ISSN 2007-9737

Leveraging Machine Learning to Unveil the Critical Role of Geographic Factors in COVID-19 Mortality in Mexico

Christian E. Maldonado-Sifuentes[1], Mariano Vargas-Santiago*,[1], Diana A. Leon-Velasco[2], M. Cristina Ortega-García[3], Yoel Ledo-Mezquita[2], Francisco A. Castillo-Velasquez[4]

[1] Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT), Mexico

[2] Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM), Campus Ciudad de México, Mexico

[3] Transdisciplinary Research for Augmented Innovation Laboratory (TRAI-L), Mexico

[4] Universidad Politécnica de Querétaro, Mexico

{christian.maldonado,mariano.vargas}@conahcyt.mx, assaely.leon@tec.mx, cristina.ortega@trai-l.com, yledo@tec.mx, francisco.castillo@upq.edu.mx

Abstract. In this paper, we present an in-depth analysis leveraging several renowned machine learning techniques, including Snap Random Forest, XGBoost, Extra Trees, and Snap Decision Trees, to characterize comorbidity factors influencing the Mexican population. Distinct from existing literature, our study undertakes a comprehensive exploration of algorithms within a defined search space, conducting experiments ranging from coarse to fine granularity. This approach, coupled with machine learning-driven feature enhancement, enables us to deeply characterize the factors most significantly affecting COVID-19 mortality rates within the Mexican demographic. Contrary to other studies, which obscure the identification of primary factors for local populations, our findings reveal that geographical factors such as residence location hold greater significance than even comorbidities, indicating that socioeconomic factors play a pivotal role in the survival outcomes of the Mexican population. This research not only contributes to the targeted understanding of COVID-19 mortality drivers in Mexico but also highlights the critical influence of socioeconomic determinants, offering valuable insights for public health strategies and policy formulation.

Keywords. Diabetes, COVID-19, machine learning, SARS CoV-2, Cox, RMST.

1 Introduction

The advent of COVID-19 has instigated a global health crisis of unparalleled magnitude, prompting a concerted effort across healthcare systems worldwide to counteract its ramifications [2, 10, 12]. This pandemic has underscored the critical need for advanced medical research and data analytics to dissect and mitigate the virus's impacts efficiently. Central to this effort is the analysis of vast datasets to identify patterns and predictors of COVID-19 outcomes, with particular emphasis on the significance of patient comorbidities and geographic statistics in influencing mortality rates [2, 7, 8, 10, 12].

This scenario has propelled the development of an Automated Machine Learning (AutoML) framework, crafted to harness the latest in machine learning innovation to expedite the evaluation of COVID-19 mortality risks. By optimizing the model development process, AutoML aims to enrich our comprehension of the medical and societal variables influencing

COVID-19 mortality, offering crucial insights to both healthcare practitioners and researchers. The dynamic and evolving nature of COVID-19 data renders the adaptability and automation features of AutoML exceptionally valuable. Such capabilities facilitate rapid algorithmic adjustments and hyperparameter optimization to assimilate new findings and data, establishing AutoML as an essential asset in combating COVID-19.

Our investigation leverages a comprehensive dataset provided by the Mexican Federal Government, chronicling the pandemic's impact on the Mexican populace from January 1, 2023, to August 8, 2023. This dataset encompasses detailed information on 1,021,380 patients, including demographic, clinical outcomes, and mortality data, thereby offering a unique lens through which to examine the multifaceted influences on COVID-19 mortality.

Utilizing various validated methodologies for COVID-19 diagnosis, including antigen testing and clinical epidemiological association, our study utilizes AutoML to dissect this dataset, aiming to unearth pivotal patterns and predictors of mortality. This endeavor aligns with the urgent global requirement for innovative analytical tools capable of pacing with the swiftly evolving pandemic landscape, marking a significant stride in applying AutoML for comprehensive data analytics in confronting the COVID-19 health crisis.

Emerging studies highlight the utility of machine learning in scrutinizing COVID-19 data and AutoML's potential to revolutionize this analysis by enhancing the accessibility and adaptability of advanced data analytics [5, 9, 11]. Building upon these insights, our research endeavors to offer valuable perspectives on the determinants of COVID-19 mortality, showcasing AutoML's utility in pandemic response and preparedness. This paper makes the following contributions:

1. We establish a comprehensive experimental framework that divides into two core components: traditional statistical analysis and a machine learning-based approach. This framework, detailed in Sections 4.1 and 4.2 for statistical analysis and Section 4.4 for machine learning, facilitates a nuanced exploration of the comorbidity factors affecting COVID-19 mortality in the Mexican population.
2. Through our rigorous analysis employing cutting-edge machine learning techniques—specifically Snap Random Forest, XGBoost, Extra Trees, and Snap Decision Trees—we provide an in-depth characterization of the comorbidity factors influencing COVID-19 mortality.

   Our study distinguishes itself by executing a comprehensive exploration of algorithms within a defined search space, ranging from coarse to fine granularity. This method, enhanced by machine learning-driven feature selection, allows for a deepened understanding of the critical factors affecting mortality rates.
3. Contrary to prevailing studies that predominantly focus on comorbidities as the key mortality determinants, our findings reveal the greater significance of residential location, pointing to socioeconomic factors as pivotal in determining survival outcomes in the Mexican context.

   This novel insight emphasizes the need for public health strategies and policy formulation to consider socioeconomic determinants alongside medical factors.
4. By leveraging a combination of traditional statistical and modern machine learning methodologies, our research contributes a unique perspective to the body of knowledge.

   It not only provides a targeted analysis of COVID-19 mortality drivers in Mexico but also underscores the crucial influence of socioeconomic factors on health outcomes. Our study paves the way for informed public health interventions aimed at reducing mortality within socioeconomically diverse populations.

The rest of this work is organized as follows: Section 2 outlines the methodologies employed in our study and presents the dataset we are using. Section 3 discusses some of the related work to our study. Section 4 describes the experimental setup for our study, which is divided into two subsections statistical or traditional analysis 4.2 and machine learning based analysis 4.4.

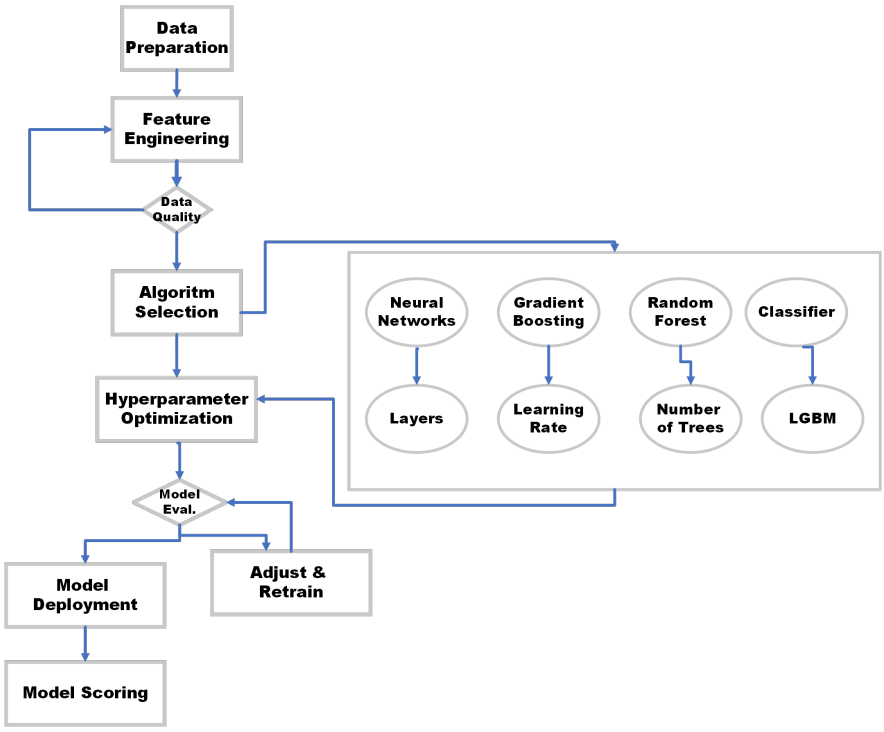

Fig. 1. Experimental setup for machine learning based analysis

Section 5 presents the findings from our experiments, for traditional statistical methods and those after using the machine Learning based analysis. Finally, conclusions and future work are given in Section 6.

2 Methods

This retrospective study is predicated on data from the Mexican Federal Government, endorsed by the Epidemiological Surveillance System for Viral Respiratory Diseases under the Mexican Ministry of Health's purview. Ethical consent for this data's utilization was comprehensively secured from the pertinent health ethics committees.

2.1 Dataset

Employing the COVID-19 Mexican Patients Dataset, our study scrutinizes the demographic profiles, characteristics, and clinical outcomes of the Mexican demographic amid the COVID-19 pandemic. This dataset, curated and disclosed by the Mexican Federal Government and the Ministry of Health, spans January 1, 2023, to August 8, 2023. It aggregates data from 475 Viral Respiratory Disease Monitoring Units, detailing individuals hospitalized following a positive COVID-19 test, totaling 1,021,380 patients with comprehensive mortality data.

2.2 COVID-19 Determination

COVID-19 diagnoses were ascertained through SARS-CoV-2 antigen detection via nasal swabs, conducted across various government-affiliated surveillance and healthcare establishments. The validation of COVID-19 cases employed three methodologies: clinical epidemiological association, a deliberation committee's verdict, or antigen testing. Conversely, a negative result signified the antigen's absence in the sample.

2.2.1 Repeatability Across Databases

Our methodology, characterized by its repeatability across diverse healthcare settings, benefits from the flexible and user-oriented nature of codeless platforms like Uber Ludwig, which significantly enhance and simplify the creation of Linux shell level scripting to allow for automation of ML tasks. This approach incorporates insights from recent studies on machine learning in COVID-19 analysis, spanning techniques from functional and sentiment analysis to causal learning and mental health data examination. Such breadth in machine learning application highlights AutoML's potential to navigate the intricate dynamics of COVID-19 mortality influences adeptly.

Table 1. Patient Demographics and Covariates

	Diabetic	**Non-diabetic**	**Total Individuals**
Total Individuals	80,346	940,035	1,021,380
Male	30,533 (38%)	391,404 (45%)	422,344 (41%)
Female	49,813 (62%)	548, 631 (58%)	599,036 (59%)
Native	80,199 (8%)	935,950 (92%)	1,017,140 (99%)
Diabetes	80,346 (100%)	0 (0%)	80,346 (8%)
Hypertension	44,851 (56%)	67,248 (7%)	112,151(11%)
Obesity	15,989 (20%)	61,978 (7%)	78,002 (8%)
Smoking	5,007 (6%)	36,773 (4%)	41,798 (4%)
Pneumonia	6,241 (8%)	22,061 (2%)	28,373 (2%)
ICU	508 (0.7%)	2,181 (0.23%)	2,704 (0.26%)
Intubation	809 (1%)	2,607 (0.28%)	3,436 (0.33%)
Death	2,198 (2.7%)	4,371 (0.46%)	6,581 (0.64%)

3 Related Work

The onset of COVID-19 has precipitated a flux of research employing machine learning to elucidate the virus's outcomes and impacts. Our study aligns with and extends this corpus of work by emphasizing the Mexican demographic and integrating socioeconomic variables into our analysis, distinguishing our research within the burgeoning field of machine learning applications in COVID-19 analysis.

Several recent studies have employed machine learning models to predict COVID-19 outcomes based on patient data. For example, He et al. [3] developed a generalizable and easy-to-use COVID-19 severity stratification model utilizing immune-phenotyping and machine learning, underscoring the importance of a comprehensive approach in determining patient outcomes.

Similarly, Badiola-Zabala et al. [1] conducted a systematic literature review of clinical decision support approaches during the pandemic, highlighting the effectiveness of ML- and AI-based models in predicting mortality among COVID-19 patients.

Moreover, Lages dos Santos et al. [4] provided a comparative analysis of machine learning algorithms for predicting COVID-19 mortality in children and adolescents using a large public dataset in Brazil, indicating the predictive power of various factors, including demographic data and comorbidities.

These studies exemplify the significant potential of machine learning in enhancing COVID-19 prognosis and management through the analysis of patient data. In contrast to these studies, our research expands the scope of analysis by employing advanced techniques such as Snap Random Forest, XGBoost, Extra Trees, and Snap Decision Trees. Our methodological approach involves a more nuanced exploration of the algorithmic search space, allowing for a detailed characterization of the influences on COVID-19 mortality.

This granularity surpasses the typical scope of existing literature by refining feature selection and optimization processes to better capture the complex interplay of factors affecting mortality rates.

Furthermore, while most studies concentrate on medical and biological predictors, our investigation reveals the paramount importance of residential location as a determinant of COVID-19 mortality in the Mexican context.

This finding aligns with research by Mendez-Astudillo [6], which points to socioeconomic factors and economic inequalities as critical determinants of health outcomes during the pandemic, underscoring the influence of social determinants on public health.

Our work contributes a unique perspective by integrating socioeconomic considerations with comorbidity analysis, thereby offering a more holistic understanding of the drivers behind COVID-19 mortality. This approach not only fills a gap in the current literature but also serves as a foundation for future research and policy-making aimed at mitigating the impacts of the pandemic on vulnerable populations.

By examining these recent works in conjunction with our study, it becomes evident that while there is a consensus on the importance of comorbidities and demographic data in predicting COVID-19 outcomes, the role of socioeconomic factors, particularly in the context of Mexico, remains underexplored. Our research aims to bridge this gap, offering insights into the significance of residential location and socioeconomic status in shaping COVID-19 mortality rates.

4 Experimental Setup

The experimental framework is delineated into two principal components. Initially, we delineate the methodology referred to as traditional statistical analysis, as depicted in Sections 4.1 and 4.2. Subsequently, the foundational setup for the analysis predicated on machine learning techniques is presented, as specified in Section 4.4. This investigation constituted a retrospective study that drew upon data sourced from the Mexican Federal Government. The data set used in this study had been publicly disseminated and subjected to validation procedures by the Epidemiological Surveillance System for Viral Respiratory Diseases under the auspices of the Mexican Ministry of Health.

Ethical approval for the use of this data set was obtained in full from the ethics committees associated with the Ministry of Health

4.1 Determination of COVID-19

The diagnosis of COVID-19 was established by detecting the SARS-CoV-2 antigen through nasal swab testing. This diagnostic procedure was conducted at various surveillance and healthcare facilities under the jurisdiction of the Mexican Government, with readily available results.

We utilize three distinct approaches to validate positive COVID-19 cases, which encompass the following: validation via clinical epidemiological association, validation through a deliberation committee, or validation through antigen testing. On the contrary, a negative status indicated the absence of this antigen in the tested samples.

4.2 Statistical Analysis

The demographic and diabetes-related characteristics of individuals testing positive for the SARS-CoV-2 antigen were subjected to analysis employing descriptive statistical methods. Comparative assessments among patients, taking into account relevant covariates, were performed using $T-$tests and X^2 tests. The primary endpoint under investigation was patient survival, defined as the duration from the onset of COVID-19 symptoms to the point of mortality, with censoring applied at the final enrollment date for adult COVID-19 patients admitted to hospitals.

To estimate the survival curve, Kaplan-Meier curves were generated and the statistical significance of the survival durations between hospitalized adult patients with and without diabetes was assessed using the log-rank test. A Cox proportional hazards model was used to calculate the hazard ratio and establish a confidence interval (CI) 95% to gauge the effect of treatment.

All statistical tests were two-sided, and $p-$value less than 0.05 was considered indicative of statistical significance. In addition to overall survival analysis, the calculation of restricted mean survival time (RMST) was executed for both diabetic and non-diabetic adult COVID-19 patients admitted to hospitals, following propensity matching to mitigate the impact of confounding variables. This approach entailed fitting a parametric survival model to the dataset to estimate the mean survival time for the two distinct groups under investigation.

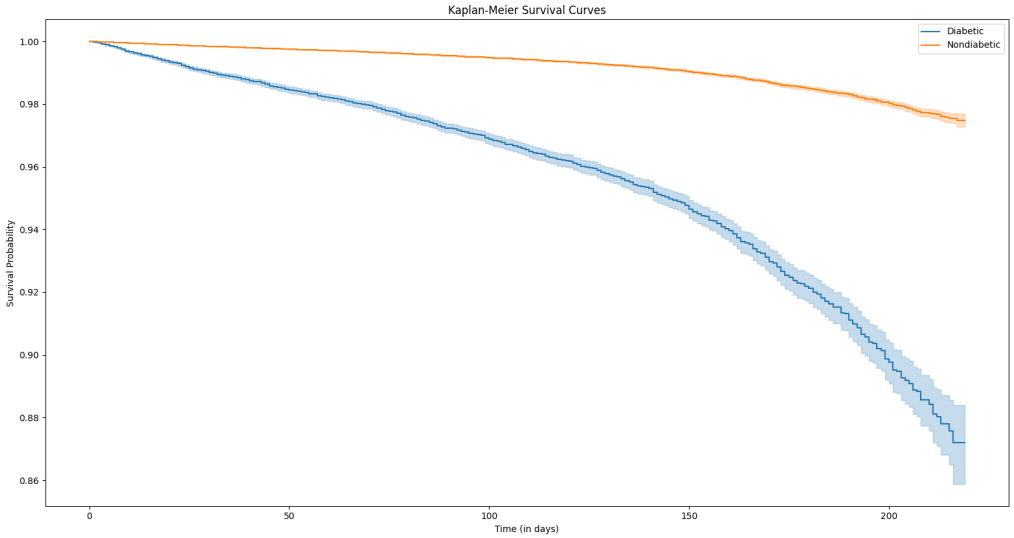

Fig. 2. Multi-variable Kaplan–Meier survival plots comparing diabetic and non-diabetic patients

4.3 Variables

The objective of this study is to investigate the influence of diabetes on the prognosis of COVID-19 in Mexican patients, specifically focusing on the likelihood of in-hospital mortality among those with diabetes. This research constitutes a substantial and nationwide retrospective cohort study, which is poised to offer valuable information on the interplay between diabetes and the outcomes of COVID-19. Such insights have the potential to guide the formulation of effective mitigation strategies and improve the patient triage process.

Key demographic information, including sex, age, country of origin, pre-existing health conditions (such as hypertension, diabetes,obesity and immunosuppression), smoking habits, and pregnancy status, was systematically documented for each individual. Data related to COVID-19 status included records of antigen test results and antigen sample collection. It is worth noting that during their hospitalization, no publicly accessible information was disclosed regarding the patients' clinical progression.

Table 2. Confusion Matrix

Observed	**Predicted**	
	Dead	Alive
Dead	**559**	**93**
Alive	**141**	**936**
% Correct	Overall % Correct: 86.5%	

4.4 Machine Learning Experimental Setup

The experimental setup for our machine learning model development consists of a multi-stage process, as illustrated in the figure 1. The procedure is initiated with Data Preparation, where raw data is collected and pre-processed to ensure it is in a suitable format for analysis. This stage is crucial for the subsequent steps as it directly affects the quality of the insights derived from the data.

Following this, we engage in Feature Engineering, which involves creating new features from the existing data to improve the model's predictive power. This step also includes assessing Data Quality to ensure the integrity and appropriateness of the data for the machine learning algorithms. The next phase is Algorithm Selection, where we choose appropriate machine

learning algorithms based on the nature of the data and the problem statement.

This decision impacts the model's ability to learn from the data and make accurate predictions. After selecting the algorithms, Hyperparameter Optimization is conducted to find the optimal settings for each algorithm, enhancing the model's performance. This involves tuning various parameters that govern the learning process of the models.

Once the models are trained with the best hyperparameters, Model Evaluation is performed using appropriate metrics to assess their performance. If the models do not meet the desired performance criteria, they are subject to Adjustment and Retraining to improve their accuracy and reliability. Upon achieving satisfactory evaluation metrics, the model is then moved to Model Deployment, where it is integrated into the production environment to make predictions on new data.

This step is critical for translating the model's capabilities into practical applications. Lastly, Model Scoring is performed on the deployed model, where it is continuously monitored and scored based on its performance in the live environment. This ensures that the model remains accurate and relevant over time.

Within the Algorithm Selection stage, a range of machine learning algorithms is considered. These include Neural Networks, with a focus on the configuration of their Layers; Gradient Boosting methods, with an emphasis on the Learning Rate; Random Forest, where the Number of Trees is a significant parameter; and a generic Classifier, which, in this context, appears to be specified as LGBM (Light Gradient Boosting Machine), a type of gradient boosting framework [1].

5 Results

The results section is systematically organized to reflect the bifurcated approach of the experimental framework. Initially, outcomes stemming from the traditional statistical analysis are elucidated, corresponding to the methodologies outlined in Sections 4.1 and 4.2, refer to 5.1. Subsequently, we present the findings derived from the machine learning-based analysis, adhering to the foundational setup delineated in Section 4.4, refer to 5.2.

This sequential presentation facilitates a comprehensive understanding of the experimental results, allowing for a direct comparison between traditional statistical methodologies and modern machine learning approaches in addressing the research objectives.

5.1 Results for Traditional Statistical Methods

In Table 1, we present a comprehensive summary of demographic variables and relevant covariates for people who tested positive for COVID-19, using data sourced from the Mexican Patient Data Set, which is publicly available through the Mexican Ministry of Health.

This study covers a total of 1,021,380 adult patients who were hospitalized due to COVID-19, all of whom have complete records of their mortality outcomes. Within this patient cohort, 38% were male, while 62% were female, with an average age of 38.41 years (standard deviation = 19.50). The main comorbidities prevalent in this population included hypertension (11%), diabetes (7.9%), and obesity (8%), while 4% were identified as smokers. It is noteworthy that the term "Nondiabetic" is used to describe individuals without a diagnosis of diabetes, as no cases of diabetes were identified within this group.

These demographic and clinical characteristics provide a foundational understanding of the patient population under investigation in this study. Figure 2 shows the Kaplan-Meier survival graphs, presenting a comparative analysis of survival probabilities among two groups: COVID-19 patients with and without diabetes. The study encompasses a comprehensive cohort of 1,021,380 individuals, among which 6,581 individuals experienced a fatal outcome related to the disease.

Statistical analysis, specifically the log-rank test, revealed a statistically significant disparity in survival rates between these two groups of

[1] Project files can be found at https://github.com/christianemaldonadomti/MLCovidMexico

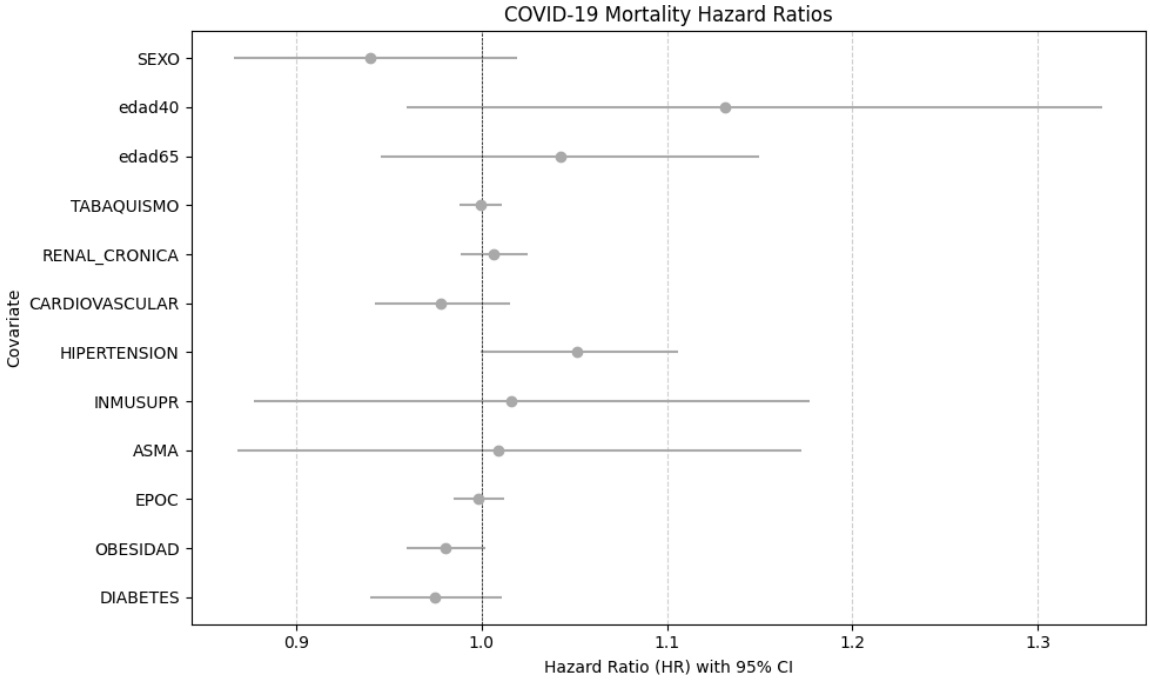

Fig. 3. COVID-19 Cox Mortality Hazard Ratios without location

hospitalized COVID-19 patients (p ¡ 0.01). It is noteworthy that the depicted survival curves adhere to the assumption of proportional hazards, as they do not intersect within the examined time frame.

Although the dataset presents a higher mortality rate among individuals without diabetes, those afflicted with this condition exhibit an acceleration in the progression towards mortality. In other words, they experience a shorter survival time compared to individuals with diabetes.

This phenomenon can be substantiated by referring to Figure 2. Figure 3 depicts the utilization of the Cox proportional hazards model to assess the impact of various covariates on the risk of mortality, with statistical significance determined through the examination of p-values. Our analytical findings yield noteworthy insights into the relationship between comorbidities and the risk of COVID-19 mortality. Specifically, individuals with diabetes displayed a modestly reduced hazard of mortality (Hazard Ratio: 0.975) in comparison to those without diabetes, although this reduction did not attain statistical significance

Table 3. Model Evaluation Measures

Measures	Holdout sc.	X-validation sc.
Accuracy	0.865	0.864
	0.876	0.876
Precision	0.799	0.798
	0.807	0.806
Recall	0.857	0.857
	0.883	0.882
F1	0.827	0.827
	0.843	0.843
Avg. precision	0.892	0.892
	0.904	0.904

(p-value: 0.167). Similarly, obesity was associated with a slight reduction in mortality hazard (Hazard Ratio: 0.981), although this reduction was only marginally significant (p-value: 0.080).

Conversely, hypertension was linked to a minor increase in mortality hazard (Hazard Ratio: 1.051), with a p-value that approached statistical significance at 0.051. Other covariates, including

Chronic Obstructive Pulmonary Disease (COPD), asthma, immunosuppression, cardiovascular conditions, renal chronic conditions, smoking habits, various age categories, and gender, did not demonstrate statistically significant effects on the risk of COVID-19 mortality.

5.2 Results for Machine Learning Based Analysis

Our analysis's effectiveness was quantitatively assessed using a confusion matrix, which provides insights into the model's predictive accuracy by comparing actual outcomes against predictions. Table 2 presents the confusion matrix derived from the model evaluation.

The confusion matrix reveals that out of the cases predicted to result in mortality (Dead), 559 were correctly identified (true positives), while 93 were misclassified (false negatives). Conversely, for the cases predicted to result in survival (Alive), 936 were accurately predicted (true negatives), with 141 instances being incorrectly forecasted as mortalities (false positives). This performance yields an overall prediction accuracy of 86.5%, demonstrating the models' robust ability to discern between survival and mortality outcomes based on the assessed comorbidities. Such a high level of accuracy underscores the potential of employing these machine learning techniques for predictive purposes in medical settings, specifically in prognosticating COVID-19 outcomes.

The differential performance across models, as inferred from the confusion matrix, emphasizes the nuanced understanding these algorithms provide regarding the critical factors affecting mortality rates. Notably, the high overall accuracy achieved across different models suggests that machine learning-driven feature selection significantly contributes to identifying the most impactful predictors of COVID-19 mortality.

Moreover, the detailed evaluation of model performance through various metrics highlights the robustness and reliability of the employed machine learning methodologies. The consistent predictive precision across models indicates the efficacy of the selected algorithms in capturing the complexities inherent in the comorbidity factors of COVID-19 patients.

We delve into the results obtained from our comprehensive analysis, which leverages advanced machine learning techniques to illuminate the intricate dynamics of comorbidity factors affecting COVID-19 mortality rates. The algorithms employed—namely, Snap Random Forest, XGBoost, Extra Trees, and Snap Decision Trees—were rigorously evaluated to ensure a robust exploration of the predictive capabilities pertaining to COVID-19 mortality.

The evaluation metrics presented in Table 3 encapsulate the performance of these models across various dimensions, including accuracy, precision, recall, F1 score, and average precision. These metrics were assessed through both holdout and cross-validation methods to validate the models' consistency and reliability. Our analysis showcased the models' commendable performance in characterizing the comorbidity factors influencing COVID-19 mortality, as evidenced by the evaluation scores tabulated in Table 3. The accuracy of the models, as denoted by the holdout score (Holdout sc.) and cross-validation score (X-validation sc.), was observed to be consistently high, with accuracy scores reaching up to 0.876. This high level of accuracy underscores the effectiveness of the machine learning techniques applied in capturing the complexities inherent in COVID-19 mortality risk factors.

Precision, a measure of the models' ability to correctly identify positive instances among the predicted positives, also demonstrated high performance, with scores up to 0.807. This indicates a significant strength in the models' capability to discern true cases of high mortality risk amidst a plethora of potential predictors.

Recall scores, which reflect the models' capacity to identify all relevant instances, were notably high as well, reaching up to 0.883. This suggests that the models are exceptionally adept at capturing the majority of significant cases, thereby reducing the risk of overlooking critical comorbidity factors.

The F1 score, a harmonic mean of precision and recall, further solidifies the models'

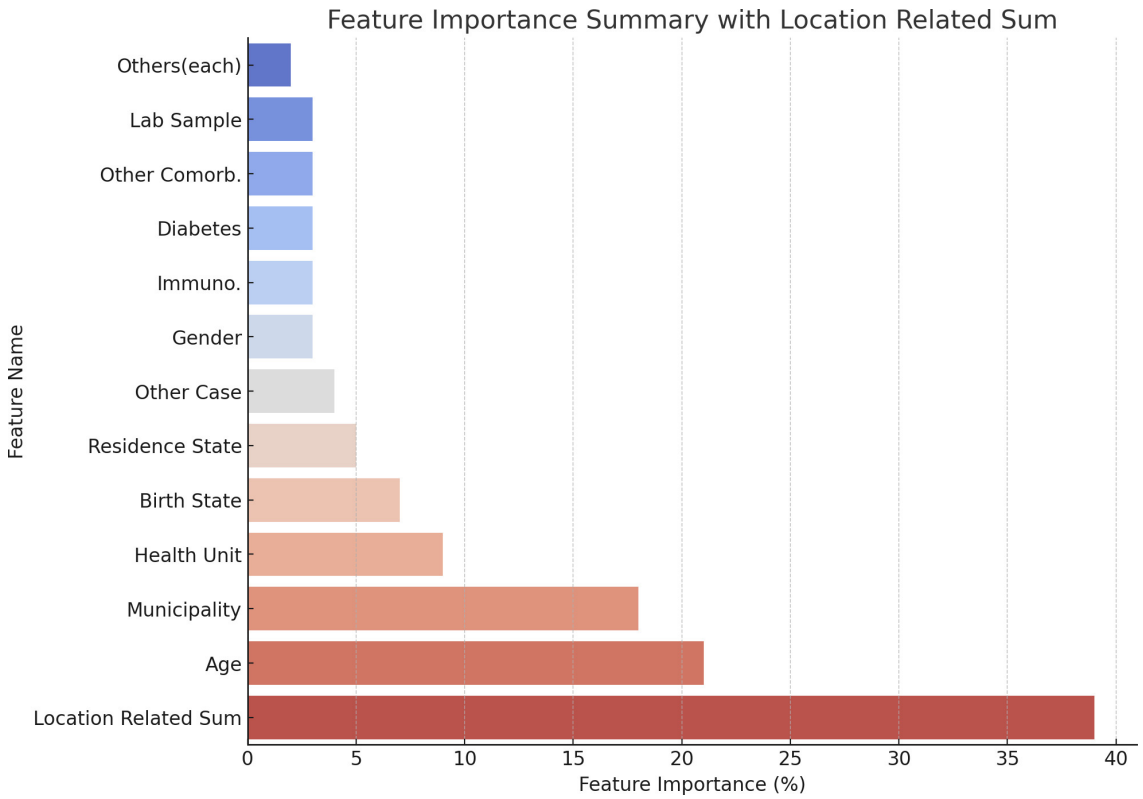

Fig. 4. Leveraging Machine Learning to Unveil the Critical Role of Geographic and Sociodemographic Factors in COVID-19 Mortality Across Mexico

robustness, with scores peaking at 0.843. This balance between precision and recall illustrates the models' overall efficacy in identifying true positives while minimizing false negatives and positives.

Moreover, the average precision score, which provides an aggregate measure of precision across varying thresholds, reached an impressive 0.904. This underscores the models' outstanding precision-recall balance, a critical aspect in the context of medical predictive modeling, where the cost of false negatives can be particularly high.

The performance metrics presented affirm the substantial capability of the employed machine learning models to discern the critical comorbidity factors influencing COVID-19 mortality. The high scores across all evaluated measures attest to the models' precision, recall, and overall accuracy, highlighting their potential utility in aiding public health efforts by providing deepened insights into COVID-19 mortality risk factors.

This detailed exposition not only underscores the predictive prowess of the deployed models but also emphasizes the significance of machine learning-driven feature selection in enhancing our understanding of the key determinants of COVID-19 mortality rates. The findings elucidated herein lay a solid foundation for further research into predictive modeling for infectious diseases, potentially guiding more targeted and effective public health interventions.

5.3 Feature Importance Analysis

A critical component of our study involved the assessment of feature importance, which highlights the relative impact of various factors on the predictive models' outcomes.

Table 4. Feature summary

Feature Name	Feature Importance
- Municipality	18.00%
- Health Unit	9.00%
- Birth State.	7.00%
- Residence State.	5.00%
Sum Loc. Related	**39.00%**
Age	21.00%
Other Case	4.00%
Lab Sample	3.00%
Other Comorb.	3.00%
Diabetes	3.00%
Immuno.	3.00%
Gender	3.00%
Others(each)	less than 2.00%

Table 4 showcases the summarized results of this analysis, revealing the percentage contribution of each feature towards the model's predictive capability.

Fig. 4 clearly underscores the paramount importance of geographic and sociodemographic factors in influencing mortality rates in Mexico. As delineated through the comprehensive analysis of various features, it is evident that the collective weight of location-based attributes—encompassing Municipality, Health Unit, Birth State, and Residence State—surpasses even the influence of age, traditionally considered one of the most significant predictors of mortality risk.

This revelation not only highlights the geographical variance within the country but also brings to light the profound impact of sociodemographic elements on health outcomes.

The overarching dominance of geographical factors in determining mortality rates suggests a complex interplay between environmental, economic, and social determinants of health. In Mexico, disparities in healthcare access, differences in environmental exposure, and varying socioeconomic conditions across different regions amplify the mortality risk associated with geographical and sociodemographic characteristics.

The aggregated importance of location-related variables, serves as a compelling testament to the critical need for targeted public health strategies and interventions that are finely tuned to the unique challenges and vulnerabilities of each region. This approach is crucial for mitigating the risks associated with these factors and for paving the way toward more equitable health outcomes across the diverse landscapes of Mexico.

Age emerged as the most significant predictor of COVID-19 mortality, accounting for 21.00% of the model's predictive power. This finding underscores the heightened vulnerability of older populations to severe outcomes following COVID-19 infection.

Following closely, Municipality with an 18.00% importance, indicates the significance of geographical and possibly socio-economic factors in mortality risk. The Health Unit and Birth State features also demonstrate considerable influence, with contributions of 9.00% and 7.00%, respectively, highlighting the role of healthcare access and regional health disparities.

Further down the list, Residence State, Other Case, Lab Sample, and various comorbidities including Diabetes and conditions affecting the immune system (Immuno.), each contribute to the model's ability to predict mortality, albeit to a lesser extent. Notably, each of these features adds valuable insights into the complex interplay between patient characteristics, healthcare system factors, and comorbid conditions in determining COVID-19 outcomes.

Gender, with a 3.00% contribution, and other less impactful features (Others(each)) accounting for less than 2.00% each, further delineate the nuanced landscape of risk factors associated with COVID-19 mortality. This granularity in feature importance not only highlights the predominant role of certain demographics and health-related factors but also underscores the multifaceted nature of COVID-19's impact on various segments of the population.

5.4 Discussion

The derived feature importance from our analysis illuminates the complex and multifactorial nature of COVID-19 mortality risk. Age stands out as a critical determinant, corroborating global observations regarding the disproportionate impact of the virus on older individuals.

The significant role of geographic and healthcare accessibility factors further emphasizes the need for targeted public health interventions and resource allocation to mitigate mortality risks effectively.

The granularity achieved through our machine learning-driven exploration enables a more nuanced understanding of the interactions between various factors and COVID-19 mortality. Such insights are invaluable for guiding clinical decision-making, optimizing healthcare resource distribution, and tailoring public health strategies to address the needs of the most vulnerable populations.

Our study's findings contribute a novel perspective to the ongoing discourse on COVID-19 mortality, providing a comprehensive analysis of comorbidity factors that influence outcomes. This enhanced understanding is pivotal for informing future research, policy-making, and clinical practices aimed at reducing the mortality burden of the pandemic.

6 Conclusion and Future Work

In our investigation, it was discovered that the demographic and socioeconomic areas significantly influence mortality rates due to COVID-19. This finding diverges from common assumptions that comorbidities alone are the primary determinants of mortality outcomes in the Mexican population. Through an exhaustive analysis utilizing advanced machine learning techniques, including Snap Random Forest, XGBoost, Extra Trees, and Snap Decision Trees, our study delved into the myriad factors impacting COVID-19 mortality.

Unlike prior research that often fails to discern the paramount factors affecting mortality rates in localized populations, our approach allowed for a nuanced exploration of the interplay between various determinants.

Our comprehensive examination extended across algorithms within a defined search space, implementing experiments with varying degrees of granularity. This methodology, augmented by machine learning-driven feature enhancement, facilitated a deepened understanding of the elements most critically affecting COVID-19 mortality rates.

The findings underscore the predominance of residential location over comorbidities in determining mortality outcomes, pointing to the substantial role of socioeconomic factors in influencing survival chances. This revelation underscores the necessity for public health strategies and policy-making to prioritize socioeconomic determinants alongside medical considerations in combatting COVID-19 mortality.

The implications of our research are twofold: firstly, it contributes to a more targeted comprehension of the drivers behind COVID-19 mortality in the Mexican context; and secondly, it accentuates the vital impact of socioeconomic conditions on health outcomes. By highlighting these insights, our study provides a foundational basis for the development of informed and effective public health interventions aimed at mitigating COVID-19 mortality within socioeconomically diverse populations.

Acknowledgments

The authors wish to express their profound gratitude to the Consejo Nacional de Ciencias, Humanidades y Tecnología (CONAHCYT), acknowledging the institution's significant support, as two of the authors are esteemed affiliated researchers within this esteemed organization. It is recognized that the synergistic contributions from all involved entities have substantially enhanced the quality and validity of the present study.

References

1. **Badiola-Zabala, G., Lopez-Guede, J. M., Estevez, J., Graña, M. (2024).** Machine learning first response to COVID-19: A systematic literature review of clinical decision assistance approaches during pandemic years from 2020 to 2022. Electronics, Vol. 13, No. 6. DOI: 10.3390/electronics13061005.
2. **Balakrishnan, K. N., Yew, C. W., Chong, E. T. J., Daim, S., Mohamad, N. E., Rodrigues, K., Lee, P. C. (2023).** Timeline of SARS-CoV-2 transmission in Sabah, Malaysia: Tracking the molecular evolution. Pathogens, Vol. 12, No. 8, pp. 1047. DOI: 10.3390/pathogens12081047.
3. **He, X., Cui, X., Zhao, Z., Zhang, H., Ge, Q., Leng, Y. (2024).** A generalizable and easy-to-use COVID-19 stratification model for the next pandemic via immune-phenotyping and machine learning. Frontiers in Immunology, Vol. 15, pp. 1372539. DOI: 10.3389/fimmu.2024.1372539.
4. **Lages-dos-Santos, A., Oliveira, M., Colosimo, E. A., Pinhati, C., Galante, S. C., Martelli-Júnior, H., Simões Silva, A. C., Oliveira, E. (2024).** Comparative analysis of machine learning algorithms for predicting COVID-19 mortality in children and adolescents using a large public dataset in Brazil. Social Science Research Network. DOI: 10.2139/ssrn.4740297.
5. **Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M. N., Weller, A., Crowcroft, J. (2020).** Leveraging data science to combat COVID-19: A comprehensive review. IEEE Transactions on Artificial Intelligence, Vol. 1, No. 1, pp. 85–103. DOI: 10.1109/TAI.2020.3020521.
6. **Mendez-Astudillo, J. (2024).** The impact of comorbidities and economic inequality on COVID-19 mortality in Mexico: A machine learning approach. Frontiers in Big Data, Vol. 7. DOI: 10.3389/fdata.2024.1298029.
7. **Padilla-Rivas, G. R., Delgado-Gallegos, J. L., Garza-Treviño, G., Galan-Huerta, K. A., Buentello, Z. G., Roacho-Pérez, J. A., Santoyo-Suarez, M. G., Franco-Villareal, H., Leyva-Lopez, A., Estrada-Rodriguez, A. E., Moreno-Cuevas, J. E., Ramos-Jimenez, J., Rivas-Estrilla, A. M., Garza-Treviño, E. N., Islas, J. F. (2022).** Association between mortality and cardiovascular diseases in the vulnerable mexican population: a cross-sectional retrospective study of the COVID-19 pandemic. Front Public Health, Vol. 10, pp. 1008565. DOI: 10.3389/fpubh.2022.1008565.
8. **Quenzer, F. C., Coyne, C. J., Ferran, K., Williams, A., Lafree, A. T., Kajitani, S., Mathen, G., Villegas, V., Kajitani, K. M., Tomaszewski, C., Brodine, S. (2023).** ICU admission risk factors for latinx COVID-19 patients at a U.S.–Mexico border hospital. J Racial Ethn Health Disparities, Vol. 6, pp. 3039–3050. DOI: 10.1007/s40615-022-01478-1.
9. **Rahman, M. M., Khan, N. I., Sarker, I. H., Ahmed, M., Islam, M. N. (2023).** Leveraging machine learning to analyze sentiment from COVID-19 tweets: A global perspective. Engineering Reports, Vol. 5, No. 3, pp. e12572. DOI: 10.1002/eng2.12572.
10. **Shi, J., Chen, F., Chen, S., Ling, H. (2023).** COVID-19 over the last 3 years in China, what we've learned. Frontiers in Public Health, Vol. 11, pp. 1209343. DOI: 10.3389/fpubh.2023.1209343.
11. **Syeda, H. B., Syed, M., Sexton, K. W., Syed, S., Begum, S., Syed, F., Prior, F., Yu-Jr, F. (2021).** Role of machine learning techniques to tackle the COVID-19 crisis: Systematic review. JMIR medical informatics, Vol. 9, No. 1, pp. e23811. DOI: 10.2196/23811.
12. **Wolf, J. M., Wolf, L. M., Bello, G. L., Maccari, J. G., Nasi, L. A. (2023).** Molecular evolution of SARS-CoV-2 from december 2019 to august 2022. Journal of Medical Virology, Vol. 95, No. 1, pp. e28366. DOI: 10.1002/jmv.28366.

ISSN 2007-9737

Article received on 03/10/2023; accepted on 05/01/2024.
** Corresponding author is Mariano Vargas-Santiago.*