MÉTODOS DE LIMPIEZA Y PREPROCESAMIENTO DE DATOS APLICADOS A COVID-19

Jorge Zavaleta

jorge@un.edu.pe

Profesor e Investigador en Ciencia de Datos

Investigador de postdoctorado en la Universidad del Estado de Rio de Janeiro (UERJ)

RESUMEN

La falta de datos es un problema en el análisis de las informaciones contenidas en los datos, afectando un real reconocimiento de patrones y la toma de decisiones en todas las áreas del conocimiento. La colecta y el preprocesamiento de datos son las dos primeras fases del ciclo de vida del análisis de datos, de un total de seis fases. Este ciclo de vida del análisis de datos define un conjunto de mejores prácticas (metodología) a realizar en cada fase, el flujo de ejecución y el resultado en el proceso de análisis [1], [2]. Independientemente de cómo se recopilen los datos, los datos suelen tener errores, lo que significa que es necesario limpiarlos (prepararlos) antes de ser preprocesados. Estos errores pueden ser causados por varios factores como errores humanos, problemas de modelo, problemas de equipos informáticos y electrónicos, valores inesperados, información incompleta, resolución, relevancia de campos, formatos de datos, interferencia del entorno, error de configuración en el proceso de registro de datos, etc., introduciendo lagunas en los datos para la siguiente fase de procesamiento de estos.

La segunda fase es el preprocesamiento o preparación de datos donde los datos son procesados, explorados y acondicionados antes de modelarlos, realizando los procesos de extracción, transformación, carga y/o transformación (ETL/ETLT) para realizar las pruebas y análisis de datos [3], utilizando una infraestructura computacional adecuada, tanto para almacenamiento de alta capacidad como para alta capacidad de entrada/salida. La preparación de datos implica usar métodos para limpiar, combinar, agregar datos o conjuntos de datos, así como elegir algunas muestras apropiadas para el entrenamiento y las pruebas. Esta fase demanda mucho tiempo y es la más laboriosa, gastando más de la mitad del tiempo de un proyecto [4].

Debido a la amplitud del tema, este tema será restringido al problema de falta de datos, que es una ocurrencia común en el análisis de datos y su teoría está estrechamente relacionada con modelos estadísticos y puede tratarse fácilmente utilizando modelos matemáticos simples y/o complejos. El problema de falta de datos es tratado usando métodos de limpieza y preprocesamiento del lenguaje Python y sus bibliotecas numpy y pandas en datasets de COVID-19 relacionados al estado de Rio de Janeiro con el objetivo de tener datos con mayor calidad y confiabilidad resultando en *toma d*e decisiones más exactas a la realidad.

Palabras claves: Preprocesamiento, covid-19, dataset, modelos matemáticos, pyhon

Referencias

[1] E. E. Services, *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, Inc., 2015.

[2] N. A. Mirza, *Practitioner's Guide to Data Science Streamlining: Streamlining Data Science Solutions Using Python, Scikit-Learn, and Azure ML Service Platform*. BPB Publications, India, 2022.

[3] A. Campbell, *Data Science for Beginners: Comprehensive Guide to Most Important Basics in Data Science*. Independently published, 2021.

[4] S. J. Wagh, M. S. Bhende, and A. D. Thakare, *Fundamentals of Data Science*. Boca Raton: Chapman and Hall/CRC, 2021. doi: 10.1201/9780429443237.