FACULTAD DE INGENIERÍA, DISEÑO Y CIENCIAS APLICADAS

ESCUELA DE TECNOLOGÍA, DISEÑO E INNOVACIÓN.

DEPARTAMENTO DE TECNOLOGÍAS DE INFORMACIÓN Y

COMUNICACIONES (TIC)

Actividad: Análisis Univariado de una Variable en un Conjunto de Datos

Análisis de datos I

Maestría en Inteligencia artificial aplicada

Integrantes:

- Juan Bautista
- Sandra Orozco
- Felipe Guerra Sáenz
- Mavelyn Sterling Londoño

Introducción y selección de variable

En la producción industrial de penicilina a través de fermentación en biorreactores, múltiples variables afectan la eficiencia del proceso y la calidad del producto final. Entre estas variables, la tasa de alimentación de ácido fenilacético (PAA), representada como PAA Flow (L/h), juega un papel crucial en la regulación de la biosíntesis de penicilina.

La selección de PAA Flow como variable de interés se basa en su función como precursor clave en la síntesis de penicilina. En el proceso de fermentación con Penicillium chrysogenum, el PAA actúa como un componente esencial que se incorpora en la estructura química de la penicilina. Sin embargo, su concentración debe mantenerse en un rango óptimo, ya que una dosificación inadecuada puede resultar en toxicidad celular o en una producción subóptima del antibiótico [1].

Además, el PAA Flow es una variable manipulable, lo que significa que puede ser ajustada en tiempo real a través de sistemas de control automatizado, a diferencia de otras variables de estado como la biomasa o la concentración de penicilina, que dependen de la evolución del proceso. Su impacto en la producción de penicilina ha sido ampliamente documentado en estudios industriales, lo que justifica su selección como una de las variables críticas para el análisis y optimización del proceso [1].

Explicación de su importancia

El ácido fenilacético (PAA) es el precursor metabólico esencial en la ruta biosintética de la penicilina. Su disponibilidad en el medio de fermentación determina la tasa de producción del antibiótico, pero su regulación es un desafío debido a los siguientes factores:

Concentraciones insuficientes de PAA (<600 mg/L) limitan la producción de penicilina porque el hongo no tiene suficiente sustrato para la biosíntesis [1].

Concentraciones excesivas de PAA (>2000 mg/L) pueden ser tóxicas para el hongo, inhibiendo su crecimiento y reduciendo la producción de penicilina [1].

La alimentación de PAA debe ser controlada dinámicamente a lo largo del tiempo, ya que los requerimientos del microorganismo varían en función de la fase de fermentación [1].

La tasa de flujo de PAA (PAA Flow) se convierte entonces en un parámetro crítico de control, ya que permite regular la concentración de este precursor en el medio y, por ende, optimizar la producción de penicilina.

La variabilidad en la alimentación de PAA puede causar inconsistencias entre los lotes de producción, afectando la estabilidad del proceso. Sin embargo, con el uso de modelos de machine learning y técnicas de optimización de parámetros, es posible diseñar un esquema de alimentación más inteligente y adaptativo, capaz de minimizar estas fluctuaciones y mejorar la confiabilidad del sistema [1].

Además, el PAA es uno de los insumos más costosos en la producción de penicilina, representando cerca del 11% del costo total de materias primas. Un manejo ineficiente de su flujo no solo puede elevar los costos operacionales, sino también generar desperdicio de recursos. En cambio, un ajuste preciso en su dosificación permite optimizar el uso de materia prima y mejorar la rentabilidad del proceso [1].

Desde el punto de vista ambiental, regular adecuadamente PAA Flow no solo impacta en la eficiencia económica, sino que también ayuda a reducir la generación de subproductos no deseados, contribuyendo a un proceso de fermentación más sostenible y eficiente.

Código en Python con análisis univariado y visualizaciones

Link del repositorio: https://github.com/MavelSterling/reto_insulina_INRAE

```python
def validar_nulos(serie):
    nulos = serie.isnull().sum()
    porcentaje_nulos = (nulos / len(serie)) * 100


    resultado = pd.DataFrame(
        {'Valores Nulos': [nulos], 'Porcentaje (%)': [porcentaje_nulos]},
        index=[serie.name]
    )
    return resultado
```

✓ 0.0s

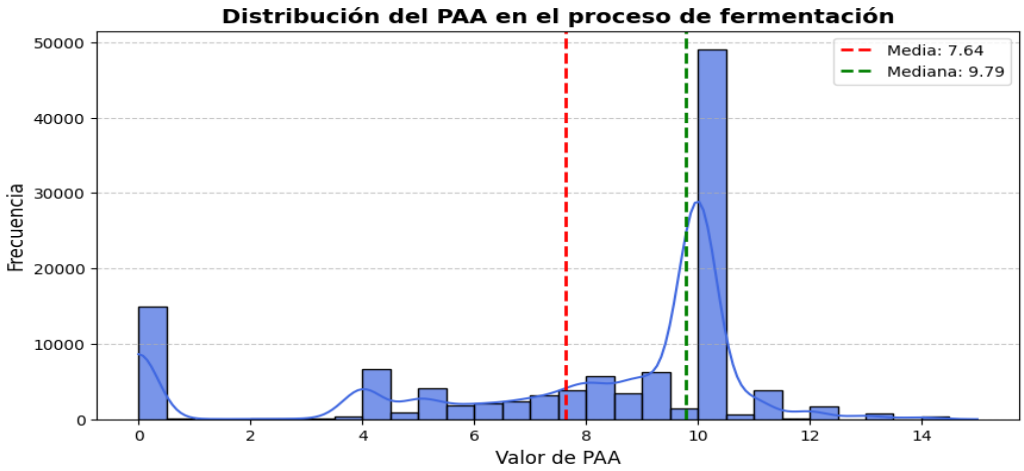
```python
resultado_nulos = validar_nulos(dataset["PAA flow(Fpaa:PAA flow (L/h))"])
print(resultado_nulos)
```

✓ 0.0s

```
                            Valores Nulos  Porcentaje (%)
PAA flow(Fpaa:PAA flow (L/h))             0             0.0
```

Analisis univariado de Fluido de PAA

```python
plt.figure(figsize=(10, 5))

# Histograma con mayor personalización
sns.histplot(dataset["PAA flow(Fpaa:PAA flow (L/h))"], bins=30, kde=True, color="royalblue", edgecolor="black", alpha=0.7)

# Agregar líneas de media y mediana
mean_value = dataset["PAA flow(Fpaa:PAA flow (L/h))"].mean()
median_value = dataset["PAA flow(Fpaa:PAA flow (L/h))"].median()

plt.axvline(mean_value, color='red', linestyle='dashed', linewidth=2, label=f'Media: {mean_value:.2f}')
plt.axvline(median_value, color='green', linestyle='dashed', linewidth=2, label=f'Mediana: {median_value:.2f}')

# Mejorar etiquetas y título
plt.title("Distribución del PAA en el proceso de fermentación", fontsize=14, fontweight="bold")
plt.xlabel("Valor de PAA", fontsize=12)
plt.ylabel("Frecuencia", fontsize=12)
plt.legend()
plt.grid(axis='y', linestyle="--", alpha=0.7)

# Mostrar el gráfico
plt.show()
```

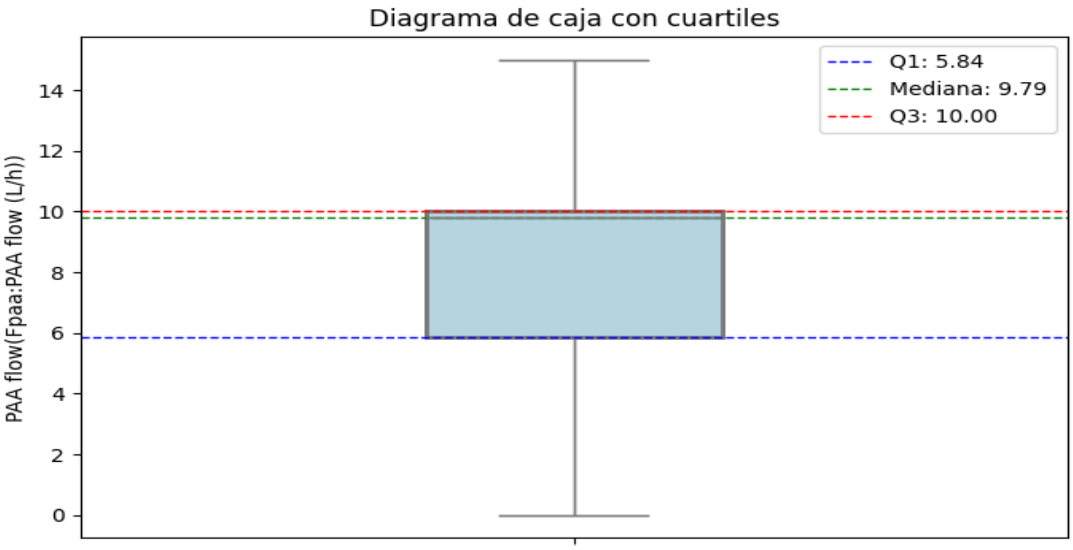



```
skewness = dataset["PAA flow(Fpaa:PAA flow (L/h))"].skew()
print("Coeficiente de asimetría:", skewness)
```

[21] ✓ 0.0s

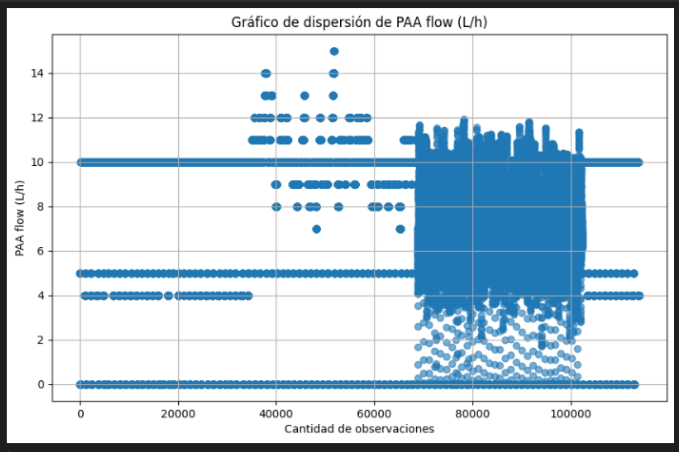
```
Coeficiente de asimetría: -1.1605246649265606
```

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8, 5))
sns.boxplot(y=dataset["PAA flow(Fpaa:PAA flow (L/h))"], color="lightblue", width=0.3, boxprops=dict(linewidth=2.0))

# Añadir líneas de cuartiles
# Calcular cuartiles de la variable
q1 = dataset["PAA flow(Fpaa:PAA flow (L/h))"].quantile(0.25)
q2 = dataset["PAA flow(Fpaa:PAA flow (L/h))"].median()
q3 = dataset["PAA flow(Fpaa:PAA flow (L/h))"].quantile(0.75)

sns.boxplot(y=dataset["PAA flow(Fpaa:PAA flow (L/h))"], color="lightblue", width=0.3, boxprops=dict(linewidth=2.0))

# Añadir líneas de cuartiles
plt.axhline(y=q1, color='blue', linestyle='dashed', linewidth=1, label=f'Q1: {q1:.2f}')
plt.axhline(y=q2, color='green', linestyle='dashed', linewidth=1, label=f'Mediana: {q2:.2f}')
plt.axhline(y=q3, color='red', linestyle='dashed', linewidth=1, label=f'Q3: {q3:.2f}')
plt.legend()
plt.title("Diagrama de caja con cuartiles")
plt.show()
```




```python
import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))
plt.scatter(dataset.index, dataset["PAA flow(Fpaa:PAA flow (L/h))"], alpha=0.6)
plt.xlabel("Cantidad de observaciones")
plt.ylabel("PAA flow (L/h)")
plt.title("Gráfico de dispersión de PAA flow (L/h)")
plt.grid(True)
plt.show()
```

✓ 0.4s



En el gráfico de dispersión (cada punto representa el valor de `PAA flow (L/h)` en el índice correspondiente) se observan **varios "bloques" o concentraciones de datos**:

1. Un grupo numeroso de puntos alrededor de **0 L/h**.
2. Un grupo más disperso en la zona de **3–4 L/h**.
3. Un bloque muy denso alrededor de **9–10 L/h** (donde parece haber la mayor cantidad de observaciones).
4. Algunos valores puntuales que superan **12 L/h**.

Interpretación de los resultados

Métrica	Valor	Interpretación
Media	7.64 L/h	En promedio, el flujo de PAA en el proceso de fermentación es de 7.64 L/h, lo que indica un nivel de alimentación moderado. Sin embargo, debido a la asimetría de la distribución, este valor puede no representar con precisión el comportamiento central de la variable.
Primer Cuartil (Q1, Percentil 25%)	5.84 L/h	El 25% de los datos se encuentran por debajo de este valor, lo que indica que existe una proporción significativa de observaciones en niveles bajos de PAA Flow.
Mediana (Q2, Percentil 50%)	9.79 L/h	La mediana es superior a la media, lo que indica que más del 50% de los valores se encuentran por encima de la media. Esto confirma la asimetría negativa en la distribución, causada por valores bajos que reducen el promedio.
Tercer Cuartil (Q3, Percentil 75%)	10.00 L/h	El 75% de los valores se encuentran por debajo de este umbral, lo que sugiere que la mayoría de los registros tienden a concentrarse en valores cercanos a 10 L/h.
Desviación Estándar	3.56 L/h	Existe una dispersión moderada en los datos, lo que indica que los valores de PAA Flow pueden variar significativamente entre diferentes registros, oscilando entre niveles bajos y altos.
Mínimo	0 L/h	Se identificaron casos en los que el flujo de PAA es nulo, lo que puede estar asociado con fases del proceso en las que la alimentación se detiene temporalmente o no es necesaria.
Máximo	15 L/h	Este es el límite superior observado en los datos. No se identificaron valores extremos que sugieran inconsistencias o errores en la medición.

El análisis de la media (7.64 L/h) y la mediana (9.79 L/h) confirma que la distribución presenta un sesgo hacia la izquierda, lo que significa que hay una mayor concentración de valores altos, pero con una proporción de observaciones en niveles bajos que reducen la media.

Este sesgo también se refleja en el coeficiente de asimetría (skewness = -1.16), lo que indica una distribución donde los valores bajos son más frecuentes de lo que se esperaría en una distribución normal.

El histograma muestra dos picos principales, lo que sugiere que la alimentación del PAA no sigue un único patrón, sino que hay dos posibles regímenes operativos:

- Un grupo de datos en valores bajos (cercanos a 0-5 L/h).
- Una alta concentración de registros cerca de 10 L/h, lo que indica que este podría ser el nivel más común de operación.

Este comportamiento puede deberse a un ajuste dinámico de la alimentación, donde en ciertos momentos la cantidad de PAA se reduce o detiene, mientras que en otros se mantiene en niveles cercanos a 10 L/h.

La desviación estándar de 3.56 L/h indica que los valores de PAA Flow varían de manera moderada, aunque la distribución no es completamente uniforme. El rango intercuartílico (IQR = Q3 - Q1 = 10.00 - 5.84 = 4.16 L/h) sugiere que la mayoría de los datos están dentro de un intervalo relativamente compacto, aunque los valores cercanos a 0 L/h muestran que existen algunas fluctuaciones más marcadas.

Existe un grupo con valores bajos (~0-5 L/h), posiblemente asociado a períodos en los que la alimentación de PAA se detiene o es mínima. Otro grupo con valores cercanos a 10 L/h, lo que sugiere que este podría ser el nivel de flujo más utilizado durante la fermentación.

El rango intercuartílico muestra que la mayor parte de los valores están comprendidos entre 5.84 y 10.00 L/h. La dispersión dentro de este rango es moderada, aunque hay valores mínimos de 0 L/h que podrían representar fases específicas del proceso donde no se requiere alimentación de PAA.

Si el flujo de PAA es demasiado bajo en ciertos momentos, podría afectar la síntesis de penicilina al limitar la disponibilidad del precursor. Si el flujo es muy alto de manera sostenida, se debe evaluar si esto tiene efectos negativos, como toxicidad para el hongo productor.

El análisis univariado de PAA Flow (L/h) indica que la variable presenta una distribución sesgada a la izquierda, con una mediana superior a la media y una posible distribución bimodal. La mayor parte de los valores se encuentran cercanos a 10 L/h, pero hay una proporción significativa de registros en valores bajos, lo que sugiere una estrategia de alimentación variable en función de la etapa del proceso.

Conclusiones

El análisis univariado de PAA Flow (L/h) permitió comprender su comportamiento dentro del proceso de fermentación de penicilina y su impacto en el bioproceso. Se identificó que la alimentación de PAA no sigue un patrón completamente uniforme, sino que presenta variabilidad estructurada, con momentos de flujo alto y periodos en los que la dosificación se reduce o se detiene.

Uno de los hallazgos más relevantes es que la distribución de PAA Flow está sesgada hacia la izquierda, con una mediana mayor que la media, lo que indica que hay registros con flujo reducido que afectan el promedio. Además, la posible bimodalidad en la distribución sugiere que el proceso opera bajo diferentes estrategias de nivel de alimentación.

Desde una perspectiva operativa, la variabilidad observada en la serie de datos puede influir en la estabilidad del proceso de fermentación. Un flujo de PAA inconstante o inadecuado podría afectar la eficiencia en la producción de penicilina, ya sea por limitación del precursor o por efectos negativos relacionados con su acumulación excesiva.

Bibliografía

[1] Goldrick, S., Duran-Villalobos, C. A., Jankauskas, K., Lovett, D., Farid, S. S., & Lennox, B. (2019). Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process. Computers & Chemical Engineering, 130, 106471. https://doi.org/10.1016/j.compchemeng.2019.05.037