Análisis Estadístico y Predictivo de los Factores Relacionados con el COVID-19

***Abstract*—We propose a method of data analysis of a Dataset that contains data related to the COVID-19. This article shows step by step the process followed to make a transformation of the data, clustering and a linear regression. The data presented contains cases of flu and coronavirus detected within the last months. We use this statistical analysis to find key characteristics of the cases, finding groups with clear differences and correlations between them. The main groups containing high number of COVID and flu showing certain similarities between the symptoms like fever, dyspnea, and cough. While the correlations showing an emphasis in other variables unrelated to the symptoms like being in a high-risk zone. We train the data and then we try to predict new cases and then comparing the real diagnosis. The results obtained can be reflected upon and some of them show us some interesting information.**

***Index Terms*—Dataset, clustering, regresión, COVID-19, atributos, relaciones**

I. Introducción

Ciertamente los tiempos que estamos viviendo son cuanto menos inquietantes, la pandemia sin duda ha cambiado nuestras vidas en los últimos meses y nos guste o no estará presente por tiempo considerable. Nuestro deber como ciudadanos es acatar las normas de distanciamiento social y el uso de mascarilla, pero también como estudiante de Ciencias de la Computación siento la necesidad de tratar hacer un aporte sin importar lo pequeño que resulte. A lo largo de este semestre como estudiantes hemos adquirido una serie de conocimientos estadísticos que nos han ayudado a tener una mejor visión de la importancia de los datos en todas las ramas de la ciencia. La importancia de sacar el mayor provecho a la cantidad de datos acerca del COVID-19 que encontramos, utilizando distintas técnicas que van desde simplemente describir los datos hasta modelos de Machine Learning.

Que estos conjuntos de datos estén a nuestro fácil alcance, abiertos completamente al público, facilita la creación de diversos sistemas relacionados con dichos datos, cuyos objetivos pueden ser muy diversos. En nuestro caso los datos adquiridos nos han facilitado el entendimiento de ciertos aspectos relacionados a la enfermedad, incluso sin entrar en mayores detalles de terminología médica. Comprender con ciertas variables influyen en el aumento de casos del virus diariamente, es cuanto menos interesante de echar un vistazo por motivos meramente informativos. Modelos como este realizado por mi persona pueden surgir a diario, ayudando a comprender mejor los fenómenos e incluso puede llegar a salvar algunas vidas.

El artículo está estructurado en seis secciones y sus referencias. Como primer punto detalamos el estado del arte. Como segundo punto tenemos el método propuesto seleccionado para el uso de los datos, esto con sus respectivos diagramas y algoritmos. El tercer punto se dedica a describir el conjunto de datos, el porcentaje usado para test y los parámetros utilizados para el análisis. En el punto cuatro concretamente presentamos los resultados en forma de figuras y tablas con su respectiva descripción. El último punto son las conclusiones que inferimos al terminar el proceso de análisis.

II. Trabajos relacionados

Análisis de datos relacionado con COVID-19 han surgido recientemente, cada uno teniendo un enfoque distinto a la temática, pero por lo general utilizando conjuntos de datos que se han ido liberando al largo de los meses. La situación actual no s ha llevado a buscar formas de ayudar a que ésta mejore, existen varios modelos de análisis que se basan en técnicas actuales como Machine Learning, análisis predictivo, análisis estadístico, entre otros. Todos estos con un solo objetivo en mente, ayudar a la comunidad científica con información útil resultante obtenida de los datos abiertos al público.

Hay un numero de artículos relacionados con el tema del COVID-19 tomando distintas vías para obtener información. Tenemos un sistema [2] que propone analizar las actuales olas de coronavirus en Estados Unidos que se basa en un novedoso método de detección y concatenación de datos guiados. Esto separando los casos de los datos obtenidos en distintos modelos o grupos que representan las mayores olas de infección. Tenemos otro enfoque similar [3], pero con datos tomados de la ciudad en donde se originó la pandemia, Wuhan. Éste enfoque teniendo un Dataset con similitudes al usado en este artículo. Tomando casos de muertes por el virus y casos donde el paciente re recuperó. Para posteriormente realizar un análisis estadístico de los atributos y sus relaciones.

Tenemos otros enfocados a diferentes utilizando análisis textuales [4] del público en general utilizando datos extraídos de Twitter. Esto para demostrar la reacción de las personas hacia la pandemia y las opiniones acerca de la reapertura de actividades cotidianas. Estos datos ayudando a encontrar descontentos y aciertos en dicha reapertura. Otros enfoques utilizan los datos clínicos [5] de los pacientes como en un artículo centrado en el análisis de la fatalidad, casos descartados y meta-análisis. Siendo el objetivo final analizar cada síntoma presente no solo en los casos actuales, sino también tomar en cuenta la presencia de coronavirus a lo largo de los años.

También se puede tener un enfoque similar a este artículo y realizar un análisis de los datos, pero a una escala mucho

mayor. Como es el caso [6] de este artículo que busca encontrar proporcionar parámetros epidemiológicos, centrándose en proporcionar estimaciones de la fatalidad y la recuperación de los pacientes. Todo esto utilizando datos de todos los meses de la pandemia. O puede ser muy específico, como puede ser el uso de una droga específica. En este artículo [7] se estudia el uso de la hidroxicloriquina en pacientes con el virus.

III. Método

A continuación, procederemos a explicar de mejor manera el método que estamos siguiendo para el análisis de nuestro conjunto de datos:

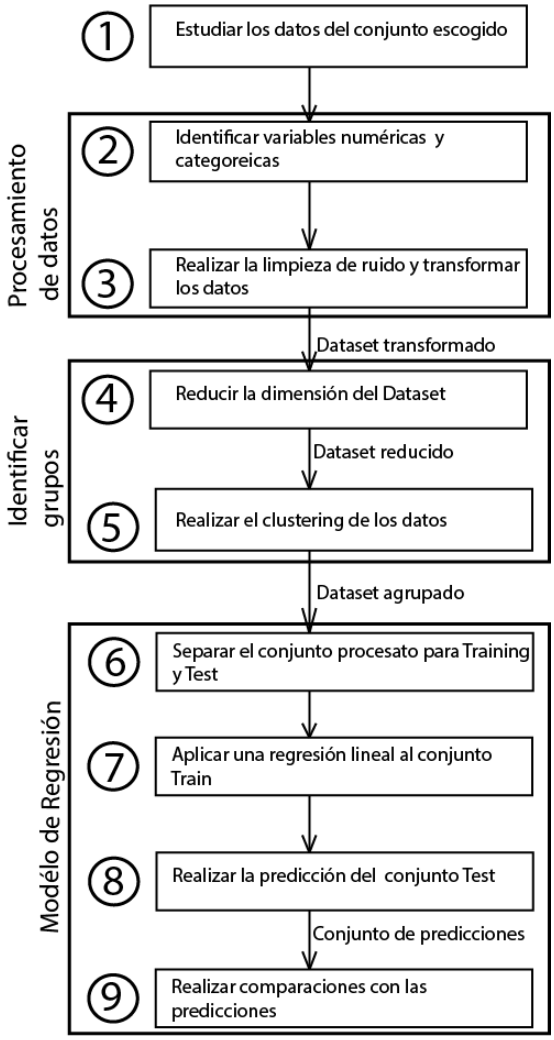

Fig. 1. Método implementado

A. Descrpción del Método

1) La primera tarea que se debería realizar al momento que se trabaja con un conjunto de datos es dedicarse un tiempo a analizar los datos que se nos presentan, la cantidad de atributos que tiene, que clase de información se nos presenta (numérica o categórica), etc. En ocasiones puede resultar compleja si estamos hablando de un Dataset con n alto número de atributos, pero en este caso son limitados y concretos. Podemos incluso detectar ciertas relaciones entre las variables si no trabajamos con un conjunto de datos con demasiados atributos.
2) Este trabajo se realiza en conjunto con el punto anterior, pero aquí se lo realiza de una manera más formar. Directamente nos dirigimos a la descripción del Dataset que usualmente se encuentra en el servicio web del que se extrajo, esto para buscar información relacionada con los valores categóricos y numéricos. Es este caso la información no estaba disponible, por lo que su identificación fue manual con un editor de tablas como Excel (al ser un Dataset con pocas variables).
3) Una vez que se haya realizado la identificación de los valores podemos preceder a hacer un procesamiento de los datos. Primeramente, se realiza la limpieza de ruido, don de se eliminan datos que podrían afectar de manera negativa nuestra salida. Para la limpieza aquellos datos numéricos que no contenían respuesta se los reemplazó con la moda, la razón fue que ya existían datos predefinidos para cada campo y de esta manera no afectar el conjunto resultante. La segunda es más simple y consta de transformar todos estos datos que contienen caracteres, a números.
4) Para realizar un proceso de clustering es recomendable realizar un proceso de reducción de dimensionalidad, esto optimizará la utilización de recursos del ordenador y también se podrá graficar los distintos grupos si lo reducimos a dos o tres dimensiones. Utilizamos el análisis de componentes principales (PCA) para reducir la dimensión del conjunto de datos.
5) Al momento de realizar el clustering, primero es necesario determinar la cantidad óptima de grupos (el mejor K). Para esto utilizaremos el método del codo en donde realizaremos un clustering con cada cantidad de componentes posibles, desde 2 hasta la cantidad de columnas que tenga nuestro Dataset (en este caso 17). Por cada cálculo de clustering obtendremos un valor de distorsión que nos da a conocer que tan dispersos están los datos. En este método graficaremos la distorsión junto con la cantidad de grupos y tomaremos el valor en donde notemos que la línea comienza a hacerse horizontal (Fig. 2).

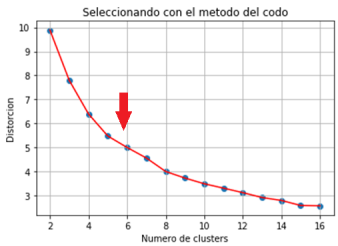

Fig. 2. Método del codo

6) Cuando se realiza un modelo de predicción se debe tomar parte del Dataset para el "training", que son los datos que se le da al sistema para que realice un aprendizaje (Machine Learning), mientras de "test" que la otra parte es destinada para hacer pruebas de predicción. En este caso se ha destinado un 20% del Dataset para realizar las pruebas. Se separan los datos y el resultado, en este caso serían los antecedentes y la posible precencia del virus.
7) Una regresión nos ayuda a aproximar la relación entre los datos, determina si estos siguen un patron visible. En este caso estamos usando una regresión lineal que nos ayuda a encontrar una recta que ayude a aproximar los dqatos. En este caso usamos nuestro Dataset previamente creado de "training" y esto nos dará como resultado coeficientes óptimos para cada uno de los atributos que ayudaran a predecir posteriormente.
8) Ahora es momento de predecir, lo que se hace es cargar el segundo Dataset de "test" donde los coeficientes antes generados nos ayudarán a realizar la predicción. Aquí es donde el análisis empieza a tener forma, aquí se realizará una predicción en donde la presencia del COVID-19 se representa con un 0 y la presencia de una simple gripe es representada con un 1.
9) Por último, nos queda comparar los distintos datos que hemos compilado a lo largo del proceso. Tenemos primeramente la salida de datos de del "test" con esto podemos comparar con el resultado de la salida de la predicción con la regreión lineal.

B. Análisis de grupos

Al momento de realizar el clustering obtendremos nuestros grupos, cada grupo compartirá características en común relacionadas con distintas variables. Todo esto nos ayuda a visualizar relaciones entre los casos de COVID-19 con los casos de la influenza común.

Algorithm 1 Proceso de Cálculo de clusters

Input: $Grupos(K)$
Output: $caracteristicasdelosgrupos$

$K = n$
repeat
  Por cada i hasta K numero de grupos
    Calcular cantidad de casos de COVID-19
    Calcular casos con Fiebre
    Calcular casos con Dyspnea
    Calcular casos con Tos
    Calcular casos con Astheniay
    Calcular casos con Leukopenia
until se haya recorrido todos los valores de K

Cuando hayamos encontradola cantidad óptima de grupos habremos precedido a realizar la separación de los grupos. En ese momento lo que se plantea es tomar cada grupo y separarlo en un Dataset diferente por motivos de comodidad. Cada grupo posee cierto número de casos positivos y cierto número de casos negativos (coronavirus o influenza), al estar separados podremos centrarnos en analizar ciertas variables que creemos pertinentes. En este caso analizaremos algunos síntomas que pueden ser detectados en casos de pacientes con coronavirus. Extraemos los casos donde están presentes estos síntomas y tenemos una fuente de información importante para nuestro análisis.

IV. Diseño de experimentos

Este Dataset contiene casos de COVID-19 al cuál podemos probar distintos algoritmos de clasificación. Está compuesto de 68 casos del virus en Italian Society of Medical and Radioligy Intervention (SIRM) y 62 casos de gripe común directamente de The Influenza Research Society (IRD); están mezclados de manera aleatoria donde se les ha agregado una columna más que nos dice si es un caso de COVID-19 o de influenza. A continuación tenemos la descripción de cada uno de los atributos presentes en el Dataset para brindar el contexto necesario:

Atributos	**Descripción**
Edad	Edad de la persona en cuestión
Género	Genero de la persona
Fiebre	Existencia de fiebre
Dispnea	Dificultad para respirar
Nasal	Congestión nasal
Tos	Existencia de tos
PO2	Presión parcial de oxigeno
CRP	Niveles de proteína C reactiva
Astenia	Existencia de cansancio o debilidad
Leucopenia	Bajo recuento de glóbulos blancos
Exposure to Covid-19 patients	Exposición a pacientes con el virus
Zona de alto riesgo	Se encuentra en zona de alto riesgo
Temperatura	Toma de temperatura
Prueba de sasngre	Resultado de prueba de sangre
RT-PCR	Reacción en cadena de la polimerasa con transcriptasa reversa
Historial	Condiciones médicas importantes (historial médico)
Decisión	Posible caso de COVID-19 o un resfriado

TABLE I
Atributos del Dataset

Al momento de estudiar el Dataset es posible darse cuenta de atributos redundantes o que son completamente irrelevantes para el análisis que se está realizando. Por ejemplo, en este caso el conjunto de datos contenía un atributo que hacía referencia al número de fila, el cual es un dato completamente irrelevante por lo cuál se lo ha eliminado del análisis.

V. Resultados

Esta sección se dividirá en dos partes donde se mostrarán los resultados en distintos aspectos. La primera sección mostrará los resultados de la estadística descriptiva como las relaciones entres variables y la distribución de los distintos grupos después del clustering. La segunda parte nos mostrará los resultados del modelo utilizando regresión para realizar predicciones.

A. Estadística Descriptiva

Ciertas medidas de la nos permiten ver cómo se comportan los datos en el conjunto de manera muy general, también podemos detectar qué variables están más relacionadas a otra y entre otras propiedades interesantes de los datos. Lo primero que puede pasar por la mente es la media, a pesar de ser un valor muy simple nos puede presentar información interesante. De acuerdo con la media tenemos que la edad media en las muestras es de 28 años, que la mayoría de personas presentaban síntomas como fiebre y tos, también tenemos que la temperatura media tomada fue de 25 grados y que la mayoría de personas no estuvo expuesta al virus.

Variable	Correlación
Zona de alto riesgo	91%
Asthenia	61%
Tos	39%

TABLE II
CORRELACIONES CON LA DECISIÓN

Algo muy importante al buscar relaciones entre los datos es sus correlaciones, nos ayuda a determinar qué tan relacionado esta un atributo de otro en una escala de más relacionado (1) o sin ninguna relación (0). En la tabla superior se encuentran las correlaciones más relevantes con la variable Decisión (ver Tabla I). Se puede notar como la decisión final que nos indica si es un caso de coronavirus o influenza está altamente relacionada con la zona de alto riesgo, ya que la probabilidad de contagio es mucho más alta, este es un factor determinante para el resultado final. En segundo puesto tenemos que la astenia a cansancio también cumple un papel importante en determinar la decisión final. Como dato general tenemos que el cansancio es uno de los síntomas más habituales en casos de coronavirus, pero también lo es en casos de influenza común. Por último, tenemos la variable Tos como un factor influyente, siendo una vez más un síntoma muy común entre las dos enfermedades. Todo esto no lleva a pensar que los factores influyentes son comunes para las dos enfermedades ya que independientemente de la decisión final, es muy probable que estos síntomas estén presentes. En tanto a la zona de alto riesgo, tiene sentido su alta influencia ya que muchas de las personas que fueron diagnosticadas para obtener estos datos seguramente se realizaron la prueba después de sospechas de haber estado en contacto con el virus.

B. Clustering

Otro punto importante para encontrar relaciones entre los datos es el clustering, al momento de realizar un proceso de clustering los datos son segmentados en distintos grupos (en este caso 5 grupos) dependiendo de una medida de distancia entre ellos. Los grupos también se traducen en relaciones ya que cada uno tendrá características en común existentes en cada miembro de ese grupo. En la Fig. 3 podemos observar los distintos grupos fácilmente gracias a la ayuda de la reducción de dimensionalidad, esto solo con fines didácticos. Existen datos que considerar en cada uno de estos grupos, pero ahora nos centraremos más en dos, el que contiene más casos de COVID-19 y el que contiene exclusivamente casos de influenza común:

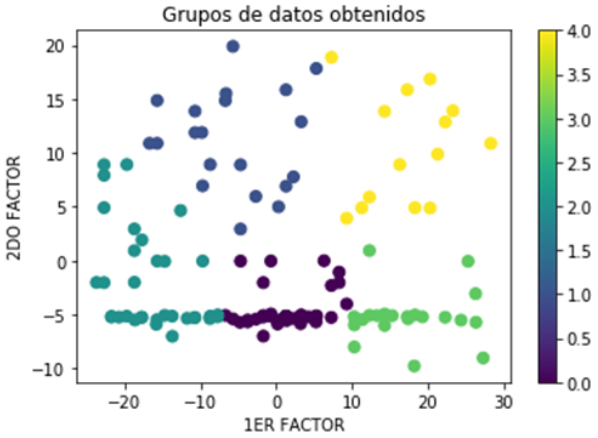

Fig. 3. Distribucuón de casos en los clusters

1) En este grupo se concentran la mayor cantidad de casos de COVID-19 en comparación con el resto de grupos. En el gráfico que se muestra el la Fig. 4 podemos ver algunas de las variables importantes que representan algunos síntomas comunes en el virus. Tenemos fiebre, dispnea (dificultad para respirar), tos, astenia (debilidad o cansancio) y leucopenia (bajo conteo de glóbulos blancos). Tenemos una alta presencia de tos y fiebre al ser síntomas muy comunes presentes en casos confirmados.

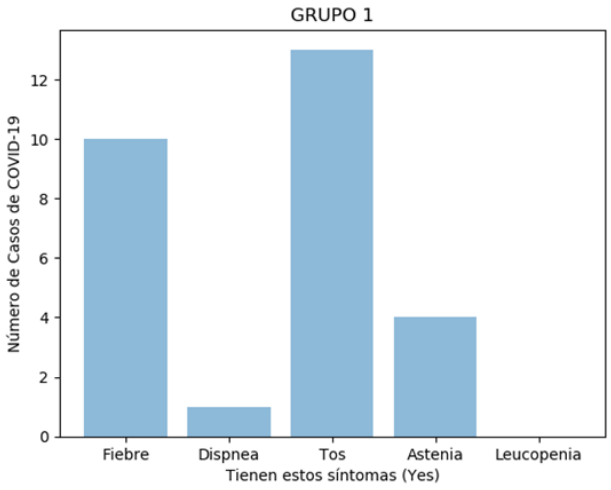

Fig. 4. Síntomas presentes en el grupo 1

2) Por otro lado, tenemos el quinto en que se encuentran la mayor cantidad de casos de influenza común, a tal punto de no existir casos de no existir ningún caso de coronavirus en el grupo. Visualizamos en la Fig. 5 que los casos poseen síntomas como fiebre y tos comúnmente, mientras que casos con dispnea y leucopenia son muy raros o inexistentes.

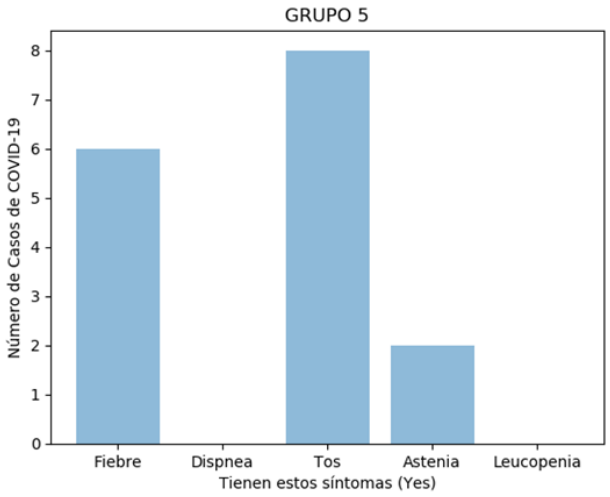

Fig. 5. Síntomas presentes en el grupo 5

Los datos en los grupos presentan similitudes muy notorias en estas variables seleccionadas para analizar, indicando que ambos casos presentan mayormente fiebre y tos. Esto se podría deber a las increibles similitudes que existen entre los síntomas de ambas enfermedades, estan dos variables simplemente no son suficientes para determinar la presencia de coronavirus o no. La variable de dispnea también es muy similar, pero mucho menos común que las anteriores pudiendo deberse al temprano diagnóstico de cualquiera de las enfermedades o simplemente que el síntoma no se manifestó en lo absoluto. Otro dato curios recide en la variable de leucopenia, porque esto indica que muchos de los casos expuesto en el Dataset no estan muy relacionados con la falta de respuesta inmune del cuerpo, esta más ligado a variables como la zona de alto riesgo de la que se habló en el punto anterior.

C. Modélo de predicción con rehgresión

Se han obtenido los coeficientes para la predicción y la predicción se puede ver a continuación el la Fig. 4.

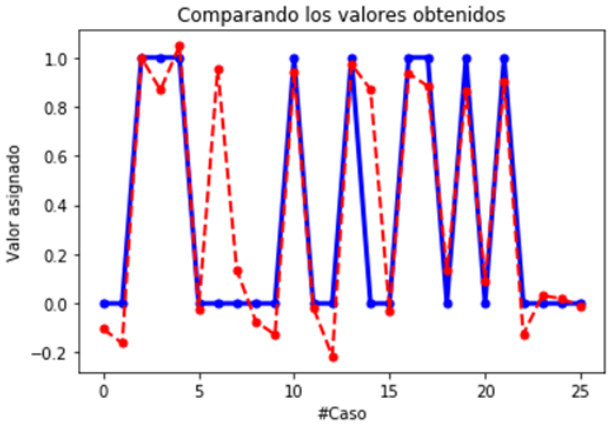

Fig. 6. Resultado de las predicciones

Al momento de realizar la regresión lineal se ha obtenido resultados mayormente favorables usando el Dataset creado para el Training.En la Fig. 4 la línea azul representa los valores reales que se encuentran en la parte de Test, estos valores pueden ser 0 o 1 (COVID-19 o Influenza respectivamente). Mientras que la línea roja representa los valores obtenidos por la predicción que se a logrado gracias a la regresión lineal implementada sobre los datos. Podemos notar que los resultados obtenidos no están muy alejados del valor real, tomando como ejemplo, nos ubicamos en el caso numero 5 y notaremos que existe una diferencia insignificante con la predicción. Aunque también tenemos lo opuesto a una buena predicción, entre el caso 5 y 10 existe un notorio cambio en la gráfica, donde la predicción fallo completamente dejándonos con un valor muy alejado del real. Este problema puede ser evitar teniendo un Dataset mucho más grande con el cual el proceso de Training será más eficiente al aumentar los datos.

Ahora es factible ingresar tus propios datos y realizar una predicción, sin embargo, el sistema sería mucho más viable si el conjunto de datos fuera extenso y que existan más variables que tomar en cuenta. A pesar de todo es un buen ejemplo de como podría ser de ayuda un sistema como este.

VI. Conclusiones

La realización de este método ha sido estimulante en mi opinión, tener una situación tan cercana para estudiarla no solo por el hecho de cumplir mi deber de estudiante, sino también mi deber como miembro de la sociedad. Es de cierta manera satisfactorio estudiar los datos y darse cuenta que hay mucho más en ellos de lo que se piensa en un comienzo, empezamos a encontrar relaciones que desencadenan en distintas inferencias referentes al significado de los datos.

En este caso en particular he podido notar mucha información importante que ha surgido del análisis estadístico. Podemos ver como ciertos síntomas se agrupan con otros en formas muy diversas, también como pueden estar tan separados entre ellos al punto de no tener ninguna relación el uno con el otro. Como una sola variable como ser parte de una zona de riesgo puede influir tanto en la decisión final de determinar si es un nuevo caso del virus o solo una gripe. Admirar como al momento de hacer un clustering y visualizar los datos de los distintos grupos pueden arrojar datos tan interesantes acerca de los casos y sus características en común con otros casos.

A final del análisis hemos podido sacar provecho de un conjunto de datos referente a un tema muy importante de actualidad, la posibilidad de sacar nueva información está a nuestro alcance. El simple hecho de poder obtener un Dataset como éste para su uso sin ninguna restricción es impresionante. Estamos sin duda en una situación muy complicada, pero la posibilidad de contribuir a esa causa está completamente abierta sin importar la escala de esa contribución.

Referencias

[1] Hamed, Ahmed, 2020, ”COVID-19 Dataset”, https://doi.org/10.7910/DVN/LQDFSE, Harvard Dataverse, V1, UNF:6:RAlD/Ta6J+9xN/Ok+6Cr7A== [fileUNF]

[2] V. Marmarelis, "Predictive modeling of Covid-19 data in the US: Adaptive phase-space approach," in IEEE Open Journal of Engineering in Medicine and Biology, doi: 10.1109/OJEMB.2020.3008313.

[3] Ruan, Q., Yang, K., Wang, W., Jiang, L., & Song, J. (2020). Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive care medicine, 46(5), 846-848.

[4] J. Samuel et al., "Feeling Positive About Reopening? New Normal Scenarios from COVID-19 US Reopen Sentiment Analytics," in IEEE Access, doi: 10.1109/ACCESS.2020.3013933.

[5] Li, L. Q., Huang, T., Wang, Y. Q., Wang, Z. P., Liang, Y., Huang, T. B., ... & Wang, Y. (2020). COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. Journal of medical virology, 92(6), 577-583.

[6] Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PloS one, 15(3), e0230405.

[7] Geleris, J., Sun, Y., Platt, J., Zucker, J., Baldwin, M., Hripcsak, G., ... & Sobieszczyk, M. E. (2020). Observational study of hydroxychloroquine in hospitalized patients with Covid-19. New England Journal of Medicine.

[8] Xu, X., Han, M., Li, T., Sun, W., Wang, D., Fu, B., ... & Zhang, X. (2020). Effective treatment of severe COVID-19 patients with tocilizumab. Proceedings of the National Academy of Sciences, 117(20), 10970-10975.

[9] Escobar, G., Matta, J., Ayala, R., & Amado, J. (2020). Características Clínicoepidemiológicas de pacientes fallecidos por COVID-19 en un hospital nacional de Lima, Perú. Revista de la Facultad de Medicina Humana, 20(2), 180-185.

[10] Mejia, C. R., Rodriguez-Alarcon, J. F., Garay-Rios, L., de Guadalupe Enriquez-Anco, M., Moreno, A., Huaytan-Rojas, K., ... & Curioso, W. H. (2020). Percepción de miedo o exageración que transmiten los medios de comunicación en la población peruana durante la pandemia de la COVID–19. Revista Cubana de Investigaciones Biomédicas, 39(2).

[11] Bastos, M. L., Tavaziva, G., Abidi, S. K., Campbell, J. R., Haraoui, L. P., Johnston, J. C., ... & Menzies, D. (2020). Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. bmj, 370.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[13] Garreta, R., & Moncecchi, G. (2013). Learning scikit-learn: machine learning in python. Packt Publishing Ltd.

[14] Jolly, K. (2018). Machine Learning with scikit-learn Quick Start Guide: Classification, regression, and clustering techniques in Python. Packt Publishing Ltd.

[15] [1] Evans, M. Rosenthal, J. (2013) Probabilidad y estadística la ciencia de la incertidumbre. Reverté.

[16] Forsyth, D. (2018). Probability and Statistics for Computer Science. Springer.

[17] Jolly, K. (2018). Machine Learning with scikit-learn Quick Start Guide: Classification, regression, and clustering techniques in Python. Packt Publishing Ltd.