

Índice

1. Breve Resumen del Trabajo Desarrollado
2. Resumen del Análisis Exploratorio
3. Manipulación de Variables y Argumentación
4. Justificación de la Selección del Modelo
5. Parámetros importantes
 - 5.1. Explicabilidad
 - 5.2. Transparencia
 - 5.3. Justicia
 - 5.4. Sostenibilidad medioambiental

1. Breve Resumen del Trabajo Desarrollado

El objetivo principal de este proyecto fue **desarrollar un modelo predictivo** capaz de estimar la concentración del producto 1 en el antígeno final después de todas las fases del proceso de producción. Para lograr esto, se utilizaron prácticamente todos los datos proporcionados, de las diferentes fases del proceso de producción: inóculo, cultivo productivo y las centrifugaciones primera y segunda.

Se llevó a cabo un extenso **análisis exploratorio** de datos y tras una comprensión exhaustiva del significado de ellos, nos dimos cuenta de que todos los conjuntos de datos guardaban una relación entre ellos y podían juntarse en un solo conjunto.

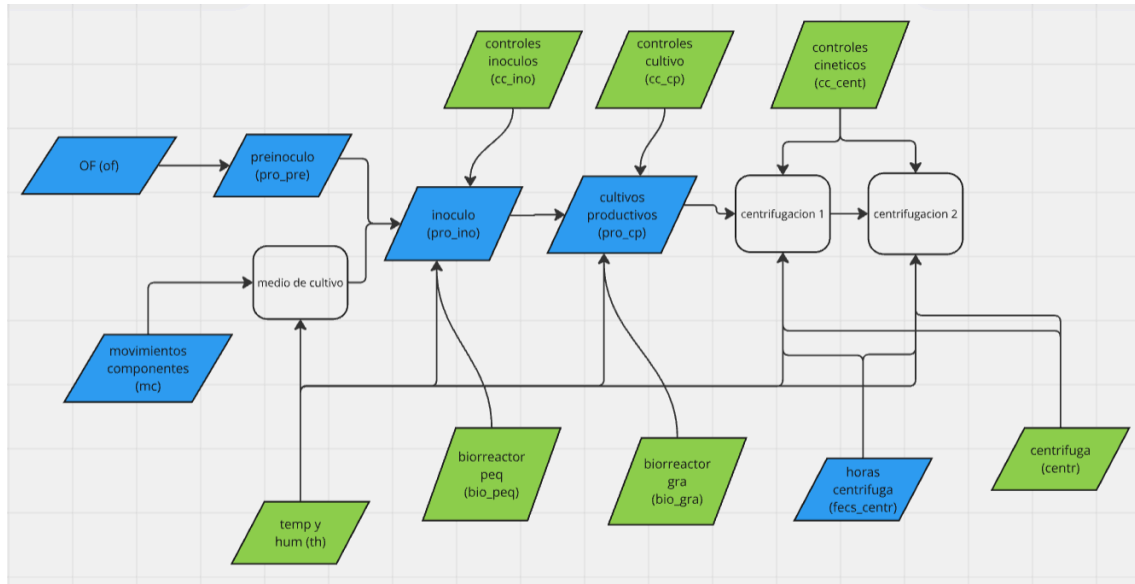


Figura 1. Boceto de la relación entre los conjuntos de datos

Así pues, nuestro primer objetivo fue **unir todos los datos** en un solo conjunto de datos. Esta parte ha sido significativamente costosa en cuestión de tiempo pero finalmente logramos unir todo a excepción del conjunto *Dataset Materias Primas Utilizadas*. En esta etapa se requirieron de algunas consideraciones especiales que explicaremos en la sección 2.



Tras la unión de los datos y dada la alta cantidad de variables, realizamos la **selección de variables importantes**. En esta etapa utilizamos primero técnicas más comunes: varianza, correlación, colinealidad; y posteriormente algunas más sofisticadas: reducción de variables según el modelo Lasso e importancia de las variables según el modelo árbol de decisión. Explicaremos el enfoque y razón de cada una en la sección 3.

En la etapa de **selección del modelo**, si bien hicimos pruebas con más de una decena de modelos distintos siempre hemos procurado que aquel elegido fuera el que mejor logra un **compromiso entre su capacidad predictiva y sus cualidades de eficiencia y explicabilidad**. Por ello, finalmente decidimos utilizar el modelo de regresión Lasso, cuyas propiedades explicaremos en la sección 4.

2. Resumen del Análisis Exploratorio

En el análisis exploratorio nos hemos centrado mucho en cómo se relacionan los datos, buscando la forma de lograr un conjunto final lo más enriquecido posible. Esto ha sido posible gracias a los identificadores de lote, centrifugadora, biorreactor, orden, número de centrifugación, y fechas, entre otros.

Los puntos clave de lo que nosotros en el script consideramos el **preprocesado** de los datos se pueden resumir en:

- **Preinóculo:** aquí se mostraban los datos correspondientes a tres frascos pero en las fases posteriores solo se utilizaban dos de ellos, por lo que era necesario descartar al tercero. Algunas veces no solo se indicaba la elección de uno por lo que generamos una selección artificial en base al frasco con segundo menor pH.
- **Cultivo productivo:** en el caso de que fueran lotes encadenados, los valores de las fases anteriores se heredan del lote parental. Esto era algo que debimos tener en cuenta y diseñamos un procesado que permitía unir al lote aquellos datos de su lote padre en las fases anteriores.
- **Horas centrifuga:** la forma en la que venían dados los datos este conjunto era muy particular y distinta a las otras. Por lo que se tuvo que estandarizar para poder unirlos con el resto de conjuntos. Dado que este proceso fue sencillo pero extenso, nos limitaremos a indicar que su explicación detallada viene dada en la función `preproces_cent_fecs` del `script_exploración.py`

Para la etapa que en el script hemos denominado **procesado**, el punto común más importante ha sido como unir lo que en el reto se conocen como *tablas de datos de lote* y las *tablas de datos de evolución*. Siendo las cuestiones más relevantes las siguientes:

- En los datos de lote teníamos un identificador y una fecha de inicio y final
- En los datos de evolución teníamos un identificador y una serie de observaciones a lo largo del tiempo
- Así pues, lo que hacemos es: para cada identificador en concreto seleccionamos su franja de observaciones en los datos de evolución y calculamos una serie de estadísticos con intención de mantener la mayor cantidad de información posible en este proceso de sintetización de los datos.
- Los estadísticos calculados para cada una de las variables numéricas de los datos de evolución han sido: media, mediana, máximo, mínimo y desviación estándar.
- Esto conlleva que para cada variable numérica en los datos de evolución, generamos cinco. Lo cual aumenta aún más el número de variables en el conjunto final de datos. (Haciendo aún más importante la etapa de manipulación y selección de variables)

Cabe destacar que para la exploración visual de los datos, todos los conjuntos de datos fueron llevados al esquema de una base de datos para posteriormente importarlas a PowerBI y realizar allí las visualizaciones pertinentes a lo largo del desarrollo.

3. Manipulación de Variables y Argumentación

Una vez que logramos unificar los datos en un solo conjunto, nos enfrentamos al desafío de manejar una gran cantidad de variables. Para construir un modelo predictivo eficiente y preciso, era esencial identificar y seleccionar las variables más relevantes. Este proceso de selección de variables se llevó a cabo en varias etapas, combinando técnicas estadísticas y algoritmos de aprendizaje automático:

1. **Eliminación de Variables con Baja Varianza:** la varianza mide cuánto se dispersan los valores de una variable respecto a su media. Las variables con varianza muy baja aportan poca información al modelo, ya que sus valores son casi constantes y no ayudan a distinguir entre diferentes resultados.
2. **Alta correlación con la variable a predecir:** las variables con alta correlación con la variable objetivo son más propensas a tener un impacto significativo en las predicciones del modelo. Cabe destacar que este método solo encuentra relaciones lineales.
3. **Eliminación de variables con alta colinealidad:** La colinealidad ocurre cuando dos o más variables predictoras están altamente correlacionadas entre sí. Esto puede causar problemas en los modelos, como inestabilidad en los coeficientes y dificultad para interpretar los resultados. A la hora de eliminar una de las dos, se elimina aquella con menor correlación con la variable a predecir.
4. **Selección de Variables mediante Regresión Lasso:** el modelo realiza una selección automática de variables al penalizar la suma de los valores absolutos de los coeficientes.
5. **Selección basada en importancia según árboles de decisión:** Los modelos de árboles de decisión, al calcular cómo dividir los datos, lo hacen evaluando la entropía (que puede interpretarse como una medida de impredecibilidad o desorden) de cada variable. De esta manera, obtienen una medida de la importancia de cada variable en función de su capacidad para reducir la entropía y mejorar la predicción. Cabe destacar que este método solo encuentra relaciones no lineales.

Tras las etapas 1, 2 y 3 pasamos de 298 a 29 variables. Tras la etapa 4 reducimos el número a 26 variables, y al final de la etapa 5 nos quedamos con un total de 18 variables.

Por último, **se imputan los datos faltantes** según la distribución de cada variable. Si una variable sigue una distribución normal, se imputa con la media de los datos. Si presenta una distribución asimétrica, se imputa con la mediana.

La razón de que el análisis exploratorio y selección de variables se haya hecho de esta manera se debe a la gran cantidad de variables. Esto nos impidió realizar un análisis pormenorizado de cada variable y centrar nuestros esfuerzos en aplicar los **principios teóricos de la ciencia de datos**.

Por último, el hecho de renombrar las columnas en la etapa de importación de los datos nos permite, una vez seleccionadas aquellas más importantes, tener una trazabilidad de que conjuntos de datos de origen han terminado siendo más relevantes:



| ID | Origen | Descripción | Estadístico |
|-----------------------------|-----------------------|---------------------------------------|---------------------|
| bio_gra_aire_spar_mean | Bioreactor grande | Aporte de aire por sparger | Media |
| bio_gra_pp_oxi1_min | Bioreactor grande | Presión parcial oxígeno 1 | Mínimo |
| bio_gra_th_hum | Temperatura y humedad | Humedad sala biorreactores | |
| bio_gra_tot_antiesp_mean | Bioreactor grande | Total antiespumante | Media |
| bio_peq_pp_oxi1_min | Bioreactor pequeño | Presión parcial oxígeno 1 | Mínimo |
| bio_peq_tot_antiesp_mediana | Bioreactor pequeño | Total antiespumante | Mediana |
| bio_peq_tot_antiesp_min | Bioreactor pequeño | Total antiespumante | Mínimo |
| cc_cent_tur_max_y | Controles cinéticos | Valor de turbidez | Máximo |
| cc_cent_tur_min_x | Controles cinéticos | Valor de turbidez | Mínimo |
| cc_cp_gluc_min | Controles cultivo | Valor de glucosa | Mínimo |
| cc_cp_gluc_std | Controles cultivo | Valor de glucosa | Desviación estándar |
| cc_cp_tur_mean | Controles cultivo | Valor de turbidez | Media |
| cc_cp_tur_std | Controles cultivo | Valor de turbidez | Desviación estándar |
| cc_ino_viab_median | Controles inóculo | Valor de viabilidad | Mediana |
| cent_ape_valv_median_x | Centrifuga | Apertura válvula agua maniobra | Mediana |
| cent_contrapes_max_x | Centrifuga | Contrapresión | Máximo |
| pro_cp_cent2_tur | Cultivos productivos | Turbidez de la segunda centrifugación | |
| pro_ino_viab_fin | Cultivos inóculos | Indicador de células vivas al final | |

Tabla 1. Estudio de las variables finalmente seleccionadas.

4. Justificación de la Selección del Modelo

Se han probado los siguientes modelos:

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 las regresiones Lineal, Ridge y Elastic Net

Cabe destacar que si bien en este punto habíamos conseguido reducir el número de variables de forma considerable (de 298 a 18) el número de muestras es escaso, teniendo en esta última etapa de la competición tan sólo 157 registros. Cuestión por la que, por ejemplo, hemos descartado probar a entrenar y evaluar modelos de Deep Learning.

Tras un estudio del rendimiento y funcionamiento de cada modelo, seleccionamos la regresión Lasso debido a su equilibrio óptimo entre capacidad predictiva, eficiencia computacional y explicabilidad.

La regresión Lasso (Least Absolute Shrinkage and Selection Operator) es una variante de la regresión lineal que incorpora una penalización L1 en la función de pérdida. Esto ayuda a prevenir el sobreajuste penalizando los coeficientes grandes, ayudando pues a que el modelo generalice bien a datos nuevos. Además, para mejorar esto último hemos utilizado técnicas de validación cruzada y búsqueda iterativa de los mejores hiperparámetros.

5. Parámetros importantes

5.1. Explicabilidad

La regresión Lasso es altamente explicable porque es un modelo lineal que permite interpretar directamente cómo cada variable influye en la variable objetivo. Al aplicar una penalización L1, Lasso reduce a cero los coeficientes de las variables menos importantes, simplificando el modelo y dejando solo las variables más relevantes. Esto facilita la comprensión de la relación entre las variables predictoras y la concentración del producto 1, ya que los coeficientes indican la dirección y magnitud del impacto de cada variable. En nuestro caso, los pesos del modelo final son:

| Variable | Valor del coeficiente |
|----------------------------|-----------------------|
| bio_gra_th_hum | 14,3206 |
| bio_gra_aire_spar_mean | 11,7831 |
| cc_cent_tur_max_y | 0 |
| cent_ape_valv_median_x | 0 |
| cc_cp_tur_std | 0 |
| cc_cp_gluc_std | 0 |
| pro_cp_cent2_tur | 0 |
| cc_cent_tur_min_x | 0 |
| cc_cp_tur_mean | 0 |
| pro_ino_viab_fin | 0 |
| cent_contrapas_max_x | 0 |
| bio_gra_tot_antiesp_mean | 0 |
| bio_peq_tot_antiesp_min | -2,639 |
| cc_ino_viab_median | -6,6798 |
| bio_peq_tot_antiesp_median | -14,6428 |
| bio_peq_pp_oxi1_min | -18,0497 |
| bio_gra_pp_oxi1_min | -30,2968 |

| | |
|----------------|-----------|
| cc_cp_gluc_min | -107,3019 |
|----------------|-----------|

Tabla 2. Coeficientes del modelo Lasso utilizado

Así pues, podemos saber en qué medida y dirección un aumento en el valor de cada variable afecta al valor a predecir (concentración del producto 1). Cabe destacar que algunas tienen coeficiente 0 porque como se ha explicado previamente, Lasso realiza por defecto selección de características al penalizar coeficientes grandes.

Gracias a haber elegido este modelo podemos implementar técnicas más sofisticadas para conocer la importancia de las variables, como por ejemplo los Índices SHAP. El valor SHAP (SHapley Additive exPlanations) es una medida que indica cuánto contribuye cada variable a la predicción de un modelo. El siguiente gráfico nos ayuda a conocer el efecto global (independientemente de su dirección positiva o negativa) de las variables usadas.

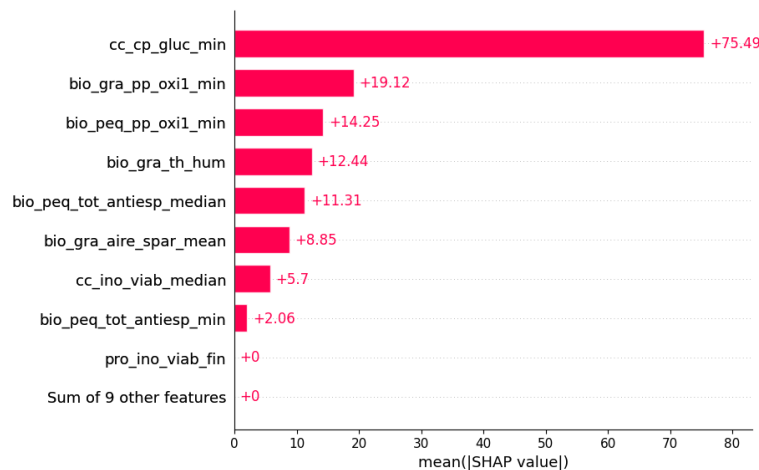


Figura 2. Gráfico de Importancia de Características según su índice SHAP

5.2. Transparencia

- **Instrucciones de uso:** una vez instaladas las librerías, basta con ejecutar el script `python script_prediccion.py` para ejecutar todo el proceso de tratamiento de datos, selección de variables y predicción. Los conjuntos de datos proporcionados deben estar al mismo nivel que el script.
- **Tratamiento sobre los datasets de datos:** a lo largo de esta memoria hemos explicado los puntos clave del tratamiento realizado, el detalle y explicación de cada paso viene recogido a lo largo de los scripts `script_exploracion.py` y `script_prediccion.py`
- **Elección de la muestra de entrenamiento y validación:** viene dada por la organización de la competición.
- **Argumento de la tipología del modelo a desarrollar:** punto desarrollado en la sección 4.
- **Criterios aplicados para la selección del ganador:** punto desarrollado en la sección 4.
- **Visualización y explicación de los resultados:** punto desarrollado en la subsección 5.1.

5.3. Justicia

Si bien nos habría encantado, dada la naturaleza técnico-científica de los datos no ha sido posible aportar en esta cuestión. Es cierto que el modelo da buenos resultados por ser insesgado, es decir, por su capacidad de generalizar. Pero entendemos que en el requisito de que sea un modelo justo, se hace referencia a sesgos en el sentido social.

5.4. Sostenibilidad ambiental

La razón por la que elegimos Regresión Lasso ha sido por su capacidad excepcional de encontrar un compromiso entre capacidad predictiva, explicabilidad y eficiencia computacional.

La búsqueda de la mayor eficiencia computacional ha condicionado nuestra solución desde el principio, y hemos procurado optimizar en todo momento tanto el procesamiento de los datos como el entrenamiento del modelo.

Los tiempos en concreto son:

- Fase de tratamiento de los datos: 87,30 segundos
- Fase de selección de variables: 10,26 segundos
- Fase de entrenamiento del modelo: 0,06 segundos
- Tiempo total del script predicción: 97,63 segundos

Si bien nos gustaría que la fase de entrenamiento de los datos tuviera un tiempo más reducido. Debemos considerar que hay un tiempo relativamente irreducible, ya que al fin y al cabo muchos conjuntos de datos son del orden de decenas de miles y en ocasiones varios cientos.

La fase de selección de variables requiere un poco de tiempo, pero podemos considerar que una vez identificadas las más importantes, de ese momento en adelante bastaría con su selección directa en lugar de el uso de métodos estadísticos, por lo tanto esta parte sería prácticamente inmediata.

Por último, considerando que el tiempo de entrenamiento del modelo es prácticamente instantáneo, nos alegra mucho haber conseguido buenos resultados con un modelo tan completo como ha resultado ser la Regresión Lasso.