**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**

* **Objetivo:** revelar patrones colectivos en 544 actividades de Strava y diseñar un dashboard que facilite su exploración.
* **Preguntas clave:** ¿Cómo varía la velocidad media según el tipo de actividad? ¿Qué rangos de frecuencia cardíaca son más frecuentes?

**Comprensión de los datos**

* Revisar las 10 variables del CSV (fecha, tipo, distancia, tiempo, desnivel, velocidades y pulso).
* Validar rango temporal (abril 2020–nov 2022), detectar Outliers y missing en average\_heartrate y max\_heartrate.

**Preparación de los datos**

* Limpiar timestamps, convertir unidades (m → km, s → min).
* Imputar o descartar registros con datos faltantes críticos.
* Generar nuevas columnas si aplica (ritmo = distancia/tiempo).

**Modelado**

* Estadística descriptiva: medias, medianas, desviaciones por tipo de actividad.
* Comparativos: boxplots de velocidad, histogramas de distancia, correlación velocidad–desnivel.

**Evaluación**

* Verificar que los gráficos respondan a las preguntas de negocio.
* Ajustar filtros e interacciones en Power BI para asegurar claridad y utilidad.

**Despliegue**

* Publicar el informe Power BI con enlaces de acceso.
* Documentar pasos y código en un repositorio para reproducción.

**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.