BORRADOR - Método para clasificación

*Note: Sub-titles are not captured in Xplore and should not be used

***Abstract*—Se propone un método de análisis de datos para un Dataset con datos relacionados con el COVID-19. El este artículo se muestra paso a paso el proceso de transformación, clustering y utilización de regresiones para predicción de datos. Se describe brevemente los datos presentes en el conjunto de datos y se hace inferencias con ayuda de estadística descriptiva. Para al final mostrar los resultado del modelo junto con gráficas, descripciones y conclusiones acerca del tema.**

***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 este caso los datos adquiridos me han facilitado entender ciertos aspectos que se relacionan a la enfermedad sin entrar en mayores detalles de terminología médica. 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 cuatro secciones y sus referencias. Como primer punto tenemos el método propuesto seleccionado para el uso de los datos, esto con sus respectivos diagramas y algoritmos. El segundo punto se dedica a describir el conjunto de datos, el porcentaje usado para test y los parámetros utilizados para el análisis. El punto tres es bastante concreto, aquí 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. Método

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

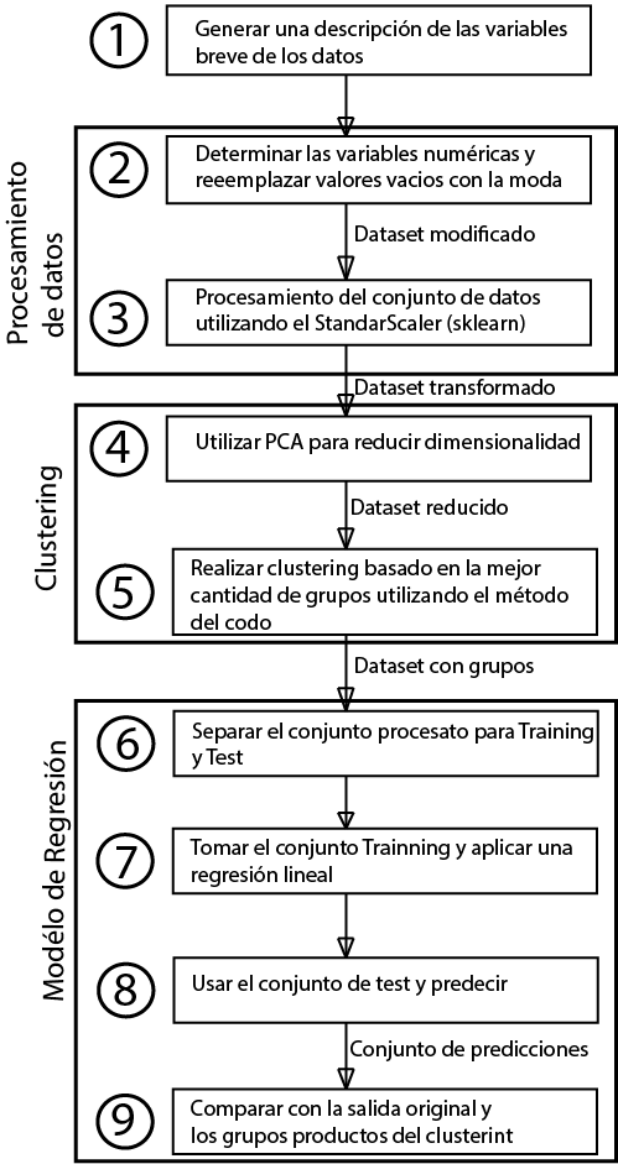

Fig. 1. Método implementado

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) Al momento de trabajar con el conjunto de datos noté que los valores numéricos contenían espacios que no contenían ninguna información más que un carácter que determinaba que no existía respuesta. Habiendo revisado de antemano nuestro conjunto de datos notamos estas irregularidades que posteriormente nos causaran problemas cuando transformemos y hagamos nuestro análisis. La mejor opción en estos casos es encontrar la media de los datos que están a su alrededor y reemplazar estos espacios con la misma, pero en este caso los valores que se colocan en estos espacios son predefinidos, entonces optamos por utilizar el valor que más se repite en cada columna con este problema.
3) Al momento de estudiar el conjunto de datos también se determinan los valores numéricos y categóricos, esto nos ayudará a determinar el mejor método de transformación de los datos. En el caso de este Dataset la mayor parte de sus atributos categóricos y los que son numéricos tienen un valor determinado, no continuo. Al momento de transformar estos datos a valores numéricos se debe t6ener en cuenta que no existan valores vacíos, de ser el caso se tendrá que asignar un valor. Se utilizó la librería "sklearn" de Python junto con el StandarScaler para transformar los datos.
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 nos podremos 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).
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. También utilizando los grupos de clustering podremos sacar más inferencias del conjunto de datos y determinar relaciones. También es posible ingresar nuevos datos y realizar la predicción sin ningún problema.

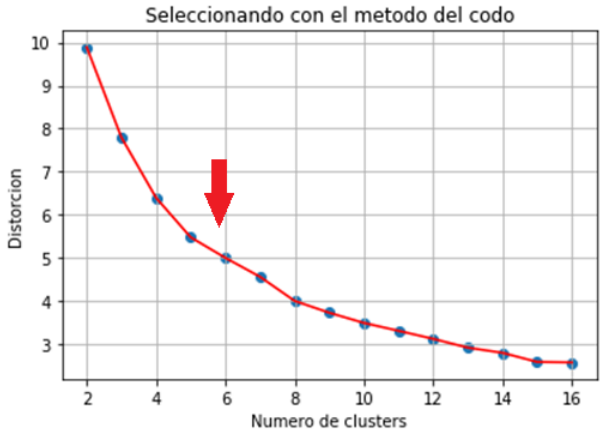

Fig. 2. Método del codo

Algorithm 1 Proceso de Cálculo de clusters

Input: K
Output: $cluster$ y $distorsion$

1: $K = [2..., 17]$
2: **repeat**
3: **Por cada** i en el rango de grupos K
4: **Calcular** i cantidad de grupos
5: **Calcular** la distorsión en i cantidad de grupos
6: **until** se haya recorrido todos los valores de K

B. Descripción general del Dataset

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
Age	Edad de la persona en cuestión
Gender	Genero de la persona
Fever	Existencia de fiebre
Dyspnea	Dificultad para respirar
Nasal	Congestión nasal
Cough	Existencia de tos
PO2	Presión parcial de oxigeno
CRP	Niveles de proteína C reactiva
Astheniay	Existencia de cansancio o debilidad
Leukopenia	Bajo recuento de glóbulos blancos
Exposure to Covid-19 patients	Exposición a pacientes con el virus
High risk zone	Se encuentra en zona de alto riesgo
Temp	Toma de temperatura
Blood Test	Resultado de prueba de sangre
RT-PCR	Reacción en cadena de la polimerasa con transcriptasa reversa
History	Condiciones médicas importantes (historial médico)
Decision label	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.

III. 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

La estadística descriptiva nos ha permitido ver como se comportan los datos en el conjunto, 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 1	Variable 2	Correlación
High risk zone	Fever	0.075294
Decision label	High risk zone	0.911765
Fever	Nasal	0.052411
Dyspnea	Leukopenia	0.093618

TABLE II
CORRELACIONES NOTORIAS

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 este caso podemos observar correlaciones notorias (TABLE II), podemos observar que la fiebre no tiene prácticamente ninguna relación con la zona de alto riesgo, esto porque la presencia de síntomas como éste no depende de la ubicación sino de la existencia de otros síntomas o posteriores estudios. En este caso se puede notar como la decisión final que nos indica si es un caso de coronavirus o influenza esta altamente relacionada con la zona de alto riesgo, ya que la probabilidad de contagio es mucho más alta. Por último, tenemos dos ejemplos de algunos síntomas que no están relacionados entres si, la fiebre y congestión nasal, no tienen ninguna relación relevante, por otro lado, también la debilidad y el ajo recuento de glóbulos blancos tampoco parecen tener ninguna relación, al menos en este conjunto de datos. Que síntomas como éste no tengan ninguna relación puede deberse a que son casos aislado o que directamente no exista información otorgada por las personas al momento de recolectar los datos (espacio vacío o sin responder).

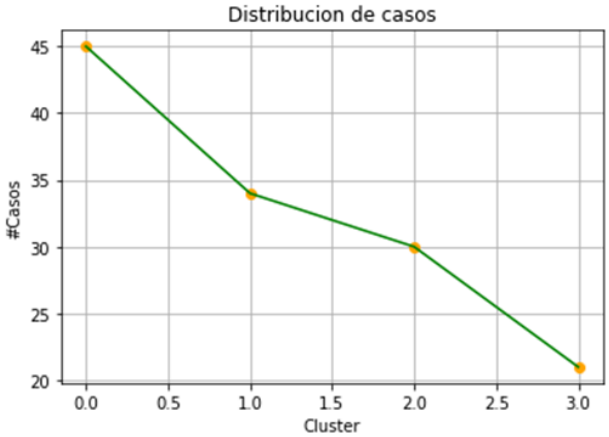

Fig. 3. Distribucuón de casos en los clusters

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 4 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

como se distribuyen los casos en los distintos grupos, se puede apreciar como el primer grupo tiene la mayor cantidad de datos, mientras que el cuarto tiene la menor cantidad. Ahora analicemos estos grupos y sus datos en busca de encontrar relaciones entre ellos. A continuación, mencionaremos las características más importantes de cada grupo:

1) En el primer grupo hay un factor muy importante que nos podría ayudar a sacar algunas conclusiones, este grupo (grupo 0.0) existen la mayor cantidad de casos determinados como influenza común, mientras que los casos de COVID-19 son muy escasos. Cada atributo en este grupo tiene muchas características en común como la presencia de congestión nasal, niveles de la proteína C reactiva, ningún problema con el recuento de glóbulos blancos, no han tenido ninguna exposición a pacientes con el virus, entre otras.
2) En cuanto a los demás grupos son mucho más variados en casos de influenza y coronavirus, pero los casos de coronavirus siempre son más numerosos que los de influenza. En el segundo grupo (grupo 1.0) tenemos características como: Dificultades para respirar, son mayores a 38 años, cansancio y debilidad presente, etc.
3) En el tercer grupo tenemos que hay existencia de fiebre en los casos, por lo general son personas menores a 30 años, su temperatura está alrededor de 25 grados, por lo general no tienen un historial médico, existen una gran parte de personas con dificultad para respirar, etc.
4) Finalmente tenemos el cuarto grupo (grupo 3.0) en donde encontramos características tale que no presentan problemas en el recuento de los glóbulos blancos, la mayoría provienen de zonas de riesgo, su cuenta de la proteína C activa es alta, etc.

B. 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.

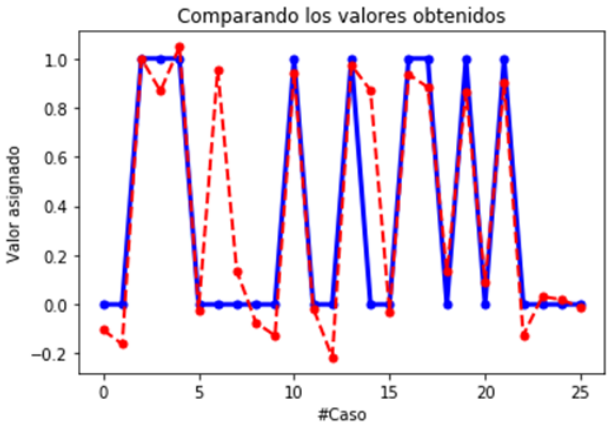

Fig. 4. Resultado de las predicciones

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.

IV. 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.