Estructura de Prueba Técnica de Conocimientos
AREA DE BIOECONOMIA Y DOCUMENTACIÓN

Postulante a cargo: Estefania Aracena

Parte 1: Análisis de Datos y Modelación Bio-económica

El archivo "(PRUEBA TÉCNICA_AREA DE BIOECONOMIA.XLSX/HOJA PARTE_1" cuenta con las siguientes características:

- Contiene 35 atributos y 684 filas
- Cuenta con 2 atributos de fecha, 5 categóricos y 28 numéricos
- Existen datos Nan, los que corresponden a datos de la fila 188. Se reemplazaron por la media general del set de datos para no eliminarlos
- No se observan valores atípicos, todos dentro de rango (según criterio propio)
- Se grafican los histogramas de las variables numéricas: 'Ciclo (Días)', 'Número Smolt', 'Peso Smolt (g)', 'Número Final', 'Peso Final (g)', 'N°Muertos', 'Temperatura (°C)', 'Acc. GF3'], en donde se observa una distribución normal
- Se observa la matriz de correlación para las variables mencionadas en donde se rescata:
    - No existe correlación entre la variable Temperatura y Acc.GF3
    - Podemos destacar la correlación entre el Ciclo(días) y el peso final, peso final y Acc.GF3

Los modelos predictivos se realizarán con las variables: 'Grupo', 'Estación', 'Ciclo (Días)', 'Número Smolt', 'Peso Smolt (g)', 'Número Final', 'Peso Final (g)', 'N°Muertos', 'Temperatura (°C)', 'Acc. GF3':

1. Se codificarán las variables categóricas 'Grupo' y 'Estación'
2. Se divide el set de datos para el entrenamiento (70%) y prueba del modelo (30%)
3. Los modelos elegidos son: SVM y Random Forest. Debido a que el set de datos cuenta con 684 filas, no se consideró necesario realizar una red neuronal para la predicción ya que el requerimiento computacional es mayor

Discusión de los resultados obtenidos y argumentos sobre cómo se podrían mejorar de dichos resultados:
El modelo que mejores predicciones entregó es el *Random Forest Regressor*. Podemos observar que el valor r^2 para el entrenamiento es de 0.97 y para testeo de 0.87. Esto indica que los valores tienen una alta correlación entre, lo que además, visualmente, indica lo mismo. La raíz del error cuadrático medio (RMSE) es de 0.04 y 0.09 para los datos de entrenamiento y prueba, los cuales están cerca a 0, indicando que el modelo es bueno.

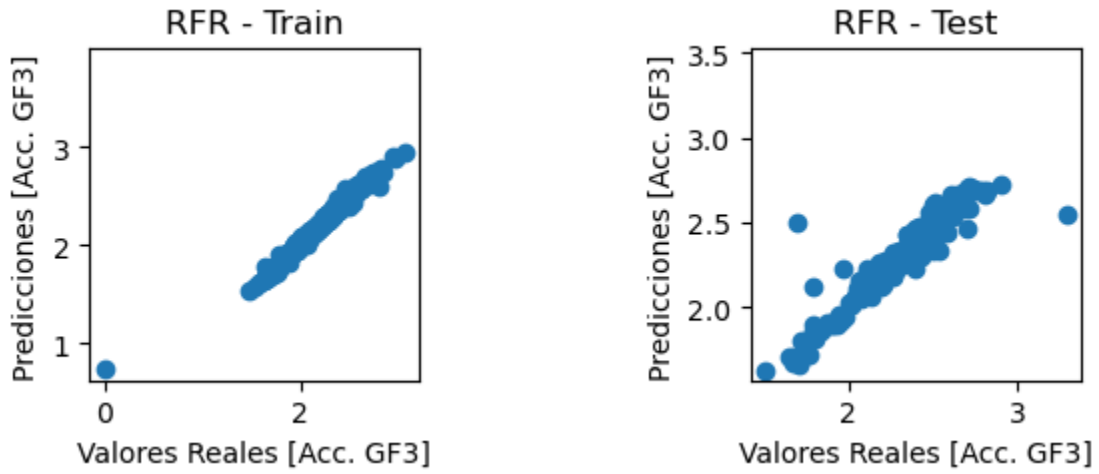

Figura 1: Gráfico de correlación para RFR de datos de prueba y predictivos

Para el modelo de *Support Vector Machine*, los indicadores no son buenos y se descarta el modelo. r^2 para el entrenamiento es de -291216 y para testeo de -297185. Claramente nos encontramos con un modelo no confiable y visualmente, se observan la no correlación entre los datos de prueba y entrenamiento.

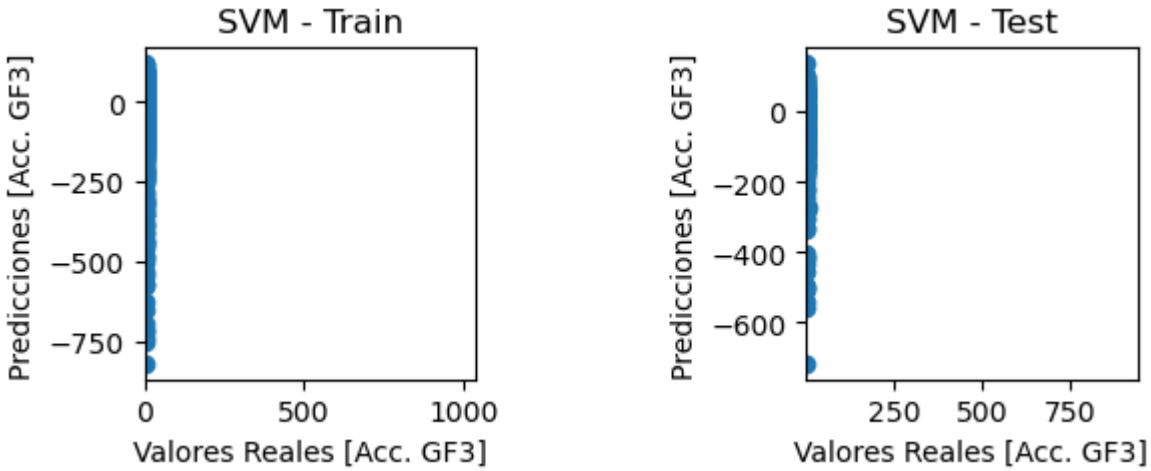

Figura 2: Gráfico de correlación para RFR de datos de prueba y predictivos

- Recomendaciones:
    - Verificar otros modelos de predicción, como redes neuronales
    - Utilizar un set de datos con mayor número de instancias y subdividirlo con una estrategia 60-20-20 (%) o 70-20-10 (%) para entrenamiento, testeo y prueba.

Parte 2: Automatización de Modelos de Crecimiento con Machine Learning

El archivo "(PRUEBA TÉCNICA_AREA DE BIOECONOMIA.XLSX/HOJA PARTE_2" cuenta con las siguientes características:

- Contiene 5 atributos y 4952 filas
- 2 atributos de fecha, 2 numéricos y 1 categórico, el cual cuenta con 1 sola categoría
- Cuenta con datos atípicos, como el número de días negativo en algunos valores. Estos se corrigen
- Se realizó un análisis de regresión entre las variables "Tiempo = Días" y "Peso Inicial Periodo (g)", indicando lo siguiente:
    - El valor r^2 es de 0.835, lo cual indica que hay una alta correlación entre las variables

Los modelos de optimización seleccionados son: modelo de regresión logística y modelo Gompertz.

Primero se definieron las ecuaciones y sus respectivos parámetros, los cuales se optimizarán. Se trabajó con las columnas "Tiempo = Días" y "Peso Inicial Periodo (g)".

Se tienen los siguientes casos:

- Datos utilizados en el set de datos recibido (ejemplo)
    - La regresión logística arrojó r^2 =0, o cual observamos gráficamente, no se ajustó a los datos de entrada. Por otro lado, el modelo Gompertz se ajustó bien al modelo con un r^2 de 0.997.
    - Para este caso, el punto óptimo es el máximo, ambos coinciden

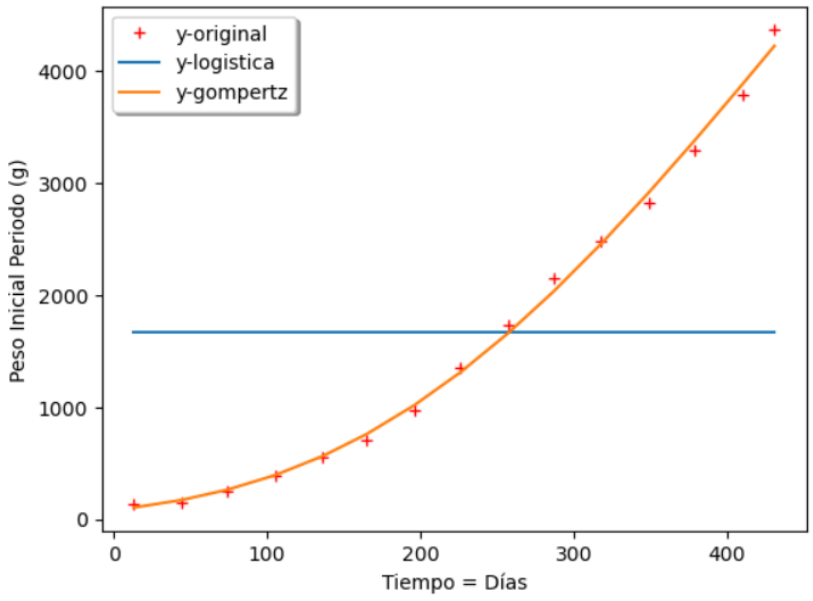

Figura 3: Regresión para el primer grupo de datos

- Set de datos completo:
    - Al utilizar todos los datos del set, observamos la siguiente figura:

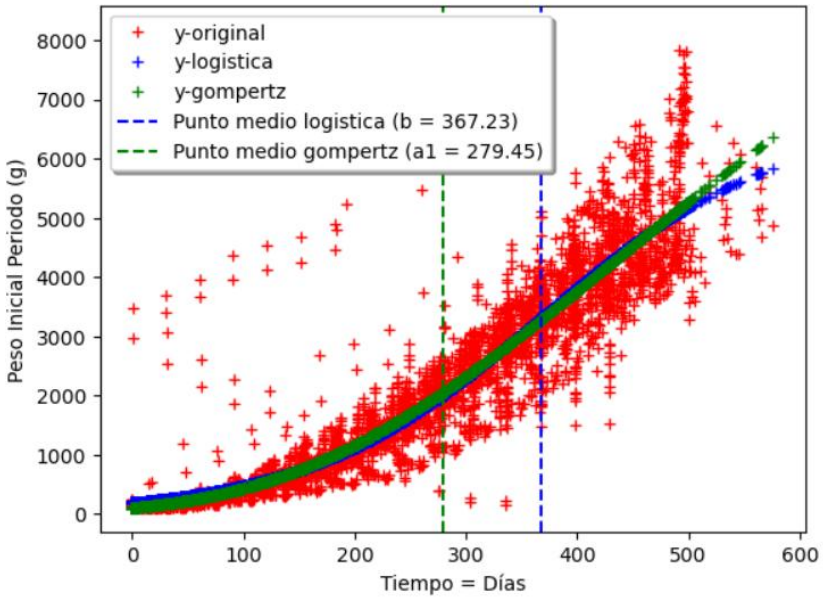

Figura 4: Regresión para el set de datos

    - Y las siguientes métricas:

Tabla 1: Métricas modelos

	Parámetro	y-logistic	y-gompertz
0	a	104.8992	279.4520
1	b	367.2298	4.6881
2	c	6635.8372	11553.1644
3	RMSE	531.1173	531.8616
4	R2 score	0.8933	0.8931

    - El objetivo es determinar el punto óptimo de días para el crecimiento de los peses (cuantificado en 'Peso Inicial Periodo (g)'). Para ello, se realizaron 2 modelos, un modelo logístico y de Gompertz. Las métricas de ambos modelos muestras similaridad y gráficamente también se confirma.
    - El r^2 para ambos modelos son similares, indicando alta correlación entre los datos reales y los datos predichos. El RMSE es similar entre los modelos, sin embargo, es alto, considerando que el rango de los valores reales para el peso va desde 97.1 (g) a 8183.24 (g). Si se observan la figura 4, se observan los datos reales (en rojo) y se ve una dispersión considerable, esto hace que RSME aumente.

    - Si bien al optimizar se busca el máximo o el mínimo, también se busca la mejor forma de realizar una actividad, y para este caso, el máximo de la función no es el punto óptimo. Si observamos el gráfico, el punto máximo se encuentra cercano a 600 días, pero ¿es rentable ese tiempo para una piscicultura? Pues no.
    - Si graficamos el parámetro "b" para la función logística, el cual representa al punto óptimo. En este caso es 367.23, por lo que se puede recomendar que el tiempo óptimo para el crecimiento del salmón es de 367 días.
    - En el caso del modelo de Gompertz, el valor es a=279.45, sim embargo, al observar la figura 4, es evidente que este punto no es correcto, el peso de los salmones es muy bajo en comparación de los datos a la derecha de esta recta.

- Recomendaciones:
    - Crecimiento óptimo de 367 días
    - La regresión logística se ajusta de mejor forma al set de datos
    - Nueva simulación eliminando extremos para acotar el rango y verificar si el ciclo (en días) aumenta y a su vez, si aumenta el peso de los salmones.
    - Nueva simulación eliminando atípicos, como los datos en donde el ciclo de días es alrededor de 100-200 pero el peso rondea los 4000-5000 (g).
    - Nueva simulación sin eliminar datos, sino que estandarizando.

Parte 3: Gestión de Bases de Datos y Herramientas de Análisis

Se crearon 4 scripts para la restructuración de la tabla, pueden utilizarse todos u omitir alguno:

- QuitarEspaciosPrimeraFila: Se observa que en la primera fila de la Hoja2 (en los títulos de columna), hay espacios en blanco al comienzo y al final y esto entorpece copiar los nombre, por lo que, al ejecutar el script, estos espacios se eliminan
- RellenarEspaciosEnBlancoColumnasABC: Se observa que las columnas A, B y C tienen celdas en blanco, esto debido a que los datos entregados vienen en ese formato. Con el script podemos completar estos espacios de acuerdo con el nombre del grupo
- CopiarDatos: este script solo funciona si los datos de la columna de destino coinciden con los nombres de las columnas de origen, por lo que en este caso se hizo un cambio manual para que coincidan
- CopiarDatosEspecificos: Para la columna "UTAs" de la "Hoja2" se probó este script que hace coincidir con la columna "Grados día desde el ingreso [°C]" de la hoja "Datos". Es similar al proceso manual

Recomendaciones:

- Procurar que el nombre de la hoja de destino y origen tengan los títulos de columna coincidentes, sin espacios en blanco al comienzo y final
- El script para rellenar las columnas con espacios en blanco se puede utilizar para todas las columnas en ese estado (no solo la A, B y C), solo debe hacerse la modificación necesaria