**Estructura de Presentación de Propuesta de Proyecto para Análisis de Datos**

**Grupo #8 Tester ACE** 

**Integrantes:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Identificación | Apellidos | Nombres | Correo | Teléfono |
|  | **Ruiz Acosta** | **Jorge Luis** | **Jorgeruizacosta117@gmail.com** |  |
|  | **De La Espriella Gándara** | **Eric David** | **ericddg12.23@gmail.com** |  |
|  | **Rojas Acevedo** | **Gissella** | **Rojasacevedo18@gmail.com** |  |
|  | **Martínez Contreras** | **Diana Karina** | **dianamcontreras22@gmail.com** |  |
|  |  |  |  |  |

**Título: Identificación de patrones colectivos en datos públicos de Strava mediante análisis estadístico y dashboards interactivos**

**1. Introducción**

La propuesta aplica CRISP-DM al análisis de 544 actividades públicas y anónimas de Strava (abril 2020–noviembre 2022), que incluyen métricas de distancia, tiempo en movimiento, velocidad y frecuencia cardíaca. A diferencia de los paneles individuales de Strava, buscamos una visión agregada que revele patrones colectivos y tendencias generales.

Emplearemos Python (pandas, matplotlib) para limpieza, estadística descriptiva y exploración, y Power BI para dashboards interactivos. Cada fase (comprensión, preparación, análisis, evaluación y despliegue) estará diseñada para ser plenamente reproducible.

A futuro, este marco soportará la creación de paneles dinámicos con filtros personalizables y el desarrollo de modelos predictivos, consolidando un producto analítico de uso real en entornos deportivos y de salud.

**2. Definición del Problema**

Strava ofrece métricas y visualizaciones individuales, pero no cuenta con herramientas para analizar de forma conjunta sus 544 actividades anónimas ni comparar sus distribuciones por tipo (“Ride” 54.6 %, “Walk” 21.3 %, “Hike” 12.3 %). Esta carencia imposibilita detectar tendencias colectivas; por ejemplo, cómo varía la velocidad media (13.5 km/h) o la frecuencia cardíaca (115.7 bpm) según el tipo de actividad, información valiosa para estudios deportivos, planificación estratégica o desarrollos futuros.

Aplicar análisis agregado a registros públicos y estructurados de Strava permitirá generar insights más amplios y comprensibles sobre la actividad física de la comunidad.

**3. Área del Proyecto**

Se ubica en el análisis de datos deportivos, con foco en estadística descriptiva, visualización avanzada y storytelling de datos. Emplea Python (pandas, matplotlib) para procesamiento y análisis, y Power BI para dashboards interactivos, siguiendo la metodología CRISP-DM para garantizar trazabilidad y reproducibilidad en cada fase.

**4. Justificación**

Contamos con un dataset de 544 actividades anónimas de Strava —que cubren 11 tipos de ejercicio (54.6 % Ride, 21.3 % Walk, 12.3 % Hike, entre otros)— y métricas como distancia, tiempo en movimiento, velocidad y frecuencia cardíaca. Esta muestra es suficiente para detectar correlaciones (por ejemplo, entre velocidad media y desnivel) y tendencias generales reproducibles.

Al analizar de forma agregada estos registros públicos, podemos llenar el vacío de visión comunitaria que deja Strava y generar insights de valor para investigadores deportivos, entrenadores y desarrolladores de productos de salud. Además, la metodología basada en CRISP-DM y herramientas como Python y Power BI garantizan un proceso técnico, exploratorio y reproducible, sentando las bases para modelos predictivos y dashboards interactivos más avanzados.

**5. Objetivo General**

Identificar patrones colectivos en un conjunto de datos públicos de Strava mediante técnicas de análisis estadístico y presentarlos en dashboards interactivos reproducibles.

**6. Objetivos Específicos**

1. Examinar la estructura, calidad y variables del dataset para asegurar su idoneidad.
2. Aplicar estadísticas descriptivas y comparativas para revelar tendencias según tipo de actividad.
3. Diseñar visualizaciones interactivas claras que destaquen los hallazgos clave.
4. Interpretar los resultados para entender el comportamiento agregado de los usuarios.
5. Proponer mejoras y extensiones futuras, incluyendo modelos predictivos y filtros dinámicos en los dashboards.

**7. Alcance del Proyecto**

Este proyecto se enfoca en el análisis exploratorio y estadístico de un dataset de 544 actividades públicas y anonimizadas de Strava, con énfasis en métricas de rendimiento (distancia, tiempo en movimiento, velocidad media y máxima, desnivel y frecuencia cardíaca).

Se seguirá la metodología CRISP-DM en sus fases de comprensión de datos, limpieza, análisis descriptivo, visualización de patrones e interpretación de resultados, empleando Python (pandas, matplotlib) para procesamiento y Power BI para la creación de informes interactivos.

No incluye el desarrollo de una aplicación o interfaz final, pero se ofrecerán recomendaciones para futuros paneles dinámicos y la implementación de modelos predictivos.

Todas las etapas respetan la naturaleza pública y anonimizada de los datos, asegurando la confidencialidad de los usuarios.

**8. Sugerencias para desarrollo futuro**

Para evolucionar este análisis exploratorio hacia una solución más completa y de uso real, se proponen:

* Modelos predictivos de rendimiento: aplicar técnicas de machine learning para estimar métricas clave (distancia, velocidad, frecuencia cardíaca) a partir de variables históricas.
* Dashboards interactivos avanzados: incorporar filtros dinámicos y comparaciones personalizables que faciliten la exploración de tendencias por tipo de actividad, rango temporal o segmentos geográficos.
* Aplicación web o móvil integrada: centralizar paneles y modelos en una interfaz accesible para entrenadores, clubes y comunidades deportivas, con énfasis en la visión colectiva.
* Análisis longitudinal: aprovechar la serie temporal de start\_date\_local para estudiar la evolución del rendimiento a lo largo de semanas, meses o años.
* Ampliación de fuentes de datos: incorporar otras APIs o datasets complementarios (meteorología, rutas, perfil demográfico) manteniendo la anonimización y el respeto a la privacidad.

Estas iniciativas potenciarán la generación de insights accionables y el valor de la herramienta analítica en entornos académicos y profesionales del deporte.

**9. Metodología (Basada en CRISP-DM)**

**Comprensión del negocio**

Para dar solidez a este apartado, proponemos estructurarlo en subsecciones que recojan el contexto, los objetivos de negocio, los interesados y los criterios que determinarán el éxito del proyecto.

1. **Contexto y oportunidad**

Strava es una red social deportiva que registra millones de actividades individuales, pero carece de un módulo nativo para análisis agregado. Al explotar un dataset de 544 actividades públicas y anónimas (11 modalidades distintas), existe la oportunidad de:

* Descubrir patrones colectivos que no son visibles en la vista individual,
* Apoyar a entrenadores y gestores con datos objetivos para planificar entrenamientos o eventos grupales,
* Guiar desarrollos de producto en plataformas de fitness basados en tendencias reales de la comunidad.

1. **Stakeholders (interesados)**

* Investigadores y académicos: requieren evidencia empírica de comportamientos deportivos.
* Entrenadores y clubes: buscan comparar rendimientos promedio y diseñar planes de grupo.
* Equipos de producto/marketing de apps de salud: precisan insights para nuevas funcionalidades enfocadas en el rendimiento colectivo.

1. **Objetivos de negocio**
   1. Revelar patrones de rendimiento (velocidad, distancia, frecuencia cardíaca) según tipo de actividad.
   2. Optimizar la toma de decisiones en entrenamiento grupal y eventos deportivos.
   3. Sentar las bases para un producto analítico escalable (dashboards y modelos predictivos).
2. **Preguntas clave**

* ¿Cómo varía la velocidad media entre “Ride” (54,6 % de las actividades), “Walk” (21,3 %) y “Hike” (12,3 %)?
* ¿Qué rangos de frecuencia cardíaca son los más comunes en cada modalidad?
* ¿Existen patrones estacionales (mensuales/trimestrales) en distancia o tiempo en movimiento?

1. **Criterios de éxito**

* Interactividad: dashboard con filtros dinámicos por tipo de actividad y rango temporal.
* Reproducibilidad: código Python y reportes generados programáticamente.
* Utilidad: validación con al menos un stakeholders (ej. entrenador) que confirme la aplicabilidad de los insights.

1. **Entregables de esta fase**

* Documento de alcance y objetivos de negocio.
* Lista de preguntas de análisis y métricas a extraer.
* Aprobación del plan de proyecto por los stakeholders.

**Comprensión de los datos**

El dataset utilizado contiene 544 registros de actividades físicas extraídas de la plataforma Strava, con datos públicos, anónimos y estructurados. Cada registro representa una sesión deportiva, e incluye 10 variables clave como fecha, tipo de actividad, distancia recorrida, duración, velocidad media y máxima, elevación, y frecuencia cardíaca promedio y máxima.

Durante esta fase se analizó la estructura del dataset con los siguientes hallazgos:

* Cobertura temporal: actividades registradas entre abril de 2020 y noviembre de 2022.
* Tipos de actividad: 11 modalidades distintas, siendo las más frecuentes: Ride (54.6 %), Walk (21.3 %) y Hike (12.3 %).
* Variables numéricas: métricas de rendimiento físico como distance, moving\_time, average\_speed, average\_heartrate, que permiten aplicar estadística descriptiva, comparativa y correlacional.
* Calidad de los datos: se detectaron algunos valores nulos en variables como average\_heartrate, lo que implica la necesidad de limpieza o imputación según el análisis requerido.
* Estructura y formato: datos en formato CSV con etiquetas claras y consistentes, listos para ser manipulados mediante herramientas como Python (pandas) y Power BI.

Esta comprensión inicial permite validar que el dataset es adecuado para el análisis exploratorio propuesto y que contiene información suficiente para responder a las preguntas clave del proyecto (variaciones por tipo de actividad, tendencias en velocidad y frecuencia cardíaca, etc.). Además, proporciona la base para aplicar transformaciones posteriores como la conversión de unidades, normalización y creación de nuevas métricas derivadas.

**Preparación de los datos**

Esta fase consiste en transformar el dataset original en una versión limpia, coherente y lista para el análisis, asegurando que los datos respondan adecuadamente a los objetivos del proyecto.

Las acciones realizadas incluyen:

**Limpieza de datos:**

1. Identificación y tratamiento de valores nulos, especialmente en variables como average\_heartrate y max\_heartrate.
2. Eliminación o imputación de registros incompletos según el caso.

**Transformación de unidades:**

1. Conversión de distancia de metros a kilómetros.
2. Transformación de tiempo en movimiento (moving\_time) de segundos a minutos.
3. Conversión de velocidad de m/s a km/h para facilitar la interpretación.

**Generación de nuevas variables derivadas:**

1. Cálculo de ritmo promedio (min/km) como una medida adicional de rendimiento.
2. Agrupación por tipo de actividad para facilitar el análisis comparativo.

**Normalización y formato:**

1. Conversión de fechas (start\_date\_local) al tipo datetime.
2. Homogeneización de categorías en la variable type.

Este proceso garantiza que los datos estén preparados de forma consistente y estructurada para su uso en análisis estadístico, visualizaciones y construcción de dashboards. Además, se asegura la trazabilidad de cada transformación para mantener la reproducibilidad del proyecto.

**Modelado**

En esta fase se aplicaron técnicas de análisis estadístico y visual exploratorio con el fin de identificar patrones relevantes en el comportamiento físico registrado en el dataset. A diferencia de modelos predictivos complejos, el enfoque se centró en métodos interpretables y adecuados al objetivo exploratorio del proyecto.

Las acciones realizadas incluyen:

**Estadística descriptiva general:**

1. Cálculo de promedios, medianas, máximos y mínimos para variables como distancia, tiempo, velocidad y frecuencia cardíaca.
2. Distribuciones globales y por tipo de actividad (por ejemplo, diferencias de velocidad entre “Ride” y “Walk”).

**Comparaciones entre categorías:**

1. Análisis de variables según la modalidad deportiva (type), evidenciando patrones diferenciados en rendimiento.
2. Boxplots, histogramas y diagramas de dispersión para contrastar grupos.

**Identificación de relaciones entre variables:**

1. Estudio de correlaciones entre variables como average\_speed y elevation\_gain o entre moving\_time y heart rate.
2. Evaluación de posibles tendencias temporales utilizando la variable start\_date\_local.

**Visualización de hallazgos:**

1. Gráficos diseñados con librerías de Python como matplotlib y seaborn, apoyando la narrativa visual del comportamiento colectivo.

Este modelado permite generar insights claros y medibles, orientados a responder las preguntas clave planteadas en la fase de negocio, y sienta la base para la evaluación final y el despliegue en dashboards.

**Evaluación**

En esta fase se revisaron los resultados obtenidos durante el modelado para verificar su coherencia con los objetivos iniciales del proyecto y asegurar que los análisis y visualizaciones realmente aportaran valor desde una perspectiva colectiva y reproducible.

Las acciones realizadas incluyen:

**Validación de hallazgos clave:**

* Verificación de que las diferencias entre modalidades deportivas (por ejemplo, en velocidad y frecuencia cardíaca) fueran estadísticamente consistentes y visualmente claras.
* Confirmación de que los patrones detectados responden a las preguntas formuladas en la etapa de comprensión del negocio.

**Revisión de la narrativa visual:**

* Evaluación de la claridad de las visualizaciones creadas, su legibilidad y su capacidad para comunicar los hallazgos de forma accesible.
* Aseguramiento de que cada gráfico responda a una pregunta específica y no genere ambigüedad.

**Retroalimentación del equipo:**

* Validación cruzada de los hallazgos por parte de los integrantes del grupo para reforzar la interpretación colectiva y detectar oportunidades de mejora o enfoque.

**Revisión de reproducibilidad:**

* Comprobación de que el código utilizado puede ser ejecutado sin errores y que los pasos de limpieza, transformación y análisis están debidamente documentados.

Esta evaluación permitió ajustar algunos enfoques de visualización, mejorar la presentación de resultados y garantizar que el análisis tenga valor técnico, comunicacional y aplicado.

**Despliegue**

La fase de despliegue consiste en presentar los resultados del análisis en una forma accesible, interactiva y útil para los usuarios finales. Aunque este proyecto se desarrolla en un entorno académico, se diseñaron productos y recomendaciones orientados a una aplicación práctica.

Las acciones y entregables incluyen:

**Construcción de un dashboard interactivo en Power BI:**

* Visualizaciones dinámicas de variables clave como velocidad media, distancia, frecuencia cardíaca y distribución por tipo de actividad.
* Inclusión de filtros por modalidad, fecha y otras variables relevantes, permitiendo una exploración personalizada de los datos.

**Documentación del proceso analítico:**

* Estructura reproducible basada en Python, con celdas organizadas por fases CRISP-DM.
* Explicación paso a paso de cada transformación, análisis y visualización, asegurando trazabilidad del trabajo.

**Guía de interpretación de resultados:**

* Indicaciones para leer e interpretar los gráficos del dashboard.
* Recomendaciones para futuros desarrollos (como integración en aplicaciones móviles o paneles comparativos entre usuarios).

**Recomendaciones para escalabilidad:**

* Ampliar el dataset con nuevas fuentes (manteniendo la anonimización).
* Incorporar modelos predictivos para anticipar métricas de rendimiento.
* Integrar el dashboard en plataformas de gestión deportiva o entrenamiento grupal.

El despliegue, aunque limitado en esta fase a un entorno académico, está concebido como la base para una solución analítica más robusta, aplicable en contextos reales y escalable a diferentes audiencias del ecosistema deportivo.

**10. Plan de Trabajo**

|  |  |  |
| --- | --- | --- |
| Semana | Fases CRISP-DM | Entregable |
| 1 | Comprensión del negocio y datos | Informe breve de objetivos y perfil del dataset (estructura, calidad). |
| 2 | Preparación de datos y modelado | Dataset limpio, nuevas variables calculadas y análisis estadístico preliminar. |
| 3 | Modelado (visualizaciones) y evaluación | Prototipo de dashboard en Power BI con gráficas clave; feedback inicial. |
| 4 | Despliegue | Dashboard finalizado, guía de uso y repositorio con código/documentación. |

**11. Recursos Necesarios**

**Técnicos:**

* Python 3.x con pandas, matplotlib/seaborn. (Colab – VS Code)
* Power BI Desktop para creación de dashboards.
* Git/GitHub para control de versiones y documentación.

**Datos:**

* Archivo strava\_data.csv (544 registros, 10 variables).

**Humanos:**

* Analista de datos (configuración, limpieza y modelado).
* Diseñador de dashboards (UX e interacción en Power BI).
* Revisor académico o compañero para validación de hallazgos.

**12. Evaluación y Métricas de Éxito**

* ***Criterios de Éxito:*** Basados en la fase 5 de CRISP-DM:
  + El dashboard muestra datos claros y responde al problema.
  + El storytelling es comprensible y atractivo.
* ***Métricas:*** Precisión de los datos, facilidad de uso del dashboard, feedback del curso.

**13. Conclusión**

* ***Resumen:*** Recapitular el uso de CRISP-DM para crear un dashboard útil en el área elegida, con storytelling efectivo.