Análisis de factores de riesgo en pacientes con COVID-19 - México

García, Irenka López, Eva Olivares, Montserrat Mejía, Marko

Diciembre 2020

Resumen

Durante el último año y a raíz del surgimiento de una enfermedad totalmente desconocida, todos los países del mundo se han visto sujetos a cambios radicales en el día a día, cambios que van desde el uso de cubrebocas y la desinfección constante de manos, hasta el confinamiento en casa durante meses. Se ha observado que el 18.79 %, 14.77 % y 17.12 % de los pacientes que han tenido un resultado positivo en la prueba COVID-19 en México sufren de hipertensión, diabetes y obesidad, respectivamente. Esto nos da un panorama más amplio sobre los posibles factores que pueden agravar los efectos del COVID-19 en algunas personas. En este estudio se utilizan métodos de minería de datos para encontrar, en individuos mexicanos atendidos en territorio nacional, variables significativas en la posible defunción de un paciente infectado, siendo los bósques aleatorios la mejor opción con una exactitud del 91.5 %.

1. Introducción

Durante el último año y a raíz del surgimiento de una enfermedad totalmente desconocida, todos los países del mundo se han visto sujetos a cambios radicales en el día a día, cambios que van desde el uso de cubrebocas y la desinfección constante manos, hasta el confinamiento en casa durante meses. La enfermedad del *coronavirus*, causada por el virus SARS-Cov2 fue identificada por vez primera en Wuhan, China a finales del año 2019 [7] y ha permeado la vida de millones de personas, desde sus hábitos sociales hasta la economía de los países mismos; además, desgraciadamente muchas familias han padecido la consecuencia más lamentable de la enfermedad que es el fallecimiento de sus seres queridos.

En aras de entender esta nueva enfermedad, muchas instituciones, universidades, académicos, investigadores y personas en general han invertido su tiempo en aplicar técnicas que van desde la estadística descriptiva hasta la inteligencia artificial para describirla y tratar de entender su comportamiento en general y bajo condiciones muy particulares como es la preexistencia de otros padecimientos, o factores sociodemográficos como son la edad, el sexo, nacionalidad, el país de residencia, entre otros. Instituciones como la Universidad John Hopkins [6] ofrecen recursos donde constantemente se van actualizando las estadísticas del número de contagios y defunciones, entre otros indicadores, alrededor del mundo; sin embargo, también muchas investigaciones nos proponen métodos basados en deep/machine learning (véanse [1] y [4] como ejemplos).

Entre las distintas investigaciones que han surgido, Arthur Simonnet [8] y colegas realizaron un estudio en donde observaron la relación que mantenían las características clínicas y la necesidad de recibir ventilación mecánica en 124 pacientes que ingresaron a un centro francés por el virus SARS-CoV2. Al considerar el índice de masa corporal, el estudio arrojó que la obesidad es una condición con alta presencia en la población estudiada y es un factor de riesgo que agrava la enfermedad del coronavirus, lo que sugiere que sería prudente tomar otras medidas preventivas en individuos con esta condición. Por su parte, Bo Li y colegas [5] encontraron que las personas con preexistencia de enfermedades cardiovasculares pueden presentar un mayor riesgo de tener una enfermedad (por Covid) más grave; adicionalmente, concluyen que las comorbilidades tienen una gran influencia en el pronóstico del Covid-19 y que este nuevo virus puede agravar el daño al corazón.

Como las anteriores, hay muchas investigaciones que nos hablan de cómo se comporta esta enfermedad desde enfoques muy particulares o en poblaciones específicas. Aunque la enfermedad tenga un comportamiento general, las afectaciones entre poblaciones de distintos países (incluso entre regiones de un mismo país) no es igual debido a la heterogeneidad que existe en las condiciones económicas, demográficas, de salud, entre otras. Es por ello y lo antes mencionado que este estudio se enfoca en la situación en México y busca agrupar a la población mexicana en función de variables como la *entidad*, la *edad* y el *número de días que transcurren entre la aparición de síntomas y el ingreso a una unidad médica*, para posteriormente identificar las características comunes en la población

que fallece y la que sobrevive a la enfermedad; para estos efectos emplearemos técnicas de clustering y técnicas de clasificación como la regresión logística y los árboles de decisión.

2. Metodología

El conjunto de datos utilizado fue obtenido a través de la página oficial del Gobierno Federal de México [3] y consta de más de 2.6 millones de registros con fecha de corte al 19 de septiembre del año en curso. Esta base de datos nos da información que va desde el sexo, edad, entidad de nacimiento hasta el padecimiento de enfermedades como diabetes, neumonía, asma, entre otras, de pacientes que se presentaron en alguna unidad médica en el territorio nacional y que se tiene sospecha de tener COVID-19.

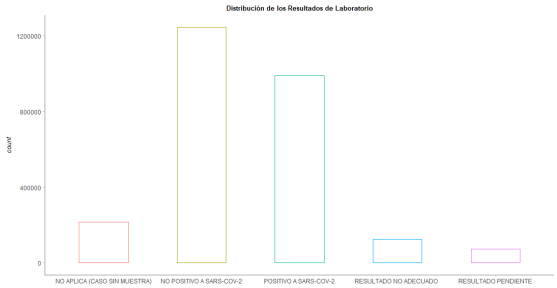

Figura 1: Resultados de Laboratorio

De acuerdo con las definiciones de *caso* de la Dirección General de Epidemiología [2], las observaciones corresponden a casos sospechosos (sin muestra), confirmados y negativos a la prueba de COVID-19. Como muestra la figura 1, la mayor parte de los pacientes tuvieron un resultado negativo; sin embargo, como se mencionó previamente, nuestra población objetivo es aquella con resultado positivo.

En total, la base nos ofrece 38 características por cada observaciones, de las cuales tomamos únicamente el sector, la entidad de la unidad médica a la que asistió, el sexo, tipo de paciente (hospitalizado o ambulatorio), edad y las variables asociadas a enfermedades (neumonía, diabetes, EPOC, asma, hipertensión, enfermedades cardiovasculares, obesidad, problemas renales crónicos, tabaquismo y otra complicación), así como la variable "Otro caso" la cual nos indica si el paciente tuvo contacto con otros infectados de coronarivus y se prescindió del resto de las variables.

Como muestran los histogramas en la figura 2 y las tablas 1 y 2, muchas de las variables presentan una opción de tipo **No especificado**; no obstante, en la mayoría de ellas esta opción no representa mucho dentro del total, por lo que se decidió omitir tales observaciones, quedándonos al final con un total de 859,218 pacientes observados.

row.names	NEUMONIA	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION
Si	16.7243257	14.7667	1.3478	2.5182	1.0204	18.787
No	83.2747641	84.9134	98.3617	97.1955	98.6772	80.912
Se Ignora	0.0009101	0.3199	0.2904	0.2864	0.3025	0.301

Tabla 1: Distribución % Enfermedades

row.names	HIPERTENSION	OTRA_COM	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO
Si	18.787	2.2760	1.8554	17.1264	1.7993	7.380	43.40
No	80.912	97.2229	97.8506	82.5837	97.9110	92.314	44.02
Se Ignora	0.301	0.5012	0.2941	0.2899	0.2896	0.306	12.58

Tabla 2: Distribución % Enfermedades (cont.)

Figura 2: Algunas características de los pacientes observados

Además, vemos que la distribución por edad toma valores de hasta 125 años, sin embargo la cola de la distribución nos muestra que son realmente pocas observaciones para edades por encima de los 100 años, por lo que se decide topar la variable hasta esta edad.

Adicionalmente, se crearon dos variables nuevas:

- Fallecimientos: Variable dicotómica que indica si el paciente falleció por COVID (1) o sobrevivió a la enfermedad (0). Esta variable se construyó a partir de la variable FECHA_DEF de la base original, donde al encontrarse un valor ausente se entiende que la persona sobrevivió a la enfermedad.
- Días con Síntomas: Variable numérica que mide el número de días que transcurren entre la fecha de inicio de síntomas y el ingreso a unidad médica.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	2.0	4.0	4.2	6.0	80.0

Tabla 3: Resumen de Días con Síntomas

Como podemos observar, en promedio transcurren dos días entre el inicio de síntomas y el ingreso a unidad médica, lo que nos habla de la velocidad de desarrollo de la enfermedad en donde en algunos pocos días se vuelve necesario buscar asistencia médica. Por otro lado, la tasa de letalidad (porcentaje de casos confirmados que falleció) es del 7 % y son, como lo muestra la figura 3 las edades por encima de los 50 años las cuales comienzan a tener una mortalidad más elevada.

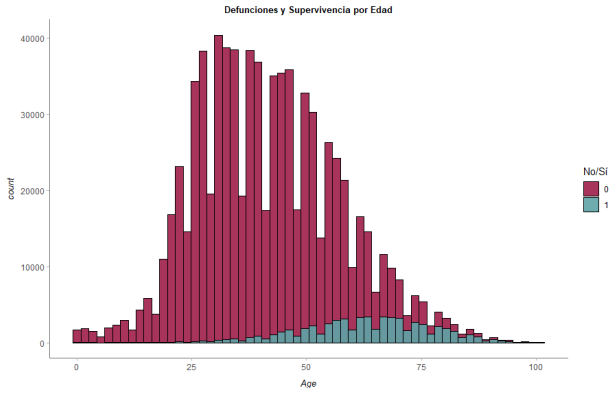

Figura 3: Defunciones y Supervivencia por Edad

Ahora bien, previo a aplicar los métodos de clasificación y dado que las variables de *Entidad*, *Edad* y *Días con Síntomas* tienen una alta dimensionalidad, se realizó un análisis de clustering jerárquico para determinar el óptimo de clusters asociados a cada variable. De este análisis y aplicando la técnica del codo, las variables se agrupan en 3, 4 y 3 grupos, respectivamente (véanse figuras 4, 5, y 6).

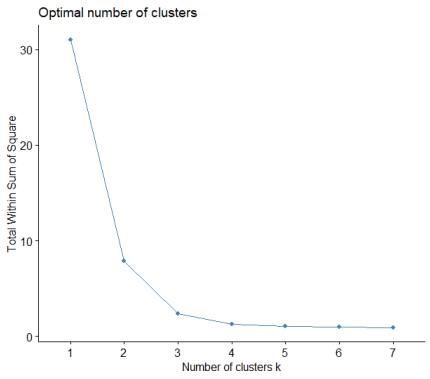

Figura 4: Número de Cluster por Entidad

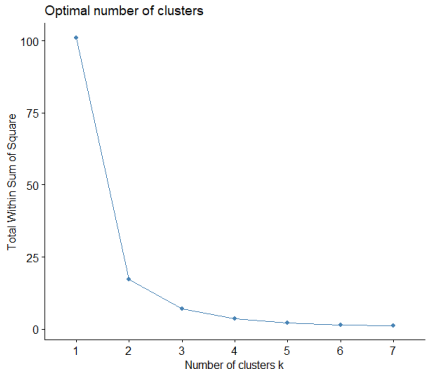

Figura 5: Número de Clusters por Edad

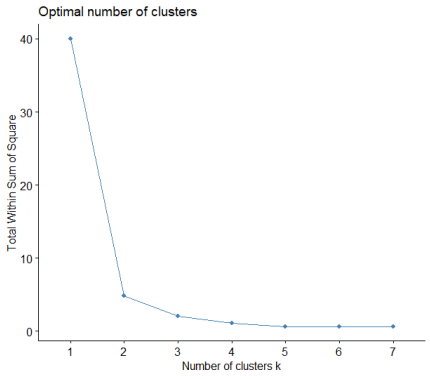

Figura 6: Número de Cluster por Días de Síntomas

Por su parte, la agrupación por entidad junta las observaciones del Estado de México y la CDMX lo cual hace total sentido en función de que ambos estados tienen un estilo de vida similar. Además, los otros dos conglomerados agrupan a entidades del sur por un lado y a entidades del norte-centro del país (véase figura 7)

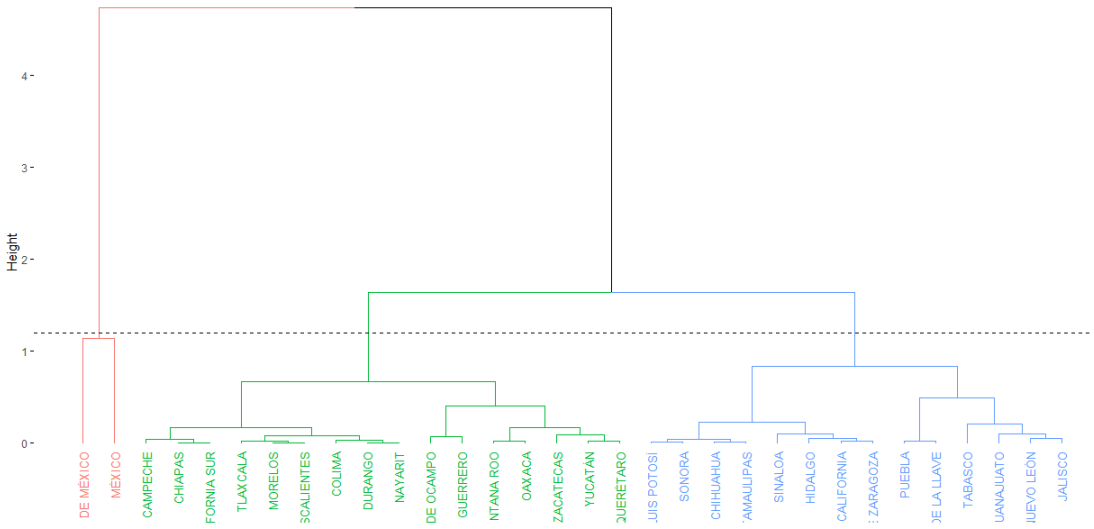

Figura 7: Dendograma por Entidad

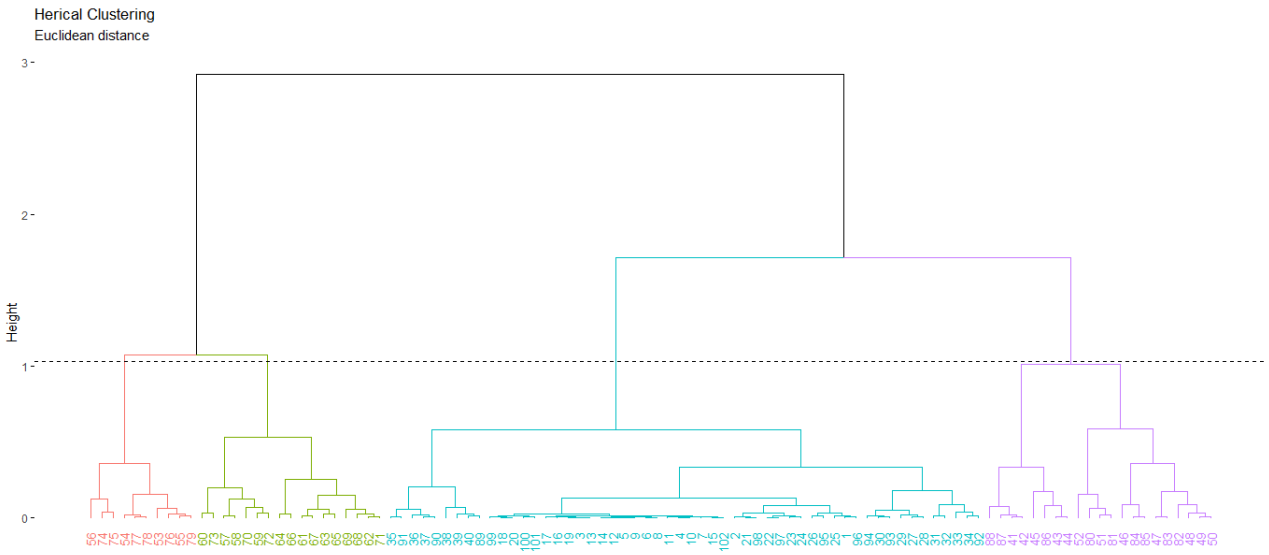

Figura 8: Dendograma por Edad

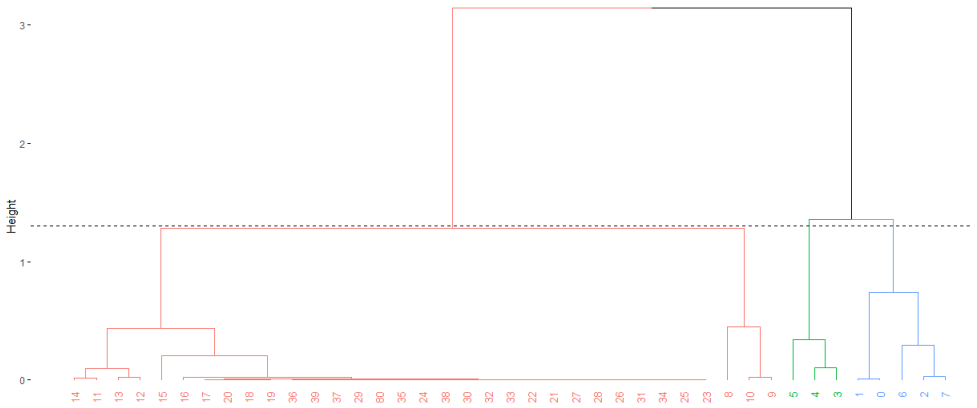

Figura 9: Dendograma por Días de Síntomas

La agrupación por edad identifica: un grupo para edad alrededor de los 40‘s, otro para edades alrededor de los 60‘s, un tercer grupo en donde se ven edades entre los 50‘s o 70‘s y un último (el más grande de ellos) en donde se encuentran los menores y los pacientes más longevos.

En lo que respecta a *Días con Síntomas*, tenemos un óptimo de 3 clusters que se identifican en el dendograma de la figura 9. Esta agrupación nos sugiere que la enfermedad tiene un grupo de días intermedios (ni pocos ni muchos) que puede resultar representativo en lo que respecta al desarrollo de la enfermedad.

Como lo muestra el histograma 3, la proporción de fallecimientos es, por mucho, inferior a la supervivencia la variable de fallecimiento estará fuertemente cargada al valor 0. Dado que clasificadores de aprendizaje automático como los bosques aleatorios son sensible a la distribución de las clases (es decir, al estar cargado a una clase en particular, la clase dominante se ve favorecida y toma mayor peso al momento de la clasificación), se realizó un balanceo sobre los datos para que los métodos le dieran un peso equilibrado tanto la ocurrencia como la no ocurrencia del evento.

2.1. Selección de Características

Adicional a lo anterior, se hizo una selección de características a través de una prueba chi-cuadrada la cual nos arroja los resultados mostrados en la figura 10.

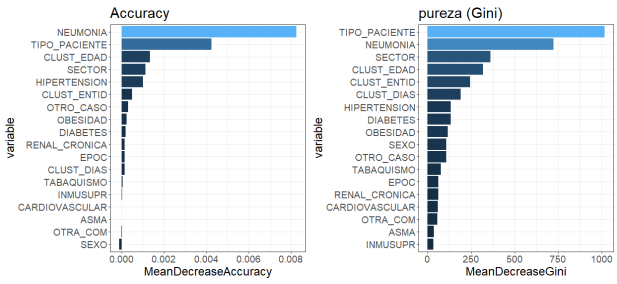

Figura 10: Accuracy e Índice de Gini

En función del gráfico anterior, seleccionamos como relevantes las siguientes 8 características:

- Neumonía
- Tipo_Paciente
- Sector
- Clust_Edad
- Clust_Entidad
- Diabetes
- Hipertensión
- Clust_Días

Para el ajuste de modelos se dividió el conjunto de datos en un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %); además, sobre cada modelo se realizó una validación cruzada con 10 grupos y 3 repeticiones [1].

2.2. Comparación de Modelos

Los modelos aplicados fueron **Árbol de decisión**, **Regresión Logística** y **Bósques aleatorios**, mismos que fueron aplicados sobre el conjunto de entrenamiento y que nos arrojan una media de precisión de:

Método de Aprendizaje	Media de Precisión
Árbol de decisión	90.1 %
Regresión Logística	89.5 %
Bósques aleatorios (1000)	91.5 %

mismos que pueden verse graficados en la siguiente figura:

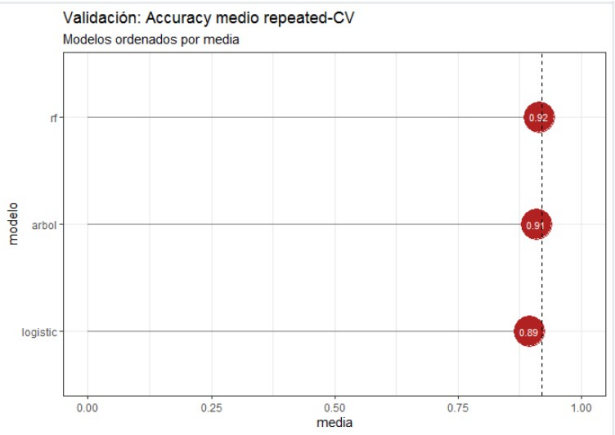

Figura 11: Accuracy Medio por Método de Aprendizaje

La figura 12 nos indica el mayor accuracy arrojado por cada método y es claro que el método de random forest alcanzó una precisión superior a la máxima alcanzada por la regresión y el árbol de decisión.

	Resample	arbol~Accuracy	logistic~Accuracy	rf~Accuracy
1	Fold01.Rep1	0.9147619	0.9019048	0.9171429
2	Fold01.Rep2	0.9085279	0.8889948	0.9128156
3	Fold01.Rep3	0.9070986	0.8837542	0.9094807
4	Fold02.Rep1	0.8976190	0.8757143	0.9028571
5	Fold02.Rep2	0.9066667	0.8871429	0.9119048
6	Fold02.Rep3	0.9123810	0.8990476	0.9142857
7	Fold03.Rep1	0.9143265	0.8986197	0.9219419
8	Fold03.Rep2	0.9086149	0.8900524	0.9143265
9	Fold03.Rep3	0.9023344	0.8847070	0.9104335
10	Fold04.Rep1	0.9080515	0.8913768	0.9161505

Figura 12: Máx.Accuracy por Método de Aprendizaje

[1] El número de grupos y repeticiones puede incrementarse si la capacidad del equipo así lo permite

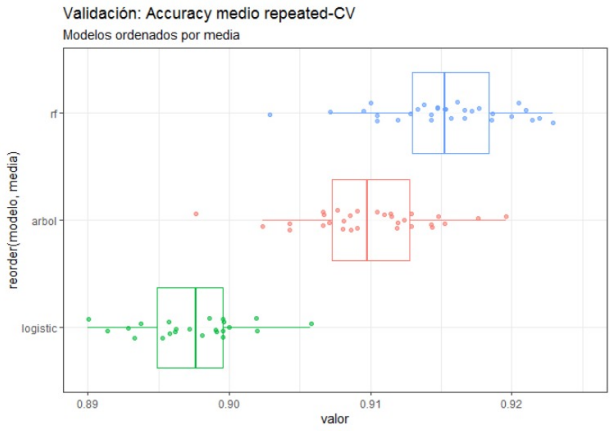

Figura 13: Diagrama de caja de la precisión de los modelos

Las gráficas de caja de la figura 12 muestran el nivel medio y la variabilidad de la precisión de cada modelo ajustado mostrando en el bosque aleatorio un comportamiento relativamente estable. Por todo lo anterior y aunque ninguno de los métodos tiene un mal ajuste, se toma el bosque aleatorio como el mejor método para clasificar la supervivencia de un paciente dadas las características seleccionadas.

2.3. Validación Bosque Aleatorio

Una vez seleccionado el bosque aleatorio como método de clasificación, se aplica el mismo análisis de 1000 árboles al conjunto de prueba.

La figura – nos muestra la evolución del accuracy de la combinación de los 1000 árboles, de donde se obtiene un accuracy medio de 0.9097 sobre el conjunto de prueba:

Conjunto de Datos	Media de Precisión
Entrenamiento	91.50 %
Prueba	90.97 %

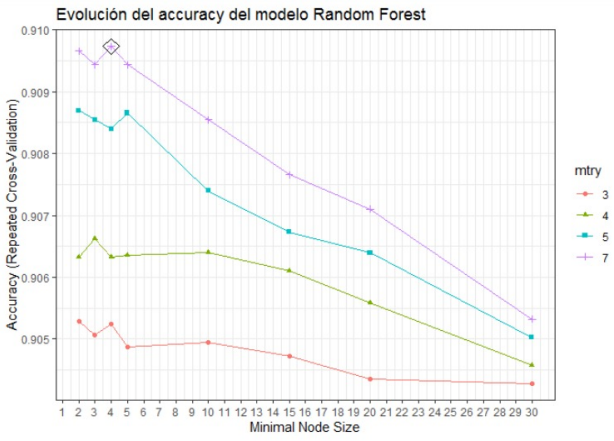

Figura 14: Accuracy Bósques Aleatorios

Si obtenemos la matriz de confusión 15 puede observarse cómo la proporción de falsos positivos y falsos negativos son realmente pequeñas. En particular, al estar hablando de la supervivencia de las personas es de especial interés lograr que la proporción de falsos negativos (2.5 % para este modelo) sea lo más pequeña posible.

Prediction \ Reference	0	1
0	43.2	2.5
1	6.6	47.8

Figura 15: Matriz de Confusión - Bósques Aleatorios

Calculando la especificidad y sensibilidad del modelo a través de:

$$Sensibilidad = \frac{VP}{VP + FN}$$

$$= \frac{47.8}{47.8 + 2.5}$$

$$= 0.9503$$

$$Especificidad = \frac{VN}{VN + FP}$$

$$= \frac{43.2}{43.2 + 6.6}$$

$$= 0.86.75$$

De donde la sensibilidad del modelo para detectar las defunciones que realmente ocurrieron es del 95 %, por lo que, en general, el modelo clasifica como defunciones las que realmente ocurrieron.

Finalmente la curva ROC de la figura 16 genera un área bajo la curva de 0.9383 lo que clasifica como un buen diagnóstico.

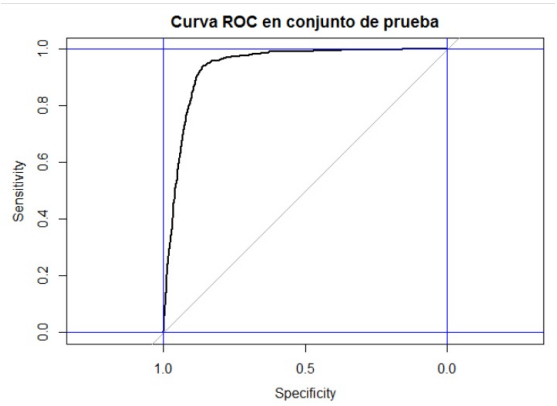

Figura 16: Curva ROC para el Bósque Aleatorio

3. Conclusiones y Discusión

Lo expuesto a lo largo de este trabajo permite arribar que es posible definir un modelo que ayude a predecir si un paciente es más propenso a fallecer o a sobrevivir dadas las características analizadas, lo cual podría ser sustento para dar mayor prioridad a pacientes más vulnerables. Es importante no perder de vista que este análisis no es exhaustivo ya que se trata de un fenómeno nuevo para el mundo y que puede ser abordado desde distintos puntos de vista.

De esta forma, son muchos los aspectos que pueden influir en la evolución de la enfermedad y que vale la pena ir agregando a los ya considerados, tales características pueden ser la ausencia de equipo médico y la calidad de servicios con la que cuenta la unidad a la que la persona acudió. Adicionalmente, este tipo de análisis puede llevarse a un caso más específico como la presencia de embarazo durante el padecimiento de esta enfermedad.

Si bien la estadística descriptiva y el seguimiento que con ella se le puede dar a la evolución de la enfermedad (que mucho sirve para entender cómo se comporta) son el primer paso para comprender un fenómeno desconocido, también es importante aplicar a ese conocimiento técnicas que permitan dar un paso más en el entendimiento del fenómenos, de tal forma que adicional al monitoreo del virus seamos capaces de reconocer variables y factores que agravan o influyen significativamente en la enfermedad; así, tendremos mayor claridad sobre qué condiciones son desfavorables para la pronta recuperación y qué medidas pueden tomarse para poblaciones que son más vulnerables.

En México aún se necesita un mayor número de gente que guste de la investigación y aplique técnicas de minería de datos para el entendimiento y comprensión del comportamiento de este (y muchos otros) fenómeno que ha revolucionado al mundo durante el último año. Así, este trabajo es un buen punto de partida para profundizar más en este objeto de estudio aplicando técnicas más avanzadas como deep learning.

Referencias

[1] Talha Burak Alakus and Ibrahim Turkoglu. Comparison of deep learning approaches to predict covid-19 infection. *Chaos, Solitons & Fractals*, 140:110120, 2020.

[2] Dirección General de Epidemiología. Lineamiento estandarizado para la vigilancia epidemiolÓgica y por laboratorio de la enfermedad respiratoria viral, 2020.

[3] Gobierno de México. Secretaría de salud, 2020.

[4] Samuel Lalmuanawma, Jamal Hussain, and Lalrinfela Chhakchhuak. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. *Chaos, Solitons & Fractals*, page 110059, 2020.

[5] Bo Li, Jing Yang, Faming Zhao, Lili Zhi, Xiqian Wang, Lin Liu, Zhaohui Bi, and Yunhe Zhao. Prevalence and impact of cardiovascular metabolic diseases on covid-19 in china. *Clinical Research in Cardiology*, 109(5):531–538, 2020.

[6] John Hopkins University of Medicine. Coronavirus resource center, 2020.

[7] World Health Organization, World Health Organization, et al. Report of the who-china joint mission on coronavirus disease 2019 (covid-19), 2020.

[8] Arthur Simonnet, Mikael Chetboun, Julien Poissy, Violeta Raverdy, Jerome Noulette, Alain Duhamel, Julien Labreuche, Daniel Mathieu, Francois Pattou, Merce Jourdain, et al. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (sars-cov-2) requiring invasive mechanical ventilation. *Obesity*, 2020.