COVID-19 morbidity and mortality forecast in megalopolis: a data approach to public health management in São Paulo, Brazil

Demian da Silveira Barcellos (demian.barcellos@gmail.com)
Pontifical Catholic University of Paraná
Giovane Matheus Kayser Fernandes
Pontifical Catholic University of Paraná
Fabio Teodoro de Souza
Pontifical Catholic University of Paraná

Research Article

Keywords: data mining, clustering, statistics, public health, COVID-19

Posted Date: November 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-111897/v1

Abstract

The need for a scientific approach that can provide subsidies to governments and public health authorities in decision making to face pandemics, epidemics and endemics was one of the aspects recognized worldwide with the first wave of COVID-19. This article presents a methodology for the application of data mining as a support tool for coping with epidemic diseases. The methodological approach was applied in the city of São Paulo, Brazil, with the aim of predicting the evolution of COVID-19 in the metropolis and identifying air quality and meteorological variables correlated with confirmed cases and deaths from the coronavirus. Forecasting public health conditions is useful for preparing health teams in advance for a pandemic to prevent the system from collapsing. The statistical analyzes indicated the most important explanatory environmental variables, while the cluster analyzes showed which are the best input variables for the forecasting models. The forecast models were built by two different algorithms, J48 (C4.5) and CBA, and their results have been compared. The models developed can be used to predict new cases and deaths by COVID-19 in São Paulo. The methodological approach can be applied in other cities and for other epidemic diseases.

Background

The new 2019 coronavirus (COVID-19) is the biggest health challenge that humanity has faced since the Spanish flu outbreak of 1918[1]. Its rapid transmission caused the virus to spread to all continents in a short period of time. In the absence of drugs or vaccines, non-pharmaceutical interventions were the first strategies governments adopted. Asian countries such as China, Taiwan, Singapore, South Korea and Japan with good experience and training in epidemic management, have employed social isolation and mass screening measures that have succeeded locally in controlling the spread of the virus[2,3,4,5,6] but with economic and social impacts that cannot yet be fully dimensioned[7]. Even with the worldwide race for the development of the vaccine and the search for appropriate treatments for quarantine, lockdown and circulation restrictions have been the most adopted strategies by governments. The main basis of this approach was the comparative forecasting models, between the evolution of contagions and deaths with and without isolationmeasures[1]. These mathematical models were so important that they changed the direction of the response of some countries. In the United Kingdom, for example, Imperial College London modeling was able to make the government change its position, not to adopt any intervention measure, and to enforce quarantine[8].

Isolation measures seek to delay major outbreaks and level the demand for hospital beds in order to prevent the collapse of health systems[9], a tragedy that has been well observed by the world in the region of Lombardy in Italy[10]. Numerous studies of modeling and data analysis have been carried out to provide support to managers. A first challenge for data scientists has been to predict the evolution of cases and deaths from the disease in different social, environmental and economic contexts[11,12,13,14,15,16,17,18,19,20,21]. A second challenge is to estimate in each region what is the flattening of the contagion curve necessary for the health system to not collapse and what measures are necessary

to achieve this hypothetical scenario[1,22]. A third challenge has been to understand which social, economic and environmental variables are correlated with viral dynamics[23,24,25,26,27,28]. Temperature, for example, was a variable that has been speculated since the virus emerged as important, since experts pointed out that low temperatures are more conducive to viral transmission, and subsequently, studies confirmed this hypothesis[23,24,25,26,27,28]. The rapid response from Asian countries was a global exception in a context where most countries acted too late. This fact demonstrates the lack of global preparation to face diseases where there is an absence of an agile, fast and efficient methodology to provide subsidies to managers.

This article proposes a methodological approach that has been used to support decision making in several areas[30,31,32], including recently in urban health[33]. Although such an approach has been little used, it is a potential tool in the context of the current health crisis caused by COVID-19 and can contribute to decision-making in facing the virus and other epidemic diseases. This approach makes it possible to identify variables (environmental, climatic, social, etc.) correlated to the disease and allows the prediction of its evolution in a coordinated and agile way with a high degree of accuracy. The scientific community has been working intensively to identify variables related to COVID-19[23,24,25,26,27,28] and to predict the evolution of the disease[11,12,13,14,15,16,17,18,19,20,21,22] however, these efforts have taken place separately. This article proposes an approach that integrates the prediction and characterization of explanatory variables quickly and contributes important information to managers. The imminent need to improve preparedness to face outbreaks of epidemic diseases is a spoil that this first wave of COVID-19 left. To apply the proposed approach in a fruitful way, the data studied were from the city of São Paulo, Brazil, the most populous in the Southern Hemisphere, being in population terms the eighth largest city in the world with more than 12 million inhabitants[29]. The contributions of this study are:

- Provide a methodology for using data mining as a tool for public health management in metropolises;
- Identify climatic and air quality variables that are correlated with the number of COVID-19 cases and deaths;
- Present and compare the first forecasting models for COVID-19 based on association rules;
- Provide forecasting models of new cases and daily deaths by COVID-19 in São Paulo.

Study area context. Until the second half of September, USA, India and Brazil, respectively, are the 1st, 2nd and 3rd countries in the world with more cases of COVID-19[34,36]. The spread of the virus occurs in the urban fabric, consequently, large cities are the first focus of outbreaks. Brazil has more than 139 thousand deaths caused by the disease and 4.6 million reported cases [35]. The city of São Paulo is the epicenter of COVID-19 in Brazil, with more than 285 thousand cases and 12.4 thousand deaths according to official data[36]. In Brazil, the second city with the most deaths from the virus is Rio de Janeiro, with 10.7 thousand deaths, but a much smaller number of confirmed cases, about 99 thousand cases[36]. The significant difference in cases is apparently the result of underreporting which is higher in the city of Rio

de Janeiro, the lethality rate of the virus (proportion of infected people who died) in the city is the highest in Brazil, about 10.7%[36], and provides subsidies for this hypothesis. The remaining Brazilian cities have a much lower number of deaths and cases of COVID-19 than São Paulo and Rio de Janeiro. Fortaleza, for example, the third city in the country with the most cases, has about 3,800 deaths and 48,700 confirmed cases[36].

In the city of São Paulo, the total number of confirmed cases and deaths of COVID-19 in 2020 surpassed in July all cases and deaths from compulsory notification diseases in 2019. The disease with the highest number of cases and deaths in 2019 were, respectively, dengue with about 17 thousand cases and severe acute respiratory syndrome (SARS) with 235 deaths[37]. These data allow to contextualize the relevance of the new coronavirus in the context of the metropolis of São Paulo. Due to low testing and technical and political problems regarding the counting of cases in Brazil, underreporting is quite high. The number of cases can be up to 12 times greater than that reported as indicated by investigations[38]. The notified cases of SARS in the city of São Paulo at the beginning of July 2020 were already around 22 thousand, about ten times higher than in the whole year of 2019, and the deaths already exceed 7 thousan [37], about 30 times more than those killed by the disease in the previous year, which shows that these cases are probably COVID-19.

Research Methodology

Data acquisition and preparation. The data analyzed for the city of São Paulo are from February 25, 2020 to July 1, 2020. The patterns between climatic, air quality and epidemiological data (related to COVID-19) were analyzed. The epidemiological database used is made available in real time by the state health departments and can be accessed directly through its electronic address (https://brasil.io/COVID-19/). Data on the isolation index are also available in real time by the state government at its own electronic address (https://www.saopaulo.sp.gov.br/coronavirus/isolamento/). While climatic data were obtained on the online platform of the Agrometeorological Monitoring System (Agritempo) of the National Institute of Meteorology (INMET) (https://www.agritempo.gov.br/agritempo/produtos.jsp?siglaUF=SP) and Air quality data were collected from the international platform “Air Quality Historical Data Platform” (https://aqicn.org/data-platform/register/) and the platform is provided by CETESB (State of São Paulo Environmental Company). The climatic data are from the automatic Meteorological Station of INMET (23 K 333498.53 m E; 7405721.27 m S) which among the six stations of Agritempo in SP is the one with the highest data continuity and number of monitored variables. The air quality data are from the CETESB Parque D. Pedro II automatic Air Quality Station (23 K 333573.00 m E; 7394924.00 m S) which among the 18 air quality stations in the city is in a strategic position, which represents the central region of the city is relatively close to the chosen weather station, and with good data continuity.

A total of 26 variables were studied over 97 consecutive days. The variables studied were: new deaths (ND), new confirmed (NC), total confirmed (total_conf), total deaths (total_deaths), mortality rate (death_rate -%), number of confirmed per 100 thousand inhabitants (case_per100k_inh), isolantion index

(isol_avg_index -%), minimum (tMin - Cº), average (tAvg - Cº) and maximum (tMax– Cº) temperature, drought (drought - days without rain), agricultural drought (dry_drought - days without rain) at 10 mm), wind speed (wind_spe - kmh-^{1}), minimum dew point (dew_pointMin - Cº) and maximum (dew_pointMax - Cº), minimum atmospheric pressure (atm_pressMin - HpA) and maximum (atm_pressMax - HpA), potential evapotranspiration (pot_eva - mmd-^{1}), real evapotranspiration (real_evapo - mmd-^{1}), minimum (urMin -%) and maximum (urMax -%) humidity, soil water availability (soil_wat_avail -%), particulate material in the air (pm25 and pm10), ozone (o3) and nitrogen dioxide (no2). The air quality variables (pm25, pm10, o3 and no2) are normalized in the format of the air quality index (AQI), for each pollutant, using the United States Environmental Protection Agency (USEPA) standard.

In the stage of data preparation, in addition to the unification of the database and standardization, according to the requirements of the statistical and modeling software, the data was discretized, since the modeling software has this requirement. Discretization, which consists of transforming a continuous variable into a categorical one, was done using the statistical tertile (low / 0-33%, medium / 34-66% and high / 67-100%) for each variable.

Multivariate analyzes. Statistical analysis and data grouping have the function of characterizing the database and identifying patterns of association between variables[32]. This step, in addition to guiding the development of the models[31], can indicate the most important variables from the management point of view. Four analyzes were performed (linear correlation, factor analysis, similarity dendrograms and k-means) using the Statistica software (developed by StatSoft). For factor analysis and linear correlation, strong correlations were considered to be positive or negative values greater than or equal to 0.6[39]. The similarity dendrograms were constructed from the Euclidean distance.

Data modeling. The modeling step consisted of developing predictive models for deaths and new cases on seven consecutive days (t+1, t+2, t+3, t+4, t+5, t+6, t+7). The two tools used in this stage were CBA (Classification Based on Associations) from the School of Computing, National University of Singapore[40] and J48, open implementation of the C4,5[41,42] algorithm in the Waikato Environment for Knowledge Analysis (Weka) tool developed by New Zealand University of Waikato[43]. These two modeling tools have similar principles as they use association rules to generate a classifier. However, the essential difference is that J48 is just a classifier expressed as decision trees (set of rules that make the classification of the target variable) [41]. The algorithm of this tool, C4.5, is one of the most important and widespread in the field of data mining[42]. However, because it is a classifier, a predetermined target is needed to generate the predictive model. While the CBA is a tool that integrates classification and association rules, this allows the analysis of existing standards in the database, the association rules, to also guide the construction of classifiers[40]. The CBA only works with discrete intervals, while the J48 decision tree can work with continuous data, but there is a significant reduction in accuracy, which can reach up to 20% in these cases[44]. Therefore, the J48 also opted for the use of the same categorical intervals developed for the CBA.

The classification rules have the format IF (A) / THEN (B) where from an interval (categorical variable) or a set of intervals it is possible to predict (classification rules). Therefore, the rules express an antecedent value (A) and a consequent value (B). So IF “A>1 THENà B>3”. The rules have a support (S%) that corresponds to the percentage of records, A and B, which were classified correctly, in relation to all records in the database. Accuracy or reliability indicates the percentage of records that the rule forecasting was correct. The confusion matrix is a table that presents the classification frequencies for each class of the model (true positive, false positive, false true and false negative). The classifier's accuracy is the sum of the main diagonal (correct classifiers) of the matrix, divided by the total of values and multiplied by 100.

Results

Methodological approach.

The data approach proposed by this research to support public health in coping with pandemics, epidemics and endemics can be seen in the sequential diagram in Figure 1. Association rules are also useful tools in discovering patterns between the variables involved, and can be used, if necessary, if those patterns are not yet discovered by statistical analysis. In applying this approach in São Paulo, the statistical analyzes were conclusive and revealed the associations between the variables through factor analysis and linear correlation. The analysis of the input variables of the models can also be supported by the association rules, if it is not clear in the data cluster analysis (k-means and dendrograms).

Relevance variables.

The statistical analysis of the linear correlation coefficient showed a correlation relationship between cases of COVID-19 and climatic variables (Table 1). The same did not happen with air quality variables. Air temperature showed an inversely proportional correlation to new confirmed cases and deaths caused by the virus, in line with several studies that have already been carried out. But it was the climatic variables related to humidity that were most prominent in this statistical analysis. The agricultural drought and the total amount of water available in the soil were the main highlights, but the actual evapotranspiration also proved to be important.

In the same direction as the linear correlation coefficient, the PCA analysis showed that exactly the same climatic variables remain grouped in the main component that are the new confirmed cases and deaths by COVID-19 (Table 1). It is also the factorial loads of these variables that have the greatest correlation with the main components, as can be seen in the values highlighted in gray in Table 1. Therefore, it can be said that factor 1 corresponds to the main component of COVID-19 variables and in this component they have significant factor loads, greater than or equal to 0.6, the most important climatic variables. While the main component that corresponds to factor 2 is represented by climatic and air quality variables.

Table 1. Factor load coefficients and linear correlation.

<_- 0.6_ (Marked) _> + 0.6	Factor 1	Factor (2)	Factor (3)	Coefficient r (environmental vs. epidemiological variables with p <0.05)	
pm25	0.00	0.37	0.46	tMin X total_conf	-0,68
PM10**	-0.29	0.84	0.13	tMin X conf_per100k inh	-0.67
O3	0.32	0.60	0.50	tAvg X total_conf	0.65
NO2	-0.22	0.75	-0.34	tAvg X total_deaths	-0.64
tmin	0,79	0,21	-0.41	dry_droughtX total_conf	0.97
tAvg	0.75	0.58	0.13	dry_droughtX conf_per100k inh	0.94
Tmax	0.58	0.74	0.08	dry_droughtX NC	0.70
drought	0.37	0.38	0.09	dry_droughtX total_deaths	0.98
dry_drought	0.97	0.067	-0.06	dry_droughtX ND	0.64
urmin	0,05	-0,69	-0.55	dry_droughtX death_rate	0.64
urMax	0.14	-0.19	-0.12	dry_droughtX NC1	0.69
pot_eva	0.57	0.70	0.19	dry_droughtX NC2	0.70
wind_spe	-0.01	0.54-	-0.10	dry_droughtX NC3	0.71
dew_poitMin	0.58	-0.16	0.59	dry_droughtX NC4	0.72
dew_poitMax	0.71	0.11	-0.44	dry_droughtX NC5	0.76
pres_atmMin	-0.44	0.00	0,27	dry_droughtX NC6	0.76
pres_atmMax	-0.48	0.03	0,24	dry_droughtX NC7	0.72
real_evapo	0.85	0.14	-0.09	dry_droughtX ND1	0.65
soil_wat_avail	0.97	0.13	0.01	dry_droughtX ND2	0.67
total_conf	0.94	0.01	-0.10	dry_droughtX ND3	0.67
conf_per100k_inh	0.91	0.09	-0.05	dry_droughtX ND4	0.68
NC	0.70	0.13	0.51	dry_droughtX ND5	0.71
total_deaths	0.94	0,04	-0.10	dry_droughtX ND6	0.69
ND	-0.67	0,21	-0.45	real_evapo X death_rate	-0.79
isol_avg_index	-0.12	-0.17	0.36	real_evapo X total_deaths	-0,68
death_rate	-0,77	0,10	0.09	soil_water_avail X total_conf	0.89
NC1 +	0.71	0,20	0.32	soil_water_avail X conf_per100k_inh	-0.86
NC2	0.71	0.23	0.12	soil_water_avail X NC	-0,69
NC3	0.70	0,10	0.36	soil_water_avail X total deaths	0.90
nC4	- 0.72	0.03-	0.33	soil_water_avail X ND	-0,68
nC5	0.74	-0.18	0,10	soil_water_avail X death rate	0.81
NC6	-0.75	0.13	- 0.26%/ ℃	soil_water_avail X NC1	0.70
NC7	- 0.72	0.06	0.51	soil_water_avail X NC2	0.70
ND1	-0,68	0.18	-0.14	soil_water_avail X NC3	-0,68
ND2	-0,68	0,20	0,26	soil_water_avail X NC4	-0,68
ND3	-0.67	0.14	0.31	soil_water_avail X NC5	0.70
ND4	-0,69	0.07	0.23	soil_water_avail X NC6	0.70
ND5	0.73	0.13	0.01	soil_water_avail X NC7	-0,68
ND6	0.70	0.02	-0.30	soil_water_avail X ND1	-0,68
ND7	0.61	0.22	-0.45	soil_water_avail X ND2	-0,69
				soil_water_avail X ND3	-0,69
				soil_water_avail X ND4	-0,69
				soil_water_avail X ND5	0.71
				soil_water_avail X ND6	0.70
				soil_water_avail X ND7	-0.64

The two cluster analyzes, in addition to complementing the characterization of the data initiated with the statistical analyzes, can indicate which are the best input variables for the generation of models. By the similarity dendrogram, constructed with the Euclidean distance between the variables, it is possible to identify 4 clusters, represented by the dashed lines in Figure 2. Cluster 3, which groups total deaths, minimum and maximum atmospheric pressure, new cases (NC) and the forecast of confirmed for the seven consecutive days (NC1, NC2, NC3, NC4, NC5, NC6 and NC7) shows the input variables for NC forecasting models. However, as the variables, deaths (ND), ND1, ND2, ND3, ND4, ND5, ND6 and ND7 that are focal variables of interest, are in a grouping of the dendrogram (cluster 4) that makes it difficult to visually separate. In the k-means method, which is interactive and which classifies the distance between variables in constant spaces, the variables were divided into five groups, instead of four. Thus, the variables in cluster 4 (Figure 2) were divided into two groups, where in one of them it was formed by pm25, maximum relative humidity, isolation index, new deaths (DN) and the death forecast for the seven days (ND1, ND2, ND3, ND4, ND5, ND6, ND7), the other variables were in the other cluster. And by the k-means method, groupings 1, 2 and 3 of dendrogram variables (Figure2) remained identical. In this way, it was possible to identify that, in order to predict new confirmed ones, the model's input variables may be in addition to NC and total number of deaths at minimum and maximum atmospheric pressure. While it was possible to visualize that to predict deaths by COVID-19, as input variables in addition to the ND the isolation index, maximum relative humidity and pm25 variables can be used.

Predictive models. Figure 3 shows the accuracy of the two modeling tools used to predict new cases and deaths by COVID-19 in seven consecutive days in SP.The forecasts allow to identify the NC and ND of the next days based on the categorical intervals of the tertiles. Thus it is possible to know for the next few days if the number of new cases (patients) and deaths will correspond to the low (33%), medium (66%) or high (100%) value of the tertiles (NC: 115, 650 and 3500 patients; ND: 10, 45 and 150 deaths). As a general result, CBA's performance is slightly higher than J48. However, to predict NC on the 4th day, the accuracy of the two models is the same, 85%, on the other days the CBA has superior performance. While to predict ND it is only on the 5th day that these two modeling tools have equivalent accuracy of 89%.

The CBA generated 166 classifiers for predicting deaths within a week and 152 classifiers for predicting new cases of COVID-19. These rules were assessed for their support, accuracy and environmental and epidemiological coherence in their relationship. The choice of classifiers also prioritized those that combined variables related to COVID-19 with environmental variables (climatic and air quality), especially the input variables pointed out by the cluster analysis (atmospheric pressure, relative humidity, insulation index and pm25).

In the selection of the rules in Table 2, a diversification of the exit intervals was sought, so that there was a good representation of the three tertiles, small, medium and large. A total of thirty-eight models were selected to predict new cases and deaths within seven days ahead (Table 2). All selected predictive rules can be used as a decision support tool by managers and authorities in the city of São Paulo.

Table 2. Predictive models generated by the CBA.

Day	New canfirmed (NC)	S%	New deaths (ND)	S%
1st	IF: wind_spe _> _ 2 and NC _ <_ 115 THENà NC1 _ <_ 115	16.5	IF: 928 _ <_ pres_atmMax__930 and total_conf__1044 THENà ND1 _ <_ 10	19.6
			IF: total_conf _ <_ 1044 and wind_spe _> _ 2 THENà ND1 _ <_ 10	17.5
	IF: 928 _ <_ pres_atmMax__930 and 1044__total_conf__18000 THENà 115 _ <_ NC1__650	10.3	IF: 115 _ <_ NC__650 and pm25__43 and 10__ND__45 THENà 10 _ <_ ND1__45	11.3
	IF: 45 _ <_ isol_avg_index__55 and conf_per100k_inh__16000000 and no2 _> _ 12 THENà NC1 _> _ 650	13.4	IF: real_evapo _ <_ 1 and 15__tMin__17 and ND _> _ 45 THENà ND1 _> _ 45	10.3
2nd	IF: pres_atmMin _ <_ 928.5 and soil_wat_avail _> _ 60 and wind_spe _> _ 2 and total_conf _ <_ 1044 THENà NC2 _ <_ 115	14.6	IF: 928 _ <_ pres_atmMax__930 and soil_wat_avail__60 and total_conf__1044 THENà ND2 _ <_ 10	16.7
	IF: pres_atmMin _ <_ 928.5 and 1044__total_conf__18000 and isol_avg_index__55 THENà 115 _ <_ NC2__650	10.4		
	IF: ND _> _ 45 and 14 _ <_ dew_poitMax__16 THENà NC2 _> _ 650	12.5	IF: 1044 _ <_ total_conf__18000 and dew_poitMin__12 THENà 10 _ <_ ND2__45	8.3
3rd	IF: conf_per100k_inh _> _ 16000000 and dry_drought _> _ 40 and dew_poitMax _ <_ 14 THENà NC3 _> _ 650	13.7	IF: o3 _ <_ 23 and 1044__total_conf__18000 THENà 10 _ <_ ND3__45	8.3
	IF: pres_atmMin _ <_ 928.5 and 1044__total_conf__18000 and dew_poitMax__16 and tAvg__22 THENà 115 _ <_ NC3__650	8.4		
	IF: total_conf _ <_ 1044 and pres_atmMax__928 THENà NC3 _ <_ 115	12.6	IF: pres_atmMax _> _ 930 and urMin _ <_ 33 and 115 _ <_ NC__650 THENà ND3 _> _ 45	8.3
4th	IF: 1044 _ <_ total_conf__18000 and tMin__17 THENà 115 _ <_ NC4__650	8.5	IF: 928 _ <_ pres_atmMax__930 and soil_wat_avail__60 and total_conf__1044 THENà ND4 _ <_ 10	17
	IF: 115 _ <_ NC__650 and dew_poitMin__7.8 and urMin__33 and wind_spe__1.5 THENà NC4 _> _ 650	8.5	IF: pres_atmMax _> _ 930 and 115 _ <_ NC__650 and dew_poitMin__7.8 THENà ND4 _> _ 45	9.6
5th	IF: no2 _ <_ 8 and conf_per100k_inh _ <_ 4000000 and tAvg _> _ 22 THENà 115 _ <_ NC5__650	8.6	IF: 928 _ <_ pres_atmMax__930 and soil_wat_avail__60 and NC__115 THENà ND5 _ <_ 10	16.1
	IF: conf_per_100k_inh _> _ 16000000 and NC _ <_ 115 THENà NC5 _ <_ 115	21.5	IF: no2 _ <_ 8 and 115__NC__650 and conf_per100k_inh _ <_ 4000000 THENà 10 _ <_ ND5__45	8.6
	IF: soil_wat_avail _> _ 60 and o3 _> _ 30 and NC _ <_ 115 THENà NC5 _ <_ 115	11.9	IF: 115 _ <_ NC__650 and dew_poitMin__7.8 and tAvg _ <_ 20 THENà ND5 _> _ 45	9.7
6th	IF: 928 _ <_ pres_atmMax__930 and soil_wat_avail__60 and NC _ <_ 115 THENà NC6 _ <_ 115	16.3	IF: soil_wat_avail _> _ 60 and wind_spe _> _ 2 and NC _ <_ 115 THENà ND6 _ <_ 10	16.3
	IF: 0 _ <_ soil_wat_avail__60 and 115__NC__650 and urMin__33	10.9	IF: 7.8 _ <_ dew_poitMin__12 and pm25__55 and ND _> _ 45	12

	THENà 115 _ <_ NC6__650		THENà ND6 _> _ 45	
	IF: 7.8 _ <_ min_dew_poit__12 and pm25__55 and ND _> _ 45 THENà NC6 _> _ 650	12		
7th	IF: conf_per100k_inh _> _ 16000000 and NC _ <_ 115 and THENà NC7 _ <_ 115	22	IF: conf_per_100k_inh _> _ 16000000 and soil_wat_avail _> _ 60 THENà ND7 _ <_ 10	22
	IF: 928 _ <_ pres_atmMax__930 and soil_wat_avail__60 and NC__115 THENà NC7 _ <_ 115	16.5	IF: drought _ <_ 10 and 10__ND__45 and 115__NC__650 THENà 10 _ <_ ND7__45	16.5
	IF: 15 _ <_ tMin__17 and no2__8 and 115__NC__650 and 14__dew_poitMax__16 THENà 115 _ <_ NC7__650	8.8	IF: soil_wat_avail _ <_ 30 and ND _> _ 45 THENà ND7 _> _ 45	19.8

Fourteen decision trees were generated by J48, seven for new cases and seven for new deaths, within seven days ahead. Considering the accuracy, support and consistency of the classification rules, two decision trees were selected, one to predict new cases, NC1, and one to predict deaths, ND1 (Table 3). The support of each rule that makes up the decision tree can be seen in the parenthesis after the exit interval of each rule.

Table 3. Predictive models generated by the J48.

New confirmed (NC)	New deaths (ND)
(1)IF:total_conf_<_1044 (2)NC_<_115 (3)THENàNC1_<_115(29.90) (2)115_<_NC_<650 (3)THENà115_<_NC1_<_650(2.06) (2)NC_>_650 (3)THENàNC1_<_115(0) (1)IF:1044_<_total_conf_<_18000 (2)ND_<_10 (3)THENà115_<_NC1_<_650(1.03) (2)10_<_ND_<_45 (3)THENà115_<_NC1_<_650(17.53) (2)ND_>_45 (3)23_<_o3_<30 (4)THENàNC1_>_650(4.12) (3)o3_<_23 (4)THENàNC1_>_650(2.06) (3)o3_>_30 (4)THENà115_<_NC1_<_650(2.06) (1)IF:total_conf_>_18000 (2)isol_avg_index_<_45 (3)THENàNC1_>_650(0) (2)isol_avg_index_>_55 (3)THENà115_<_NC1_<_650(2.06) (2)45_<_isol_avg_index_<_55 (3)THENàNC1_>_650(23.71)	(1)IF:total_conf_<_1044 (2)THENàND1_<_10 (32.99) (1)IF:1044_<_total_conf_<_18000 (2)pm10_<_17 (3)THENà10_<_ND1_<_45 (9.28) (2)17_<_pm10_<_25 (3)tMax_>_27 (4)THENà10_<_ND1<_45(5.15) (3)tMax_<_25 (4)THENà10_<_ND1_<_45(4.12) (3)25_<_tMax_<_27 (4)THENàND1_>_45(5.15) (2)pm10_>_25 (3)THENàND1_>_45(6.19) (1)IF:total_conf_>_18000 (2)isol_avg_index_<_45 (3)THENàND1_>_45(0) (2)isol_avg_index_>_55 (3)THENà10_<_ND1_<_45(2.06) (2)45_<_isol_avg_index_<_55 (3)THENàND1_>_45(22.68)

Discussion

The proposed methodology (Figure 1) can be used to face the COVID-19 pandemic and other important diseases such as dengue, malaria and different generations of influenza. This approach is agile, efficient and can quickly provide subsidies for coping with these diseases, with an integrated perspective capable of covering the main issues that have been the subject of quantitative research on COVID-19, therefore, it can be an important support tool for public health authorities and governments. The approach has a reliability of support and forecasting proportional to the reliability of the data. Therefore, comprehensive tracking initiatives and assiduity and authenticity in notifications are essential.

The air temperature showed an inversely proportional correlation with the new cases and deaths by COVID-19, in line with the several studies that have been carried out trying to understand the correlation of these variables[23,24,25,26,27,28]. The minimum and average temperature showed linear correlation coefficients, greater than -/+ 0.60, with the numbers of deaths and new daily cases of coronavirus in São Paulo (Table 1). This indicates that there is a greater number of people affected by COVID-19 when temperatures are lower, the same pattern as other respiratory diseases. Agricultural drought (number of days without precipitation greater than 10 mm) and the percentage of water available in the soil also

showed significant linear correlation coefficients (greater than -/+ 0.60 with p <0.05) with the epidemiological variables of COVID-19 (Table 1). However, the correlation between water availability in the soil was inversely proportional to the epidemiological variables, while these variables had a directly proportional correlation with the agricultural drought. The relationship of these two meteorological variables with the epidemiological variables indicates that the dry climate is a factor that exacerbates the number of COVID-19 infections. This can be explained by the dryness of the airways, which compromises the nasal function of preventing the entry of viruses and bacteria in the human body. These variables need to be considered by managers when making decisions, since the total number of deaths and confirmed had the best correlations identified in this research with the agricultural drought (0.98, 0.97 respectively) and the availability of water in the soil (respectively -0.90, -0.89).

In this same direction, real evapotranspiration presented a high correlation coefficient, and inversely proportional, with the total number of deaths from the coronavirus. This fact indicates that the lower the actual evapotranspiration, loss of water by evaporation of the soil and transpiration of plants, the greater the total number of deaths from the virus. It is difficult to infer the reason for this correlation due to the complexity of factors involved in the evapotranspiration process (solar radiation, wind, temperature and humidity), but a viable possibility is the temperature, the greater the capacity of the air to contain water steam. Thus, the lower the temperature, the lower the evapotranspiration and the greater the number of deaths by COVID-19, considering that the total number of deaths was also inversely proportional to the average temperature (Table 1). The same meteorological variables, which presented significant linear correlation coefficients (greater than -/+ 0.60 with p <0.05), with the epidemiological variables of interest, were those which presented correlations greater than 0.60 with factor 1, in the factor analysis (Table 1). Therefore, it is these meteorological variables (average and minimum temperature, agricultural drought, maximum dew point, real evapotranspiration and availability of water in the soil) that need to be monitored and studied to understand the dynamics of COVID-19 under the bias of modeling in megacities.

The data cluster analyzes made it possible to identify other environmental variables that also need to be measured. The total number of deaths and atmospheric pressure are the best input variables for new confirmed forecasting models. While the isolation index, relative humidity and pm25 are the best input variables for predictive models of new deaths (figure 2). In tables 2 and 3, it can be seen that, in fact, these variables were recurrently selected by the algorithms as input variables of the models. This occurred mainly with the new confirmed cases, in which all the rules of the J48 used the variable the total of confirmed as input and among the twenty selected rules of the CBA eight used these variables as the total of confirmed or atmospheric pressure.

The predictive models generated by the two algorithms had a close performance in terms of accuracy, but the support of the J48 decision tree rules was lower than the statistically representative values (greater than 8%) to validate most of the rules. Therefore, it is recommended to use only the CBA models, for the case of São Paulo. However, for future studies it is recommended to reuse these two modeling tools in a

comparative way and to use other data mining algorithms in an exploratory way. Discretization using the quartile is another possibility that can be tested to verify the behavior of the results by other studies.

Uncertainty about future outbreaks is a problem in recurrent epidemics and pandemics, especially in developing countries with few resources and poor health systems. An accurate predictive tool, capable of anticipating the levels of hospitalizations and deaths, can be useful for health managers.

This work elucidates the temporal patterns of morbidity and mortality by COVID-19 in São Paulo, a Brazilian megalopolis. A history of confirmed cases and deaths was organized with meteorological and air quality variables. Multivariate analyzes were performed to understand the relationships between the variables involved. Predictive models with high and satisfactory precision were built to predict morbidity and mortality.

Forecasting public health conditions is useful for preparing health teams in advance for an outbreak and prevents the system from collapsing. In addition, prior information can optimize the resources invested in COVID-19 or other outbreaks of other urban diseases.

Declarations

Acknowledgment

Special thanks to the Araucária Foundation (Araucária Foundation - 001/2017 project) by the financial support of research.

Competing interests

The authors declare that there are no competing interests.

References

1. Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, Bhatia S, Boonyasiri A, Cucunubá Z, Cuomo-Dannenburg G, Dighe A. *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand* (Tech. Rep, Imperial College London, 2020).
2. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA.*;323(13):1239–1242. (2020). doi:10.1001/jama.2020.2648
3. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. *JAMA.*;323(14):1341–1342. (2020). doi:10.1001/jama.2020.3151
4. Park S, Choi GJ, Ko H. Information Technology–Based Tracing Strategy in Response to COVID-19 in South Korea—Privacy Controversies. *JAMA.*;323(21):2129–2130. (2020). doi:10.1001/jama.2020.6602

5. Wong JEL, Leo YS, Tan CC. COVID-19 in Singapore—Current Experience: Critical Global Issues That Require Attention and Action. *JAMA.*;323(13):1243–1244. (2020). doi:10.1001/jama.2020.2467.
6. Shaw R, Kim Y, Hua J. Governance, technology and citizen behavior in pandemic: Lessons from COVID-19 in East Asia. *Progress in Disaster Science*, v. 6, n. 100090, (2020). https://doi.org/10.1016/j.pdisas.2020.100090.
7. Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., & Agha, R. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery (London, England)*, *78*, 185–193. (2020). https://doi.org/10.1016/j.ijsu.2020.04.018
8. Hunter DJ. Covid-19 and the Stiff Upper Lip - The Pandemic Response in the United Kingdom. *N Engl J Med.*;382(16):e31. (2020). doi:10.1056/NEJMp2005755
9. Bedford, J., Enria, D., Giesecke, J., Heymann, D. L., Ihekweazu, C., Kobinger, G., Lane, H. C., Memish, Z., Oh, M. D., Sall, A. A., Schuchat, A., Ungchusak, K., Wieler, L. H., & WHO Strategic and Technical Advisory Group for Infectious Hazards. COVID-19: towards controlling of a pandemic. *Lancet* (London, England), 395(10229), 1015–1018. (2020). https://doi.org/10.1016/S0140-6736(20)30673-5
10. Rudan I. A cascade of causes that led to the COVID-19 tragedy in Italy and in other European Union countries. *Journal of global health*, *10*(1), 010335. (2020). https://doi.org/10.7189/jogh-10-010335
11. Al-qaness, M.A.A.; Ewees, A.A.; Fan, H.; Abd El Aziz, M. Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. *J. Clin. Med.9*, 674. (2020).
12. Perc M, Gorišek Miksić N, Slavinec M and Stožer A (2020) Forecasting COVID-19. *Front. Phys.*8:127. (2020). doi: 10.3389/fphy.2020.00127
13. Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., Wang, D., Chen, G., Zhang, J., Peng, H., & Shao, Y. Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling*, *5*, 282–292. (2020). https://doi.org/10.1016/j.idm.2020.03.002
14. Chakraborty, T., & Ghosh, I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, solitons, and fractals*, *135*, 109850. (2020). https://doi.org/10.1016/j.chaos.2020.109850
15. Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *Int J Infect Dis.*;93:201-204. (2020). doi:10.1016/j.ijid.2020.02.033
16. Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Sci Total Environ.*;728:138762. (2020). doi:10.1016/j.scitotenv.2020.138762
17. Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, solitons, and fractals*, *134*, 109761.(2020). https://doi.org/10.1016/j.chaos.2020.109761
18. Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M., Damen, J., Debray, T., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohman, A., Luijken, K., Ma, J., Andaur, C. L., Reitsma, J. B., … van Smeden, M. Prediction models for diagnosis

and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ (Clinical research ed.)*, *369*, m1328. (2020). https://doi.org/10.1136/bmj.m1328

19. Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS one*, *15*(3), e0230405. (2020). https://doi.org/10.1371/journal.pone.0230405
20. Ribeiro, M., da Silva, R. G., Mariani, V. C., & Coelho, L. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, solitons, and fractals*, *135*, 109853. (2020). https://doi.org/10.1016/j.chaos.2020.109853
21. Djilali, S., & Ghanbari, B. Coronavirus pandemic: A predictive analysis of the peak outbreak epidemic in South Africa, Turkey, and Brazil. *Chaos, solitons, and fractals*, *138*, 109971. (2020). https://doi.org/10.1016/j.chaos.2020.109971
22. Bouchnita, A., & Jebrane, A. A hybrid multi-scale model of COVID-19 transmission dynamics to assess the potential of non-pharmaceutical interventions. *Chaos, Solitons, and Fractals*, 109941. (2020). https://doi.org/10.1016/j.chaos.2020.109941
23. Şahin M. Impact of weather on COVID-19 pandemic in Turkey. *Sci Total Environ.*;728:138810. (2020). doi:10.1016/j.scitotenv.2020.138810
24. Auler AC, Cássaro FAM, da Silva VO, Pires LF.Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: A case study for the most affected Brazilian cities. *Sci Total Environ*, V.729, 139090, (2020). https://doi.org/10.1016/j.scitotenv.2020.139090.
25. Tosepu R, Gunawan J, Effendy DS, Ahmad OAI, Lestari H, Bahar H, et al. Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Sci Total Environ.*; 725:138436. (2020)
26. Bashir, M. F., Ma, B., Bilal, Komal, B., Bashir, M. A., Tan, D., & Bashir, M. Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci Total Environ.*, *728*, 138835. (2020) https://doi.org/10.1016/j.scitotenv.2020.138835
27. Sobral, M., Duarte, G. B., da Penha Sobral, A., Marinho, M., & de Souza Melo, A. Association between climate variables and global transmission oF SARS-CoV-2. *Sci Total Environ.*, *729*, 138997. (2020). https://doi.org/10.1016/j.scitotenv.2020.138997
28. Ma Y, Zhao Y, Liu J, He X, Wang B, Fu S, Yan J, Niu J, Zhou J, Luo B.Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. *Sci Total Environ.*, V.724,138226. (2020). https://doi.org/10.1016/j.scitotenv.2020.138226.
29. IBGE. *Panorama São Paulo* 2019 https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama
30. Souza, F. T.; Koerner, T. C. ; Chlad, R. A data-based model for predicting wildfires in Chapada das Mesas National Park in the State of Maranhão. *Environmental Earth Sciences* (Print), v. 74, p. 3603-3611, 2015.
31. Pinzón, Daniel Felipe Del Busto, & Fabio Teodoro de Souza. A data based model as a metropolitan management tool: The Bogotá-Sabana region case study in Colombia*. Land Use Policy* 54: 253-263. 2016.

32. Duarte, F., Gadda, T., Luna, C. A.M.,& Souza, F. T. (2016). What to expect from the future leaders of Bogotá and Curitiba in terms of public transport: opinions and practices among university students. *Transport Res F: Traffic Psychol Behav*, 38, 7–21.
33. Souza, F. T. Morbidity Forecast in Cities: A Study of Urban Air Pollution and Respiratory Diseases in the Metropolitan Region of Curitiba, Brazil. *Journal of Urban Health*, v. 1, p. 1, 2018.
34. Globo. *COVID-19 in the world* https://especiais.g1.globo.com/bemestar/coronavirus/mapa-coronavirus/#/mundo
35. Globo. *COVID-19 in the Brazil* https://especiais.g1.globo.com/bemestar/coronavirus/estados-brasil-mortes-casos-media-movel/#/
36. Brazil. *Especial COVID-19 - Data by Municipality* https://brasil.io/COVID-19/
37. São Paulo. *Health Surveillance in the city of São Paulo - Epidemiological data* https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/vigilancia_em_saude/index.php?p=245263
38. Center for Operations and Health Intelligence (NOIS). *Technical note: Analysis of underreporting of the number of confirmed cases of COVID-19 in Brazil* http://www.supersuporte.com/myRpubs/NT7_Subnotificacao_notaDia11-abr-2020.pdf
39. Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2010. *Multivariate data analysis*. 7 ed. New Jersey: Prentice-Hall.
40. Liu, B., Hsu, W., Chen, S., and Ma, Y. 1998. Integrating classification and association rule mining. *KDD-98*, New York. 27-31 August. AAAI. pp 80-86.
41. Quinlan, J. R. (1993). C4.5: Programing for machine learning, Morgan Kauffmann.
42. Wu, X., Kumar, V., Ross Quinlan, J. et al. Top 10 algorithms in data mining. Knowl Inf Syst 14, 1–37 (2008). https://doi.org/10.1007/s10115-007-0114-2
43. Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J.(2016). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.
44. Drakakis, G., Moledina, S., Chomenidis, C., Doganis, P. and Sarimveis, H. (2016). Decision trees for continuous data and conditional mutual information as a criterion for splitting instances. *Combinatorial chemistry & high throughput screening* 19(5): 423–428.

Figures

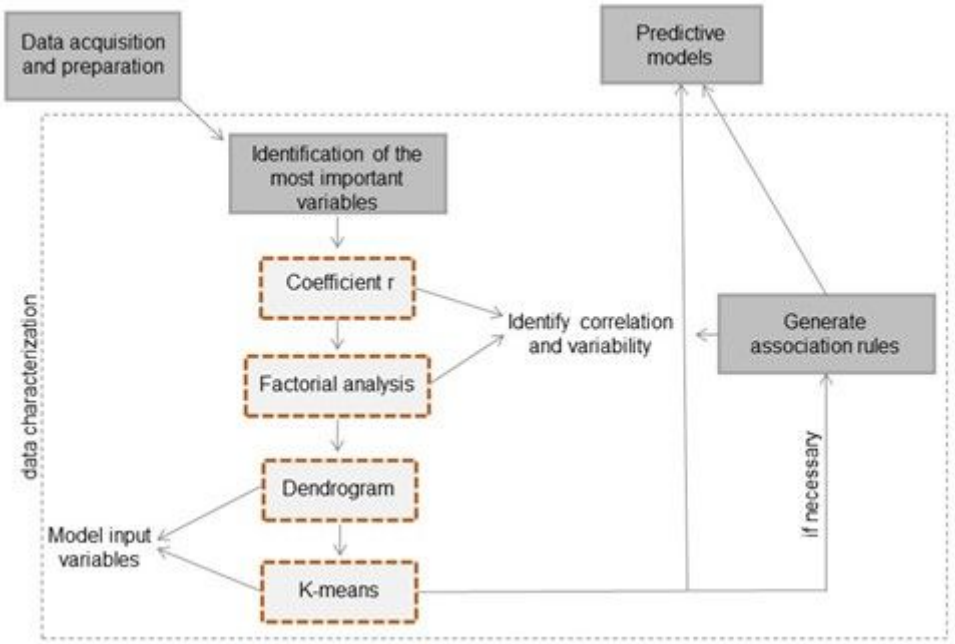


Figure 1

Sequential diagram of the proposed data approach for public health management.

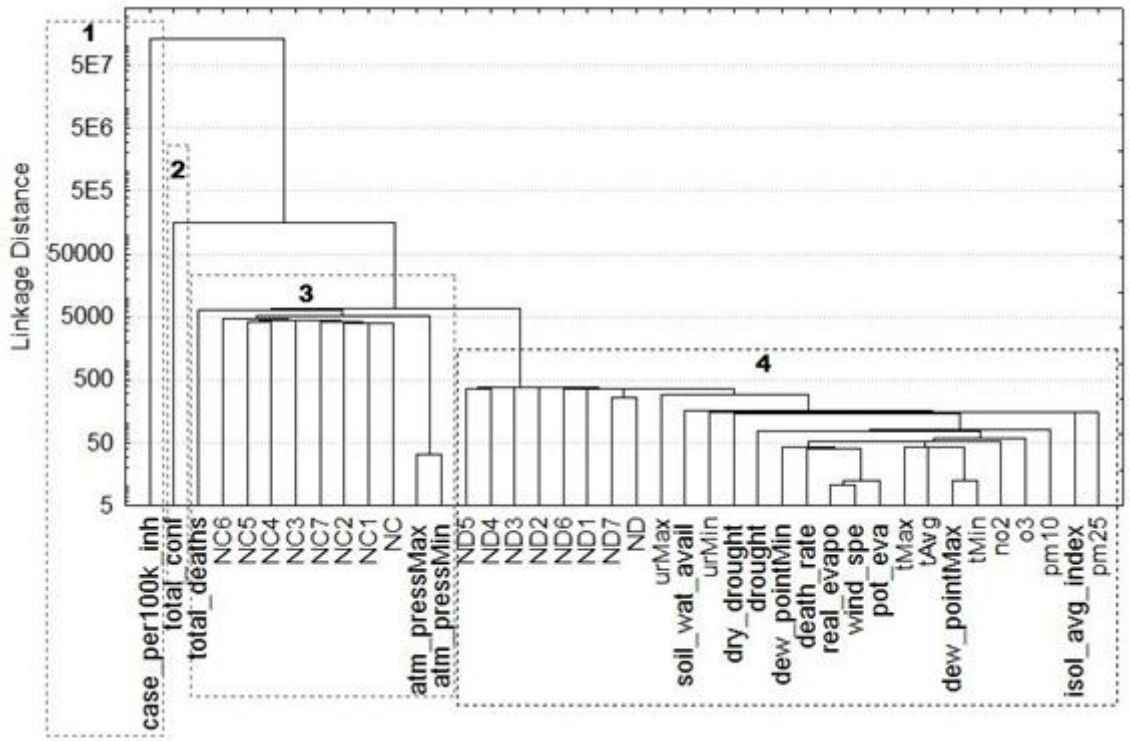
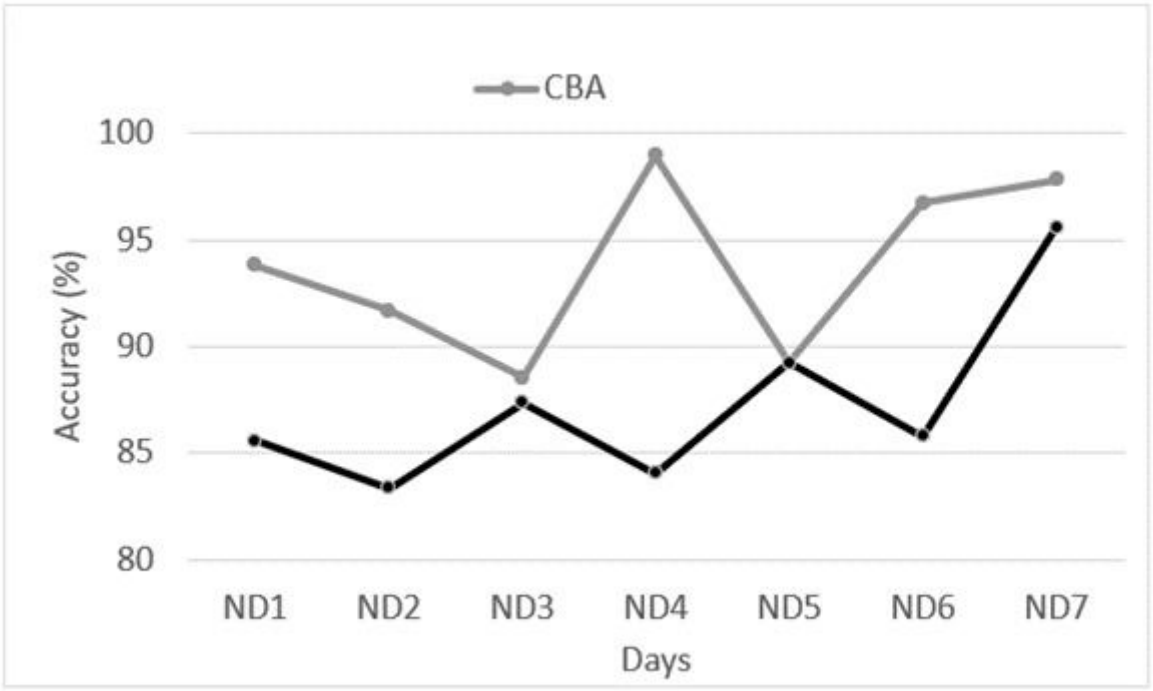


Figure 2

Similarity dendrogram.



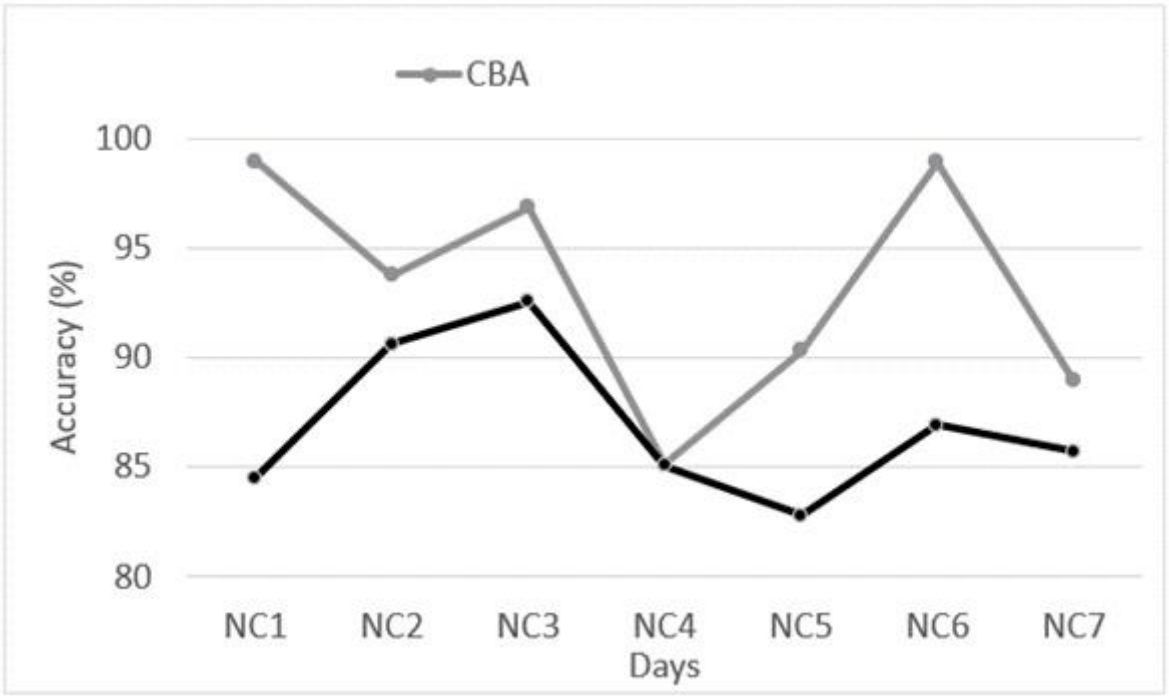


Figure 3

Comparison of the accuracy of the CBA and J48 models.