Analysis of Correlation among Previous Health Conditions and Deaths for COVID-19 in the Mexican Population to Create a Prediction Model.

Erick Jesús Ríos González

November 7, 2023

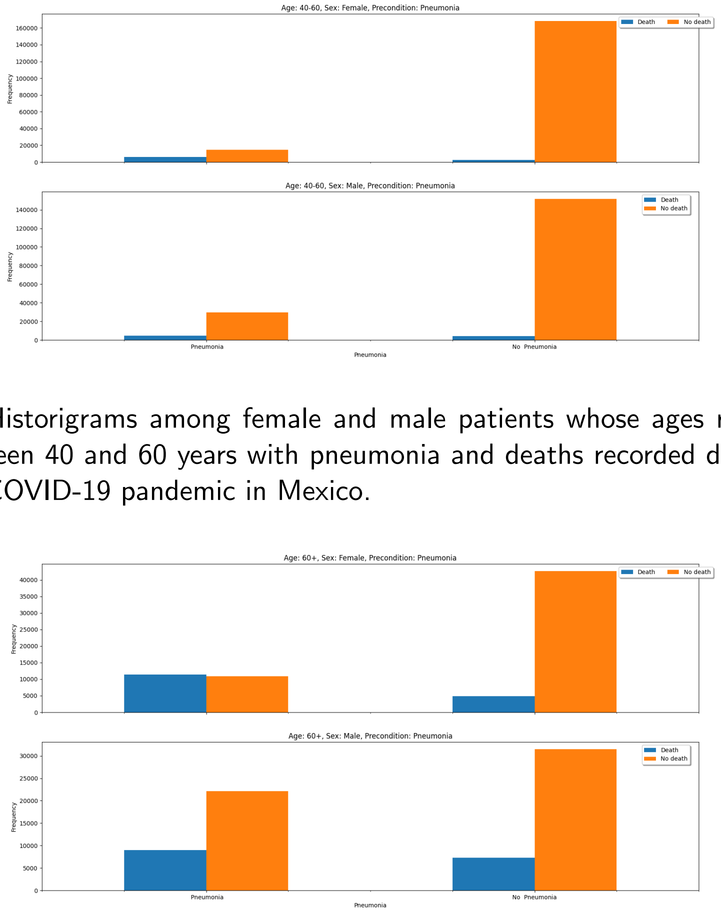
Abstract

Using Machine Learning to determine the possibility of a person dying or surviving due to pre-medical conditions in the Mexican population represents an important issue for national health, since the impact of these studies can raise awareness among the population to have better health. Using a database provided by the Epidemiological Surveillance System for Respiratory Viral Diseases of the Government of Mexico, we trained different classification models using characteristic preconditions of patients entering hospitals diagnosed with COVID-19. With a single input, the models can make predictions that closely match those reported by the Mexican health system.

1 Introduction

1.1 Data

We used data on medical preconditions in the Mexican population provided by the Epidemiological Surveillance System for Viral Respiratory Diseases of the Government of Mexico [Sistema de Vigilancia Epidemiológica de Enfermedades Respiratorio Viral (2020)]. The data was divided into four age subsets according to the diagnosed cases: 0 to 20 years, 20 to 40 years, 40 to 60 years, and 60 or older. The diagnosed patients are Mexicans of all ages and with different medical preconditions. A record of 19 pre-medical conditions was taken from each patient, as well as their sex and final status. The final status was a date of death or a default date, which indicated that the patient had survived the disease.



(a) Historigrams among female and male patients whose ages range between 40 and 60 years with pneumonia and deaths recorded during the COVID-19 pandemic in Mexico.

(b) Historigrams among female and male patients whose ages range from 60 years and older with pneumonia and deaths recorded during the COVID-19 pandemic in Mexico.

Figure 1. Examples of Historograms between the patient's age, sex, medical preconditions, and recorded deaths. The other examples can be found in González (2023)

2 Models

Our overall process involved extracting features from patients and their prior medical conditions to iteratively train and use different prediction models (Szymański and Kajdanowicz, 2017) such as `Neural Networks, Random Forest, KNN, Decision Tree` and `Naive Bayes`. We use grid search to obtain the best hyperparameters and model selection.

2.1 Features

In the `Boolean features`, 1 means ”yes” and 2 means ”no”.

- sex: 1 for female and 2 for male.
- pneumonia: whether the patient already have air sacs inflammation or not.

- copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
- asthma: whether the patient has asthma or not.
- inmsupr: whether the patient is immunosuppressed or not.
- hypertension: whether the patient has hypertension or not.
- cardiovascular: whether the patient has heart or blood vessels related disease.
- renal chronic: whether the patient has chronic renal disease or not.
- obesity: whether the patient is obese or not.
- tobacco: whether the patient is a tobacco user..
- set: 1 for 0-20 years, 2 for 20-40 years, 3 for 40-60 years and 4 for 60+ years.

2.2 Data division for training the artificial neural network

The database used for training the Artificial Neural Network was divided into 70% for the training set and 30% for the test set. We looked for the number of labels between death and survivor to be equally distributed in both sets as can be seen in Figure 2

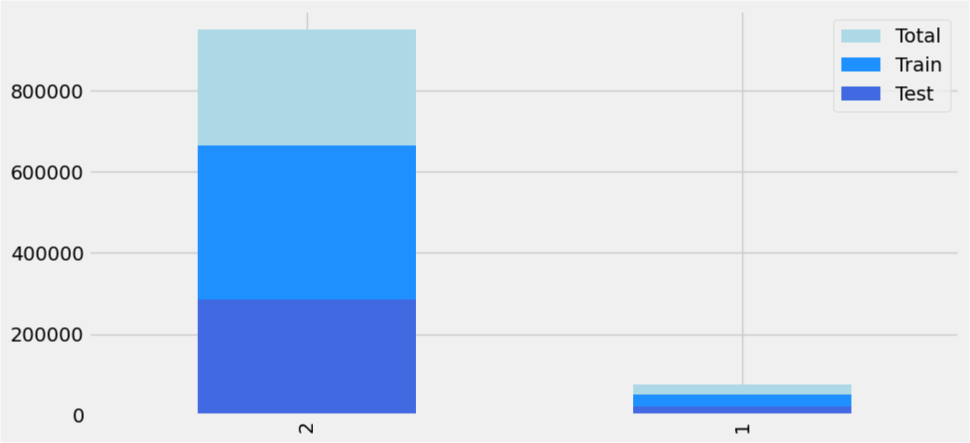
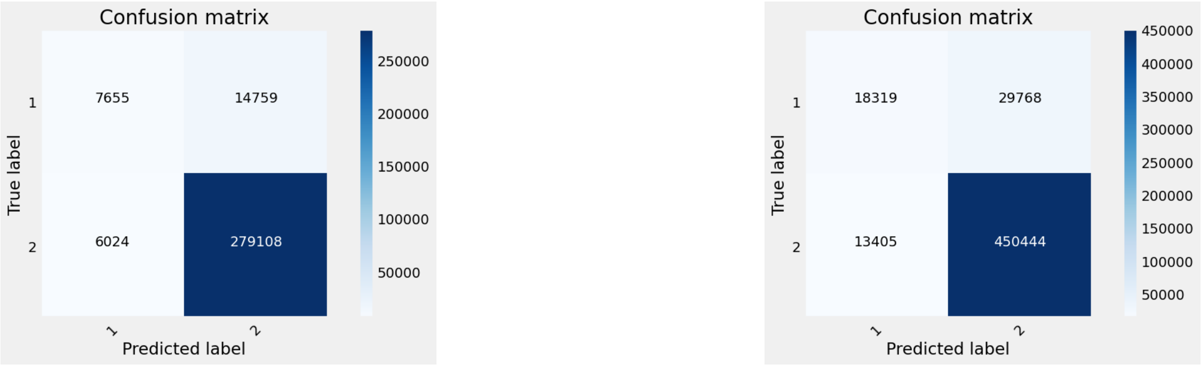


Figure 2. Distribution of validation and training data of the total data set.

3 Results

In order to efficiently classify the final state of the patient, we only take as relevant characteristics `'sex'`, `'diabetes'`, `'copd'`, `'asthma'`, `'inmsupr'`, `'hypertension'`, `'set'`, `'obesity'`, `'pneumonia'`, `'renal_chronic'`, `'tobacco'` and `'cardiovascular'`. We found that the `decision tree` uses all the features described above, but not the inter-patient `pregnancy` feature. The classification made by the model closely coincided with the classifications reported by the Mexican health system, however there was a classification problem for those patients who died according to the entries given to the decision tree (see Figrue 3a and Figure 3b). We used cross-validation to rectify that the neural network outputs closely matched those reported. The pregnancy characteristic had less precision when classifying the final state of the patient, which is why it was removed as a main characteristic. For some data sets, the classification made is not in accordance with what is reported in reality, but since it is smaller as a report, a certain relationship can be inferred between it and the data housed in our database. We tried using a simpler model for classification, but it doesn't seem to get better results than the one described in this section.



(a) Confusion matrix for the test data assuming the database as a total set.

(b) Confusion matrix for the test data assuming the database as a subset of only the female sex.

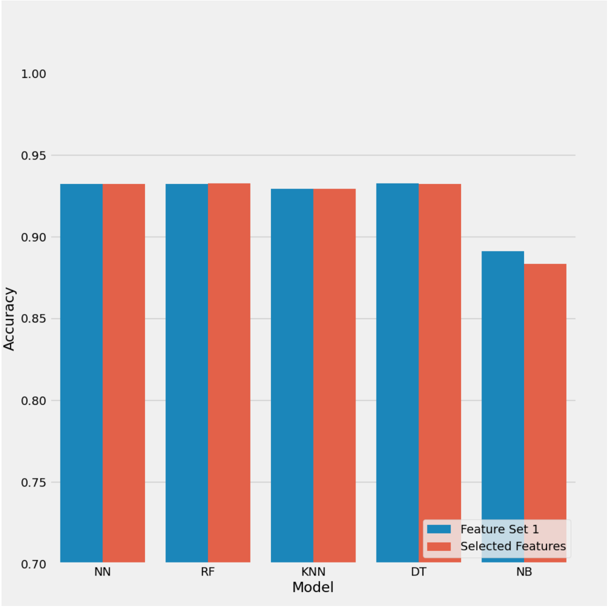


Figure 4. Comparison of the accuracy of different models trained using different feature sets. We can see that Naive Bayes has the lowest accuracy.

4 Discussion

4.1 Features

We wanted to know which characteristics were most useful in predicting the death or survival of a patient diagnosed with COVID-19 given their preconditions. To do this we eliminate features one at a time. Figure 5 shows the results received from 2 different sets of features: the first one just with the features mentioned in 2 and the second one including the feature of `pregnancy`. As can be seen, the precision improves as the number of features for training the model increases. The characteristic of `pregnancy` was not a factor as important as originally believed.

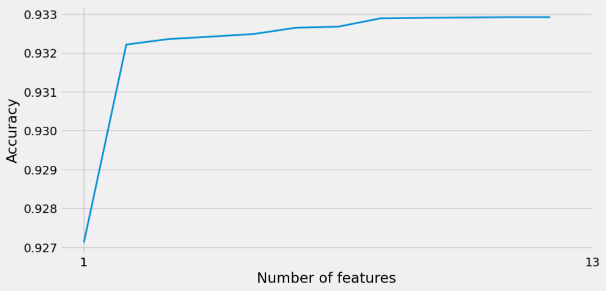


Figure 5. Comparison of the accuracy of different models trained using different feature sets.

4.2 Limitations

Due to the large number of missing values present in the database, it is not possible to use the 19 characteristics of medical preconditions in the patients. This may be affecting the accuracy of classifying patients with the chosen medical conditions. That is why a better database could help us improve the classification accuracy.

5 Conclusions

Using a decision tree, grid search and cross-validation to optimize the hyperparameters, we managed to classify patient survivals close to that reported by the Mexican health system. Our research shows that for a patient aged 60 or older and with medical preconditions such as pneumonia, we can classify their final state quite accurately according to their sex and other characteristics. But the model presents certain errors for different age ranges than the one mentioned above, where the main classification is in those who survive. This is largely due to the fact that the database reported deaths linked to the aforementioned medical preconditions.

This is because from the database the reported deaths linked to the aforementioned medical conditions do not appear to be highly correlated for ages under 60 years of age. Well, from the historigrams obtained we can see that it seems that those who have these medical conditions suffer fewer deaths than those who do not suffer from them. This part of the database needs to be reviewed. Another action in the future may be to resample the data set. Change the metric space in the characteristics and results, with this we intend to avoid undersampling or oversampling.

References

de Vigilancia Epidemiológica de Enfermedades Respiratorio Viral, S. (2020). Información referente a casos de covid-19 en méxico. Consultado 27 octubre, 2023 de https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico.

González, E. J. R. (2023). Historigrams. Consultado 27 octubre, 2023 de https://github.com/erick-rios/COVID19-erick-rios/tree/main/reports/figures.

Szymański, P. and Kajdanowicz, T. (2017). A scikit-based python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460*.