ANÁLISIS DE CLASIFICADORES SUPERVISADOS EN MINERÍA DE DATOS

Julio Cesar Aguirre Rojas, Universidad Nacional de Colombia, jucaguirrero@unal.edu.co
Basil Esteban Borda Avila, Universidad Nacional de Colombia, bbordaa@unal.edu.co

Julio 19, 2025

Abstract

This article presents a case study of a large public ultramarathon dataset, with over 7.4 million individual records collected between 1798 and 2022. The dataset includes information on events, distances, speeds, gender, age, and performance of over 1.6 million unique runners. Exploratory data analysis (EDA) and preprocessing techniques were applied, including data cleaning, variable transformation, outlier detection, and novel feature creation. The objective of the study is to identify patterns of performance and participation based on demographic and contextual variables. The results reveal historical trends in ultramarathon practice, as well as relevant differences by gender, age, and race type. This analysis highlights the value of open and longitudinal data for the study of human performance and the evolution of extreme sports disciplines.

Keywords

Data mining, data mining, ultramarathon, exploratory data analysis, sports performance, preprocessing, outliers, visualization, longitudinal data, Naive Bayes.

Resumen

Este artículo presenta un estudio de caso sobre un extenso conjunto de datos públicos de ultramaratones, con más de 7.4 millones de registros individuales recolectados entre 1798 y 2022. El dataset incluye información sobre eventos, distancias, velocidades, género, edad y rendimiento de más de 1.6 millones de corredores únicos. Se aplicaron técnicas de análisis exploratorio (EDA) y preprocesamiento, incluyendo limpieza de datos, transformación de variables, detección de valores atípicos, y creación de nuevas características. El objetivo del estudio es identificar patrones de rendimiento y participación en función de variables demográficas y contextuales. Los resultados revelan tendencias históricas en la práctica del ultramaratón, así como diferencias relevantes por género, edad y tipo de carrera. Este análisis destaca el valor de los datos abiertos y longitudinales para el estudio del rendimiento humano y la evolución de disciplinas deportivas extremas.

Palabras clave

Minería de datos, minería de datos, ultramaratón, análisis exploratorio de datos, rendimiento deportivo, preprocesamiento, outliers, visualización, datos longitudinales, Naive Bayes.

1 - Introducción

En las últimas décadas, la disponibilidad masiva de datos ha impulsado el desarrollo y aplicación de técnicas de minería de datos en múltiples dominios. El deporte, y en particular las disciplinas de resistencia extrema como el ultramaratón, no son la excepción. Estos eventos, que superan la distancia tradicional del maratón (42.195 km), reúnen una gran cantidad de variables de contexto que, al ser recopiladas de forma sistemática a lo largo de años, constituyen una fuente valiosa para el análisis exploratorio y predictivo.

Este trabajo presenta un estudio aplicado de minería de datos sobre un conjunto de datos públicos de ultramaratones, que incluye más de 7.4 millones de registros individuales correspondientes a eventos realizados entre 1798 y 2022. El dataset contiene información detallada de los participantes —como edad, género, nacionalidad, club, ritmo promedio y distancia recorrida—, permitiendo aplicar técnicas tanto supervisadas como no supervisadas para identificar patrones ocultos, segmentar corredores y generar modelos de predicción.

La metodología seguida cubre todas las fases principales del ciclo de minería de datos, iniciando con un exhaustivo proceso de preprocesamiento y análisis exploratorio, seguido por técnicas de agrupamiento mediante el algoritmo K-Prototypes, reglas de asociación extraídas con Apriori, y finalmente un modelo de clasificación supervisado utilizando Naïve Bayes. A lo largo de este proceso se construyen variables derivadas,

se reduce la granularidad de ciertos atributos y se emplean visualizaciones para facilitar la interpretación.

El objetivo de este proyecto no es solo obtener conocimiento útil sobre la práctica del ultramaratón, sino también demostrar la aplicabilidad y el valor de distintas técnicas de minería de datos sobre un dataset complejo, masivo y altamente heterogéneo. En última instancia, se busca ilustrar cómo la combinación de ingeniería de características, algoritmos adecuados y análisis crítico puede transformar datos abiertos en conocimiento significativo, tanto para la investigación deportiva como para el aprendizaje académico.

2- Origen del Dataset:

El presente estudio se basa en un conjunto de datos público disponible en la plataforma Kaggle, que contiene registros históricos de resultados en carreras de ultramaratón realizados entre los años 1798 y 2022. Este dataset reúne más de 7.4 millones de participaciones individuales correspondientes a más de 1.6 millones de corredores únicos, lo que lo convierte en una fuente de información amplia y diversa para el análisis de patrones deportivos a lo largo del tiempo.

La **Tabla 1** resume las principales variables incluidas en el conjunto de datos, su tipo y una breve descripción de su contenido.

Descripción del conjunto de datos

Nombre de Columna	Tipo de dato	Descripción
Year of event	int64	Año del evento
Event dates	object	Fecha específica del evento
Event name	object	Nombre del evento (incluye país)
Event distance/length	object	Distancia recorrida
Event number of finishers	int64	Número de finalistas
Athlete performance	object	Tiempo registrado por el atleta
Athlete club	object	Club al que pertenece el atleta
Athlete country	object	País de origen del atleta
Athlete year of birth	float64	Año de nacimiento
Athlete gender	object	Género del atleta
Athlete age category	object	Rango de edad por categoría
Athlete average speed	object	Velocidad promedio en km/h
Athlete ID	int64	ID único del atleta (anonimizado)

Tabla 1. Descripción de las variables del conjunto de datos

2.1 - Resultados del análisis exploratorio de los datos

Exploración Inicial y Generación de Variables Derivadas

En cuanto al preprocesamiento de datos, se llevaron a cabo varias tareas fundamentales para homogeneizar y depurar el conjunto. En primer lugar, se parametrizan algunas columnas: la distancia, originalmente expresada en distintas unidades (millas, metros, etc.), se unificó completamente en kilómetros; la edad de cada atleta se calculó a partir de su fecha de nacimiento y la fecha del evento; y, a partir del tiempo de finalización y la distancia recorrida, se obtuvo la velocidad en km/h para, a continuación, transformarla a ritmo (min/km), métrica más representativa de las variaciones de velocidad en pruebas de ultramaratón. Asimismo, se estandarizó el formato de las fechas al patrón AAAA-MM-DD y, finalmente, se eliminaron registros con valores erróneos o excesivamente atípicos—probablemente derivados de fallos de digitación o captura—reduciendo la muestra en alrededor de un millón de observaciones.

2.1.1 -distribución de los datos base

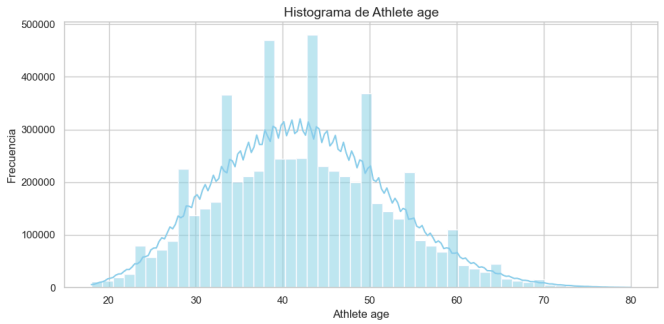

Figura 1:histograma edades

En la figura 1 tenemos la distribución de la edad de los atletas representados en un histograma donde vemos que el valor medio es de 42 años, aunque también tenemos noción de algunos picos de valor en [28, 34, 44, 49, 60], posiblemente resultado de algún ajuste manual.

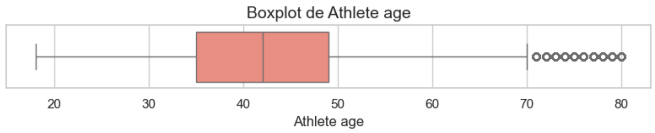

Figura 2: boxplot edades

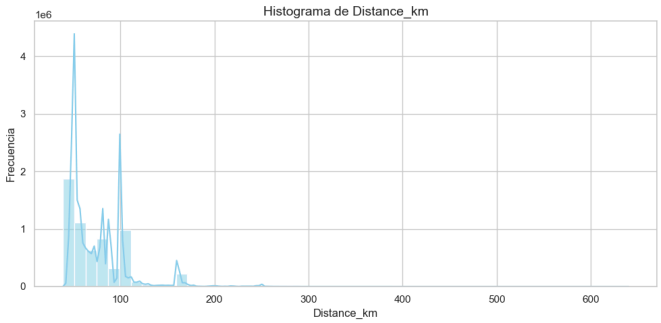

Figura 3:histograma distancias

En la figura 3 tenemos la distribución de la distancia de las carreras informadas, vemos que la mayoría de las carreras son

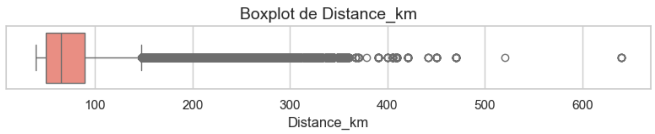

Figura 4: boxplot distancias

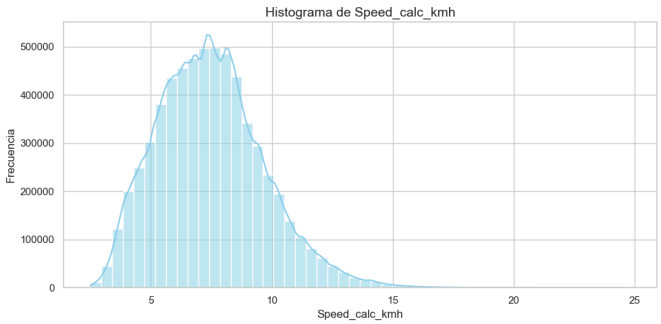

Figura 3:histograma velocidades calculadas (km/h)

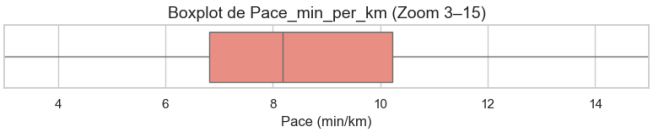

Figura 4: boxplot pace (min/km)

Debido al tamaño del conjunto de datos (~6 millones tras limpieza inicial), se trabajó en chunks durante el análisis exploratorio. En esta fase se identificaron problemas de granularidad, redundancia y heterogeneidad en varias columnas clave.

Como parte del preprocesamiento, se generaron las siguientes variables categóricas a partir de atributos numéricos:

- **grouped_distances**: Bins de distancia (≤70, 71–100, 101–160, 161–200, >200 km).
- **grouped_pace**: Bins de ritmo promedio (<6, 6–8, 8–10, …, >20 min/km).
- **age_without_gen**: Edad numérica agrupada por décadas (20, 30, 40…), extraída de la categoría original.

Análisis de cardinalidades y calidad de datos

Se realizó análisis de cardinalidad por variable (p. ej. número de eventos únicos, países, clubes), detección de valores nulos y valores atípicos. Dado el gran volumen de datos se eliminaron columnas con valores vacíos y se homogeneizó el formato de las restantes.

3 - Agrupamiento (Clustering)

El objetivo de esta etapa fue identificar patrones de agrupamiento entre corredores de ultramaratón combinando variables numéricas y categóricas. Inicialmente se evaluaron métodos clásicos como K-Means y K-Modes, pero cada uno presentó limitaciones específicas. K-Means no maneja adecuadamente variables categóricas, mientras que K-Modes, aunque trabaja bien con variables no numéricas, produce clusters poco representativos cuando se usan transformaciones categóricas de edad o ritmo.

Como solución, se empleó el algoritmo K-Prototypes, que extiende a K-Means y K-Modes, permitiendo trabajar con variables mixtas. Este método fue implementado con el paquete kmodes en Python, configurando el algoritmo con init='Cao', normalización previa de las variables numéricas

(pace_min_per_km y athlete_age), y etiquetas categóricas como athlete_gender y grouped_distances.

Se trabajó sobre una muestra aleatoria estratificada de 500 000 registros, seleccionada con semilla fija para reproducibilidad. Las variables categóricas fueron transformadas con codificación ordinal y las numéricas estandarizadas con StandardScaler.

3.1 Visualización Comparativa de Métodos

En la Figura 1 se observa la diferencia entre los resultados generados por K-Modes y K-Prototypes. Mientras que el primero tiende a forzar agrupaciones rígidas por etiquetas, el segundo logra una separación más realista de los grupos según el ritmo y la edad.

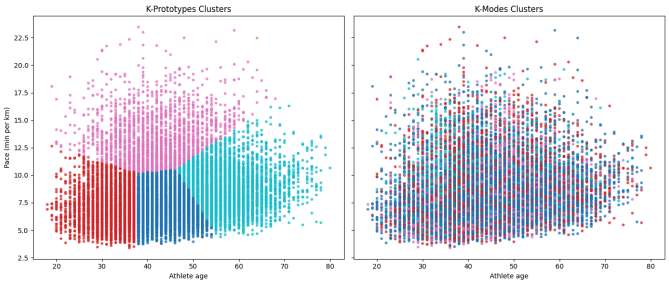

Figura 1: Comparación entre K-Prototypes y K-Modes (pace vs age).

3.2 Determinación del Número de Clusters (k)

Se realizaron múltiples pruebas visuales con distintos valores de k (número de clusters) para observar el comportamiento y coherencia de los grupos. En la **Figura 2** se presenta el resultado con k=6, que muestra cierta separación pero también redundancia. En la Figura 3 (k=8) se aprecia un exceso de fragmentación, mientras que en la Figura 4 (k=4) se logra un balance entre interpretabilidad y coherencia.

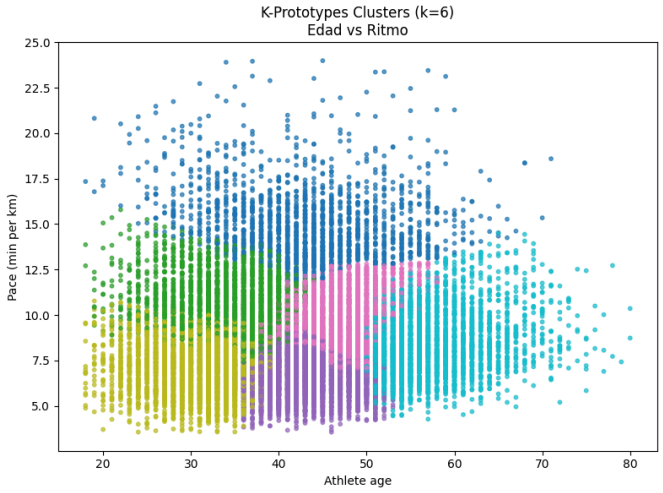

Figura 2: K-Prototypes con k=6 (sample 100000).

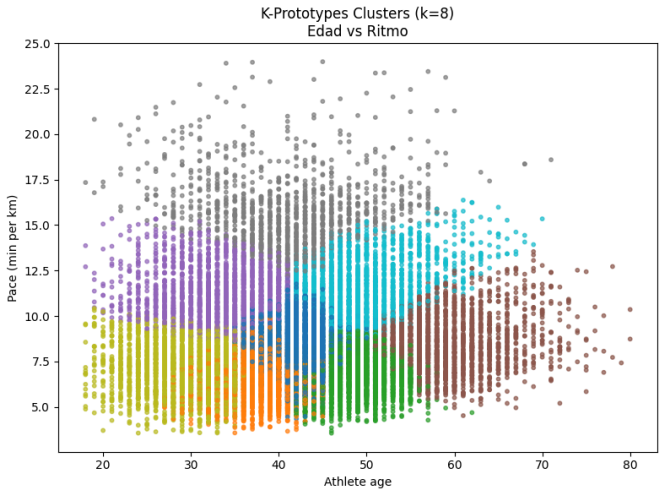

Figura 3: K-Prototypes con k=8 (sample 100000).

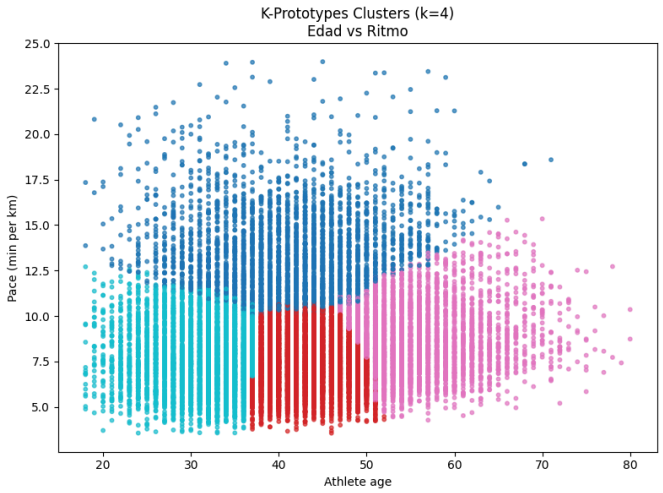

Figura 4: K-Prototypes con k=4 (sample 100000).

3.3 Perfilamiento de Clusters Finales (k=4)

El modelo final elegido fue k=4, ya que presenta una segmentación clara y consistente con el análisis posterior de asociación y clasificación. A continuación, se describen brevemente los perfiles de cada cluster, obtenidos a partir de observaciones cruzadas entre edad, ritmo, distancia y género:

Cluster 0 – "Recreativos-avanzados"

Corredores de edad media (~45 años), ritmo sólido (6–8 min/km) y distancia media. Predominan atletas hombres sin club formal.

Cluster 1 – "Ultras de media distancia"

Atletas en torno a los 35 años, ritmo élite (<6 min/km) o moderado (6–8), participando en carreras de 71–100 km. Alto rendimiento.

Cluster 2 – "Ritmos lentos"

Ritmos de 12–16 min/km, sin club, mayoría hombres. Representan corredores noveles, rutas técnicas o enfoque recreativo.

Cluster 3 – "Veteranos de ritmo medio"

Corredores de 55 a 65 años, con ritmos de 8–10 min/km. Pertenecen mayoritariamente al grupo de veteranos, usualmente sin club.

3.4 Visualización de Clusters Finales

Las siguientes gráficas muestran la separación lograda con el modelo final (k=4), proyectando las observaciones sobre dos pares de variables clave:

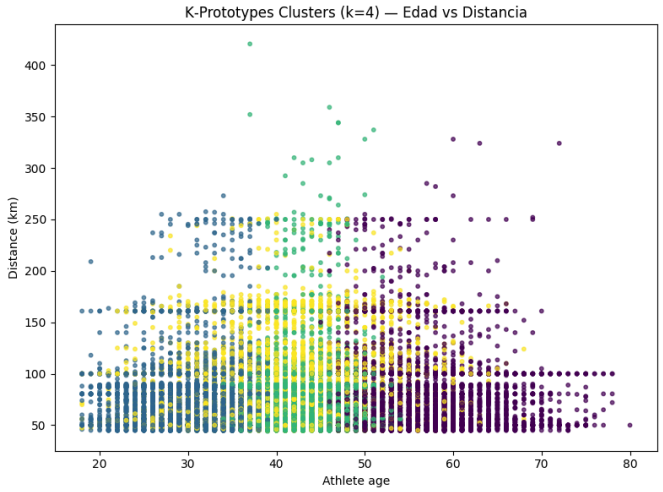

Figura 5:. Relación entre edad y distancia por cluster

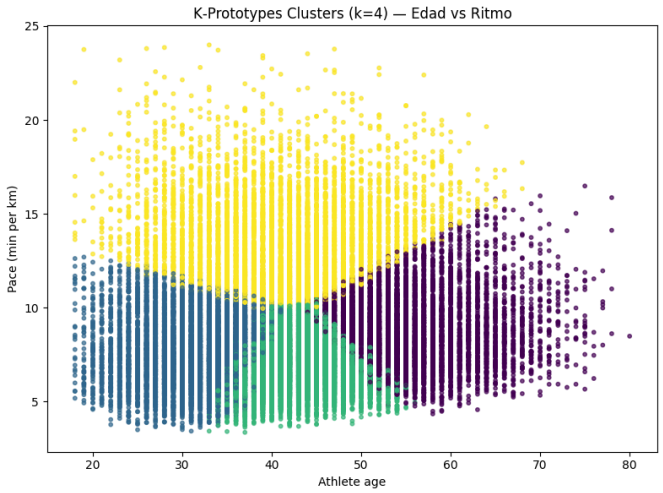

Figura 6: Relación entre edad y ritmo promedio por cluster

3.5 Conclusión del Agrupamiento

El algoritmo K-Prototypes con k=4 permitió obtener una clasificación natural de los corredores basada en sus características de rendimiento y demografía. Esta segmentación fue validada posteriormente mediante reglas de asociación y clasificación supervisada, que arrojaron resultados coherentes con la estructura descubierta en esta etapa.

4 - Reglas de asociación

El análisis de reglas de asociación tuvo como objetivo identificar patrones frecuentes dentro de los grupos formados por el modelo de clustering (K-Prototypes). Para ello, se aplicó

el algoritmo Apriori sobre una versión codificada del dataset, utilizando como variable objetivo la etiqueta cluster.

Se trabajó exclusivamente con variables categóricas derivadas o transformadas previamente:

- grouped_distances
- grouped_pace
- age_without_gen
- athlete_gender
- athlete_country (Top 10 + "Other")
- athlete_club (Top 10 + "Other")

Cada valor fue representado como un ítem del tipo variable=valor, generando transacciones compatibles con el algoritmo. El soporte mínimo se estableció en 0.02 y la confianza mínima en 0.70, lo que permitió identificar asociaciones con representatividad y precisión aceptables.

4.1 Ejemplos de Reglas por Cluster

La siguiente tabla resume las reglas más destacadas para cada cluster, ordenadas por su valor de lift, el cual indica cuán más probable es encontrar el consecuente (ej. cluster=2) en presencia del antecedente, comparado con una selección aleatoria.

Cluster 0 – "Recreativos-avanzados"

grouped_pace=6-8, age_without_gen=45 ⇒ cluster=0 (lift ≈ 2.82, conf = 1.00)

grouped_distances=71–100, age_without_gen=50 ⇒ cluster=0 (lift ≈ 2.95, conf = 0.86)

Estas reglas sugieren que los corredores de mediana edad, con ritmos sólidos y distancias intermedias, son característicos del cluster 0.

Cluster 1 – "Ultras de media distancia"

grouped_pace=6–8, age_without_gen=35 ⇒ cluster=1 (lift ≈ 3.72, conf = 1.00)

grouped_pace=<6, age_without_gen=35 ⇒ cluster=1 (lift ≈ 4.09, conf = 0.89)

Cluster asociado a atletas de alto rendimiento, con ritmos rápidos y edad promedio más baja.

Cluster 2 – "Ritmos lentos, sin club"

grouped_pace=14–16, athlete_club=Sin club ⇒ cluster=2 (lift = 5.39, conf = 0.98)

grouped_pace=12–14, athlete_gender=M, athlete_club=Sin club ⇒ cluster=2 (lift ≈ 4.91, conf = 0.89)

Este grupo representa corredores más recreativos o en contextos técnicos, con escasa afiliación institucional.

Cluster 3 – "Veteranos de ritmo medio"

age_without_gen=55, grouped_pace=8-10 ⇒ cluster=3 (lift ≈ 5.12, conf = 1.00)

age_without_gen=60, athlete_club=Other ⇒ cluster=3 (lift ≈ 4.88, conf = 0.95)

Grupo predominantemente masculino, con edad entre 55 y 65 años, ritmo medio y sin pertenencia a clubes principales.

4.2 Análisis de reglas por Lift

Las reglas generadas por el algoritmo Apriori fueron ordenadas según su valor de lift, una métrica que indica cuán más probable es que ocurra el consecuente (por ejemplo, pertenecer al cluster=2) en presencia del antecedente, en comparación con una ocurrencia aleatoria. A continuación, se presentan algunas de las reglas más representativas:

- grouped_pace=14–16, athlete_club=Sin club ⇒ cluster=2

  (support=0.020, confidence=0.98, lift=5.39)

- age_without_gen=55, grouped_pace=8–10 ⇒ cluster=3

  (support=0.031, confidence=1.00, lift=5.12)

- age_without_gen=65 ⇒ cluster=3, athlete_gender=M

  (support=0.021, confidence=0.85, lift=5.22)

- grouped_pace=12–14, athlete_club=Sin club ⇒ cluster=2

  (support=0.029, confidence=0.90, lift=4.94)

- age_without_gen=60, grouped_distances=≤70 ⇒ cluster=3

  (support=0.028, confidence=0.96, lift=4.93)

Las reglas con mayor lift se concentraron principalmente en los clusters 2 y 3. Esto es consistente con los perfiles más extremos observados en la etapa de agrupamiento:

Cluster 2: Asociado a corredores con ritmos muy lentos (14–16 min/km), generalmente hombres y sin club. Las reglas relacionadas a este grupo presentan lift > 5 y confianza cercana al 1.00, lo que sugiere que el comportamiento de estos corredores es altamente predecible a partir de sus características.

Cluster 3: Asociado a atletas veteranos (55–65 años), con ritmos medios (8–10 min/km). Las reglas indican que el género masculino y la pertenencia a clubes fuera del top 10 (Other) refuerzan la probabilidad de pertenencia a este grupo.

Los clusters 0 y 1 presentaron reglas válidas pero con menor lift (<3.5), lo que sugiere que su composición es más variada o difusa en comparación con los perfiles más definidos de los clusters 2 y 3.

Además de su utilidad descriptiva, las reglas de asociación resultan ser una herramienta efectiva para validar y complementar los resultados obtenidos mediante clustering. La aparición de reglas con lift significativamente alto, especialmente en los clusters 2 y 3, confirma que dichos grupos presentan patrones conductuales claramente definidos, lo cual sugiere una buena cohesión interna.

En contraste, los clusters 0 y 1, si bien muestran algunas asociaciones relevantes, poseen perfiles más dispersos y heterogéneos. Esto puede deberse a la superposición de atributos entre corredores de niveles intermedios o a variaciones no capturadas por las variables incluidas.

Cabe resaltar que este análisis no sólo revela regularidades, sino que también identifica *subgrupos ocultos* que podrían ser invisibles mediante métodos tradicionales. Por ejemplo, el grupo de hombres mayores sin club, con ritmos de 12–16 min/km, representa una fracción menor del total, pero altamente consistente en su comportamiento competitivo.

5. Clasificación Supervisada

En la etapa final del proyecto se implementaron modelos de clasificación supervisada con el objetivo de predecir el grupo (cluster) al que pertenece un atleta, en función de sus características personales y de desempeño. Esta tarea permite evaluar la coherencia de los clusters obtenidos previamente y construir modelos capaces de anticipar el comportamiento esperado de un corredor, dada su edad, género, ritmo y tipo de carrera.

5.1 Preparación del Dataset

Para esta tarea se utilizó la misma muestra balanceada de 500 000 registros empleada en el modelo de agrupamiento. Se conservaron las siguientes variables como predictores:

- athlete_age (numérica)
- grouped_pace (categórica)
- grouped_distances (categórica)
- athlete_gender (categórica)

La variable objetivo fue cluster, generada previamente mediante K-Prototypes (k=4).

Se aplicó codificación one-hot a las variables categóricas y normalización a la variable numérica. El conjunto se dividió en entrenamiento y prueba en una proporción 70/30, manteniendo la distribución estratificada de clases.

5.2 Implementación y Evaluación del Modelo

Se utilizó un clasificador Naïve Bayes (versión GaussianNB) por su eficiencia, interpretabilidad y bajo costo computacional. El entrenamiento se realizó sobre 350 000 muestras, y la evaluación se centró en las 150 000 restantes. Los resultados fueron los siguientes:

- Accuracy: 0.88
- F1-score macro: 0.86
- Precision macro: 0.87
- Recall macro: 0.86

Cluster	Precision	Recall	F1-score	Soporte
0	0.88	0.93	0.91	53 139
1	0.94	0.94	0.94	40 339
2	0.74	0.81	0.77	27 248
3	0.93	0.76	0.83	29 274

5.3 Interpretación del Modelo

Cluster 0: Muy buen desempeño con recall de 0.93, aunque con cierto margen de falsos positivos (precisión 0.88). El modelo detecta correctamente casi todos los casos reales de este grupo.

Cluster 1: Es el más fácil de identificar, con precisión y recall de 0.94. Esto valida su fuerte cohesión interna observada también en la fase de asociación.

Cluster 2: El más difícil de clasificar. Aunque el recall es aceptable (0.81), la precisión es baja (0.74), indicando una alta proporción de falsos positivos. Este grupo comparte atributos con otros clusters y presenta mayor variabilidad.

Cluster 3: Alta precisión (0.93) pero recall moderado (0.76). El modelo tiende a identificar bien los ejemplos de cluster 3 cuando los predice, pero no logra detectar todos los casos reales.

El cluster 2 sufre confusiones frecuentes con los clusters 0 y 1.

El cluster 3 es confundido con cluster 0 (4 329 casos) y cluster 2 (2 790 casos).

El cluster 1 tiene la matriz más limpia, sin confusión con cluster 0.

5.4 Conclusiones

El modelo Naïve Bayes logró un rendimiento muy sólido (88 % de accuracy), especialmente teniendo en cuenta su simplicidad y velocidad de entrenamiento. Es útil como línea base para tareas de clasificación posteriores y proporciona una validación externa para los clusters definidos previamente.

Se identificaron oportunidades de mejora:

Cluster 2 podría beneficiarse de feature engineering adicional, incorporando variables como distance_km en su versión continua o transformaciones basadas en país/club.

Cluster 3 podría mejorar si se redefinen los bins de ritmo y distancia para hacer el grupo más homogéneo.

Podría explorarse el uso de CategoricalNB si se trabaja exclusivamente con variables categóricas, o bien modelos más complejos como Random Forest o XGBoost para capturar no linealidades.

Referencias

[1] Kaggle: UltraRunning Race Data — https://www.kaggle.com/datasets/piterfm/ultramarathon-running-race-results

[2] Scikit-learn: Machine Learning in Python — https://scikit-learn.org/

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.