Primera entrega - Data Science II (CoderHouse)

Autor: Sebastián Díaz
Curso: Data Science II - CoderHouse
Fecha: Octubre 2025
Curso: 74260

Abstracto y motivación

Este proyecto continúa el análisis iniciado en Data Science I, en el que se exploró el rendimiento de atletas en competencias HYROX. El objetivo de esta entrega es aplicar técnicas de Machine Learning y enriquecimiento de datos para mejorar la capacidad predictiva de los modelos, utilizando métricas como MAE y R^2 para evaluar el rendimiento de los atletas. Dirigido a entrenadores, gestores de desempeño y analistas deportivos interesados en optimizar estrategias de entrenamiento.

Resumen de metadata

Dataset trabajado: hyrox_enriquecido_opt.csv, que es producto del dataset original Hyrox_procesado_SDP proveniente de DS I.
Filas válidas para modelado: 5000.
Columnas totales: 38.
Variables principales: tiempos por estación, parciales de carrera, divisiones, género, nacionalidad y clusters de rendimiento.

Preguntas e hipótesis

Entre otras, las preguntas que motivaron este análisis fueron las siguientes:

1. Qué factores tienen mayor impacto en el tiempo total de competencia?

2. Existen patrones comunes entre atletas agrupados por clusters de rendimiento?

3. El enriquecimiento con nuevas variables mejora el rendimiento predictivo del modelo base?

4. Cuáles son las estaciones más determinantes en el resultado final?

Visualizaciones

Se realizaron heatmaps interactivos de correlación por género y evento, gráficos comparativos de importancia de variables y análisis de error mediante validación cruzada. Estas visualizaciones permitieron detectar relaciones clave entre estaciones y rendimiento total.

Resultados principales

Resultados en mismo test set:
Base: MAE = 298.93 s | R^2 = 0.8285
Enriquecido: MAE = 198.07 s | R^2 = 0.9164

GroupKFold CV (por event_id):
Baseline MAE = 311.07 ± 14.70
Enriquecido MAE = 204.34 ± 16.86

Top features (por permutation importance): cluster_0, pct_last_3_stations, station_7, station_4, cluster_1.

Insights

El enriquecimiento del dataset generó una mejora sustancial en el modelo. El MAE se redujo casi 100 segundos y el R^2 aumentó significativamente, demostrando que las nuevas variables (clusters, proporciones de estaciones y métricas derivadas) aportan información relevante. Las estaciones 7 y 4 destacan como puntos críticos del rendimiento, mientras que la inclusión del perfil de cluster arrojó patrones multivariados que los tiempos individuales no reflejaban.

Conclusión y posible trabajo futuro

El enriquecimiento de los datos y la aplicación de técnicas avanzadas de limpieza y modelado demostraron un impacto real en la predicción de rendimiento. Los resultados son más estables y generalizables entre eventos. En próximas etapas se pueden explorar modelos más complejos (como Gradient Boosting o XGBoost) y la incorporación de datos externos (como clima o frecuencia cardíaca).

Repositorio GitHub

https://github.com/SDP-data/Proyecto_ML_Parte1_Diaz