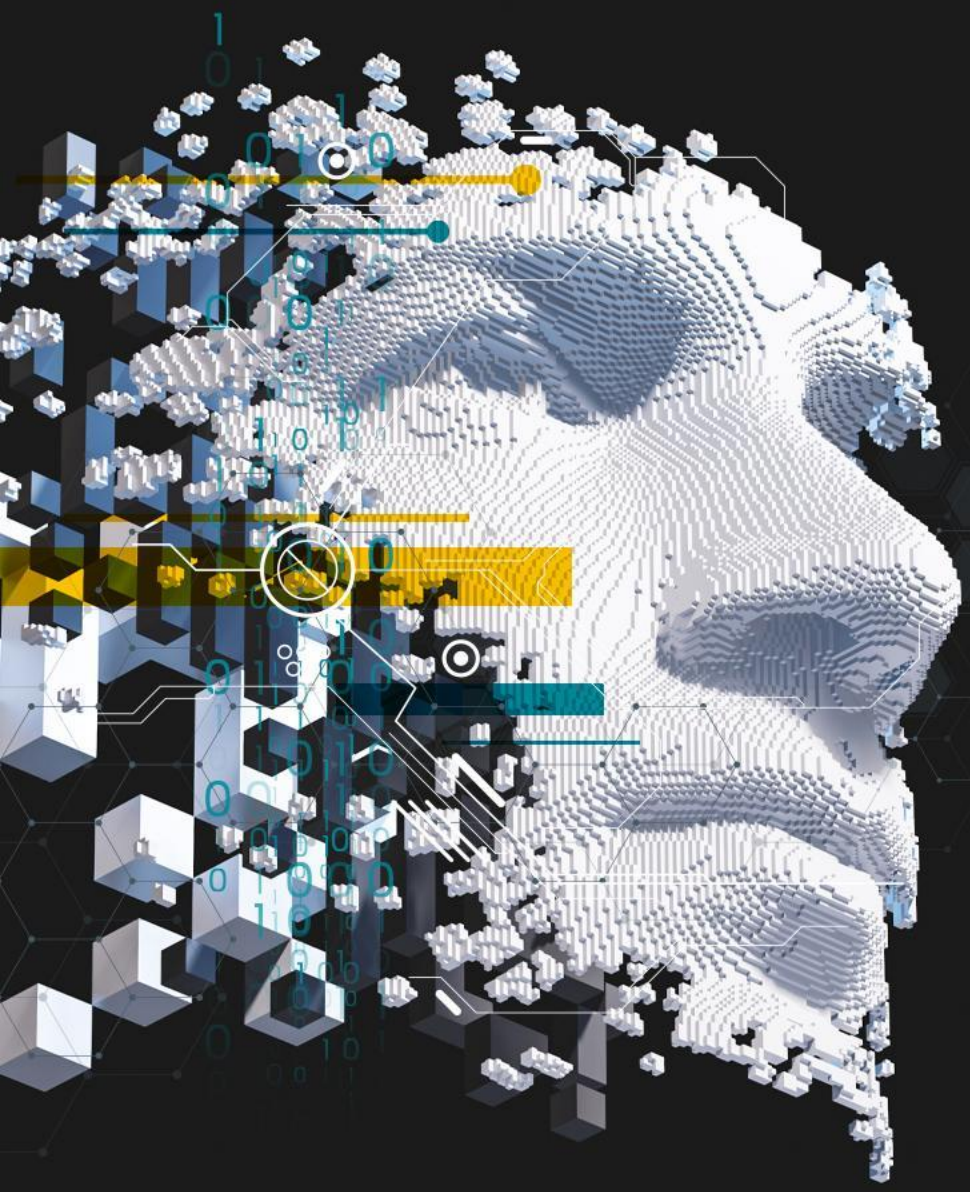


# UNIVERSITYHACK 2024®
# DATATHON

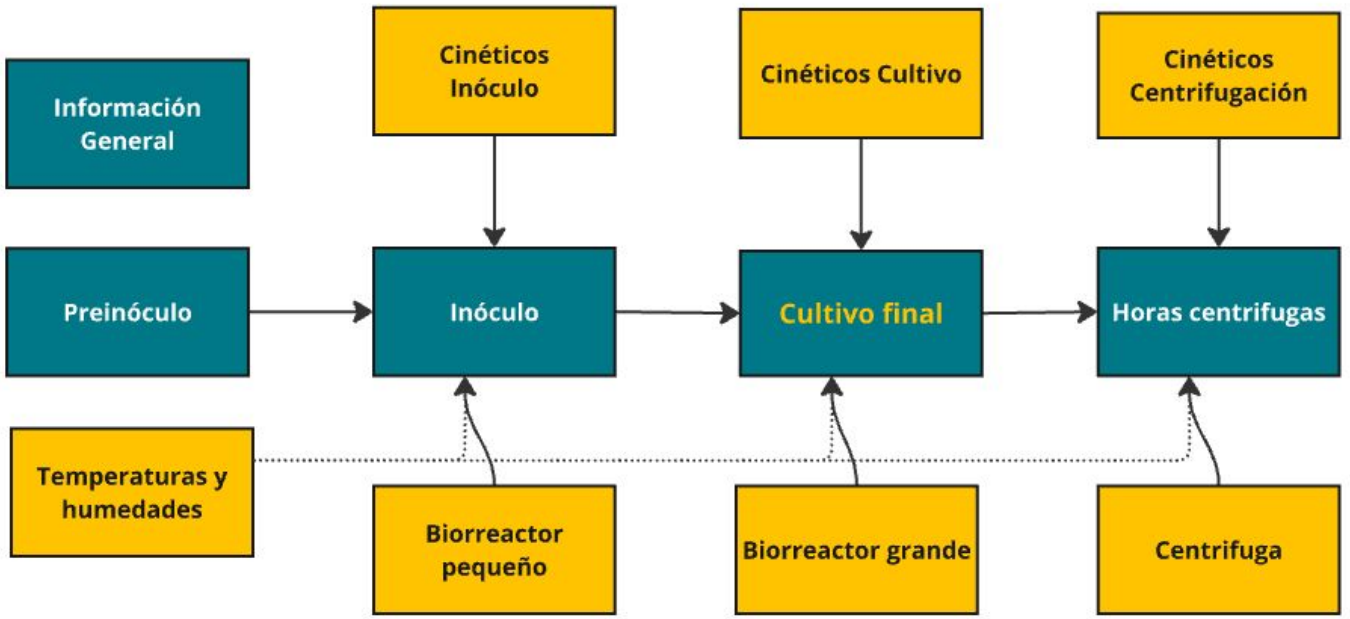
VNIVERSITAT DE VALÈNCIA

## TITOS

# 1. Exploración de los datos

*Objetivo:*
en un proceso productivo, estimar la concentración de la variable *Producto 1 en el antígeno final*

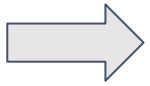
- Tablas de datos de lote: un registro por cada lote.
- Tablas de datos de evolución: una serie temporal por cada lote.



# 2. Preparación de los datos

| Lote | Variable 1 | Producto 1 |
|---|---|---|
| 001 | 90 | 10 |
| 002 | 80 | 20 |
| 003 | 110 | 0 |

| Lote | Timestamp | Variable 2 |
|---|---|---|
| 001 | 9:00:00 | 60 |
| 001 | 9:15:00 | 10 |
| 001 | 9:30:00 | 40 |
| 001 | 9:45:00 | 50 |



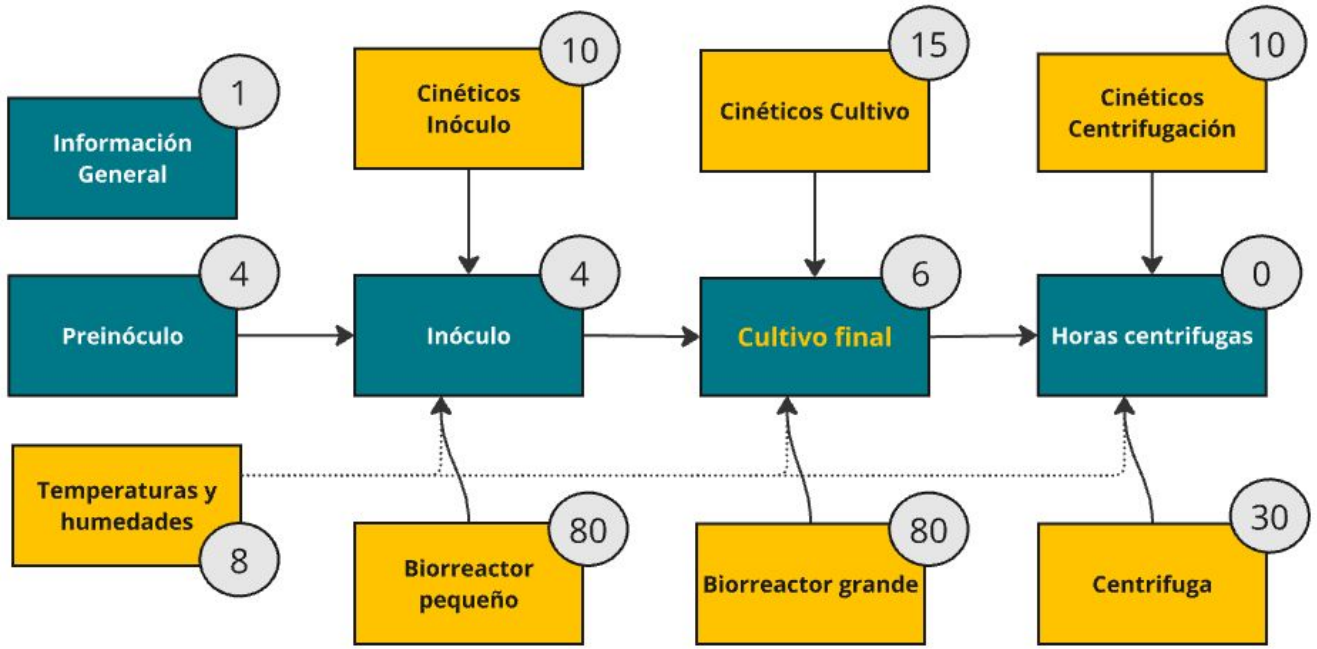
| Lote | Var 2 (media) | Var 2 (mínimo) |
|---|---|---|
| 001 | 40 | 10 |

Nota: En el proyecto calculamos la media, mediana, máximo, mínimo y desviación estándar.

| Lote | Var 1 | Var 2 (media) | Var 2 (mínimo) | Producto 1 |
|---|---|---|---|---|
| 001 | 90 | 40 | 10 | 10 |
| 002 | 80 | n | n | 20 |
| 003 | 110 | n | n | 0 |

Ejemplo: si sabemos que un alumno ha hecho dos exámenes y su media es de 7.5 con desviación estándar de 2.5, podemos inferir que las notas eran 5 y 10.

# 2. Preparación de los datos




n número de variables

Aspectos clave:

- Identificadores: Lote, ID_Bioreactor, ID_centrifuga, OF, fechas…
- Elección de frascos en *Preinóculo*
- Lotes parentales en *Inóculo*
- Estandarización de *Horas centrifugas*

Nota: Pasamos de 67 variables a 298

# 3. Selección de variables

298 variables → 259 → 43 → 29 → 26 → 18

| Varianza | Correlación | Colinealidad | Lasso | Random Forest | Valores Faltantes |
|---|---|---|---|---|---|
| A más varianza más información.<br><br>Descartamos aquellas con varianza menor a 0.001 | Una correlación con la variable a predecir indica que están relacionadas.<br><br>Eliminamos aquellas con correlación menor a 0.3 | Las variables que aportan la misma información pueden confundir al modelo.<br><br>Entre aquellas con correlación mayor a 0.95, nos quedamos con la que presente mayor correlación con la variable objetivo. | Penaliza la cantidad de variables, forzando a que solo las más relevantes mantengan un coeficiente distinto de cero.<br><br>Eliminamos las variables cuyo coeficiente resulte en cero. | Evalúa la importancia de cada variable basándose en cuánto reduce la mezcla de distintos tipos de datos en cada división.<br><br>Tras validación cruzada, nos quedamos con las que mejor resultado dan. | Imputamos los valores faltantes de cada variable según su distribución estadística.<br><br>Media para distribución normal y mediana si la distribución es asimétrica. |

Nota: correlación encuentra relaciones lineales y random forest encuentra relaciones no lineales.

# 4. Modelado y predicción de los datos

## *Modelos explorados*

1. Regresión Lineal
2. Regresión Ridge
3. **Regresión Lasso**
4. Regresión Elastic Net
5. Maquina de vector soporte regresiva
6. Regresión de los K vecinos más cercanos
7. Arbol de decisión regresivo
8. Random Forest regresivo
9. Regresión Gradient Boosting
10. Regresión XGBoost
11. Regresión LGBM
12. Regresión CatBoost
13. Ensemble de Lineal, Ridge y Elastic Net

## *¿Por qué elegimos Regresión Lasso?*

- Comúnmente usado desde 1996
- Muy habitual en problemas de genética (pocas muestras y muchas variables, y en el conjunto de entrenamiento de la fase dos tenemos 151 registros)
- Poco tiempo de computación (en nuestro caso se entrenó en 0.06 segundos)
- Perfectamente explicable
- Excelentes resultados

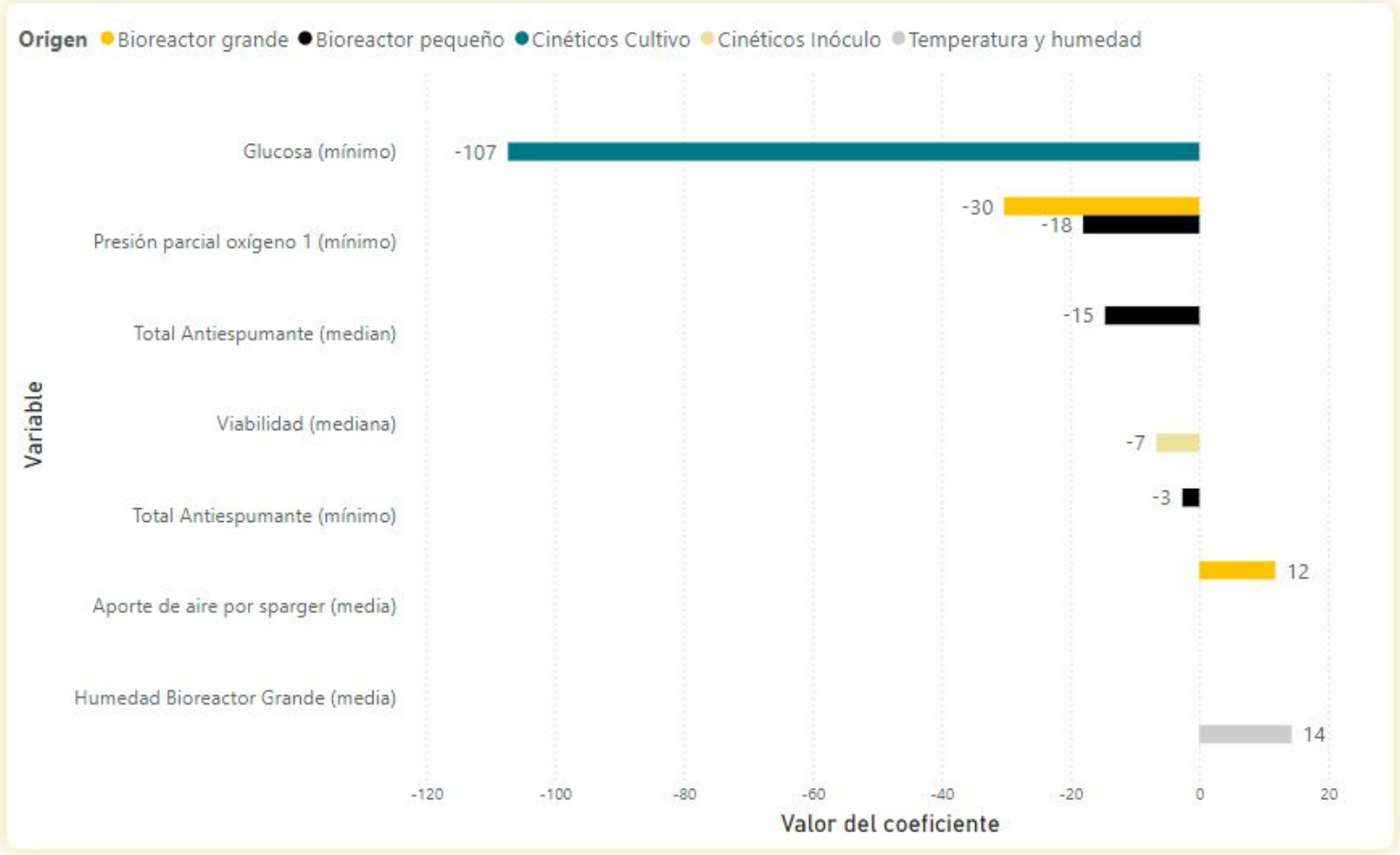
# 5. Regresión Lasso

Regresión Lineal $$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}))^2$$

Regresión Lasso $$\min_{\beta_0,\beta_1,\ldots,\beta_p} \left(\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)$$

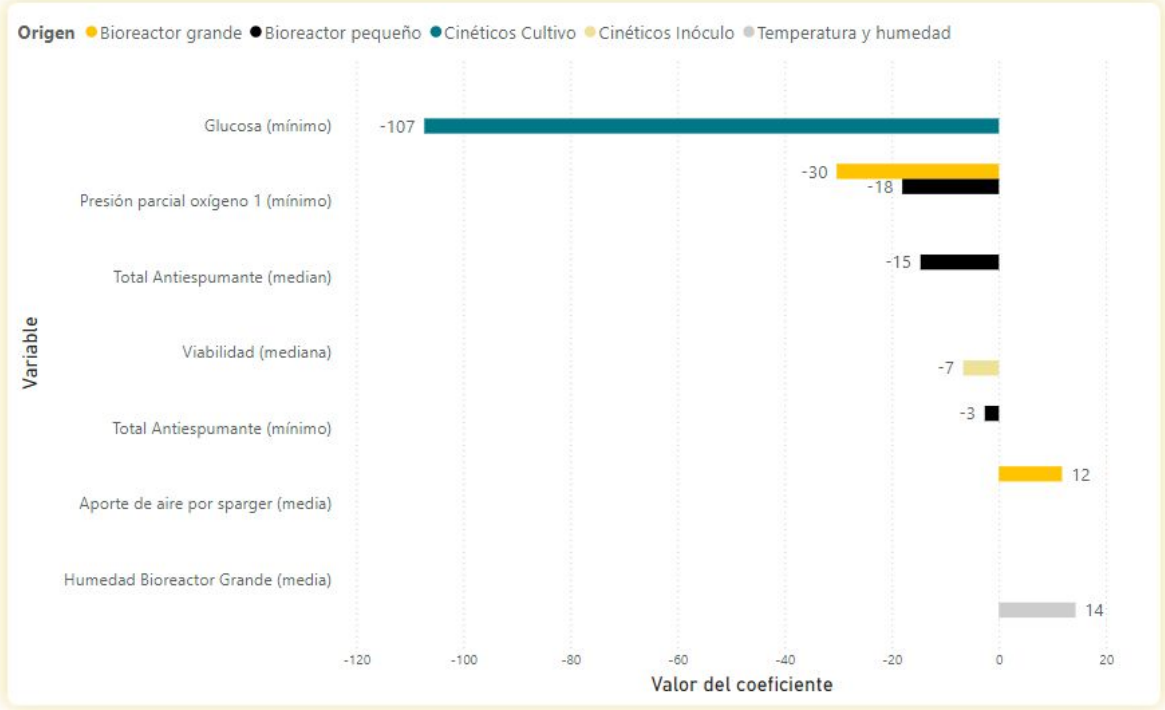
- Aplica una penalización que puede llevar los coeficientes a cero, simplificando el modelo.
- De forma general, reduce los coeficientes de todas las variables.
- Todo esto tiende a aumentar el sesgo y disminuir la varianza, mejorando la generalización sobretodo en datos ruidosos.

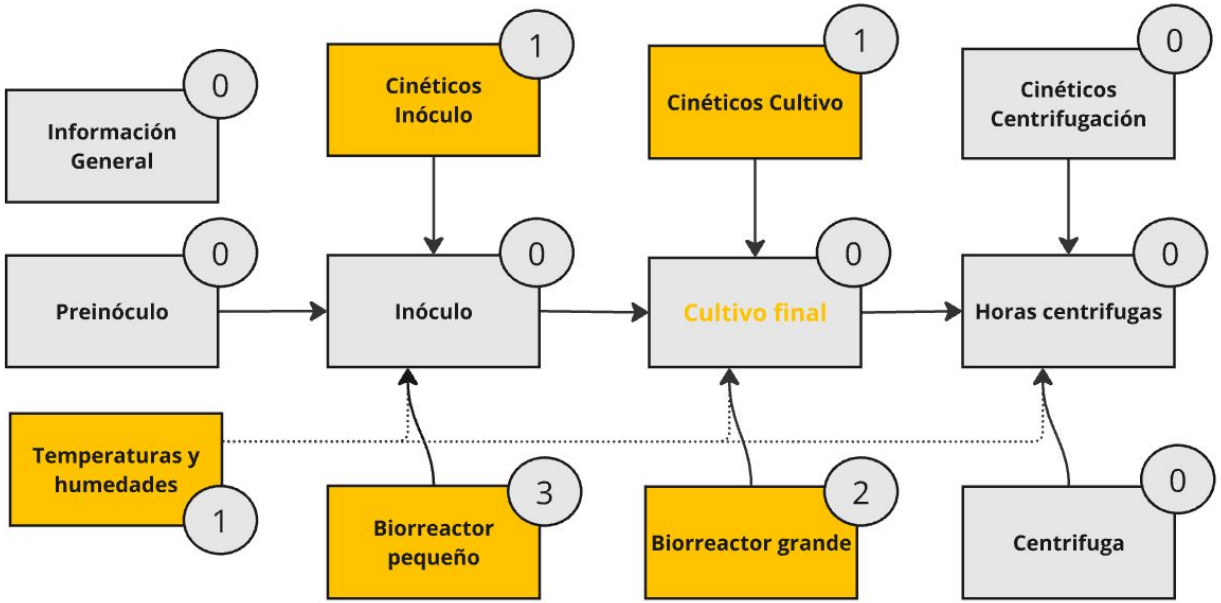
# 6. Visualización y explicación de los resultados




Nota: Un coeficiente de -107 nos indica que por cada unidad que se aumente, manteniendo el resto de variables constantes, el Producto 1 se reducirá en 107 unidades.

# 6. Visualización y explicación de los resultados

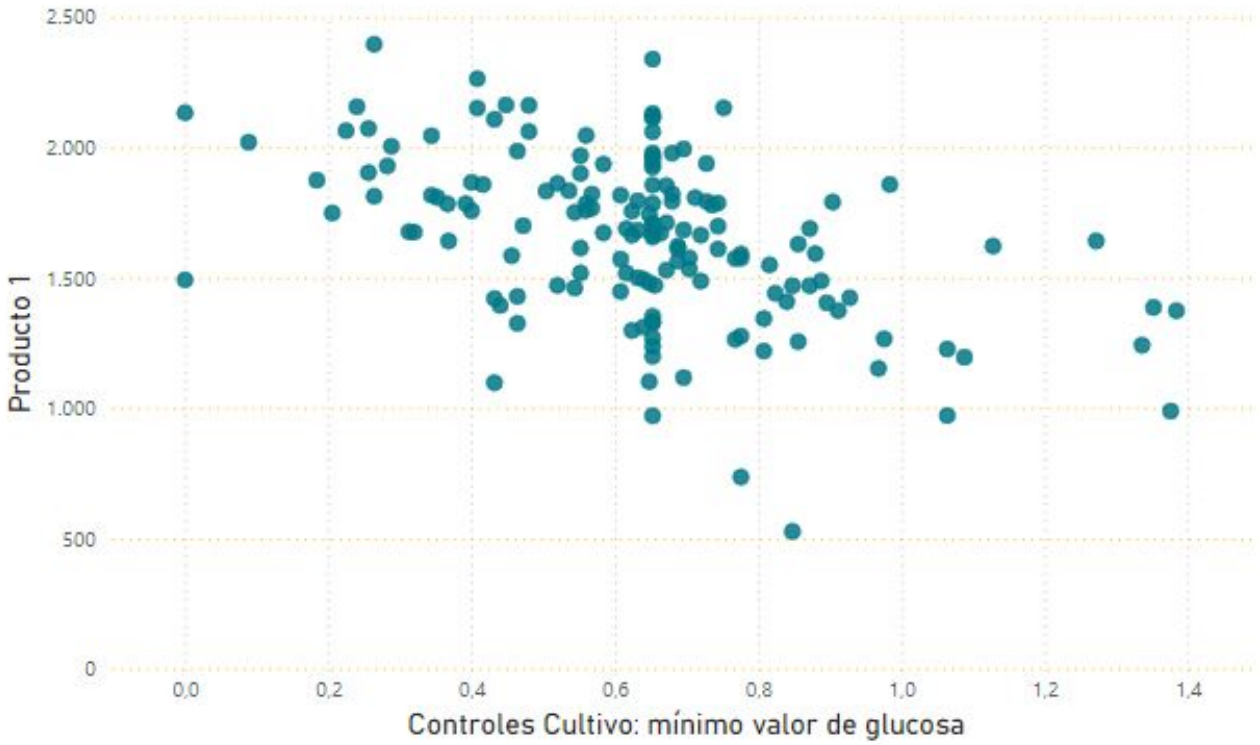



## Puntos clave

- Tablas de datos de evolución.
- Bioreactores.
- Glucosa y Presión parcial oxígeno (mínimo).

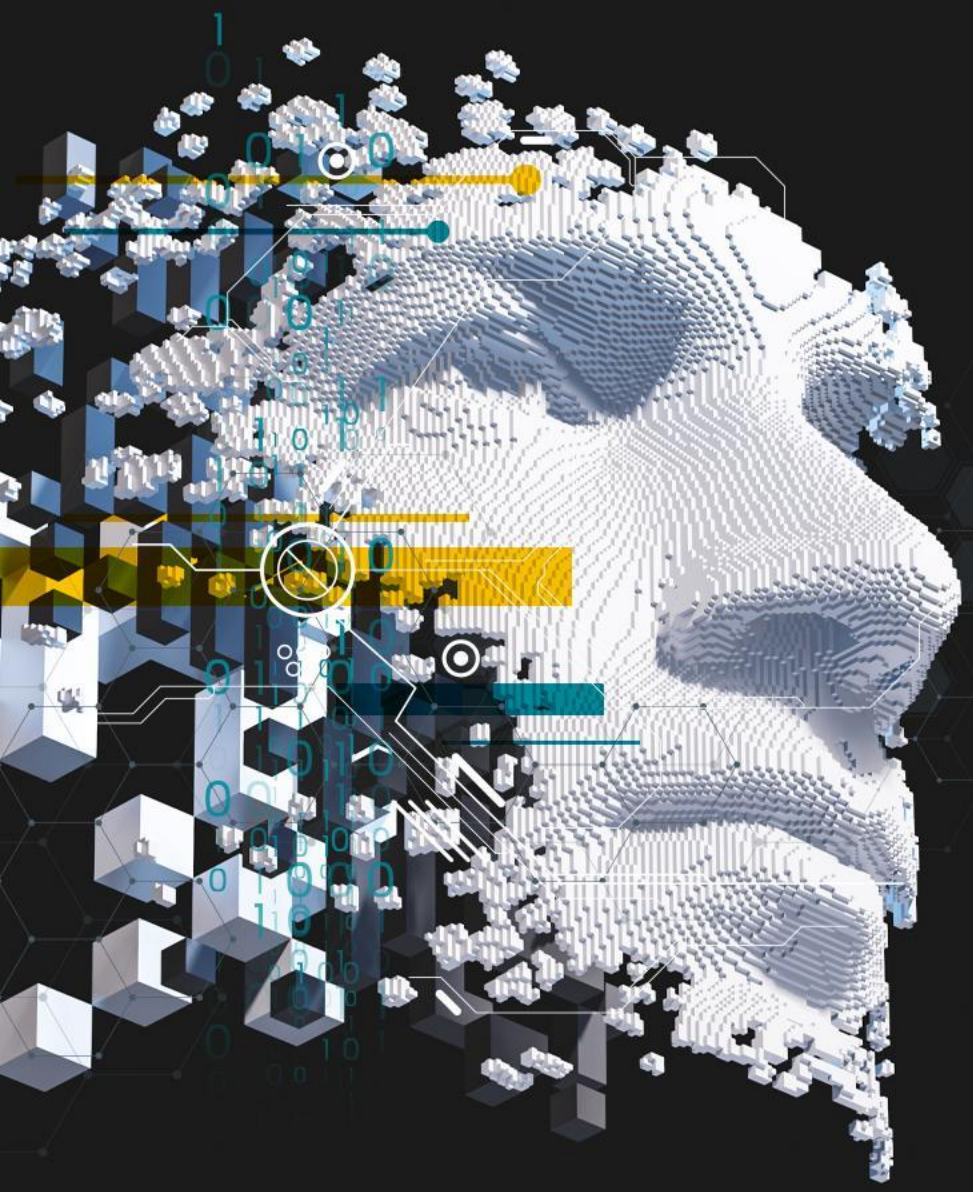
# 6. Visualización de los resultados




- Si proyectamos la variable más importante vemos una clara relación con el Producto 1

# 7. En resumen

- En inóculo tenemos 6 variables (tabla donde se encuentra Producto 1).
- Uniendolo todo y calculando estadísticos (media, mediana, máximo, mínimo y desviación estándar) obtenemos 298 variables.
- Seleccionamos variables: varianza, correlación, colinealidad, lasso, random forest. Reduciendo las variables a 18.
- Imputamos los valores faltantes, la media o mediana según su distribución.
- Usamos la Regresión Lasso para realizar la predicción (usando 8 variables).



# UNIVERSITYHACK 2024®
# DATATHON

PATROCINADORES TERABYTE

atmira · cajamar CAJA RURAL · EY · Flower · INTELYGENZ

PATROCINADORES GIGABYTE

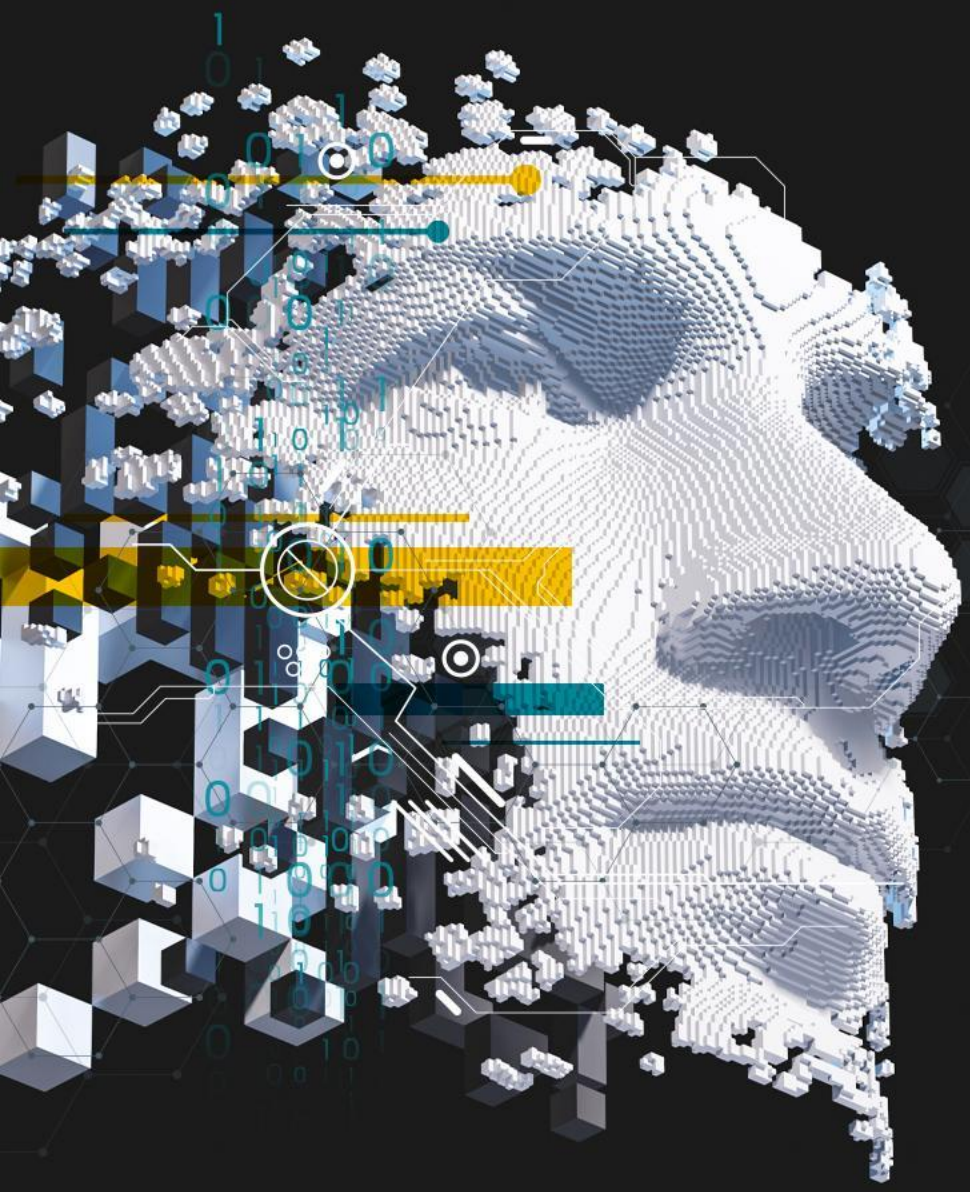
avanade · DXC TECHNOLOGY · innova-tsn · meta enlace · minsait An Indra company · NTT DATA · teradata. · VIEWNEXT AN IBM SUBSIDIARY

DATA PARTNERS

aggity

ORGANIZAN

CAJAMAR DATALAB · SPAIN AI



UNIVERSITYHACK 2024®

# DATATHON

VNIVERSITAT DE VALÈNCIA

## TITOS

Una de las competiciones de analítica de datos más importantes de España.

Muchas gracias.