Subgroup classification model identifying the most influential factors in the mortality of patients with COVID-19 using data analysis

1st Carlos Peña
Universidad Politécnica Salesiana
Cuenca, Ecuador
cpenaf@est.ups.edu.ec

2nd Florencio Peralta
Universidad Politécnica Salesiana
Cuenca, Ecuador
cperaltab1@est.ups.edu.ec

3rd Remigio Hurtado
Universidad Politécnica Salesiana
Cuenca, Ecuador
rhurtadoo@ups.edu.ec

***Abstract*—This research assesses the health conditions of the people in the study and determines the reason why a person dies after being infected with COVID-19. In this study, 538 sample groups that provided medical data from people in different locations were analyzed. The biggest challenge in this study was to carry out 2 different criteria within the same data set to conclude that the mortality of the persons inside a group depends more than anything on the age of the person at risk and the presence of one or more other health disorders of the primary disease, which in this case is COVID-19.**

For this study, the public data set "COVID analytics" was used, which provided all the necessary medical information and the classification of the groups, which are then interpreted as useful labels to better deduce the degree of mortality of the affected person. After completing the data analysis, it is determined that the factors that aggravate the condition of a patient with COVID-19 are: hypertension, advanced age and any other disease.

Covid-19, Linear regression, Quadratic regression, Polynomial regression, Clustering, Multi-linear regression

I. Introducción

Para comenzar con nuestro estudio es primordial tener conocimiento del estado actual del tema que se va a desarrollar, es por eso que empezaremos primero explicando la problemática del COVID-19, luego explicaremos todos los antecedentes del estudio para poder empezar el procesamiento de los datos.

II. Problemática actual por la pandemia COVID-19

El COVID-19 es una enfermedad infecciosa causada por el coronavirus [17], el cual atacó al mundo de manera desprevenida; teniendo como epicentro el país de China en la provincia de Wuhan en diciembre del 2019, esta pandemia ha afectado a la mayor parte de países del mundo, dejando a su paso varios muertos y contagiados.

La importancia de esta investigación radica en informar a la población sobre los factores más influyentes correlacionados con la mortalidad debido al COVID-19. Se realizó este artículo con el fin de analizar un conjunto de datos públicos [2] que contenga grupos de análisis provenientes de China, ya que este fue el primer país en estar expuesto al virus.

A. Limpieza de datos

La limpieza de datos es uno de los procesos más importantes cuando se requiere asegurar la calidad de los mismos. Aquí se eliminan datos inconsistentes o nulos que pueden afectar al análisis. Este tratamiento de ruido mejora de una manera más adecuada los procesos del aprendizaje de máquina [12].

B. Conceptos previos sobre aprendizaje de máquina

Uno de los principales objetivos en utilizar el aprendizaje de máquina, es poder desarrollar un sistema que pueda cambiar su comportamiento de manera automática en base a su experiencia.

Existen varios tipos de aprendizaje de máquina, entre los principales tenemos al supervisado y al no supervisado:

Aprendizaje supervisado: este tipo de método tiene como objetivo la predicción de variables categóricas basándose en observaciones pasadas mejorando de esta manera el conocimiento que se obtiene [16].

Aprendizaje no supervisado: su objetivo principal es el análisis y clasificación de datos en grupos "clusters" y así como la reducción de dimensionalidad de un conjunto de datos del cual no se conoce su salida [7]. En este artículo se usa **clustering** [20] como método de clasificación con la ayuda de K-means [4] para saber su K óptimo, y así poder tener un mejor análisis partiendo de grupos creados a partir de los subgrupos analizados en el conjunto de datos original. Se utiliza un análisis **PCA** para corroborar la decisión del K óptimo mediante un número de componentes igual a K.

C. Datos orientados a la medicina

Al estudiar el conjunto de datos encontramos variables sobre padecimientos que ya han sido referenciados en otros trabajos similares, reforzando la idea de cuales son factores que se consideran mortales al contraer la enfermedad del COVID-19, entre los más destacables tenemos:

Cualquier comorbilidad, hace referencia a cuando se encuentra más de una enfermedad en la misma persona [18]. Esta presencia puede constituir varios factores de riesgo potenciales para los resultados más graves. La edad [13], el virus afecta mayormente a las personas de avanzada edad, debido a que su sistema inmunológico se ha visto comprometido por enfermedades previas. Enfermedades cardiovasculares [14], ya que estas patologías están relacionadas con las vías respiratorias. La diabetes [9], la hiperglucemia es capaz de aumentar la gravedad de las infecciones virales. Cualquier tipo de cáncer [10], las personas que lo padecen e incluso aquellos que lo han sobrevivido tienen un sistema inmunitario que se considera más débil en comparación con un adulto sano promedio.

D. Ventajas de usar un aprendizaje no supervisado

Una de las mayores ventajas de este proceso es que puede distinguir entre la inducción predictiva y descriptiva. La inducción predictiva tiene como objetivo principal el poder descubrir el conocimiento necesario para clasificar o predecir datos. La inducción descriptiva, cuyo principal objetivo es extraer el conocimiento más influyente, sin la necesidad de conocer las etiquetas del conjunto de datos [5].

En la sección III se revisan las investigaciones que se han realizado con respecto a la problemática planteada para tener un mejor enfoque; después de realizar las comparaciones, se procede a efectuar los procesos, los mismos se pueden ver en la secciónIV. Luego, se analizan los resultados más importantes en la sección VI. Finalmente, se revisan las conclusiones y propuestas para futuros trabajos que tendrán como base esta investigación, esto, en la sección VII.

III. Trabajos relacionados

Existen varios estudios que se enfocan en descubrir que factores son los que más perjudican a el estado actual de una persona contagiada con COVID-19, el presente artículo tiene la finalidad de enfocarse en las poblaciones que fueron afectadas desde un inicio cuando ocurrió el brote de esta pandemia; en este caso, los grupos más correlacionados con la mortalidad [3].

En la actualidad el mundo intenta descubrir cómo actúan las enfermedades que más agravan el estado de salud de una persona que haya contraído COVID-19, esto impulsó a que se realizaran muchos estudios que intentan predecir y analizar las distintas variables que se correlacionan con el coronavirus, provocando una mayor tasa de mortalidad. Como algunos métodos tenemos:

Mediante un análisis con clustering se logra determinar 3 grupos de turistas chinos contagiados con COVID-19, gracias a esto se pudo realizar acciones de vigilancia y medidas de respuesta óptimas para evitar la propagación del virus [19]. La aplicación de la regresión polinomial para determinar la caracterización de la curva del COVID-19, esta caracterización permite ver las curvas de casos de muertes y de recuperados del COVID-19 [8]. Evaluación de la asociación a nivel de país entre la tasa de natalidad y el número acumulado de infectados mediante regresión lineal [11]. El uso de radiografías de tórax para aplicarlas en tres modelos diferentes basados en redes neuronales convolucionales (ResNet50, InceptionV3 e Inception-ResNetV2) para la detección de COVID-19, en donde la más acertada se obtiene al utilizar el modelo ResNet50 [15]. Uso de datos epidemiológicos, clínicos y genéticos para comprender mejor el patrón de propagación viral, mejorando aún más la velocidad y precisión del diagnóstico, desarrollando nuevos enfoques terapéuticos efectivos y potencialmente identificar a las personas más susceptibles a contagiarse en función de características genéticas y fisiológicas [1]. Análisis de datos sociodemográficos y epidemiológicos para facilitar la planificación de estrategias para reducir la transmisión de la enfermedad [6].

Combinar características de datos recolectados durante un día que sean similares a otros días (basada en la región) utilizando Xgboost, K-Means y redes neuronales de memoria a largo y corto plazo (LSTM) para construir un modelo de predicción (es decir, K-Means-LSTM) para la previsión de casos de COVID-19 a corto plazo en el estado de Louisana, EE. UU [21]. Recopilación de datos epidemiológicos y clínicos de personas con COVID-19 confirmado para evaluar las interacciones y los posibles modos de transmisión del coronavirus, de estos se obtuvieron 3 grupos los cuales estaban epidemiológicamente vinculados a los tres primeros grupos de transmisión local en Singapur concluyendo que el virus es transmisible en entornos comunitarios [19].

En la siguiente sección, se plantea el método que hemos propuesto para el análisis de los factores más influyentes partiendo desde dos enfoques distintos.

IV. Método Propuesto

En esta sección se plantea la estructura del proceso 1 que hemos seguido para el análisis del conjunto de datos, con el objetivo de encontrar las variables más correlacionadas con la mortalidad en base a determinados grupos y provincias.

Hemos creado un repositorio en GitHub: https://github.com/Charly1590/Classification-of-covid-mortality-by-groups-and-by-provinces donde se encuentran todos los recursos utilizados y requeridos para la realización de este estudio.

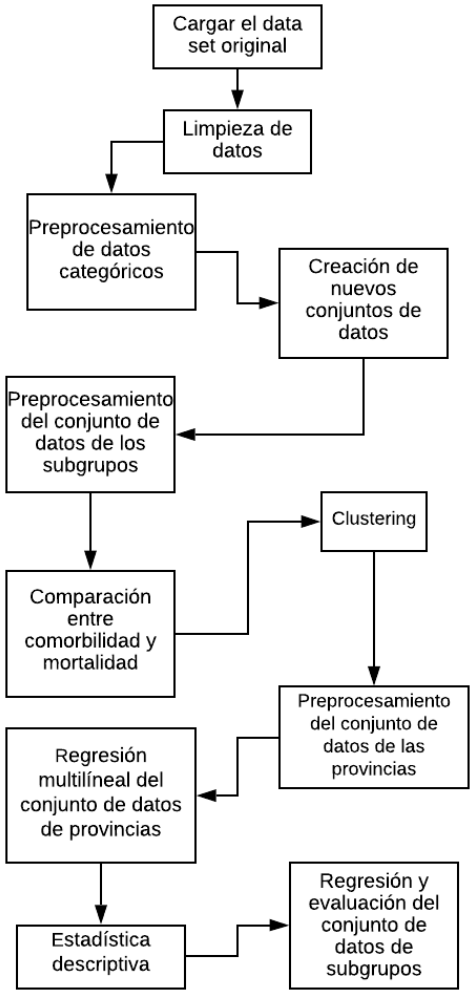

Fig. 1. Proceso del modelo propuesto

A. Limpieza de datos

Ya hemos expuesto la importancia de la eliminación del ruido, es por eso que hemos seguido el siguiente proceso: se eliminaron los datos no referentes directamente al COVID-19, para los datos categóricos nominales se cambiaron las celdas nulas por el valor de la media de la variable analizada, por otra parte, en los datos categóricos nominales se cambiaron los datos nulos por la etiqueta "Unknown" esto debido a la incertidumbre que genera.

B. Preprocesamiento de datos categóricos

Una vez que hemos tratado el conjunto de datos para que quede libre de ruido, clasificamos los datos en categóricos nominales y ordinales.

Se aplica estandarización para los datos categóricos ordinales teniendo así datos más limpios y generalizados, y para los categóricos nominales se realiza codificación para que el agrupamiento de datos se cumpla de manera correcta.

C. Creación de nuevos conjuntos de datos

A partir del conjunto de datos preprocesado se crea uno nuevo donde consta la media de los datos en función de los subgrupos; el mismo proceso se realiza para el agrupamiento de los datos por provincia.

D. Procesamiento y análisis del conjunto de datos de los subgrupos

En base al conjunto de datos relacionados a grupos específicos se realiza nuevamente la estandarización de datos con la excepción de la variable mortalidad ya que la hemos tomado como salida. También se vuelve a realizar el proceso de categorización aplicado a la variable de los subgrupos. Después de realizar el preprocesamiento del conjunto de datos se obtienen las correlaciones de las variables con respecto a la mortalidad, y así saber que variables son las que tienen un mayor impacto en la salida. Se realiza la elección de un K óptimo a través del método del codo con base a la distorsión, y se procede para aplicar el método de clustering. Posteriormente en base a la distribución de usuarios por cluster se evalúa la correlación con la salida.

E. Procesamiento del conjunto de datos de las provincias

Para este estudio se carga otro conjunto de datos relacionado específicamente a provincias, generado previamente, igualmente se divide en variables ordinales, nominales y se designa como salida a la variable mortalidad.

Luego de esto se aplica un MinMaxScaler a los datos para un mejor procesamiento de los mismos. Se aplica una regresión multilineal usando el conjunto de datos de Train esto es para el entrenamiento, y luego se realizan las predicciones comparando los resultados con el conjunto de datos de Test.

En la sección de diseño de experimento V se presentan las medidas de calidad y la descripción del conjunto de los datos utilizados en esta investigación.

V. Diseño de experimentos

En esta sección se muestran la descripción general de los conjuntos de datos y los parámetros de las técnicas utilizadas en nuestro modelo.

A. Parámetros de las técnicas utilizadas

En esta tabla se pueden ver los parámetros más óptimos obtenidos después de una serie de pruebas dentro de nuestro modelo.

TABLE I
Parámetros del Proceso

Técnica	Se utiliza
Clustering	4 grupos
PCA	4 Componentes
Regresión Polinomial	Función polinómica

B. Información de los conjuntos de datos

Todos los experimentos realizados fueron llevados a cabo usando el conjunto de datos de COVID19 Analytics y a partir de este, se generaron dos conjuntos de datos, de los cuales hacemos uso en nuestro experimento. En la tabla II se muestran el número de grupos de pacientes analizados en cada conjunto de datos.

TABLE II
INFORMACIÓN BÁSICA DEL CONJUNTO DE DATOS

Dataset	Número de grupos	Número de pacientes
COVID19 analytics clinical data	538	170918
csvDatasetGrupos	242	113239
csvDatasetProvincias	58	14190

C. Descripción general del conjunto de datos

En las tablas III y IV, se muestran los casos positivos y negativos respectivamente, los cuales dan a conocer cómo se relacionan estos factores con las variables estudiadas y de las cuáles nos hemos guiado para encaminar el de nuestro modelo de análisis.

TABLE III
SUBGRUPOS DE CASOS POSITIVOS DEL CONJUNTO DE DATOS CSVDATASETGRUPOS

Tamaño de la población de estudio	Edad Media	Porcentaje de Comorbilidad	Mortalidad
46	74,97	0,503	0,09
78,6	57	0,979	0,26
50,7515906	75	0,6133	0,01
58,65	17,5	0,52	0,16995662
50,6171969	110	0,65734689	0,44663774
53,2	75,4	0,49032033	0,16995662

TABLE IV
SUBGRUPOS DE CASOS NEGATIVOS DEL CONJUNTO DE DATOS CSVDATASETGRUPOS

Tamaño de la población de estudio	Edad Media	Porcentaje de Comorbilidad	Mortalidad
52,8	79,5	0,5665	0,16995662
73,8	28	1	0
50,8	26,1	0,49032033	0,16995662
50,1	37,59	0,49032033	0,16995662

En la tabla V mostramos las provincias/estados con el mayor número de personas afectadas dentro del estudio lo cual nos ayuda a fijar los estudios realizados al primer brote de COVID-19.

Entre los resultados, presentados en la sección VI, se expone toda la información resultante de la investigación que hemos venido realizando; se interpreta la misma para darle un sentido y contribuir así con conocimiento que ayude a enfrentar esta situación.

TABLE V
PROVINCIAS CON EL MAYOR NÚMERO DE HABITANTES ESTUDIADOS

Provincia/ Estado	Tamaño de la población de estudio	Edad Media	Porcentaje de Comorbilidad	Mortalidad
Henan	524	45	0,49032033	0,16995
Hunan	246,333333	48,270	0,46021355	0,14663
Shandong	219,6	51,369	0,49032033	0,13596
Hubei	206,857143	56,714	0,5496	0,08427
Yichang	197	55,94	0,49032033	0,14
Guizhou	162	37	0,49032033	0,16995
Zhejiang	151,25	42,957	0,41148146	0,04248
Shanghai	146,153846	50,779	0,29479087	0,1469
Jiangsu	145	43,202	0,51982423	0

VI. RESULTADOS

En esta sección, se revisan los resultados obtenidos, así como el rendimiento en las fases específicas de modelo de análisis. Además, se interpretan los resultados en base a: las variables, sus relaciones y el impacto de las mismas en los pacientes.

A. Resultados entre cualquier comorbilidad y mortalidad

En la figura 2, se muestra la relación entre cualquier comorbilidad y el riesgo de mortalidad que genera; cabe resaltar que, el eje de las "X" representa el porcentaje de pacientes en dicho grupo, se resalta además, que las personas que presentan enfermedades secundarias al COVID-19 generan una mayor mortalidad, es por ello que hemos realizado distintos tipos de regresiones para intentar predecir el riesgo de mortalidad de un determinado grupo de pacientes teniendo como resultado un error absoluto de 0.15 con una regresión polinomial.

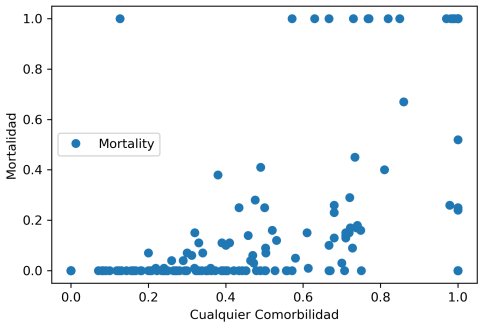

Fig. 2. Variable de cualquier comorbilidad vs. variable de la mortalidad

B. Resultados del clustering

También en nuestro modelo de análisis contemplamos el método del clustering, para ello utilizamos el método del codo para encontrar el mejor número de grupos con base a la distorsión, tal como se plantea en la figura 3. Así se determina que el mejor número de grupos es 4.

A fin de validar el proceso realizado por el clustering, se evalúa al conjunto de datos, también con una reducción de dimensionalidad a través del PCA. Así que se realiza este

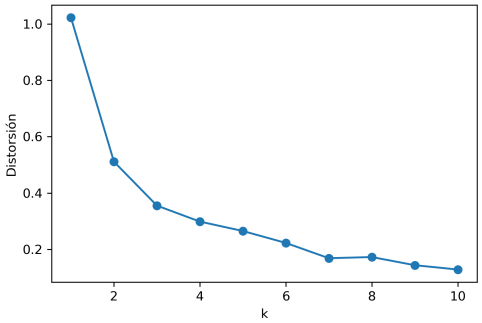

Fig. 3. Proceso del método del codo para saber el K óptimo

nuevo proceso a través del mismo criterio del método del codo, en este caso para seleccionar el número correcto de componentes. Lo anterior mencionado se lo expone en la figura 4 y 5.

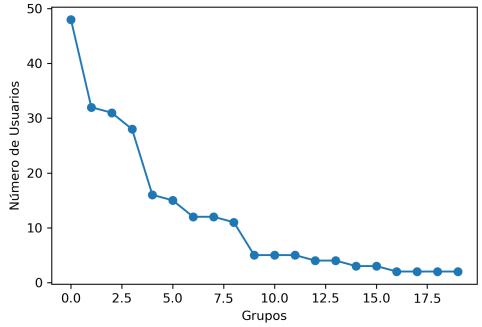

Fig. 4. Distribución de usuarios por grupo

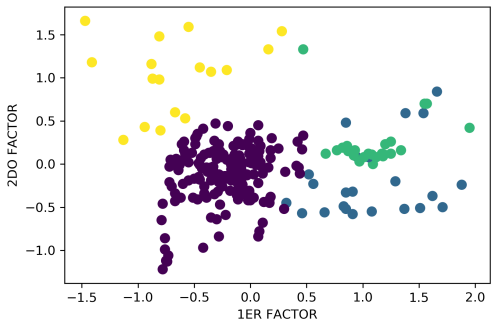

Fig. 5. Grupos con clustering

C. Resultados del análisis de provincias

Hemos realizado también un análisis predictivo para los grupos referentes a las provincias. En la tabla VI se muestran las medidas de calidad resultantes.

TABLE VI
MEDIDAS DE CALIDAD DE PREDICCIÓN PARA LAS PROVINCIAS

Medida	Resultado
MAE	0.06328
MSE	0.00632
RMSE	0.07951

D. Subgrupos con más correlación con el riesgo de mortalidad

En la tabla VII y VIII, presentamos las correlaciones existentes entre los subgrupos más destacables y su mortalidad. Se asegura que, mientras mayor es la mortalidad del grupo, mayor es la prioridad con la que deberán ser tomados en cuenta.

TABLE VII
SUBGRUPOS CON MÁS CORRELACIÓN AL RIESGO DE MORTALIDAD

Subgrupo	Correlación Mortalidad
Edad 91-100	1
AGI III-IV	1
CCRT	0,55
AGI II	0,54
Edad 81-90	0,52
Lesiones cardiacas	0,51
ARDS	0,50
AGI I	0,47
ARDS	0,50
Edad 61-70	0,29
Admitidos a cuidados intensivos	0,26
Basado en contacto	0,16

TABLE VIII
VARIABLES CON MÁS CORRELACIÓN AL RIESGO DE MORTALIDAD

Variables	Correlación Mortalidad
Cualquier cormovilidad	0,620151
Edad Media	0,539737
Hipertensión	0,403877
Cáncer	0,296525
Diabetes	0,286573
Enfermedad Cardiovascular	0,253761
Enfermedad Cerebrovascular	0,253761
Conteo de glóbulos blancos (Media)	0,2367
Fiebre	0,100683
Respiración Corta	0,199727

E. Rendimiento del Modelo

Para el análisis hemos utilizado una separación de 70-30 para los tamaños de conjuntos de entrenamiento y pruebas, respectivamente.

TABLE IX
ERRORES OBTENIDOS DE CADA TIPO DE REGRESIÓN, ENTRE CUALQUIER COMORBILIDAD Y MORTALIDAD

Variables	Tipo de regresión		
	Lineal	Cuadrática	Polinomial
MAE	0.1915	0.1617	0.1560
MSE	0.0729	0.0552	0.0567
RMSE	0.2700	0.2351	0.2381

TABLE X
ERRORES OBTENIDOS DE CADA TIPO DE REGRESIÓN, ENTRE EDAD MEDIA Y MORTALIDAD

Variables	Tipo de regresión		
	Lineal	Cuadrática	Polinomial
MAE	7.1396	7.1211	7.0364
MSE	83.0179	82.7738	81.0739
RMSE	9.1114	9.0980	9.0041

A continuación, se muestran las conclusiones sobre las variables analizadas y trabajos a futuro que podrían desarrollarse tomando como base la presente investigación.

VII. CONCLUSIONES Y TRABAJO FUTURO

En esta investigación se logra demostrar que a partir de los 60 años la tasa de mortalidad aumenta considerablemente, así mismo, si el paciente tiene alguna enfermedad o conflicto pulmonar el riesgo de mortalidad es más crítico. Acerca del análisis de los médicos residentes, se determina que son un grupo vulnerable, esto debido a que tienen un contacto prolongado con personas contagiadas; por lo tanto, su riesgo de contagio aumenta.

En conclusión, con este procedimiento se observa que las variables más correlacionadas con la mortalidad son: la edad media, hipertensión, díabetes, enfermedades cardiovasculares cualquier tipo de cáncer y un bajo nivel de glóbulos blancos.

Como trabajos futuros proponemos y motivamos a tomar en cuenta los factores que hemos resaltado en este estudio para profundizar más acerca de su relación con la mortalidad. Además, se incentiva a proponer, y complementar nuevos modelos o técnicas para mejorar el proceso de hemos planteado. También, se puede tomar como un grupo de análisis los grupos de personas fallecidas, con el fin de encontrar factores realmente influyentes y críticos que llevaron a dichas personas a perecer.

BIBLIOGRAFÍA

[1] Ahmad Alimadadi et al. *Artificial intelligence and machine learning to fight COVID-19*. 2020.

[2] Dimitris Bertsimas et al. *An Aggregated Dataset of Clinical Outcomes for COVID-19 Patients*. 2020. URL: http://www.covidanalytics.io/dataset_documentation.

[3] Rodolfo Bojorque, Remigio Hurtado, and Andrés Inga. "A comparative analysis of similarity metrics on sparse data for clustering in recommender systems". In: *International Conference on Applied Human Factors and Ergonomics*. Springer. 2018, pp. 291–299.

[4] Cristina Garcıa Cambronero and Irene Gómez Moreno. "Algoritmos de aprendizaje: knn & kmeans". In: *Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid* 23 (2006).

[5] CJ Carmona et al. "Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes". In: *Proceedings of the VII Simposio Teorıa y Aplicaciones de Mineria de Datos*. 2015, pp. 685–694.

[6] Viswa Chandu. "Identification of spatial variations in COVID-19 epidemiological data using K-Means clustering algorithm: a global perspective". In: *medRxiv* (2020).

[7] Margarita Gallardo Campos. "Aplicación de técnicas de clustering para la mejora del aprendizaje". MA thesis. 2009.

[8] Gabriel El'ıas Chanch'ı Golondrino, Wilmar Yesid Campo Mu noz, and Luz Marina Sierra Martinez. "Aplicaci'on de la regresi'on polinomial para la caracterizaci'on de la curva del COVID-19 en Colombia, mediante t'ecnicas de machine learning". In: *Investigaci'on e Innovaci'on en Ingenier'ıas* 8.2 (2020), pp. 87–105.

[9] Weina Guo et al. "Diabetes is a risk factor for the progression and prognosis of COVID-19". In: *Diabetes/metabolism research and reviews* (2020), e3319.

[10] Timothy P Hanna, Gerald A Evans, and Christopher M Booth. "Cancer, COVID-19 and the precautionary principle: prioritizing treatment during a global pandemic". In: *Nature Reviews Clinical Oncology* 17.5 (2020), pp. 268–270.

[11] Chris Kenyon. "Flattening-the-curve associated with reduced COVID-19 case fatality rates-an ecological analysis of 65 countries". In: *Journal of Infection* 81.1 (2020), e98–e99.

[12] Beatriz López Porrero. "Limpieza de datos: reemplazo de valores ausentes y estandarización". PhD thesis. Universidad Central "Marta Abreu" de Las Villas. Facultad de Matemática …, 2011.

[13] Andrés Losada-Baltar et al. "Diferencias en función de la edad y la autopercepción del envejecimiento en ansiedad, tristeza, soledad y sintomatologıa comórbida ansioso-depresiva durante el confinamiento por el COVID-19". In: *Revista Española de Geriatrıa y Gerontologıa* (2020).

[14] Mandeep R Mehra et al. "Cardiovascular disease, drug therapy, and mortality in COVID-19". In: *New England Journal of Medicine* (2020).

[15] Ali Narin, Ceren Kaya, and Ziynet Pamuk. "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks". In: *arXiv preprint arXiv:2003.10849* (2020).

[16] Bac Nguyen Cong, Jorge Luis Rivero Pérez, and Carlos Morell. "Aprendizaje supervisado de funciones de distancia: estado del arte". In: *Revista Cubana de Ciencias Informáticas* 9.2 (2015), pp. 14–28.

[17] *OMS*. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses. Accessed: 2020-07-24.

[18] Thais M Plasencia-Urizarri, Raúl Aguilera-Rodrıguez, and Luis E Almaguer-Mederos. "Comorbilidades y gravedad clınica de la COVID-19: revisión sistemática

y meta-análisis". In: *Revista Habanera de Ciencias Médicas* 19 (2020).

[19] Rachael Pung et al. "Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures". In: *The Lancet* (2020).

[20] Antonio F Gómez Skarmetta. "Modelado difuso de sistemas mediante aprendizaje por clasificación con técnicas de agrupamiento (clustering)". PhD thesis. Universidad de Murcia, 1995.

[21] Shashank Reddy Vadyala et al. "Prediction of the Number of COVID-19 Confirmed Cases Based on K-Means-LSTM". In: *arXiv preprint arXiv:2006.14752* (2020).