



Análisis del Rendimiento Académico Estudiantil mediante Técnicas de Procesamiento y Preparación de Datos

Steven Alipio Berrio

Facultad de Ingeniería, Universidad de Antioquia

Resumen

Este informe presenta un análisis comparativo de técnicas estadísticas para la detección de valores atípicos en un conjunto de datos estudiantiles. El objetivo fue evaluar la efectividad del método del Rango Intercuartílico (IQR) en la variable *absences* y del método Z-Score en la variable *G3* (calificación final), considerando las propiedades de distribución de cada una. Se utilizaron visualizaciones como boxplots, histogramas y gráficos comparativos que muestran claramente las diferencias en la sensibilidad de ambos métodos. Los resultados evidencian que IQR detecta un mayor número de atípicos en distribuciones sesgadas, mientras que Z-Score es más adecuado cuando los datos presentan normalidad o baja variabilidad relativa. Este análisis aporta fundamentos para la selección de técnicas apropiadas de preprocesamiento en contextos educativos.

Palabras clave:

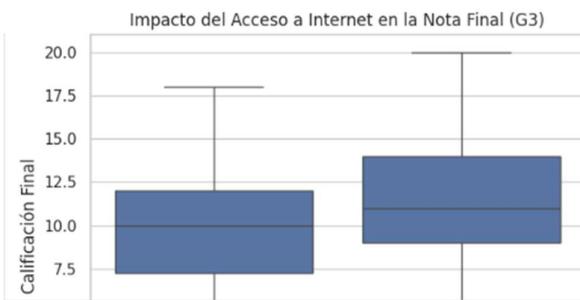
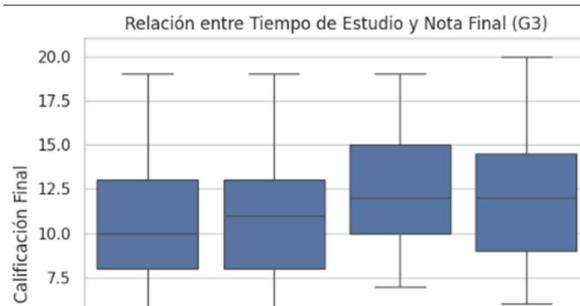
valores atípicos, IQR, Z-Score, preprocesamiento, análisis exploratorio.

1. Introducción

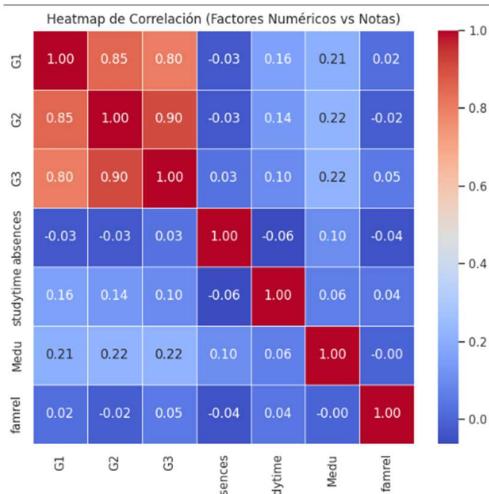
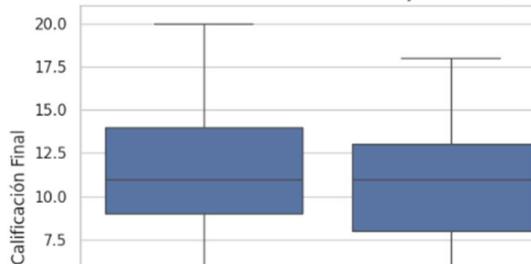
El análisis de datos en contextos educativos requiere especial atención al tratamiento de valores extremos, ya que estos pueden afectar la validez de modelos predictivos, análisis estadísticos o procesos de toma de decisiones. Variables como las ausencias (*absences*) y las calificaciones finales (*G3*) suelen presentar comportamientos anómalos o distribuciones no convencionales que requieren técnicas de detección específicas.

Detectar valores atípicos permite mejorar la calidad del dataset, reducir sesgos y garantizar una interpretación más precisa. La literatura reconoce métodos como el Rango Intercuartílico (IQR) como una técnica robusta frente a distribuciones sesgadas, mientras que el Z-Score destaca en datos con comportamiento cercano a la normalidad.

El propósito de este informe es realizar un análisis comparativo empleando ambos métodos, evaluando su aplicabilidad y resultados en dos variables con características estadísticas distintas. Esto permite justificar la elección metodológica en procesos de preprocesamiento y contribuir a una comprensión más profunda de los fenómenos registrados en datos estudiantiles.



Relación entre Situación Romántica y Nota Final (G_3) valores extremos.



2. Marco teórico.

La detección de valores atípicos (outliers) constituye un componente fundamental del preprocesamiento de datos, ya que la presencia de valores extremos puede distorsionar estadísticas descriptivas, afectar la calidad de modelos predictivos y generar interpretaciones erróneas. Según Hawkins (1980), un valor atípico es una observación que se desvía marcadamente del patrón general del conjunto de datos.

Existen múltiples enfoques para su detección; entre los más utilizados se encuentran los métodos basados en la dispersión estadística, particularmente el **Rango Intercuartílico (IQR)** y el **Z-Score**.

El **IQR** es una medida robusta frente a distribuciones asimétricas, ya que se basa en los cuartiles Q1 y Q3, reflejando la dispersión central del 50% de los datos. Se considera valor atípico todo punto que excede los límites:

$$Q1 - 1.5 \cdot IQR \text{ o } Q3 + 1.5 \cdot IQR.$$

Este método es recomendado cuando las distribuciones presentan sesgo positivo o negativo significativo, ya que no depende de la media ni de la desviación estándar, las cuales se ven afectadas por

Por otro lado, el **Z-Score** estandariza los datos transformándolos a una distribución con media 0 y desviación estándar 1. Un valor se considera atípico si su Z-Score excede el umbral habitual de $|Z| > 3$. Este método asume que la distribución subyacente sigue una forma aproximadamente normal y puede resultar poco sensible en presencia de sesgo o varianza inflada.

Ambas técnicas son ampliamente utilizadas en análisis exploratorio y modelado estadístico. Su aplicación depende de las características de la variable estudiada: forma de la distribución, presencia de colas largas, y nivel de desviación estándar. En contextos educativos, donde variables como ausencias o calificaciones pueden presentar patrones heterogéneos, es crucial seleccionar el método adecuado para garantizar una limpieza de datos coherente y validada metodológicamente.

3. Metodología o desarrollo experimental.

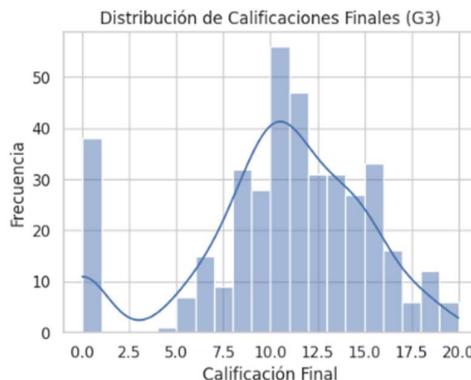
3.1 Fuente y características del conjunto de datos

El dataset analizado corresponde a registros estudiantiles utilizados comúnmente en estudios de desempeño académico. Incluye variables numéricas como *absences* (días de ausencia) y G_3 (calificación final), las cuales presentan variaciones y comportamientos estadísticos relevantes para la detección de outliers.

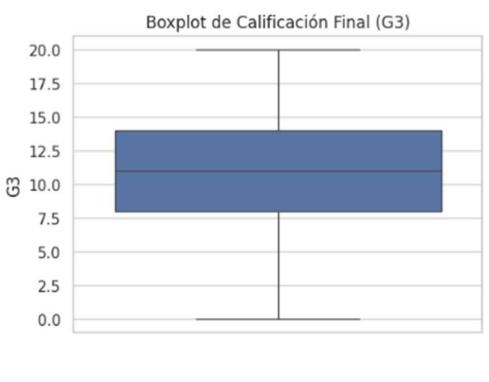
3.2 Técnicas empleadas

3.2.1 Análisis exploratorio (EDA)

- Evaluación de distribuciones, sesgo, frecuencia y dispersión.
- Gráficas iniciales mediante histogramas y boxplots.



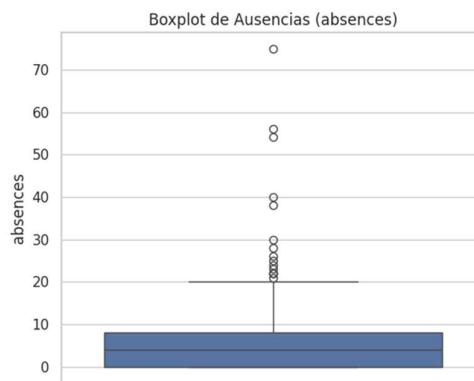
Distribucion de Calificaciones Finales (G3)



Boxplot Calificacion Final

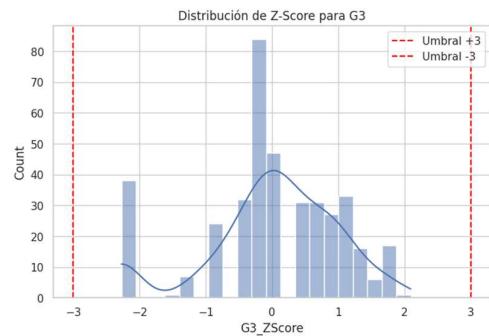
3.2.2 Detección de valores atípicos

- **IQR para absences:**
Se calcularon Q1, Q3 y el rango intercuartílico para definir límites superior e inferior. Valores más allá de $1.5 \cdot \text{IQR}$ fueron marcados como atípicos.



Detección y Análisis de Atípicos en ABSENCES

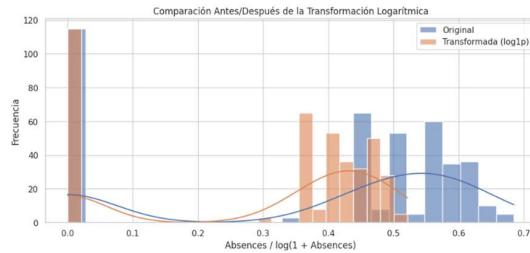
- **Z-Score para G3:**
Se estandarizó la variable y se consideró como atípico todo valor con $|Z| > 3$.



Detección de Atípicos con Z-Score en G3

Imputación y transformación

Definición de lineamientos para tratamiento posterior dependiendo del análisis.



Transformación Logarítmica en 'absences'

3.3 Herramientas utilizadas

- Python 3.10
- Librerías: **pandas, numpy, seaborn, matplotlib**
- Entorno: Google Colab / Jupyter Notebook

3.4 Criterios de tratamiento aplicados

- Variables sesgadas → técnicas robustas como IQR.
- Variables cercanas a la normalidad → estandarización y Z-Score.
- Priorización de gráficos comparativos para interpretar sensibilidad.

4 Resultados

4.1. Resultados del método IQR (Absences)

El cálculo del rango intercuartílico reveló límites que identificaron un número significativo de estudiantes con ausencias extremas. La distribución altamente sesgada de esta variable provocó que el método fuera más sensible a detectar puntos alejados del comportamiento típico. El boxplot permitió visualizar claramente estos valores extremos y su distancia respecto al grueso de la población.

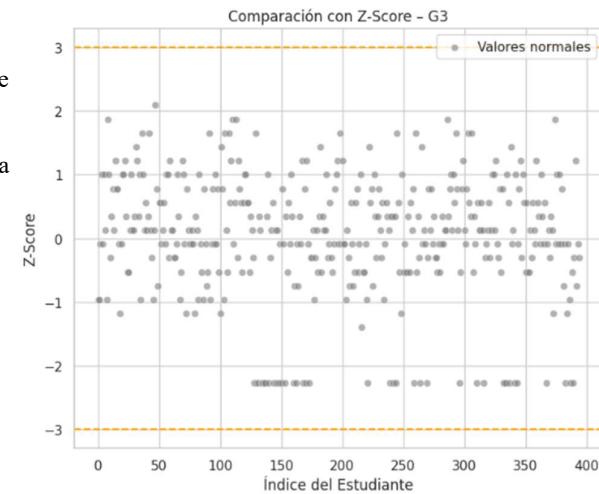
4.2. Resultados del método Z-Score (G3)

La estandarización mediante Z-Score mostró que la variable *G3* presenta un comportamiento más cercano a la normalidad, evidenciado en un histograma centrado y con colas moderadas. Bajo el umbral de $|Z| > 3$, se detectaron pocos o ningún valor atípico, demostrando que este método es menos sensible cuando la variabilidad de los datos es baja o cuando la distribución es estable.

4.3. Comparación directa entre IQR y Z-Score

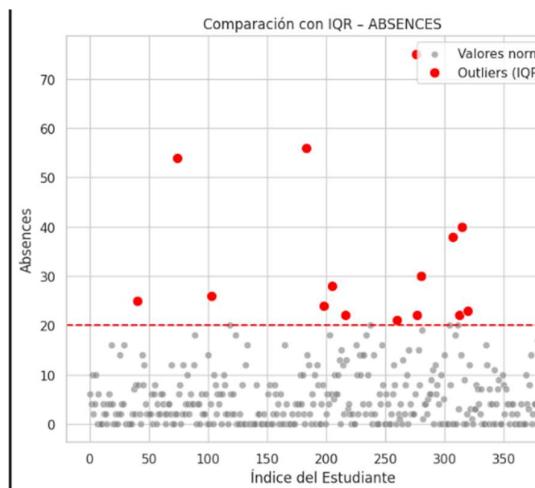
Los gráficos comparativos mostraron que:

- IQR detectó un mayor número de valores atípicos**, debido al sesgo positivo de *absences* y a la presencia de colas largas.



- La sensibilidad de cada método depende de la forma de la distribución:
 - IQR → apropiado para datos no normales
 - Z-Score → apropiado para datos simétricos o normalizados

Esta comparación evidencia la necesidad de seleccionar técnicas de acuerdo con la estructura estadística de cada variable y no aplicarlas indiscriminadamente.



- **Z-Score detectó menos outliers**, pues la desviación estándar y la forma de la distribución de *G3* mantienen la mayoría de los valores dentro de los límites ± 3 .

5. Análisis de Resultados

La aplicación del método IQR sobre la variable *absences* reveló un comportamiento altamente sesgado hacia la derecha, caracterizado por la presencia de colas largas y valores significativamente superiores al rango típico. Los cuartiles mostraron un rango intercuartílico estrecho en comparación con la amplitud máxima de la variable, lo que llevó a la detección de un número considerable de valores atípicos. El boxplot permitió visualizar de manera inmediata la distancia entre la mayoría de los estudiantes y aquellos con niveles de ausencias extraordinariamente altos. Esto sugiere que las ausencias no siguen una distribución normal, sino que presentan concentración en valores bajos y casos aislados significativamente elevados, posiblemente asociados a situaciones excepcionales o comportamientos irregulares de asistencia.

En contraste, el análisis mediante Z-Score aplicado a *G3* (calificación final) evidenció una distribución mucho más estable y simétrica. La mayoría de los

valores se situaron dentro de ± 2 desviaciones estándar y el cálculo mostró que pocos o ningún estudiante alcanzó valores absolutos superiores al umbral de $|Z| > 3$. El histograma de Z-Scores confirmó la concentración alrededor de la media y la ausencia de colas pronunciadas. Esto indica que el rendimiento académico final presenta una variabilidad controlada y que la desviación estándar no está siendo afectada por valores anómalos.

La comparación directa entre ambos métodos señala que **IQR es mucho más sensible en variables sesgadas**, mientras que **Z-Score es más adecuado para distribuciones cercanas a la normalidad**. En *absences*, Z-Score habría pasado por alto valores extremos debido a la alta desviación estándar causada por su sesgo, mientras que IQR detectó adecuadamente estos casos. En *G3*, donde la dispersión es baja y la distribución más simétrica, el método Z-Score no identifica outliers porque efectivamente no existen valores extremadamente alejados del comportamiento común.

En términos prácticos, los resultados demuestran que **la selección del método de detección debe basarse en la estructura estadística de cada variable**. Aplicar un solo método de manera indiscriminada podría llevar a conclusiones incorrectas o a una limpieza deficiente del conjunto de datos. Por ello, la combinación de ambos métodos permitió evaluar con mayor precisión la naturaleza de los outliers presentes en el dataset estudiantil.

6. Conclusiones

El análisis realizado demuestra que la detección de valores atípicos debe adaptarse al comportamiento estadístico de cada variable. El método IQR mostró una alta sensibilidad para identificar valores extremos en distribuciones sesgadas como la de *absences*, mientras que Z-Score resultó más adecuado para la variable *G3*, cuya distribución presenta mayor estabilidad y menor dispersión relativa.

La principal contribución del presente trabajo es la comparación visual y analítica de ambos enfoques, lo cual proporciona criterios sólidos para seleccionar métodos de limpieza de datos en etapas de preprocessamiento.

Entre las limitaciones se encuentra el análisis aislado de variables individuales (univariado), sin considerar correlaciones entre atributos.

Referencias

- [1] Iglewicz, B., & Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.
- [2] Scikit-Learn Developers. (2024). *Scikit-Learn Documentation: Preprocessing*. <https://scikit-learn.org/stable/modules/preprocessing.html>
- [3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.