Classification of covid mortality by groups and by provinces

August 5, 2020

Abstract

In this study, a total of 538 sample groups that provided medical data about groups of people in different locations were analyzed. The mayor's challenge in this study was to carry out 2 different criteria within the same data set to be able to conclude that the mortality of the person depends more than anything on the age of the person at risk and the presence of one or more health disorders in addition to the primary disease, which in this case is COVID-19 disease. To carry out this study, the COVID Analytics dataset [2] 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. If the patient's disease is mostly correlated with hypertension and any disease the patient has, while with the coronavirus, age is also an influencing factor in patient mortality. This research allows evaluating the conditions in which a person is found and determining in the most specific way the reason why a person dies after being infected by COVID-19.

Keywords— COVID-19, Regresion Lineal, Clustering, Regresion multi-lineal

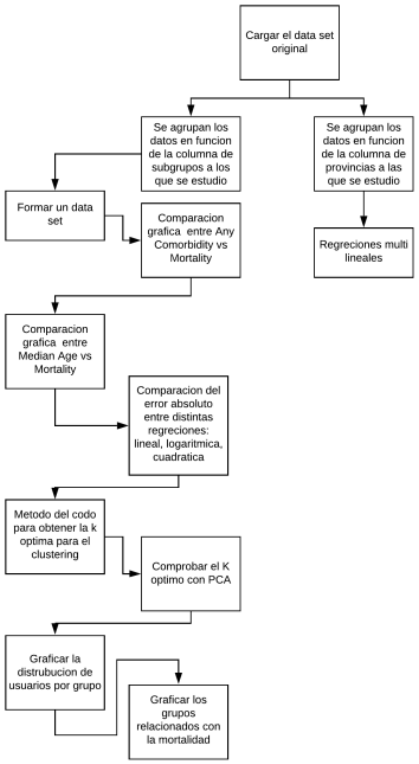

Figure 1: Proceso realizado

I Introducción

En este articulo lo que se hace es un preprocesamiento de los datos para que se pueda comenzar un análisis bajo dos enfoques distintos los cuales serían el análisis de los datos a partir de los distintos grupos analizados y un análisis según la provincia en la que se llevaron los estudios. Para poder llevar a cabo esta clasificacion se siguieron los siguiente pasos: 1)Limpieza de datos, 2) Preprocesamiento, 3) Creacion de nuevos dataset.

II Resultados

I Limpieza de datos

Para comenzar con el análisis del dataset se realizó una limpieza de los datos la cual nos ayudara a que el data set no tenga espacios en blanco o datos que no se entiendan y así obtener datos de calidad.

En esta limpieza de datos lo que se hizo fue cambiar los espacios en blanco de los datos categóricos ordinales por la media de la columna analizada, en el caso de los datos categóricos nominales se campearon los datos

vacíos por Unknown ya que no se sabe con certeza a que clasificación de los datos pertenecen las casillas en blanco.

II Preprocesamiento

En el preprocesamiento partimos desde la clasificación de los datos en ordinales y nominales para poder hacer un tratamiento de los datos con la librería sktlearn [11].

Para los datos ordinales se aplicó un procesamiento con StandardScaler()

$$z = (x - \mu)/s$$

Para los datos nominales se aplicó un procesamiento con OneHotEncoder() para poder formar columnas con valores únicos por cada característica que haya en la respectiva columna del dato categórico nominal.

III Creacion de nuevos dataset

A partir del dataset preprocesado se creó un data set donde se agrupa la media de los datos con relación a los subgrupos de estudio que están dentro del mismo. Para la creación del dataset de las provincias se hizo una agrupación de la media de los datos con relación a las provincias que se analizaron en el dataset.

En la siguiente sección III presentamos el trabajo relacionado a la extracción de datos, en la sección IV y V se aclaran los procesos seguidos, finalmente presentamos la sección VI las conclusiones y el trabajo futuro.

Actualmente ya existen varios estudios que se enfocan a descubrir que factores son los que más perjudican a el estado actual de una persona contagiada con COVID-19 [7] [4] en este trabajo tratamos de hacer lo mismo, pero enfocándonos a las poblaciones que sufrieron de esto cuando recién empezó a expandirse la enfermedad.

Todos estos trabajos estan enfocados en ayudar a saber cuales grupos dar mayor prioridad en su recuperacion para evitar mas muertes causadas por el virus.

III Trabajo relacionado

Primero se empieza con la limpieza de los datos, lo cual nos ayudara a que este sea más liviano y mejorar su procesamiento, además de que gracias a este se podrán evitar problemas debido a datos nulos o con valores extraños, para conseguir los datos necesarios para este estudio se eliminó toda la información sobre las publicaciones incluidas en el dataset para que no interfieran con el tema de estudio principal.

Una vez terminada la limpieza se separaron los datos del dataset en ordinales y nominales. A los datos categorizados como ordinales le aplicamos el método de preprocesamiento StandardScaler() y a los categorizados como nominales se les aplico el OneHotEncoder(). Gracias a estos procesos se puedo llevar a cabo el proceso de correlación donde se descubrió la alta correlación que la mortalidad del coronavirus tiene con las comorbilidades que tenga el paciente y su edad.

Para lograr un mejor enfoque en el estudio del dataset se tomaron 2 enfoques, el primero consiste en una agrupación de la media de los datos con relación a las provincias y otro donde se agrupa la media de los datos con relación a los subgrupos de estudio.

I Tabla: Any Comorbidity

	Lineal	Cuadratica	Polinomial
Mean Absolute Error	0.1915	0.1617	0.1560
Mean Squared Error	0.0729	0.0552	0.0567
Root Mean Squared Error	0.2700	0.2351	0.2381

II Tabla: Median Age

	Lineal	Cuadratica	Polinomial
Mean Absolute Error	7.1396	7.1211	7.0364
Mean Squared Error	83.0179	82.7738	81.0739
Root Mean Squared Error	9.1114	9.0980	9.0041

IV Resultados del análisis de subgrupos

Se ingresa el data set csvDatasetGrupos.csv generado anteriormente el cual se dividió en variables categóricas y nominales donde las nominales se procesaron con el StandardScaler a excepción de la variable de "Mortality" la cual ocuparemos como salida y las categóricas con OneHotEncoder.

Después de realizar el preprocesamiento del dataset se saca las correlaciones de las variables con la mortalidad para poder saber cuáles son las que tienen más correlación con esta y así saber que variables son las que tienen un mayor impacto en relación con la mortalidad.

En conclusión con este procedimiento se vio que de las variables más correlacionadas con la mortalidad

son cualquier comorbilidad [5] , edad [10], hipertensión [3], Diabetes [6], enfermedades cardiovasculares [9], cualquier tipo de cáncer [1], Un bajo nivel de cedulas blancas [8].

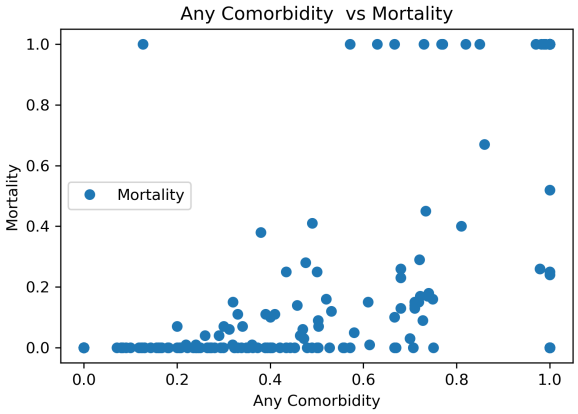

Figure 2: Cualquier comorbilidad vs Mortalidad

Como podemos ver en esta grafica las personas que tenían enfermedades aparate de la principal que sería el COVID-19 tienen más posibilidades que la mortalidad aumente.

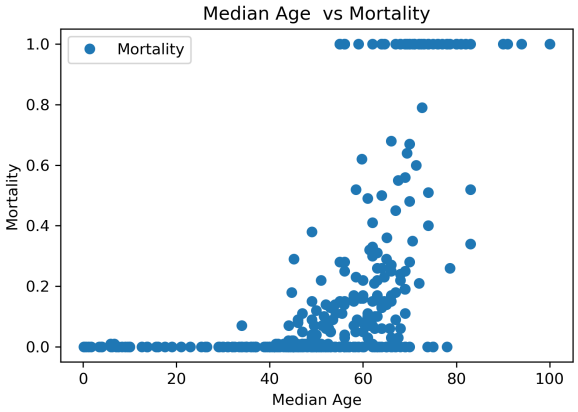

Figure 3: Edad vs Mortalidad

En la gráfica de la Edad vs Mortalidad podemos ver que a partir de los 40 años la tasa de mortalidad aumenta ya que mientras mayor edad tenga el paciente el riesgo de mortalidad aumenta.

Después de este análisis se vio que la variable "Any Comorbidity" puede englobar a las enfermedades que tenga el paciente a parte del COVID-19 por su alta correlación con la variable de "Mortality" así que se decidió hacer una regresión lineal para intentar predecir el riesgo de mortalidad de un grupo de personas mediante la regresión lineal.

$$Y = \alpha + \beta X + \varepsilon$$

Con esta única variable se logró sacar un error absoluto de 0.2052 en la regresión lineal entre Any "Comorbidity" y "Mortality".

Para poder sacar los grupos de variables más relevantes en relación con la variable de "Mortality" se aplicó la técnica de Clustering partiendo del Clusterning con Kmeans para poder sacar la K optima mediante el método del codo.

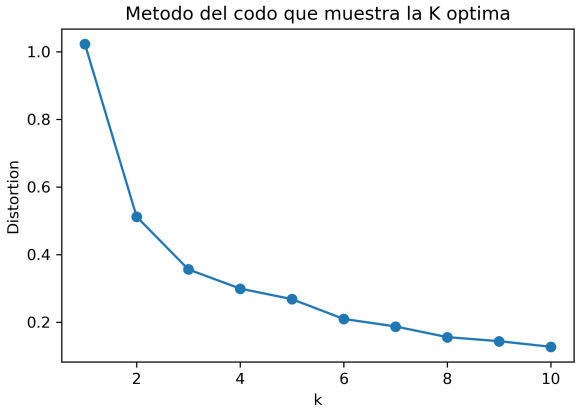

Figure 4: K optima

Como podemos ver en la gráfica hemos visto que la k más óptima para nuestro Clustering es de 4 y lo que haremos para comprobar esta decisión se sacó un PCA con 4 componentes para sacar su varianza y de esta manera asegurar nuestra decisión.

$$[0.106 \;\; 0.054 \;\; 0.049 \;\; 0.038]$$

En las varianzas que nos arrojó el PCA podemos ver que nuestra decisión del k optimo es correcta ya que

las varianzas entre los componentes son alejadas entre unas y otras.

Una vez que comprobamos nuestro k optimo se procedió a ver la distribución de usuarios que habría por cada grupo en el Clustering como podemos ver en la siguiente gráfica:

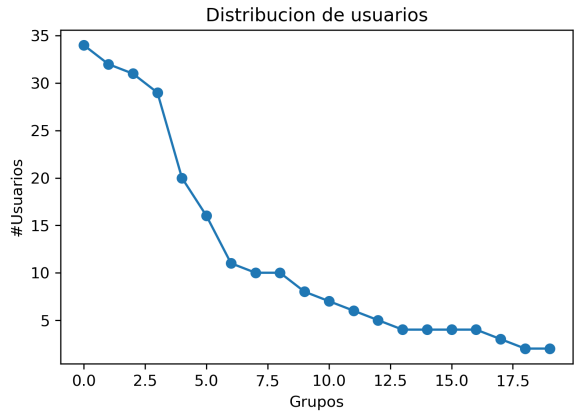

Figure 5: Distribucion de usuarios

Se crearon los 4 grupos con el proceso de Clustering y gracias a estos se puede hacer estudios independientes para cada grupo en concreto y se podrá ver que grupos tienen más relación con la mortalidad en nuestro data set.

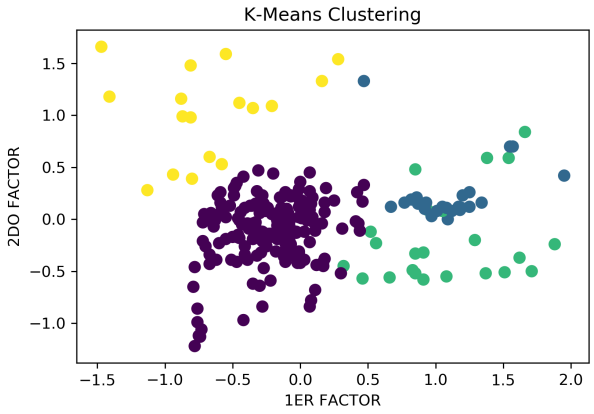

Figure 6: Grupos con Clustering

V Resultados del análisis de provincias

Para este estudio se cargó el otro dataset llamado csvDatasetProvincias que también se generó anteriormente, se dividió en variables categóricas y nominales, para las categóricas se utilizo el OneHotEncoder y para las nominales se usó el StandardScale.

Completado el preprocesado del dataset de provincias se designó a la variable Mortality como salida y se prepararon los datos de Train y Test con un tamaño del 70% y del 30% respectivamente. Luego de esto se aplicó un MinMaxScaler a los datos de Train y Test para que los datos no difieran tanto entre ellos, a los datos de salida no se les aplico porque se dañarían los datos.

Luego de eso se le aplica una regresión multilíneal usando los datos de Train esto es para luego realizar las predicciones con los datos de test y someterlos a una comparación.

$$Y = \beta 0 + \beta 1X1 + \ldots + \beta pXp + \varepsilon$$

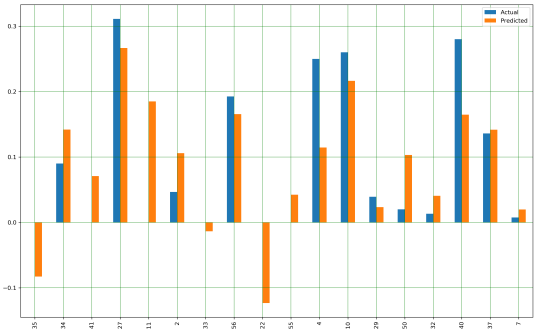

Figure 7: Regrecion lineal (Any Comorbidity, Mortality)

Obtuvimos la medida de los errores de las predicciones con el test para una calificación general de las predicciones.

Error absoluto medio: 0.06328468403276438.

Error cuadrático medio: 0.006322767478601251.

Error cuadrático medio de raíz: 0.07951583162239612.

A continuación se muestra las conclusiones y trabajos a futuro.

VI Conclusiones y trabajo futuro

Se logro determinar 4 grupos de factores que afectaban a la tasa de mortalidad que produce del covid sobre los grupos del dataset analizados.

Para la parte de predicciones en vez de grupos usamos las provincias en base a su tasa de mortalidad.

La razón por la cual se dividió al dataset en 2 más es para poder obtener una mejor respuesta de la evaluación de los datos.

Se quiso usar un análisis de PCA pero dado a que existía una complicación con la variable de salida por el tipo de datos que se obtenían se tomó la decisión de hacer una regresión multilínea que permita analizar la tasa de mortalidad de acuerdo a los grupos establecidos en el dataset.

Se puede definir los grupos de mortalidad del coronavirus y analizarlos también con la mortalidad para nuevamente encontrar entre estos cual es el que más se consideraría peligroso.

Bibliografía

[1] *American Cancer Society*. https://www.cancer.org/es/quienes-somos/coronavirus-covid-19-y-cancer.html. Accessed: 2020-07-24.

[2] *COVID Analytics*. https://www.covidanalytics.io/dataset. Accessed: 2020-07-24.

[3] *COVID-19 e hipertension arterial*. https://revistanefrologia.org/index.php/rcn/article/view/405. Accessed: 2020-07-24.

[4] *el-periodico*. https://www.elperiodico.com/es/sociedad/20200226/coronavirus-infectados-peligroso-7863722. Accessed: 2020-07-24.

[5] *infoMed*. https://temas.sld.cu/coronavirus/tag/comorbilidades/. Accessed: 2020-07-24.

[6] *Massachusetts General Hospital*. https://www.massgeneral.org/es/coronavirus/la-relacion-entre-la-diabetes-y-el-covid-19. Accessed: 2020-07-24.

[7] *mayo-clinic*. https://www.mayoclinic.org/es-es/diseases-conditions/coronavirus/in-depth/coronavirus-who-is-at-risk/art-20483301. Accessed: 2020-07-24.

[8] *Modern Healthcare*. https://www.modernhealthcare.com/clinical/low-white-blood-cell-counts-linked-severe-covid-19-cases. Accessed: 2020-07-24.

[9] *Organizacion Mundial de la Salud*. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses?gclid=CjwKCAjwsO_4BRBBEiwAyagRTS-lsEOC0unG-yiiL_N79JOZQeLS7iwVwzIFuZeEq5dG5tB3CuO7ohoCnAIQAvD_BwE. Accessed: 2020-07-24.

[10] *PRIMICIAS*. https://www.primicias.ec/noticias/sociedad/edad-importa-muertos-covid-quito-mas-50-anos/. Accessed: 2020-07-24.

[11] *scikit-learn*. https://scikit-learn.org/stable/. Accessed: 2020-07-24.