UNIVERSITYHACK 2024 - Proceso y metodología aplicada

Con el fin de seguir un orden y una metodología adecuada, dividimos el proceso de depuración de los datos y entreno de los modelos en 3 etapas:

- Clean
- Structured
- Analytics

ETAPA 1: CLEAN

Durante la primera etapa, partimos de todos los Excel que teníamos, y los "limpiamos", asegurándonos de que las columnas tuvieran los formatos y nombres necesarios para poder realizar cruces entre ellos y trabajar correctamente.

En segundo lugar, pivotamos varias tablas y juntamos otras, para que las tablas fueran más sencillas de unir y así trabajar de manera más sencilla. Realizamos cambios en las siguientes tablas:

- Biorreactores pequeños: Unimos todas las tablas que pertenecían a datos de los biorreactores pequeños, ya que contenían las mismas columnas. Asociando cada fila con el "id_biorreactor" al que pertenecía.
- Biorreactores grandes: Del mismo modo que con los pequeños, obtuvimos una única tabla para los biorreactores grandes.
- Centrífugas: De manera similar a los biorreactores, unimos las tablas de centrífugas en una, tomando el "id_centrífuga" al que pertenecía cada fila.
- Materiales (Movimientos Componentes): Pivotamos la tabla de "Movimientos Componentes" con el fin de tomar una única fila para cada lote. Creamos 13 columnas para las 13 componentes distintas que aparecían, y tomamos la cantidad total de cada componente en cada lote.
- Centrifugación: Pivotamos la tabla de "Horas inicio-fin centrífugas". Creando 4 columnas con las fechas de inicio-fin de centrifugación 1 o 2, obteniendo así una única fila para cada lote.

En tercer lugar, para realizar nuestro esquema, la tabla de Cinéticos IPC fue dividida en 2 tablas. Por un lado, tomamos Cinéticos IPC – Inóculo; y, por otro lado, Cinéticos IPC – Cultivo.

Finalmente, el resto de tablas las dejamos como estaban (Fases producción [Preinóculo, inóculo, cultivo final], OF123456 ["Orden de Fabricación"] y Temperaturas y humedades).

ETAPA 2: STRUCTURED

Durante esta etapa, creamos un esquema con todas las tablas que finalmente teníamos y las relaciones que existían entre ellas, con el fin de realizar "JOINS" y obtener una tabla final. La idea era que esta tabla final constara de una única fila para cada lote, y que tuviera todas las variables de todas las tablas asociadas.

Adjunto una imagen del esquema final:

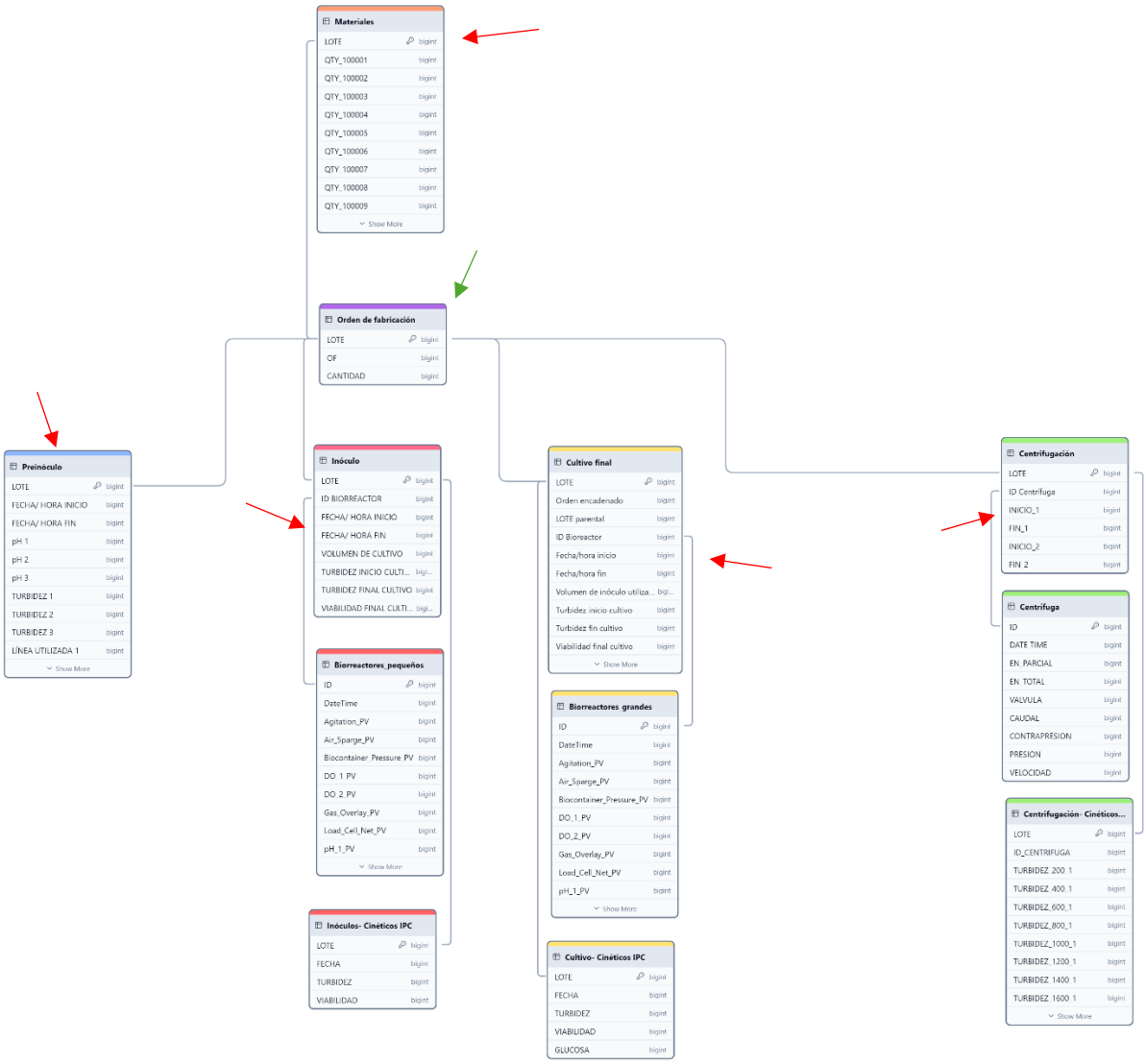

Imagen 1: Esquema de las tablas

Y, comenzamos a unir cada una de las tablas mediante JOINS.

La idea principal era, ir de “abajo a arriba”. Es decir, obtener primero las 5 tablas señaladas en rojo tras unirlas con las que tienen relación. Estas tablas son Preinóculo, Inóculo, Cultivo final, Centrifugación y Materiales. Todo ello para, posteriormente, unificar todas con Orden de Fabricación (señalada en verde). Por tanto, finalmente, obtenemos una única tabla con una fila para cada lote y 203 variables.

Cabe destacar que, a la hora de unir las tablas tuvimos que utilizar varias técnicas:

Por un lado, había tablas en las que no teníamos el número de lote explícito, por lo que a la hora de realizar los JOIN, tuvimos que utilizar las columnas con fechas y horas. Para ello, calculamos distintas métricas en función de las horas que correspondían a cada lote:

- Unión Biorreactores Pequeños e inóculo: En biorreactores tenemos distintas mediciones en distintos momentos, representados en una columna de fecha-hora.

Lo que hicimos fue, a la hora de unir estas tablas, como teníamos el intervalo de tiempo que pasaba el lote en el biorreactor, calculamos la media, el mínimo y el máximo de todas las variables durante el tiempo que cada lote estuvo en el biorreactor. Para así, terminar uniendo estos resultados con sus lotes asociados.

- Unión Biorreactores Grandes y Cultivo Final: Trabajamos de la misma manera que antes.
- Unión Centrífuga y Centrifugación: Trabajamos de la misma manera que antes.

Por otro lado, había algunas tablas que no contenían una única fila para cada lote. Entonces, a la hora de realizar la unión, decidimos agrupar las filas por lote obteniendo la media, desviación típica, mínimo y máximo de cada una de las variables en función del lote. Éstas son:

- Cultivos - Cinéticos IPC (Unión con Cultivo Final)
- Inóculos – Cinético IPC (Unión con Inóculo)

Entonces, conseguimos tener las 5 tablas con una única fila para cada lote y todas las variables asociadas. Y, finalmente, realizamos una unión de todas las tablas con la tabla de Orden de Fabricación, obteniendo un único Excel con el que trabajar.

Este mismo proceso lo repetimos tanto para el Cultivo Final del Train como para el Cultivo Final del Test. Obteniendo tablas similares, pero con distintos lotes, con las que comenzar a trabajar en los modelos predictivos.

ETAPA 3: ANALYTICS

Una vez obtuvimos la tabla final, comenzamos realizando un EDA para ver cómo se comportaba la target (Producto 1) y las variables. Para ello, dibujamos matrices de correlación e histogramas, con el fin de observar qué distribución seguía la target y con qué variables estaba más relacionada. También, para observar la relación que había entre las variables y sus distribuciones.

Debido a que ninguna variable tenía mucha correlación con Producto 1 y teníamos un número alto de columnas para empezar con una regresión lineal, descartamos la idea de comenzar con ello. Por otro lado, también pensamos que había un número muy bajo de datos y muchos nulos para trabajar con una red neuronal. Entonces, nos decantamos por profundizar en otro tipo de métodos.

Realizamos muchas pruebas creando un código que “probara” los modelos con distintos hiperparámetros, obteniendo métricas con validación cruzada, con el fin de quedarnos con el que nos aportara un RMSE medio menor. Durante este proceso aplicamos Árboles de decisión, RandomForest, XGBoost, Catboost, SVR y Gradient Boosting. Con la mayoría de los mejores modelos obteníamos un RMSE medio entre 220 y 280.

Tras realizar muchas pruebas, observamos que lo que mejor estaba funcionando eran los XGBoost, por lo que seguimos explotando este modelo, probando con más y más hiperparámetros. Finalmente, tras realizar unas entregas de los mejores resultados en CodaBench, decidimos quedarnos con el siguiente:

- **xgb = XGBRegressor(n_estimators=250, max_depth=3, learning_rate=0.05, random_state=73).fit(X, y)**

Con él, en la prueba de CodaBench obtuvimos un RMSE = 448.64.

Y, tras haber obtenido los resultados con los datos de test, hemos intentado analizar cómo funciona el modelo y qué variables son más significativas. Para ello, hemos tomamos los Shap Values. Podemos destacar algunas variables como IPC_CULTIVO_GRUCOSA_MIN (cuando mayor es esta variable, menor es el contenido de Producto 1 en el lote) o IPC_CULTIVO_TURBIDEZ_MAX (cuando mayor es esta variable, mayor es el contenido de Producto 1).

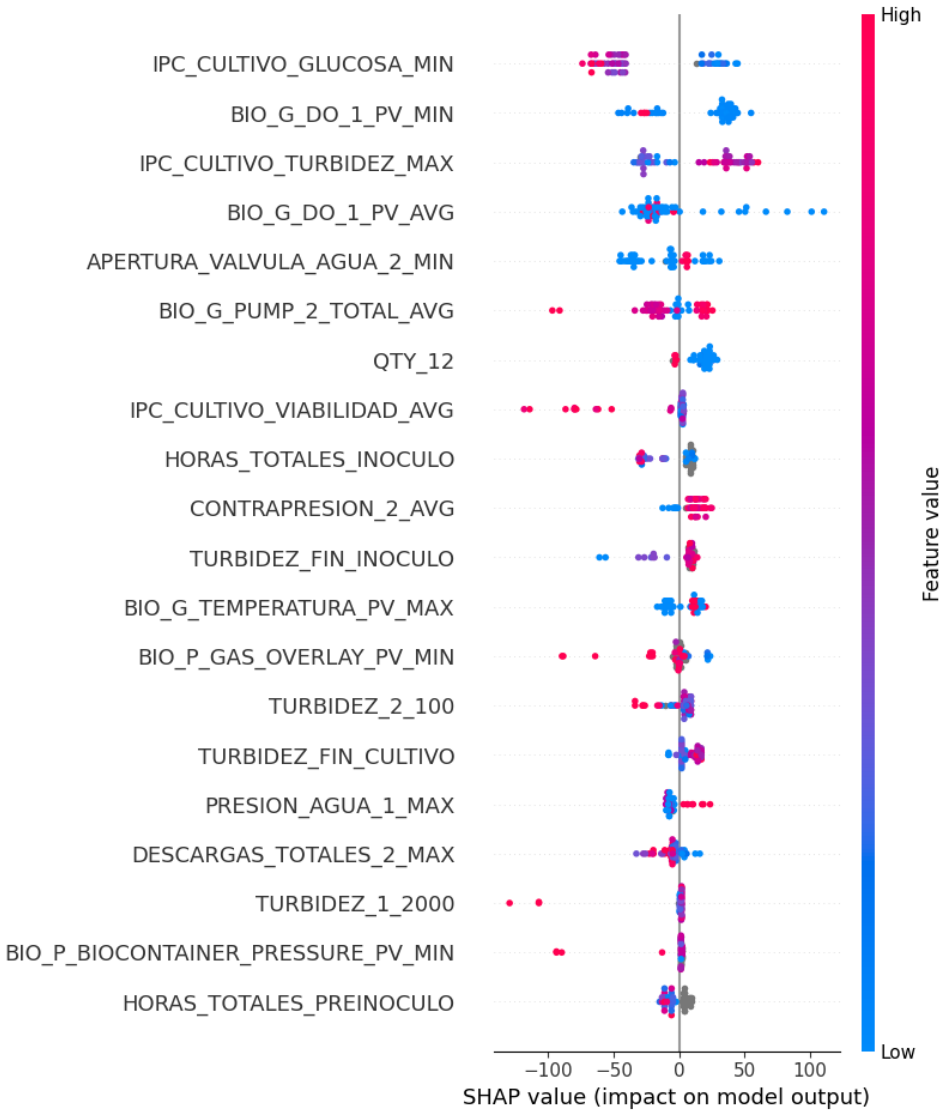

Imagen 2: SHAP Values del modelo final