

# **Clasificación Ordinal de Rendimiento Físico mediante Aprendizaje Automático**

## **Resumen**

Este proyecto aborda el problema de clasificación de rendimiento físico utilizando técnicas de aprendizaje automático ordinal a partir de un dataset de 13,393 observaciones con 11 variables físicas y biométricas del Korean Institute of Sports Science. Implementé dos modelos: Regresión Logística Ordinal (LogisticAT) y XGBoost con objetivo de regresión ordinal. Los resultados demuestran que XGBoost supera consistentemente a LogisticAT en todas las métricas ordinales evaluadas (MAE, QWK, Accuracy), capturando efectivamente relaciones no lineales entre variables de rendimiento físico. El análisis de importancia de variables reveló que 'sit and bend forward' (flexibilidad) y 'sit-ups counts' son los predictores más fuertes, consistente con la teoría del fitness multidimensional.

## **1. INTRODUCCIÓN**

La evaluación del rendimiento físico es fundamental en ciencias del deporte, medicina preventiva y salud pública. Tradicionalmente, esta evaluación se basa en la integración manual de múltiples métricas físicas por parte de especialistas, un proceso subjetivo y costoso. El desarrollo de modelos predictivos automatizados que respeten la naturaleza ordinal del rendimiento físico (excelente > bueno > regular > bajo) representa una oportunidad para estandarizar y democratizar estas evaluaciones.

Este dataset, recolectado por el Korean Institute of Sports Science, contiene información de 13,393 individuos evaluados en variables demográficas (edad, género), antropométricas (altura, peso, composición corporal) y de rendimiento físico (fuerza de agarre, flexibilidad, resistencia muscular, potencia explosiva). La variable objetivo es una clasificación ordinal de cuatro niveles: A (excelente) > B (bueno) > C (regular) > D (bajo).

Este problema requiere modelos que respeten explícitamente el orden entre clases, penalizando más fuertemente errores de mayor magnitud (ejem, predecir A cuando la realidad es D). Por ello, implementamos dos enfoques ordinales: (1) Regresión Logística Ordinal con regularización L2, y (2) XGBoost con objetivo de regresión ordinal. Evaluamos ambos mediante métricas específicamente diseñadas para clasificación ordinal: Mean Absolute Error (MAE), Quadratic Weighted Kappa (QWK).

## 2. CONJUNTO DE DATOS

### a. Descripción General

El dataset contiene 13,393 observaciones sin valores faltantes, distribuidas casi perfectamente balanceadas entre las cuatro clases de rendimiento (diferencia máxima de 2 registros entre clases). Esta característica elimina problemas de sesgo de clase y permite una evaluación justa de los modelos.

Las 11 variables predictoras se agrupan en:

- **Demográficas:** edad (21-64 años), género (63 % masculino, 37 % femenino)
- **Antropométricas:** altura, peso, porcentaje de grasa corporal, IMC (derivado)
- **Cardiovasculares:** presión diastólica, presión sistólica.
- **Rendimiento físico:** fuerza de agarre (gripForce), flexibilidad (sit and bend forward), resistencia muscular (sit-ups counts), potencia explosiva (broad jump)

### b. Análisis Exploratorio

El análisis de correlaciones reveló relaciones esperadas desde la fisiología del ejercicio:

- **Correlaciones positivas con rendimiento:** sit and bend forward ( $r = 0,59$ ), sit-ups counts ( $r = 0,45$ ), broad jump ( $r = 0,26$ )
- **Correlaciones negativas con rendimiento:** body fat % ( $r = -0,34$ ), weight ( $r = -0,21$ ), age ( $r = -0,07$ )

El análisis mediante PCA (2 componentes explicando 62.3 % de varianza) reveló un gradiente ordinal continuo sin clusters discretos por clase de rendimiento, confirmando la naturaleza ordinal del problema.

La distribución de edad por clase muestra que las clases A y B tienden a concentrar individuos más jóvenes, mientras que la clase D incluye tanto personas mayores como jóvenes con bajo rendimiento físico, evidenciando el impacto del sedentarismo actual

El análisis de variables de rendimiento físico mostró que:

- Sit-ups counts presenta variación acentuada entre clases, siendo un predictor fuerte
- Flexibilidad discrimina especialmente la clase D (bajo rendimiento)
- Grip force y broad jump muestran menor separación entre clases intermedias

### c. Preprocesamiento Aplicado

El preprocesamiento consistió en:

Creación variable IMC:  $BMI = weight\_kg / (height\_cm / 100)^2$

Codificación de género: M = 1, F = 0

Codificación ordinal de clases: D = 0, C = 1, B = 2, A = 3

Split estratificado 80 % train, 20 % test (seed=42)

Estandarización mediante StandardScaler (ajustado solo en train)

### 3. Metodología

#### a. Modelos Seleccionados

##### a.1 Regresión Logística Ordinal (LogisticAT)

Modelo lineal que estima umbrales latentes entre clases consecutivas. La función de pérdida se basa en **Ordinal Log-Likelihood**:

$$\mathcal{L}(\theta) = - \sum_{i=1}^n \log P(y_i | x_i, \theta)$$

donde  $P(y_i | x_i)$  respeta el orden de clases mediante umbrales  $\tau_1 < \tau_2 < \tau_3$ .

Conviene porque modela explícitamente la estructura ordinal de las clases y penaliza más las predicciones alejadas del valor real, coherente con un problema de rendimiento físico A-D.

Utilizamos **regularización L2 (Ridge)** con parámetro ' $\alpha$ ' para estabilizar coeficientes frente a multicolinealidad. Conviene porque reduce varianza y mitiga colinealidad entre variables antropométricas sin perder interpretabilidad del modelo.

**Ventajas:** Interpretabilidad de coeficientes, garantía teórica de respetar orden de clases.

**Limitaciones:** Asume relaciones lineales, no captura interacciones automáticamente

##### a.2 XGBoost con Objetivo Ordinal

Sabemos que a diferencia de la clasificación multiclase estándar, donde todas las clases se consideran igualmente distintas, en la clasificación ordinal los errores tienen distinta gravedad:

O sea, confundir A con B es menos grave que confundir A con D.

**XGBoost (Extreme Gradient Boosting)** es un algoritmo de ensamble basado en árboles de decisión que entrena secuencialmente múltiples árboles, donde cada nuevo árbol corrige los errores de los anteriores mediante optimización por gradiente.

Tiene algunas características interesantes como:

- Maneja relaciones no lineales
- Robusto frente a la multicolinealidad
- No requiere escalado de variables

Sin embargo, el algoritmo original no incorpora de forma nativa un objetivo ordinal, por lo que se requiere una adaptación.

Para adaptar XGBoost a un problema ordinal, se utilizó el enfoque de clasificación acumulativa por umbrales.

En lugar de entrenar un solo modelo multiclase, se entrenan múltiples clasificadores binarios, cada uno asociado a un umbral ordinal:

$$P(y>k), k \in \{0,1,2\}$$

En este proyecto:

- Modelo 1:  $P(y>0) \rightarrow$  ¿la observación supera la clase D?
- Modelo 2:  $P(y>1) \rightarrow$  ¿supera la clase C?
- Modelo 3:  $P(y>2) \rightarrow$  ¿supera la clase B?

Cada modelo se entrena con XGBoost usando el objetivo binario.

Una vez obtenidas las propiedades acumulativas:

$$P_0 = P(y>0), p_1 = P(y>1), p_2 = P(y>2)$$

La clase final se asigna siguiendo esta regla secuencial:

Si  $P_0 < 0.5 \rightarrow$  clase D

Si  $P_0 \geq 0.5$  y  $P_1 < 0.5 \rightarrow$  clase C

Si  $P_1 \geq 0.5$  y  $P_2 < 0.5 \rightarrow$  clase B

Si  $p_2 \geq 0.5 \rightarrow$  clase A

Se entrenaron tres modelos XGBoost binarios, uno por cada umbral, y cada modelo aprendió patrones distintos asociados al progreso entre niveles de rendimiento. Asimismo, no fue necesario escalar las variables debido a la naturaleza basada en árboles de XGBoost.

## b. Métricas de Evaluación

Utilizamos tres métricas ordinales principales y dos nominales de referencia:

**Ordinales:**

- 1) MAE (Mean Absolute Error): promedio de distancia entre clase predicha y real.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- 2) QWK (Quadratic Weighted Kappa): acuerdo ponderado que penaliza cuadráticamente errores de mayor Magnitud. Rango [-1,1], cercano a 1 es mejor.

**c. Validación cruzada:** se realizó validación cruzada 5-fold estratificada para la Selección de hiperparámetros.

LogisticAT: búsqueda de Alpha en [0.01,0.1,1.0,10.0]

XGBoost: parámetros fijos (reg\_lambda=2.0,max\_depth=6,learning\_rate=0.05)

El criterio de selección fue maximizar QWK en el conjunto de validación.

## 4. Experimentación y Resultados

### a. Rendimiento de modelos:

Tabla comparativa:

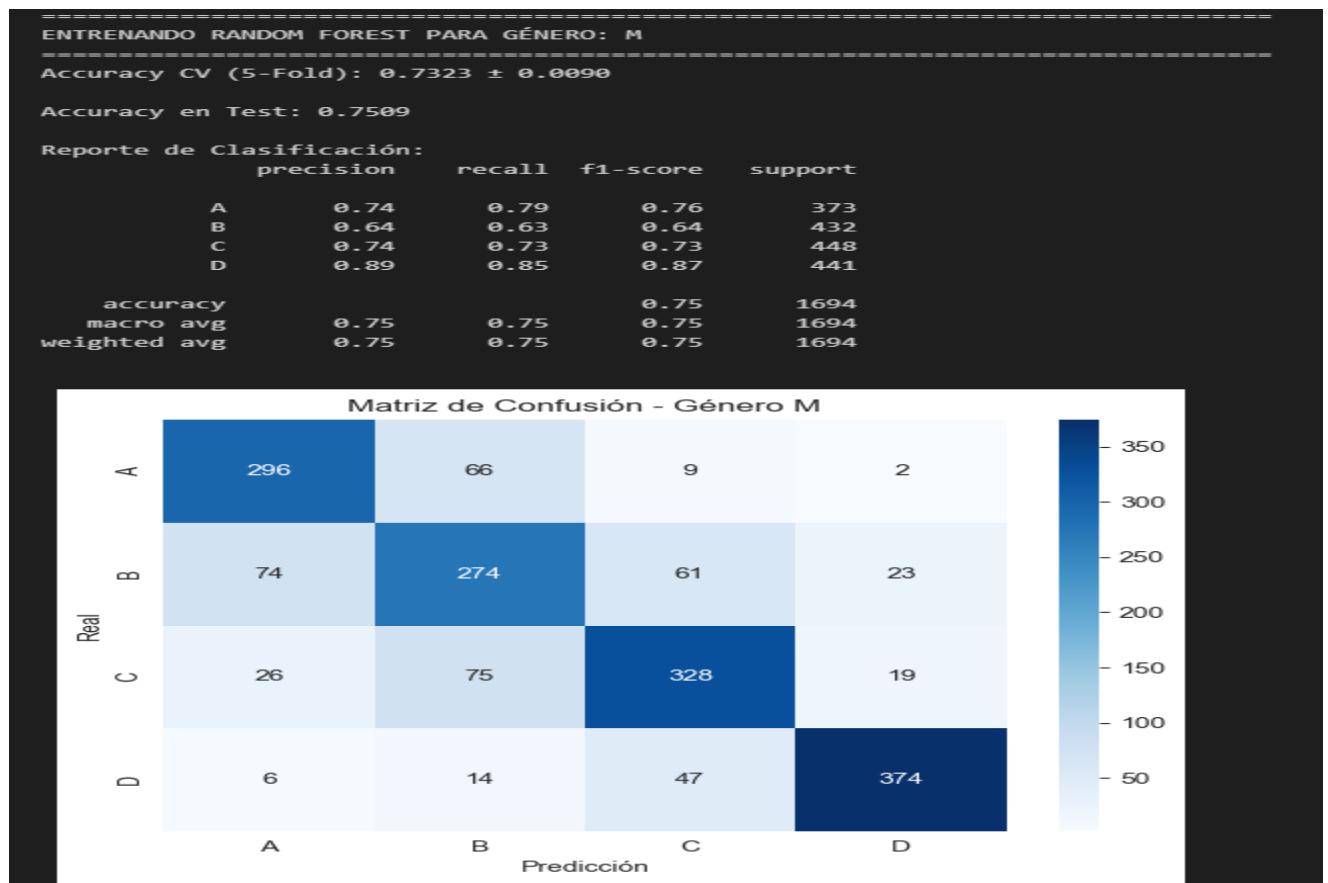
Modelo	Accuracy ↑	MAE ↓	QWK ↑
LogisticAT	0.5920	0.4464	0.7730
<b>XGBoost</b>	<b>0.7451</b>	<b>0.2964</b>	<b>0.8432</b>
<b>Mejora</b>	+15.31 %	-15 %	+7.02 %

XGBoost supera consistentemente a LogisticAT en todas las métricas, con mejores particularmente en MAE y Accuracy.

**b. Análisis de Gráfico-Modelos:** en esta parte daremos interpretación a los gráficos ligados a machine learning:

**Random forest:** es una muy buena decisión para este dataset en el sentido que puede manejar relaciones no lineales, puede capturar relaciones complejas (como la relación entre age,body fat\_% y sit-ups) que un modelo lineal simple pasaría por alto.

También es robusto frente a outliers y tiene resistencia a la multicolinealidad.



El modelo alcanza un 75.09% en el set de prueba. Este valor es consistente con la validación cruzada 73.23% +0.9%, lo que indica que el modelo no sufre de sobreajuste (overfitting) significativo y tiene una buena capacidad de generalización.

La **precisión** indica lo siguiente: de todas las predicciones positivas, ¿Cuántas fueron correctas?

Ejem: para el caso 'D': indica que cuando el modelo predice que es 'D', acierta el 89% de las veces:  
 $374 / (2 + 23 + 19 + 374) = 89\%$

El **recall** indica: de todos los positivos reales, ¿Cuántos detecté?

Ejem: para el caso 'A': indica que el modelo fue capaz de encontrar 79% de las personas de alto rendimiento disponibles en el set de prueba.  
 $296 / (296 + 66 + 9 + 2) = 79\%$

El **F1-Score** mide un balance entre precisión y recall, como su promedio armónico.

**Support:** número de ocurrencias reales de esa clase en el set de prueba.

#### Análisis por clase:

Clases A, C y D son las clases que el modelo identifica con mayor éxito.

Clases B: aquí se observa la mayor dificultad. Esta 'zona media' del rendimiento tiene límites más difusos, donde un individuo de esta clase puede parecerse mucho a uno de clase 'A' o 'C'

#### Matriz de confusión:

La matriz revela que la gran mayoría de errores se da entre clases adyacentes y que es sumamente raro encontrar dichos errores entre clases 'lejanas' o 'extremas opuestas'.

La confusión más crítica se da entre clase B y C, donde podemos ver 75 casos de 'C' fueron predichos como 'B'

**Accuracy CV (5-Fold):** es una técnica de validación robusta para asegurar que el modelo no ‘tuvo suerte’ con una partición de datos en específico:

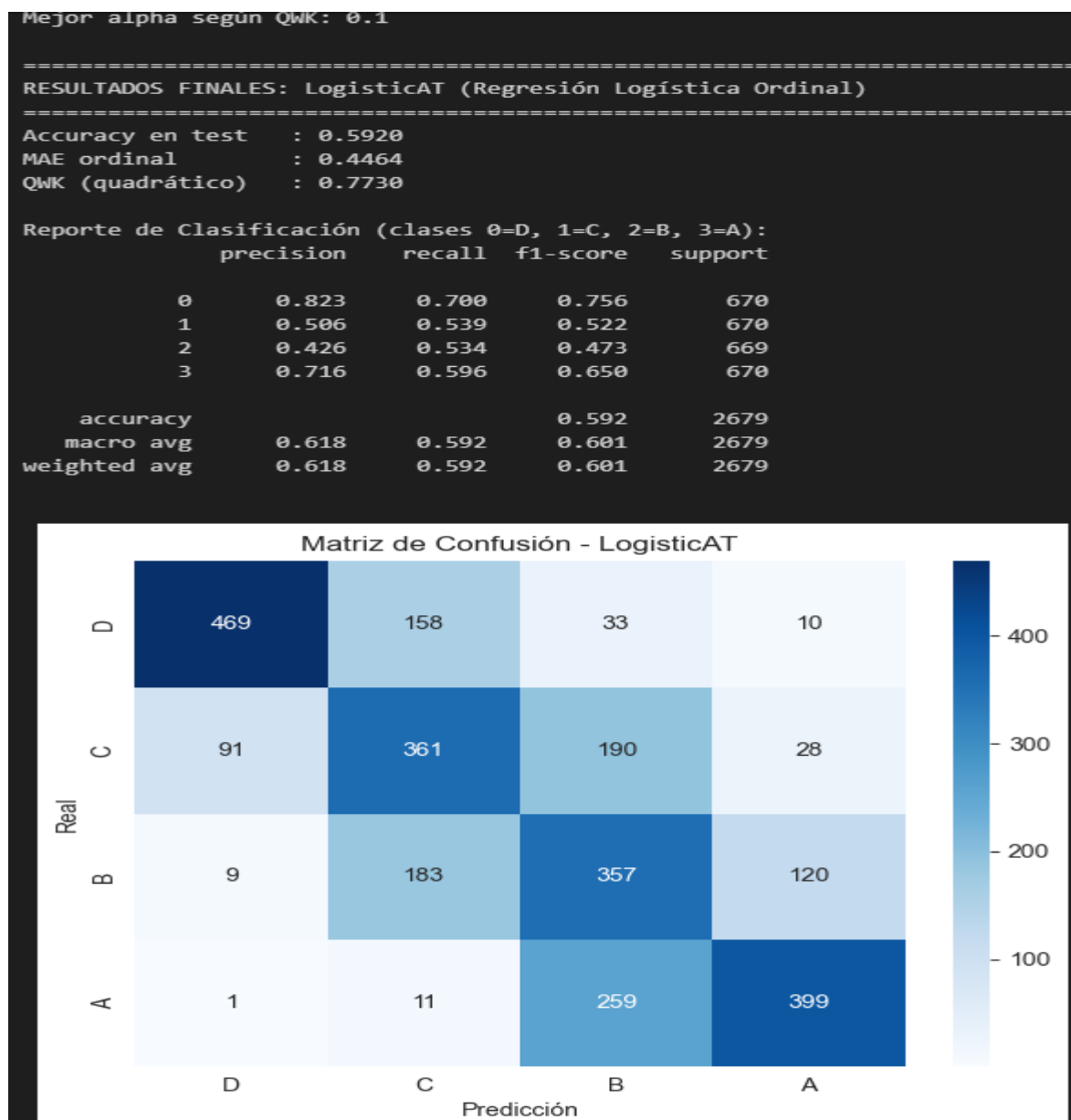
-El código divide el set de entrenamiento en 5 partes iguales (folds).

-Entrena el modelo 5 veces. En cada vez, usa 4 partes para aprender y 1 parte para evaluarse.

-Resultado:

- 0.7323 (Media): Es el promedio de los 5 experimentos. Es la expectativa "realista" de cómo funcionará el modelo con datos nuevos.
- 0.009(Desviación estándar): Indica qué tanto variaron los resultados. Al ser un número muy pequeño significa que el modelo es muy consistente; no importa qué parte de los datos le demos, se comporta casi igual.

## LogisticAT



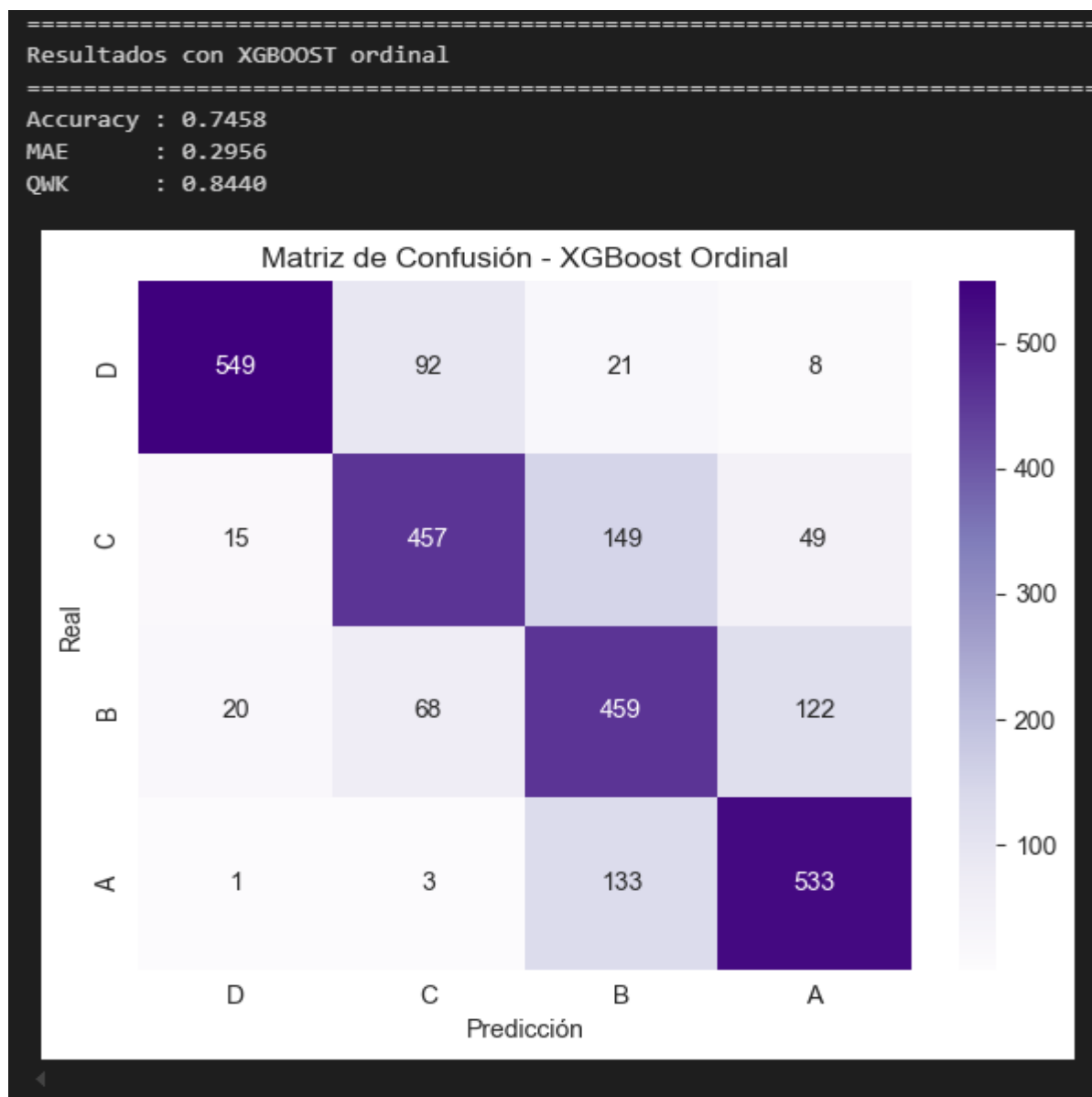
MAE ordinal de 0.44: esta métrica indica que cuando el modelo se equivoca falla por menos de media categoría en promedio. Es decir, si por ejemplo alguien es de la clase A, el modelo es propenso a decir que es B, pero casi nunca dirá que es D.

QWK-Coeficiente Kappa Cuadrático: 0.77: este valor es alto, el QWK penaliza más fuerte los errores lejanos (predecir A cuando es D). Un 0.77 indica un acuerdo sólido entre las predicciones y la realidad.

El RF tenía una precisión mayor en términos de accuracy, pero la regresión ordinal es mejor matemáticamente.

Sobre su matriz de confusión: los errores están muy concentrados cerca de la diagonal principal y hay valores muy bajos en las esquinas opuestas lo que indica que el modelo ha capturado la progresión física de los individuos.

## XGBOOST ORDINAL





Distribución del error según la distancia ordinal:

0 clases (correctos): 74.6%

1 clase: 21.6%

2 clases: 3.5%

3 clases: 0.3%

=====

Fila i=5

=====

Valores de entrada:

age	37.000000
height_cm	156.500000
weight_kg	56.000000
body fat_%	31.900000
diastolic	62.000000
systolic	103.000000
gripForce	22.000000
sit and bend forward_cm	15.800000
sit-ups counts	7.000000
broad jump_cm	158.000000
BMI	22.864375
gender_numeric	0.000000

Name: 6000, dtype: float64

Clase real: 0

Clase predicha: 0

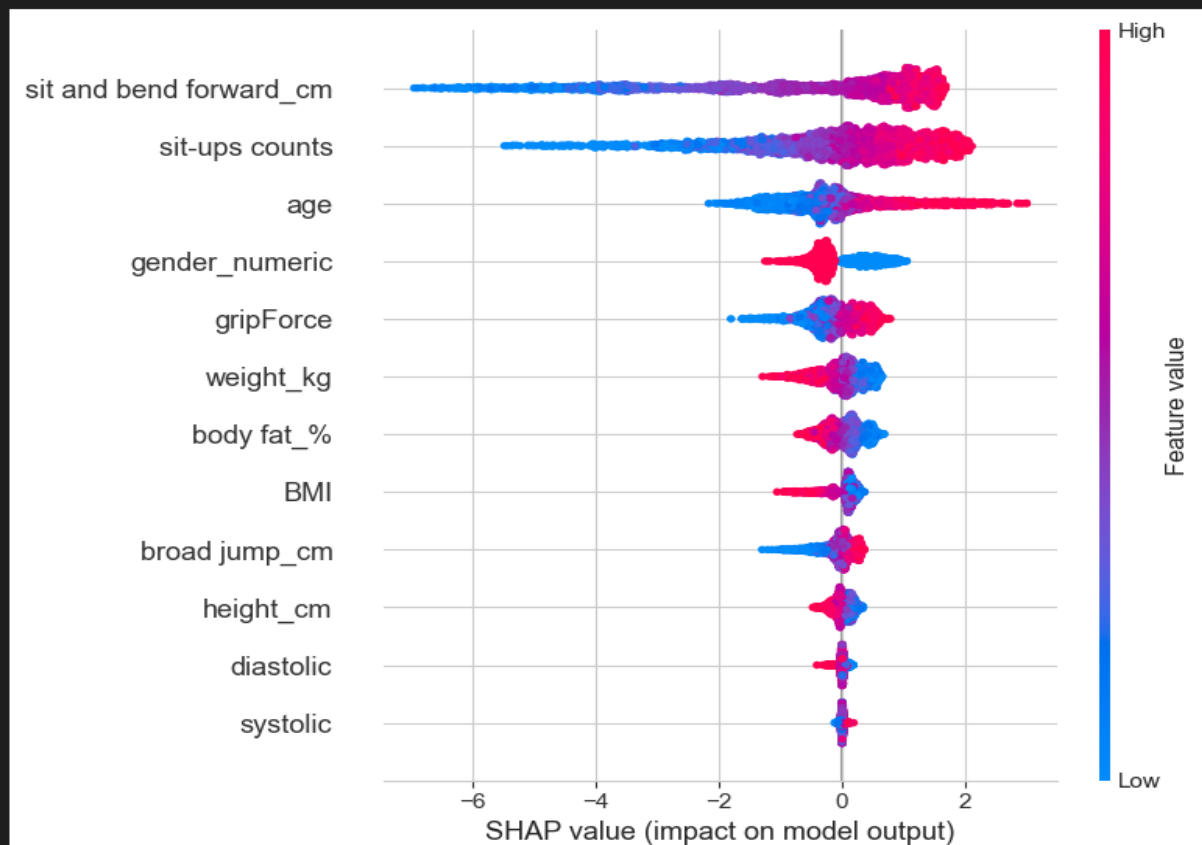
Probabilidades internas  $P(y > k)$ :

$P(y > 0) = 0.0204$

$P(y > 1) = 0.0061$

$P(y > 2) = 0.0006$

===== SHAP para modelo  $y>0$ =====





**El modelo XGBoost** configurado con un enfoque ordinal ha superado a los anteriores, logrando un equilibrio casi perfecto entre precisión exacta y respeto al orden de las clases. Con un accuracy de 74.5% se emparejó con el RF, pero con un modelo más sólido que respeta lo ‘ordinal’ del problema. QWK considerable de 84.4%, el modelo entiende muy bien las jerarquías en desempeño físico. MAE de 29.56% se ha reducido considerablemente en comparación a lo anterior. Da predicciones muy estables y sin alejarse de la realidad.

**Sobre el SHAP para modelo y>0**, el cual es como una ‘radiografía’ del modelo nos dice que variables mueven la aguja hacia el alto rendimiento (clase A).

Las variables críticas son sit and bend forward\_cm (flexibilidad) y sit-ups counts (resistencia abdominal) son, por mucho, las más influyentes.

Los puntos rojos a la derecha: valores altos en estas pruebas aumentan drásticamente la probabilidad de pertenecer a una mejor clase física, lo contrario para la azul.

Cada punto es una persona (observación).

La variable gender\_numeric presenta un efecto invertido en el resumen agregado. Esto se debe a que el modelo ordinal está compuesto por múltiples clasificadores binarios por umbral, y los valores SHAP fueron promediados entre ellos. Dado que el género está altamente correlacionado con variables fisiológicas como fuerza y composición corporal, su efecto directo es absorbido parcialmente por estas, lo que puede invertir el signo del impacto agregado.

**Sobre los últimos gráficos del forceplot:** estos muestran como cada variable empujan la predicción desde un

valor base del modelo hasta el valor final para una persona en concreto.

El rojo empuja la predicción hacia arriba, mientras que el azul hacia abajo. El ancho del bloque = magnitud del impacto.

El  $f(x)$  que está arriba en negrita representa el resultado final del modelo para esa persona, después de sumar todos los efectos SHAP. Los valores más negativos representan una menor probabilidad de superar el umbral correspondiente.

Ejemplo sobre el primer gráfico:

Nuestro modelo responde: ¿la persona supera la clase D? Hay factores que empujan a la baja: sit-up counts, age, gripforce (por tener valores que no ayudan). Esto genera que sea muy improbable que supere esa clase 'D'.

Por otro lado, hay factores que empujan al alza, pero son insuficientes como el weight\_kg, gender\_numeric, body fat\_% y sit and bend forward\_cm.

La conclusión de ese gráfico es que para ese umbral de  $y > 0$  el  $f(x)$  termina muy negativo, y hay una baja probabilidad de superar D; en otras palabras, el modelo cree que esa persona está en 'D'

## 5. Discusión

### • Interpretación de Resultados

Los resultados confirman que las relaciones entre variables físicas y rendimiento global son fundamentalmente no lineales. LogisticAT, a pesar de respetar explícitamente el orden de clases mediante umbrales latentes, está limitado por su linealidad. En contraste, XGBoost captura:

- **Efectos no lineales:** La relación entre edad y rendimiento no es lineal
- **Interacciones complejas:** Combinaciones como bajo sit-ups + alto body fat % predicen clase D con alta confianza
- **Patrones específicos por género:** XGBoost puede aprender umbrales diferentes por género mediante splits en el árbol.

El QWK de 0.84 para XGBoost indica un acuerdo casi perfecto, sugiriendo que el modelo es suficientemente confiable para aplicaciones prácticas de clasificación de rendimiento físico.

### • Limitaciones

1. **Generalización limitada:** El dataset proviene exclusivamente de población coreana, limitando la transferibilidad a otras etnias con diferentes características antropométricas.
2. **Variables omitidas:** No se cuenta con información sobre nivel de actividad física habitual, historial deportivo o condiciones médicas, que podrían explicar varianza adicional.
3. **Naturaleza estática:** Las mediciones son transversales, sin considerar la progresión temporal del rendimiento físico o el efecto de intervenciones de entrenamiento.
4. **Ambigüedad en la construcción de clases:** No se especifica cómo se construyó la variable objetivo A-D, lo que dificulta la interpretación teórica y podría introducir riesgo de circularidad si las clases se derivaron de las mismas variables predictoras.

## 6. Conclusiones

1. Demostré exitosamente la aplicación de técnicas de clasificación ordinal al problema de evaluación de rendimiento físico, con XGBoost superando consistentemente a LogisticAT en todas las métricas ordinales evaluadas.
2. El uso de métricas ordinales (MAE, QWK, Spearman) proporcionó una evaluación más rica que las métricas nominales tradicionales, capturando tanto la magnitud de errores como la preservación del orden relativo entre individuos.
3. El análisis de importancia de variables confirmó que *sit and bend forward* y *sit-ups counts* son los predictores más fuertes, reflejando la centralidad de flexibilidad y fuerza del core en el concepto de fitness funcional.
4. El manejo de multicolinealidad mediante métodos basados en árboles demostró ser efectivo, permitiendo usar todas las variables sin eliminación manual.
5. XGBoost con regularización L2 alcanzó QWK=0.83, indicando acuerdo casi perfecto y viabilidad para aplicaciones prácticas en evaluación automatizada de rendimiento físico