

Informe del Modelo de Clasificación de Condición Física

1. Introducción

Este informe detalla el desarrollo y evaluación de un modelo de clasificación para predecir la condición física de personas según datos relacionados con edad, género y desempeño en pruebas físicas. El dataset original contiene 13,393 registros con 12 columnas.

2. Descripción del Dataset

El conjunto de datos incluye las siguientes características:

- Edad (20 a 64 años)
- Género (F/M)
- Altura (cm)
- Peso (kg)
- Porcentaje de grasa corporal
- Presión arterial diastólica y sistólica
- Fuerza de agarre
- Flexibilidad (sit and bend forward)
- Abdominales (sit-ups)
- Salto horizontal (broad jump)
- Clase objetivo: A, B, C, D (A = mejor condición física)

3. Selección y Optimización del Modelo

Se evaluaron diferentes algoritmos de clasificación como Random Forest y XGBoost. Sin embargo, Gradient Boosting ofreció el mejor desempeño global. Para la búsqueda de hiperparámetros se utilizó RandomizedSearchCV, con validación cruzada estratificada (StratifiedKFold) y una métrica principal de precisión (accuracy). No se usó Optuna en la optimización, se probó pero nos daba peor precisión.

4. Hiperparámetros Óptimos del Modelo

Los mejores parámetros encontrados fueron:

- n_estimators: 300
- learning_rate: 0.03
- max_depth: 3
- min_samples_leaf: 5

- min_samples_split: 15
- subsample: 0.75
- max_features: log2

5. Resultados del Modelo

Comparación de las métricas entre entrenamiento y test:

Métrica	Entrenamiento	Test
Accuracy	0.7377	0.7048
Precision (weighted)	0.7437	0.7110
Recall (weighted)	0.7377	0.7048
F1 Score (weighted)	0.7370	0.7046
ROC AUC (multi-class)	0.9176	0.8996

-A pesar de que las métricas generales como la accuracy y el F1 score rondan valores en torno al 70%, es importante tener en cuenta que se trata de un problema de clasificación multiclase con cuatro posibles categorías (A, B, C, D), lo que naturalmente reduce la probabilidad base de acierto por azar (25%). En este contexto, obtener una exactitud superior al 70% representa un rendimiento significativamente superior al azar y evidencia que el modelo ha aprendido patrones relevantes del conjunto de datos.

Además, el modelo presenta un equilibrio adecuado entre precisión, recall y F1 score, lo que sugiere que no está sesgado hacia ninguna clase en particular. El ROC AUC de 0.8996 en test refuerza aún más esta afirmación, indicando una buena capacidad de discriminación entre las clases.

Por tanto, aunque las métricas no sean excepcionalmente altas, son sólidas y consistentes para un problema de esta naturaleza y tamaño, y respaldan la elección de Gradient Boosting como mejor modelo frente a alternativas como Random Forest o XGBoost.

6. Análisis de Overfitting

La diferencia de accuracy entre entrenamiento y test es de 0.0329, lo que indica un ligero overfitting del 4.46%. Este nivel es considerado aceptable dada la estabilidad de las métricas y la complejidad del problema.

La pequeña diferencia de precisión entre entrenamiento y test sugiere una generalización correcta, sin síntomas graves de overfitting.