Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Estimation of the main conditions in (SARS-CoV-2) Covid-19 patients that cause their death using Machine learning, the case of Mexico

José A. Guzmán-Torres, Elia M. Alonso-Guzmán, Francisco J. Domínguez-Mota, Gerardo Tinoco-Guerrero

PII: S2211-3797(21)00596-9
DOI: https://doi.org/10.1016/j.rinp.2021.104483
Reference: RINP 104483

To appear in: Results in Physics

Received Date: 6 November 2020
Accepted Date: 18 June 2021

Please cite this article as: Guzmán-Torres, J.A., Alonso-Guzmán, E.M., Domínguez-Mota, F.J., Tinoco-Guerrero, G., Estimation of the main conditions in (SARS-CoV-2) Covid-19 patients that cause their death using Machine learning, the case of Mexico, Results in Physics (2021), doi: https://doi.org/10.1016/j.rinp.2021.104483

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 The Author(s). Published by Elsevier B.V.
Estimation of the main conditions in (SARS-CoV-2) Covid-19 patients that cause their death using Machine learning, the case of Mexico

José A. Guzmán-Torres\textsuperscript{a*}, Elia M. Alonso-Guzmán\textsuperscript{a}, Francisco J. Domínguez-Mota\textsuperscript{a,b}, Gerardo Tinoco-Guerrero\textsuperscript{b}.

\textsuperscript{a}Faculty of Civil Engineering, UMSNH.
\textsuperscript{b}Faculty of Mathematical and Physics Science, UMSNH.
*Correspondence author: jaguzman@umich.mx

June 17, 2021

Keywords: Covid-19; Data Science; Diseases; Machine learning.

Abstract

Nowadays, society faces a catastrophic problem related to respiratory syndrome due to the coronavirus SARS-CoV-2: the Covid-19 disease. This virus has changed our coexistence rules and, in consequence, has reshaped the daily activities in modern societies. Thus, there are many efforts to understand the virus behaviour in order to reduce its negative impact, and these efforts produce an incredible amount of information and data sources every week. Data scientists, which use techniques such as Machine learning, are focusing their abilities to develop mathematical models for analysing this critical situation. This paper uses Machine Learning techniques as tools to help understand some specific new patterns in Covid patients that arise from unknown complex interactions in the transmission-dynamic models of the SARS-CoV-2 virus, and their relation with the corresponding social contact patterns which are often known or can be inferred from populations variables. One of the main motivations of this research is to find the diseases that cause an increase in the risk of death in infected people with the Covid-19 virus. Mexico is the case of study in this research. The general conditions of health that cause death are well known generally in the world. However, these conditions in each country can differ depending on different factors such as the general health status of people. The results show that the principal causes of death in Mexico are related to age, bad eating habits, chronic diseases, and contact with infected people having not proper care. Results from the analysis show a remarkable accuracy of 87\%, which is considered satisfactory.

1 Introduction

Coronaviruses (CoVs) belong to a large family of viruses that cause severe illness to human beings. Crucial epidemic have occurred recently, like the Severe Acute Respiratory Syndrome (SARS)
occurred in 2003 [1], and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) [2]. Due to the quick spread of such virus, all of these events were responsible for many deaths, and the biggest problem with the new coronavirus, the severe acute respiratory syndrome coronavirus 2 SARS-CoV-2 [3], is the fact that it seems to be even more contagious than other coronavirus and therefore has quickly spread globally in few weeks. Currently, the newspapers give account of the quick spread of the Covid-19 around the globe and of its brutal impact on the economy.

Covid-19 arrived to Mexico and affected the performance of the health institutions and required a huge amount of their available material and human resources. A total of 623,000 cases, 66,851 deaths and 434,667 cases of recovered people are reported up to August 2020, according to data provided by the Ministry of Health. This study presents the exploration of the dataset up to 2020 September, from the portal www.gob.mx/salud, where an analysis of the information was carried out to determine which are the first ten conditions in patients infected with Covid-19 that cause their death at the present time in Mexico, as well as predict the percentage of mortality based on these conditions.

The analysis made in this research was done by applying Supervised Machine learning algorithms, such as the wrapping technique to indicate the variables with greater weight that cause this problem and using logistic regression analysis to predict the patient’s mortality based on their conditions. Machine learning is an evolving branch of computational algorithms that have been designed to emulate human intelligence by learning from the surrounding circumstances, as is widely detailed in [4]. There are several studies involving Machine learning techniques to predict and evaluate the behaviour of Covid-19. There is a known approach for the prediction of the epidemic spread of the coronavirus, promoted by the transport of the spring festival in China which, was proposed by Fan et al. [5]. A bioinformatic approach that can predict candidate targets for immune responses to SARS-CoV-2 was presented by Grifoni et al. [6]. Today it is still analyzed what could be done to control the occurrences of Covid-19 according to [7]. Other research, such as Huang’s, described a spatial-temporal distribution of Covid-19 in China and its prediction [8]. Li and et al. Described the propagation analysis and the prediction of the actual time series of Covid-19 [9]. Artificial Neural Networks (ANNs) have also been involved in these studies as described in [10], where an ANN has been used to predict infections over time.

Supervised machine learning is akin to the type of model-fitting that is standard in epidemiologic practice: The value of the outcome (the dependent variable), often called its “label” in machine learning, and is known for each observation according to [11]. Supervised learning methods include standard epidemiologic approaches such as linear and logistic regression. The generalized logistic model or logistic regression has been successfully applied in other studies to describe previous epidemics [12]. It should be noted that GPU hardware was used in the data exploration and modelling process due to the amount of data. The data set was segmented into two groups, training and test.

The limitations of this study are the data collection bias, the features amount obtained in the hospitals, the number of available tests, the cases that are not reported and the information that is not published.

The most important motivation for this analysis is to contribute to a better understanding of the crucial diseases that produce a high mortality risk in infected patients with Covid-19. Since there are studies that provide information on the preexisting conditions and characterize its progress in Mexico, the intention is to provide information to refine the prognosis and encourage more efficient actions by policymakers to optimize resources in response to Covid-19 [13].
2 Research significance

The present study aims to develop a predictor model for estimating the top factors (predictive variables) that cause mortality in Covid-19 patients in Mexico using Machine learning. Several factors that contribute to the patients mortality are well known around the globe, but depend strongly on complex unknown patterns which are different from one place to another; for this analysis, the aim is to prove that the determinant factors can be different depending on the geographical area and on intrinsic social behaviours linked to idiosyncrasy. Codes, dataset, and model corresponding to this research are available at github.com/JaGuzmanT/Logistic-Regression-to-predict-the-risk-of-death-in-Covid-19-Patients.

3 Methodology

3.1 Dataset

This study analyses a large open access dataset. The dataset is available at the world web site www.gob.mx/salud and was published by Hector Paredes. It contains general information of about 1,048,575 patients such as nationality, sex, age, origin, entity, the health sector in which the patient was attended, knowledge of the patient’s current health condition and diseases, such as pneumonia, COPD (Chronic obstructive pulmonary disease), asthma, and their respective Covid-19 test results. Each feature involved in the dataset was considered as a predictive variable or regressor. Information such as demographic, laboratory results, and diseases were collected to create the dataset at the time of patient’s admission. The dataset did not show missing data, and any observation was removed. The information was presented in numerical and categorical variables. The categorical variables were addressed with one-hot encoding.

In the step of data preprocessing in Machine learning, the data needs to be prepared in specific ways before feeding it into a predictive model. One-hot encoding is a processing the data processing that is applied to categorical data, in order to convert it into a binary vector representation for use in Machine learning algorithms. The one-hot encoding process is a necessary step for establishing the features or attributes in numerical variables. This process transforms from categorical to numerical variables, this is a common practice in Data science analysis [14, 15].

The present study has included 35 regressors and after to discretize the information, only 18 regressors were used. The other regressors were dropped due to provide not crucial information such as nationality, ID, place of residence and others. The 18 regressors were encoding trough one hot a pre-processing, and 32 regressors were obtained. The new 32 regressors were analysed and discretized in order to find the main conditions that increase the probability of mortality in Covid-19 patients. The target known as 'MORTALITY', considers 2 clasess, DEAD and LIVE. Prior to the modelling process, the dataset was standardized. This process is a type of normalization. Normalization is a technique often applied as part of data preparation for Machine learning. Normalizations aims to change the values of numeric columns in the dataset to a common scale without distorting differences in the ranges of values. The data analysis and the logistic regression model was performed in Python’s environment, using a personal computer and supported on different libraries such as Scikit-Learn, Pandas and Numpy.

For this study, the 'MORTALITY' feature was considered as the target variable. This attribute was used to develop the model that predicts the risk of death in Covid-19 patients. In the beginning, the dependent variable was considered qualitative, dead (nonsurvivor), and live (survivor) and later
was recognised quantitative, DEAD=0 and LIVE=1. The success in the logistic regression model presented in this paper was achieved when the dependent variable took the numerical values.

The dataset contains a difference between the established classes since there are 78,536 deaths registered and 970,039 living cases. For this reason, a balancing process of the information was required. In the Data Science field, this process is also known as the resampling process. This process consists of segmenting the information in the same amount of samples for the established classes to avoid problems in the algorithm process. In this case, an undersampling process was used, i.e., the data were resampled according to the number of deaths.

Prior to the modelling, the dataset was segmented into two groups, training and test, distributed in 70% and 30%, respectively. Likewise, 78,536 cases of deaths and 78,536 cases of living people were considered, having a total of 157,072 samples to feed the model.

### 3.2 Wrapper method and Logistic Regression

The wrapper method evaluates a specific Machine learning algorithm in order to find the optimal regressors or attributes in a given dataset. It follows an ambitious search approach by estimating all possible combinations of features with respect to the assessment criterion. The assessment criterion is simply the performance measure that depends on the type of problem. For instance, for regression, the assessment criterion may be the p-values, Adjusted R-Squared. Similarly, for classification, the assessment criterion might be accuracy, precision, recall, f1-score and others. Eventually, it selects the combination of attributes that provides the optimal results for the specified machine learning algorithm. The most commonly used techniques under wrapper methods are listed below:

- Forward selection
- Backward elimination

For this study, Backward elimination was selected. It works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains. It is possible to achieve this by fitting the given machine learning algorithm used in the model core, ranking features by importance, discarding the least important features, and re-fitting the model. This process is replicated until a specified number of features remains [16]. Finally, the least important features are displaced from the current set of data. Figure (1) shows the Backward diagram.

![Figure 1: Flow chart - Backward diagram process.](image-url)
The Recursive Feature Selection (RFE) contained in the Scikit-learn library is the algorithm that works for employing the Backward process. One of the principal advantages of this process is the dimensionality reduction in large datasets. The dataset employed in the present research is considered a high dimensional. The RFE method performs well in the cycle of cross-validation to find the optimal number of attributes [17]. The RFE method requires an estimator, which is given by logistic regression, and at the same time, it implements an lbfgs regularization for multi-class problems and here, the regularization is applied by default, where the regularization is $L2$ with a primary formulation [18]. The lbfgs is an optimization algorithm that approximates the Broyden-Fletcher-Goldfarb-Shanno algorithm, which belongs to the quasi-Newton methods. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt), or the log-linear classifier [19]. In this model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. The regularization is applied by default, which is common in Machine learning but not in statistics. Another advantage of regularization is that it improves numerical stability. The absence of regularization is equivalent to setting $C$ at a quite large value [20].

As an optimization problem, the binary class $L2$ penalized by logistic regression minimizes the cost function given by Eq. (1)

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Similarly, $L1$ regularized logistic regression solves the Eq. (2) for optimization problems

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Once the Backward elimination method has been applied is possible to identify the top ten principal diseases that cause death in Covid-19 patients. After defining the top diseases, a predictive model was designed based on the Logistic regression to evaluate the nonsurvivor probability in infected patients.

### 3.3 Evaluation metrics

In this research, the model was evaluated by comparing the actual value ($y_{real}$) against the predicted value ($y_{predicted}$). For predicting the outcome, the test values of the regressor were put in the prediction model to obtain a probability. The accuracy of the model was computed with classification metrics. It is a common practice when the target is considered categorical. In this case, the target presents two outcome possibilities DEAD=0 and LIVE=1. ROC (Receiver Operating Characteristic) plots are two-dimensional plots in which the $TP$ (true positive) rate is represented on the $y$ - axis, and the $FP$ (false positive) rate is plotted on the $x$ - axis.

A ROC graph represents the relative compensations between the ratio (true positives) and costs (false positives). A discrete classifier produces only one class label. Each discrete classifier provides a pair ($FPR$ (rate), $TPR$ (rate)) that corresponds to a single point in the ROC space [21]. The construction of the ROC curves for this study was represented to evaluate the accuracy of the model. The classifiers are given by the next Eq., and all of them are discrete.

$$FPR \text{ (rate)} = \frac{FP}{N}$$

(3)
\[ TPR (rate) = \frac{TP}{P} \] (4)

\[ \text{precision} = \frac{TP}{TP + FP} \] (5)

\[ \text{recall} = \frac{TP}{P} \] (6)

\[ \text{accuracy} = \frac{TP + TN}{P + N} \] (7)

\[ F - \text{measure} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} \] (8)

where if the instance is positive and at the same time is classified as positive, it is classified as a true positive (\( TP \)); if the instance is negative, it is identified as a false negative (\( FN \)). Otherwise, whether the case is negative and is classified as negative, it is counted as a true negative (\( TN \)); if it is marked as positive, it is counted as a false positive (\( FP \)). All of these parameters can be correctly represented in a confusion matrix, which shows the assessment of the predicted labels.

4 Results and discussions

In this section, the results of the dataset analyzed are presented. As was described in section 3.1, 18 attributes or regressors were considered to be analyzed to find the top ten critical diseases that increase mortality in Covid-19 patients. The 18 regressors were transformed into 32 using the Backward propagation encoding, and these regressors are listed in Table (1).

| Disease number | Disease                         | Disease number | Disease                        |
|----------------|---------------------------------|----------------|--------------------------------|
| 1              | AGE                             | 17             | INMUSUPR-IGNORED               |
| 2              | RESULT-POSITIVE Covid-19        | 18             | INMUSUPR-YES                   |
| 3              | RESULT-PENDING RESULT           | 19             | HYPERTENSION-IGNORED           |
| 4              | SEX-WOMAN                       | 20             | HYPERTENSION-YES               |
| 5              | KIND OF PATIENT-HOSPITALIZED    | 21             | ANOTHER COM-IGNORED            |
| 6              | PNEUMONIA-NOT SPECIFIED         | 22             | ANOTHER COM-YES                |
| 7              | PNEUMONIA-YES                   | 23             | CARDIOVASCULAR-IGNORED         |
| 8              | PREGNANCY-NOT APPLIED           | 24             | CARDIOVASCULAR-YES             |
| 9              | PREGNANCY-IGNORED               | 25             | OBESITY-IGNORED                |
| 10             | PREGNANCY-YES                   | 26             | OBESITY-YES                    |
| 11             | DIABETES-IGNORED                | 27             | RENAL CHRONIC-IGNORED          |
| 12             | DIABETES-YES                    | 28             | RENAL CHRONIC-YES              |
| 13             | COPD-IGNORED                    | 29             | TABACISM-IGNORED               |
| 14             | COPD-YES                        | 30             | TABACISM-YES                   |
| 15             | ASTHMA-IGNORED                  | 31             | ANOTHER CASE-NOT SPECIFIED     |
| 16             | ASTHMA-YES                      | 32             | ANOTHER CASE-CASE-YES          |

From Table (1), it can be seen that all variables are categorical except the first one, which belongs to the patient’s age. It is possible to appreciate that, in each category, one variable was
dropped per group to avoid multicollinearity issues. This problem occurs when one attribute is linearly dependent on the other. For instance, feature number 4, SEX-WOMAN, appears individually without the attribute SEX-MAN since one variable depends on the other, and indeed they are linearly dependent, causing multicollinearity issues and producing critical errors in the accuracy estimation. Table (2) shows the top ten diseases obtained from the Backward propagation, and these are considered the main conditions that increase the risk of mortality in Covid-19 patients in Mexico. The conditions listed in Table (2) are ordered in descending way, starting with the variable of highest impact to the variable with lower impact.

Table 2: Top ten conditions causing mortality in Covid patients in Mexico.

| Disease number | Disease                                      |
|----------------|----------------------------------------------|
|                | KIND OF PATIENT-HOSPITALIZED                 |
| 1              | AGE                                          |
| 2              | ANOTHER CASE-NOT SPECIFIED                   |
| 3              | DIABETES-YES                                 |
| 4              | PNEUMONIA-YES                                |
| 5              | ANOTHER CASE-YES                             |
| 6              | RESULT-POSITIVE Covid-19                     |
| 7              | CARDIOVASCULAR-YES                           |
| 8              | HYPERTENSION-YES                             |
| 9              | INMUSUPR-YES                                 |
| 10             | ANOTHER COM-YES                              |

In the present study, 157,072 patients were analysed with a median age of 51 (IQR 38-65 years). The number of males analysed is 88,602, and the number of females is 68,470. The standard deviation is lower than 0.49 for all the features. The first results show that condition number one is the hospitalized patient. It could be considered the principal cause of death in Mexico. This statement makes sense because there is no record for non-hospitalized patients. However, it is not considered a disease or condition. For this reason, this attribute is omitted because this attribute is not a fundamental cause to predict the death possibility in patients infected with Covid-19.

It was found that the patient’s age is the main factor and is considered the principal condition of death in Covid-19 patients for Mexico. This indicator has been one of the main attributes to increase the mortality risk as is detailed in [22]. The attribute ANOTHER CASE-NOT SPECIFIED is related to patients who do not know if they had contact with Covid-19 infected persons. The attributes DIABETES and PNEUMONIA occupy the 3rd and 4th place respectively, the condition of diabetes is one of the major ailments in the Mexican people. It is responsible for the high number of registered deaths, and Gupta et al. reported that diabetes is a significant risk factor for mortality in patients with other types of influenza [23]. Also, the attribute ANOTHER COM-YES is connected to patients who admit to having been in contact with Covid-19-infected persons. Thus, this analysis proves that exposure to infected people can be considered a condition to increase the risk of suffering complications and death. It is a critical problem in Mexico because society ignores preventive social distancing norms. The sixth attribute, RESULT-POSITIVE Covid19, surprisingly is not the 1st cause of death in patients with Covid-19 since it seems that this disease is only aggravated if the patients suffer another condition.

Disease 7 and 8 are connected with cardiovascular and hypertension problems, respectively. These diseases have a significant contribution as indicators of mortality in Covid-19 patients. These risk factors have been identified in other researches as is mentioned in [24].

The attribute INMUSUPR-YES (immunosuppressed) indicates that the patient has a depressed immune system that aids in having a higher risk of death because the patient’s immune system is
not working optimally. Finally, ANOTHER COM-YES attribute identifies if the patient has been diagnosed with other diseases. In addition, it is possible to mention that the attributes RESULT-PENDING RESULT, TABACISM-YES, ASTHMA-YES, OBESITY-YES, OTHER COM-YES, are the following conditions described in Table (2) since these factors contribute significantly to increase the death risk. In this study, it was found that obesity is not the most crucial disease for increasing the risk of death in Mexico, as some studies propose.

4.1 Model analysis

Initially, the data analysis determined that out of 1,048,575 people studied, 970,039 were registered alive. This amount represents 92.51% of the population tested. Otherwise, 78,536 observations were registered as death, and this corresponds to 7.489% of the sampled population. The dataset considered for the analysis has 157,072 observations and was segmented in both train and test sets.

Figure (2) shows the ROC curve that evaluates the model performance in the testing phase. Of note is that the accuracy accomplished by the model was the same in the training stage. The ROC curve indicates the correctly predicted cases (TPR) precision with the actual cases as correct and those that are not as FPR. In Figure (2), the green curve reflects the accuracy of the model, with the figure of a right triangle being the ideal shape indicating 100% model accuracy. The red dotted line indicates the limit to establish the low accuracy of the model equivalent to an accuracy of 50%. The accuracy obtained in the training and testing phases is equal to 87%, which is considered a satisfactory accuracy regarded the given inputs in the model. For the test phase, the sensitivity of the model and the other classifiers are equal to 87%, and these values are reported in Table (3). It gives the certainty that the data have a uniform behaviour and there are not variations in the predicted cases by the model.

![Figure 2: Performance of ROC curve.](image.png)
The classification report shows that the model has an accuracy of 87%. The model predicts with high precision the survival probabilities that the patients have considering their health conditions. It means that whether the patient has all the diseases listed in Table (2), the patient has an 87% chance of dying. Or by otherwise has the 13% chance of living. A detailed description of the classes estimated is reported in a confusion matrix, which is shown in Figure (3) represented as a heat map.

![Confusion Matrix](image)

Figure 3: Predicted vs Actual diagnostics.

5 Conclusions

This study presents an Artificial Intelligence approach for estimating the top ten critical diseases or conditions for Covid-19 patients in Mexico. This approach is based on both a feature selection process and a Logistic Regression model. The model aims to evaluate the patient’s survival probabilities, and it is considered a suitable alternative for predicting death risk at the time the patient is hospitalized. One of the main challenges to examine the Covid-19 patients is the limited available information. For this reason, this research makes an effort to extract more information from standard laboratory results and develop a beneficial tool that guides health professionals care workers intake crucial decisions and use specialized diagnostic procedures for the group of patients with the highest risk of mortality. In particular, the multivariate ordinal regression generated in this research demonstrated that the diseases listed in Table (2) are independent predictors of Covid-19 patient’s severity.

For this research, the Machine learning approach allows estimating the most influential variables that cause the death of patients hospitalized by Covid-19 with an accuracy of 87% in the training and testing phases. This analysis is considered multivariate, and its behaviour is more complex compared to univariate problems. The accuracy achieved by the model is estimated high compared
to other similar researches. For instance, a satisfactory accuracy of 70% was obtained by Bhandari Sudhir et al. [25]. Some researchers used different Machine learning algorithms to find the most relevant diseases that increase the mortality risk. Some of them are based on Artificial neural networks, and Deep Neural Networks achieving an accuracy of about 94%, as is detailed in [26]. However, these implementations are more complex and require a sophisticated hyperparameters fine-tuning process according to the given dataset. Therefore, the outcomes show that Logistic regression is an excellent alternative for analyzing the given dataset and make predictions with low computational cost. Likewise, the model presented in this research reveals a probability of 87% of death for the cases where the people have the principal diseases founded by the feature selection process. The outcomes obtained by the model proposed to allow promoting actions that improve the patient’s treatment, such as taking preventive measures for patients with the top ten critical diseases and taking care of the conditions that cause complications in Covid-19 patients. The precision of the probabilities of survival or death was satisfactory considering the high-order multivariable problem.

Future work will consider the model performance improvement by testing other Machine Learning models such as Artificial Neural Networks, Deep Neural Networks, and Support Vector Classification. Further, the approach used in this study can be extended to predict other diseases or features.

6 Acknowledgments

The authors thank the Ministry of Health for the information shared on the www.gob.mx/salud website, as well as the people who have kept the detailed record of the data presented in this research, to Aula CIMNE Morelia and the Universidad Michoacana de San Nicolás de Hidalgo.

References

[1] Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. Covid-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv, 2020.

[2] LJ Muhammad, Ebrahem A Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty, and Ibrahim Alh Mohammed. Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. SN computer science, 2(1):1–13, 2021.

[3] Edison Ong, Mei U Wong, Anthony Huffman, and Yongqun He. Covid-19 coronavirus vaccine design using reverse vaccinology and machine learning. Frontiers in immunology, 11:1581, 2020.

[4] Issam El Naqa and Martin J Murphy. What is machine learning? In machine learning in radiation oncology, pages 3–11. Springer, 2015.

[5] Changyu Fan, Linping Liu, Wei Guo, Anuo Yang, Chenchen Ye, Maitixirepu Jilili, Meina Ren, Peng Xu, Hexing Long, and Yufan Wang. Prediction of epidemic spread of the 2019 novel coronavirus driven by spring festival transportation in china: A population-based study. International journal of environmental research and public health, 17(5):1679, 2020.
[6] Alba Grifoni, John Sidney, Yun Zhang, Richard H Scheuermann, Bjoern Peters, and Alessandro Sette. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to sars-cov-2. *Cell host & microbe*, 2020.

[7] Zhenjian He. What further should be done to control covid-19 outbreaks in addition to cases isolation and contact tracing measures? *BMC medicine*, 18(1):1–3, 2020.

[8] Rui Huang, Miao Liu, and Yongmei Ding. Spatial-temporal distribution of covid-19 in china and its prediction: A data-driven modeling analysis. *The Journal of Infection in Developing Countries*, 14(03):246–253, 2020.

[9] Lixiang Li, Zihang Yang, Zhongkai Dang, Cui Meng, Jingze Huang, Haotian Meng, Deyu Wang, Guanhua Chen, Jiaxuan Zhang, Haipeng Peng, et al. Propagation analysis and prediction of the covid-19. *Infectious Disease Modelling*, 5:282–292, 2020.

[10] Rado Doričić, Tanja Ćorić, Morana Tomljenović, Danijela Lakseljac, Amir Muzur, and Branko Kolarić. International journal of environmental research and public health. *International Journal of Environmental Research and Public Health*, 15(11), 2018.

[11] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, and Justin Lessler. What is machine learning? a primer for the epidemiologist. *American journal of epidemiology*, 188(12):2222–2239, 2019.

[12] Gerardo Chowell, Amna Tariq, and James M Hyman. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. *BMC medicine*, 17(1):164, 2019.

[13] Daniela A Gomez-Cravioto, Ramon E Diaz-Ramos, Francisco J Cantu-Ortiz, and Hector G Ceballos. An exploration and forecast of covid-19 in mexico with machine learning. 2020.

[14] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8):1477–1494, 2018.

[15] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

[16] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[17] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[18] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[19] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18, 2014.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[21] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[22] Qiurong Ruan, Kun Yang, Wenxia Wang, Lingyu Jiang, and Jianxin Song. Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china. *Intensive care medicine*, 46(5):846–848, 2020.

[23] Ritesh Gupta, Amerta Ghosh, Awadhesh Kumar Singh, and Anoop Misra. Clinical considerations for patients with diabetes in times of covid-19 epidemic. *Diabetes & metabolic syndrome*, 14(3):211, 2020.

[24] Rong-Hui Du, Li-Rong Liang, Cheng-Qing Yang, Wen Wang, Tan-Ze Cao, Ming Li, Guang-Yun Guo, Juan Du, Chun-Lan Zheng, Qi Zhu, et al. Predictors of mortality for patients with covid-19 pneumonia caused by sars-cov-2: a prospective cohort study. *European Respiratory Journal*, 55(5), 2020.

[25] Sudhir Bhandari, Ajit Singh Shaktawat, Amit Tak, Bhoopendra Patel, Jyotsna Shukla, Sanjay Singhal, Kapil Gupta, Jitendra Gupta, Shivankan Kakkar, Amitabh Dube, et al. Logistic regression analysis to predict mortality risk in covid-19 patients from routine hematologic parameters. *Ibnosina Journal of Medicine and Biomedical Sciences*, 12(2):123, 2020.

[26] Farzaneh Mohammadi, Hamidreza Pourzamani, Hossein Karimi, Maryam Mohammadi, Mohammad Mohammadi, Nahid Ardalan, Roya Khoshravesh, Hassan Pooresmaeiel, Samaneh Shahabi, Mostafa Sabahi, et al. Artificial neural network and logistic regression modelling to characterize covid-19 infected patients in local areas of iran. *Biomedical Journal*, 2021.

**List of Tables**

|   |   |
|---|---|
| 1 | Dummy variables generated as a result of the one-hot encoding. | 6 |
| 2 | Top ten conditions causing mortality in Covid patients in Mexico. | 7 |
| 3 | Classification report. | 9 |

**List of Figures**

|   |   |
|---|---|
| 1 | Flow chart - Backward diagram process. | 4 |
| 2 | Performance of ROC curve. | 8 |
| 3 | Predicted vs Actual diagnostics. | 9 |