Comorbidities in patients with COVID-19,case study:Baja California, using ANN

Bogart Yail Márquez , ElizabethAguilar-Calderón, Arnulfo Alanís, Maribel Guerrero-Luis
Instituto Tecnológico de Tijuana
Tijuana, Baja California

bogart@tectijuana.edu.mx

***Abstract*—This project's main objective is to discover which are those comorbidities that could lead to a fatal outcome in a patient diagnosed with COVID-19 in the state of Baja California through a classification algorithm using neural networks. For this, a database obtained on the federal government portal by the General Directorate of Epidemiology with a cutoff date of June 8, 2020 was used. Only the records of the residents of Baja California were kept and only the following data: Sex, Municipality, Date of death, Age, all those variables referring to morbidities, Result (Confirmed cases of COVID-19), ICU (If they needed to enter the intensive care unit); also, from the variable of the date also, from the date variable of death, another variable called "Deceased" was generated to categorize whether the patient died or not. The resulting database was imported into the software where the model of the neural network, data preparation was performed and built the neural network model (multilayer perceptron). The dependent variable "Deceased" was selected, as variables the variables referring to the patient's comorbidities and as a covariate the variable of the scalar type Age. For this model, a random partition of the data was carried out, where 70% of the data was assigned for training and the remaining 30% for tests, obtaining a success rate of 82% and an 18 % error.**

Keywords—COVID-19; comorbidities; fatal outcome; neural networks; classification algorithms…

I. Introduction

According to the World Health Organization [13] the current pandemic due to the disease called COVID-19 has caused more than 450,000 deaths worldwide. According to various studies such as the one carried out in the city of Wuhan, China [3] the majority of patients who have died from COVID-19 have been found to have at least one other underlying condition.

In our country, for approximately 20 years there has been an increase in the number of deaths due to morbidity, chronic non-communicable diseases and externally caused diseases[5].

II. Theoretical Framework

A. COVID-19

Coronaviruses are an extensive family of viruses that can cause disease in both animals and humans. In humans, it has been known so far that they can cause respiratory infections ranging from the common cold to more serious illnesses such as Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS), both with a mortality rate of 37% and 10% respectively [3], [6].

According to Pérez, Manuel[6],"COVID-19 (coronavirus disease 2019) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), its shape is round or oval and often polymorphic, it has a diameter of 60 to 140 nm, the spike protein that is found on the virus surface and forms a rod-shaped structure, is the structure used for typing, the nucleocapsid protein encapsulates the viral genome and can be used as a diagnostic antigen".

B. COVID-19 clinical phenotype

The main initial symptoms of COVID-19 include fever, cough, fatigue, muscle pain, and dyspnea. Some patients showed atypical symptoms, such as diarrhea and vomiting. Sudden loss of smell and taste have also been observed (without mucus being the cause). In severe cases it is characterized by pneumonia, acute respiratory distress syndrome, sepsis, and septic shock. In the same way, there are people who can be non-symptomatic[6], [11].

C. Comorbidities

According to Rosas Oscar,*et al.*[8] "The term comorbidity was introduced to medicine by Alvan Feinstein (1970) when he observed that errors in classifying and analyzing comorbidity had led to many difficulties in the results in medical studies. Therefore, he defined comorbidity as the existence of a different additional clinical entity that occurs during the clinical course of a patient with an indexed disease under study".

According to studies conducted in Wuhan, China, the clinical phenotype was confused by the fact that 25.2% of patients had at least one other underlying medical condition.

The higher mortality rate in this region was due to more to people with morbid conditions[11].

D. Mortality rate in Mexico

According to the Pan American Health Organization [5] "Mexico for approximately 20 years has presented an epidemiological transition due to a decrease in communicable and parasitic diseases and an increase in the morbidity and mortality of chronic non-communicable diseases and diseases of external cause" as can be seen in the graph from Fig. 1 obtained from the information system platform of the Secretaría de Salud (Ministry of Health) [9].

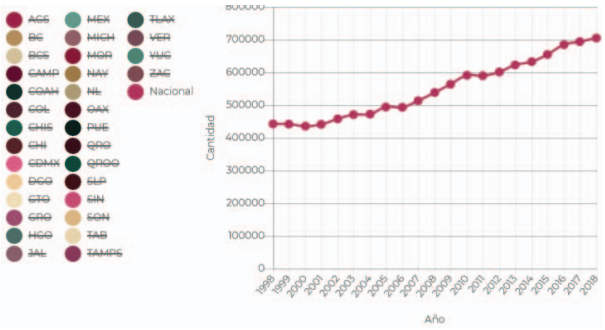


Fig. 1. Deaths per year nationwide [9].

According to information provided by the Information System of the Secretaría de Salud (Ministry of Health), in 2018 the main pathologies related to deaths in the country were heart disease, other unspecified causes, diabetes mellitus, malignant tumors and liver diseases.

In the Fig. 2 the complete list of pathologies can be observed[9].

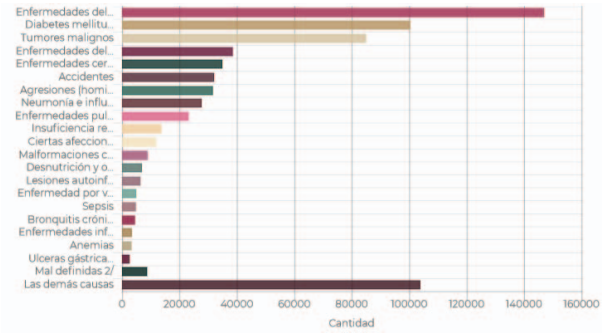


Fig. 2. Main causes of deaths in 2018 nationwide [9].

In the state of Baja California in 2018, the main pathologies related to deaths were heart disease (22.55%), malignant tumors (14.61%) and diabetes mellitus (13.43%), among others more mentioned in the Fig. 3 [9].

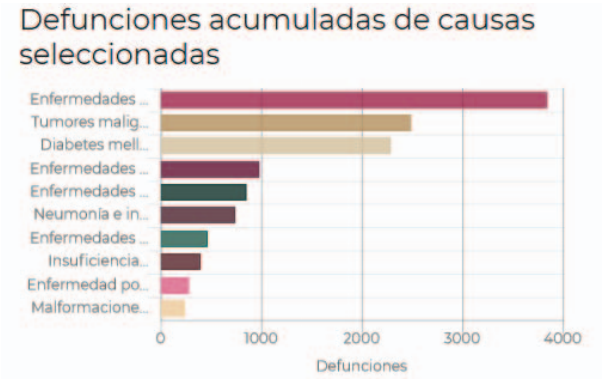


Fig. 3. Main causes of deaths in 2018 in Baja California [9].

51.22% of deaths occurred in people older than 65 years, 31.62% in people 45 to 64 years old, and 11.07% in people 25 to 44 years old.

E. *Biomedical computing*

Data mining aims to process data to find useful behavior the field of applied research, work is being done to discover patterns of data behavior in different areas of medicine, an activity known as biomedical informatics[4].

According to Capurro, Daniel,*et al.*[1]"Biomedical informatics is a discipline that ranges from the biological sciences to public health, and includes the following subdomains:

- a.Bioinformatics: studies biological information, especially cellular and molecular.
- b.Image processing: study the storage and processing of biomedical images.
- c.Clinical informatics: studies the information generated by direct patient care.
- d.Public health informatics: study the health information of the population.
- e.Patient computing: study the interaction between patients and information.
- f.Translational computing: seeks to integrate the areas already described through the integration of biomedical information".

III. Methodology

A. *CRISP-DM Methodology*

The CRISP-DM methodology from the acronym Cross-Industry Standard Process for Data Mining [2] is usually a very accepted guide in research lines that have data mining oriented works.

The model life cycle consists of 6 phases:

- Exploratory: This includes an understanding of the line of research on which the data analysis will be carried out.
- Descriptive: This phase covers the understanding of the data, the exploration between the relationship of the variables and the preparation of the information for the mining process.
- Relational: In this phase the transformation of the data from numeric to categorical, typing of numerical variables, determination of dependent and independent variables for the construction of the model is performed.
- Explanatory: The creation of the model is carried out with a predictive purpose.
- Predictive: Evaluation of the predictive capacity of the model and its efficiency when predicting.
- Application: Use of the model in daily and professional practice.

Below in Fig. 4the phases of the aforementioned CRISP-DM methodology model are shown in a more visual way[10].

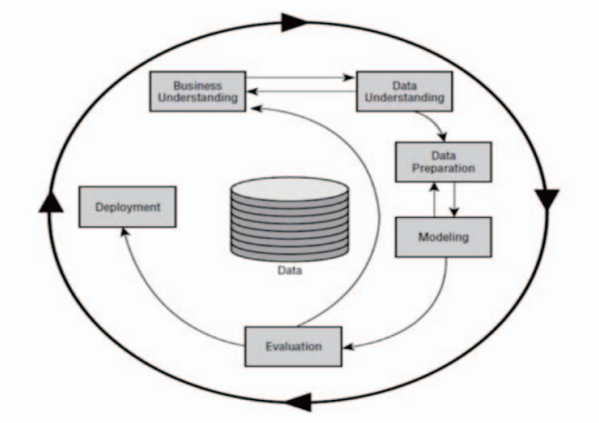


Fig. 4. Phases of the CRISP-DM methodology process model [10].

B. Neural Networks

Neural networks or also called artificial neural networks are processing algorithms that allow us to recognize patterns in data and are inspired by the functioning of biological neural networks.

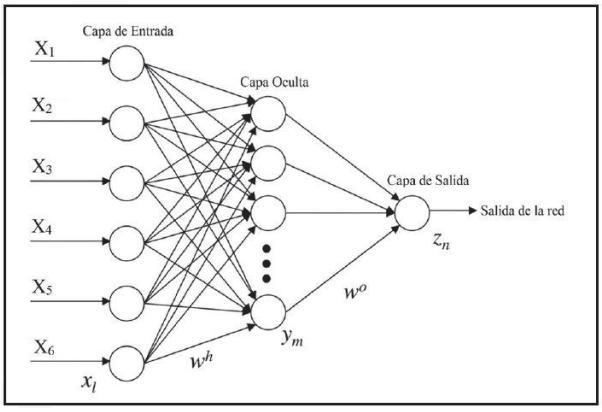
A neural network is made up of processing units called neurons, distributed in different layers, connected to each other by a series of weights that establish the relationships between them[12].

In general, neural networks have shown a classification capacity equal to or greater than statistical techniques, with the advantage that they can be used independently of compliance with the theoretical assumptions regarding these techniques.

C. Multilayer perceptron

The multilayer perceptron model consists of an input layer, one or more hidden layers, and an output layer as shown in Fig. 5[7].

The input layer is the one that receives the data provided by the independent variables of our model, "the hidden layers are responsible for representing the level of complexity that may exist between the input layer and the output layer" [12] and finally in the output layer it shows us the result of the classification made by the neural network, it is worth mentioning that the number of neurons that we have in this output layer corresponds to the number of classes that you want to identify.



Starting from the composition of a multilayer perceptron we can define the mathematical model of a hypothetical example like the one shown in Fig. 6 where 2 nodes are shown in the input layer, 2 in the only hidden layer that our model has and 2 nodes in the output layer.

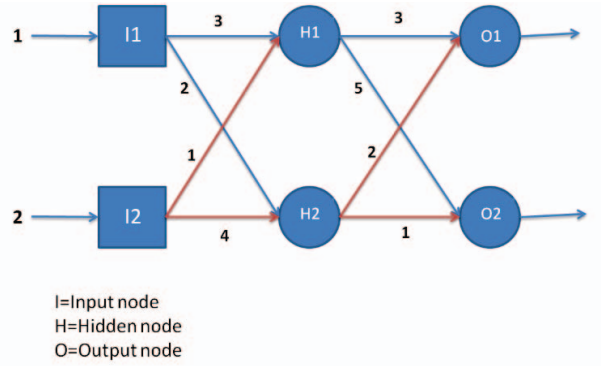


Fig. 6. Example of a multilayer perceptron model.

As you can see in this example, weights are assigned between the connections, although in practice they are usually assigned randomly.

1) Mathematical modeling of the multilayer perceptron using equations

- The values of the input layer are stored in a vector that we will call R^d, where "d" is the number of nodes the input layer has.
- In the output layer we will obtain a vector R^d belonging to the number of output nodes.
- Having this, then we can define the neural network in a function

$$f=||R^2 \rightarrow ||R^2$$

whose domain is the input vector R^d and the arrival set would be the vector R^d.

For this example we will assign the same activation function in both the hidden layer and the output layer(φ).

For this activation function you will have the identity activation function

$$\varphi(x)=x$$

Due to this, a weight will be produced in each node, but applying the identity activation function to said weight will give us the same output.

We proceed to calculate the output of the first hidden node, for this the following weighted sum is calculated using the equation (1)

$$V=(W_1)(E_1)+(W_2)(E_2)+b \quad (1)$$

Where W_n represents the incoming weight to the node, E_n corresponds to the value of the input from which said weight originates and b to the bias value of the node being calculated.

Then the activation function is applied as shown in the equation (2)

$$\varphi(V)=Y \quad (2)$$

Once having the output of the first hidden node the

respective input values to the node. Once having the outputs of the nodes of the hidden layer, we proceed to calculate the values of the nodes of the output layer, for this we will repeat the process of the nodes of the hidden layer, only this time, the resulting outputs of the nodes of the hidden layer will be taken as input values and the weights to consider will be those that connect the nodes of the hidden layer with the nodes of the output layer.

2) Mathematical modeling of the matrix multilayer perceptron: Another way to model the multilayer perceptron is through matrix calculations, for this we have the equation (3)

$$V=WX+b \tag{3}$$

Where W is the vector of weights, X the input vector and \$b\$ the bias vector associated with the layer.

For this example, our equation would be as follows (see equation (4))

$$V=\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}+\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}+\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{4}$$

Once the weight vector has been calculated, the activation function shown in the equation is applied (5) to get the resulting output

$$y=\varphi(V) \tag{5}$$

Remembering that the activation function in our example is the identity function, the result would be as follows (see equation (6))

$$y=\varphi\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}=\begin{bmatrix} \varphi V_1 \\ \varphi V_2 \end{bmatrix}=\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \tag{6}$$

This multilayer perceptron model can be simplified as shown in Fig. 7.

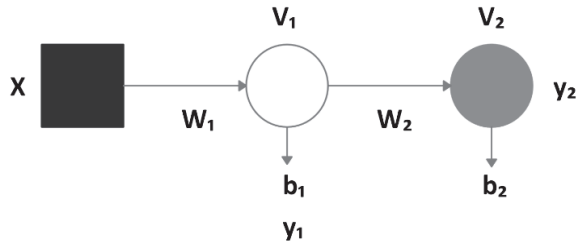


Fig. 7. Simplified multilayer perceptron model.

Where X represents the input vector, V_1 y V_2 are the vectors of the weighted sums of the hidden layer and output layer respectively, W_1 y W_2 corresponds to the vectors of the weights, b_1 y b_2 to the bias value vectors and finally Y_1 and Y_2 are the resulting output vectors.

Being this way, then we can calculate this perceptron with the equations (7) to calculate the hidden layer and (8) to calculate the output layer.

$$Y_1=\varphi(V_1)=\varphi(W_1X+b_1)=W_1X+b_1 \tag{7}$$

$$Y_2=\varphi(V_2)=\varphi(W_2Y_1+b_2)=W_2Y_1+b_2 \tag{8}$$

IV. Results Obtained

A. Collection of data

The database that has been used for this case study belongs to the open data published on the federal government portal by the General Directorate of Epidemiology. In this portal you can download the database with .csv extension as well as a spreadsheet called "descriptores" which is an explanatory manual of the variables shown in the database and also another called "Catalogos" in which show the rankings of each variable in the database.

As clarification in the portal of the daily federal government, the updated database of COVID-19 records is uploaded, which we will use for this study has the cutoff date corresponding to June 8 of this year.

B. Descriptive analysis and exploration between the relationship of the variables

In this step, all the records of other entities in the country were discarded and only those that belonged to residents of Baja California and that were also treated at medical centers in the region were preserved.

Only the following data was kept: Sex, Municipality, Date of death, Age, all those variables referring to morbidities, Result (Confirmed cases of COVID-19), ICU (If they needed to enter the intensive care unit); In addition, from the variable of the date of death, another variable was generated called "Deceased" where it has the condition of showing (1/Yes) in case the patient's record has a death date and shows (2/No) in case the death date is not registered for the patient.

C. Data preparation

We proceeded to import the resulting database into the software where the neural network model will be built, the corresponding typing of each variable was performed (all variables are numeric except DATE-DEAD which is string),all variables will be input variables and their respective categorical values were assigned as shown in the Table. I.

TABLE I. DATA PREPARATION.

Name	Label	Values	Measure
SEX	Sex	{1, Woman...	Nominal
MUNICIPALITY_RES	Municipality	{1, Ense...	Nominal
DATE_DEA	Date ofdea...	None	Nominal
DECEASED	Deceased	{1, Yes}..	Nominal
PNEUMONIA	Pneumonia	{1, Yes}..	Nominal
AGE	Age	None	Scale
PREGNANCY	Pregnancy	{1, Yes}..	Nominal
DIABETES	Diabetes	{1, Yes}..	Nominal
COPD	COPD	{1, Yes}..	Nominal
ASTHMA	Asthma	{1, Yes}..	Nominal
IMMUSUPP	Immunosupp..	{1, Yes}..	Nominal
HYPERTENSION	Hypertension	{1, Yes}..	Nominal
OTHER_COM	Othercom...	{1, Yes}..	Nominal
CARDIOVASCULAR	Cardiovas ...	{1, Yes}..	Nominal
OBESITY	Obesity	{1, Yes}..	Nominal
CHRONIC_RENAL	Chronic renal...	{1, Yes}..	Nominal
SMOKING	Smoking	{1, Yes}..	Nominal
RESULT	Result	{1, Posit...	Nominal
ICU	Intensivecare...	{1, Yes}..	Nominal

Once the data was prepared, the construction of the multilayer perceptron model began. As can be seen in the

the variables referring to the patient's comorbidities as factors, and the variable of the scalar type Age as covariate.

TABLE II. SELECTION OF DEPENDENT AND INDEPENDENT VARIABLES.

Dependent Variables	Covariates
Deceased [DECEASED]	Age [AGE]
Factors	
Pneumonia [PNEUMONIA]	Pregnancy [PREGNANCY]
Diabetes [DIABETES]	COPD [COPD]
Asthma [ASTHMA]	Hypertension [HYPERTENSION]
Immunosuppression [IMMUNOSUPP]	Obesity [OBESITY]
Othercomorbidity [OTHER_COM]	Smoking [SMOKING]
Chronic renal failure [CHRONIC_RENAL]	
Cardiovascular disease [CARDIOVASCULAR]	

D. Model Creation

For this model, a random partition of the data was carried out, where 70% of the data was assigned for neural network training and the remaining 30% for tests.

Regarding the selection of the multilayer perceptron architecture, the automatic selection option was left active so that the system assigns the architecture that can give us a better prediction.

E. Model Evaluation

As can be seen in the Table. IV of the Appendix, the system used the hyperbolic tangent activation function for the hidden layers and the Softmax activation function for the output layers.For this calculation the system excludes the bias unit.The resulting neural network had 37 inputs, a hidden layer and 2 outputs that correspond to the Yes and No values of the variable to be predicted (Deceased), (see Appendix, Fig. 8).According to the results obtained, a predictive accuracy percentage of 82% was obtained (see Appendix, Table. V) and an 18% error (see Appendix, Table. VI).

V. Discussion

According to the data obtained in the table of importance of the independent variables shown in the Table. III, we noticed that the variables or factors with the greatest weight to define whether a person diagnosed with COVID-19 may die or not are mainly age with an importance of 22.7%, pneumonia follows with 16.2% and chronic renal failure with 11.2%.

TABLE III. IMPORTANCE OF INDEPENDENT VARIABLES.

	Importance	Normalizedimportance
Pneumonia	.162	58.4%
Pregnancy	.054	19.5%
Diabetes	.063	22.6%
COPD	.053	19.3%
Asthma	.056	20.4%
Immunosuppression	.045	16.3%
Hypertension	.021	7.4%
Othercomorbidity	.037	13.5%
Cardiovascular disease	.040	14.5%
Obesity	.041	14.7%
Chronic renal failure	.112	40.4%
Smoking	.039	14.0%
Age	.277	100.0%

VI. Conclusions

In conclusion, this report shows the main characteristics and pathologies that can lead a COVID-19 patient to have greater complications or even have a fatal outcome.

The mortality rate in Baja California and the rest of the country is increasing every year and has morbidities and chronic non-communicable diseases as the main causes of death.

Furthermore, according to our predictive analysis, the most important variable is the age of the patient;remembering the deaths registered during the year 2018, they tell us that 51.22% of these occurred in people over 65.

According to these results, it can be said that there is a "dark panorama" for the population and it is very likely that the health services may be saturated when attending to so many patients infected with complications and requiring even intensive care.

It is for this reason that it is hoped that this study may be useful for the health sector to see areas of opportunity in strengthening prevention and health care programs or to disseminate existing campaigns in order to reduce rates of mortality from morbidities or chronic non-communicable diseases.

VII. Limitations and future directions of research

As this is a recently discovered disease, there is still a lot of unknown information and uncertainty in the scientific field,the study related to COVID-19 seems to be very broad and different research directions can be taken according to new discoveries that are obtained over time.

References

[1] Daniel Capurro, Mauricio Soto, Macarena Vivent, Marcelo Lopetegui, and Jorge R. Herskovic. Informática biomédica. Revista Medica de Chile, pages 1611–1616, 2011.

[2] Javier Jesús Espinoza. Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ingeniería Investigación y tecnología, 11(1):1–17, 2020.

[3] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, GuangfaWang,RongmengJiang,ZhanchengGao,QiJin,JianweiWang, and Bin Cao. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, 395(10223):497–506, 2020.

[4] Sergio Monserrat and Omar Chiotti. Minería de Datos en Base de Datos de Servicios de Salud Monserrat,. Technical report, Universidad Tecnológica Nacional-Facultad Regional Santa Fe, 2014.

[5] Organización Panamericana de la Salud. Estrategia de Cooperación de la Organización Panamericana de la Salud/Organización Mundial de la Salud con México. 2015-2018.Technical report, Organización Panamericana de la Salud, México, D.F., 2015.

[6] Manuel Ramón Pérez, Jairo Jesús Gómez, and Ronny Alejandro Dieguez. Características clínico-epidemiológicas de la COVID-19. Revista Habanera De Ciencias Medicas, 19(2):1–15, 2020.

[7] Nelson Marcelo Romero and Eustaquio Alcides Martínez. Aplicación de Redes Neuronales Artificiales en la Orientación Vocacional.

[8] Oscar Rosas-Carrasco, Eduardo González-Flores, Ana M. BritoCarrera, Odín E. Vázquez-Valdez, Emma Peschard-Sáenz, Luis Miguel Gutiérrez-Robledo, and Emilio José García-Mayo. Evaluación de la comorbilidad en el adulto mayor., 2011.

[9] Secretaría de Salud, Subsecretaría de Integración y Desarrollo del

información oficial de defunciones INEGI, and SS 1979-2017. Sistema de Información de la Secretaría de Salud.

[10] Umair Shafique and Haseeb Qaiser. A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1):217–222, 2014.

[11] Jiumeng Sun, Wan Ting He, Lifang Wang, Alexander Lai, Xiang Ji, Xiaofeng Zhai, Gairu Li, Marc A. Suchard, Jin Tian, Jiyong Zhou, Michael Veit, and Shuo Su. COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives, 2020.

[12] Guillermo Antonio Toro and Iván Alberto Lizarazo. Evaluación de las Redes Neuronales Artificiales Perceptron Multicapa y Fuzzy-Artmap en la Clasificación de Imágenes Satelitales. Ingeniería, 17(1):61–72, 2012.

[13] World Health Organization. Coronavirus disease (COVID-19) Situation Report-151. Technical report, World Health Organization, Slovenia, 2020.

APPENDIX

TABLE IV. INFORMATION FROM THE RESULTING NEURAL NETWORK.

Network Information			
Input Layer	Factors	1	Pneumonia
		2	Pregnancy
		3	Diabetes
		4	COPD
		5	Asthma
		6	Immunosuppression
		7	Hypertension
		8	Other comorbidity
		9	Cardiovascular disease
		10	Obesity
		11	Chronic renal failure
		12	Smoking
	Covariates	1	Age
	NumberofUnits		37
	RescalingMethod forCovariates		Standardized
Hidden Layer(s)	NumberofHidden Layers		1
	Number of Units in Hidden Layers		8
	ActivationFunction		Hyperbolic tangent
Output Layer	Dependent Variables		Deceased
	NumberofUnits		2
	ActivationFunction		Softmax
	Error Function		Cross-entropy

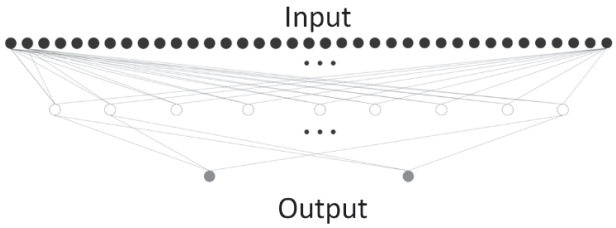


Fig. 8. Neural network architecture of 37 inputs and 2 outputs.For thedeceased state (yes and no).

TABLE V. TABLE OFCLASSIFICATION.

Classification				
	Predicted			
Sample	Observed	Yes	No	Percent Correct
Training	Yes	370	498	42.6%
	No	240	3165	93.0%
	OverallPercent	14.3%	85.7%	82.7%
Testing	Yes	152	214	41.5%
	No	113	1342	92.2%
	OverallPercent	14.6%	85.4%	82.0%

TABLE VI. MODEL SUMMARY

ModelSummary		
Training	Cross Entropy Error	1502.661
	PercentIncorrect Predictions	17.3%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error
	Training Time	0:00:00.95
Testing	Cross Entropy Error	644.395
	PercentIncorrect Predictions	18.0%