Predicción del grado de vulnerabilidad por COVID-19 en E.E.U.U.

Salazar Carazas, Renzo Osmar
PEAAP - DSRP
Lima, Perú
reocs10@gmail.com

Astete, Leonardo
PEAAP - DSRP
Lima, Perú
leonardo.astetar@gmail.com

Davila Rojas, Mireille
PEAAP - DSRP
Lima, Perú
mdavilarojas@gmail.com

Casusol Casma, Alexis Jose
PEAAP - DSRP
Lima, Perú
alexiscasusol30@gmail.com

***Abstract*—La pandemia ocasionada por la COVID-19 desencadenó un efecto devastador en la economía y salud en la población mundial, cuyas implicaciones sociales para los próximos años es aún incierta. En este trabajo se muestra como se creo un modelo de predicción con una técnica de inteligencia artificial y aprendizaje automático que nos permita determinar que factores conllevan a un mayor grado de vulnerabilidad en los pacientes con COVID-19. Se analizó la base de datos del sistema de vigilancia de casos COVID-19 que incluye datos a nivel individual reportados en Estados Unidos, incluyendo 11 rasgos (físicos,demográficos y clínicos) sobre cerca de 7 millones de casos positivos (más de 183 mil fallecidos), se emplearon técnicas de preparación y visualización de datos para el análisis explotarorio con el foco en identificar patrones que predigan un desarrollo fatal de la enfermedad. El modelo encontrado muestran un alto indíce de prediccción de casos de personas muy vulnerables al COVID-19, dicha información puede apoyar a la toma de decisiones y la planificación logística en los sistemas de salud.**

***Index Terms*—Covid-19, machine learning, artificial intelligence**

I. Introducción

La COVID-19 es la enfermedad causada por el nuevo coronavirus conocido como SARS-CoV-2. La OMS tuvo noticia por primera vez de la existencia de este nuevo virus el 31 de diciembre de 2019, al ser informada de un grupo de casos de neumonía vírica que se habían declarado en Wuhan (República Popular China).

El primer caso confirmado de la pandemia de COVID-19 en Estados Unidos se anunció el 21 de enero de 2020. Los CDC (Centers for Disease Control and Prevention) advirtieron que la transmisión generalizada de la enfermedad puede obligar a un gran número de personas a buscar hospitalización y otros servicios de salud, lo que puede sobrecargar los sistemas de salud [1].

A partir del 26 de marzo de 2020, Estados Unidos se convirtió en el país con más casos de COVID-19 en el mundo, superando a China. Posteriormente,el 11 de abril de 2020 se convirtió en el país con más muertes en el mundo, superando a Italia.

El segundo aumento de infecciones comenzó en junio de 2020, luego de restricciones relajadas en varios estados[2]. La propagación descontrolada de la comunidad llevó a algunas instalaciones médicas a rechazar a nuevos pacientes o comenzar a transferirlos.

Al cierre de noviembre con aproximadamente 180 mil fallecidos en poco más de 8 millones de infectados es la duodécima tasa más alta entre las naciones[3].

El gobierno de EEUU se ve desbordado por una pandemia fuera de control, es importante determinar las condiciones de pacientes que pueden conllevar mayores factores de vulnerabilidad para asi poder tomar acciones de prevencion y cuidado dirigidas a ciertos sectores sociodemograficos de la poblacion en EEUU.

Con base en la dinámica comportamental del COVID-19, se requieren soluciones prontas para el monitoreo, detección y diagnóstico de las enfermedades generadas por su causa El aprendizaje automático es una disciplina de la IA que se vale de algoritmos que permiten la identificación de patrones, efectuar predicciones, aprender de los datos y toma de decisiones.

El propósito de esta investigación ha sido adquirir información útil en base a la predicción del grado de vulnerabilidad de una persona con COVID-19 utilizando técnicas de IA. El propósito de la investigación es similar a la desarrollada por Li Yan et al.[4] pero los rasgos usados para describir los casos son diferentes: mientras en aquella se utilizan resultados de exámenes clínicos de laboratorio, en el presente trabajo se usan otros tipos de atributos, como la edad, el sexo, si el paciente presenta una condición clínica preexistente, razgos étnicos; también los datos utilizados son de origen distinto. Los datos utilizados han sido un Conjunto de datos de uso público de vigilancia de casos Covid-19 en EE.UU.[5]

Se han aplicado técnicas de aprendizaje automático para identificar los rasgos más importantes para realizar el pronóstico y extraído información para esta problemática,el cual puede ser útil para confirmar patrones de comportamiento ya conocidos o eventualmente generar algunos novedosos. Los resultados son el primer paso en la generación de información;su validación y utilidad solo es posible por parte de los especialistas en el campo de la medicina y gobierno que puedan tomar acciones ante la información presentada.

II. Métodos

Para desarrollar la prediccción de grado de vulnerabilidad de una persona se utilizaron los datos públicos de vigilancia de casos COVID-19 provistos por los CDC (Centers for Disease Control and Prevention) [5], donde se encuentra la información de varios millones de pacientes incluyendo

numerosos rasgos sobre estos, el diagnóstico realizado de la enfermedad y la fecha de muerte, en los casos de desarrollo fatal. Considerando el propósito de este trabajo los datos fueron depurados dejando de lado los registros con valores faltantes, como resultado se tiene una base de 8404990 de registros de los cuales se tiene 7910118 casos positivos a la Covid-19 y se sabe que 183577 de estos casos fallecen, lo cual indica que los datos tienen un alto desbalance, es decir, si nuestra clase de decisión fuese el fallecimiento está mucho menos representada en los datos cabe mencionar que el desbalance en los datos afecta los procesos de predicción, lo cual se tuvo en cuenta para realizar el proceso.

Cada caso queda conformado por 11 atributos presentados en la siguiente tabla:

TABLE I
DESCRIPCIÓN DE LOS ATRIBUTOS DEL DATASET

Nº	*Nombre del atributo*	*Descripción*
1	cdc_report_dt	Fecha en que se informó el CDC (Center for Disease Control and Prevention)
2	pos_spec_dt	Fecha de la primera recolección de muestras positivas
3	onset_dt	¿Cuál fue la fecha de inicio?
4	current_status	¿Cuál es el estado actual de esta persona?
5	sex	género
6	age_group	Categorías de grupos por edad
7	Race and ethnicity (combined)	Caso demográfico, razas y etnias
8	hosp_yn	¿Fue hospitalizado el paciente?
9	icu_yn	¿El paciente fue ingresado en una unidad de cuidados intensivos (UCI)?
10	death_yn	¿Murió el paciente como consecuencia de esta enfermedad?
11	medcond_yn	¿Tenían alguna condición médica subyacente y/o conductas de riesgo?

De los cuales observamos 3 atributos de fechas (cdc_report_dt, pos_spec_dt, onset_dt), 6 atributos nominales (sex, current_status, Race and ethnicity,hosp_yn,icu_yn,medcond_yn), un atributo númerico que corresponde a la edad y el atributo de decisión death_yn.

Como primera acción en el estudio, se emplearon técnicas de visualización de información para el análsis exploratorio de los datos y así mostrar la distribución de los datos según los casos analizados. Esta etapa de visualización es muy útil para entender los datos, se recomienda empezar por ella en los enfoques más actuales de ciencia de datos.

Despues de ello, con esta base de casos se realizó el proceso de descubrimiento de información, en el cual se ejecutaron las siguientes tareas principales:

- Analisis en el contexto de los EE.UU.
- Selección de rasgos importantes para predecir una evolución desfavorable del paciente.
- Adquisición de información en base al modelo de predicción.

El proceso de análisis en el contexto de los EE.UU. nos sirvió para analizar los datos y como estos fueron evolucionando en el tiempo comparándolos a otras fuentes de información.

El proceso de selección de atributos consiste en determinar cuáles de los atributos descritos en la Tabla 1 tienen una mayor incidencia para realizar un pronóstico eficaz. Existen diferentes métodos para realizar este proceso. En este trabajo se utilizó el método PCA (análisis de componentes principales)[6].

Como el propósito de este trabajo es predecir el grado de vulnerabilidad de una persona frente a la COVID-19 y teniendo en cuenta la naturaleza de nuestros atributos se propuso un método de arbóles de decisión para realizar dicha predicción.

III. RESULTADOS

A. Analisis Exploratorio

A continuación se presentan los resultados obtenidos del análisis exploratorio de los datos, comenzando con la visualización de la información más relevante, fundamentalmente relacionada con las Defunciones.

En el gráfico de la figura 1 muestra la cantidad de defunciones por rango de edad y sexo. Se visualiza como la tasa de mortalidad va en aumento conforme el rango de edad es mayor También podemos visualizar que en el rango de 70-79 años los varones fallecen en mayor proporción sin embargo de 80 a mas son las mujeres quienes fallecen en mayor cantidad.

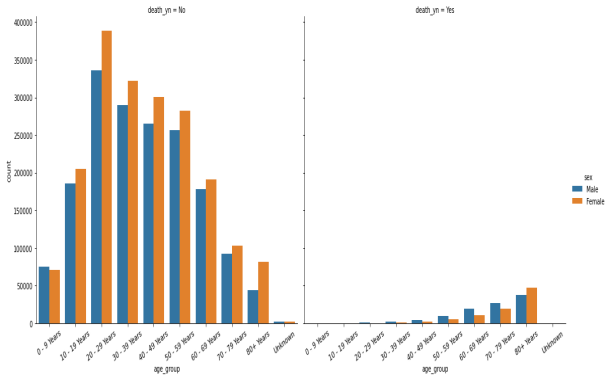

Fig. 1. Defunciones por rango de edad.

En la figura 2 muestra las defunciones por raza dividos en grupos como Blancos, Hispanos, Negros y äsitaticos. Se puede visualizar en gran medidad que los blancos son los que tienen mayor predisposición a fallecer, seguido de los negros pero esta información no es tan trascendental debido a que la población mayoritaria son blancos o en todo caso de raza desconocida.

A partir de los gráficos mostrados en las figuras 1 y 2, se puede observar una tendencia a aumentar la letalidad con la edad.

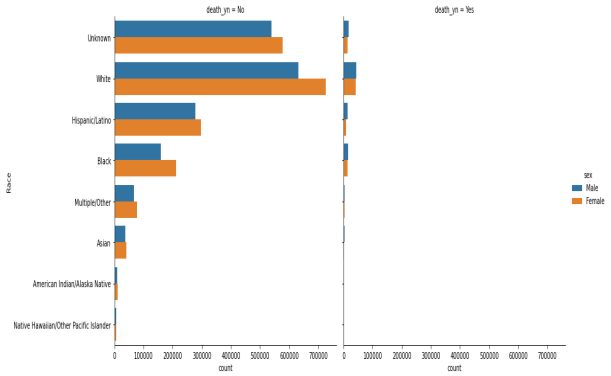

Fig. 2. Defunciones por raza.

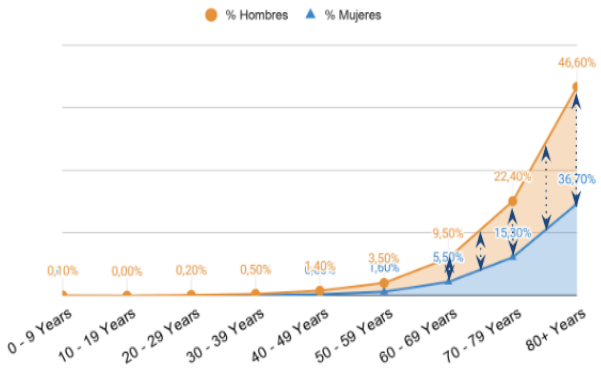

Fig. 4. Influencia de la edad y sexo en el % de fallecidos.

B. *Analisis en el contexto de los EE.UU.*

PLanteamos en función a los datos una curva que representa los casos reportados, los casos positivos al Covid y los fallecimientos.

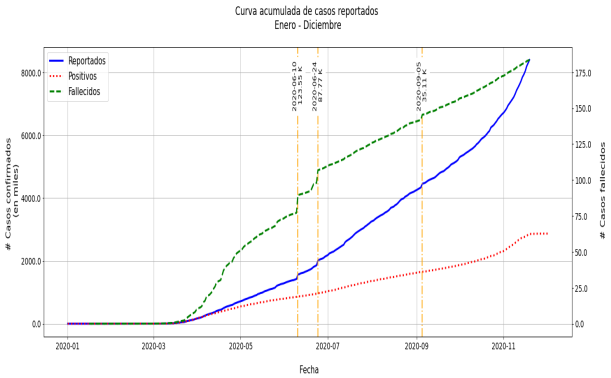

Fig. 3. Evolución de los casos positivos en el tiempo.

Se observa un crecimiento lineal del número de fallecidos (desde junio), mientras que los reportados crecen en curva. Además,la cantidad de positivos no sigue la tendencia (reducción de número exámenes realizados en los últimos meses. Lo que nos muestra la data es que va correlacionada a la crisis sanitaria que ocurre en los EE.UU.

C. *Selección de Atributos*

Despúes de realizar el análisis explotario de los datos y visualizar que las variables cdc_report_dt (Fecha en que se informó el CDC) y onset_dt (¿Cuál fue la fecha de inicio?) presentan muchos valores faltantes. Se decide a no tomarlas en cuenta para el estudio.

Analizamos algunos atributos que pueden influir como condiciones para aumentar el grado de vulnerabilidad de la persona frente a la COVID-19.

En la figura 4 Se observa un aumento en la brecha de las curvas por sexo del porcentaje de fallecidos a medida que pasamos los 50 años de edad, llegando a una diferencia máxima de 10% en el rango de 80 a más.

Podemos observar en la figura 5 la tasa de mortalidad condicional respecto a la variable medcond_yn donde el paciente tien alguna condición médica subyacente y/o conductas de riesgo.

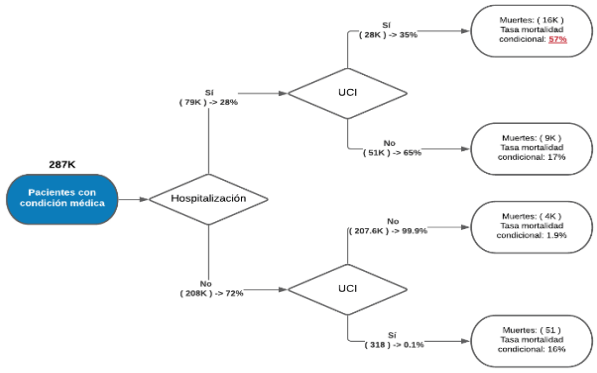

Fig. 5. Mortalidad condicional respecto a condición médica precedente.

La selección de los atributos para nuestro modelo son current_status, sex, age_group, hosp_yn ,icu_yn, medcond_yn, Race. Finalmente a estos atributos se decide aplicar un PCA(Análisis de componentes principales), con PCA Para más del 86 % de la muestra sugiere que esta se puede representar por los 5 primeros componentes,siendo los rasgos más importantes.

D. *Modelo de predicción*

Random forest. En Random forest, cada árbol en el conjunto se construye a partir de una muestra extraída con reemplazo (es decir, una muestra de arranque) del conjunto de entrenamiento.

Además, al dividir cada nodo durante la construcción de un árbol, la mejor división se encuentra entre todas las características de entrada.

El propósito de estas dos fuentes de aleatoriedad es disminuir la varianza del estimador forestal. De hecho, los árboles de decisión individuales suelen presentar una gran variación y tienden a sobreajustarse. La aleatoriedad inyectada en los bosques produce árboles de decisión con errores de

predicción algo desacoplados. Al tomar un promedio de esas predicciones, algunos errores pueden anularse. Los bosques aleatorios logran una variación reducida al combinar diversos árboles, a veces a costa de un ligero aumento del sesgo. En la práctica, la reducción de la varianza suele ser significativa, por lo que se obtiene un modelo mejor en general.

Después de entrenar el modelo obtenemos una precisión del 88% de acierto en predecir el grado de vulnerabilidad de la persona frente al COVID-19.

E. Conclusiones

Resulta de interes que el análisis de los atributos ratitifica que algunos atributos inciden con mayor fuerza en la vulnerabilidad de la persona como es la edad.

El modelo encontrado muestra un alto indíce de prediccción de posibles casos muy vulnerables dicha información puede ayudar a mejorar la comprensión de la enfermedad y muestran las capacidades de las técnicas de inteligencia artificial para analizar datos desde diferentes perspectivas, como apoyo a la toma de decisiones y la planificación logística en los sistemas de salud.

REFERENCES

[1] CDC (26 de marzo de 2020). "Coronavirus Disease 2019 (COVID-19)" Situation Summary. Centers for Disease Control and Prevention (en inglés estadounidense). Consultado el 7 de abril de 2020.

[2] "The Trump administration is eyeing a new testing strategy for coronavirus, Anthony Fauci says" https://www.washingtonpost.com/news/powerpost/paloma/the-health-202/2020/06/26/the-health-202-the-trump-administration-is-eyeing-a-new-testing-strategy-for-coronavirus-anthony-fauci-says/5ef4f629602ff1080718f308/

[3] "Mortality Analyses". Johns Hopkins Coronavirus Resource Center (en inglés). Consultado el 20 de diciembre de 2020.

[4] Li Yan et al. "Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan". 2020. DOI: https://doi.org/10.1101/2020.02.27.20028027

[5] "COVID-19 Case Surveillance Public Use Data" https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

[6] Skiena S.S. "The Data Science Design Manual". Springer. 2017.

[7] María Matilde García Lorenzo y Yanela Rodríguez y Alejandro Ramón Hernández y Beatriz Bello García y Yaima Filiberto y Alejandro Rosete y Yaile Caballero Mota y Rafael Bello, "Adquisición de conocimiento sobre la letalidad de la COVID-19 mediante técnicas de inteligencia artificial",Anales de la Academia de Ciencias de Cuba, volume 10,2020, url = http://www.revistaccuba.sld.cu/index.php/revacc/article/view/891,