

Informe Final – UrbanFlow: Divvy Trips + Weather + ML

Integrantes:

Jorge Alberto Roncancio

Daniel Sebastián Bello Hernández

1. Introducción

UrbanFlow es un proyecto orientado al análisis y predicción de la duración de viajes de bicicletas compartidas (Divvy), incorporando variables meteorológicas. El objetivo principal fue identificar patrones temporales, evaluar el impacto del clima sobre la demanda y construir un modelo base de predicción de duración de viaje, con la posibilidad de extenderlo a futuras optimizaciones locales y visualizaciones interactivas.

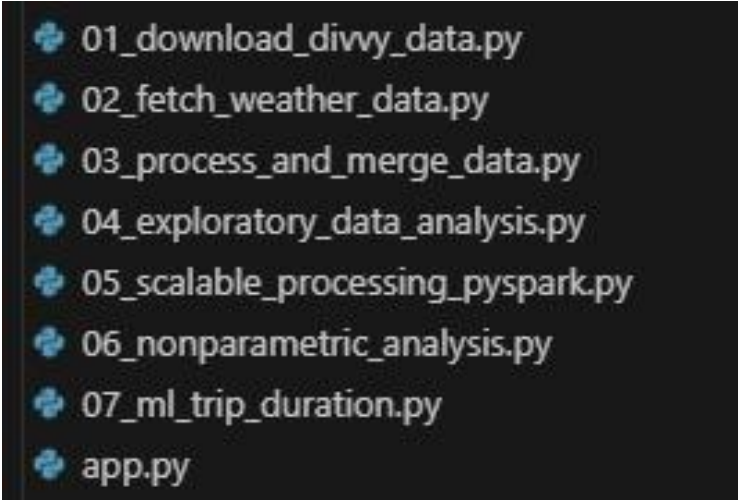
2. Objetivos

- Comprender patrones temporales de uso de bicicletas (hora, día, fin de semana).
- Analizar la influencia de condiciones climáticas (temperatura, precipitación, viento).
- Construir un modelo predictivo para la duración de los viajes (trip_duration_minutes).
- Presentar resultados en un dashboard interactivo (Streamlit).

3. Código Fuente

Se aplicaron buenas prácticas como separación clara de responsabilidades (ingesta, limpieza, EDA, ML, dashboard), uso de Parquet para eficiencia, y caching en Streamlit para optimizar la ejecución.

Repositorio: [GitHub - UrbanFlow](#)



```
01_download_divvy_data.py
02_fetch_weather_data.py
03_process_and_merge_data.py
04_exploratory_data_analysis.py
05_scalable_processing_pyspark.py
06_nonparametric_analysis.py
07_ml_trip_duration.py
app.py
```

Consideraciones de buenas prácticas:

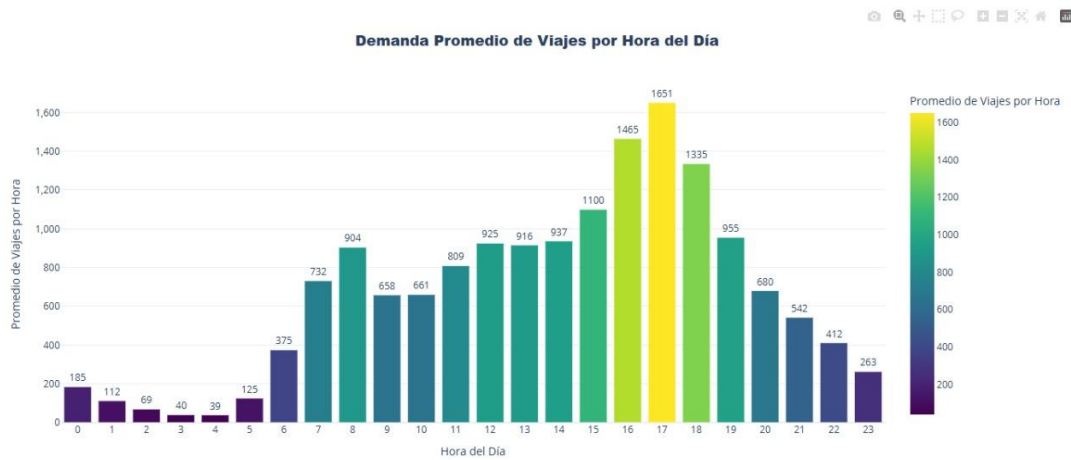
- Separación de scripts según función.
- Uso de Parquet para eficiencia en lectura/escritura.
- Estructura de carpetas organizada para datos, scripts y salidas.
- Variables y funciones nombradas de manera descriptiva.

4. Visualizaciones y Predicciones

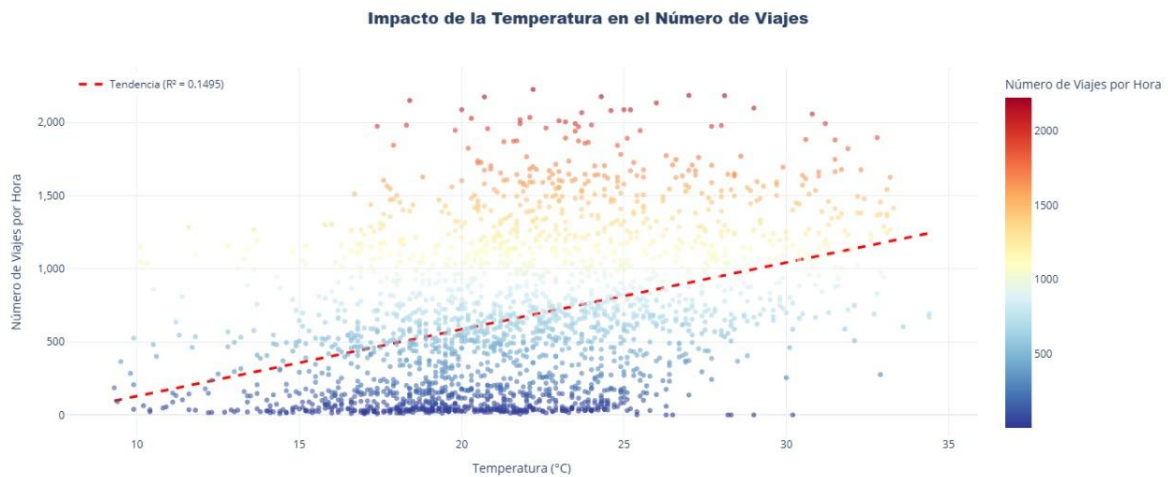
4.1 Análisis Exploratorio

Se generaron gráficos de:

- Distribución de viajes por hora y día



- Viajes según tipo de usuario (miembro vs casual).
- Relación entre duración de viaje y temperatura/precipitación.



4.2 Modelos Predictivos

Target: trip_duration_minutes

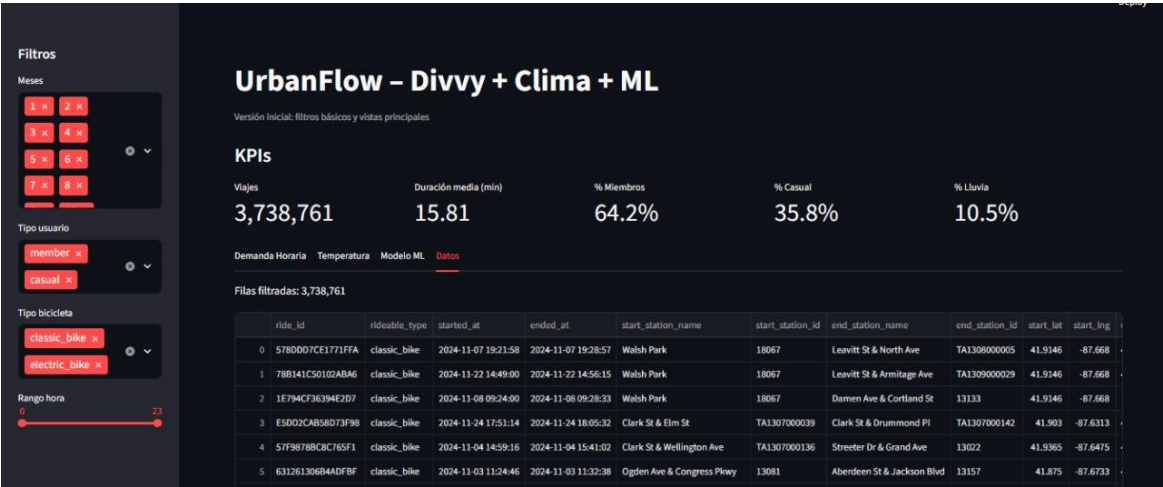
Modelos evaluados:

- Baseline (media)
- LinearRegression
- RandomForestRegressor
- HistGradientBoostingRegressor

4.3 Dashboard Streamlit

El dashboard interactivo permite:

- Filtrar por meses, tipo de usuario, tipo de bicicleta, rango horario.
- Visualizar KPIs: total viajes, duración media, % miembros vs casual, % lluvia.
- Explorar gráficamente demanda horaria y relación viajes vs temperatura.



5. Optimización Local

- Particionado de datos por año/mes.
- Compresión Snappy de Parquet.
- Selección de columnas específicas para ML.
- Scripts reutilizables para refresco y re-entrenamiento de modelos.

6. Análisis Ético y Recomendaciones

Aunque los datos utilizados no contienen información sensible, se evaluó la potencial exposición de:

- Diferencias de duración entre usuarios miembros y casuales.
- Bias en predicciones según condiciones climáticas.

Recomendaciones éticas:

- Mantener anonimización de usuarios.
- Revisar posibles sesgos en predicciones que afecten decisiones de negocio o planificación de estaciones.

7. Conclusiones

- UrbanFlow proporciona un pipeline completo: ingesta, limpieza, EDA, ML y dashboard.
- Se obtuvieron resultados preliminares de predicción con $R^2 \sim 0.14$, útiles como baseline para iteraciones futuras.
- Visualizaciones y KPIs permiten identificar patrones de uso y evaluar impacto del clima.
- El proyecto está listo para ampliaciones: mapas de estaciones, tuning de modelos, validación temporal y optimización de features.