FACULTAD DE INGENIERÍA, DISEÑO Y CIENCIAS APLICADAS

ESCUELA DE TECNOLOGÍA, DISEÑO E INNOVACIÓN.

DEPARTAMENTO DE TECNOLOGÍAS DE INFORMACIÓN Y

COMUNICACIONES (TIC)

Actividad: Análisis Bivariado

Análisis de datos I

Maestría en Inteligencia artificial aplicada

Integrantes:

- Felipe Guerra Sáenz
- Mavelyn Sterling Londoño

Taller: Análisis Bivariado

Justificación de la selección de variables

La elección de la variable objetivo "Penicillin concentration (P: g/L)" como la variable dependiente en el análisis del proceso de fermentación de penicilina se fundamenta en su importancia dentro de la producción industrial. La concentración de penicilina es el principal indicador de rendimiento del bioproceso, ya que refleja la eficiencia metabólica de Penicillium chrysogenum en la conversión de sustratos en el antibiótico deseado. Esta variable es clave para la optimización del proceso, ya que cualquier fluctuación en su concentración puede impactar directamente la calidad y cantidad del producto final. Además, su control adecuado es esencial para cumplir con los estándares regulatorios y garantizar la rentabilidad del proceso de producción a escala industrial. [1]

Por otro lado, la variable predictora seleccionada, "PAA flow (Fpaa: PAA flow (L/h))", representa el flujo de ácido fenilacético (PAA), el cual es un precursor esencial en la biosíntesis de penicilina. La adición controlada de PAA en el biorreactor influye

directamente en la tasa de producción del antibiótico, ya que proporciona la cadena lateral necesaria para la síntesis de penicilina G. Una concentración inadecuada de PAA puede resultar en una disminución de la productividad o en efectos tóxicos que afecten el crecimiento del microorganismo. Por ello, su monitoreo y ajuste en tiempo real es crucial para maximizar el rendimiento del proceso. La relación entre estas dos variables permite desarrollar estrategias de control avanzadas que optimicen la eficiencia del bioproceso y reduzcan variabilidad en la producción de penicilina. [1]

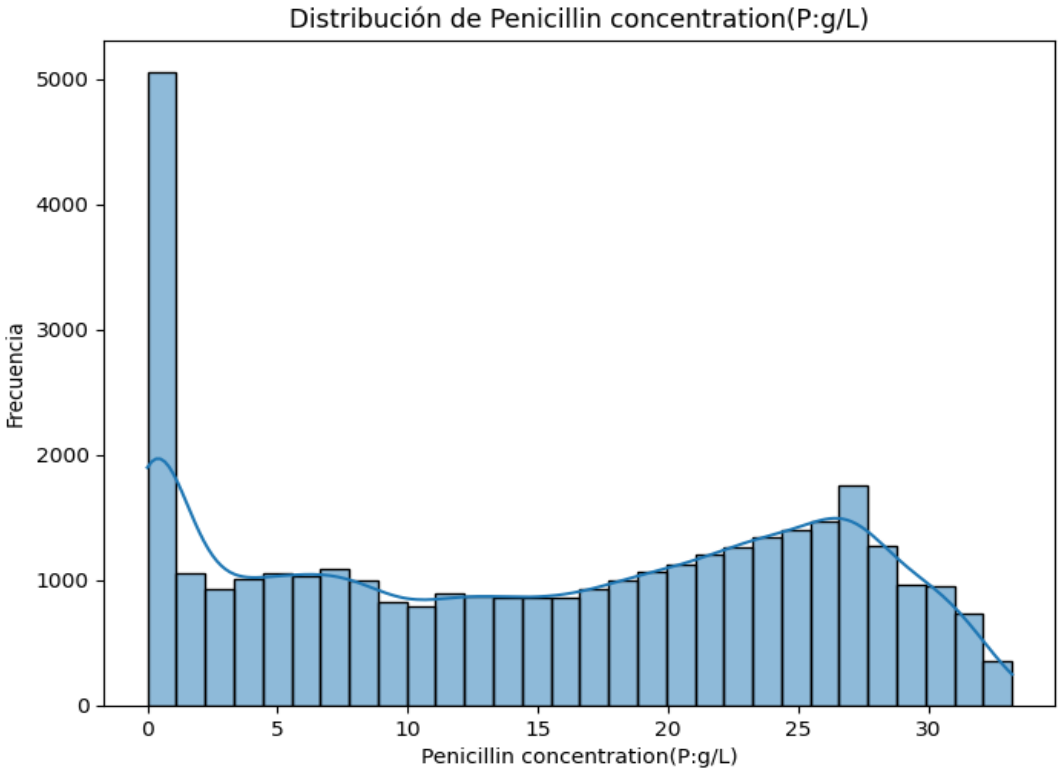
Código en Python con análisis univariado y visualizaciones

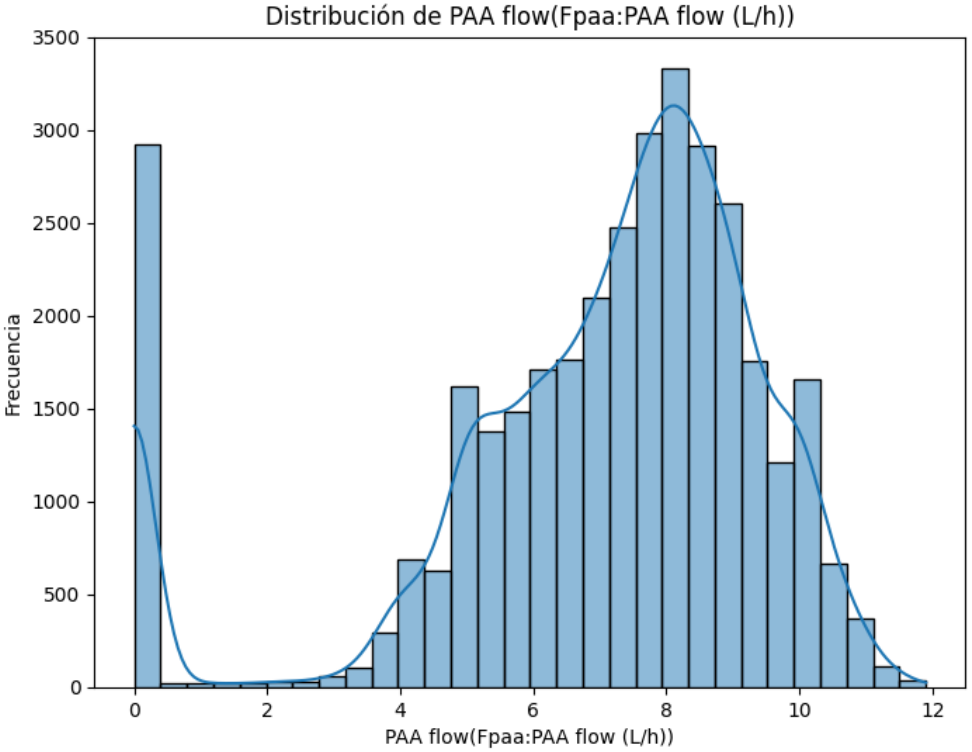
Link del repositorio:
https://github.com/MavelSterling/proyecto_reto_insulina_INRAE/tree/feature/analisis_bivariado

```python
# ================================================================================
# 1. Selección de variables: Variable objetivo y predictor
# Se seleccionan:
# - Penicillin concentration(P:g/L): Indicador final de rendimiento.
# - PAA flow(Fpaa:PAA flow (L/h)): Variable de control crítica para la dosificación del precursor.
# ================================================================================
target_variable = 'Penicillin concentration(P:g/L)'
predictor = 'PAA flow(Fpaa:PAA flow (L/h))'

# Extraer el subconjunto con las dos variables y eliminar valores faltantes
df_biv = batch_61_90_df[[target_variable, predictor]].dropna()

# ================================================================================
# 2. Estadísticas descriptivas y cálculo de correlaciones
# ================================================================================
desc_target = df_biv[target_variable].describe()
desc_predictor = df_biv[predictor].describe()

print("Estadísticas descriptivas de Penicillin concentration:")
print(desc_target)
print("\nEstadísticas descriptivas de PAA flow:")
print(desc_predictor)

# Correlación de Pearson (lineal) y Spearman (monótona)
corr_pearson = df_biv[target_variable].corr(df_biv[predictor], method='pearson')
corr_spearman = df_biv[target_variable].corr(df_biv[predictor], method='spearman')

print(f"\nCorrelación de Pearson entre {target_variable} y {predictor}: {corr_pearson:.3f}")
print(f"Correlación de Spearman entre {target_variable} y {predictor}: {corr_spearman:.3f}")
```

```
Estadísticas descriptivas de Penicillin concentration:
count    3.499000e+04
mean     1.500956e+01
std      1.036009e+01
min      3.694700e-26
25%      5.160875e+00
50%      1.573250e+01
75%      2.446950e+01
max      3.320900e+01
Name: Penicillin concentration(P:g/L), dtype: float64

Estadísticas descriptivas de PAA flow:
count    34990.000000
mean         6.915300
std          2.666898
min          0.000000
25%          5.816325
50%          7.555400
75%          8.677600
max         11.901000
Name: PAA flow(Fpaa:PAA flow (L/h)), dtype: float64

Correlación de Pearson entre Penicillin concentration(P:g/L) y PAA flow(Fpaa:PAA flow (L/h)): 0.406
Correlación de Spearman entre Penicillin concentration(P:g/L) y PAA flow(Fpaa:PAA flow (L/h)): 0.278
```

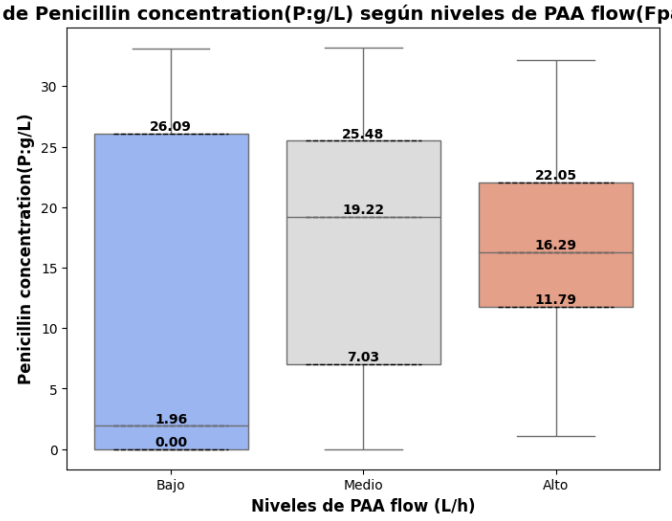
```python
# ================================================================================
# 6. Histogramas individuales para cada variable
# ================================================================================
plt.figure(figsize=(8,6))
sns.histplot(df_biv[target_variable], bins=30, kde=True)
plt.title(f"Distribución de {target_variable}")
plt.xlabel(target_variable)
plt.ylabel("Frecuencia")
plt.show()

plt.figure(figsize=(8,6))
sns.histplot(df_biv[predictor], bins=30, kde=True)
plt.title(f"Distribución de {predictor}")
plt.xlabel(predictor)
plt.ylabel("Frecuencia")
plt.show()
```


Distribución de Penicillin concentration(P:g/L)
Frecuencia
Penicillin concentration(P:g/L)
0
1000
2000
3000
4000
5000
0
5
10
15
20
25
30

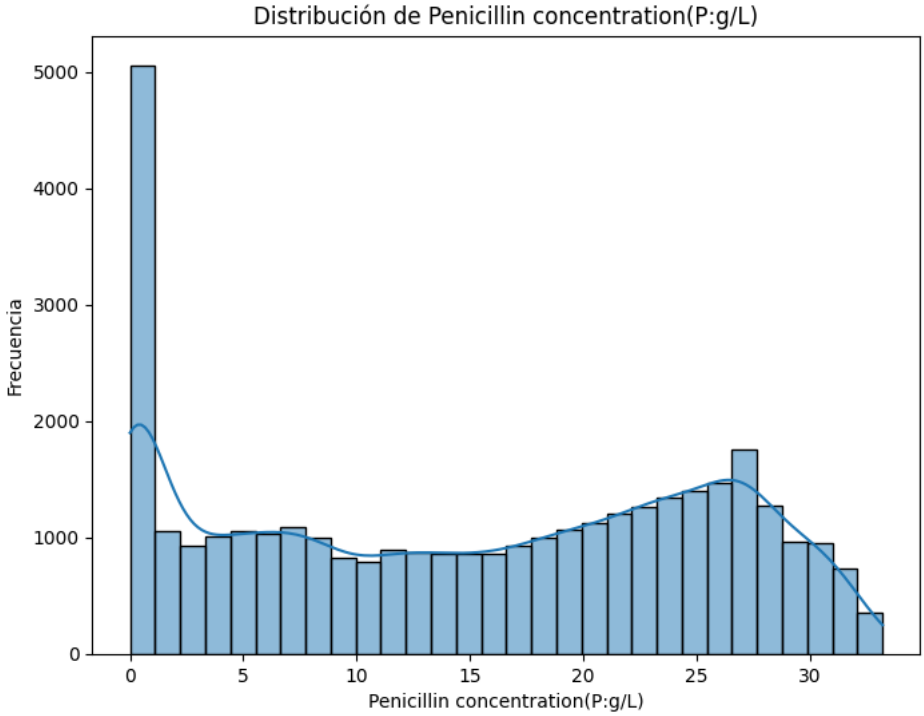

Distribución de PAA flow(Fpaa:PAA flow (L/h))
Frecuencia
PAA flow(Fpaa:PAA flow (L/h))
0
500
1000
1500
2000
2500
3000
3500
0
2
4
6
8
10
12

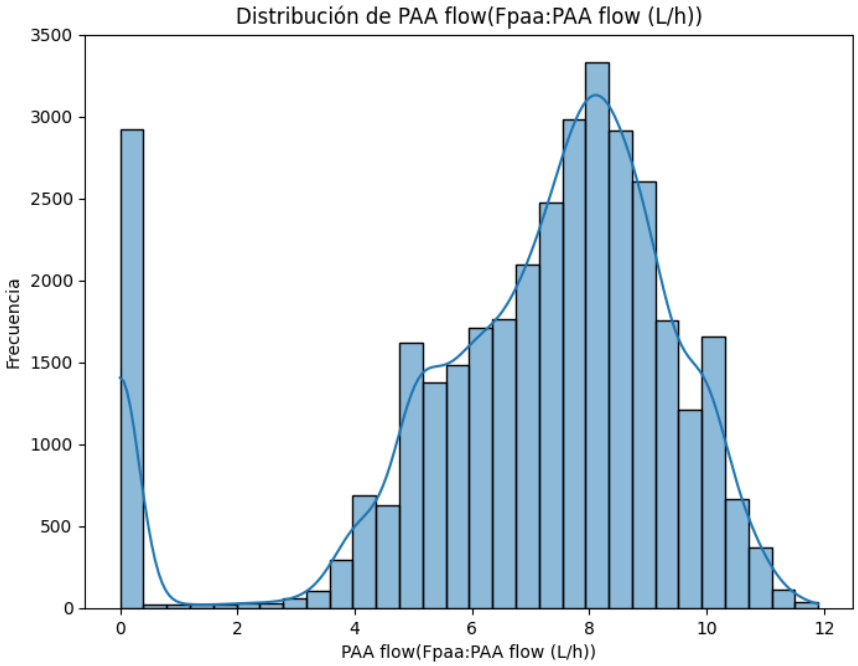
```python
# ========================================================================
# 5. Boxplot comparativo: Analizar el impacto del flujo de PAA en la producción
# ------------------------------------------------------------------------
# Se crea una variable categórica dividiendo PAA flow en terciles
df_biv['PAA_flow_bin'] = pd.qcut(df_biv[predictor], q=3, labels=["Bajo", "Medio", "Alto"])

plt.figure(figsize=(8,6))
sns.boxplot(x='PAA_flow_bin', y=target_variable, data=df_biv)
plt.title(f"Distribución de {target_variable} según niveles de {predictor}")
plt.xlabel("Niveles de PAA flow (L/h)")
plt.ylabel(target_variable)
plt.show()

# Crear la variable categórica dividiendo PAA flow en terciles
df_biv['PAA_flow_bin'] = pd.qcut(df_biv[predictor], q=3, labels=["Bajo", "Medio", "Alto"])

# Crear el boxplot
plt.figure(figsize=(8,6))
ax = sns.boxplot(x='PAA_flow_bin', y=target_variable, data=df_biv, palette="coolwarm")

# Calcular los cuartiles y agregarlos al gráfico
for i, level in enumerate(["Bajo", "Medio", "Alto"]):
    quartiles = df_biv[df_biv['PAA_flow_bin'] == level][target_variable].quantile([0.25, 0.5, 0.75])

    # Marcar los cuartiles en la gráfica
    for q in quartiles:
        plt.hlines(y=q, xmin=i-0.3, xmax=i+0.3, colors='black', linestyles='dashed', linewidth=1)
        plt.text(i, q, f'{q:.2f}', ha='center', va='bottom', fontsize=10, fontweight='bold', color='black')

# Etiquetas y título
plt.title(f"Distribución de {target_variable} según niveles de {predictor}", fontsize=14, weight='bold')
plt.xlabel("Niveles de PAA flow (L/h)", fontsize=12, weight='bold')
plt.ylabel(target_variable, fontsize=12, weight='bold')

# Mostrar la gráfica
plt.show()
```

Distribución de Penicillin concentration(P:g/L) según niveles de PAA flow(Fpaa:PAA flow (L/h))

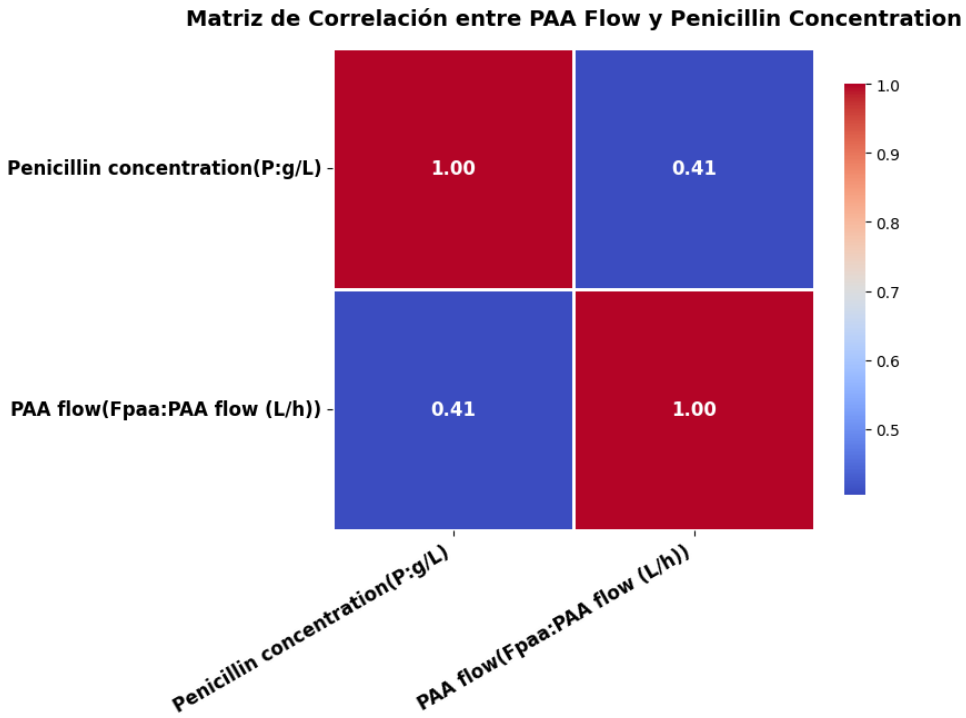


```python
# ============================================================================
# 6. Histogramas individuales para cada variable
# ============================================================================
plt.figure(fi  (variable) df_biv: DataFrame
sns.histplot(df_biv[target_variable], bins=30, kde=True)
plt.title(f"Distribución de {target_variable}")
plt.xlabel(target_variable)
plt.ylabel("Frecuencia")
plt.show()

plt.figure(figsize=(8,6))
sns.histplot(df_biv[predictor], bins=30, kde=True)
plt.title(f"Distribución de {predictor}")
plt.xlabel(predictor)
plt.ylabel("Frecuencia")
plt.show()
```






```python
# ============================================================================
# 7. Matriz de correlación y heatmap
# ============================================================================
# Filtrar solo columnas numéricas
df_numeric = df_biv.select_dtypes(include=['number'])

# Crear matriz de correlación
correlation_matrix = df_numeric.corr(method='pearson')

# Ajustar la figura y el heatmap
plt.figure(figsize=(7, 6))
sns.heatmap(
    correlation_matrix,
    annot=True,
    cmap="coolwarm",
    fmt=".2f",
    linewidths=1,
    square=True,
    cbar_kws={"shrink": 0.8},
    annot_kws={"size": 12, "weight": "bold"}
)

# Ajustar las etiquetas de los ejes
plt.xticks(rotation=30, ha="right", fontsize=12, weight="bold")
plt.yticks(rotation=0, fontsize=12, weight="bold")

# Título más claro
plt.title("Matriz de Correlación entre PAA Flow y Penicillin Concentration",
          fontsize=14, weight="bold", pad=15)

# Mostrar el gráfico
plt.show()
```



Interpretación de los resultados

Estadísticas de Penicillin concentración		
Métrica	**Valor**	**Interpretación**
Media	15.00956	En promedio, la concentración de penicilina en el proceso de fermentación es de 15.01 g/L, indicando un nivel de producción moderado.
Desviación Estándar	10.36009	Existe una dispersión relativamente alta en la concentración de penicilina, lo que sugiere variabilidad en el proceso de fermentación.
Mínimo	3.6947e-26	El valor mínimo es cercano a cero, indicando la presencia de muestras sin producción de penicilina.

Primer Cuartil (Q1, Percentil 25%)	5.160875	El 25% de los valores se encuentran por debajo de este umbral, lo que sugiere que en ciertos momentos la producción de penicilina es baja.
Mediana (Q2, Percentil 50%)	15.7325	La mediana es 15.73 g/L, similar a la media, lo que indica que la distribución es aproximadamente simétrica.
Tercer Cuartil (Q3, Percentil 75%)	24.4695	El 75% de los valores están por debajo de 24.47 g/L, lo que sugiere que la mayoría de las muestras se concentran en valores menores a este.
Máximo	33.209	El valor máximo observado es 33.21 g/L, representando el nivel más alto registrado en la producción.

Estadísticas de PAA Flow		
Métrica	**Valor**	**Interpretación**
Media	6.9153	En promedio, el flujo de PAA en el proceso de fermentación es de 6.92 L/h, indicando un nivel moderado de alimentación.
Desviación Estándar	2.666898	Existe una variabilidad moderada en los valores de PAA flow, lo que sugiere fluctuaciones en la alimentación del proceso.
Mínimo	0.0	Se identificaron casos con flujo de PAA nulo, lo que indica momentos en los que la alimentación se detiene o no es necesaria.
Primer Cuartil (Q1, Percentil 25%)	5.816325	El 25% de los valores son menores a 5.82 L/h, lo que sugiere que existe una proporción significativa de observaciones en niveles bajos de alimentación.
Mediana (Q2, Percentil 50%)	7.5554	La mediana es 7.56 L/h, lo que indica que la mayoría de los valores están por encima de la media.

Tercer Cuartil (Q3, Percentil 75%)	8.6776	El 75% de los valores están por debajo de 8.68 L/h, lo que sugiere que la mayoría de los registros se concentran en valores cercanos a este umbral.
Máximo	11.901	El valor máximo observado es 11.90 L/h, indicando el nivel más alto de alimentación registrado.

Análisis de la Concentración de Penicilina (Penicillin concentration)

1. **Media (15.01 g/L) y Mediana (15.73 g/L):**
    a. La media y la mediana son relativamente cercanas, lo que sugiere que la distribución de la concentración de penicilina es aproximadamente simétrica.
    b. Sin embargo, la ligera diferencia indica que existen valores bajos que reducen el promedio, aunque la mayoría de los valores se concentran en niveles intermedios.
2. **Desviación Estándar (10.36 g/L):**
    a. Existe una dispersión significativa en los datos, lo que indica que la producción de penicilina varía de manera notable entre diferentes mediciones.
    b. Esta variabilidad podría deberse a ajustes en el proceso de fermentación, fluctuaciones en la alimentación de precursores o condiciones operativas cambiantes.
3. **Mínimo (≈0 g/L) y Máximo (33.21 g/L):**
    a. Se identificaron valores muy bajos, lo que puede indicar momentos donde no se produjo penicilina o la medición captó una fase inicial de fermentación.
    b. El valor máximo sugiere que el sistema puede alcanzar niveles de producción elevados en condiciones óptimas.
4. **Distribución de los Cuartiles:**
    a. **Q1 (5.16 g/L):** El 25% de los datos tienen una concentración menor a 5.16 g/L, lo que sugiere periodos de baja producción.
    b. **Q3 (24.47 g/L):** El 75% de los datos están por debajo de 24.47 g/L, indicando que solo una fracción de las muestras logra valores más altos de producción.
    c. **IQR (Interquartile Range = Q3 - Q1 = 19.31 g/L):** Un rango Inter-cuartil amplio indica que hay una dispersión significativa en los datos.
5. **Interpretación:**

a. La alta variabilidad y la presencia de valores bajos sugieren que la producción de penicilina no es completamente estable.
b. Esto puede deberse a cambios en la disponibilidad de nutrientes, ajustes en el proceso de fermentación o factores ambientales.
c. Un análisis más detallado de las condiciones del bioproceso permitiría identificar patrones más precisos.

Análisis del Flujo de PAA (PAA flow)

1. **Media (6.92 L/h) y Mediana (7.56 L/h):**
    a. La mediana es mayor que la media, lo que sugiere una ligera asimetría hacia la izquierda en la distribución de los datos.
    b. Esto indica que hay una proporción de valores bajos que afectan la media.
2. **Desviación Estándar (2.67 L/h):**
    a. Existe una variabilidad moderada en los valores de PAA flow, lo que indica que la alimentación de PAA fluctúa en diferentes momentos.
    b. Esto sugiere que el proceso de alimentación no es completamente constante y se ajusta según las necesidades del sistema.
3. **Mínimo (0 L/h) y Máximo (11.90 L/h):**
    a. Se identificaron momentos en los que el flujo de PAA es nulo, lo que podría estar relacionado con fases específicas donde la alimentación no es necesaria o es detenida temporalmente.
    b. El valor máximo de 11.90 L/h representa el nivel más alto de alimentación registrado, lo que indica la capacidad máxima del sistema en términos de suministro de PAA.
4. **Distribución de los Cuartiles:**
    a. **Q1 (5.82 L/h):** El 25% de los valores están por debajo de este nivel, indicando momentos de baja alimentación.
    b. **Q3 (8.68 L/h):** El 75% de los valores están por debajo de este umbral, lo que sugiere que la mayoría de los registros se concentran en valores cercanos a este.
    c. **IQR (Q3 - Q1 = 2.86 L/h):** El rango inter-cuartil indica una variabilidad moderada en la alimentación de PAA.
5. **Interpretación:**
    a. La distribución de los datos sugiere la existencia de dos posibles regímenes operativos:
        i. Un grupo de valores bajos (~0-5 L/h), probablemente asociado a momentos en los que la alimentación de PAA es mínima o está detenida.

ii. Un grupo con valores más cercanos a 10 L/h, lo que sugiere que este podría ser el nivel más común de alimentación.

b. Esta variabilidad puede deberse a ajustes dinámicos en la alimentación para optimizar el proceso de fermentación.

Análisis de la Correlación entre las variables Penicillin concentration y PAA flow

1. **Correlación de Pearson (0.406):**
    a. Indica una correlación positiva moderada entre el flujo de PAA y la concentración de penicilina.
    b. En términos prácticos, esto sugiere que aumentar el flujo de PAA puede estar relacionado con un incremento en la producción de penicilina, pero la relación no es completamente lineal.
2. **Correlación de Spearman (0.278):**
    a. La correlación de Spearman, que mide relaciones no lineales, también indica una relación positiva pero más débil.
    b. Esto sugiere que, aunque existe una tendencia general de que mayor flujo de PAA favorece la producción de penicilina, hay otros factores que también influyen significativamente.

La matriz de correlación muestra la relación entre el flujo de PAA y la concentración de penicilina, donde los valores en la diagonal (1.00) indican una correlación perfecta de cada variable consigo misma. El coeficiente de correlación entre "Penicillin concentration" y "PAA flow" es de 0.41, lo que representa una correlación positiva moderada. Esto significa que a medida que el flujo de PAA aumenta, la concentración de penicilina tiende a incrementarse, aunque no de manera determinante, ya que otros factores también pueden influir en la producción. El color rojo oscuro en la diagonal refuerza la relación perfecta dentro de cada variable, mientras que el azul oscuro en la correlación entre las dos variables indica que la relación existe, pero no es lo suficientemente fuerte como para considerarse una dependencia directa.

Conclusiones

El análisis de la relación entre el flujo de ácido fenilacético (PAA) y la concentración de penicilina en el proceso de fermentación industrial permite extraer conclusiones clave sobre la optimización del bioproceso. Basándonos en estudios previos sobre fermentación a escala

industrial y modelización de la producción de penicilina, queda claro que el control preciso de la alimentación de PAA es un factor determinante en el rendimiento del proceso [1].

Los resultados obtenidos en este estudio revelan una correlación positiva moderada entre el flujo de PAA y la concentración de penicilina (Pearson: 0.406, Spearman: 0.278). Esto indica que un mayor flujo de PAA puede favorecer la producción de penicilina, pero no de manera completamente lineal, ya que otros factores también influyen en la eficiencia del proceso. En particular, los datos muestran dos patrones operativos bien diferenciados:

Un grupo con valores bajos de flujo de PAA (~0-5 L/h), asociado a períodos en los que la alimentación se reduce o se detiene.

Un grupo con valores cercanos a 10 L/h, que parece representar el nivel de alimentación más frecuente durante la fermentación.

Desde una perspectiva operativa, esta variabilidad en la alimentación de PAA puede tener implicaciones importantes. Estudios previos han demostrado que es esencial mantener la concentración de PAA dentro de un rango óptimo para evitar efectos negativos en el crecimiento de la biomasa y en la producción de penicilina [1]. Mientras que una alimentación insuficiente puede limitar la síntesis de penicilina, niveles excesivamente altos pueden generar toxicidad en el cultivo y afectar la viabilidad celular [1].

Además, la dispersión observada en los valores de concentración de penicilina sugiere que la producción no es completamente estable. La desviación estándar relativamente alta (10.36 g/L) indica fluctuaciones en el rendimiento del proceso, lo que podría estar relacionado con variaciones en la disponibilidad de nutrientes o con ajustes en la dosificación de precursores.

Este análisis confirma que la alimentación de PAA juega un papel clave en la producción de penicilina, aunque su impacto está mediado por otros factores del proceso.

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