Proyecto Big Data - FULL DATA

Selección del Dataset asignado

Usamos el dataset Gym Members Exercise Tracking, que tiene 973 registros y 15 variables. Este dataset contiene datos de personas que entrenan en el gimnasio, con las siguientes variables: Age, Gender, Weight (kg), Height (m), Max_BPM, Avg_BPM, Resting_BPM, Session_Duration (hours), Calories_Burned, Workout_Type, Fat_Percentage, Water_Intake (liters), Workout_Frequency (days/week), Experience_Level y BMI.

Definición del problema

Queremos predecir el Experience_Level de un miembro del gimnasio (principiante, intermedio o avanzado) a partir de sus hábitos y características físicas.
Este análisis es importante porque en muchos gimnasios, al momento de la inscripción, no se conoce con precisión el nivel de experiencia real del nuevo miembro. Esto puede llevar a asignarle rutinas demasiado exigentes o, por el contrario, demasiado simples, generando frustración, desmotivación y un alto riesgo de abandono.
Contar con un modelo que prediga el nivel de experiencia permite a la empresa personalizar los planes de entrenamiento desde el primer día, ofrecer una mejor experiencia inicial y aumentar la retención de clientes. Además, esta información puede utilizarse para diseñar programas de fidelización más efectivos y optimizar la asignación de entrenadores y recursos dentro del gimnasio.

Análisis Exploratorio de Datos:

Las hipótesis que planteamos son las siguientes:

- **Hipótesis 1:** La variable Workout_Frequency (days/week) y la Session_Duration (hours) influyen directamente en el nivel de experiencia.
- **Hipótesis 2:** El Fat_Percentage y el BMI permiten diferenciar a principiantes de avanzados, ya que los principiantes suelen tener valores más altos y los avanzados más controlados.
- **Hipótesis 3:** El Workout_Type (Cardio, HIIT, Strength, Yoga) es un buen predictor del nivel de experiencia, porque los más avanzados tienden a llevar a cabo y combinar rutinas más intensas.

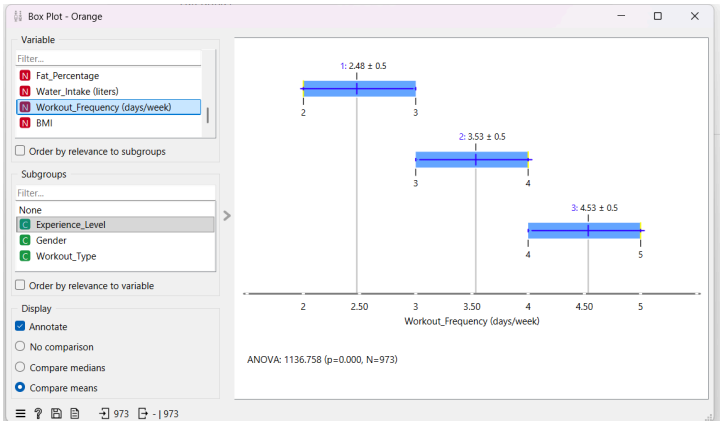
Antes de modelar, realizamos un análisis exploratorio para entender el comportamiento general del dataset y detectar patrones relevantes.
Primero observamos que la mayoría de los registros corresponden a principiantes e intermedios, mientras que los usuarios avanzados son minoría. Esto implica que el modelo podría tener más ejemplos de los primeros niveles, algo a tener en cuenta al evaluar su rendimiento.
También identificamos que las variables Workout_Frequency y Session_Duration son las que más diferencian a los grupos. Los avanzados entrenan más días por semana y durante más tiempo, mientras que los principiantes tienen rutinas más cortas y menos frecuentes. Este patrón confirma que la constancia y el tiempo de dedicación son factores claves en la progresión del usuario.

En cuanto a la composición corporal, los principiantes presentan un mayor porcentaje de grasa corporal, mientras que los avanzados muestran valores más bajos y estables. Sin embargo, el BMI no evidencia grandes diferencias, lo que sugiere que no es un indicador tan confiable del nivel de experiencia por sí solo.
Desde una mirada práctica, estos hallazgos ofrecen insights relevantes para una empresa de fitness:

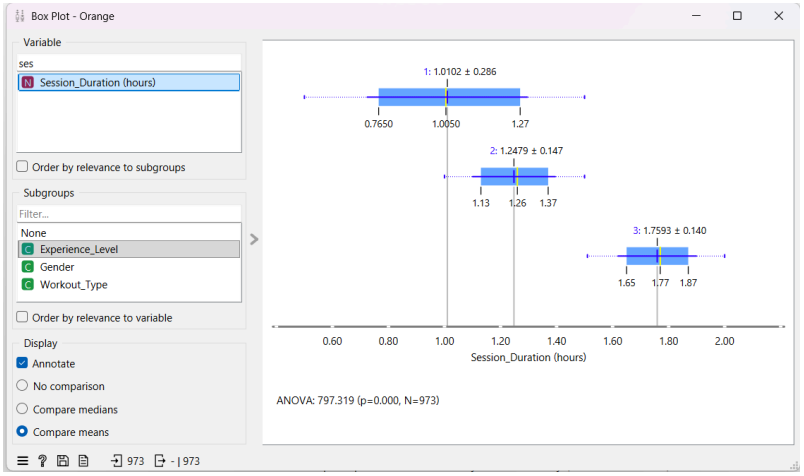
- Un nuevo miembro con baja frecuencia y sesiones cortas podría necesitar mayor acompañamiento para evitar el abandono.
- Observar el progreso en la duración o frecuencia de las rutinas puede servir como indicador temprano de fidelización.
- Las métricas corporales deben complementarse con hábitos de entrenamiento para definir mejor el nivel y las recomendaciones personalizadas.¡

En resumen, el EDA no solo ayudó a validar nuestras hipótesis, sino también a entender qué comportamientos predicen mejor la constancia y la evolución del usuario, lo que es clave para mejorar la retención en gimnasios o apps de entrenamiento.

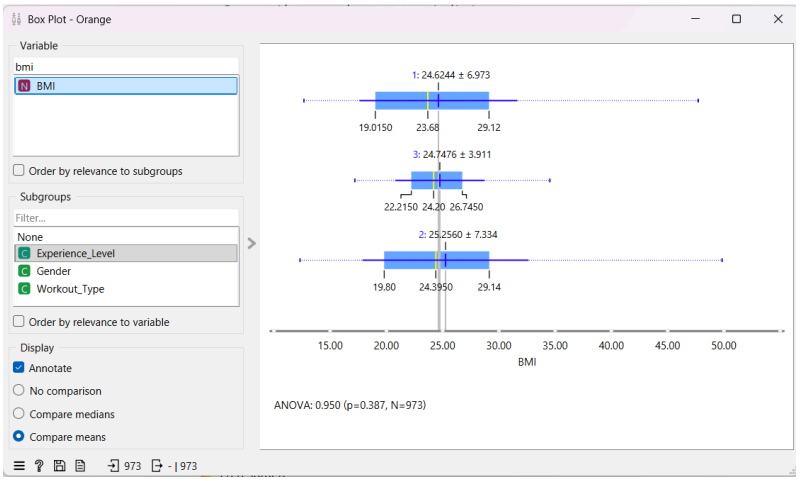
Boxplot - Workout_Frequency (days/week)



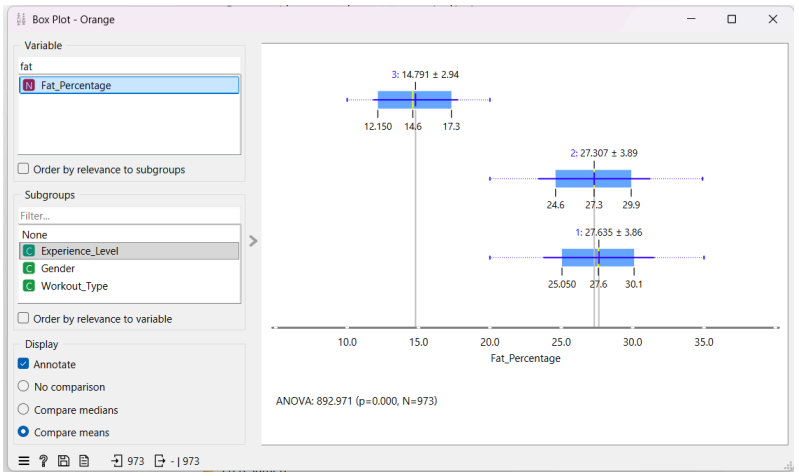
Boxplot - Session_Duration (hours)



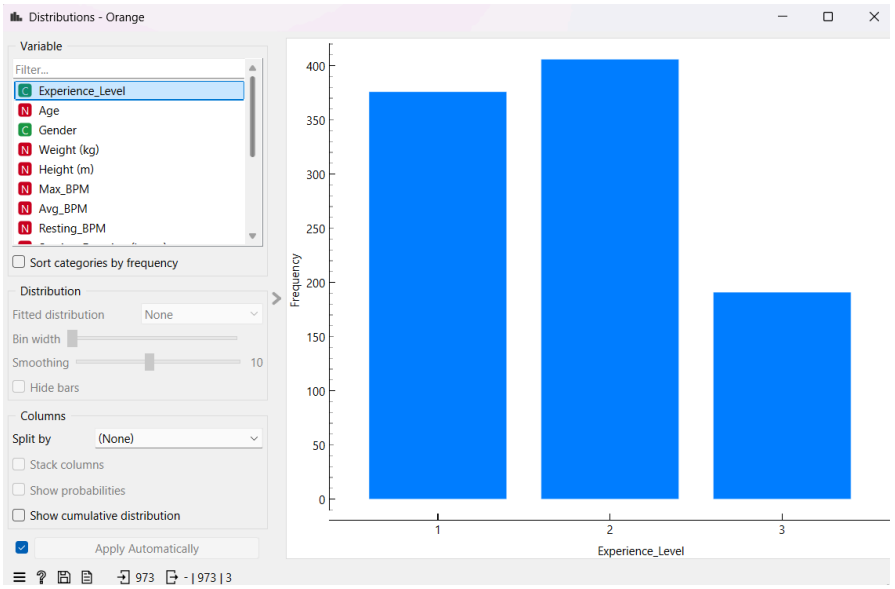
Boxplot - BMI



Boxplot - Fat_Percentage



Distribución de Experience_Level



Preprocesamiento y Selección de Variables

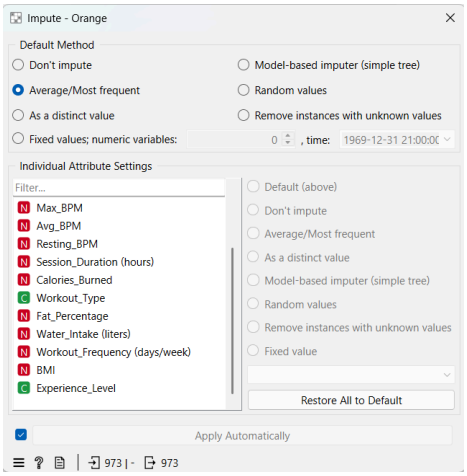
Antes de aplicar los modelos de clasificación, realizamos un preprocesamiento completo del dataset para garantizar la calidad de los datos y hacer más sencilla su interpretación para los algoritmos. En este proceso usamos distintos widgets de visualización como Data Table, Feature Statistics, Distributions y Box Plots, para ir verificando los cambios que fuimos haciendo en cada etapa y poder controlar la consistencia de los datos.

Select Columns

En el widget Select Columns establecimos la variable Experience_Level como Target.
El resto de las variables se mantuvieron como Features, y no nos pareció necesario descartar ninguna, ya que todas aportaban información que podría resultar relevante para el modelo.

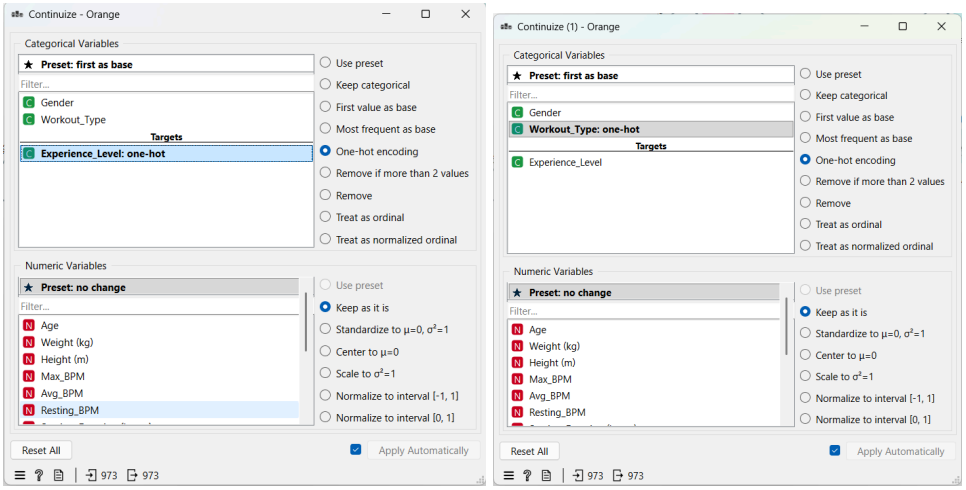
Impute

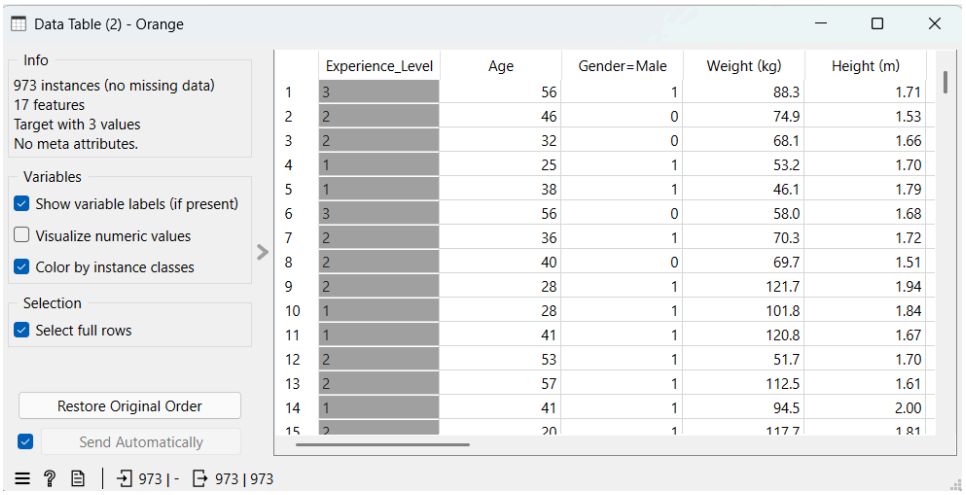
Aunque no registramos ningún valor nulo en el dataset, utilizamos por precaución el widget Impute con la configuración por defecto, para asegurar que no haya datos vacíos y mantener la integridad del conjunto.



Continuize

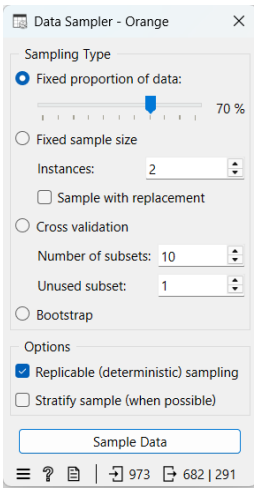
En este paso utilizamos dos widgets Continuize, cada uno con un propósito distinto dentro del flujo.
El primero se aplicó antes del Data Table (2), para preparar los datos del modelado. En este caso, realizamos un One-hot encoding sobre la variable Workout_Type, transformándola en columnas numéricas (una por cada tipo de entrenamiento) para que los modelos pudieran procesarla correctamente. La variable Gender se mantuvo categórica binaria, mientras que las variables numéricas conservaron la opción Keep as it is, sin normalización ni estandarización.
El segundo Continuize se utilizó solo para el análisis de correlaciones. Allí aplicamos un One-hot encoding a la variable Experience_Level, con el fin de convertirla en variables binarias (una por cada nivel de experiencia) y poder calcular las correlaciones entre esta y el resto de las variables.





Data Sampler

Usamos el widget de Data Sampler para dividir el dataset en dos subconjuntos: un 70% de los datos para entrenamiento y el otro 30% para prueba.



Ya en esta instancia, el flujo de trabajo ya se encontraba preparado para avanzar con la etapa de modelado.

Modelado y Evaluación del Modelo

Para este caso, el error más costoso sería clasificar como avanzado a un usuario que en realidad es principiante, ya que recibiría rutinas demasiado exigentes y podría abandonar el gimnasio. Por eso, la métrica más importante para la empresa es la Precision, que mide la capacidad del modelo para evitar falsos positivos.

El modelo Random Forest fue el que mejor desempeño tuvo en este sentido, con una Precision de 0.897, lo que indica que logra identificar correctamente la mayoría de los usuarios avanzados sin confundir a principiantes. Además, mantiene un buen equilibrio con un Recall de 0.890 y un AUC de 0.953, mostrando una alta capacidad para distinguir entre los diferentes niveles de experiencia.

Por eso, seleccionamos Random Forest como el modelo más confiable para predecir el nivel de experiencia de los usuarios y apoyar decisiones que reduzcan el riesgo de abandono en los nuevos miembros del gimnasio.

Model	AUC	CA	F1	Prec	Recall
Logistic Regression	0.959	0.863	0.862	0.863	0.863
Naive Bayes	0.949	0.845	0.844	0.845	0.845
Random Forest	0.953	0.890	0.888	0.897	0.890

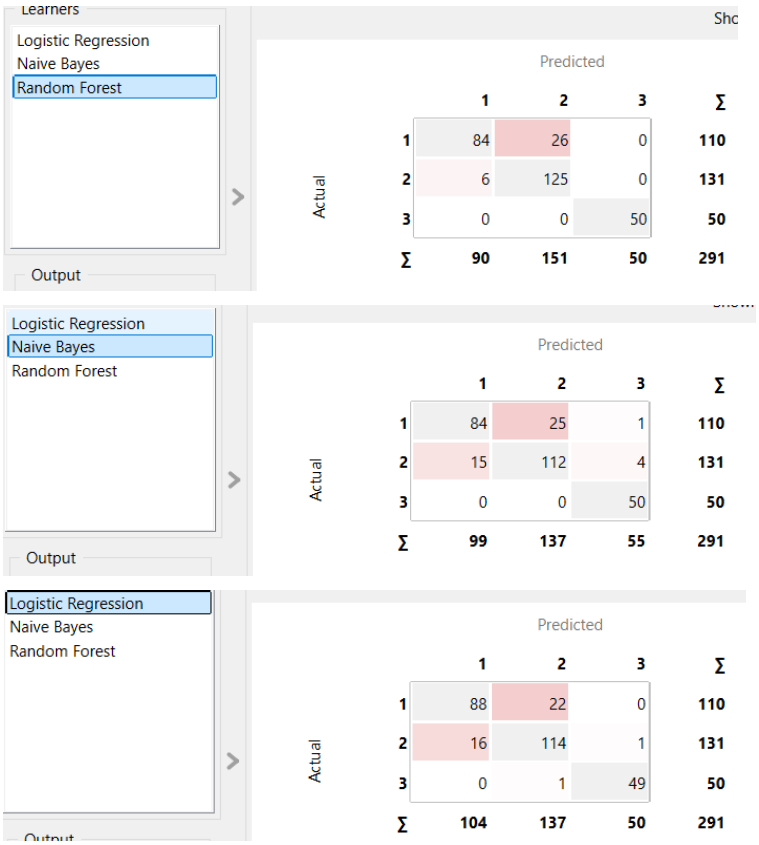
Matriz de Confusión

Al analizar las matrices de confusión, se observa que el modelo Random Forest logró la clasificación más equilibrada entre los distintos niveles de experiencia. La mayoría de los casos fueron correctamente identificados, y los errores se concentran principalmente entre clases contiguas (por ejemplo, confundir a un intermedio con un principiante), lo cual es esperable y menos grave.

Desde el punto de vista del negocio, el error más importante a evitar es clasificar como avanzado a alguien que no lo es, ya que podría recibir rutinas demasiado exigentes y abandonar el gimnasio. En ese sentido, Random Forest presenta el menor número de falsos positivos, lo que refuerza su alta Precision y su confiabilidad para predecir correctamente a los usuarios más experimentados sin sobreestimar su nivel.

En comparación, los modelos Naive Bayes y Logistic Regression muestran una mayor cantidad de errores cruzados entre niveles, lo que implica menor control sobre este tipo de equivocaciones.

Por eso, además de tener buenas métricas generales, Random Forest destaca por su distribución de errores más coherente y por minimizar los errores más costosos para la empresa.



Interpretación de las Predicciones

A partir de las correlaciones obtenidas con el modelo de Random Forest, se comprobó que las variables más influyentes para predecir el nivel de experiencia están relacionadas con la frecuencia, duración e intensidad del entrenamiento, tal como se anticipaba en las hipótesis iniciales.

Hipótesis 1:

La frecuencia de entrenamiento (Workout_Frequency) y la duración de las sesiones (Session_Duration) fueron las variables más relevantes. En los niveles iniciales se observan correlaciones negativas fuertes (alrededor de −0.8 y −0.7), lo que indica que quienes entrenan poco suelen ser principiantes. En cambio, a medida que aumenta la experiencia, estas correlaciones se vuelven positivas (en torno a +0.6 y +0.7), reflejando que los usuarios más avanzados entrenan con mayor constancia y por más tiempo.

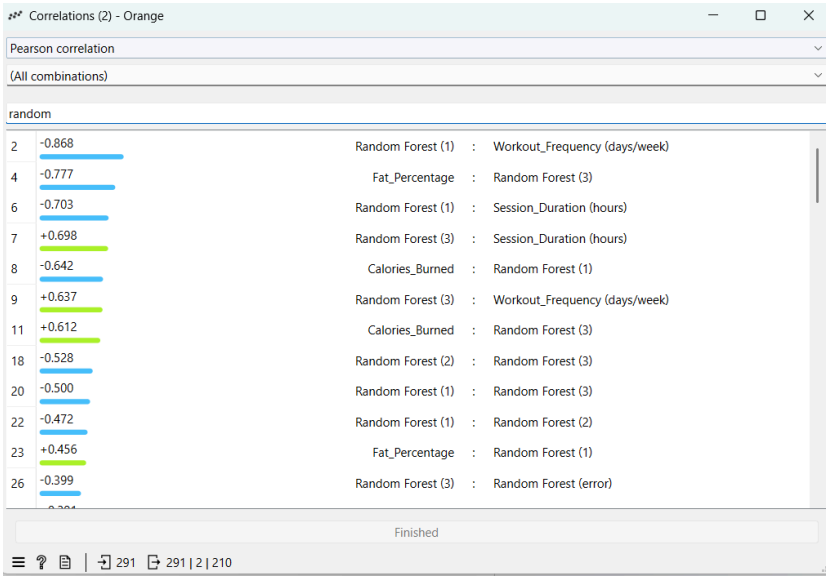
Hipótesis 2:

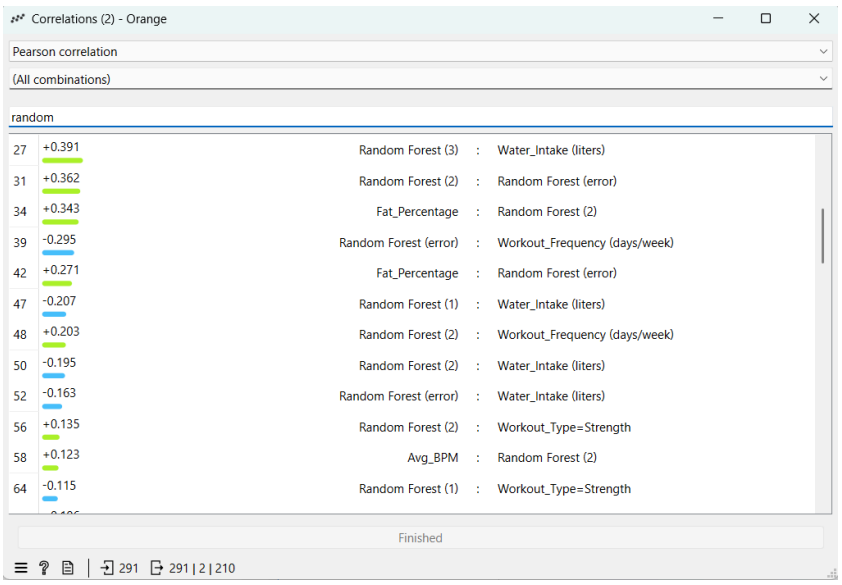
El porcentaje de grasa corporal (Fat_Percentage) también tuvo un papel destacado. Se registraron correlaciones negativas elevadas (cercanas a −0.75), lo que significa que a mayor nivel de experiencia, menor porcentaje de grasa corporal. Esto respalda la idea de que el progreso técnico suele ir acompañado de una mejora en la composición corporal.

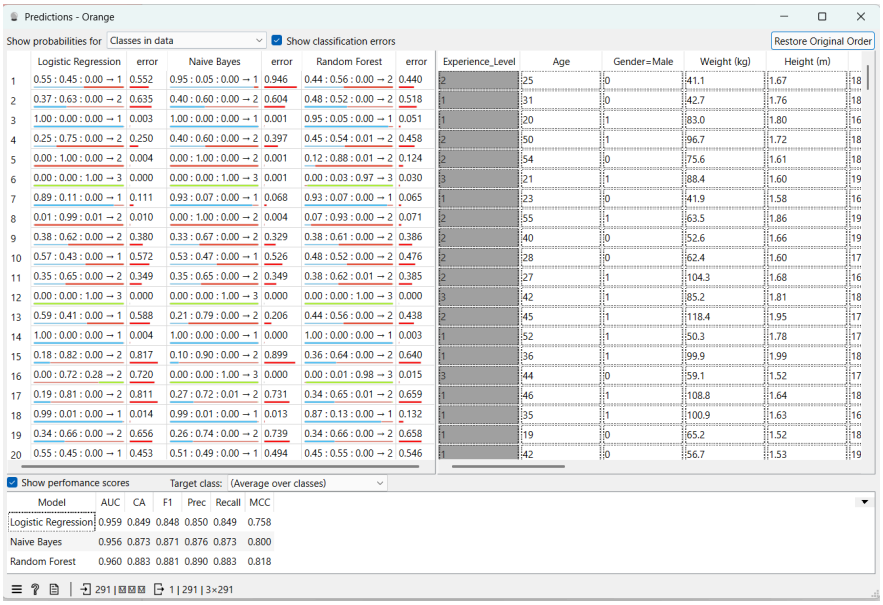
Hipótesis 3:

El tipo de entrenamiento (Workout_Type) no mostró una relación marcada con el nivel de experiencia. Las correlaciones fueron débiles (por ejemplo, alrededor de −0.16 en el caso de Strength), lo que sugiere que tanto principiantes como avanzados suelen realizar tipos de ejercicio similares, y que el progreso depende más de la frecuencia, la duración y la regularidad que del tipo de rutina.

En conjunto, el modelo confirma que la frecuencia de entrenamiento, la duración de las sesiones y el control corporal son los factores que mejor explican el avance en el nivel de experiencia, coherentes con las tendencias detectadas en el análisis exploratorio previo.







Conclusiones y Lecciones Aprendidas

A lo largo del proyecto aprendimos que, con un buen tratamiento de los datos, es posible predecir de forma bastante precisa el nivel de experiencia de cada usuario a partir de su comportamiento dentro del gimnasio. El modelo de Random Forest fue el que mejor captó estas diferencias, mostrando relaciones claras entre las variables de entrenamiento y el nivel de experiencia.

Los resultados confirmaron que la frecuencia de entrenamiento, la duración de las sesiones y el gasto calórico son los factores que más pesan al momento de determinar si una persona es principiante, intermedia o avanzada. También aprendimos que, aunque ciertas métricas (como la AUC o la accuracy) son útiles, en este tipo de problema importa más reducir los falsos negativos, es decir, evitar clasificar a un usuario avanzado como principiante o viceversa, porque eso impacta directamente en la calidad de las rutinas que se le asignan.

Desde una mirada más práctica, el trabajo muestra cómo este tipo de modelo podría ayudar a un gimnasio a ajustar las rutinas según el nivel real de cada persona. Por ejemplo, al identificar correctamente a los principiantes, se podrían evitar entrenamientos demasiado exigentes que generen frustración y abandono. En cambio, reconocer a los usuarios avanzados permitiría ofrecerles desafíos adecuados a su capacidad, mejorando su experiencia y motivación.