Análisis Exploratorio de Datos (EDA) para el Área de Ingeniería de Sistemas en la Universidad del Sinú

Geovany deavila medrano,

Ingeniería de Sistemas

Universidad del sinu Elías bechara zainum cartagena, Colombia

Geovanydeavila@unisinu.edu.co

Recibido: Abril, 2024. Accepted: 08/04, 2024.

Resumen

Esta propuesta de proyecto propone realizar un Análisis Exploratorio de Datos (EDA) para el área de Ingeniería de Sistemas en la Universidad del Sinú. Utilizaremos la herramienta Plotly para visualizar y analizar los datos, lo que permitirá identificar patrones, tendencias y relaciones que podrían no ser evidentes inicialmente. Se obtendrá todo tipo de información que puedan proporcionar información valiosa para la mejora de procesos y la toma de decisiones en el ámbito académico y administrativo.

Palabras Clave: EDA, Plotly, Ingeniería de Sistemas, Universidad del Sinú, análisis de datos, visualización, patrones, tendencias, toma de decisiones.

Abstract

The objective of this study is to apply EDA techniques to analyze and understand data related to the area of Systems Engineering at the University of Sinú. The Plotly tool will be used to create interactive visualizations that facilitate this analysis.

I. Introducción

El análisis de datos se ha convertido en una parte integral de cualquier campo, incluyendo la Ingeniería de Sistemas. Este proyecto busca utilizar el EDA para analizar los datos del área de Ingeniería de Sistemas en la Universidad del Sinú.

El objetivo de este estudio es aplicar el EDA utilizando la herramienta Plotly en el contexto de la Ingeniería de Sistemas en la Universidad del Sinú. Con este análisis, se buscará identificar patrones, tendencias y relaciones en los datos relevantes para el área, para generar conocimiento útil para la toma de decisiones.

Al comprender mejor los datos relacionados con el rendimiento académico de los estudiantes y otros aspectos clave, la universidad podrá optimizar sus recursos y estrategias, promoviendo así la excelencia académica y la eficiencia institucional.

II. Objetivo General

- Analizar el rendimiento académico de los estudiantes del área de Ingeniería de Sistemas en la Universidad del Sinú mediante un Análisis Exploratorio de Datos (EDA), utilizando la herramienta Plotly para visualizar y entender los patrones y tendencias que pueden influir en su desempeño académico.

III. Objetivos especificos

- Recolectar y preprocesar los datos necesarios relacionados con el rendimiento académico de los estudiantes de Ingeniería de Sistemas en la Universidad del Sinú.
- Realizar un análisis estadístico de los datos para identificar patrones y tendencias que puedan influir en el rendimiento académico de los estudiantes.
- Utilizar Plotly para visualizar los datos y los resultados del análisis, facilitando la comprensión de los factores que afectan el rendimiento académico.
- Identificar áreas de mejora o intervención basándose en los resultados del análisis para proponer estrategias que puedan mejorar el rendimiento académico de los estudiantes.

IV. Metodología

La metodología comprende la recolección de datos de diferentes fuentes relevantes para el área de Ingeniería de Sistemas, utilizando técnicas como encuestas, registros académicos y bases de datos institucionales. Posteriormente, se realizará un proceso de limpieza y preparación de datos para su análisis. La herramienta principal para la visualización y análisis exploratorio será Plotly, que permite crear gráficos interactivos y dinámicos. Se aplicarán técnicas estadísticas y de minería de datos para identificar patrones, tendencias y relaciones significativas en los datos.

V. Preguntas de la investigación

- ¿Qué factores influyen más en el rendimiento académico de los estudiantes de Ingeniería de Sistemas en la Universidad del Sinú?
- ¿Cómo se pueden visualizar estos factores y su impacto en el rendimiento académico utilizando Plotly?
- ¿Existen patrones o tendencias en el rendimiento académico de los estudiantes que se puedan identificar a través del Análisis Exploratorio de Datos?
- ¿Cómo se pueden utilizar los hallazgos de este análisis para proponer estrategias de mejora en el rendimiento académico de los estudiantes?

VI. Conclusión de lo que supone hacer este análisis:

Realizar este análisis exploratorio de datos (EDA) para el área de Ingeniería de Sistemas en la Universidad del Sinú supone un esfuerzo significativo en términos de recolección, limpieza y análisis de datos. Sin embargo, este esfuerzo tiene el

potencial de proporcionar información valiosa sobre el rendimiento académico de los estudiantes. Al utilizar la herramienta Plotly, podemos visualizar estos datos de una manera más intuitiva y accesible, lo que facilita la identificación de patrones y tendencias.

VII. Conclusión de lo que se espera del análisis:

A partir de este análisis, esperamos obtener una comprensión más profunda de los factores que influyen en el rendimiento académico de los estudiantes de Ingeniería de Sistemas. Estos datos pueden informar estrategias de mejora e intervención, para mejorar el rendimiento académico de los estudiantes.

VI. Análisis descriptivo:

Para este análisis descriptivo realizaremos los siguientes dos pasos:

- **Resumen estadístico**: En donde Calcularemos medidas estadísticas como la media, mediana, moda, rango, varianza, desviación estándar, etc., para cada variable en mi conjunto de datos.

- **Visualización de datos**: Utilizare gráficos como histogramas, gráficos de barras, gráficos de caja (boxplots), gráficos de dispersión, etc., para visualizar la distribución y las relaciones entre las variables de mi data.

Desarrollo.

Ejecutando el comando ***df.dtypes*** nos damos cuenta de que tipo de datos tiene nuestra data.

```
df.dtypes
```

```
ID_Estudiante                int64
Promedio_Calificaciones    float64
Edad                         int64
Num_Cursos                   int64
dtype: object
```

Como podemos Observar en la Data, hay cuatro columnas con sus respectivos tipos de datos listados a continuación:

- ID_Estudiante: el tipo de dato es int64, lo que indica que esta columna contiene datos enteros de 64 bits.

- Promedio_Calificaciones: el tipo de dato es float64, lo que sugiere que esta columna almacena números decimales de 64 bits, posiblemente promedios de calificaciones que pueden tener valores fraccionarios.
- Edad: al igual que la primera columna, el tipo de dato es int64, indicando que contiene edades expresadas en números enteros.

- Num_Cursos: también es de tipo int64, lo que implica que esta columna representa la cantidad de cursos que cada estudiante ha tomado, y se expresa en números enteros

El siguiente comando nos proporciona un resumen de la data, incluyendo el tipo de clase, el rango de índice e información sobre sus columnas.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   ID_Estudiante           200 non-null    int64
 1   Promedio_Calificaciones  191 non-null    float64
 2   Edad                    200 non-null    int64
 3   Num_Cursos              200 non-null    int64
dtypes: float64(1), int64(3)
```

RangeIndex: El DataFrame tiene un RangeIndex que va de 0 a 199, lo que indica que hay un total de 200 entradas (filas).

Columnas: Hay cuatro columnas en el DataFrame, y la información de cada columna se muestra en una fila separada dentro de la salida:

Columna 0: La primera columna se llama ID_Estudiante. Tiene 200 entradas no nulas, lo que significa que no hay valores faltantes en esta columna. El tipo de datos (Dtype) de esta columna es int64, lo que significa que contiene valores enteros de 64 bits.

Columna 1: La segunda columna se llama Promedio_Calificaciones. Tiene 200 entradas no

nulas, lo que indica que no hay valores faltantes. Su tipo de datos es float64, lo que sugiere que contiene números de punto flotante, que se utilizan típicamente para representar valores decimales.

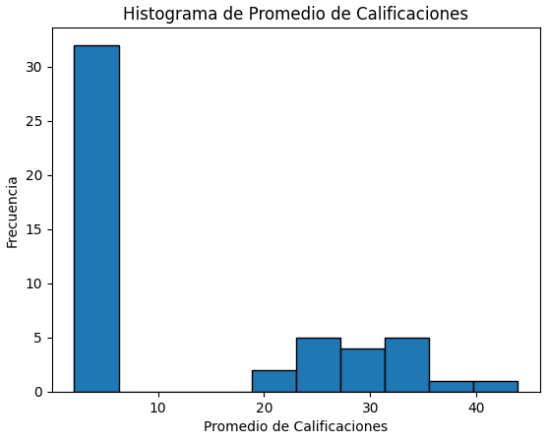
Columna 2: La tercera columna es Edad, con 191 entradas no nulas. Esto sugiere que hay 9 valores faltantes en esta columna (200 entradas totales menos 191 entradas no nulas). El tipo de datos para esta columna también es float64.

Columna 3: La cuarta columna es Num_Cursos, que tiene 200 entradas no nulas, indicando que no hay valores faltantes. El tipo de datos para esta columna es int64.

	ID_Estudiante	Promedio_Calificaciones	Edad	Num_Cursos
count	200.000000	197.000000	200.00000	200.000000
mean	99948.815000	27.424721	23.42000	5.025000
std	299.405207	69.495150	3.54775	2.595778
min	99437.000000	2.020000	18.00000	1.000000
25%	99671.000000	4.000000	20.00000	3.000000
50%	99984.000000	4.800000	24.00000	5.000000
75%	100199.000000	29.590000	27.00000	7.000000
max	100431.000000	567.190000	29.00000	9.000000

1. **ID_Estudiante**:
    - Esta variable es un identificador único para cada estudiante, ya que el conteo es de 200 y no hay desviación estándar, lo que indica que no hay variación en los datos. El valor mínimo es 99437 y el máximo es 100199, lo que sugiere que los ID de los estudiantes están en este rango.
2. **Promedio_Calificaciones**:
    - Esta variable representa el promedio de calificaciones de los estudiantes. El conteo es de 197, lo que indica que hay algunos valores faltantes en los datos. La media es de aproximadamente 27.42, lo que sugiere que el promedio general de calificaciones de los estudiantes es de alrededor de 27.42. La desviación estándar es de aproximadamente 69.50, lo que indica que las calificaciones varían en un rango de alrededor de 69.50 puntos alrededor de la media. Sin embargo, el valor máximo es de 567, lo que parece ser un error ya que normalmente las calificaciones no son tan altas.
3. **Edad**:
    - Esta variable representa la edad de los estudiantes. El conteo es de 200, por lo que no hay valores faltantes. La media es de aproximadamente 23.42 años, lo que sugiere que la edad promedio de los estudiantes es de alrededor de 23 años. La desviación estándar es de aproximadamente 3.55, lo que indica que las edades varían en un rango de alrededor de 3.55 años alrededor de la media.
4. **Num_Cursos**:
    - Esta variable representa el número de cursos que cada estudiante ha tomado. El conteo es de 200, por lo que no hay valores faltantes. La media es de aproximadamente 5.03, lo que sugiere que los estudiantes han tomado un promedio de alrededor de 5 cursos. La desviación estándar es de aproximadamente 2.60, lo que indica que el número de cursos varía en un rango de alrededor de 2.60 cursos alrededor de la media.

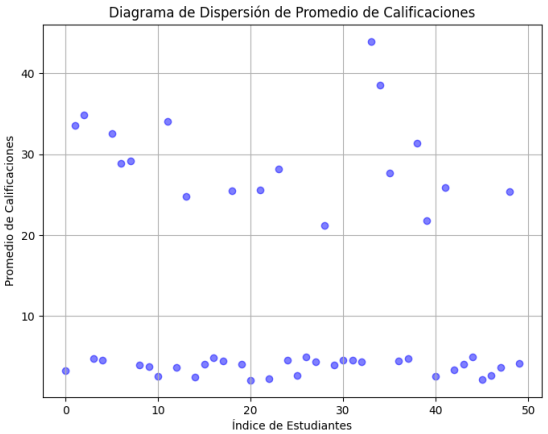
Visualización de datos.



Tomando los primeros 50 datos de la data tenemos que:

Como podemos observar el rango de las calificaciones es de 0 a 5 ya que la mayoría de los estudiantes frecuentemente se encuentran ahí, pero observamos que hay datos fuera de ese rango.

Estos datos pueden ser **valores atípicos** o **outliers**. Los outliers pueden ser causados por errores en la recopilación de datos, variaciones naturales en los datos, o pueden indicar una anomalía o una situación especial.



Como podemos observar gracias al grafico de dispersión podemos ver con mayor claridad estos valores atípicos que se encuentran por fura de la línea de tendencia.

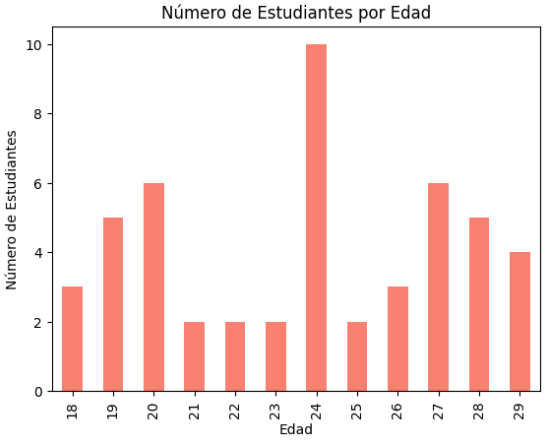
¿en que afectan estos valores a mi data?

Estos tienen un impacto significativo en mi análisis ya que, distorsionan la media, afectan a la desviación estándar e indican que hay algún error en estos datos.

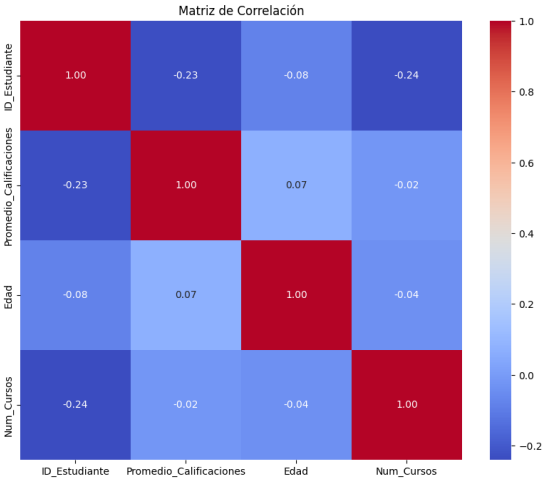
¿Qué decisión tomar?

Eliminación de Valores Atípicos:

eliminar los valores atípicos de mí de data. Esto puede ser apropiado si los valores atípicos son el resultado de errores y no representan la variabilidad natural de tus datos. Sin embargo, la eliminación de valores atípicos puede introducir sesgos.



En este grafico vemos que el pico de edad mas alto es de 24 años esto significa que la mayoría de estudiantes oscilan en esta edad.



parece que no hay una correlación fuerte entre ninguna de las variables. Los coeficientes de correlación varían desde -0.24 hasta 1.00, pero los valores de 1.00 son solo para la correlación de cada variable consigo misma (la diagonal principal de la matriz).

Las correlaciones más fuertes parecen ser entre "ID_Estudiante" y "Promedio_Calificaciones" (-0.23), y entre "ID_Estudiante" y "Num_Cursos" (-0.24). Sin embargo, estos valores son bastante bajos, lo que indica que la correlación es débil.

Por lo tanto, mi conclusión es que, basándome en esta matriz de correlación, no parece haber una correlación fuerte entre ninguna de las variables.

CONCLUSIONES.

- Como ya habíamos identificado anteriormente en el análisis estadístico me di cuenta que teníamos algunos datos con valores faltantes. Por decisión propia borre estos datos ya que no afectaban como tal en el análisis de la data.
- Identifique outliers en el promedio de las calificaciones de los estudiante mediante la visualización de datos ya que habían datos que sobrepasaban la línea de tendencia(rango de notas 1-5) por ende estos datos