Publications of the Astronomical Society of Australia (2021), 1–4
doi:10.1017/pasa.xxxx.xx

CAMBRIDGE
UNIVERSITY PRESS

Research Paper

Minería de datos y aprendizaje automático aplicados al estudio de comorbilidades y mortalidad por COVID-19 en México

Gustavo Escobar[1]

[1]Facultad de Ciencias Físico Matematicas, Universidad Autonoma de Nuevo León, Nuevo León , México.

Abstract

This guide is for authors who are preparing papers for the *Publications of the Astronomical Society of Australia* journal using the LaTeX document preparation system and the CUP PAS style file.

Keywords: Covid-19, Machine Learning, México

(Received xx xx xxxx; revised xx xx xxxx; accepted xx xx xxxx)

1. Introducción

Descripción de la investigación en general.

2. Descripción de los datos

Según la Secretaría de Salud (De Salud, s.f.):

> *Conforme al Decreto publicado en el Diario Oficial de la Federación el 20 de febrero del 2015, que establece la regulación en materia de Datos Abiertos, la Dirección General de Epidemiología, con base en los ordenamientos aplicables en dicha materia, pone a disposición de la población en general, la información contenida en los Anuarios Estadísticos de Morbilidad 2015–2017, así como la información referente a los casos asociados a Influenza, COVID-19 y otros virus respiratorios con el propósito de facilitar a todos los usuarios que la requieran, el acceso, uso, reutilización y redistribución de la misma.*

En este artículo trabajaremos con la base de datos de los casos asociados a Influenza, COVID-19 y otros virus respiratorios publicada (). El conjunto de datos cuenta con un total de () registros y con 42 variables temporales, numericas y categoricas.

2.1. Variables temporales

En la Tabla 1 podemos apreciar las variables temporales junto con su descripción del conjunto de datos. Para las variables `Fecha_Actualizacion`, `Fecha_Ingreso` y `Fecha_Sintomas` no se encontraron valores nulos, caso contrario para `Fecha_Def` en donde estos valores nulos indican que el paciente no falleció.

Table 1. Descripción de las variables de tipo fecha.

Variable	Descripción
`Fecha_Actualizacion`	Permite identificar la última actualización.
`Fecha_Ingreso`	Fecha de ingreso del paciente.
`Fecha_Sintomas`	Fecha en que inició la sintomatología.
`Fecha_Def`	Fecha en la que el paciente falleció.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (s.f.).

2.2. Variables categoricas

Las variables categóricas del conjunto de datos se encuentran codificadas conforme al *Catálogo de variables y descriptores* publicado por el Gobierno de México, el cual define las categorías asociadas a cada valor numérico. Dichas variables pueden clasificarse en cinco grupos principales: *tipo de atención y estado clínico*, *pruebas de laboratorio y resultados*, *comorbilidades y factores de riesgo*, *características demográficas* y *origen y ubicación*.

2.3. Tipo de atención y estado clínico

En la Tabla 2 se encuentran las variables categoricas relacionadas al tipo de atención y estado clínico del paciente y su descripción.

Author for correspondence: G. de J. Escobar Mata, Email: gustavo.escobarma@uanl.edu.mx

Cite this article: Escobar Mata, G. de J., Minería de datos y aprendizaje automático aplicados al estudio de comorbilidades y mortalidad por COVID-19 en México. *Revista universitaria* **00**, 1–12. https://doi.org/10.1017/pasa.xxxx.xx

Table 2. Descripción de las variables categoricas correspondientes al tipo de atención y estado clínico del paciente.

Variable	Descripción
`Tipo_Paciente`	Identifica el tipo de atención que recibió el paciente en la unidad.
`Intubado`	Identifica si el paciente requirió de intubación.
`Neumonia`	Identifica si al paciente se le diagnosticó con neumonía.
`UCI`	Identifica si el paciente requirió ingresar a una Unidad de Cuidados Intensivos.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (s.f.).

2.3.1. Pruebas de laboratorio y resultados

En la Tabla 3 se muestra las variables categoricas correspondientes a las pruebas de laboratorio y sus resultados de los pacientes y su descripción.

Table 3. Descripción de las variables categoricas correspondientes a pruebas de laboratorio y resultados del paciente.

Variable	Descripción
`Toma_muestra_lab`	Identifica si al paciente se le tomó muestra de laboratorio.
`Resultado_PCR`	Identifica el resultado de la muestra del paciente a los diferentes virus respiratorios.
`Resultado_PCR_coinfeccion`	Identifica una coinfeccion encontrada en la muestra.
`Toma_muestra_antigeno`	Identifica si al paciente se le tomó muestra de antígeno para SARS-COV-2.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (s.f.).

2.3.2. Comorbilidades y factores de riesgo

En la Tabla 4 encontramos las variables categoricas correspondientes a las comorbilidades y factores de riesgo con los que cuentan los pacientes y su descripción.

2.3.3. Características demográficas

En la Tabla 5 se muestra las variables categoricas correspondientes a las características demográficas de los pacientes y su descripción.

2.3.4. Origen y ubicación

En la Tabla 6 se muestra las variables categoricas correspondientes al origen y ubicación de los pacientes y su descripción.

2.4. Variables númericas

Se identificó a la variable `Edad` como la única variable numérica continua presente en el conjunto de datos analizado con registros desde 0 hasta 111 años.

Table 4. Descripción de las variables categoricas correspondietnes a comorbilidades y factores de riesgo del paciente.

Variable	Descripción
`Neumonia`	Indica si el paciente tiene un diagnostico de neumonia.
`Diabetes`	Indica si el paciente tiene un diagnostico de diabetes.
`Asma`	Indica si el paciente tiene un diagnostico de asma.
`Hipertencion`	Indica si el paciente tiene un diagnostico de hipertencion.
`Epoc`	Indica si el paciente tiene un diagnostico de EPOC
`Inmuspr`	Indica si el paciente tiene un diagnostico de inmunosupresión
`Cardiovascular`	Identifica si el paciente tiene un diagnostico de enfermedades cardiovasculares
`Obesidad`	Identifica si el paciente tiene un diagnostico de obesidad
`Renal_Cronica`	Identifica si el paciente tiene un diagnostico de insuficiencia renal cronica.
`Tabaquismo`	Identifica si el paciente tiene hábito de tabaquismo.
`Otra_com`	Indica si el paciente tiene diagnóstico de otras enfermedades

Fuente: Catálagos y Descriptores de la Secretaría de Salud (s.f.).

Table 5. Descripción de las variables categoricas correspondietnes a las características demográficas del paciente.

Variable	Descripción
`Sexo`	Identifica al sexo del paciente.
`Nacionalidad`	Identifica si el paciente es mexicano o extranjero.
`Pais_Nacionalidad`	Identifica la nacionalidad del paciente.
`Pais_Origen`	Identifica el país del que partió el paciente rumbo a México.
`Migrante`	Identifica si el paciente es una persona migrante.
`Habla_Lengua_Indig`	Identifica si el paciente habla lengua índigena.
`Indigena`	Identifica si el paciente se autoidentifica como una persona indígena.
`Embarazo`	Identifica si la paciente está embarazada.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (s.f.).

3. Antecedentes

Treviño (2020) señala que, de acuerdo con los datos recopilados el 10 de junio de 2020 —que comprenden un total de 362,362 casos confirmados—, las comorbilidades más prevalentes entre los pacientes hospitalizados por COVID-19 en México fueron la

Table 6. Descripción de las variables categoricas correspondietnes al origen y ubicación del paciente.

Variable	Descripción
`Origen`	Identifica el Sistema de unidades de salud (USMER) que monitorean enfermedades respiratorias mediante hospitales y clínicas de distintos niveles de atención.
`Sector`	Identifica el tipo de institución del Sistema Nacional de Salud que brindó la atención.
`Entidad_UM`	Identifica la entidad donde se ubica la unidad medica que brindó la atención.
`Entidad_Nac`	Identifica la entidad de nacimiento del paciente.
`Entidad_Res`	Identifica la entidad de residencia del paciente.
`Municipio_Res`	Identifica el municipio de residencia del paciente.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (s.f.).

hipertensión (16.5 %), la obesidad (14.4 %) y la diabetes (12.6 %). Además ($\chi^2 = 162.51$; $p < 0.05$) el autor demuestra la existencia de una asociación estadísticamente significativa entre el hábito de fumar y la probabilidad de contraer COVID-19.

En el estudio realizado por Becerril-Gaitán *et al.* (2021), se analizaron los datos de 29,416 mujeres embarazadas en México, de las cuales el 39 % resultó positivo al SARS-CoV-2. Los resultados mostraron que el riesgo de mortalidad materna fue 3.24 veces mayor en las mujeres positivas en comparación con las negativas ($p < 0.01$). Desde julio de 2020, COVID-19 se identificó como la principal causa de muerte materna, representando más del 50 % de las defunciones en 2021. Asimismo, la enfermedad renal crónica (RM = 4.11; $p < 0.01$) y la diabetes (RM = 2.53; $p < 0.01$) se identificaron como las comorbilidades más fuertemente asociadas con la mortalidad materna por COVID-19. Los autores concluyen que las comorbilidades presentes durante el embarazo —particularmente aquellas que incrementan la respuesta inflamatoria o alteran la función inmunitaria— aumentan significativamente el riesgo de muerte materna por COVID-19 en mujeres gestantes mexicanas.

4. Metodología

4.1. Clustering

Sea un conjunto de datos de n puntos en un espacio d-dimensional,$D = \{X_i\}_{i=1}^{n}$, y dado un número de clústeres k, $\mathbb{C} = \{C_1, C_2, ..., C_k\}$ tal que

$$\bigcup_{i=1}^{k} C_r = \{X_i\}_{i=1}^{n}$$
$$C_\zeta \bigcap C_\xi = \varnothing \quad \text{para} \quad \zeta \neq \xi \tag{1}$$

Un clúster C_i queda representado por un punto representativo llamado *centroide*, que en muchas ocaciones es μ_i.

4.1.1. Density-Based Clustering

El método de agrupamiento basado en densidad (Density-Based Clustering) utiliza la densidad local de los puntos para determinar los clústeres, en lugar de considerar únicamente la distancia entre ellos.De acuerdo con la descripción presentada por Zaki y Meira (2014), se define una esfera de radio ε alrededor de un punto $x \in \mathbb{R}^d$, denominada ε-vecindad de x, como el conjunto:

$$N_\varepsilon(x) = \{y \quad | \quad \delta(x,y) \leq \varepsilon\} \tag{2}$$

donde $\delta(x,y)$ representa la distancia entre los puntos x e y, en nuestro caso de estudio utilizaremos la metrica de (incertar metrica).Para cualquier punto $x \in D$, se dice que x es un punto núcleo (core point) si existen al menos minPts puntos dentro de su ε-vecindad, es decir, si $|N_\varepsilon(x)| \geq$ minPts. Un punto frontera (border point) es aquel que no cumple el umbral minPts, es decir, $|N_\varepsilon(x)| <$ minPts, pero pertenece a la vecindad de algún punto núcleo z, es decir, $x \in N_\varepsilon(z)$. Finalmente, si un punto no es ni núcleo ni frontera, se le denomina punto de ruido (noise point) o atípico (outlier).

Mencionan ademas que un punto x es *directamente alcanzable por densidad* desde otro punto y si $x \in N_\varepsilon(y)$ y y es un *punto núcleo*. De manera más general, se dice que x es *alcanzable por densidad* desde y si existe una cadena de puntos $x_0, x_1, \ldots, x_l$ tal que $x = x_0$ y $y = x_l$, y cada x_i es directamente alcanzable por densidad desde x_{i-1} para todo $i = 1, \ldots, l$.

4.1.2. DBSCAN

Una de las limitaciones de *DBSCAN* es su sensibilidad a la elección del parámetro ε, especialmente cuando los clústeres presentan diferentes densidades. Si el valor de ε es demasiado pequeño, los clústeres menos densos pueden ser clasificados como ruido. Por el contrario, si ε es demasiado grande, los clústeres más densos pueden fusionarse entre sí. En otras palabras, cuando existen regiones con distintas densidades locales,un único valor de ε puede no ser suficiente para identificar correctamente todas las estructuras presentes en los datos.

4.2. Redes neuronales

5. Resultados

Se muestran y explican los resultados obtenidos.

6. Conclusiones y discusión

Se exponen las conclusiones de la investigación

Data Availability

We encourage authors to include a Data Availability Statement in their manuscript. This statement should include information on where resources such as data, materials, protocols and software code can be accessed. If data sharing is not applicable, authors should state that 'Data sharing is not applicable to this article as no new data were created or analysed in this study.

References

De Salud, S. (s.f.). Datos abiertos Dirección General de Epidemiología. gob.mx. Disponible en: https://www.gob.mx/salud/documentos/datos-abiertos-152127

Treviño, Jesús A. 2020, *Demografía, comorbilidad y condiciones médicas de los pacientes hospitalizados por Covid-19 en México. Middle Atlantic Review of Latin American Studies*, 4(1), 49–70. Disponible en: https://doi.org/10.23870/marlas.317

Becerril-Gaitán, A., Matías-García, B., Cruz-Domínguez, M. del P., Cruz-Domínguez, M. del P., Machorro-Lazo, M. V., León-Juárez, M., y

Mancilla-Ramírez, J. 2021, *La pandemia de COVID-19 y su relación con la mortalidad materna en México*. *Gaceta Médica de México*, 157(6), 618–624. Disponible en: https://www.scielo.org.mx/scielo.php?pid=S0016-38132021000600618&script=sci_arttext

Zaki, Mohammed J., y Meira Jr., Wagner. 2014, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press.