El cáncer hepático antes y después de la pandemia de COVID 19. Forward Selection en la predicción de supervivencia utilizando algoritmos supervisados.

Alen Francisco Luévano Lara*
Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Carretera Zacatecas-Guadalajara km 6, Ejido La Escondida, C.P. 98160, Zacatecas, Zac., México.
* *alen_lara@uaz.edu.mx*

Abstract— **Since the start of the COVID-19 pandemic there have been over 643 million confirmed infections and 6.6 million deaths reported worldwide and the situation wouldn't be worst for liver cancer in Mexico, was summarized at an annual rate in 2015 of 5.2 with 6,333 deaths. The prediction of diseases has gained importance in 20th century, however, the prediction of survival from those diseases has been less studied. Liver cancer were studied in several studies related to COVID-19, it has been sought to statistically analyze this disease and the behavior it has obtained in the year of the pandemic 2021. In the present work it has been found that there is a decrease in death from cancer in the pandemic period, in addition to an analysis of measures of central tendency and dispersion. Finally, a prediction model was designed using Forward Selection as a method to reduce the characteristics to some particular data, the models were designed with Random Forest and Linear Regression with an area under the ROC curve of 98.6% and 97.2% respectively.**

Palabras clave— **Cáncer hepático, Carcinoma Hepatocelular, Random Forest, Regresión Logística Forward Selection, Pandemia, Covid 19, HCC.**

I. Introducción

La enfermedad de coronavirus 2019 (COVID-19) es una enfermedad global que se volvió pandémica en el año 2020 [1]. Según la Organización Mundial de la Salud, más de 643 millones de personas han sido diagnosticadas de esta enfermedad con un estimado de 6,600,000 muertes [2]. El COVID-19 es una enfermedad de mayor interés en los últimos años, esto por ser una enfermedad crónica que ha puesto en crisis a muchos países debido a la facilidad de contagio. Hasta el momento se han encontrado varios síntomas o patologías causadas por las secuelas de la enfermedad [3]; en el presente artículo se relaciona esta enfermedad con un tipo de cáncer hepático conocido como Carcinoma Hepatocelular. La situación del cáncer hepático en México se resume a una tasa anual en el 2015 de 5.2 con 6,333 muertes. Es una enfermedad también de carácter crónico que se ha estudiado en diversos artículos, donde los pacientes infectados por COVID-19 también presentan afecciones en el hígado. En el artículo "Impacts of COVID-19 on Liver Cancers: During and after the Pandemic" se menciona que entre el 15% y 54% de los pacientes positivos a COVID-19 tienen una lesión hepática, además se señala el aumento de riesgo a sufrir COVID-19 en los pacientes con cáncer hepático [1,4,5,6]. En "Assessing the impact of COVID-19 on liver cancer management (CERO-19)" se le dio un enfoque diferente al relacionar las muertes o tratamientos de cáncer de hígado con el protocolo de administración y el manejo situacional en la pandemia debido a la saturación de hospitales [7].

La predicción se realizará con los datos recopilados prospectivamente en todos los pacientes remitidos al equipo multidisciplinario hepatopancreatobiliar (HPB MDT) de Newcastle-upon-Tyne NHS Foundation Trust (NUTH) en los primeros 12 meses de la pandemia (marzo de 2020 a febrero de 2021), en comparación con un estudio observacional retrospectivo en los 12 meses inmediatamente anteriores (marzo 2019-febrero 2020) [8]. Antes de la predicción, se realizó un análisis estadístico para obtener el contexto en el que se está trabajando, mostrarlo gráficamente con ayuda de tablas de frecuencia y obteniendo medidas de tendencia central y de dispersión para las variables cuantitativas.

Para escoger las características más importantes de la base de datos se pueden utilizar diferentes métodos, en este caso se utiliza forward selection, un algoritmo que consiste básicamente en encontrar las características que mejor puntaje de exactitud nos den en un modelo de regresión lineal [9]. En este trabajo se presenta un conjunto de modelos de algoritmos de aprendizaje supervisado: bosques aleatorios y regresión logística, se utilizan todas las características para la predicción; por último, se propone realizar una selección de características para mejorar la exactitud de dichos modelos, además de disminuir el costo computacional de dicha predicción. Estos algoritmos se pueden implementar como herramientas de diagnóstico asistido, los cuales consisten en procedimientos que ayudan a los profesionales en la interpretación de distintos datos. Los sistemas de diagnóstico asistido utilizan algoritmos para reconocer patrones en datos de pacientes, de esta manera se proporciona un soporte a los especialistas al momento de realizar un diagnóstico [10].

II. Metodología

Se han desarrollado técnicas de clasificación automática con el fin de modelar el riesgo de padecer cierta enfermedad, estas técnicas se han estado utilizando por los investigadores de los últimos años, teniendo una gran aceptación [11]. Ha llegado el punto en la que estás técnicas por si solas no son lo suficientemente exactas, por lo que se ha optado en utilizar

forward selection, lo que nos podría permitir mejorar la exactitud, sensibilidad y especificidad [12]. Para el desarrollo del trabajo se llevó a cabo la metodología mostrada en la Fig. (1).

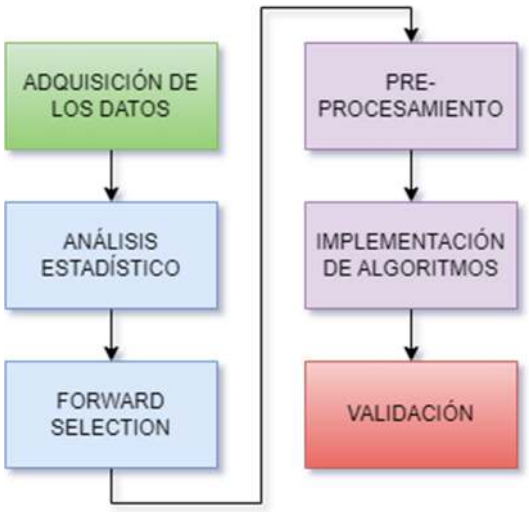

Figura 1. Diagrama general de la metodología.

Adquisición de los datos: La base de datos contiene la información requerida para evaluar el impacto del COVID-19 en relación de las personas diagnosticadas con cáncer hepático entre marzo de 2020 y febrero de 2021, siendo 450 pacientes los evaluados en el año más significativo de la pandemia por COVID-19. Se pretende comparar con el año anterior (periodo prepandémico entre marzo de 2019 y febrero de 2020) e inferir un posible cambio en alguna de las características, el incremento o decremento de casos de cáncer y analizar la posibilidad de supervivencia en pacientes relacionados con el cáncer y el COVID-19.

La información presente en la base de datos ha sido recopilada en Newcastle-upon-Tyne NHS Foundation Trust (NUTH) en el norte de Inglaterra [8]. En la base de datos se tiene registrada la etiología de la enfermedad subyacente, la etapa en la que se diagnostica la enfermedad, la infección por COVID-19 y el tratamiento que se le da a cada paciente. La fecha de referencia fue la fecha de la primera discusión del equipo multidisciplinario hepatopancreatobiliar, esta fecha se tomó como si fuera la fecha del diagnóstico [8]. La supervivencia (survival) se registró hasta el 29 de noviembre de 2021, por lo que la información se toma como parte del periodo pandémico. Se tiene registro de cánceres asintomáticos en la parte incidental, los sintomáticos y los que se detectaron en vigilancia.

Las diferencias entre variables de tipo cuantitativa continua se evaluaron previamente con pruebas tipo t para conjuntos de datos paramétricos y pruebas U de Mann-Whitney para los no paramétricos; en cuanto a las variables categóricas, se realizó en el estudio la prueba de frecuencias conocida por el nombre de chi cuadrada de Pearson para evaluar la distribución de variables en etapa prepandémica y pandémica.

A continuación, la Tabla (1) contendrá las características a estudiar y su significado.

Variable	Descripción
Cancer	Yes-No [Y/N]
Year	Prepandemic/Pandemic
Month	Month of the year 1-12
Bleed	Spontaneous tumour haemorrhage [Y/N]
Mode Presentation	Surveillance, Incidental, or Symptomatic
Age	Age of the patitent
Gender	Male or Female [M/F]
Etiology	Manner of causation of a disease or condition
Cirrhosis	Yes-No [Y/N]
Size	Tumour diameter in mm
HCC TNM Stage	Hepatocellular carcinoma Tumour node metastasis Stage
HCC BCLC Stage	Hepatocellular carcinoma Barcelona Clinic for Liver Cancer Stage
Treatment grps	First-line treatment received
Survival from MDM	Survival from Multidisciplinary meeting
Alive Dead	"Alive", "Dead"
Surveillance programme	Patient in a formal surveillance programme ("Y", "N")
PS	Performance status [0, 1, 2, 3, 4]
Prev known cirrhosis	Yes-No [Y, "N"]

Tabla 1. Descripción de características de la base de datos recopilada en NewCastle-upon-Tyne NHS Foundation Trust (NUTH) [8].

Análisis estadístico: Las medidas de tendencia central nos resumen las variables numéricas del conjunto de datos con los valores en una ubicación específica, éstas fueron calculadas con la función summary de R studio. El principal propósito de estas medidas es darnos una idea de qué tan típico o común es un valor para una variable.

En la Tabla (2), se pueden mostrar las medidas de tendencia central, como el valor mínimo, el valor que indica el primer cuartil o 25%, la mediana que indica el 50%, el 3er cuartil donde se encuentran el 75% de los datos y el valor máximo, así como el promedio de los datos [13]. La moda se calculó con la función "mlv" de la librería "multimode". Con los valores de dicha tabla, podemos inferir que el mes más

Característica	Min	1er Cuartil	Mediana	Media	3er cuartil	Max	Moda
Month	1	4	7	6.758	10	12	12
Age	27	65	72	70.37	78	96	73
Size	10	24	40	53.35	70.5	220	20
Survival from MDM	-0.03	4.032	10.785	12.697	21.282	32.77	10.50, 29.77

Tabla 2. Medidas de tendencia central de las variables en la base de datos.

frecuente en la base de datos es diciembre; el 50% de las personas tienen la edad entre 65 y 78 años, solo el 25% de la gente que se estudió está entre los 27 a 65 años, y el otro 25% se encuentra entre los 78 y 96 años; el tamaño (Size) de los tumores puede variar desde los 10 mm hasta los 220 mm, el tamaño que más se repite es de 20 mm, en específico 11 casos de 127 en etapa I de metástasis en ganglios tumorales de carcinoma hepatocelular (CHC), y la mediana indica que hay un sesgo a la derecha; la característica "Survival from MDM" es una medida que representa la supervivencia del paciente, obtenida a partir de la historia clínica electrónica.

En la Tabla (3) se observan las medidas de dispersión, las cuales nos indica qué tanto varían los datos. La varianza de los datos está definida como el promedio del cuadrado de las desviaciones respecto a la media, la desviación estándar es el promedio de la desviación entre los datos respecto a la media, y el coeficiente nos indica la relación entre la desviación estándar y el promedio [14].

Variable	Varianza	Desviación estándar	Coeficiente de variación
Month	11.89	3.45	51%
Age	111.78	10.57	15%
Size	1622.74	40.28	75%
Survival from MDM	94.15	9.7	76%

Tabla 3. Medidas de dispersión de las variables en la base de datos

El coeficiente de variación de "Size" y "survival" son bastante altos debido a que existe mucha diferencia entre los datos, por lo que se tienen que escalar para no tener problemas con el modelo de aprendizaje automático, la variable "Month" se representa como una variable numérica de 1 a 12 representando el mes del año, por lo que se puede considerar como una variable categórica.

La distribución del cáncer en la base de datos se puede observar en la figura (2), con una frecuencia de 140 diagnosticados sin cáncer y 310 diagnosticados con cáncer, teniendo una proporción de 31% y 69% respectivamente.

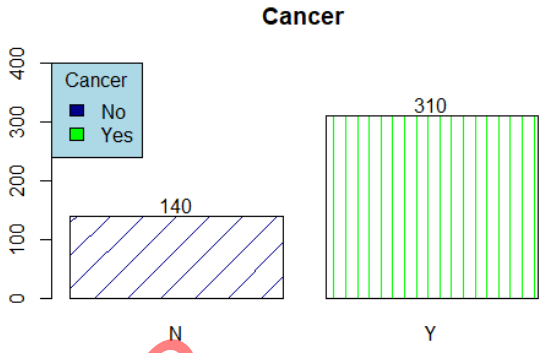

Figura 2. Gráfica de distribución de pacientes con/sin cáncer en la base de datos.

En la figura (3), el periodo prepandémico sobresalió a comparación del periodo pandémico, teniendo 266 pacientes el año prepandémico y 184 pacientes en el año pandémico diagnosticados con cáncer hepático.

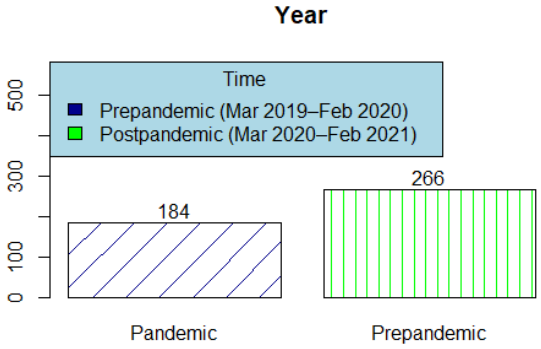

Figura 3. Gráfica de distribución de pacientes en periodo de pandemia/prepandemia.

Se tiene una proporción de 59% en el primer año y el 41% para el segundo, lo cual es contraproducente al ver en la figura (4) el comportamiento de los nuevos casos diagnosticados con CHC en años anteriores.

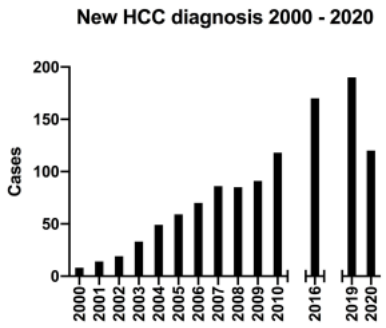

Figura 4. Casos de diagnóstico de CHC de los años 2000-2020 [8].

Esto se puede deber al verdadero decremento de CHC en los últimos años debido a factores externos, sin embargo, existe la posibilidad de que en las medidas sanitarias de prohibición de contacto físico, la idea de ir a un hospital a diagnosticarse no era una opción, ya que la contingencia creada por la pandemia de COVID-19 indicaba quedarse en la casa para no contagiarse, lo que resultó contraproducente para los pacientes diagnosticados con CHC. En el principio de la primer y segunda ola se registró un descenso en el número de casos de CHC como se puede observar en la figura (5).

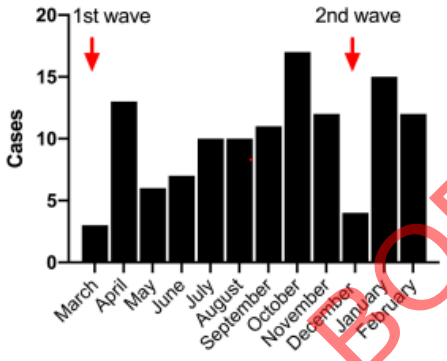

Figura 5. Número de casos diagnosticados con CHC por mes en el periodo pandémico [8].

La mortalidad con relación al cáncer en el periodo pandémico y prepandémico es un factor a considerar, ya que pareciese que los casos de muerte han disminuido en el año 2021, como se muestra en la Tabla (4).

		Pandemic		
		Alive	Dead	Total
cancer	N	18	46	64
	Y	66	54	120
	Total	84	100	184

		Prepandemic		
		Alive	Dead	Total
cancer	N	21	55	76
	Y	81	109	190
	Total	102	164	266

Tabla 4. Frecuencias en relación al cáncer, la muerte y el periodo de pandemia/prepandemia.

Para calcular si la probabilidad de mortalidad disminuyó en el año pandémico, se evaluaron las siguientes probabilidades con la fórmula de probabilidad condicional [15]:

- ✓ La probabilidad de que el paciente tenga cáncer dado que se haya diagnosticado en periodo pandémico:

$$P(c|Pa) = \frac{P(C \cap Pa)}{P(Pa)} = \frac{\frac{120}{450}}{\frac{184}{450}} = \frac{120}{184} = 65.2\% \quad (1)$$

- ✓ La probabilidad de que el paciente se haya diagnosticado en periodo prepandémico:

$$P(c|Pre) = \frac{P(C \cap Pre)}{P(Pre)} = \frac{\frac{190}{450}}{\frac{266}{450}} = \frac{190}{266} = 71.4\% \quad (2)$$

- ✓ La probabilidad de que el paciente haya fallecido dado lo anterior:

$$P(M|c|Pa) = \frac{P(M \cap c|Pa)}{P(c|Pa)} = \frac{\frac{54}{184}}{0.652} = 45\% \quad (3)$$

$$P(M|c|Pre) = \frac{P(M \cap c|Pre)}{P(c|Pre)} = \frac{\frac{109}{266}}{0.714} = 57.4\% \quad (4)$$

Se obtiene por lo tanto que la probabilidad de que el paciente haya fallecido en el año de pandemia donde se le diagnosticó cáncer es menor a la probabilidad de que haya muerto en prepandemia, lo que indica una disminución de los casos de muerte en la pandemia, donde: c es cáncer, M es muerte, Pre es prepandémico y Pa es pandémico.

Forward selection: Se utiliza este método con el fin de seleccionar las características a estudiar. Consiste en agregar variables al modelo una a la vez basándose en el nivel de significancia alta con respecto a la variable dependiente. Este proceso continúa con las siguientes variables hasta que se haya minimizado el valor de p-value y aumentado el valor de R^2, ya que se basa en el algoritmo de regresión lineal [16]. La base de datos tiene alrededor de 140 datos faltantes, por lo que se eliminarán para poder utilizar todas las características disponibles y seleccionar las características con más relevancia. Al momento de eliminar los datos faltantes, se observa que en la columna de cáncer no se encuentra ningún paciente que haya sido diagnosticado sin cáncer, por lo que se infiere que todos los pacientes fueron diagnosticados con cáncer y se elimina la columna. Nuestra variable dependiente a utilizar es la columna de Alive_Dead, y las demás características se relacionarán con la misma. Como resultado se obtuvieron 8 variables independientes que nos ayudaran a la predicción de supervivencia más la variable dependiente, un total de 9 variables:

Variables
Alive_Dead
Survival_fromMDM
Year
Month
Treatment_grps
PS
Size
HCC_TNM_Stage y Gender

Tabla 5. Variables seleccionadas con el método de Forward Selection para la creación de los modelos.

Pre-procesamiento: El conjunto de datos es escalado debido a la variación que existe entre sus datos, después es dividido en dos subconjuntos, el primero pertenece al entrenamiento con un 75% de datos del conjunto total y el segundo corresponde al de prueba y contiene el 25% restante. Posteriormente, se normalizan los subconjuntos de datos por medio del método del puntaje Z, que en lenguaje de programación R está dado por la función scale(), el cual consiste en transformar los datos a una distribución con una media 0 y una desviación estándar de 1, este método tiene el propósito de definir una misma escala numérica para todos los datos.

Implementación de algoritmos: En esta investigación se implementaron tres algoritmos de aprendizaje automático los cuales se describen a continuación:

- ✓ Regresión logística: Se centra en encontrar las relaciones entre la variable dependiente y la independiente utilizando la función logística para las probabilidades [19].
- ✓ Bosques aleatorios: Se trata de un algoritmo utilizado constantemente en el área médica, consiste en crear múltiples árboles de decisión para así generar el llamado bosque aleatorio [20].

Validación: El algoritmo debe ser calificado en cuanto al desempeño se refiere, por lo que se usaron métricas de significancia estadística para evaluar la forma en la que predice la probabilidad de supervivencia del paciente; estos métodos son:

El área bajo la curva, que representa la probabilidad de que una muestra aleatoria positiva se clasifique correctamente, utilizada en complemento con la curva ROC (Receiver Operating Characteristic Curve), la cual nos permite visualizar el rendimiento de un algoritmo. Para calcular el área de la curva ROC y graficar esta curva, es necesario definir dos valores de su trayectoria: sensibilidad y especificidad [21].

La sensibilidad o tasa de verdaderos positivos se refiere a la proporción de sujetos con un estado positivo que se clasifican correctamente y se calcula usando la Ecuación (1) donde VP es verdadero positivo y FN es falso negativo [21].

$$Sensibilidad = \frac{VP}{VP+FN} \quad (5)$$

Los sujetos con una condición negativa a los cuales se clasificaron correctamente corresponden a la especificidad, conocida también como tasa de verdaderos negativos. Es calculada con la Ecuación (2), donde VN son los verdaderos negativos y FP son los falsos positivos [21].

$$Especificidad = \frac{VN}{VN+FP} \quad (6)$$

La medida de exactitud calcula el rendimiento promedio del algoritmo como se muestra en la Ecuación (4), el propósito de esta métrica es calcular el porcentaje de muestras clasificadas correctamente [22].

$$Exactitud = \frac{VP+VN}{VP+VN+FP+FN} \quad (7)$$

La curva ROC se grafica tomando en cuenta la unión de distintos puntos de corte en la que la especificidad y la sensibilidad se va iterando, donde el eje Y representa a la sensibilidad y el eje X a (1-especificidad) de cada uno de los puntos de corte [20].

III. RESULTADOS

En la Tabla (6) se muestra la comparativa de los resultados en los modelos de bosques aleatorios. En primera instancia, se tiene el modelo de las 17 variables originales del modelo, teniendo en cuenta que las variables con menos del 50% de la información fueron previamente eliminadas para el estudio, así como la variable "Cancer" debido a que todas las personas en cuestión estaban diagnosticadas con cáncer. En cuanto al modelo de 9 variables, las características utilizadas se seleccionaron con el método de Forward Selection y están descritas en el apartado II. Metodología, Tabla 5. Otro aspecto a considerar, es que, debido a los datos faltantes, se tienen en total 294 instancias en lugar de los 450 pacientes en la base de datos original.

Métricas	**17 variables**	**9 variables**
Sensibilidad	72.73%	97.30%
Especificidad	70.00%	100%
Exactitud	71.23%	98.63%
Área bajo la curva ROC	71.20%	98.60%

Tabla 6. Métricas en el modelo de Bosques Aleatorios.

Los resultados muestran que el conjunto de datos que cuenta con 17 variables obtuvo un porcentaje de 71.20% de área bajo la curva; en contraste tenemos al modelo con 9 variables, el cual obtuvo un área de 98.6%. Esto nos indica

que la capacidad de clasificar del modelo de 9 variables es mejor al de 17, lo que significa que con una cantidad menor de variables se logran obtener resultados mayormente significativos al del conjunto de datos completo, permitiendo así, disminuir los costos computacionales y de tiempo al momento de entrenar los algoritmos. Se puede afirmar que este algoritmo mejoró un aproximado del 28% en la predicción de supervivencia al cáncer hepático durante el periodo de pandemia y prepandemia.

Después se implementó regresión logística a ambos modelos en la Tabla (7), para así obtener una comparativa del desempeño de los 2 algoritmos.

Métricas	**17 variables**	**9 variables**
Sensibilidad	75.00%	100.00%
Especificidad	75.68%	94.87%
Exactitud	69.44%	97.26%
Área bajo la curva ROC	75.3%	97.20%

Tabla 7. Métricas en el modelo de Regresión Logística.

Se puede observar claramente que las métricas mejoraron en el modelo con 9 variables, la especificidad tiene un 100%, por lo que los VN se clasificarán de maneara idónea, en cuanto a la sensibilidad se tiene un 97.3%, por lo que no se queda atrás con la predicción de los VP. Una vez comparados los modelos de 17 y 9 variables, se compararán los modelos de 9 variables de Regresión Logística y Bosque Aleatorio en la Tabla (8).

Métricas	**Regresión Logística**	**Bosque Aleatorio**
Sensibilidad	100.00%	97.30%
Especificidad	94.87%	100%
Exactitud	97.26%	98.63%
Área bajo la curva ROC	97.20%	98.60%

Tabla 8. Métricas en los modelos con 9 variables: Comparación de Regresión Logística y bosque aleatorio.

Se lograron resultados estadísticamente significativos en ambos algoritmos con porcentajes superiores al 90% en todas las métricas, además, se observa que el modelo de bosques aleatorios logró un mejor desempeño en la mayoría de las métricas, excepto en la sensibilidad, teniendo un 97.30% frente a un 100%. En cuanto a la medida de especificidad se tiene un valor del 100% en el modelo de bosques aleatorios frente al 94.87% del modelo de regresión logística. La matriz de confusión se ambos modelos es la siguiente:

Resultado de Prueba	**Regresión Logística**	**Bosques Aleatorios**
Verdaderos positivos	34	36
Falsos positivos	2	0
Verdaderos Negativos	37	36
Falsos Negativos	0	1

Tabla 9. Matriz de confusión de los modelos con 9 características.

Gráficamente, se observa el comportamiento de las curvas características formadas por el valor de la sensibilidad y 1 – especificidad en las Figuras (6,7,8 y 9). Las gráficas con 17 características tienen un comportamiento más linear de las que tienen 9 características.

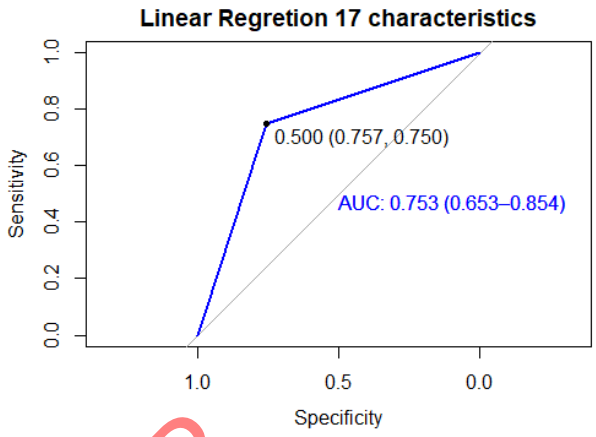

Figura 6. Curva ROC del modelo de regresión linear con 17 características.

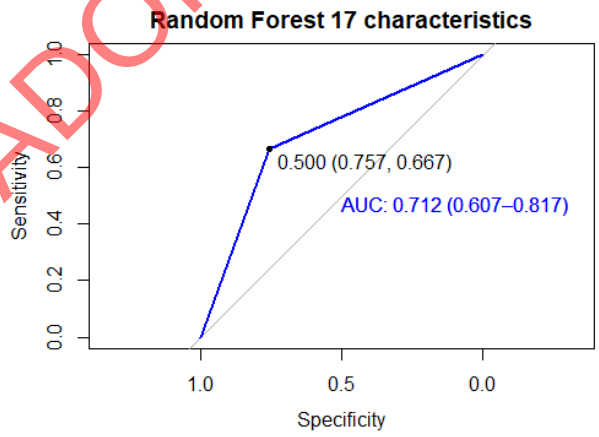

Figura 7. Curva ROC del modelo de bosques aleatorios con 17 características.

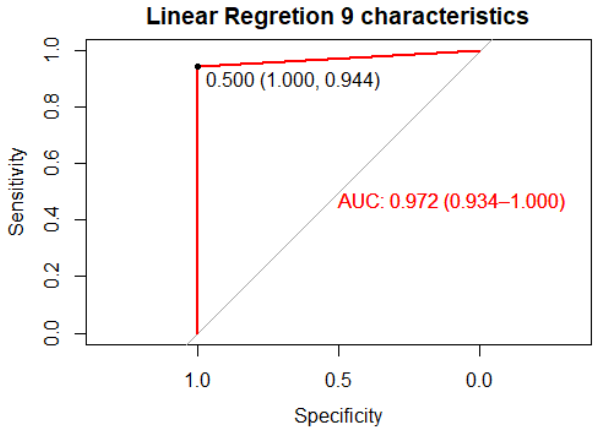

Figura 8. Curva ROC del modelo de regresión linear con 9 características.

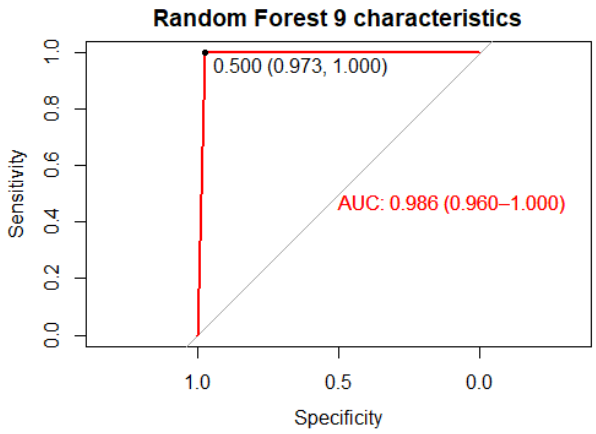

Figura 9.Curva ROC del modelo de bosques aleatorios con 9 características.

IV. Discusión

En la parte estadística se pudieron analizar sucesos interesantes, como las medidas de tendencia central y las medidas de dispersión, métricas que nos ofrecieron un panorama general al estudio. Las medidas de tendencia central nos indican que las personas se diagnosticaban más en diciembre que en cualquier otro mes, además de que se produjo un descenso en los meses donde en los que hubo más contagios por COVID-19 eran más propensos a contraer enfermedades,

La base de datos utilizada para llevar a cabo el presente trabajo fue dividida en dos subconjuntos, uno para entrenamiento que corresponde al 75% de los datos y el 25% restante para el subconjunto de prueba. Los resultados apuntan a que el método de selección de características por Forward Selection fue un acierto, debido a que se logró reducir el número de características en un aproximado del 48% con respecto al conjunto de datos original, afectando de manera significativa el desempeño del modelo obtenido, lo cual indica que al momento de entrenar o implementar algún modelo con este conjunto de datos, los costos computacionales serán menores en contraste a utilizar el conjunto de datos original.

V. CONCLUSIONES

La probabilidad de que un paciente padezca de cáncer hepático dado que se haya diagnosticado en periodo de pandemia es menor a la probabilidad de que se haya diagnosticado en prepandemia. Teniendo esto, es más probable que se el paciente haya fallecido debiod al cáncer en periodo de prepandemia que en el año de pandemia, esto podría ser por el periodo de contingencia en las casas para evitar el contagio, sobre todo en el área de hospitales que fue recurrida por gente enferma de COVID-19. Los modelos de predicción de la supervivencia al COVID-19 en relación con el tipo de cáncer hepático conocido como Carcinoma Hepatocelular han tenido métricas estadísticamente significativas, el método de Forward Selection ha sido clave para aumentar estas métricas y así obtener modelos de predicción más exactos, con mejor sensibilidad y especificidad, así como un mejor desempeño general para predecir. El modelo de Random Forest tiene el mejor desempeño general con un área bajo la curva de 0.986, una sensibilidad del 97.3%, exactitud del 98.63% y una especificidad del 100%,.

VI REFERENCIAS

1. S. L. Chan y M. Kudo, “Impacts of COVID-19 on Liver Cancers: During and after the Pandemic”, Liver Cancer, vol. 9, n.° 5, págs. 491-502, 2020. doi: 10.1159/000510765.
2. World Health Organization. WHO Coronavirus (COVID-19) Dashboard. 2022. url: https://covid19.who.int/
3. F. P. Peramo-Álvarez, M. Á. López-Zúñiga y M. Á. López-Ruz, “Secuelas médicas de la COVID-19”, Medicina Clínica, vol. 157, n.° 8, págs. 388-394, 2021. doi: https ://doi.org/10.1016/j.medcli.2021.04.023.
4. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al; China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020 Apr;382(18):1708–20.
5. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARSCoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med. 2020 May;8(5):475–81.
6. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020 Feb;395(10223):507– 13.
7. S. Muñoz-Martínez, V. Sapena, A. Forner et al., “Assessing the impact of COVID-19 on liver cancer management (CERO-19”), JHEP reports, vol. 3, n.° 3, pág. 100 260, 2021. doi: https://doi.org/10.1016/j.jhepr.2021.100260.
8. D. Geh, R. Watson, G. Sen et al., “COVID-19 and liver cancer: lost patients and larger tumours”, BMJ open gastroenterology, vol. 9, n.° 1, e000794, 2022. doi: http://dx.doi.org/10.1136/bmjgast-2021-000794.
9. F. G. Blanchet, P. Legendre y D. Borcard, “Forward selection of explanatory variables”, Ecology, vol. 89, n.°9, págs. 2623-2632, 2008. doi:https://doi.org/10.1890/07-0986.1.
10. R. L. Azpitarte, D. Cortés y D. R. P. Palacios, “Aportaciones al diagnóstico de cáncer asistido por ordenador”, Tesis doct., Tese de Doutorado. Universidad Politécnica de Valencia, 2006.
11. V. B. Kolachalama y P. S. Garg, “Machine learning and medical education”, NPJ digital medicine, vol. 1, n.° 1, págs. 1-3, 2018. doi: https://doi.org/10.1001/jama.2017.18391.
12. L. E. C. Bravo, H. J. F. López y E. R. Trujilllo, “Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble”, Revista Boletın Redipe, vol. 10, n.° 13, págs. 171-190, 2021. doi: https://doi.org/10.36260/rbr.v10i13.1737.
13. A. R. León Pirela y C. E. Pérez, “Análisis estadístico en investigaciones positivistas: medidas de tendencia central”, 2019, issn: 1856-1594.
14. F. Quevedo, “Medidas de tendencia central y dispersión”, Medwave, vol. 11, n.° 03, 2011. doi: 10.5867/medwave.2011.03.4934.
15. J. M. Contreras, C. Díaz, C. Batanero y G. Cañadas, “Definiciones de la probabilidad y probabilidad condicional por futuros profesores”, 2013.
16. J. M. Sutter y J. H. Kalivas, “Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection”, Microchemical journal, vol. 47, n.° 1-2, págs. 60-66, 1993. doi: https://doi.org/10.1006/mchj.1993.1012.

17. Hilbe, J.M. “Logistic Regression Models”, CRC Press: Boca Raton, FL, USA, 2009.
18. Speiser, J.L., Miller, Michael E., Tooze, J., Ip, E. “A comparison of random forest variable selection methods for classification prediction modeling”. Expert Syst. Appl., n.°134, págs.93-101, 2019
19. Hilbe, J.M. Logistic Regression Models; CRC Press: Boca Raton, FL, USA, 2009.
20. Speiser, J.L., Miller, Michael E., Tooze, J., Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst. Appl. 2019, 134, pp.93-101
21. J. Cerda y L. Cifuentes, “Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos”, Revista chilena de infectología, vol. 29, n° 2, págs. 138-141, 2012, doi: 10.4067/S0716-10182012000200003.
22. Alcalá-Rmz, Vanessa, et al. "Identification of People with Diabetes Treatment through Lipids Profile Using Machine Learning Algorithms." Healthcare. Vol. 9. No. 4. Multidisciplinary Digital Publishing Institute, Mexico City, Mexico April, 2021, Accessed on: July, 10, 2021, doi: 10.3390/healthcare9040422.
23. Corso, Cynthia Lorena, “Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba” Universidad Tecnológica Nacional, Facultad Regional Córdoba, 2009.