UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Clasificación de la mortalidad por COVID-19 de adultos mayores en la Ciudad de México: Un enfoque de Inteligencia Computacional

ARTÍCULO ACADÉMICO

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

Sinuhe Mazuti Osorio Rivero

ASESOR DE ARTÍCULO ACADÉMICO

Dr. Guillermo Gilberto Molero Castillo

Ciudad Universitaria, Cd. Mx., 2023

Agradecimientos

A mi hermana Erandeny, por hacer de mi infancia una etapa maravillosa de mi vida. Por enseñarme, con el ejemplo, los estándares más altos de excelencia académica y profesional.

A mi abuelita Ana María, por su cariño y apoyo incondicional. Por regalarme mi primera computadora, con la cual conocí el increíble mundo de la computación.

A mi mamá y a mi papá, por sus consejos y su apoyo. Por enseñarme que la cultura, la educación y la preparación académica son aspectos primordiales para el desarrollo personal.

A mis amigos, a los que están y a los que se han ido, por los buenos y malos momentos por los que pasamos juntos. En particular a Adrián, Alejandro y Alfredo. Las risas no faltaron.

A mis profesores, por compartir sus conocimientos y por ser parte fundamental en mi crecimiento profesional. Especialmente a los profesores: Dr. Guillermo Gilberto Molero Castillo, Ing. Gustavo Camacho Palacios, M.C.I. Jorge Alberto Solano Gálvez, Ing. Noé de Jesús Romero Serrano y al Ing. Ernesto Ramírez Sánchez.

A la Universidad Nacional Autónoma de México, por brindarme la mejor educación académica. Por permitirme conocer personas increíbles y darme momentos inolvidables.

Resumen

En la actualidad, la inteligencia computacional concentra una amplia variedad de métodos y algoritmos que se aplican para hacer frente a problemas complejos del mundo real. Es en el campo de la salud donde su uso se vuelve significativo para entender el comportamiento de una determinada enfermedad, como COVID-19. El presente documento, a manera de tesina, expone los resultados obtenidos del trabajo realizado bajo la Modalidad de Titulación por Actividad de Investigación, aprobado por el Comité de Titulación de la División de Ingeniería Eléctrica de la Facultad de Ingeniería. **Objetivo.** Implementar un método de inteligencia computacional para la clasificación de la mortalidad de adultos mayores contagiados con SARS-CoV-2 en la Ciudad de México. **Método.** La propuesta de solución para el análisis de la mortalidad de adultos mayores en la Ciudad de México, a consecuencia de COVID-19, fue dividido en cuatro etapas: i) adquisición de la fuente de datos, ii) selección de variables, iii) clasificación mediante bosques aleatorios, y iv) validación. **Resultados.** Con base en los resultados obtenidos, las comorbilidades con mayor grado de importancia en la clasificación fueron enfermedades crónicas renales, diabetes, enfermedades cardiovasculares y obesidad. Por otra parte, las comorbilidades con menor grado de importancia fueron: hipertensión, asma e inmunosupresión. Las variables relacionadas con la enfermedad pulmonar obstructiva crónica y el tabaquismo proporcionaron un porcentaje bajo de ganancia de información. **Conclusiones.** El algoritmo de bosques aleatorios obtuvo una exactitud promedio de 96.04% y una precisión de 98%, lo que significa una notable clasificación de la mortalidad de los adultos mayores contagiados con SARS-CoV-2 en la Ciudad de México.

Índice

1. Introducción **2**

1.1 Contexto de la investigación 2

1.2 Problema de investigación 3

1.3 Objetivos 4

1.3.1 Objetivo general 4

1.3.2 Objetivos específicos 4

1.4 Justificación 5

1.5 Organización del documento 6

2. Marco teórico y estado del arte **7**

2.1 Antecedentes de la Inteligencia Artificial y Computacional 7

2.2 Bosques Aleatorios 8

2.3 Trabajos relacionados 9

3. Método de solución **13**

3.1 Fuente de datos 13

3.2 Selección de variables 14

3.3 Clasificación 16

3.4 Validación 18

4. Resultados **19**

4.1 Resultados alcanzados 19

5. Conclusiones y trabajo futuro **22**

5.1 Conclusiones 22

5.2 Trabajo futuro 23

Anexo A 24

Anexo B 25

Anexo C 40

Referencias bibliográficas 45

Capítulo 1
Introducción

1.1. Contexto de la investigación

En la actualidad, la Inteligencia Artificial (IA) abarca una amplia variedad de subcampos, que van desde áreas de propósito general, como el aprendizaje y la percepción (Russell y Norvig, 2004), a otras más específicas, como aplicaciones basadas en inteligencia computacional, aprendizaje automático, aprendizaje profundo, por refuerzo, o mixto (Kaplan y Haenlein, 2019).

Precisamente, la Inteligencia Computacional (IC), concentra una amplia variedad de técnicas y algoritmos que se aplican para imitar el poder de pensamiento humano con el fin de hacer frente a problemas complejos del mundo real (Kumar *et al.*, 2021). Hoy en día, es evidente el impulso que ha tomado la IC en su aplicación sobre diferentes campos de actividad humana, como salud, seguridad, educación, biología, química, entre otros. Sin duda, en la actualidad, es en el campo de la salud donde su uso se vuelve significativo para entender el comportamiento de ciertas enfermedades, como COVID-19, que en el presente, convertido en pandemia, afectó a gran parte de la humanidad.

La pandemia por COVID-19 fue causada por un nuevo tipo de coronavirus, conocido como SARS-CoV-2. Los primeros casos de personas contagiadas se remontan a diciembre de 2019 en la ciudad de Wuhan, China. Así, desde los primeros contagios hasta mayo de 2023, se registraron más de 766 millones de casos confirmados y más de 6.9 millones de

defunciones a nivel mundial (OMS, 2023). En el caso específico de la Ciudad de México, objeto de estudio en esta investigación, se han reportado más de 1.89 millones de casos confirmados y más de 58 mil defunciones (Secretaría de Salud, 2023).

En este sentido, existen sectores de la población que pueden desarrollar fácilmente una complicación por COVID-19, incluso llegar a fallecer. A estos sectores de la población se denominan grupos vulnerables o de riesgo. Entre estos se encuentran las personas de 60 años o más, considerados como adultos mayores. De acuerdo con el Gobierno de la Ciudad de México (2020), este grupo vulnerable de adultos mayores se clasifica en dos categorías: i) con comorbilidad, que se caracteriza por ser personas mayores de 60 años, que tienen una o más enfermedades consideradas como factores de vulnerabilidad; y ii) sin comorbilidad, que se identifican como adultos mayores sin ninguna enfermedad o trastorno que se considere vulnerabilidad.

1.2. Problema de investigación

Existen ciertas características, enfermedades y padecimientos en los adultos mayores que afectan de forma considerable su estado de salud. Por tales motivos, se considera uno de los grupos con mayor vulnerabilidad ante la enfermedad COVID-19 (Vega *et al.*, 2020). Esta población adulta fácilmente puede desarrollar complicaciones e incluso morir por la enfermedad. Por lo que, es importante identificar los patrones que condicionan su estado de salud. Esto con el propósito de brindar información útil para tomar mejores decisiones sobre el tratamiento médico que reciben. Además, proporcionar un análisis reflexivo del grupo vulnerable mencionado.

Aunado a lo anterior, el contagio de los adultos mayores con el virus SARS-CoV-2, influyó directamente en la sociedad debido a la forma en que operó el semáforo de riesgo epidémico COVID-19, establecido por el Gobierno de la Ciudad de México, mediante el cual se anunció, a través de colores, el nivel de riesgo poblacional y el incremento o decremento de la actividad local, así como las medidas de seguridad sanitaria apropiadas para la

reapertura de las actividades laborales, educativas y el uso de los espacios públicos (Cortés y Dyer, 2021).

Por otro lado, es importante destacar el aumento de la población de adultos mayores en la última década, donde pasó de 9.1% en 2010 a 12.0% en 2020. Mientras que la población joven de 0 a 17 años disminuyó de 35.4% en 2010 a 30.4% en 2020 (INEGI, 2021). Esto significa que la población de adultos mayores en México va en aumento. Por tal motivo, al ser una población creciente y vulnerable, resulta importante realizar esfuerzos, desde diferentes vertientes, como es el caso de la Inteligencia Computacional, para analizar los riesgos y afectaciones que puedan presentar las personas mayores. Este tipo de análisis son útiles para identificar patrones en forma de tendencias sobre la población analizada.

1.3. Objetivos

1.3.1. Objetivo general

- Implementar un método de Inteligencia Computacional para la clasificación de la mortalidad de adultos mayores contagiados con SARS-CoV-2 en la Ciudad de México.

1.3.2. Objetivos específicos

- Hacer un análisis exploratorio de datos sobre registros de adultos mayores contagiados con SARS-CoV-2 en la Ciudad de México.
- Seleccionar las variables para definir la matriz de entrada para el funcionamiento del algoritmo seleccionado.
- Hacer ajustes en los hiperparámetros del algoritmo para su correcto funcionamiento con base en los resultados de entrenamiento y prueba.
- Interpretar los resultados obtenidos, de manera que permita clasificar nuevos registros de datos de adultos mayores contagiados con SARS-CoV-2.

1.4. Justificación

En la actualidad, es fundamental ejecutar esfuerzos multidisciplinarios para dar solución a diversos problemas de impacto social, como la vulnerabilidad de los adultos mayores ante la enfermedad COVID-19. Esto representa retos científicos y tecnológicos para su análisis, desde diferentes aristas, y así lograr un entendimiento del comportamiento de la enfermedad por el virus SARS-CoV-2 en la sociedad.

Como parte del desarrollo tecnológico actual, uno de los campos crecientes es la Inteligencia Computacional, que a través de técnicas y algoritmos especializados permite la identificación de patrones, efectuar predicciones, y servir de apoyo a la toma de decisiones. En este contexto, el algoritmo de bosques aleatorios es ideal para trabajar con una gran cantidad de datos y múltiples variables, debido a que selecciona muestras aleatorias para entrenar modelos de clasificación o pronóstico, según sea el caso (Merino y Chacón, 2017). Una aplicación de dicha tecnología, ronda en torno a la predicción de la tasa de mortalidad en pacientes infectados por el virus SAR-CoV-2, por medio de modelos entrenados con un historial clínico mundial. Con el objetivo de asignar prioridades a los pacientes con mayor riesgo de mortalidad durante su valoración clínica (Khan *et al.*, 2021).

En este sentido, es importante analizar el grupo de adultos mayores, debido a que es uno de los grupos vulnerables, que han sido gravemente afectados por la enfermedad COVID-19. El aumento de edad condiciona un descenso de la respuesta inmunológica y capacidades de regeneración, así como una disminución del índice de masa corporal, la funcionalidad y el aumento de las comorbilidades. Dadas estas situaciones, se ha evidenciado un incremento del riesgo de hospitalización y mortalidad en comparación con la población general (Rodríguez y López, 2020). Por lo tanto, a través de esta investigación se utiliza tecnología especializada para analizar el grupo vulnerable de adultos mayores en la Ciudad de México afectados por la enfermedad COVID-19.

1.5. Organización del documento

Este documento está organizado de la siguiente manera, el Capítulo 2 presenta los antecedentes de la inteligencia artificial y computacional, bosques aleatorios y los principales trabajos relacionados; el Capítulo 3 describe el método establecido como propuesta de solución; el Capítulo 4 presenta los resultados obtenidos, basados en datos de la población adulta mayor; y el Capítulo 5 resume las principales conclusiones y el trabajo futuro.

Se presenta además tres anexos, en los que se incluye información relacionada sobre el trabajo de investigación efectuado. En el Anexo A se presenta la carta de aceptación de la publicación del artículo de investigación en la revista Research in Computing Science. El Anexo B muestra el artículo de investigación aceptado para su publicación en la revista mencionada (www.rcs.cic.ipn.mx), cuyo título es 'Elderly mortality from COVID-19 in Mexico City: A Computacional Intelligence approach base on Random Forests'. En el Anexo C se presenta el código en Python de los métodos de inteligencia computacional utilizados para el análisis de la mortalidad de los adultos mayores por COVID-19 en la Ciudad de México.

Capítulo 2

Marco teórico y estado del arte

2.1. Antecedentes de la Inteligencia Artificial y Computacional

La Inteligencia Artificial como campo del conocimiento, propuesto por John McCarthy en 1956, hace referencia a la ciencia e ingeniería para la construcción de máquinas inteligentes, el cual ha enfrentado durante las últimas décadas múltiples desafíos, debido a la transición de estados con tecnologías emergentes, métodos y algoritmos (Fulcher, 2008; Raj, 2019). Esto hace que la Inteligencia Artificial tradicional fuera incompatible con las crecientes demandas en la búsqueda, optimización y resolución que los problemas actuales requieren. El camino de lo tradicional a lo moderno ha permitido el surgimiento de mejores herramientas computacionales como la Inteligencia Computacional (Raj, 2019).

A través de la Inteligencia Computacional es posible construir modelos, razonamientos, máquinas y procesos, basados en comportamientos estructurados e inteligentes (Raj, 2019). Este tipo de inteligencia adopta métodos que toleran el conocimiento incompleto, impreciso e incierto en entornos complejos. De esta forma, permiten soluciones aproximadas, flexibles, robustas y eficientes (Kruse *et al.*, 2016). Por lo tanto, la Inteligencia Computacional puede ser implementada para abordar problemas que afectan a la sociedad actual (Fulcher, 2008).

Sin duda, para construir modelos de aprendizaje inductivo, que basa su funcionamiento en el descubrimiento de patrones a partir de ejemplos, uno de los algoritmos

más utilizados en la Inteligencia Computacional son los árboles de decisión, mediante los cuales se pueden resolver problemas de pronóstico y clasificación, teniendo como objetivo construir una estructura jerárquica, eficiente y escalable en función de las condiciones (variables) establecidas en los datos. Para esto se utiliza la estrategia divide y vencerás.

2.2. Bosques aleatorios

Un árbol de decisión gráficamente se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo (variable) a partir del cual se inicia el proceso de clasificación. Los nodos internos corresponden a cada una de las condiciones de los atributos, asociadas a un determinado problema. Mientras que cada posible respuesta a las condiciones, se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con alguna clase (etiqueta) de la variable a clasificar (Martínez *et al.*, 2009).

Es importante mencionar que en ocasiones los árboles de decisión son susceptibles a desarrollar un sobreajuste (overfitting), lo cual significa que tienden a aprender muy bien de los datos de entrenamiento, pero su generalización pudiera ser no tan buena. Una forma de mejorar la generalización de los árboles de decisión es combinar varios árboles, a esto se le conoce como bosques aleatorios.

Los bosques aleatorios son ampliamente utilizados en la actualidad. Estos tienen como objetivo construir un ensamble de árboles de decisión, que al juntarlos, lo que en realidad está pasando es que estos ven distintas porciones de datos. Ningún árbol utiliza todos los datos de entrenamiento, sino cada uno se entrena con distintas muestras para un mismo problema. Al combinar los resultados, los errores se compensan con otros y se tiene una predicción (pronóstico o clasificación) que generaliza mejor al problema. La Figura 1 muestra el esquema general del funcionamiento de los bosques aleatorios para clasificación, el cual consta de cuatro pasos:

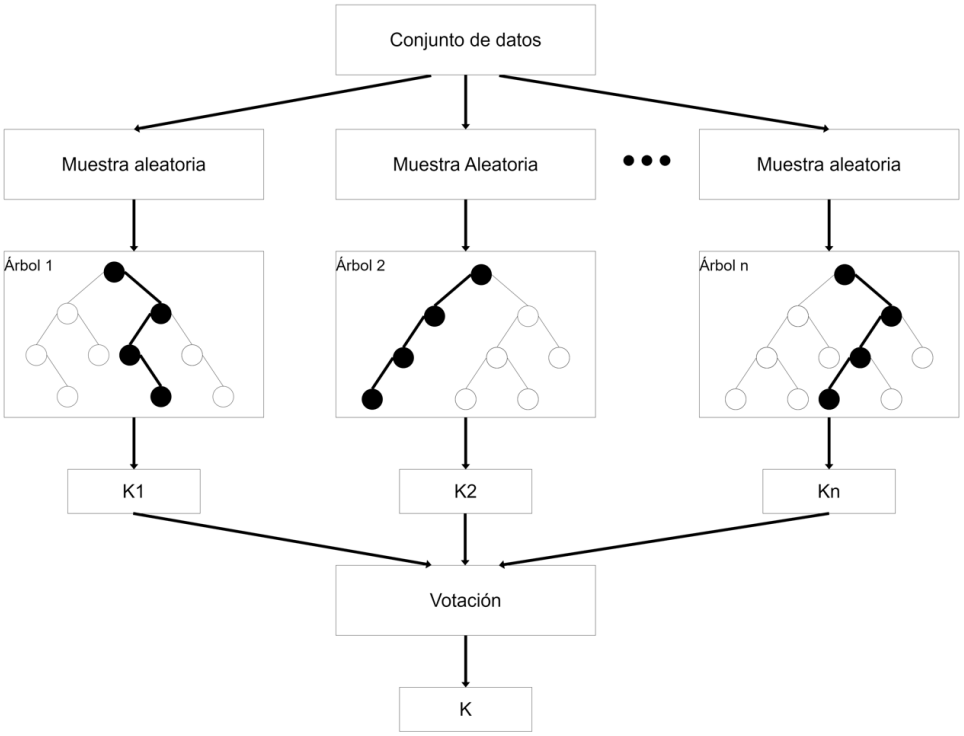

Figura 1. Esquema general de un bosque aleatorio.

1. Selección de muestras aleatorias a partir del conjunto de datos.
2. Construcción de un árbol de decisión para cada muestra y su respectivo resultado.
3. Votación (clasificación) con base en los resultados obtenidos.
4. Selección del resultado con más votos (clasificación).

2.3. Trabajos relacionados

En la actualidad, una de las aplicaciones significativas de los bosques aleatorios, debido a la pandemia por la enfermedad COVID-19, es la clasificación de la mortalidad en pacientes infectados por el virus SARS-CoV-2. El objetivo es clasificar características (variables) de pacientes con riesgo de mortalidad por dicha enfermedad (Khan *et al.*, 2021), como es el caso de grupos vulnerables, por ejemplos, los adultos mayores.

En Leandro-Astorga y Calvo (2021) se afirma que las personas mayores tienen una mayor probabilidad de contraer COVID-19 y desarrollar complicaciones. Se señala que en Estados Unidos, a través del Centro para el Control y la Prevalencia de Enfermedades (CDC, por sus siglas en inglés), se identificó que las personas mayores de 65 años, que representan el 17% del total de la población de ese país, representaron el 31% de las infecciones por SARS-CoV-2, el 45% de las hospitalizaciones, el 53% de los ingresos a unidades de cuidados intensivos, y el 80% de las muertes causadas por esta infección.

En la actualidad, se identificaron algunas investigaciones que han aportado conocimiento sobre la enfermedad de COVID-19 por medio de implementaciones de algoritmos de Inteligencia Computacional. Estos trabajos tienen diferentes enfoques y objetos de estudio. La Tabla 1 resume cinco de estos trabajos, donde se describe brevemente el trabajo realizado, el algoritmo utilizado y las limitaciones identificadas.

Tabla 1. Trabajos relacionados.

Autor	Descripción	Algoritmo utilizado	Limitaciones
Rami *et al.* (2022)	Se realizaron tres experimentos utilizando un conjunto de datos de pacientes con COVID-19. Se probaron siete modelos de clasificación. El mejor rendimiento se obtuvo con el algoritmo Bagging, con una precisión del 83.55%.	Bagging, J48, Regresión Logística, Bosque Aleatorio, Máquinas de Vectores de Soporte, Naïve Bayes, y Valor Umbral.	Se utilizaron los registros de 582 pacientes, de los cuales 15 características se utilizaron para el primer experimento, 6 para el segundo y 11 para el tercero.
Alves *et al.* (2021)	Se analizaron los casos de adultos mayores italianos hospitalizados por COVID-19. Los registros fueron divididos en dos grupos: sobrevivientes y no sobrevivientes. Posteriormente, se analizaron las comorbilidades de cada grupo. La demencia, la diabetes, la enfermedad renal crónica y la hipertensión arterial fueron las principales enfermedades implicadas en la mortalidad.	Statistical analysis (Software Stata).	El número de casos registrados utilizados osciló entre 18 y 1591 pacientes.

Khan *et al.* (2021)	Se analizó la tasa de mortalidad de los pacientes con COVID-19. Se utilizaron datos sociodemográficos y clínicos de pacientes de diferentes países y se evaluaron los modelos en cuanto a exactitud, precisión, sensibilidad y especificidad. El modelo de redes neuronales profundas logró una mejor predicción con un 97% de precisión.	Redes Neuronales Profundas, Árbol de Decisión, Regresión Logística, Bosque Aleatorio, Potenciación del Gradiente (XGBoost), K vecinos más próximos.	Se utilizaron 103888 registros de pacientes de 45 países, con el mayor número de India (98632), y Filipinas (4493). Sin embargo, el resto de los países contaban con menos de 200 registros.
Cardoso *et al.* (2021)	Se utilizaron algoritmos para predecir los casos positivos a COVID-19 y encontrar patrones en las bases de datos de seis distritos (municipios) de Argentina.	Relaciones Difusas y Redes Neuronales Artificiales.	El modelo de red neuronal artificial obtuvo un error medio del 20%.
Akinnuwesi *et al.* (2021)	Se analizaron métodos de inteligencia computacional para el diagnóstico de personas con COVID-19. El rendimiento de cada algoritmo se midió en términos de precisión, precisión, recuperación, equilibrio y exactitud. Los métodos con mejor rendimiento fueron Perceptron Multicapa, Mapa Cognitivo Difuso y Redes Neuronales Profundas.	Regresión Logística, Máquinas de Vectores de Soporte, Naïve Byes, Perceptrón multicapa, Mapa Cognitivo Difuso y Redes Neuronales Profundas.	Conjunto de datos limitado a 600 registros, de los cuales el 80% se utilizó en el entrenamiento y el 20% en las pruebas.

Varios trabajos relacionados utilizan algoritmos de inteligencia computacional para abordar los problemas derivados de la pandemia por COVID-19. Es importante destacar el uso de los métodos y herramientas que proporciona la inteligencia computacional para entender los riesgos y afectaciones que pueden sufrir ciertos grupos vulnerables, como es el caso de los adultos mayores.

Con relación a los trabajos relacionados identificados, en esta investigación se utilizó como algoritmo de clasificación a los bosques aleatorios, con el propósito de aprovechar las bondades y ventajas de este enfoque de inteligencia computacional, basado en la estrategia de divide y vencerás, con el cual se extraen reglas explicativas, ventaja que no tienen otros algoritmos al proporcionar soluciones sin justificación o explicación (Fulcher, 2008).

Además, el algoritmo de bosques aleatorios ha mostrado gran precisión para clasificar registros de personas con COVID-19 (Rami *et al.*, 2022).

Por otra parte, la mayoría de los trabajos relacionados identificados, realizan las investigaciones con fuentes de información de la población en general. En cambio, en esta investigación, se concentra el estudio en el grupo vulnerable de adultos mayores de la Ciudad de México. Permitiendo conocer factores que condicionen el estado de salud de dicha población, que actualmente va en aumento y tiene una mayor tasa de mortalidad por COVID-19 (Wang *et al.*, 2020).

Aunado a lo anterior, se utilizó una fuente de datos que contiene información de 591352 registros, que corresponden a los casos reales capturados de dos años de la población objeto de estudio. Mientras que los trabajos relacionados identificados realizaron sus investigaciones con fuentes de datos de menor periodo y tamaño.

Capítulo 3
Método de solución

El método de solución para el análisis de la mortalidad de adultos mayores en la Ciudad de México, a consecuencia de COVID-19, fue dividido en cuatro etapas (Figura 2): *i*) adquisición de la fuente de datos, *ii*) selección de variables, *iii*) clasificación, y *iv*) validación de la clasificación.

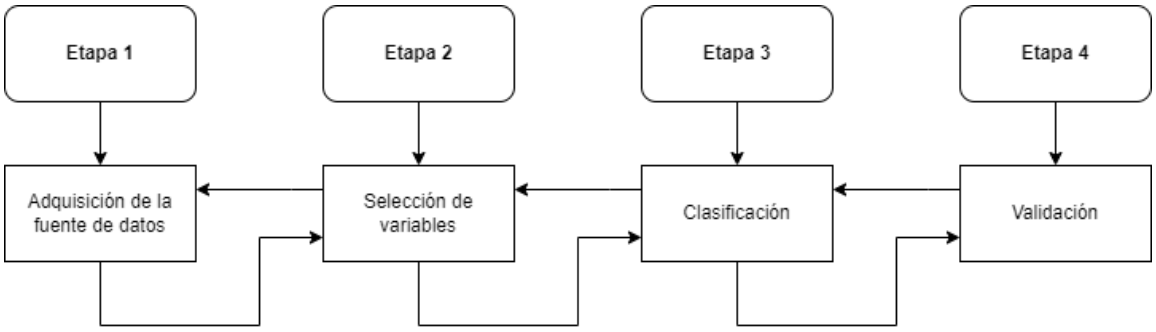

Figura 2. Esquema del método utilizado como solución propuesta.

3.1. Fuente de datos

Conforme al decreto publicado en el diario Oficial de la Federación el 20 de febrero de 2015, que establece la regulación en materia de Datos Abiertos, la Dirección General de Epidemiología, puso a disposición de la población en general las bases históricas publicadas desde el 14 de abril de 2020 sobre los casos asociados con COVID-19 (Gobierno de México, 2022).

Por lo que, en el presente, existen fuentes de datos sobre COVID-19 que elaboran diferentes entidades estatales, nacionales e internacionales; abarcando diferentes poblaciones. Por ejemplo, tan solo para Ciudad de México se encontraron 24 fuentes de datos relacionadas con COVID-19, como histórico de la capacidad hospitalaria, afluencia preliminar en el transporte público, inventarios de medidas por contingencia sanitaria, entre otras.

De manera particular, para esta investigación se utilizaron los datos de la Dirección General de Epidemiología de la Secretaría de Salud de México (Gobierno de México, 2022), los cuales son publicados periódicamente para facilitar a todos los usuarios el acceso, uso, reutilización y redistribución de los mismos. El propósito es monitorear los posibles casos de COVID-19 a nivel federal, y específicamente en la Ciudad de México. Por lo tanto, el periodo analizado en esta investigación comprende del 14 de abril de 2020 al 14 de abril de 2022, lo que representa 591352 casos reales de COVID-19 en adultos mayores en la Ciudad de México, es decir, registros de dos años consecutivos.

3.2. Selección de variables

La fuente de datos seleccionada contiene 40 variables, las cuales aportan información sobre el caso del paciente. Sin embargo, no todas las variables aportaban información significativa para esta investigación. Por lo que, se realizó un análisis exploratorio de datos (EDA, por sus siglas en inglés) con el propósito de hacer una cuidadosa selección de estas variables. Así, a partir de una selección de variables significativas desde el punto de vista médico y del análisis de datos, se obtuvo una fuente de datos compuesta por 20 variables, las cuales se listan en la Tabla 2.

Tabla 2. Variables seleccionadas.

No.	Nombre	Descripción	Valores
1	SEXO	Identifica el sexo del paciente.	1-Mujer, 2-Hombre, 99-No Especificado
2	TIPO_PACIENTE	Identifica el tipo de atención que recibió el paciente.	1-Ambulatorio, 2-Hospitalizado, 99-No Especificado
3	ESTADO	Identifica la situación (vivo o muerto) del paciente.	1-Vivo 2-Muerto

4	INTUBADO	Identifica si el paciente requirió de intubación.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
5	NEUMONIA	Identifica si el paciente se le diagnosticó con neumonía.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
6	EDAD	Identifica la edad del paciente.	Numérico
7	DIABETES	Identifica si el paciente tiene un diagnóstico de diabetes.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
8	EPOC	Identifica si el paciente tiene un diagnóstico de Enfermedad Pulmonar Obstructiva Crónica (EPOC).	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
9	ASMA	Identifica si el paciente tiene un diagnóstico de asma.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
10	INMUSUPPR	Identifica si el paciente tiene un diagnóstico de inmunosupresión.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
11	HIPERTENSION	Identifica si el paciente tiene un diagnóstico de hipertensión.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
12	OTRA_COM	Identifica si el paciente tiene diagnóstico de otras enfermedades.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
13	CARDIOVASCULAR	Identifica si el paciente tiene un diagnóstico de enfermedades cardiovasculares.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
14	OBESIDAD	Identifica si el paciente tiene diagnóstico de obesidad.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
15	RENAL_CRONICA	Identifica si el paciente tiene diagnóstico de insuficiencia renal crónica.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
16	TABAQUISMO	Identifica si el paciente tiene hábito de tabaquismo.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
17	OTRO_CASO	Identifica si el paciente tuvo contacto con algún otro caso diagnosticado con SARS-CoV-2.	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado
18	RESULTADO_ANTIGENO	Identifica el resultado del análisis de la muestra de antígeno para SARS-CoV-2.	1-Positivo SARS-CoV-2, 2-Negativo SARS-CoV-2, 97- No Aplica (Caso sin muestra)
19	CLASIFICACION_FINAL	Identifica la clasificación del resultado de la prueba Covid-19: confirmado, inválido, no realizado, sospechoso y negativo.	1-Confirmado por Asociación Clínica Epidemiológica, 2-Confirmado por comité de Dictaminación, 3-Caso confirmado, 4-Inválido por laboratorio, 5-No realizado por laboratorio, 6-Caso sospechoso, 7-Negativo a SARS-CoV-2.
20	UCI	Identifica si el paciente requirió ingresar a una Unidad de Cuidados Intensivos (UCI).	1-Sí, 2-No, 97-No aplica, 98-Se ignora, 99-No Especificado

Todas las variables seleccionadas contienen información relevante sobre el paciente, características del tratamiento médico recibido y la evolución de la enfermedad. A través de estas se pueden identificar patrones que permiten una clasificación precisa de la mortalidad de los adultos mayores infectados por el virus SARS-CoV-2. Mientras que otras variables fueron descartadas por incluir información redundante o no relevante, como es el caso de la toma de muestra (TOMA_MUESTRA_LAB), resultado de la prueba de laboratorio aplicada al paciente para confirmar la enfermedad (RESULTADO_LAB), muestra de antígeno de SARS-CoV-2 (TOMA_MUESTRA_ANTIGENO), por mencionar algunas de estas. Las cuales son redundantes debido a la existencia de otra variable que indica el resultado final de la prueba de antígeno de SARS-CoV-2 (RESULTADO_ANTIGENO).

3.3. Clasificación

El primer criterio para la clasificación utilizando bosques aleatorios es calcular la entropía para todas las clases y atributos. La entropía es una medida de incertidumbre (información) que se representa de la siguiente manera:

$$Entropía(S) = I(S) = Inf(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

Donde: *S* es una colección de elementos (objetos), y p_i es la probabilidad de posibles valores. Posteriormente, se selecciona el mejor atributo basado en la ganancia de información de cada variable, la cual se representa como:

$$(S, A) = Entropía(S) - \sum_{v \in V(A)} \frac{|Sv|}{|S|} Entropía(Sv)$$

Donde: *S* es una colección de elementos, *A* son las variables, *Sv* es un subconjunto de elementos, y *V(A)* es el conjunto de valores que *A* puede tomar.

Con base en lo anterior, para la clasificación de la mortalidad por COVID-19 en adultos mayores en la Ciudad de México, se seleccionaron los casos con base en las

siguientes condiciones: a) que la edad fuera mayor de 59 años, esto debido a que en la Ciudad de México las personas a partir de esa edad son consideradas adultos mayores; b) que la unidad médica y la residencia del paciente fueran la Ciudad de México; y c) que existiera una variable clase que permitiera identificar el estado de vida o muerte de las personas infectadas con SARS-CoV-2, es decir, casos 'vivos' o 'muertos', respectivamente.

Una vez finalizada la preparación y selección de las variables, se estableció una estructura compuesta por 19 variables independientes y una variable clase (ESTADO), descrita en la Tabla 2, como matriz de entrada para el funcionamiento del algoritmo. Posteriormente, para el proceso de clasificación y validación, la matriz de entrada se dividió en vectores de datos de entrenamiento y de prueba, como se muestra en el segmento de código escrito en Python (Figura 3).

```
X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y,
                                                                    test_size = 0.2,
                                                                    random_state = 0,
                                                                    shuffle = True)
```

Figura 3. Separación de los vectores de datos para el entrenamiento y prueba del algoritmo.

Por último, se ajustó el parámetro de profundidad máxima que pueden alcanzar los estimadores de bosques aleatorios (profundidad máxima de 8 niveles). También se configuró los criterios del número mínimo de muestras necesarias antes de dividir un nodo (al menos cuatro elementos) y el número mínimo de muestras en un nodo hoja (al menos dos elementos), como se muestra en la Figura 4.

```
ClassificationRF = RandomForestClassifier(random_state=0,
                                          max_depth=8,
                                          min_samples_leaf=2,
                                          min_samples_split=4)
ClassificationRF.fit(X_train, Y_train)
```

Figura 4. Configuración de parámetros para el funcionamiento del algoritmo.

Estos parámetros se ajustaron para evitar el sobreajuste de los estimadores del bosque aleatorio. En este sentido, con la configuración establecida y los vectores de datos de entrenamiento y prueba, se aplicó el algoritmo.

3.4. Validación

Para la validación del bosque aleatorio se utilizaron 113599 casos de prueba (el 20% de los registros, esto es, casos nuevos que no se utilizaron en el proceso de entrenamiento), de los cuales 4672 fueron clasificados erróneamente. La Tabla 3 muestra la matriz de clasificación obtenida para el caso de estudio.

Tabla 3. Matriz de clasificación.

		Clasificación	
		Muerto	Vivo
Real	Muerto	4047	2783
	Vivo	1889	109552

La matriz de clasificación muestra información acerca del desempeño del algoritmo, esto es, permite medir la exactitud de los resultados obtenidos (Inca-Balseca *et al*, 2022). La matriz de clasificación contiene cuatro tipos de resultados, los cuales son: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Capítulo 4

Resultados

4.1. Resultados alcanzados

Se identificaron las variables con mayor ganancia de información, las cuales se muestran en la Tabla 4. La variable que tiene la mayor relevancia para clasificar los casos fue INTUBADO, la cual hace referencia si fue necesario intubar al paciente. Otra variable importante fue TIPO_PACIENTE, encargada de describir la atención que recibió el paciente (ambulatorio - hospitalaria). La tercera variable fue UCI, cuyo objetivo es identificar si el paciente ingresó a una unidad de cuidados intensivos. La cuarta variable fue NEUMONIA, que tiene el propósito de identificar si el paciente fue diagnosticado con neumonía.

Tabla 4. Ganancia de información.

Variable	Importancia
INTUBADO	27.38 %
TIPO_PACIENTE	24.60 %
UCI	22.53 %
NEUMONIA	13.32 %
CLASIFICACION_FINAL	7.10 %
RESULTADO_ANTIGENO	2.37 %
OTRO_CASO	0.77 %
EDAD	0.74 %
SEXO	0.24 %
RENAL_CRONICA	0.22 %
DIABETES	0.14 %
CARDIOVASCULAR	0.10 %
OBESIDAD	0.09 %
OTRA_COM	0.09 %
EPOC	0.08 %

TABAQUISMO	0.08 %
HIPERTENSION	0.07 %
ASMA	0.05 %
INMUSUPPR	0.05 %

Las siguientes variables con menor porcentaje hacen referencia a la clasificación del resultado de la prueba COVID-19 (CLASIFICACION_FINAL), al resultado del análisis de la muestra de antígeno (RESULTADO_ANTIGENO) y si el paciente tuvo contacto con algún otro caso diagnosticado con la enfermedad (OTRO_CASO). Las variables posteriores hacen referencia a características del paciente (edad y sexo) y las comorbilidades que padece, como: insuficiencia renal, diabetes, enfermedades cardiovasculares y obesidad.

Se observó en uno de los estimadores del bosque aleatorio que el nodo principal fue la variable UCI, la cual corrobora una importante ganancia de información (22.53%). Otras variables en los siguientes niveles (nodos hijo) de los estimadores fueron las variables INTUBADO (27.38%), CLASIFICACIÓN_FINAL (7.1%), NEUMONÍA (13.32%) y RESULTADO_ANTIGENO (2.37%). Estas variables pertenecen al grupo que proporcionan más información para clasificar los datos.

Con respecto a la validación, se observó que los casos positivos fueron los registros asignados correctamente por el algoritmo en la categoría ‘Muerto’, como se pudo observar en la Tabla 3, se obtuvieron 4047 resultados de tipo verdadero positivo. Por otra parte, los verdaderos negativos fueron los registros clasificados como ‘Vivo’, cuando realmente pertenecen a dicha categoría, en este caso se obtuvieron 109552 clasificaciones correctas. Los resultados verdaderos positivos y verdaderos negativos representan una clasificación exitosa de los registros en la categoría a la cual pertenecen.

En cambio, los resultados falsos negativos y falsos positivos son los registros clasificados erróneamente por el algoritmo. Los resultados falsos negativos representan a los registros de tipo ‘Muerto’ que son clasificados en la categoría de ‘Vivo’, en este caso se obtuvieron 2783 falsos negativos. Por otro lado, los resultados falsos positivos son aquellos registros que pertenecen a la categoría ‘Vivo’ y que fueron clasificados como ‘Muerto’. En la matriz de clasificación se observan 1889 falsos positivos.

En este sentido, como resultado de la aplicación y validación del algoritmo, se observó un 96.04% de exactitud y 98% de precisión. Por otro lado, el error promedio es de 3.96%, demostrando una notable clasificación de los casos de supervivencia y mortalidad de los casos de COVID-19 en los adultos mayores de la Ciudad de México.

Como parte de la validación, se comprobó también la clasificación a través de un árbol de decisión, mediante el cual se obtuvo una exactitud sobresaliente del 95.3% y una precisión media del 97%. Sin embargo, a través del bosque aleatorio, como se observó, se obtuvo un mejor resultado tanto en exactitud (96.04%), precisión (98%) y menor error de casos mal clasificados (3.96%). Lo que representa que la solución a través del bosque aleatorio fue mejor. Además, reduce significativamente los puntos débiles del árbol de decisión, como el sobreajuste.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Un paciente de tipo hospitalario, que ingresó a la unidad de cuidados intensivos, que requirió ser intubado y fue diagnosticado con neumonía, tiene una alta probabilidad de ser clasificado como 'Muerto'. Debido a que las variables con mayor ganancia de información para la clasificación fueron INTUBADO, TIPO_PACIENTE, UCI y NEUMONIA.

Las variables relacionadas con la edad y el sexo del paciente tienen mayor grado de importancia que las variables asociadas con las comorbilidades, como: obesidad, hipertensión, asma, diabetes y otros. Para el caso de las personas clasificadas como fallecidas, la variable EDAD fue una condición de separación importante, donde los decesos fueron principalmente entre los 67 y 76 años.

Con base en los resultados obtenidos, las comorbilidades con mayor grado de importancia en la clasificación fueron enfermedades crónicas renales, diabetes, enfermedades cardiovasculares y obesidad. Por otra parte, las comorbilidades con menor grado de importancia fueron hipertensión, asma e inmunosupresión. Las variables relacionadas con la enfermedad pulmonar obstructiva crónica y tabaquismo proporcionaron un bajo porcentaje de ganancia de información.

El algoritmo de bosques aleatorios obtuvo una exactitud promedio de 96.04% y una precisión de 98%, lo que significa que la clasificación de la mortalidad de los adultos mayores

contagiados con SARS-CoV-2 en la Ciudad de México fue notable, cuyos casos fueron registrados en dos años, esto es, del 14 de abril de 2020 al 14 de abril de 2022.

5.2. Trabajo futuro

Existen diferentes grupos vulnerables que son gravemente afectados por el virus SARS-CoV-2, en este caso se concentró el esfuerzo y atención en el grupo de los adultos mayores. Sin embargo, existen otros sectores de la sociedad que se necesitan ser analizados, como los grupos indígenas, personas migrantes, personas con discapacidades, entre otras.

Como trabajo futuro, para enriquecer los resultados obtenidos, se pretende realizar un nuevo análisis con información actualizada y nuevos algoritmos utilizados en inteligencia computacional, como las máquinas de soporte de vectores (SVM, por sus siglas en inglés) y las redes neuronales profundas (DNN, por sus siglas en inglés). Esto puede ser importante y de gran interés debido al comportamiento de la pandemia por la enfermedad COVID-19, y su impacto en la población, especialmente en ciertos grupos vulnerables.

Anexo A

Carta de aceptación

En este apartado se presenta la carta de aceptación de la publicación del artículo de investigación en la revista Research in Computing Science, emitidada por el editor.

RESEARCH IN COMPUTING SCIENCE

ISSN 1870-4069

Centro de Investigación en Computación, Instituto Politécnico Nacional,
Av. Juan de Dios Bátiz, s/n, Col. La Escalera, CP 07320, DF, México
Tel.: +52-55-5729 6000, ext. 56518, 56653
http://www.rcs.cic.ipn.mx

Mexico City, September 18, 2022

Letter of acceptance

I hereby confirm that the paper

"Elderly mortality from COVID-19 in Mexico City: A Computational Intelligence approach based on Random Forests"

by Sinuhe Mazuti Osorio-Rivero, Guillermo Molero-Castillo, Everardo Bárcenas, Rocío Aldeco-Pérez

after thorough reviewing process is accepted for publication in our journal. The paper will be published in volume 152, No. 3 (2023), corresponding to March 2023.

With best regards,

Dr. Grigori Sidorov
Editor-in-Chief

Anexo B

Artículo publicado

En este anexo se presenta el artículo de investigación aceptado para su publicación en la revista Research in Computing Science, indizada en DBLP, LatIndex y Periodica.

Research in Computing Science

eISSN (applied for)
Indexing: DBLP, LatIndex, Periodica

Research in Computing Science, eISSN (applied for), is an internationally refereed open access scientific research journal published by the National Polytechnic Institute, a government-owned PhD-granting university subordinated to the Ministry of Public Education of Mexico. All papers submitted for publication are subject to rigorous international review process. Publication in this journal is free of charge. For the moment the journal is *not* indexed in WoS or EI. Contact: Prof. Grigori Sidorov, Editor-in-Chief. See the Editorial Board.

The topics of interest, number of pages per paper, submission procedure, deadlines, and contact for submissions are specified in the Call for Papers of a respective special issue or conference.

The format of papers is identical to Springer LNCS series format (though the journal is *not* published by Springer). You can find useful these formatting tips. Papers that do not follow these format requirements may be rejected without review or may be not included in the journal even if they have been accepted for publication. For the moment we do not require a copyright form, so please do not send it to us. We may contact you later for a copyright form.

Elderly mortality from COVID-19 in Mexico City: A Computational Intelligence approach based on Random Forests

Sinuhe Mazuti Osorio-Rivero, Guillermo Molero-Castillo,
Everardo Bárcenas, Rocío Aldeco-Pérez

Universidad Nacional Autónoma de México,
Facultad de Ingeniería,
México

mazuti.96@gmail.com, guillermo.molero@ingenieria.unam.edu,
{ebarcenas, raldeco}@unam.mx

Abstract. Computational intelligence encompasses a wide variety of techniques and algorithms that are applied to address complex real-world problems. It is in the field of health where its use becomes significant to understand the behavior of a given disease, such as COVID-19. In this sense, there are sectors of the population that can easily develop a complication or die from such a condition, such as people over 60 years of age. As this is a growing and vulnerable population, it is important to make efforts to analyze the risks and effects that the elderly may present. This paper presents the implementation of a computational intelligence method for the prediction of mortality in older adults infected with SARS-CoV-2 in Mexico City. Open data, published and distributed by the Government of Mexico City, were used for this analysis. The results show that the variables with the greatest contribution of information for classification were: intubated, patient care, pneumonia and intensive care unit (ICU). This concludes that a hospitalized patient, who is admitted to the intensive care unit and requires intubation, has a high probability of being classified as 'dead'. In addition, the results show that variables related to the patient's age and sex are more important than variables associated with comorbidities.

Keywords: COVID-19, Computational Intelligence, Elderly, Random Forests.

1 Introduction

Today, Artificial Intelligence (AI) encompasses a wide variety of subfields, ranging from general purpose areas, such as learning and perception [1], to more specific ones, such as applications based on computational intelligence, machine learning, deep learning, reinforcement, or mixed [2].

Specifically, Computational Intelligence (CI) concentrates a wide variety of techniques and algorithms that are applied to mimic human reasoning power in order to

cope with complex real-world problems [3]. Today, it is evident the momentum that CI has taken in its application in different fields of human activity, such as health, security, education, biology, among others. Undoubtedly, at present, it is in the field of health where its use becomes significant to understand the behavior of certain diseases, such as COVID-19, which is currently considered a pandemic affecting humanity.

The COVID-19 pandemic was caused by a new type of coronavirus, known as SARS-CoV-2. The first cases of infected people date back to December 2019 in the city of Wuhan, China. Thus, from the first infections until October 2022, there are more than 615 million confirmed cases and more than 6.5 million deaths worldwide [4]. In the specific case of Mexico City, the object of study in this research, more than 1.74 million confirmed cases and more than 57 thousand deaths have been reported [5].

In this sense, there are sectors of the population that can easily develop a COVID-19 complication and even die. These sectors of the population are called vulnerable or at-risk groups. Among them are people 60 years of age or older, considered to be the elderly. According to the Government of Mexico City, this vulnerable group of older adults is classified into two categories [6]: i) with comorbidity, which is characterized as people over 60 years of age, who have one or more diseases considered as factors of vulnerability; and ii) without comorbidity, which are identified as older adults without any disease or disorder that is considered a vulnerability.

There are certain characteristics, diseases and conditions in older adults that considerably affect their health status. For these reasons, they are considered one of the groups with the greatest vulnerability to COVID-19 disease [7]. This adult population can easily develop complications and even die from the disease. Therefore, it is important to identify the patterns that condition their health status. The purpose of this is to provide useful information to understand and make better decisions about the management of pandemic disease. In addition, to provide a reflective analysis of the vulnerable group mentioned.

In addition to the above, the infection of older adults with the SARS-CoV-2 virus has a direct influence on society due to the way in which the epidemic risk traffic light COVID-19 operates, established by the Government of Mexico City through which the level of population risk and the increase or decrease of local activity is announced through colors, as well as the appropriate health safety measures for the reopening of work and educational activities and the use of public spaces [8].

On the other side, it is important to highlight the increase in the population of older adults in the last decade, where it went from 9.1% in 2010 to 12.0% in 2020. While the young population aged 0 to 17 years decreased from 35.4% in 2010 to 30.4% in 2020 [9]. This means that the population of older adults in Mexico is increasing. For this reason, being a growing and vulnerable population, it is important to make efforts, from different perspectives, as is the case of computational intelligence, to analyze the risks and affectations that the elderly may present. This type of analysis is useful for identifying patterns in the form of trends in the population under analysis.

The aim of this research work was to implement a computational intelligence method, specifically random forests, for the classification of mortality in older adults infected with SARS-CoV-2 in Mexico City. For this, open data was used, published by the Government of Mexico City.

This paper is organized as follows: Section 2 presents the background of artificial and computational intelligence, COVID-19 in the adult population and the main related works; Section 3 describes the method established as a proposed solution; Section 4 presents the results obtained, based on a use case, such as the adult population; and Section 5 summarizes the main conclusions and future work.

2 Background

Artificial intelligence as an area of knowledge, proposed by John McCarthy in 1956, which refers to the science and engineering for the construction of intelligent machines, has faced multiple challenges during the last decades, due to the transition of states with emerging technologies, methods and algorithms [10] [11]. This makes traditional artificial intelligence incompatible with the increasing demands in search, optimization and resolution that problems require. The path from traditional to modern has enabled the emergence of better computational tools such as computational intelligence [11].

Through computational intelligence it is possible to build models, reasoning, machines and processes, based on structured and intelligent behaviors [11]. This type of intelligence adopts methods that tolerate incomplete, imprecise and uncertain knowledge in complex environments. In this way, they allow approximate, flexible, robust and efficient solutions [12]. Therefore, computational intelligence can be implemented to address problems that affect today's society [10].

Undoubtedly, to build inductive learning models, which base their function on the discovery of patterns from examples, one of the most used algorithms in computational intelligence are decision trees (DTs), through which prognosis and classification problems can be solved, aiming to build a hierarchical, efficient and scalable structure based on the conditions (variables) established in the data. The divide and conquer strategy are used for this purpose.

2.1 Random Forests

A tree is graphically represented by a set of nodes, leaves and branches. The main node or root is the attribute (variable) from which the classification process starts. The internal nodes correspond to each of the attribute conditions associated with a given problem. While each possible answer to the conditions is represented by a child node. The branches coming out of each of these nodes are labeled with the possible values of the attribute. The final nodes or leaf nodes correspond to a decision, which coincides with some class (label) of the variable to be classified [13].

It is important to mention that sometimes decision trees are susceptible to overfitting, which means that they tend to learn very well from the training data, but their generalization may not be as good. One way to improve the generalization of decision trees is to combine several trees, known as random forests (RFs).

Random forests are widely used today. They aim to build an ensemble of decision trees, which when put together, what is actually happening is that they see different portions of the data. No tree uses all the training data, but each one is trained with different samples for the same problem. By combining the results, the errors are

compared with each other, and it has a prediction (forecast or classification) that generalizes better to the problem. Figure 1 shows the general scheme of the operation of random forests for classification, which consists of four steps:

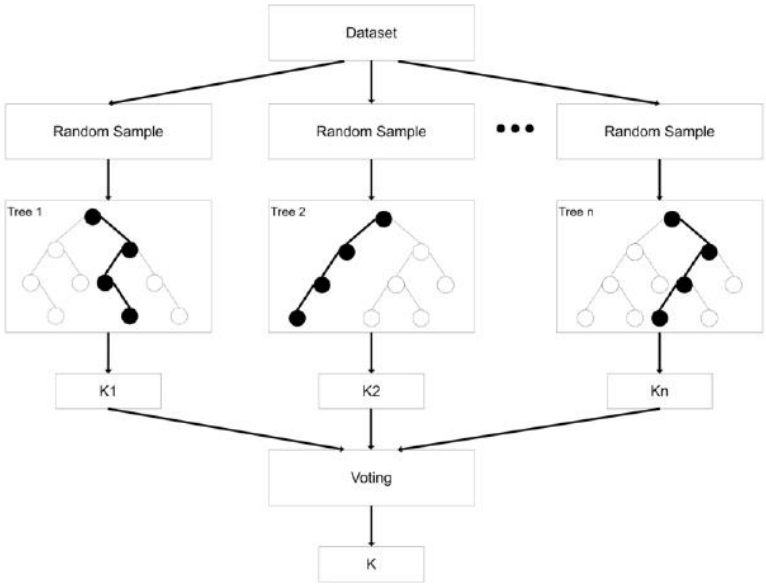

Fig. 1. Random forest general scheme.

1. Selection of random samples from the data set.
2. Construction of a decision tree for each sample and its respective result.
3. Voting (classification) based on the results obtained.
4. Selection of the result with the most votes (ranking).

2.2 Related work

At present, one of the significant applications of random forests, due to the COVID-19 pandemic, is the classification of mortality in patients infected by the SARS-CoV-2 virus. The objective is to classify characteristics (variables) of patients at risk of mortality due to this disease [14], as is the case of vulnerable groups, for example, the elderly.

In [15] it is stated that older people are more likely to contract COVID-19 and develop complications. These same authors mention that in the United States, through the Center for Disease Control and Prevalence (CDC) [16], it was identified that people over 65 years of age, accounted for 31% of SARS-CoV-2 infections, 45% of hospitalizations, 53% of admissions to intensive care units, and 80% of deaths caused by this infection.

Today, some researches have been identified that have provided knowledge about the COVID-19 disease by means of implementations of computational intelligence

algorithms. These works have different approaches and objects of study. Table 1 summarizes five of these works, where the work performed, algorithms used, and limitations identified are briefly described.

Table 1. Related work.

Author	Description	Algorithm used	Limitations
Rami *et al.* (2022) [17]	Three experiments were performed using a data set of patients with COVID-19. Seven classification models were tested. The best performance was with the Bagging algorithm, with an accuracy of 83.55%.	Bagging, J48, Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Threshold Selector.	Records from 582 patients were used, of which 15 features were used for the first experiment, 6 for the second and 11 for the third.
Alves *et al.* (2021) [18]	Cases of Italian older adults hospitalized for COVID-19 were analyzed. Subsequently, the comorbidities of each group were analyzed. Dementia, diabetes, chronic kidney disease and high blood pressure were the main diseases involved in mortality.	Statistical analysis (Stata software).	The number of registered cases used ranged from 18 to 1591 patients.
Khan *et al.* (2021) [14]	The mortality rate of patients with COVID-19 was analyzed. Sociodemographic and clinical data from patients from different countries were used, and the models were evaluated for accuracy, precision, sensitivity and specificity. Deep Neural Networks model achieved a better prediction with 97% accuracy.	Deep Neural Network (DNN), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor (KNN).	It was used 103888 patient records from 45 countries, with the largest number from India (98632), and Philippines (4493). Nevertheless, there were less than 200 records for remaining countries
Cardoso *et al.* (2021) [19]	Algorithms were used to predict COVID-19 positive cases and find patterns in the databases of six districts (municipalities) in Argentina.	Fuzzy relationships and Artificial Neural Networks.	The artificial neural network model obtained an average error of 20%.
Akinnuwesi *et al.* (2021) [20]	Computational intelligence methods for the diagnosis of people with COVID-19 were analyzed. The performance of each algorithm was measured in terms of accuracy, precision, recall, balanced and accuracy. The methods with the best performance were MLP, FCM and DNN.	Logistic Regression (LR), Support Vector Machine (SVM), Naïve Byes (NB), Multilayer Perceptron (MLP), Fuzzy Cognitive Map (FCM) and Deep Neural Network (DNN).	Dataset limited to 600 records, of which 80% were used in training and 20% in testing.

Several related works use computational intelligence algorithms to address problems arising from the COVID-19 pandemic. It is important to highlight the use of the methods and tools provided by computational intelligence to understand the risks and affectations that certain vulnerable groups may suffer, as is the case of the elderly.

In relation to the related works identified, random forests were used as a classification algorithm for this research, with the purpose of taking advantage of the benefits and advantages of this computational intelligence approach, based on the divide and conquer strategy, with which explanatory rules are extracted, an advantage that other algorithms do not have by providing solutions without justification or explanation [10]. In addition, the random forest algorithm has shown high accuracy for classifying records of people with COVID-19 [17].

On the other hand, most of the identified related works perform the investigations with general population information sources. In contrast, in this research, the study focuses on the vulnerable group of older adults in Mexico City. This allows to learn about factors that condition the health status of this population, which is currently increasing and has a higher mortality rate due to COVID-19 [21].

In addition to the above, a data source containing 591352 records was used, corresponding to real cases captured during two years of the population under study. While the related works identified carried out their research with data sources of smaller period and size.

2.3 Motivation

Random forests are ideal for working with a large amount of data and multiple variables, due to the fact that it selects random samples to train classification or prognostic (regression) models, as the case may be [22].

In this sense, it is important to analyze the group of older adults, because it is one of the vulnerable groups that have been severely affected by COVID-19 disease. Increasing age conditions, a decrease in immune response and regenerative capacities, as well as a decrease in body mass index, functionality and an increase in comorbidities. Given these situations, there is evidence of an increased risk of hospitalization and mortality compared to the general population [23]. Therefore, through this research, specialized technology is used to analyze the vulnerable group of older adults in Mexico City affected by COVID-19 disease.

3 Method

The solution method for the analysis of mortality of elderly in Mexico City, as a result of COVID-19, was divided into four stages (Figure 2): i) acquisition of data source, ii) variables selection, iii) classification, and iv) validation.

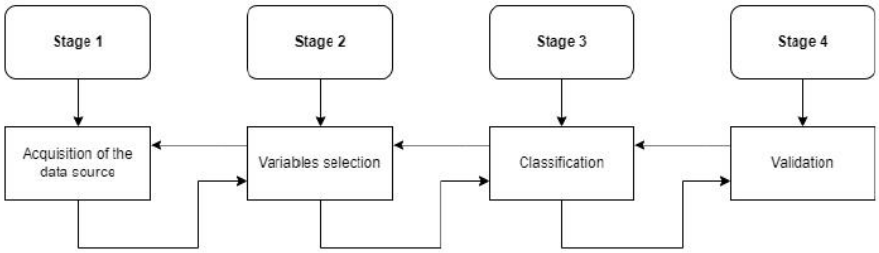

Fig. 2. Scheme of the method used as a proposed solution.

3.1 Data source

In accordance with the decree published in the Official Journal of the Federation on February 20, 2015, which establishes the regulation on Open Data, the General Directorate of Epidemiology, made available to the general population the historical bases published since April 14, 2020 on cases associated with COVID-19 [24].

Thus, at present, there are sources of data on COVID-19 produced by different state, national and international entities, covering different populations. For example, for Mexico City alone, 24 data sources related to COVID-19 were found, such as historical hospital capacity, preliminary affluence in public transportation, inventories of contingency measures, among others.

In particular, for this research, data from the General Directorate of Epidemiology of the Ministry of Health of Mexico [24] were used, which are published periodically to facilitate all users access, use, reuse and redistribution of the same. The purpose is to monitor possible cases of COVID-19 at the federal level, and specifically in Mexico City. Therefore, the period analyzed in this research comprises from April 14, 2020 to April 14, 2022, which represents 591352 real cases of COVID-19 in older adults in Mexico City, that is, records of two consecutive years.

3.2 Variables selection

The original data source contains 40 variables, which provide information about the patient's case. However, not all the variables provide significant information for this research. So, an exploratory data analysis (EDA) was performed in order to make a careful selection of these variables. Thus, from a selection of significant variables from the medical point of view and data analysis, a data source consisting of 20 variables was obtained, which are listed in Table 2.

Table 2. Selected variables.

Item	Name	Description	Values
1	SEX	Identifies the sex of the patient.	1-Female, 2-Male, 99-Not Specified
2	PATIENT_CARE	Identifies the care type that patient received.	1-Ambulatory, 2-Hospitalized, 99-Not Specified

3	STATE	Identifies the situation (alive or dead) of the patient.	1-Alive 2-Dead
4	INTUBATED	Identifies whether the patient required intubation.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
5	PNEUMONIA	Identifies whether the patient was diagnosed with pneumonia.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
6	AGE	Identifies the patient's age.	Numeric
7	DIABETES	Identifies if the patient has a diagnosis of diabetes.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
8	COPD	Identifies if the patient has a diagnosis of Chronic Obstructive Pulmonary Disease (COPD).	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
9	ASTHMA	Identifies if the patient has a diagnosis of asthma.	1-Yes, 2-No, 97-Does not apply, 98-Ignore, 99-Not Specified
10	INMUSUPPR	Identifies if the patient has a diagnosis of immunosuppression.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
11	HYPERTENSION	Identifies if the patient has a diagnosis of hypertension	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
12	OTHER_DISEASES	Identifies whether the patient has a diagnosis of other diseases.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
13	CARDIOVASCULAR	Identifies whether the patient has a diagnosis of cardiovascular disease.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
14	OBESITY	Identifies if the patient has a diagnosis of obesity.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
15	CHRONIC_RENAL	Identifies if the patient has a diagnosis of chronic renal insufficiency.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
16	SMOKING	Identifies if the patient has a smoking habit.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
17	OTHER_CASE	Identifies if the patient had contact with any other case diagnosed with SARS-CoV-2.	1-Yes, 2-No, 97-Not Applicable, 98-Ignore, 99-Not Specified
18	ANTIGEN_RESULT	Identifies the result of the SARS-CoV-2 antigen sample analysis.	1-Positive SARS-CoV-2, 2-Negative SARS-CoV-2, 97- Not Applicable (Case without sample).
19	FINAL_CLASSIFICATION	Identifies the classification of the Covid-19 test result: confirmed, invalid, not performed, suspect, and negative.	1-Confirmed by the Clinical Epidemiological Association, 2-Confirmed by the Ruling Committee, 3-Confirmed case, 4-Invalid by laboratory, 5-Not laboratory performed, 6- Suspect case, 7-Negative to SARS-COV-2
20	ICU	Identifies whether the patient required admission to an Intensive Care Unit (ICU).	1-Yes, 2-No, 97-Not applicable, 98-Unknown, 99-Not Specified

All the selected variables contain relevant information about the patient, characteristics of the medical treatment received and the evolution of the disease. Through which patterns can be identified that allow an accurate classification of the mortality of older adults infected with SARS-CoV-2. While other variables were discarded because they included redundant or irrelevant information, as is the case of sample taking (SAMPLING), result of the laboratory test applied to the patient to confirm the disease (LAB_RESULT), SARS-CoV-2 antigen sample (ANTINGEN_SAMPLING), to mention a few of these. Which are redundant due to the existence of another variable that indicates the final result of the antigen test for SARS-CoV-2 (ANTIGEN_RESULT).

3.3 Classification

The first criterion for classification using random forests is to calculate the entropy for all classes and attributes. Entropy is a measure of uncertainty (information) that is represented as follows:

$$Entropy(S) = l(S) = lnf(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

Where: S is a collection of elements (objects), and p_i is the probability of possible values. Subsequently, the best attribute is selected based on the information gain of each variable, which is represented as:

$$(S, A) = Entropy(S) - \sum_{v \in V(A)} \frac{|Sv|}{|S|} Entropy(Sv)$$

Where: S is a collection of elements, A are the variables, Sv is a subset of elements, and $V(A)$ is the set of values that A can take.

Based on the above, for the classification of COVID-19 mortality in older adults in Mexico City, the cases were selected based on the following conditions: a) that the age was greater than 59 years, this because in Mexico City people from that age are considered older adults; b) that the medical unit and residence of the patient were Mexico City; and c) there is a class variable that allows identifying the state of life or death of people infected with SARS-CoV-2, that is, 'Alive' or 'Dead' cases, respectively.

Once the preparation and selection of variables was completed, a structure made up of 19 independent variables and one class variable (STATE), described in Table 2, was established as an input matrix for the operation of the algorithm. Subsequently, for the classification and validation process, the input matrix was divided into training and test data vectors, as shown in the code segment written in Python (Figure 3).

```
X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y,
                                                                    test_size = 0.2,
                                                                    random_state = 0,
                                                                    shuffle = True)
```

Fig. 3. Separation of the data vectors for the training and testing of the algorithm.

Finally, the maximum depth parameter that random forest estimators can reach (maximum depth of 8 levels) has been adjusted. The criteria of the minimum number of samples required before splitting a node (at least 4 elements) and the minimum number of samples in a leaf node (at least 2 elements) were also adjusted, as shown in Figure 4.

```
ClassificationRF = RandomForestClassifier(random_state=0,
                                          max_depth=8,
                                          min_samples_leaf=2,
                                          min_samples_split=4)
ClassificationRF.fit(X_train, Y_train)
```

Fig. 4. Configuration of parameters for the operation of the algorithm.

These parameters were adjusted in order to avoid overfitting the random forest estimators. In this sense, with the established configuration and the training and test data vectors, the algorithm was applied.

3.4 Validation

For the validation of the random forest, 113599 test cases were used (20% of records, new cases, which were not used in the training process), of which 4672 were misclassified. Table 3 shows the classification matrix obtained for the case study.

Table 3. Classification matrix.

		Classification	
		Dead	**Alive**
Real	**Dead**	4047	2783
	Alive	1889	109552

The classification matrix shows information about the performance of the algorithm, that it, it allows measuring the accuracy of the results obtained [25]. The classification matrix contains four types of results, which are: true positives, true negatives, false positives and false negatives.

4 Results

The variables with the greatest gain of information were identified, which are shown in Table 4. The variable with the greatest relevance for classifying cases was INTUBATED, which refers to whether it was necessary to intubate the patient. Another important variable was PATIENT_CARE, which describes the care received by the patient (outpatient - inpatient). The third variable was ICU, whose objective is to identify whether the patient was admitted to an intensive care unit. The fourth variable was PNEUMONIA, which to identify whether the patient was diagnosed with pneumonia.

Table 4. Information gain.

Variable	Importance
INTUBATED	27.38 %
PATIENT_CARE	24.60 %
ICU	22.53 %

PNEUMONIA	13.32 %
FINAL_CLASSIFICATION	7.10 %
ANTIGEN_RESULT	2.37 %
OTHER_CASE	0.77 %
AGE	0.74 %
SEX	0.24 %
CHRONIC_RENAL	0.22 %
DIABETES	0.14 %
CARDIOVASCULAR	0.10 %
OBESITY	0.09 %
OTHER_DISEASES	0.09 %
COPD	0.08 %
SMOKING	0.08 %
HYPERTENSION	0.07 %
ASTHMA	0.05 %
INMUSUPPR	0.05 %

The following variables with lower percentages refer to the classification of the COVID-19 test result (FINAL_CLASSIFICATION), the result of the antigen sample analysis (ANTIGEN_RESULT) and whether the patient had contact with any other case diagnosed with the disease (OTHER_CASE). Subsequent variables refer to patient characteristics (age and sex) and comorbidities (renal failure, diabetes, cardiovascular disease and obesity).

It was observed in one of the estimators of the random forest that the main node was the variable UCI, which corroborates an important gain of information (22.53%). Other variables at the next levels (child nodes) of the estimators were the variables INTUBATED (27.38%), FINAL_CLASSIFICATION (7.1%), PNEUMONIA (13.32%) and ANTIGEN_RESULT (2.37%). These variables belong to the group that provides more information to classify the data.

Regarding validation, it was observed that the positive cases were the records correctly assigned by the algorithm in the 'Dead' category, as can be seen in Table 3, of which 4047 correct classifications were obtained, considered as true positives. On the other side, true negatives were the cases classified as 'Alive', when they really belong to that category. In this case, 109552 correct classifications were obtained, considered true negatives. The true positive and negative results represent a successful classification of the analyzed cases in the category to which they belong.

The opposite happens with the results of false negatives and false positives, which are cases misclassified by the algorithm. False negatives represent cases of type 'Dead' that were classified in the 'Alive' category, in this case 2783 false negatives were obtained. On the other side, the false positive results are those cases that belong to the 'Alive' category and were classified as 'Dead'. In the classification matrix, 1889 false positives were observed.

As a result of the application and validation of the algorithm, 96.04% accuracy and 98% precision were obtained. On the other side, the average error was 3.96%, demonstrating a remarkable classification of survival and mortality of COVID-19 cases in older adults in Mexico City.

As part of the validation, the classification was also tested through a decision tree, whereby which an outstanding accuracy of 95.3% and an average precision of 97%

were obtained. Nevertheless, through random forest, as observed, a better result was obtained both in accuracy (96.04%), precision (98%) and less error of misclassified cases (3.96%). Which represents that the solution through the random forest was better. Furthermore, it significantly reduces decision tree weaknesses such as overfitting.

5 Conclusions

A hospitalized-type patient is one who was admitted to the intensive care unit, required intubation, and was diagnosed with pneumonia. This type of patient has a high probability of being classified as 'Dead', since the variables with the greatest information gain for the classification were INTUBATED, PATIENT_CARE, ICU (Intensive Care Unit), and PNEUMONIA.

Variables related to the patient's age and sex have a higher degree of importance than variables associated with comorbidities, such as: obesity, hypertension, asthma, diabetes and others. For the case of persons classified as deceased, the variable AGE was an important separating condition, where deaths were mainly between 67.5 and 76.5 years old.

Based on the results obtained, the comorbidities with the highest degree of importance in the classification were chronic kidney disease, diabetes, cardiovascular disease and obesity. On the other side, the comorbidities with the lowest degree of importance were: hypertension, asthma and immunosuppression. The variables related to chronic obstructive pulmonary disease and smoking provided a low percentage of information gain.

The random forest algorithm obtained an average accuracy of 96.04% and a precision of 98%, which means that the mortality classification of older adults infected with SARS-CoV-2 in Mexico City was remarkable, whose cases were registered in two years, that is, from April 14, 2020 to April 14, 2022.

There are different vulnerable groups that are severely affected by the SARS-CoV-2 virus; in this case, efforts and attention were focused on the elderly. However, there are other sectors of society that need to be analyzed, such as indigenous groups, migrants, people with disabilities, among others.

As future work, to enrich the results obtained, it is intended to make a new analysis with updated information and new algorithms used in computational intelligence, such as support vector machines (SVM) and deep neural networks (DNN). This may be important and of great interest due to the behavior of the pandemic due to the disease caused by COVID-19 and its impact on the population, especially on certain vulnerable groups.

References

1. Russell, S., Norvig, P.: Inteligencia artificial: Un enfoque moderno. Pearson Education (2004)
2. Kaplan, A., Haenlein, M.: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25 (2019)

3. Kumar, G., Jain, S. Singh, U.: Stock market forecasting using computational intelligence: A survey. Archives of Computational Methods in Engineering, 28(3), 1069-1101 (2021)
4. OMS-World Health Organization-: Weekly epidemiological update on COVID-19, October 5, 2022
5. Secretaría de Salud: Informe Técnico Semanal COVID-19 México, Mexico City, October 4, 2022
6. Government of Mexico City: Criterios para las poblaciones en situación de vulnerabilidad que pueden desarrollar una complicación o morir por COVID-19 en la reapertura de actividades económicas en los centros de trabajo (2020)
7. Vega, J. A., Ruvalcaba, J. C., Hernández, I., Acuña, M. D., López, L: La salud de las personas adultas mayores durante la pandemia de COVID-19. Journal of Negative and No Positive Results, 5(7), 726-739 (2020)
8. Cortés, R., Dyer, D.: Lineamiento para la estimación de riesgos del semáforo por regiones COVID-19. Secretaria de Salud (2021)
9. INEGI: En México somos 126'014,024 habitantes: censo de población y vivienda 2020 (2021)
10. Fulcher, J.: Computational intelligence: an introduction. In Computational intelligence: a compendium, pp. 3-78. Springer, Berlin, Heidelberg (2008)
11. Raj, J. S.: A comprehensive survey on the computational intelligence techniques and its applications. Journal of ISMAC, 1(03), 147-159 (2019)
12. Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., Steinbrecher, M.: Introduction to Computational Intelligence. Springer, London (2016)
13. Martínez, R., Ramírez, N., Mesa, H., Suárez, I., Trejo, M.., León, P., Morales, S.: Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, 9(2), 19-24 (2009)
14. Khan, I., Aslam, N., Aljabri, M., Aljameel, S., Kamaleldin, M., Alshamrani, F., Chrouf, S.: Computational Intelligence-Based Model for Mortality Rate Prediction in COVID-19 Patients. International Journal of Environmental Research and Public Health, 18(12), 6429 (2021)
15. Leandro-Astorga, G., Calvo, I. B.: Infección por COVID-19 en población adulta mayor: recomendaciones para profesionales. Revista Médica de Costa Rica y Centroamérica, 86(629), 44-50 (2021)
16. Centers for Disease Control and Prevention: Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19)—United States, February 12–March 16, 2020. https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm. Accessed October 5, 2022.
17. Rami, A., Malak, M., Aboulnour, M., Mirza, S., Alshobaiki, A.: Classifying the Mortality of People with Underlying Health Conditions Affected by COVID-19 Using Machine Learning Techniques. Applied Computational Intelligence and Soft Computing. DOI: 10.1155/2022/3783058 (2022)
18. Alves, V., Casemiro, F., De Araujo, B., De Souza, M., Silva, R., Tamires, F., Campos A. Gregori, D.: Factors Associated with Mortality among Elderly People in the COVID-19 Pandemic (SARS-CoV-2): A Systematic Review and Meta-Analysis, International Journal of Environmental Research and Public Health, 15, 8008. DOI: 10.3390/ijerph18158008 (2021)
19. Cardoso, A., Talame, L., Amor, M.: Aprendizaje automático aplicado a la pandemia del virus Covid-19 en Argentina. In XXIII Workshop de Investigadores en Ciencias de la Computación (2021)
20. Akinnuwesi, B. A., Fashoto, S. G., Mbunge, E., Odumabo, A., Metfula, A. S., Mashwama, P. et al.: Application of intelligence-based computational techniques for classification and

early differential diagnosis of COVID-19 disease. Data Science and Management, 4, 10-18 (2021)
21. Wang, X., Song, G., Yang, Z., Chen, R., Zheng, Y., Hu, H., Su, X., Chen, P.: Association between ageing population, median age, life expectancy and mortality in coronavirus disease (COVID-19). Aging (Albany NY), 12(24), 24570-24578. DOI: 10.18632/aging.104193 (2020)
22. Merino, R. F., Chacón, C. I.: Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases, (10), 165-189 (2017)
23. Rodríguez, Y. L., López, L. A.: A propósito del artículo "COVID-19. De la patogenia a la elevada mortalidad en el adulto mayor y con comorbilidades. Revista Habanera de Ciencias Médicas, 19(4), 14 (2020)
24. Gobierno de México: Datos Abiertos de Bases Históricas. Dirección General de Epidemiología de la Secretaría de Salud. www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia. Last accessed in May 16, 2022
25. Inca-Balseca, C. L., Paredes-Proaño, A. M., Cornejo-Reyes, P. J., Mena-Reinoso, Á. P.: Eficiencia de modelos de predicción de COVID-19 usando curvas ROC y matriz de confusión. Dominio de las Ciencias, 8(2), 1442-1460 (2022)

Anexo C

Código fuente

En este apartado se muestra el código fuente implementado en Python para el análisis de la mortalidad de los adultos mayores por COVID-19 en la Ciudad de México, mediante el algoritmo de bosques aleatorios. Aunado a lo anterior, también se presenta el código fuente del árbol de decisión utilizado para la validación de la clasificación.

```
""" BOSQUES ALEATORIOS - CLASIFICACIÓN """


#Bibliotecas para el manejo básico de la fuente de datos.
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#Bibliotecas relacionadas con el algoritmo 'Bosques Aleatorios'.
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn import model_selection

#Bibliotecas para la grafica y reporte del estimador.
import graphviz
from sklearn.tree import export_graphviz
from sklearn.tree import export_text


""" Información general de la fuente de datos """

from google.colab import files
files.upload()
DatosCovid01 = pd.read_csv('DatosCovid_Codigo_18.csv')
DatosCovid01
DatosCovid01.info()
DatosCovid01[(DatosCovid01.DIED_DATE == 'Alive')]
DatosCovid01[(DatosCovid01.DIED_DATE == 'Dead')]


""" Definición de variables predictoras y variable clase """

print(DatosCovid01.groupby('DIED_DATE').size())

```

```
#Variables predictoras.
X = np.array(DatosCovid01[['SEX','PATIENT_TYPE','INTUBATED','PNEUMONIA','AGE','DIABETES',
                           'COPD','ASTHMA','INMUSUPPR','HYPERTENSION','OTHER_DISEASES',
                           'CARDIOVASCULAR','OBESITY','CHRONIC_RENAL','SMOKING',
                           'OTHER_CASE','ANTIGEN_RESULT','FINAL_CLASSIFICATION','ICU']])
pd.DataFrame(X)

#Variable clase.
Y = np.array(DatosCovid01[['DIED_DATE']])
pd.DataFrame(Y)


""" División de datos y aplicación del algoritmo """

#División de datos en subconjuntos aleatorios de tren y prueba.
X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y, test_size = 0.2,
                                                                    random_state = 0,
                                                                    shuffle = True)
pd.DataFrame(X_train)
pd.DataFrame(Y_train)

#Entrenamiento del modelo a partir de los datos de entrada.
ClassificationRF = RandomForestClassifier(random_state=0, max_depth=8, min_samples_leaf=2,
                                          min_samples_split=4)
ClassificationRF.fit(X_train, Y_train)


""" Clasificaciones """

#Etiquetas de clasificaciones.
Y_Clasificacion = ClasificacionBA.predict(X_validation)
pd.DataFrame(Y_Clasificacion)
Valores = pd.DataFrame(Y_validation, Y_Clasificacion)
Valores


""" Validación del modelo """

#Promedio de validación.
ClasificacionBA.score(X_validation, Y_validation)

#Matriz de clasificación.
Y_Clasificacion = ClasificacionBA.predict(X_validation)
Matriz_Clasificacion = pd.crosstab(Y_validation.ravel(), Y_Clasificacion, rownames=['Real'],
                                   colnames=['Classificación'])
Matriz_Clasificacion

#Reporte de clasificación.
print('Criterio: \n', ClasificacionBA.criterion)
print('\nVariables Importance: \n', ClasificacionBA.feature_importances_)
print("\nAccuracy: ", ClasificacionBA.score(X_validation, Y_validation))
print(classification_report(Y_validation, Y_Clasificacion))

#Importancia de variables para clasificación.
Importancia = pd.DataFrame({'Variable': list(DatosCovid01[['SEX','PATIENT_TYPE','INTUBATED',
                                                           'PNEUMONIA','AGE','DIABETES','COPD',
                                                           'ASTHMA','INMUSUPPR','HYPERTENSION',
                                                           'OTHER_DISEASES','CARDIOVASCULAR',
                                                           'OBESITY','CHRONIC_RENAL','SMOKING',
                                                           'OTHER_CASE','ANTIGEN_RESULT',
                                                           'FINAL_CLASSIFICATION','ICU']]),
'Importance': ClasificacionBA.feature_importances_}).sort_values('Importance', ascending=False)
Importancia

```

```
""" Visualización y reporte de estimador """

Estimador = ClasificacionBA.estimators_[20]
Estimador

#Objeto para visualizar estimador.
Elementos = export_graphviz(Estimador, feature_names = ['SEX','PATIENT_TYPE','INTUBATED',
                                                        'PNEUMONIA', 'AGE', 'DIABETES','COPD',
                                                        'ASTHMA','INMUSUPPR','HYPERTENSION',
                                                        'OTHER_DISEASES','CARDIOVASCULAR',
                                                        'OBESITY','CHRONIC_RENAL','SMOKING',
                                                        'OTHER_CASE','ANTIGEN_RESULT',
                                                        'FINAL_CLASSIFICATION','ICU'])
Arbol = graphviz.Source(Elementos)
Arbol

#Reporte con reglas del estimador.
Reporte = export_text(Estimador, feature_names = ['SEX','PATIENT_TYPE','INTUBATED','PNEUMONIA',
                                                  'AGE','DIABETES','COPD','ASTHMA','INMUSUPPR',
                                                  'HYPERTENSION','OTHER_DISEASES',
                                                  'CARDIOVASCULAR','OBESITY','CHRONIC_RENAL',
                                                  'SMOKING','OTHER_CASE','ANTIGEN_RESULT',
                                                  'FINAL_CLASSIFICATION','ICU'])
print(Reporte)

```

```
"""ÁRBOL DE DECISIÓN - CLASIFICACIÓN"""


#Bibliotecas para el manejo básico de la fuente de datos.
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#Bibliotecas relacionadas con el algoritmo 'Árbol de Decisión'.
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn import model_selection

#Bibliotecas para la grafica y reporte del arbol.
import graphviz
from sklearn.tree import export_graphviz
from sklearn.tree import plot_tree
from sklearn.tree import export_text


""" Información general de la fuente de datos """

from google.colab import files
files.upload()
DatosCovid01Arbol = pd.read_csv('DatosCovid_Codigo_18.csv')
DatosCovid01Arbol


""" Definición de variables predictoras y variable clase """

print(DatosCovid01Arbol.groupby('DIED_DATE').size())
```

```
35
36  #Variables predictoras
37  X = np.array(DatosCovid01Arbol[['SEX','PATIENT_TYPE','INTUBATED','PNEUMONIA','AGE','DIABETES',
38                                  'COPD','ASTHMA','INMUSUPPR','HYPERTENSION','OTHER_DISEASES',
39                                  'CARDIOVASCULAR','OBESITY','CHRONIC_RENAL','SMOKING',
40                                  'OTHER_CASE','ANTIGEN_RESULT','FINAL_CLASSIFICATION','ICU']])
41  pd.DataFrame(X)
42
43  #Variable clase
44  Y = np.array(DatosCovid01Arbol[['DIED_DATE']])
45  pd.DataFrame(Y)
46
47
48  """ División de datos y aplicación del algoritmo """
49
50  X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y,
51                                                                                  test_size = 0.2,
52                                                                                  random_state = 0,
53                                                                                  shuffle = True)
54  pd.DataFrame(X_train)
55  pd.DataFrame(Y_train)
56
57  #Entrenamiento del modelo a partir de los datos de entrada.
58  ClasificacionAD = DecisionTreeClassifier()
59  ClasificacionAD.fit(X_train, Y_train)
60
61
62  """ Clasificaciones """
63
64  #Etiquetas de clasificaciones.
65  Y_Clasificacion = ClasificacionAD.predict(X_validation)
66  pd.DataFrame(Y_Clasificacion)
67  Valores = pd.DataFrame(Y_validation, Y_Clasificacion)
68  Valores
69
70
71  """ Validación del modelo """
72
73  #Promedio de validación.
74  ClasificacionAD.score(X_validation, Y_validation)
75
76  #Matriz de clasificación.
77  Y_Clasificacion = ClasificacionAD.predict(X_validation)
78  Matriz_Clasificacion = pd.crosstab(Y_validation.ravel(), Y_Clasificacion, rownames=['Real'],
79                                     colnames=['Classificación'])
80  Matriz_Clasificacion
81
82  #Reporte de clasificación.
83  print('Criterion: \n', ClasificacionAD.criterion)
84  print('variable Importance: \n', ClasificacionAD.feature_importances_)
85  print("Accuracy", ClasificacionAD.score(X_validation, Y_validation))
86  print(classification_report(Y_validation, Y_Clasificacion))
87
88  #Importancia de variables para clasificación.
89  Importancia = pd.DataFrame({'Variable': list(DatosCovid01Arbol[['SEX','PATIENT_TYPE','INTUBATED',
90                                                                  'PNEUMONIA','AGE','DIABETES',
91                                                                  'COPD','ASTHMA','INMUSUPPR',
92                                                                  'HYPERTENSION','OTHER_DISEASES',
93                                                                  'CARDIOVASCULAR','OBESITY',
94                                                                  'CHRONIC_RENAL','SMOKING',
95                                                                  'OTHER_CASE','ANTIGEN_RESULT',
96                                                                  'FINAL_CLASSIFICATION','ICU']]),
97  'Importancia': ClasificacionAD.feature_importances_}).sort_values('Importancia', ascending=False)
98  Importancia
99
```

```
100
101 """ Visualización y reporte de estimador """
102
103 #Objeto para visualizar el Árbol.
104 Elementos = export_graphviz(ClasificacionAD, feature_names = ['SEX','PATIENT_TYPE','INTUBATED',
105                                                               'PNEUMONIA','AGE','DIABETES','COPD',
106                                                               'ASTHMA','INMUSUPPR','HYPERTENSION',
107                                                               'OTHER_DISEASES','CARDIOVASCULAR',
108                                                               'OBESITY','CHRONIC_RENAL','SMOKING',
109                                                               'OTHER_CASE','ANTIGEN_RESULT',
110                                                               'FINAL_CLASSIFICATION','ICU'],
111                                                               class_names = Y_Clasificacion)
112 Arbol = graphviz.Source(Elementos)
113 Arbol
114
115 #Reporte con reglas del estimador.
116 Reporte = export_text(ClasificacionAD, feature_names = ['SEX','PATIENT_TYPE','INTUBATED',
117                                                         'PNEUMONIA','AGE','DIABETES','COPD','ASTHMA',
118                                                         'INMUSUPPR','HYPERTENSION','OTHER_DISEASES',
119                                                         'CARDIOVASCULAR','OBESITY','CHRONIC_RENAL',
120                                                         'SMOKING','OTHER_CASE','ANTIGEN_RESULT',
121                                                         'FINAL_CLASSIFICATION','ICU'])
122 print(Reporte)
123
```

Referencias bibliográficas

Akinnuwesi, B. A., Fashoto, S. G., Mbunge, E., Odumabo, A., Metfula, A. S., Mashwama, P. et al. (2021). Application of intelligence-based computational techniques for classification and early differential diagnosis of COVID-19 disease. Data Science and Management, 4, 10-18.

Alves, V., Casemiro, F., De Araujo, B., De Souza, M., Silva, R., Tamires, F., Campos A. Gregori, D. (2021). Factors Associated with Mortality among Elderly People in the COVID-19 Pandemic (SARS-CoV-2): A Systematic Review and Meta-Analysis, International Journal of Environmental Research and Public Health, 15, 8008. DOI: 10.3390/ijerph18158008.

Cardoso, A., Talame, L., Amor, M. (2021). Aprendizaje automático aplicado a la pandemia del virus Covid-19 en Argentina. In XXIII Workshop de Investigadores en Ciencias de la Computación.

CDC-Centers for Disease Control and Prevention- (2020). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19)—United States, February 12–March 16, 2020. https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm. Accessed October 5, 2022.

Cortés, R., Dyer, D. (2021). Lineamiento para la estimación de riesgos del semáforo por regiones COVID-19. Secretaria de Salud, México.

Fulcher, J. (2008). Computational intelligence: an introduction. In Computational intelligence: a compendium, pp. 3-78. Springer, Berlin, Heidelberg.

Gobierno de la Ciudad de México. (2020). Criterios para las poblaciones en situación de vulnerabilidad que pueden desarrollar una complicación o morir por COVID-19 en la reapertura de actividades económicas en los centros de trabajo

Gobierno de México. (2022). Datos Abiertos de Bases Históricas. Dirección General de Epidemiología de la Secretaría de Salud. www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia. Last accessed in May 16, 2022.

Inca-Balseca, C. L., Paredes-Proaño, A. M., Cornejo-Reyes, P. J., Mena-Reinoso, Á. P. (2022). Eficiencia de modelos de predicción de COVID-19 usando curvas ROC y matriz de confusión. Dominio de las Ciencias, 8(2), 1442-1460.

INEGI. (20121). En México somos 126’014,024 habitantes: censo de población y vivienda 2020.

Kaplan, A., Haenlein, M. (2019). Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25.

Khan, I., Aslam, N., Aljabri, M., Aljameel, S., Kamaleldin, M., Alshamrani, F., Chrouf, S. (2021). Computational Intelligence-Based Model for Mortality Rate Prediction in COVID-19 Patients. International Journal of Environmental Research and Public Health, 18(12), 6429.

Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., Steinbrecher, M. (2016). Introduction to Computational Intelligence. Springer, London.

Kumar, G., Jain, S. Singh, U. (2021). Stock market forecasting using computational intelligence: A survey. Archives of Computational Methods in Engineering, 28(3), 1069-1101.

Leandro-Astorga, G., Calvo, I. B. (2021). Infección por COVID-19 en población adulta mayor: recomendaciones para profesionales. Revista Médica de Costa Rica y Centroamérica, 86(629), 44-50.

Martínez, R., Ramírez, N., Mesa, H., Suárez, I., Trejo, M.., León, P., Morales, S. (2009). Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, 9(2), 19-24.

Merino, R. F., Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases, (10), 165-189.

OMS-World Health Organization- (2023). Weekly epidemiological update on COVID-19, May 18, 2023.

Raj, J. S. (2019). A comprehensive survey on the computational intelligence techniques and its applications. Journal of ISMAC, 1(03), 147-159.

Rami, A., Malak, M., Aboulnour, M., Mirza, S., Alshobaiki, A. (2022). Classifying the Mortality of People with Underlying Health Conditions Affected by COVID-19 Using Machine Learning Techniques. Applied Computational Intelligence and Soft Computing. DOI: 10.1155/2022/3783058 (2022).

Rodríguez, Y. L., López, L. A. (2020). A propósito del artículo “COVID-19. De la patogenia a la elevada mortalidad en el adulto mayor y con comorbilidades. Revista Habanera de Ciencias Médicas, 19(4), 14.

Russell, S., Norvig, P. (2004). Inteligencia artificial: Un enfoque moderno. Pearson Education.

Secretaría de Salud. (2023). Informe Técnico Semanal COVID-19 México, Ciudad de México, Mayo 2, 2023.

Vega, J. A., Ruvalcaba, J. C., Hernández, I., Acuña, M. D., López, L. (2020). La salud de las personas adultas mayores durante la pandemia de COVID-19. Journal of Negative and No Positive Results, 5(7), 726-739.

Wang, X., Song, G., Yang, Z., Chen, R., Zheng, Y., Hu, H., Su, X., Chen, P. (2020). Association between ageing population, median age, life expectancy and mortality in coronavirus disease (COVID-19). Aging (Albany NY), 12(24), 24570-24578. DOI: 10.18632/aging.104193.