600,000 confirmed deaths. A clear understanding of the structure of the available Covid19 datasets gives the healthcare provider a better understanding of identifying some of the cases at an early stage can be achieved using some of the available machine learning (ML) approaches using different classifiers for the classification and or the prediction of Covid19 cases, based on the available features in their dataset for both survived and non-survived cases, such an understanding. In this article, we will be looking into a Covid19 Mexican Patients’ Dataset (Covid19MPD) that was publicly provided by the Government of Mexico, more details about the dataset will be presented in the next section. We will apply number of ML algorithms on the dataset to be able to select the best possible classification algorithm for the death and survived cases in Mexico, then we will study the performance enhancement of the specified classifiers in terms of feature selection in order to be able to predict severe, and or death, cases from the available dataset with minimal number of available features instead of using the whole dataset. Such a result can be useful to the healthcare providers in identifying and diagnose covid19 cases in a more efficient way. A feature importance approach of the provided dataset would also be presented to evaluate all features importance to Mexico region and see what was the main contributors to the severe cases of patients.

Many researchers applied ML algorithms in number of bioinformatics related datasets for the classification, prediction, feature selection, and features importance, but to the best of our knowledge applying such techniques on a Covid19 dataset, and specifically to the Covid19MPD in hands has not been done yet. In what follows, we will present a literature review and related work on the used ML algorithms in general, and present some of the few related covid19 related research. Many bioinformatics related detection and classification papers have been developed for different types...
of health related dataset in the form of disease detection and dataset classification and/or diseases predictions. In Reference 2, a classification proposed model for the electroencephalogram (EEG) of the elliptic seizure recoded waves was presented using K-mean clustering, and in References 3–5, an automatic elliptic seizure detection based on length feature and discrete wavelet transform (DWT)-based approach was proposed. EEG signal classification for elliptic seizure was presented in Reference 6, and multi-domain EEG signal’s feature extraction was shown in References 7 and 8. Time frequency and nonlinear detection analysis of EEG recorded waves was proposed in Reference 9, and detection of elliptic seizure cases using advanced ML algorithms was presented in Reference 10. Parkinson disease (PD) has its share in term of applying ML algorithms for the detection, classification and prediction of PD cases. A speech signal processing for Parkinson patients using classifications algorithms was proposed in References 11–13 and prediction of PD cases using different ML classifiers were presented in References 14–17. Also, number of chronic kidney disease (CKD) related data analysis approaches can be found in the literature, and to mention resent few, prediction of CKD using ML algorithms was presented in References 18–20. Performance evaluation for ML classification algorithm for CKD was proposed in References 21 and 22. A comparative study for the use of different classifiers to classify the CKD dataset was shown in References 23 and 24, and detection of CKD cases was presented in Reference 25. Other health related studies including the association between work-related features and coronary artery disease using ML classifiers was proposed in Reference 26, and the diagnosis of coronary artery disease was presented in References 27 and 28, heart disease classification and prediction using different ML classification algorithms and approaches were pretend in References 29–32, and prediction of cleft before birth cases using ML algorithms was proposed in Reference 33 to mention few. Dataset’s feature selections also has a fair share in the literature, such as using ++ support vector machine (SVM) for feature selection of a PD recorded wave in 34–36 as well as prediction of CKD using feature selection in Reference 37 and feature reduction and selection based on modified dominance soft set in Reference 38. Feature importance was presented in many health related studies using different ML and classifications algorithms. 39–41 Other health informatics related applications using ML and feature selection methods can also be found in References 42–47.

Covid19 disease being a newly surfacing disease has a moderate ML related approaches in the literature especially work related to classification, prediction and feature selection and reduction of available covid19 datasets. In what follow, we will present the most relevant work related to our proposed techniques. In Reference 48, the authors proposed an artificial intelligence (AI) and ML to fight COVID-19 possible diagnosed cases using the patients’ distinct respiratory pattern and by using thoracic computed tomography (CT) images for the detection and monitoring of COVID-19 patients, and using similar approach with CT images and X-ray images was presented in References 49 and 50. A classification approach for covid19 cases based on the CT images as well was shown in Reference 51 where and early phase detection of coronavirus was achieved using different methods of ML algorithms, and a case study using intrinsic genomic signatures for COVID-19 classification using ML algorithms was presented in References 52 and 53 by combining the supervised machine learning algorithms with digital signal processing (MLDSP). Covid19 vaccine design by using reverse vaccinology and ML was suggested in Reference 54. Prediction and growth of Covid19 cases using different ML algorithms by applying different mathematical model was proposed in Reference 55 and screening of covid-19 using infection size-aware random forest (RF) classification was presented in References 56 and 57. In References 58 and 59, the authors suggested a classification of covid19 dataset using X-ray and CT images by applying ML algorithms and selecting certain features.

The rest of this article is organized as follows: the preparation of the Covid19MPD is presented in Section 2, methodology is shown in Section 3, and used classifiers in Section 4, Section 5 presents the experimental results. Remarks and the conclusion were presented in Sections 6 and 7, respectively.

## 2 DATASET PREPARATION

The dataset was obtained from a publically accessible link provided by the government of Mexico for more than 500,000 cases of covid19 patients admitted to different location of the country. For the purpose of this study, only 200,000 first cases where studied, and the purpose was to evaluate the common classifiers algorithms to classify the mentioned dataset, and the samples were downsized for computational time reasons, but the selected samples have enough data to cover all possible cases as we can see in Table 1, it is worth mentioning that there was two extra features regarding the patient starting symptoms and the date he/she was admitted to hospital, but they were removed for this study for their non-relevancy to this study.

Table 1 shows the available features for the Covid19MPD, the set has 20 features, 19 distinguished features and a class attribute. Each feature has a set of possible values assigned to it and a specific reason for its selection. What is interesting about this collection of dataset that it has a fair group of attributes associated with the patient medical history, such as if the patient has suffered from diabetic illness, asthma, and other illnesses, and as well as if the patient is a smoker, or had contact with a covid19 case, and what type of healthcare the patient received since admitting to hospital. The interesting part for these wide range of features is that we can identify later on the importance of these features to the dataset in hand and identify if a specific illness would actually contribute to the death of a covid19 cases or not for the given country, in which in our case would be Mexico.

Table 2 shows the distribution values for all available feature in the dataset, based on the attribute’s possible values. We can see from the provided dataset that number 97, 98, and 99 were used in many cases for not specified cases weather a patient had a specific illness or not.
TABLE 1  Available features and their descriptions

| Feature          | Code | Value            | Note                                                                 |
|------------------|------|------------------|----------------------------------------------------------------------|
| sex              | 1, 2 | Female/male      |                                                                       |
| patient_type     | 1, 2 | PATIENT_TYPE identifies the type of care received by the patient in the unit. It is called an outpatient if you returned home or it is called an inpatient if you were admitted to hospital. |
| intubed          | 1, 2, 97, 99 | INTUBED identifies if the patient required intubation. |
| pneumonia        | 1, 2, 99 | PNEUMONIA identifies if the patient was diagnosed with pneumonia. |
| age              | Range [0:110] | Age of the tested group |
| pregnancy        | 1, 2, 98, 97 | PREGNANCY identifies if the patient is pregnant. |
| diabetes         | 1, 2, 98 | DIABETES identifies if the patient has a diagnosis of diabetes. |
| copd             | 1, 2, 98 | COPD identifies if the patient has a diagnosis of COPD. |
| asthma           | 1, 2, 98 | ASMA identifies if the patient has a diagnosis of asthma. |
| inmusrp          | 1, 2, 98 | INMUSUPR identifies if the patient has immunosuppression. |
| hypertension     | 1, 2, 98 | HYPERTENSION identifies if the patient has a diagnosis of hypertension. |
| other_disease    | 1, 2, 98 | OTRAS_COM identifies if the patient has a diagnosis of other diseases. |
| cardiovascular   | 1, 2, 98 | CARDIOVASCULAR identifies if the patient has a diagnosis of cardiovascular disease. |
| obesity          | 1, 2, 98 | OBESITY identifies if the patient is diagnosed with obesity. |
| renal_chronic    | 1, 2, 98 | RENAL_CHRONIC identifies if the patient has a diagnosis of chronic kidney failure. |
| tobacco          | 1, 2, 98 | TOBACCO identifies if the patient has a smoking habit. |
| contact_other_covid | 1, 2, 99 | OTHER_CASE identifies if the patient had contact with any other case diagnosed with SARS CoV-2. |
| covid_res        | 1, 2, 3 | RESULT identifies the result of the analysis of the sample reported by the laboratory of the National Network of Epidemiological Surveillance Laboratories (INDRE, LESP and LAVE). |
| icu              | 1, 2, 97 | ICU identifies if the patient required to enter an Intensive Care Unit. |
| class (died/survived) | 1, 2 | Indicating if the patient passed away or survived the covid19. |

Explanation of the codes given in Table 2 can be explained from Table 1. And Figure 1 shows a visual representation of the distribution of the attributes values presented in Table 2.

Figure 1 shows a visual representation of the attributes' possible values based on their available cases, and we can see that there are noticeable cases of pneumonia, diabetes, hypertension, and obesity among the tested patients, and a large number of cases were patient had a contact with other covid19 patients.

Table 3 shows the distribution of the age groups with 10 years intervals from early born to 110 years old, and we can see from Figure 2 a close to normal distribution for the tested age group, and most of the reported cases are between 30 and 50 years old. It is worth mentioning that for the selected samples of 200,000 patients, male and female patients are almost equal for most age groups.

3 | METHODOLOGY

We will present three different methods in this study:

1. The case of a direct classification exercise using common classifiers that will be explained in the next section.
2. The case of using feature selection for some of these classifiers to evaluate the classifier performance with a subset of its features.
3. Last but not least we will present a case of feature importance to study the direct effect of some of the features on the result of death for some of the patients in the presented dataset, and run a feature importance selection for attribute to evaluate the classifier classification performance using these attributes.

Figure 3 highlights the difference approached used per suggested methodology, and the process for each method from a dataset to the results.
### Table 2
The distribution of all values for the available features

| Feature Code | Female | Male |
|--------------|--------|------|
| Sex          | 98,827 | 101,173 |
| Patient type | 157,101 | 42,899 |
| INDRE        | 3427   | 39,428 |
| LESP         | 1434   | 96,842 |
| LAVE         | 25,077 | 174,235 |
| COVID Res    | 3204   | 196,182 |
| Asthma       | 6445   | 192,944 |
| COPD         | 3105   | 196,202 |
| Pregnancy    | 32,555 | 166,805 |
| COPD         | 4054   | 195,313 |
| Renal Chronic| 16,949 | 182,366 |
| Cardiovascular| 78,143 | 60,081 |
| Is ICU       | 3524   | 39,330 |
| Class (Died/Alive) | 12,714 | 187,286 |

**Figure 1** Features count distribution
TABLE 3 Age group frequency

| Age   | 0–10 | 11–20 | 21–30 | 31–40 | 41–50 | 51–60 | 61–70 | 71–80 | 81–90 | 91–100 | >100 |
|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|--------|------|
| Female| 377  | 1997  | 3711  | 18,965| 24,546| 21,666| 14,701| 7237  | 3725  | 1576   | 298  |
| Male  | 462  | 2247  | 3558  | 17,454| 23,687| 21,636| 16,182| 9299  | 4639  | 1767   | 242  |
| Total | 839  | 4244  | 7269  | 36,419| 48,233| 43,302| 30,883| 16,536| 8364  | 3343   | 540  |

FIGURE 2  Age group distribution

FIGURE 3  Used methods

4  | USED CLASSIFIERS

We will briefly discuss some of the ML classifications algorithms used in this study for the classification, feature selection, and prediction of Covid19MPD, and more details can be found in Reference 32.

4.1  | Naïve Bayes

A well-known classifier that uses the conditional probability of the occurrence of a given feature (attribute) with respect to another feature to perform the classification of the available features.
4.2 | Decision tree and random forest

In this classifier, a dataset is usually divided into smaller subsets based on provided questions (conditions) and the decision is based on the comparison of these questions between all possible classes after covering all available attributes in the dataset. A special case of the decision tree (DT) is the J48 classifier, where a unified variable will be associated with the provided dataset. RF classifier is a collection of multiple random trees classifiers and the result of the classification is taken over an average of the performance classifications of all associated trees.

4.3 | K-nearest neighbor

Nearest neighbor (NN) classifier is based on a comparison of a given sample test element to be classified to a certain class with respect to an available sample training element based on the distance between the test and the training elements, and usually “K” refers to the number of training elements that a test element can be as close as possible to be a member of their class, and this task can be done using a direct distance between two elements or for more accurate results, one can use the Euclidian distance between two samples.

4.4 | Stochastic gradient descent

Gradient descent is an algorithm that optimizes many loss functions, such as SVM and logistic regression models, and is usually used to optimize the linear function, and the stochastic concept is introduced here based on the roots finding nature of the optimization task. In stochastic gradient descent (SGD), for each iteration, samples are selected randomly using a term “batch” for number of samples, instead of the whole data set, and these batches are used to calculate the gradient for each iteration.

5 | EXPERIMENTAL RESULTS

In this section, we will go over the results obtained from using different mentioned methods and their used classifiers and compare the performance of the classifiers per method, then the comparison between methods to select the best approach for the classification and or the prediction of a Covid19MPD. Statistical parameters were used for the simulation results to compare classifiers’ performances, such as relative absolute error (RAE) for the relative error of estimation with respect to the actual value, mean absolute error (MAE) in terms of the relative error with respect to the number of instances, and area under curve (ROC) distinguish how well a given classifier is preforming in term of the identification of a specific data point, where best performance for ROC = 1.

5.1 | Direct classification using mentioned classifiers

We will present the results obtained for the direct used classifiers in term of classification accuracy, MAE, RMSE, and ROC for the purpose of performance compassion. A 10-fold cross validation technique for classification was used in this simulation.

Table 4 shows the results obtained for the direct classification of the Covid19MPD, we can see that J48 classifier gives the best classification accuracy with 94.41% and RMSE = 0.2028 and ROC = 0.919, compared to the other classifiers of accuracy of 93.64%, 93.50% and 92.71% for SGD, RF, and K-NN (K = 1), respectively. Figures 4 and 5 show the graphical representation of the results obtained in Table 4.

| Classifier used     | Accuracy (%) | MAE   | RMSE  | ROC   | Time (s) |
|---------------------|--------------|-------|-------|-------|----------|
| Naïve Bayes         | 84.23        | 0.1522| 0.3771| 0.927 | 1.37     |
| SGD                 | 93.64        | 0.0636| 0.2521| 0.50  | 23.09    |
| J48                 | 94.41        | 0.0798| 0.2028| 0.919 | 45.63    |
| Random forest       | 93.50        | 0.0774| 0.2161| 0.910 | 209.09   |
| K-NN (N = 1)        | 92.71        | 0.0771| 0.2441| 0.889 | 0.2      |
Figure 4: Accuracy percentages per classifier

Figure 5: Classifiers’ MAE and RMSE results

Table 5: Classification accuracy with feature selection

| Classifier     | Selected features (#) | Before/after feature selection (%) | Accuracy | MAE     | RMSE     | ROC       |
|----------------|------------------------|-------------------------------------|----------|---------|----------|-----------|
| K-NN (K = 1)   | 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 (17) | 92.71/92.82 | 0.0771/0.0706 | 0.2441/0.2585 | 0.889/0.837 |
| J48            | 1, 3, 4, 5, 8, 13, 14, 17, 18, 19 (10) | 94.41/94.88 | 0.0789/0.0752 | 0.2028/0.1994 | 0.919/0.895 |
| Random forest  | 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 (17) | 93.50/92.93 | 0.0774/0.0701 | 0.2161/0.2612 | 0.910/0.697 |

5.2 Feature selection classifications and prediction

In this section, we will introduce a feature selection algorithm using classifier subset evaluation to select the most contributing features (attributes) in the Covid19MPD dataset, then compute the classification accuracy for the selected classifiers with the selected feature and evaluate the performance of these classifiers based on the selected features, this method can be used for cases prediction with minimal number of attributes per dataset instead of using the entire attributes for each patient.

We see from Table 5 that the feature selected for a slightly better classification accuracy for the J48 classifier are sex, intubed, pneumonia, age, copd, cardiovascular, obesity, contact_other_covid, covid_res, and icu.
It is worth mentioning that pregnancy, diabetes, asthma, inmsupr, hypertension, other_disease, and tobacco have minimal or no effect on the classification of this specific dataset using the J48 classifier. Using J48 one can predict a surviving Covid19MPD case within 94.88% accuracy by using only 10 out of the total 19 features from the original dataset and 94.41% accuracy. And we can also see a slight improvement in the classification accuracy of the K-NN (N = 1) classifier with 92.82% accuracy with selected feature compared to 92.71% with full set. Figure 6 shows a graphical representation of accuracy results obtained in Table 5.

Figure 6 shows a graphical representation of accuracy results obtained in Table 5, and we can see an increase of the classification accuracy after feature selection for both K-NN (K = 1) and J48 classifiers.

Figure 7 shows a graphical representation of the values of MAE and RMSE for the results obtained in Table 5, and we can see decrease of the MAE value for all concerned classifiers after applying the feature selection method, as well as a slight decrease for RMSE value for the J48 after the feature selection method was applied.

5.3 Feature importance

In this section, we will introduce a feature importance method to highlight the effect of the selected features to the outcome of the results for surviving and not surviving cases of Covid19MPD based on the provided samples of patients. A feature importance method using the interaction test will be used to calculate the impact of a given feature on the resulting class of the dataset.
It is a statistical approach that measures the importance of a feature value in a dataset based on its p-value resulting from a $\chi^2$ test in a DT. Then we will do a comparison for the J48 classifier using full attributes. Feature selected using subset evaluation attributes and feature selection using interaction test attributes to evaluate and compare the classification accuracy for the three used methods.

Figure 8 shows the results obtained for positive and negative feature importance using the interaction test for feature importance on the Covid19MPD, and results show that some of the features are highly contributing to the classification of the sever and non-sever cases on Covid19. Table 6 shows the classification results in term of the accuracy, MAE and the RMSE for the three different algorithms as well as the classification result for the case where common feature was selected from the three methods. We can see that both feature selection methods outperformed the classification results for the full featured data with an accuracy of 94.88% and 94.65, and MAE = 0.0752 and 0.0778, and RMSE = 0.1995, and 0.214 respectively, compared to a full set classification of an accuracy = 94.41% and MSE = 0.789 and RMSE = 2028. Even the classification of the Covid19MPD using only the common features between the three methods outperformed the full set classification with an accuracy = 94.61%, MSE = 0.786 and RMSE = 0.200.

6 | REMARKS

Based on the presented results, the following remarks can be mentioned based on the used methodology:

1. The case of a direct classification: J48 classifier gives the best classification accuracy compared to all other used classifiers.
2. The case of using feature selection: J48 classifier can predict a surviving Covid19MPD case with 94.88%, and also it outperforms all other used classifiers.
3. The case of feature importance: Feature importance method show that this method outperformed the classification results for the full featured data, and that gives a direct importance to the major contributing feature in the classification used for the available dataset.
TABLE 6 Classification results comparison for all used methods

| Feature          | J48 all features | J48 feature selection | J48 feature selection based on feature importance | Common features |
|------------------|------------------|------------------------|--------------------------------------------------|-----------------|
| 1 sex            | sex              | sex                    | sex                                              | sex             |
| 2 patient_type   | –                | patient_type           | –                                                | –               |
| 3 intubed        | intubed          | intubed                | intubed                                          | intubed         |
| 4 pneumonia      | pneumonia        | pneumonia              | pneumonia                                        | pneumonia       |
| 5 age            | age              | age                    | age                                              | age             |
| 6 pregnancy      | –                | pregnancy              | –                                                | –               |
| 7 diabetes       | –                | –                      | –                                                | –               |
| 8 copd           | copd             | copd                   | copd                                             | copd            |
| 9 asthma         | –                | asthma                 | –                                                | –               |
| 10 inmsupr       | –                | –                      | –                                                | –               |
| 11 hypertension  | –                | hypertension            | –                                                | –               |
| 12 other_disease | –                | –                      | –                                                | –               |
| 13 cardiovascular| cardiovascular   | cardiovascular          | cardiovascular                                   | cardiovascular   |
| 14 obesity       | obesity          | –                      | –                                                | –               |
| 15 renal_chronic | –                | –                      | –                                                | –               |
| 16 tobacco       | –                | –                      | –                                                | –               |
| 17 contact_other_covid | contact_other_covid | –                | –                                                | –               |
| 18 covid_res     | covid_res        | covid_res              | covid_res                                        | covid_res       |
| 19 icu           | icu              | –                      | –                                                | –               |
| **Accuracy (%)** | 94.41            | 94.88                  | 94.65                                            | 94.61           |
| **MAE**          | 0.0789           | 0.0752                 | 0.0778                                           | 0.0786          |
| **RMSE**         | 0.2028           | 0.1994                 | 0.2014                                           | 0.200           |

7 | CONCLUSION AND FUTURE WORK

ML algorithms, such as naïve Bayes, SGD, RF, KNN (K = 1), and J48 Classifiers were applied on the Covid19MPD to select the best possible classification algorithm for the selection of the death and survived cases in Mexico, then the performance enhancement of the specified classifiers in term of features selection was performed, such a task can be useful to the healthcare providers in identifying and diagnose covid19 cases in a better efficient ways, also, a feature importance algorithm was applied on the mentioned dataset to evaluate all features importance to Mexico region based on the available dataset, and to understand what was the main contributors to the sever cases of patients.

Results show that J48 classifier gives the best classification accuracy with 94.41% and RMSE = 0.2028 and ROC = 0.919, compared to the other classifiers of accuracy of 93.64%, 93.50%, and 92.71% for SGD, RF, and K-NN (K = 1), respectively. When using the feature selection method, J48 classifier can predict a surviving Covid19MPD case with 94.88% accuracy and by using only 10 out of the total 19 available features, which can be a useful fact for healthcare providers in identifying possible infected Covid19MPD cases. Results for the classification using feature selection, based on the feature importance method, show that this method outperformed the classification results for the full featured data with an accuracy of 94.65%, MAE = 0.0778, and RMSE = 0.214.

As a future work extension to this work would be to encourage researchers to investigate and develop a feature importance evaluation of the corona virus for different countries/regions, where such an investigation would assess the extent of the possible mutation of Covid19 per country or region, and ideally if the importance of features of several regions were to be evaluated, it may be able to reflect on the possible mutation of the virus per country/region and would help vaccine developers to pin point the required treatment/vaccine based on the selection of the important features that contribute to the well-being of the Covid19 patients.

ACKNOWLEDGMENTS
The author would like to thank Prince Sultan University, Riyadh, KSA, for supporting this work.
DATA AVAILABILITY STATEMENT
Data were used from a publicly available dataset https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset/data.

ORCID
Khaled Mohamad Almustafa  
https://orcid.org/0000-0003-2129-7686

REFERENCES
1. https://www.gob.mx/salud/documentos/datos-abiertos-152127
2. Orhan U, Hekim M, Özer M. EEG signals classification using the K means clustering and a multilayer perceptron neural network model. Expert Syst Appl. 2011;38:13475-13481.
3. Guo L, Rivero D, Dorado J, Rabunal JR, Pazos A. Automatic epileptic seizure detection in EEG based on line length feature and artificial neural network. J Neurosci Methods. 2010;191:101-109.
4. Kumar Y, Dewal ML, Anand RS. Epileptic seizures detection in EEG using DWT-based ApEn and artificial neural network. Signal Image Video Process. 2014;8:1323-1334. https://doi.org/10.1007/s11760-2010012-20100362-20109
5. Sharma, A, Madan S, Srivastava K. Epilepsy detection using DWT based Hurst exponent and SVM. K-NN classifiers. Serbian J Exper Clin Res. 2019;19(4):311-319. https://doi.org/10.1515/10.1515/sjerc-20102017-20100043
6. Subasi A, Gursoy M. EEG signal classification using PCA, ICA, LDA and support vector machine. Expert Syst Appl. 2010;37:8659-8666.
7. Wang L, Xue W, Yang L, et al. Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis. Entropy. 2017;19:222. https://doi.org/10.3390/e19060222
8. Nicolaou N, Georgiou J. Detection of epileptic electroencephalogram based on permutation entropy and support vector machine. Expert Syst Appl. 2012;39:202-209.
9. Gajic D, Djurovic Z, Gilgjonjievic J, Di Gennaro S, Savic-Gajic I. Detection of epileptiform activity in EEG signals based on time-frequency and non-linear analysis. Comput Neurosci. 2015;9:38. https://doi.org/10.3389/fncom.2015.00038
10. Fergus P, Hignett D, Hussain A, Al-Jumeily D, Abdel-Aziz K. Automatic epileptic seizure detection using scalp EEG and advanced artificial intelligence techniques. BioMed Res Int. 2015;2015:986736. https://doi.org/10.1155/2015/986736
11. Sakar CO, Serbes G, Gunduz A, et al. A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable Q-factor wavelet transform. Appl Soft Comput. 2019;74:255-263.
12. Tsanas A, Little M, McSharry P, Spielman J, Ramig L. Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease. IEEE Trans Biomed Eng. 2012;59(3):1264-1271.
13. Tsanas A, Little M, McSharry P, Ramig L. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity. J R Soc Interface. 2010;8(39):842-855.
14. Idris A, Ifikhar A, Rehan Z. Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. Cluster Comput. 2017;22(3):7241-7255.
15. Sakar CO, Kursun O, Seker H, Gurgen F. Combining multiple clusterings for protein structure prediction. Int J Data Min Bioinform. 2014;10(2):162-174.
16. Sakar BE, Isenkul M, Sakar C, et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inform. 2013;17(4):828-834.
17. Peker M. A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. J Med Syst. 2016;40(5):116.
18. Vijayarani S, Dhayanand S, Phil M. Kidney disease prediction using SVM and ANN algorithms. Int J Comput Bus Res. 2015;6(2):1-12.
19. Chimwaviyike, K. R. N. C. E. N. Risk level prediction of chronic kidney disease using Neuro-fuzzy and hierarchical clustering algorithm (s). Int J Multimed Ubiquitous Eng. 2017;12(8):23-36. https://doi.org/10.14257/ijmube.2017.12.8.03
20. Zeynu S, Patil S. Survey on prediction of chronic kidney disease using data mining classification techniques and feature selection. Int J Pure Appl Math. 2018;118(8):149-156.
21. Gunaratne WHSD, Perera KDM, Kahandawarachchi K. KADCP. Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). Proceedings of the IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE); 2017:291-296. IEEE.
22. Radha N, Ramya S. Performance analysis of machine learning algorithms for predicting chronic kidney disease. Int J Comput Sci Eng. 2015;3(8):72-76.
23. Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using SVM and KNN. Expert Syst Appl. 2015;4(12):608-612.
24. Pratibha Devishi S, Ragan OR, Anisha GS. Comparative study of classification algorithms in chronic kidney disease. Int J Recent Technol Eng (IJRTE). 2019;8(1):180-184.
25. Arora M, Sharma EA. Chronic kidney disease detection by analyzing medical datasets in Weka. Int J Comput Appl. 2016;6(41):20-26.
26. Nasarian E, Abdar M, Fahami MA, et al. Association between work-related features and coronary artery disease: a heterogeneous hybrid feature selection integrated with balancing approach. Pattern Recogn Lett. 2020;133:33-40.
27. Muhammad LJ, Garba EJ, Oye ND, Wajiga GM. On the problems of knowledge acquisition and representation of expert system for diagnosis of coronary artery disease (CAD). Int J e-Serv Sci Technol. 2018;11(3):49-58.
28. Alizadehsani R, Roshanzamir M, Abdar M, et al. Hybrid genetic-discretized algorithm to handle data uncertainty in diagnosing stenosis of coronary arteries. Expert Systems. 2020. https://onlinelibrary.wiley.com/doi/10.1111/exsy.12573
29. Abdar M, Ksia˛e, R. M, Abdar M, Davarzani Z, Zhou X, Plawiak P, Acharya UR. Hybrid real-binary particle swarm optimization for rule discovery in the diagnosis of coronary artery disease; 2019;38: Wiley Online Library. https://doi.org/10.1111/exsy.12485
30. Kandala BN, Dhilli R, Plawiak P, et al. Towards real-time heartbeat classification: evaluation of nonlinear morphological features and voting method. Sensors. 2019;19(23):5079.
31. Almustafa KM. Prediction of heart disease and classifiers’ sensitivity analysis. BMC Bioinform. 2020;21(1):1-18.
33. Bukhari F, Shafi N, Asif M, Almustafa K, Iqbal W, Nawaz Z. Cleft prediction before birth using deep neural network. Health Inform J (HIJ). 2020;26(4):2568-2585.

34. Pal M, Foody G. Feature selection for classification of hyperspectral data by SVM. IEEE Trans Geosci Remote Sens. 2010;48(5):2297-2307.

35. Sakar C, Kursun O, Gorgen F. A feature selection method based on kernel canonical correlation analysis and the minimum redundancy-maximum relevance filter method. Expert Syst Appl. 2012;39(3):3432-3437.

36. Erdogdu Sakar B, Serbes G, Sakar C. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PLoS One. 2017;12(8):e0182428.

37. Tazin N, Sabab SA, Chowdhury MT. Diagnosis of chronic kidney disease using effective classification and feature selection technique. Proceedings of the 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Electronic; 2016.

38. Jothi G, Inbarani HH, Azar AT, Almustafa KM. Feature reduction based on modified dominance soft set. Proceedings of the 5th International Conference on Fuzzy Systems and Data Mining (FSDM2019); October 18-21, 2019; Kitakyushu, Japan.

39. Janitza S, Strobl C, Boulesteix A-L. An AUC-based permutation variable importance measure for random forests. BMC Bioinform. 2008;9(1):307. https://doi.org/10.1186/10.1186/1471-20102105-20109-2010307

40. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2017;27(3):659-678.

41. Too J, Mirjalili S. A hyper learning binary dragonfly algorithm for feature selection: a COVID-19 case study. Knowl Based Syst. 2021;212:106553.

42. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19; 2020:200-202.

43. Kassania SH, Kassanib PH, Wesolowski MJ, Schneidera KA, Detersa R. Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning-based approach; 2020. arXiv preprint arXiv:2004.10641.

44. Ozturk T, Talo M, Yildirim EA, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020;121:103792.

45. Barstugan M, Ozturk S, Ozkaya U. Artificial neural network classification of high-dimensional data with novel optimization approach of dimension reduction. Ann Data Sci. 2018;5(4):615-635.

46. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. Chaos Soliton Fract. 2020;139:110059.

47. Ong E, Wong MU, Hoffmann A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning; 2020. BioRxiv.

48. Shi F, Xia L, Shan F et al. Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification; 2020. arXiv preprint arXiv:2003.09860.

49. Shi F, Xia L, Shan F et al. Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification; 2020. arXiv preprint arXiv:2003.09860.

50. Loh WY. Regression trees with unbiased variable selection and interaction detection. Stat Sin. 2002;12(2):361-386.

51. Barstugan M, Ozkaya U, Ozturk S. Coronavirus (covid-19) classification using ct images by machine learning methods; 2020. arXiv preprint arXiv:2003.09424.

52. Ozkaya U, Ozturk S, Barstugan M. Coronavirus (COVID-19) classification using deep features fusion and ranking technique. Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach. Cham: Springer; 2020:281-295. https://link.springer.com/chapter/10.1007/978-3-030-55258-9_17

How to cite this article: Almustafa KM. Covid19-Mexican-Patients’ Dataset (Covid19MPD) Classification and Prediction Using Feature Importance. Concurrency Computat Pract Exper. 2022;34(4):e6675. doi: 10.1002/cpe.6675