Big Data Analytics aplicado al estudio del desempeño académico en un curso universitario

Big Data Analytics applied to the study of academic performance in a university course

Brayan Steven Ramírez Cortes

Universidad Nacional Abierta y a Distancia, Soacha, Colombia

Recibido: 30/09/22 Aprobado 20/10/22

Resumen

El Data Analytics para grandes volúmenes de información aplicado en educación puede aportar de manera importante para los planes de mejora en ámbitos de planeación y evaluación. La siguiente investigación tuvo como objetivo aplicar un EDA (Exploratory Data Analysis) por medio de un análisis multivariado de correlación a las calificaciones de una muestra de 13.490 estudiantes de la UNAD en un curso de primera matrícula. Se encontró que la definitiva de los estudiantes es directamente proporcional en un factor del 99 % con la calificación obtenida en el 75 % del curso y en un factor del 91 % directamente proporcional con la Tarea 3. En contraparte, no hay una correlación significante entre la definitiva y el ejercicio de diagnóstico o la Tarea 1.

Palabras clave: Big Data, Python, educación, análisis de datos, análisis de correlación.

Abstract

Data Analytics for large volumes of information applied in education can make an important contribution to improvement plans in the areas of planning and evaluation. The following research aimed to apply an EDA (Exploratory Data Analysis) through a multivariate correlation analysis to the grades of a sample of 13,490 UNAD students in a first enrollment course. It was found that the final score of the students is directly proportional in a factor of 99 % to the grade obtained in 75 % of the course and a factor of 91 % directly proportional to Task 3. In contrast, there is no significant correlation between the final and the diagnostic exercise or Task 1.

Keywords: *Keywords: Big Data, Python, Education, Data Analytics, correlation Analytics.*

Citación: Ramírez Cortés , B. S. . (2022). Big Data Analytics aplicado al estudio del desempeño académico en un curso universitario. Publicaciones E Investigación, 16(4). https://doi.org/10.22490/25394088.6493

brayan.ramirez@unad,edu.co, https://orcid.org/0000-0002-6039-9530

https://doi.org/10.22490/25394088.6493

1. Materiales y métodos

En la actualidad se requieren y usan nuevas tecnologías para gestionar y extraer información de los datos que se generan en grandes volúmenes a altas velocidades (Menasalvas *et al.,* 2017).

Python es una herramienta que permite realizar análisis estadístico para la toma de decisiones, a partir de librerías como Pandas, Seaborn, Matplotlib, etc. Y ha sido utilizada en investigaciones exploratorias para el análisis de datos, investigaciones de mercado, y análisis de gran cantidad de información (Sahoo *et al.,* 2019; Mittal *et al.,* 2020; McKinney, 2011; Stancin & Jovic, 2019).

Según Sahoo, Kumar Samal, Pramanik, & Kumar Pani (Sahoo *et al.,* 2019), el análisis gráfico exploratorio de datos multivariados (multivariate GEDA en inglés) se usa para entender las conexiones entre diferentes campos en el conjunto de datos o para encontrar las conexiones entre más de dos variables. En este tipo de análisis se usa el diagrama de pares, Pairplot en inglés, para mostrar la vista de todas las variables y su relación.

El proceso se inicia con la carga de los datos al libro de trabajo llamado Jupyter el cual emplea lenguaje de programción Python. se cuenta con las variables, definitiva, Tarea1, Tarea2, Tarea3, 75 %, pretarea, 25 %, entre otras que se tuvieron en cuenta a la hora de recolactar la base de datos. El total de datos recolectados es de 13.940.

Se procede a eliminar los datos en cero para el valor de la definitiva ya que corresponde a estudiantes que aplazaron o cancelaron y no son útiles para el análisis del desempeño académico. Luego de eso se centralizan los datos respecto a la media según el algoritmo presentado por Álvarez Irausquin (2010), para así poder aplicar el procedimiento de correlación multivariante (Pasha & Latha, 2021; Mora García, 2018; García Cazorla, 2016).

2. Resultados y discusión

Respecto a la definitiva se pudo observar que el promedio de calificación es de 321 y el 50 % de los estudiantes obtuvo más de 388 puntos. En la Figura 1 se puede observar la distribución de los datos y como se agrupan a la derecha del promedio, mostrando que la tendencia de los estudiantes es de aprobar el curso. También se puede observar la barra correspondiente a los valores entre 1 y 20 sobrepasa los 1.250 estudiantes, lo que es un indicador de que es necesario revisar los motivos por los cuales se da esta situación.

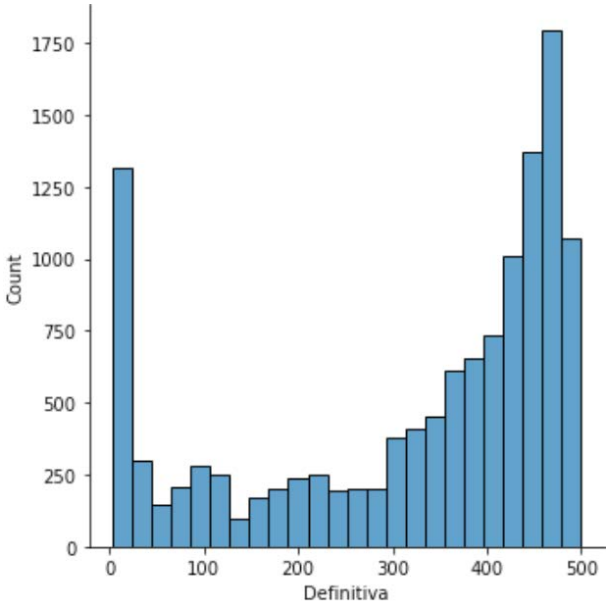

Figura 1. Tabla de frecuencias Definitiva.

En la Figura 2 se puede observar el mapa de calor de las correlaciones de las variables. Se focaliza la variable definitiva y se puede observar que tiene una correlación lineal casi perfecta del 99 % según el coeficiente de correlación de Pearson, lo que indica que un estudiante que apruebe el 75 % del curso muy seguramente aprobará el curso así repruebe el 25 % restante, correspondiente a la evaluación final.

También, se puede observar que la Tarea 3 del curso es la segunda variable con mejor correlación lineal con la definitiva, por lo tanto, el desempeño en esta tarea es un indicador fuerte del desempeño final en el curso.

En cuanto a las correlaciones más débiles, se puede observar que la Pretarea o tarea de diagnóstico, y la Tarea 1 son las que menos se ajustan con la definitiva con un porcentaje de 34 % y 53 %, respectivamente, Esto indica que el desempeño de un estudiante en la parte inicial del curso no está relacionado directamente con la definitiva, se puede observar en la Figura 3 que los datos para estas variables son dispersos, es decir, los estudiantes tienen la posibilidad de mejorar su desempeño después de las dos primeras actividades y lograr aprobar en definitiva.

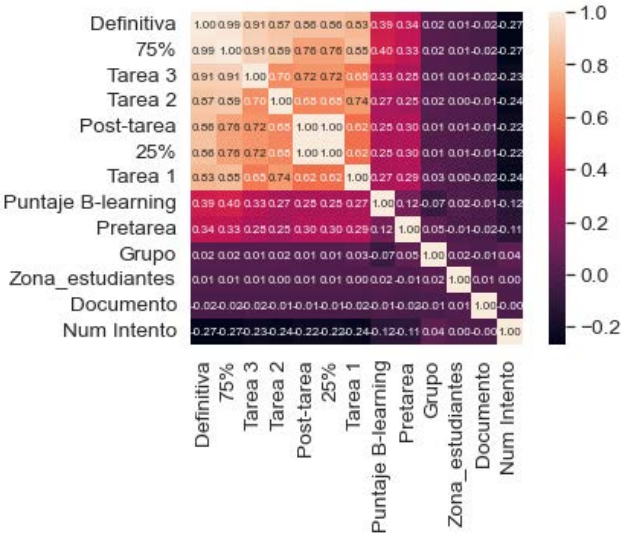

Figura 2. Mapa de calor matriz de correlaciones.

Revisando la gráfica de pares (Figura 3) se puede observar que la dispersión de los datos disminuye conforme se avanza en el curso, es así como en las primeras actividades del curso la dispersión es alta, y para las últimas la relación entre las variables es más estrecha.

Se organizaron las zonas de la universidad respecto a la cantidad de estudiantes que tienen inscritos en el curso en cuestión, se logra observar que la cantidad de estudiantes que posee la zona no está relacionada con el desempeño en la definitiva, es más, es la variable que menos relación tiene con dicha variable.

Otra observación que se puede hacer al revisar la fila de la variable Definitiva es que las dispersiones muestran un comportamiento directamente proporcional, que es débil en las tareas iniciales y se vuelve más fuerte en las últimas tareas.

3. Conclusiones

Como se pudo observar, el Big Data aplicado al análisis del desempeño de un gran volumen de estudiantes arrojó información importante que puede servir para tomar decisiones frente a la estructura del curso, la evaluación, estrategias de mejora, etc. La metodología aplicada se puede extender a otros cursos, promedios académicos, indicadores de deserción u otros campos de análisis en educación, de tal manera que se puedan crear unidades de análisis que permitan entender de mejor manera los datos para poder transformar la realidad educativa a partir de información real que puede ser difícil de rescatar sin el uso de las tecnologías.

Agradecimiento

Agradecimientos de los autores al equipo de la Escuela de Ciencias Básicas, Tecnología e Ingeniería de la Zona centro Bogotá - Cundinamarca, más específicamente al equipo de la UDR Soacha por su apoyo en la consolidación del presente escrito.

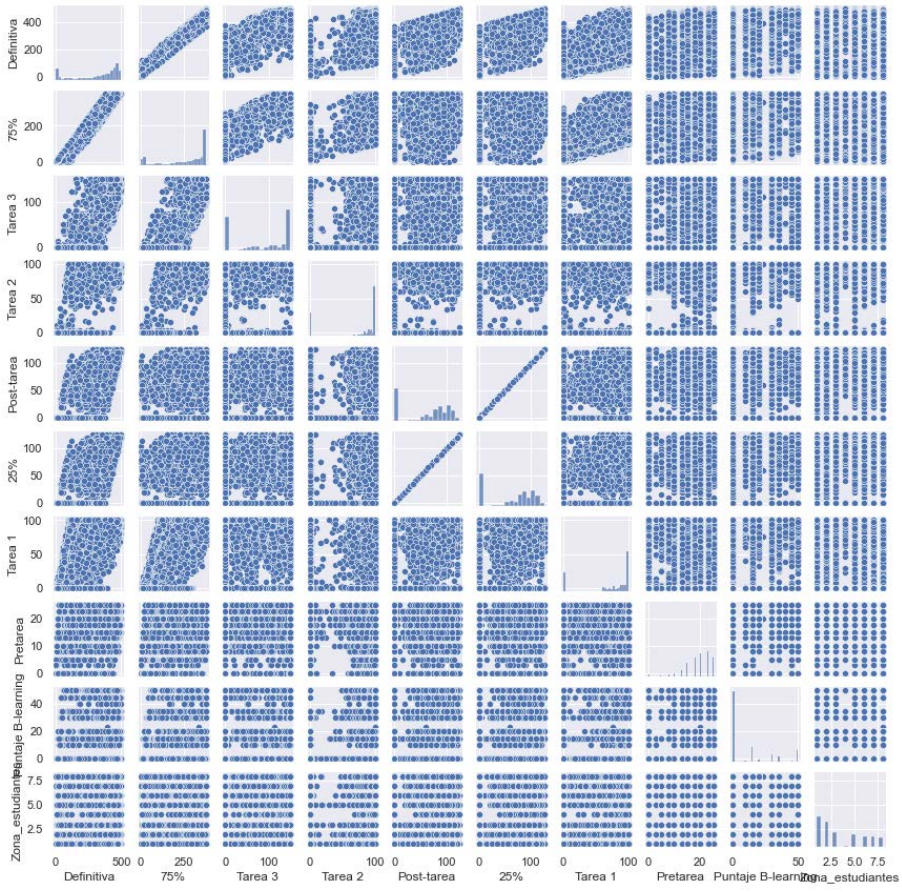

Figura 3. Gráfico de pares para las variables estudiadas.

Referencias

Álvarez Irausquin, W. (2010). Análisis multivariante del Índice de Productividad del Trabajo de América Latina y la Unión Europea por medio de Biplot usando R. *Ingeniería Industrial, 10-5*(19). http://servicio.bc.uc.edu.ve/ingenieria/revista/Inge-Industrial/volv-n19/art02.pdf

García Cazorla, V. (2016). *Valoración de startups con Aprendizaje Automático.* (Trabajo de grado). Universidad Carlos III de Madrid.

McKinney, W. (2011). Pandas: a Foundational Python Library for Data. Deutsches Zentrum für Luft- und Raumfahrt. https://www.semanticscholar.org/paper/pandas%3A-a-Foundational-Python-Library-for-Data-and-McKinney/1a62eb61b2663f8135347171e30cb9dc0a8931b5

Menasalvas, E., Gonzalo, C. & Rodríguez González, A. (2017). Big Data en salud: retos y oportunidades. *Economía industrial, 405*, 87-97. https://dialnet.unirioja.es/servlet/articulo?codigo=6207516

Mittal, P. & Et al. (2020). Sales Analysis and Prediction Using Python. *International Journal of Engineering Research and Applications, 10*(5), (Series-III) , 50-54.

Mora García, O. (2018). *Aplicación de técnicas de Data Analytics para la evaluación y mejora de la calidad en comunicaciones digitales*. (Trabajo de grado). Universidad Carlos III de Madrid.

Pasha, A. & Latha, P. (2021). Well-calibrated probabilistic machine learning classifiers for multivariate healthcare Data. *International Journal of Advanced Research in Computer Science, 12*(2), http://dx.doi.org/10.26483/ijarcs.v12i2.6696

Sahoo, K., Kumar Samal, A., Pramanik, J. & Kumar Pani, S. (2019). Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering 8*(12), 4727-4735. https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35911081219.pdf

Stancin, I. & Jovic, A. (2019). *An overview and comparison of free Python libraries for data mining and big data analysis*. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics. https://ieeexplore.ieee.org/document/8757088.