ANÁLISIS DE DATOS DEMOGRÁFICOS DE PACIENTES DIAGNOSTICADOS CON COVID-19 EN LA POBLACIÓN DE MÉXICO

Henrry I. Arias M.*, Jose E. Perez M.**y Fredy A. Huanca T.***

Abstract

The combination of advanced data analysis techniques allow us to explore the complexity of the data, identify underlying patterns and obtain valuable information for decision making. The objective of this article is to segment the data associated with COVID-19 cases in Mexico in order to find demographic patterns in the patients who attended a care center. A comparison will be made between the influence of clustering proposals with dimensionality reduction and without dimensionality reduction on the segmentation of these patterns. Also, by gaining a deeper insight into the COVID-19 data, we will be able to better understand the factors that contributed to the spread of the disease, and the disparities across different population groups (indigenous and non-indigenous). This article focuses on the analysis of information on Covid-19 in Mexico using techniques such as K-means, autoencoder and PCA.

Keywords COVID-19, Mexico, analysis of data, segmentation, patterns, autoencoder, dimensionality reduction, PCA.

1 INTRODUCCIÓN

El Covid-19, causado por el virus SARS-CoV-2, ha generado una crisis sanitaria global sin precedentes. Desde su aparición en diciembre de 2019, la enfermedad se ha propagado rápidamente en todo el mundo, afectando a millones de personas y causando un impacto significativo en la salud pública, la economía y la sociedad en general. México no ha sido ajeno a esta situación, enfrentando desafíos importantes en la contención y control del virus.

En este contexto, el análisis de la información relacionada con el COVID-19 se ha convertido en una herramienta esencial para comprender la evolución de la enfermedad, identificar patrones y tomar decisiones informadas para su gestión y mitigación. En este artículo científico, nos enfocaremos en el análisis de datos del COVID-19 en México, utilizando técnicas como K-means, autoencoder y PCA para la reducción de dimensionalidad y la búsqueda de distintos tipos de patrones.

El objetivo principal de este estudio es contribuir al conocimiento científico existente sobre el COVID-19 en México, a través del análisis de una amplia gama de variables relacionadas con la enfermedad. Estas variables pueden incluir datos demográficos, tasas de infección, hospitalizaciones, mortalidad, distribución geográfica de los casos, medidas de control implementadas y otros factores relevantes. La combinación de técnicas avanzadas de análisis de datos nos permitirá explorar la complejidad de estos datos, identificar patrones subyacentes y obtener información valiosa para la toma de decisiones.

El uso de K-means nos permitirá realizar agrupaciones de los datos, lo que facilitará la identificación de patrones similares entre diferentes grupos de casos. El autoencoder, por otro lado, nos brindará una forma de reducir la dimensionalidad de los datos, extrayendo características clave y revelando patrones más complejos y sutiles. Además, el análisis de componentes principales (PCA) nos ayudará a comprender la estructura de los datos, identificar las variables más influyentes y examinar las relaciones entre ellas.

Al obtener una visión más profunda de los datos del COVID-19 en México, podremos comprender mejor los factores que contribuyen a la propagación de la enfermedad, la eficacia de las medidas de control y las disparidades en la afectación de diferentes grupos de población. Esto, a su vez, permitirá a los responsables de la salud pública y a los tomadores de decisiones implementar estrategias más efectivas y basadas en evidencia para contener el virus y proteger a la población.

En conclusión, este artículo científico se enfoca en el análisis de información del COVID-19 en México utilizando técnicas como K-means, autoencoder y PCA. La combinación de estas herramientas avanzadas nos permitirá identificar patrones ocultos en los datos, proporcionando conocimientos esenciales para el manejo de la pandemia. A través de este estudio, esperamos contribuir al avance de la investigación en salud pública y proporcionar información relevante para la toma de decisiones en la lucha contra el COVID-19 en México y en todo el mundo.

2 TRABAJO RELACIONADOS

Los autores revisan los métodos más relevantes para el análisis de datos, incluyen reducción de dimensionalidad, A continuación, se resumen algunos de los métodos mencionados:

El artículo [4] analiza los factores de riesgo asociados con la mortalidad en pacientes con COVID-19 en México. Se encontró que la edad, el sexo y las comorbilidades como la diabetes, la obesidad y la hipertensión aumentan significativamente el riesgo de muerte. Además, las enfermedades menos frecuentes como la enfermedad pulmonar obstructiva crónica, la enfermedad renal crónica y las condiciones de inmunosupresión también aumentan el riesgo de muerte. La hospitalización y la neumonía también se identificaron como factores de riesgo significativos. El estudio destaca la importancia de comprender los factores de riesgo asociados con COVID-19 para prevenir y manejar mejor la propagación del virus.

* Universidad Nacional San Agustín de Arequipa, hariasm@unsa.edu.pe
**Universidad Nacional San Agustín de Arequipa, jperezma@unsa.edu.pe
**Universidad Nacional San Agustín de Arequipa, fhuancat@unsa.edu.pe

En el artículo [3], se utiliza el algoritmo de clustering K-means para segmentar a los clientes de un banco en función de sus hábitos de uso de tarjetas de crédito. El objetivo es ayudar al banco a ejecutar campañas de marketing efectivas dirigidas a clientes específicos. El artículo menciona que el número óptimo de clusters se puede determinar utilizando el método de la curva del codo y que la visualización y el preprocesamiento de datos son pasos importantes en el desarrollo del modelo. También se menciona que la arquitectura del autoencoder se puede utilizar para reducir el número de clusters y mejorar la precisión del modelo.

En el estudio [5] utilizó un enfoque de clustering para identificar diferentes patrones en las tasas de incidencia y mortalidad de COVID-19 en 206 países. Los autores calcularon 27 medidas resumen para cada trayectoria, utilizaron análisis factorial para capturar correlaciones entre las medidas y aplicaron un algoritmo K-means para agrupar las trayectorias. Encontraron tres patrones diferentes tanto para las tasas de incidencia como de mortalidad, y la variación parecía estar más relacionada con las políticas de salud gubernamentales que con la distribución geográfica. El estudio proporciona información importante para los responsables políticos e investigadores para comprender los patrones epidemiológicos y comportamientos de COVID-19 en las sociedades.

El artículo [6] analiza el impacto del COVID-19 en un conjunto de datos a nivel de país utilizando técnicas de aprendizaje automático no supervisado. Los autores utilizan el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos e identificar variables significativas, y luego aplican el enfoque de agrupamiento K-means para detectar estructuras de comunidad ocultas entre los países. Los resultados muestran que las comunidades logradas después de aplicar PCA son más precisas que las logradas sin ella. El artículo sugiere que el uso de PCA puede mejorar la identificación de comunidades y ser útil para investigadores, científicos, sociólogos, responsables políticos y gerentes del sector de la salud.

3 OVERVIEW

4 ANÁLISIS DE TAREAS

Descripción de los datos y preprocesamiento

El conjunto de datos obtenido de la página de *Datos Abiertos del Gobierno de México* abarca casos de COVID-19 reportados entre el 01 de enero de 2022 y el 16 de mayo de 2023. Los casos reportados contemplan casos sospechosos, negativos y positivos. El presente trabajo se basará únicamente en los casos confirmados; para ello, el primer paso es eliminar columnas con características que tienen poco valor significativo o son redundantes a nuestro objetivo; estas columnas son la siguientes: FECHA_ACTUALIZACION, ID_REGISTRO, ORIGEN, SECTOR, MUNICIPIO_RES, ENTIDAD_NAC, NACIONALIDAD, MIGRANTE, FECHA_INGRESO, FECHA_SINTOMAS, FECHA_DEF, HABLA_LENGUA_INDIG, TOMA_MUESTRA_LAB, TOMA_MUESTRA_ANTIGENO, RESULTADO_LAB, RESULTADO_ANTIGENO, PAIS_NACIONALIDAD, PAIS_ORIGEN; estas columnas no serán usadas en nuestro análisis de datos.

Con respecto a los datos de columnas de tipo SÍ o NO, éstas tienen valores 1 (SÍ) y 2 (NO), pero los valores de "no aplica", "se ignora" y "no especificado" tienen valores 97, 98 y 99 respectivamente. Estos valores se reemplazaron por valores negativos o 3, según el caso; esto con la finalidad de no generar mucha desviación en cada registro.

Nomenclatura

- *CLASIFICACION_FINAL* identifica los casos confirmados de COVID-19 (valores del 1 al 3, en caso contrario son valores negativos o inválidos).
- *ENTIDAD_UM* es la entidad donde se ubica la Unidad Médica de atención.
- *ENTIDAD_RES* es la entidad de residencia del paciente.
- *EPOC* identifica si el paciente tiene un diagnóstico de EPOC (Enfermedad Pulmonar Obstructiva Crónica).
- *MORTALIDAD_DIAS* contabiliza los días transcurridos desde el diagnóstico hasta el fallecimiento del paciente.
- *OTRO_CASO* indica si el paciente tuvo contacto con algún otro caso diagnosticado con COVID-19.
- *OTRAS_COM* identifica si el paciente tiene un diagnóstico de otras enfermedades.
- *TIPO_PACIENTE* identifica el tipo de atención que recibió el paciente, ambulatorio si regresó a su casa u hospitalizado si ingresó a hospitalización.

Visualización de los datos

Utilizaremos una matriz de correlación para un análisis multivariable. El resultado se muestra en la figura 1.

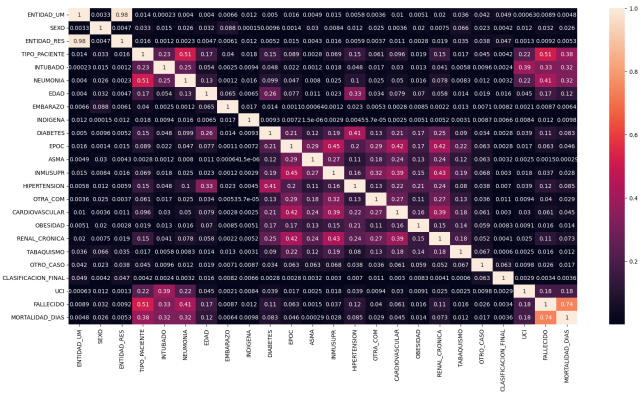

Figura 1: Matriz de correlación para los datos de COVID-19

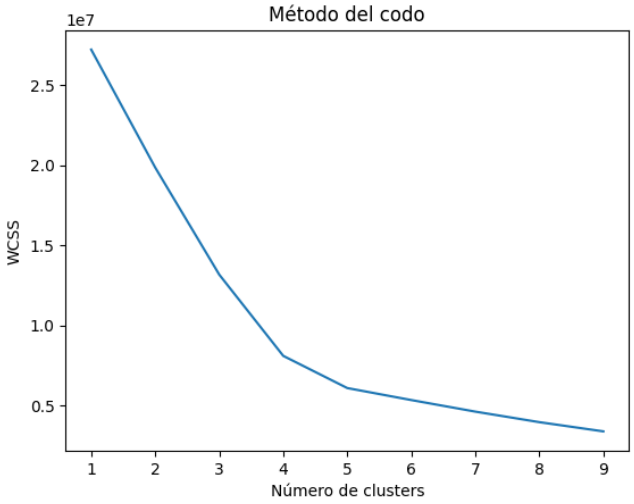

Figura 2: Aplicación del método del codo (*Elbow*) para determinar el número óptimo de clusters

.

5 METODOLOGÍA

6 CASOS DE ESTUDIO

7 DISCUSIÓN

8 CONCLUSIONES

Los detalles del entrenamiento y evaluación del modelo MIRNet para tres tareas de procesamiento de imágenes de bajo nivel: eliminación de ruido, superresolución e imagen mejorada. Para cada tarea, utilizaron cinco conjuntos de datos reales diferentes y compararon el rendimiento de MIRNet con otros métodos del estado del arte. Considerar algunas posibles limitaciones del experimento se podrían incluir:

- Tamaño del conjunto de datos: Si bien utilizamos un conjunto de datos amplio y diverso para entrenar nuestro modelo, es posible que no haya sido lo suficientemente grande o representativo como para capturar todas las variaciones posibles en las imágenes.

- Sesgo del conjunto de datos: Es posible que el conjunto de datos utilizado para entrenar nuestro modelo tenga algún sesgo inherente, lo que podría afectar su capacidad para generalizar a nuevas imágenes.

- Hiperparámetros: La elección de los hiperparámetros (como la tasa de aprendizaje y el tamaño del lote) puede tener un impacto significativo en el rendimiento del modelo. Es posible que no hayamos encontrado los mejores valores para estos hiperparámetros en nuestro experimento.

- Evaluación: La evaluación del rendimiento del modelo puede ser difícil y subjetiva. Es posible que hayamos utilizado métricas inadecuadas o insuficientes para evaluar el rendimiento del modelo.

En la figura **??**, se describe el proceso de experimentación del método propuesto para el procesamiento de imágenes utilizando una red neuronal profunda llamada MIRNet. Se entrenó la red neuronal utilizando un conjunto de datos de entrenamiento y se evaluó en cinco conjuntos de datos diferentes para tareas como denoising, super-resolución e imagen mejorada. Durante el entrenamiento, se utilizaron parches de tamaño 128x128 y operaciones de aumento de datos para mejorar la precisión del modelo. La tasa de aprendizaje disminuyó gradualmente durante el entrenamiento para mejorar la estabilidad del modelo. Los resultados muestran que el método propuesto supera a los métodos existentes en todos los conjuntos de datos evaluados y demuestra una buena capacidad generalización a través de diferentes conjuntos de datos.

Es importante tener en cuenta que el tiempo de ejecución del método dependerá del tamaño y complejidad de la imagen, así como del hardware utilizado para su procesamiento. Imágenes más grandes y complejas requerirán más tiempo para procesar que imágenes más pequeñas y simples. Además, es posible que se requiera un hardware especializado, como una tarjeta gráfica potente, para acelerar el proceso de procesamiento y obtener resultados más rápidos. En general, el proceso de ejecución del método propuesto puede ser relativamente sencillo si se tiene experiencia en el procesamiento de imágenes y acceso a los recursos adecuados.

ESTUDIOS DE POBLACIÓN

Cuadro 1: Impacto de los componentes individuales de MRB.

Skip connections		✓	✓	✓	✓
DAU	✓		✓		✓
SKFF intermediate	✓	✓			✓
SKFF nal	✓	✓	✓	✓	✓
PSNR (in dB)	27.91	30.97	30.78	30.57	31.16

Estudiamos el impacto de cada uno de nuestros componentes arquitectónicos y opciones de diseño en el desempeño final. La sección de estudios de ablación incluye varias tablas que resumen los resultados de los experimentos realizados. El cuadro 1 muestra el impacto de cada componente individual del modelo MRB en la tarea de superresolución. Los resultados indican que las conexiones "skip"son el componente más importante para el rendimiento, ya que su eliminación causa la mayor disminución en la PSNR. El cuadro **??** compara el método propuesto SKFF con otros métodos de agregación de características y muestra que SKFF utiliza menos parámetros pero produce mejores resultados. Finalmente, el cuadro **??** presenta un estudio de ablación sobre diferentes diseños de MRB, donde se varía el número de corrientes paralelas y columnas que contienen DAUs. Los resultados muestran que aumentar el número de corrientes paralelas y columnas puede mejorar el rendimiento del modelo en la tarea de superresolución. En general, estas tablas proporcionan información valiosa sobre cómo cada componente del modelo contribuye al rendimiento general

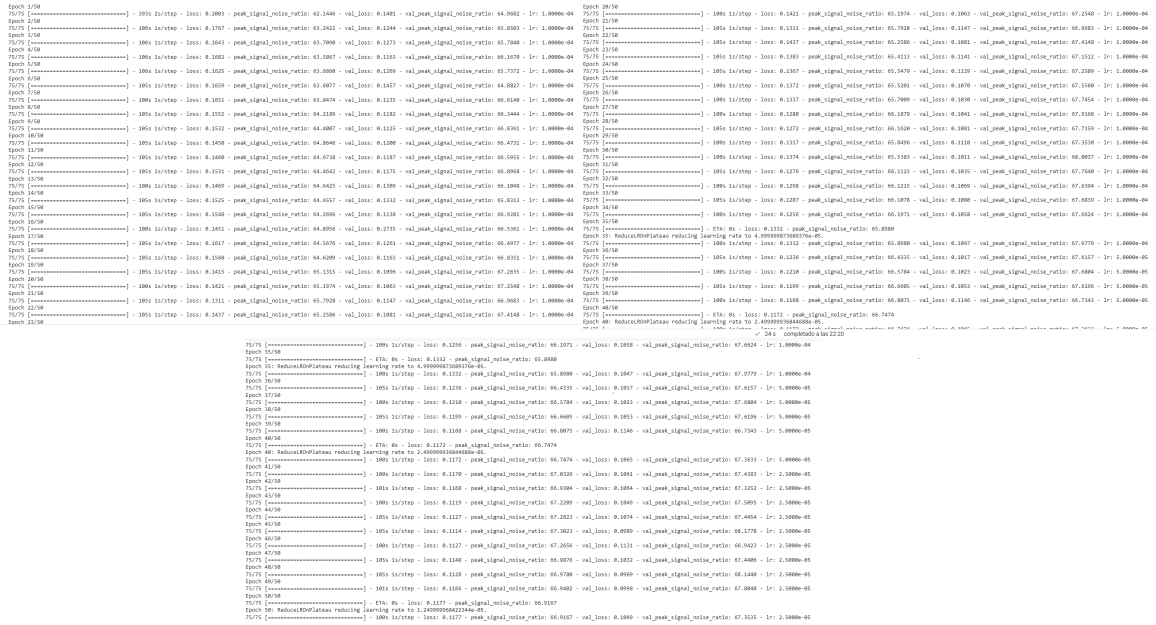

Figura 3: Ejecución del entrenamiento. La figura muestra las 50 épocas con sus respectivas pérdidas.

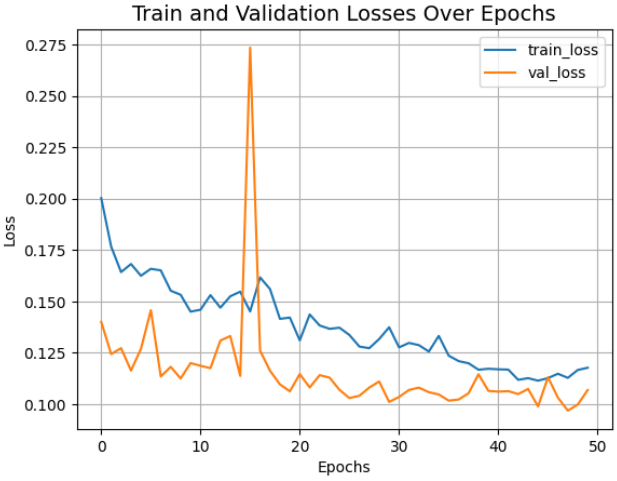

Figura 4: Pérdida de entrenamiento y validación a través de las épocas

y pueden ayudar a guiar futuras mejoras en la arquitectura del modelo.

CONCLUSIÓN

En este estudio, se realizó una revisión exhaustiva de los métodos de aprendizaje, para la eliminación dedimensionalidad. Usando PCA y luego usar Kmeas y .

Se concluyó que los métodos basados en redes neuronales convolucionales (CNN) son los más efectivos para la eliminación de ruido en imágenes [**?**, **?**, **?**]. Además, se encontró que el uso de arquitecturas profundas y técnicas de entrenamiento avanzadas, como la normalización por lotes y la regularización, puede mejorar significativamente el rendimiento del modelo [**?**]. Sin embargo, aún hay desafíos importantes en este campo. Por ejemplo, la eliminación de ruido en imágenes con texturas finas y detalles complejos sigue siendo un problema difícil. Además, muchos métodos existentes requieren grandes cantidades de datos etiquetados para el entrenamiento del modelo, lo que puede ser costoso y limitar su aplicabilidad en situaciones donde los datos son escasos.

También se destacó la importancia del conjunto de datos utilizado para entrenar y evaluar los modelos. Se recomendó el uso de conjuntos de datos grandes y diversos para garantizar que los modelos sean capaces de generalizar bien a diferentes tipos de ruido y condiciones [**?**].

En general, se concluyó que el aprendizaje profundo ha demostrado ser una herramienta poderosa para la eliminación de ruido en imágenes y que hay muchas oportunidades para futuras investigaciones en esta área.

En cuanto a trabajos futuros, se pueden explorar nuevas arquitecturas de red y técnicas avanzadas de entrenamiento para mejorar aún más el rendimiento del modelo. También se pueden investigar métodos para reducir la dependencia del modelo en grandes cantidades de datos etiquetados. Además, se pueden explorar aplicaciones prácticas para la eliminación de ruido en imágenes en campos como la medicina y la astronomía.

REFERENCES

[1] Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan.Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., et al. (2020).*Journal of Allergy and Clinical Immunology*, 22, 1650-1652. `https://doi.org/10.1093/ntr/ntaa059`.

[2] COVID-19 fatality in Mexico's indigenous populations. Argoty-Pantoja, A. D., Robles-Rivera, K., Rivera-Paredez,

B., & Salmerón, J. (2021). *Public health*, 193, 69–75. https://doi.org/10.1016/j.puhe.2021.01.023.

[3] Credit Card Holders Segmentation Using K-mean Clustering with Autoencoder. Dash, D., & Mishra, A. (2022).En *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)* (pp. 1-5). Bhubaneswar, India. https://doi.org/10.1109/ASSIC55218.2022.10088368.

[4] Clinical characteristics and risk factors for mortality of patients with COVID-19 in a large data set from Mexico. Parra-Bracamonte, G. M., Lopez-Villalobos, N., & Parra-Bracamonte, F. E. (2020). *Annals of Epidemiology*, 52, 93-98.e2. ISSN 1047-2797. https://doi.org/10.1016/j.annepidem.2020.08.005.

[5] Clustering of countries according to the COVID-19 incidence and mortality rates.Gohari, K., Kazemnejad, A., Sheidaei, A., et al. (2022). *BMC Public Health*, 22, 632. https://doi.org/10.1186/s12889-022-13086-z.

[6] Community detection using unsupervised machine learning techniques on COVID-19 dataset.Chaudhary, L., & Singh, B. (2021). *Social Network Analysis and Mining*, 11(28), 1-14. https://doi.org/10.1007/s13278-021-00734-2.