



INSTITUTO TECNOLÓGICO DE ESTUDIOS SUPERIORES DE MONTERREY

Maestría en Inteligencia Artificial Aplicada

Proyecto Integrador

Avance 1. Análisis exploratorio de datos

Estudiantes:

A01793188 - Carlos Alberto Reynoso González

A01334976 - Israel Campos Báez

¿Hay valores faltantes en el conjunto de datos? ¿Se pueden identificar patrones de ausencia?

En este caso, no habrá valores faltantes en el conjunto de datos. La razón es que estos datos ni siquiera existen. En cambio, estos datos son virtualmente generados por nosotros mediante un simulador y por lo tanto, toda la disponibilidad y estado son conocidos de antemano. El proceso consiste en varias etapas:

1. Creación de geometrías aleatorias

Empleamos un script en Python para generar rectángulos dentro de un área predefinida. Este script presenta las siguientes características principales:

- **Restricciones espaciales:** Cada rectángulo se construye dentro de una caja delimitadora específica manteniendo un espacio mínimo entre ellos y un margen con respecto a los bordes.
- **Prevención de superposiciones:** El script se asegura de que los rectángulos no se solapen ni se toquen, realizando verificaciones de colisión entre las geometrías previamente generadas.
- **Aleatoriedad controlada:** Las dimensiones de los rectángulos se determinan de manera aleatoria dentro de un rango establecido, garantizando diversidad en las configuraciones de cada modelo creado.

```
• # generate a list of rectangles
• # Define the bounding box (same as outerRectMinX, outerRectMaxX, etc.)
• bounding_box = (-10, 10, -10, 10)
•
• # Set parameters
• num_rectangles = random.randint(1, 10) # Number of rectangles to generate
• min_size = 1 # Minimum size for width/height
• max_size = 2 # Maximum size for width/height
• min_gap = 1 # Minimum gap between rectangles
• border_gap = 1 # Minimum gap from the bounding box borders
•
• # Generate rectangles
• rectangles_data = generate_rectangles_data(bounding_box, num_rectangles,
• min_size, max_size, min_gap, border_gap)
•
• # Print the output in the required format
• print("rectangles_data = [")
• for rect in rectangles_data:
•     print(f"    {rect},")
```

- `print("]")`

2. Importación de geometrías en ANSYS

Una vez que hemos generado las geometrías aleatorias, las importamos a ANSYS para llevar a cabo análisis de elementos finitos (FEA). Este procedimiento incluye los siguientes pasos:

- **Creación del modelo geométrico:** Empleamos un script en ANSYS que toma las coordenadas generadas previamente y las convierte en un modelo geométrico 2D. Cada par de coordenadas (x1, y1, x2, y2) se transforma en un rectángulo dentro de ANSYS, respetando las dimensiones y posiciones indicadas.
- **Definición de condiciones de contorno y materiales:** En ANSYS, asignamos las propiedades de los materiales y las condiciones de contorno necesarias para el análisis estructural, tales como las fuerzas aplicadas y las restricciones en los bordes del modelo.
- **Simulación de elementos finitos:** Un análisis por elementos finitos se realiza para obtener los valores del estrés en varios nodos sobre la geometría. Los resultados se guardan en archivo de texto .txt, en el que cada línea tiene el número de nodo y el valor de estrés.

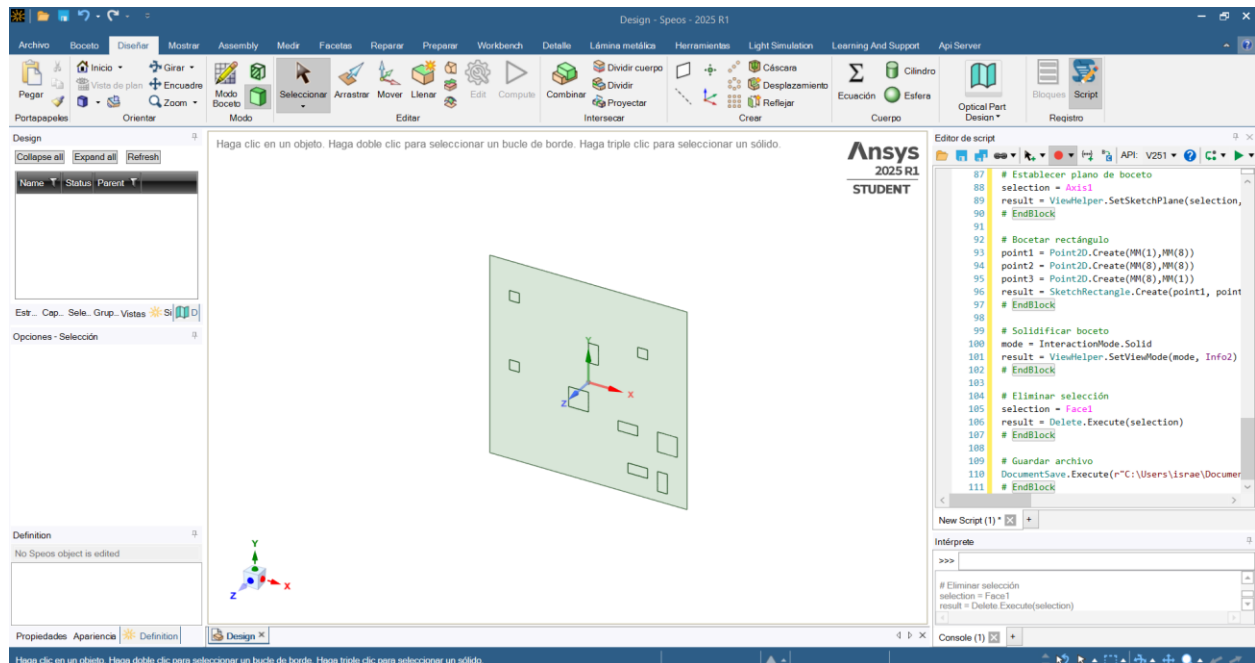
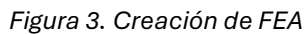


Figura 1. Creación Forma



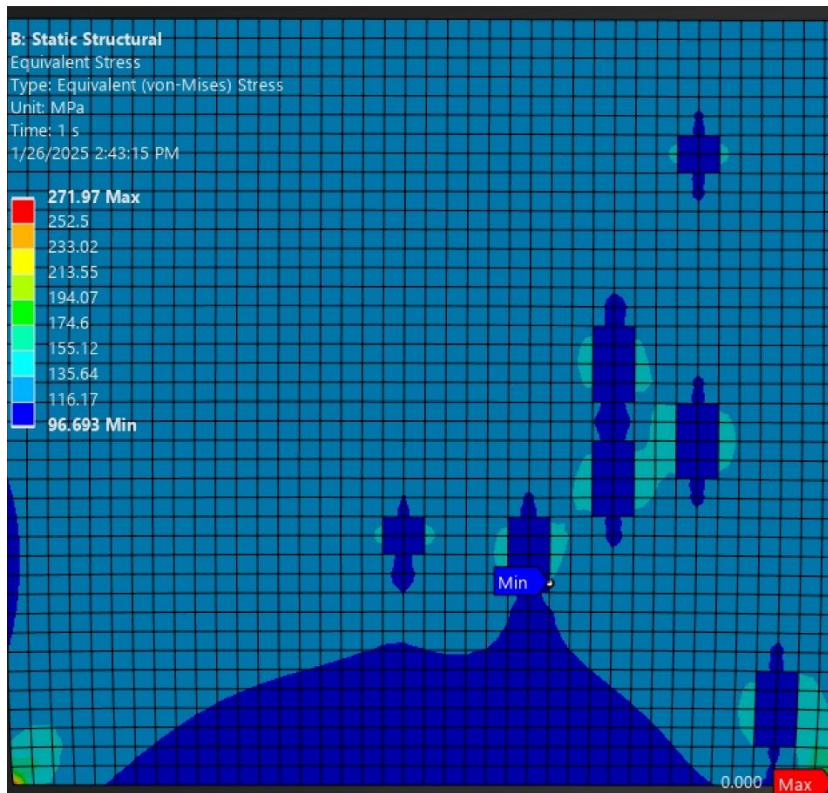


Figura 4. Resultados Simulación

3. Relación de los datos generados

Cada archivo generado por ANSYS tiene dos salidas básicas, “.png” como imagen y “.txt” como archivo de texto. Pues bien, el primero, que es una imagen binaria, es simplemente una visualización de una geometría de la pieza; el segundo es el archivo de texto que contiene los resultados de los datos de esfuerzo que obtuvimos previamente a través del cálculo de FEA. Para ligar estos archivos, los títulos de los archivos png y txt son iguales entre sí. Por lo tanto, “1.png” y “1.txt” indica que pertenecen a la misma simulación.


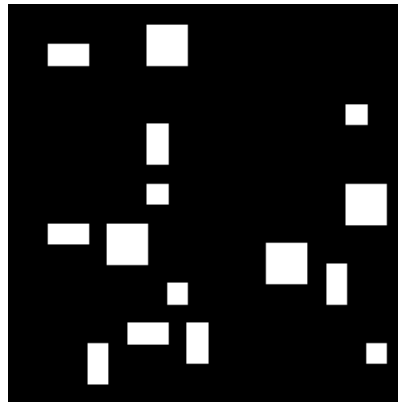
Nombre	↑
 1.png	
 1.txt	
 2.png	
 2.txt	

Figura 5. Base de datos .txt + .png

¿Cuáles son las estadísticas resumidas del conjunto de datos?

En los resultados obtenidos del esfuerzo, las estadísticas resumidas, en la *Tabla 1*, nos dan una rápida idea de las distribuciones de los valores de estrés y el tiempo correspondiente referidos a cada modelo:

- **Modelo geométrico:** Los modelos geométricos son piezas de acero estructural con inclusión colocada de manera aleatoria. El número de inclusión también es aleatorio, entre 5 y 20. La información de estos modelos se almacena como imagen binaria como se muestra en la imagen:



- **Esfuerzo Mínimo (Min Stress [MPa]):** Los valores mínimos de esfuerzo varían entre 0.59 MPa y 99.03 MPa. La desviación estándar de 20.52 sugiere una variabilidad considerable en los valores mínimos de esfuerzo entre los modelos. Esto podría indicar diferencias en las propiedades del material o en las condiciones iniciales de las pruebas realizadas.
- **Esfuerzo Máximo (Max Stress [MPa]):** El valor máximo de esfuerzo varía entre 264.53 MPa y 285.65 MPa. La desviación estándar de 4.42 indica una variabilidad relativamente pequeña en los valores máximos de esfuerzo entre los modelos, sugiriendo que el comportamiento del material fue bastante consistente bajo condiciones de carga máxima.
- **Esfuerzo Promedio (Avg Stress [MPa]):** Los valores promedio de esfuerzo tienen un rango entre 121.20 MPa y 123.03 MPa, con una desviación estándar de 0.38. Esto muestra que los esfuerzos promedio son altamente consistentes entre los modelos. El valor promedio de 122.72 MPa refuerza la idea de una distribución homogénea de los esfuerzos promedio entre los modelos analizados.

	Model	Min Stress [MPa]	Max Stress [MPa]	Avg Stress [MPa]
count	20.00000	20.000000	20.000000	20.000000
mean	10.50000	85.854642	272.929500	122.716000
std	5.91608	20.517763	4.422024	0.383494
min	1.00000	0.593850	264.530000	121.200000
25%	5.75000	86.322750	270.412500	122.685000
50%	10.50000	89.353000	273.230000	122.760000
75%	15.25000	93.382000	275.177500	122.895000
max	20.00000	99.029000	285.650000	123.030000

Tabla 1. Estadística descriptiva para base de datos

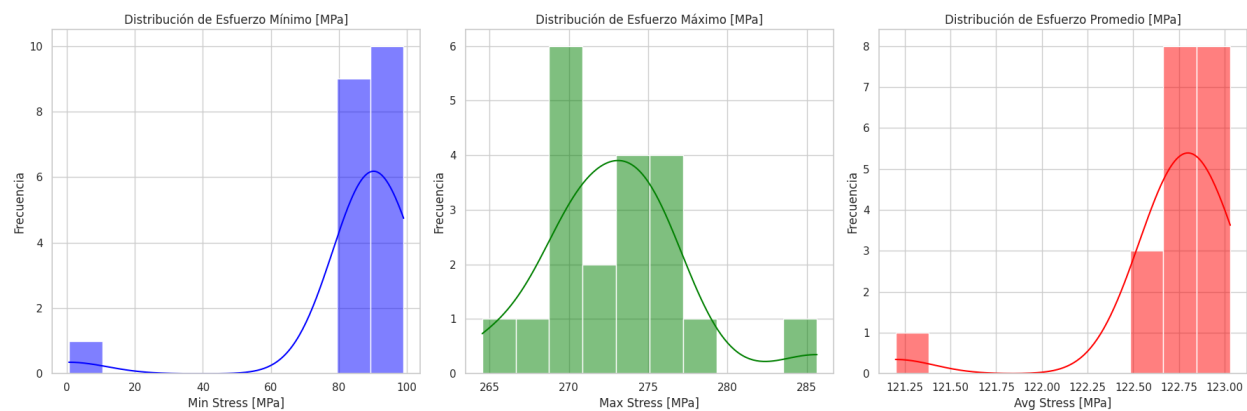


Figura 6. Histogramas de Esfuerzos/Estrés [MPa]

¿Hay valores atípicos en el conjunto de datos?

En el análisis de los valores atípicos del conjunto de datos, *Tabla 2*, se observó lo siguiente:

- **Valores atípicos de Min Stress [MPa]:** Se identificó un valor atípico en el modelo 19 con un valor de 0.59385, que es significativamente más bajo que el resto de los datos. Este resultado podría deberse a un error en la medición o a un comportamiento inusual del modelo.
- **Valores atípicos de Max Stress [MPa]:** En el modelo 12, se encontró un valor atípico de 285.65 MPa, que es considerablemente más alto que los demás valores. Esto sugiere que este modelo podría presentar un comportamiento excepcional o una medición fuera de lo común.
- **Valores atípicos de Avg Stress [MPa]:** El modelo 19 también presentó un valor atípico en esta categoría, con un promedio de esfuerzo de 121.2 MPa, siendo significativamente más bajo que el resto. Este comportamiento refuerza la posibilidad de una condición atípica o un error asociado al modelo.

```

⇒ Valores atípicos de Min Stress [MPa]:
      Model  Min Stress [MPa]  Max Stress [MPa]  Avg Stress [MPa]  \
18      19          0.59385        264.53          121.2

                                     Image Path
18  /content/drive/MyDrive/MNA/MNA - Colab/Proyect...

Valores atípicos de Max Stress [MPa]:
      Model  Min Stress [MPa]  Max Stress [MPa]  Avg Stress [MPa]  \
11      12          88.658        285.65          122.82

                                     Image Path
11  /content/drive/MyDrive/MNA/MNA - Colab/Proyect...

Valores atípicos de Avg Stress [MPa]:
      Model  Min Stress [MPa]  Max Stress [MPa]  Avg Stress [MPa]  \
18      19          0.59385        264.53          121.2

                                     Image Path
18  /content/drive/MyDrive/MNA/MNA - Colab/Proyect...

```

Tabla 2. Estadística descriptiva para base de datos

¿Cuál es la cardinalidad de las variables categóricas?

En nuestro caso, el análisis de cardinalidad no es relevante para las variables que realmente nos interesan, que son los valores de estrés, ya que estas columnas ("*Min Stress [MPa]*", "*Max Stress [MPa]*" y "*Avg Stress [MPa]*") contienen datos numéricos continuos, no categóricos. La cardinalidad es más útil para variables categóricas y la única variable categórica en el conjunto de datos es "*Model*", que simplemente representa el número de cada modelo. Esta variable no aporta información adicional significativa sobre los datos de estrés que estamos analizando, ya que su función es más identificativa que informativa.

¿Existen distribuciones sesgadas en el conjunto de datos? ¿Necesitamos aplicar alguna transformación no lineal?

Como se puede observar en la *Figura 6. Histogramas de Esfuerzos/Estrés [MPa]*, aparentemente los valores se encuentran un poco sesgados, para corroborarlo, calculamos la asimetría (skewness):

```

⇒ Asimetría de Min Stress [MPa]: -4.148709735816935
   Asimetría de Max Stress [MPa]: 0.8617845259617456
   Asimetría de Avg Stress [MPa]: -3.514051084550601

```

Figura 7. Resultados de skewness

- **Min Stress [MPa]: -4.15:** Presenta una asimetría negativa significativa, lo que indica que los datos están sesgados hacia valores bajos. Podría beneficiarse de una transformación logarítmica o raíz cuadrada para reducir este sesgo y normalizar la distribución.
- **Max Stress [MPa]: 0.86:** Tiene una asimetría moderada positiva, lo que sugiere un leve sesgo hacia valores altos. Una transformación podría ser útil, aunque no es estrictamente necesaria debido a la moderación del sesgo.
- **Avg Stress [MPa]: -3.51:** Muestra una asimetría negativa considerable, lo que indica un sesgo hacia valores bajos. Al igual que con Min Stress [MPa], una transformación logarítmica o raíz cuadrada podría ayudar a reducir esta asimetría.

¿Se identifican tendencias temporales? (En caso de que el conjunto incluya una dimensión de tiempo).

No hay dimensión de tiempo relevante para el conjunto de datos ni el propósito del proyecto.

¿Hay correlación entre las variables dependientes e independientes?

Ese es el objetivo principal de la investigación, pero podemos previsualizar con los datos y avance actual mediante una matriz de correlación. Se eliminó la columna de “*Tiempo [s]*” ya que tiene valores constantes por lo que no aporta variabilidad alguna, lo que hace que no tenga sentido incluirla en un análisis de correlación, ya que la correlación depende de la variabilidad entre los valores:

⇒ Matriz de correlación:

	Min Stress [MPa]	Max Stress [MPa]	Avg Stress [MPa]
Min Stress [MPa]	1.000000	0.348992	0.975269
Max Stress [MPa]	0.348992	1.000000	0.271530
Avg Stress [MPa]	0.975269	0.271530	1.000000

Figura 8. Resultados de matriz de correlación

- **Correlación entre Min Stress [MPa] y Max Stress [MPa]: 0.349.** Esto indica una correlación débil positiva, lo que significa que un incremento en el esfuerzo mínimo está débilmente relacionado con un incremento en el esfuerzo máximo.
- **Correlación entre Min Stress [MPa] y Avg Stress [MPa]: 0.975.** Existe una correlación muy fuerte positiva entre estas dos variables, lo que sugiere que cuando el esfuerzo mínimo aumenta, el esfuerzo promedio también tiende a incrementarse significativamente.

- **Correlación entre Max Stress [MPa] y Avg Stress [MPa]:** 0.272. Esta es una correlación débil positiva, lo que implica que un incremento en el esfuerzo máximo está ligeramente relacionado con un aumento en el esfuerzo promedio, pero esta relación no es muy significativa.

En conclusión, se observa que la variable **Min Stress [MPa]** tiene una relación mucho más fuerte con el esfuerzo promedio (**Avg Stress [MPa]**) que con el esfuerzo máximo (**Max Stress [MPa]**), lo cual podría ser relevante al evaluar el comportamiento del material o los modelos bajo análisis.

¿Cómo se distribuyen los datos en función de diferentes categorías? (análisis bivariado)

La variabilidad en los valores de estrés (Min Stress [MPa], Max Stress [MPa], Avg Stress [MPa]) entre las diferentes simulaciones es esperada, ya que cada modelo representa una geometría distinta de la estructura, lo que influye directamente en cómo se distribuye y se comporta el estrés bajo las mismas condiciones de carga. Las diferencias observadas en la distribución de los datos no son sorprendentes, ya que los modelos están diseñados con diferentes geometrías y propiedades, lo que provoca distintas respuestas mecánicas.

¿Se deberían normalizar las imágenes para visualizarlas mejor?

No es necesario, ya que al ser datos sintéticos, se tiene cuidado de generar imágenes binarias dentro del rango 0,1 desde el principio.

¿Hay desequilibrio en las clases de la variable objetivo?

En este caso al no tratarse de una tarea de clasificaciones no aplica el desequilibrio de clases.