FASE 0 — Presentación, Alineación y Viabilidad del Proyecto

SmartTrafficFlow AI: Predicción Inteligente de Congestión en Madrid

1. Título provisional del proyecto

SmartTrafficFlow AI: Sistema Predictivo de Congestión de Tráfico en Madrid mediante Big Data y Machine Learning.

2. Elevator pitch

La movilidad en Madrid dispone de datos abundantes de sensores, clima y calendario, pero se explotan sobre todo de forma descriptiva y en tiempo real. Proponemos crear SmartTrafficFlow AI, una solución que predice la congestión a corto plazo (p.ej., 30–60 minutos) que se visualize en un dashboard web. Se parte de un dataset público y real de Madrid (MTD), y dejamos abiertas varias opciones de alcance (básico, por zonas, deep learning, grafos, o analítica + baseline) para escoger la más adecuada al equipo.

3. El tema y el problema

Situación actual. Las ciudades publican datos de tráfico e incidencias (muchas veces en formatos como DATEX2), y existen portales con recursos abiertos (DGT, portales municipales, Kaggle). Sin embargo, la mayoría de las soluciones consumidas por el público son descriptivas y no anticipan la congestión futura de manera personalizada al contexto (clima, calendario, patrón horario).

Impacto.

- Atascos inesperados → pérdida de tiempo y coste económico.
- Mayor contaminación por tráfico parado.
- Planificación subóptima de logística y desplazamientos diarios.

Por qué es importante.

- Anticipar picos de tráfico permite planificar mejor salidas/rutas y coordinar servicios urbanos.

4. Objetivo principal

Construir un sistema predictivo (MVP web) que anticipe niveles de congestión/incrementos de intensidad en puntos de la ciudad con 30–60 minutos de antelación, usando datos históricos de sensores de tráfico, clima y calendario, y lo exponga en un dashboard web sencillo para su consulta. (Dataset base: Madrid Traffic Dataset).

5. Alcance IN / OUT

Con opciones para votar en equipo

IN (comunes a todas las opciones)

- Ingesta de CSV del **Madrid Traffic Dataset (MTD)**.
- Limpieza, alineación temporal y feature engineering (hora, día, festivo, clima).
- Despliegue web del MVP (Streamlit/HF Spaces/Render).
- Dashboard con: histórico, estado actual (si procede) y **predicción a corto plazo**.

OUT (fase inicial)

- **App móvil nativa** (iOS/Android).
- Navegación/ routing en tiempo real tipo “GPS”.
- Integraciones complejas con APIs en vivo de DGT/terceros (posible fase futura).

Opciones de alcance:

- **A) Predicción sencilla por sensor (baseline ML clásico).** Modelos: Regresión lineal, RandomForest, XGBoost. Horizonte: 30–60 min. Ventaja: Alta viabilidad y control de tiem**pos.**

- **B) Predicción por zonas (clustering + modelo por zona).** Añade: K-Means/Clustering, mapa por áreas.

- **C) Forecasting con secuencias (LSTM/GRU/Transformer).** Uso de: ficheros de temporal_sequences del MTD.

- **D) Modelo espacial con grafos (GNN)** usando MTD_adj_matrix.npy**.** Se valora como futura iteración.

6. Usuarios / partes interesadas

Usuarios: Conductores urbanos, empresas de logística, técnicos de movilidad, ciudadanía.

Decisores: Administraciones locales y áreas de movilidad/infraestructura.

Beneficiarios: Ciudadanía, empresas, servicios municipales (planificación).

Potencialmente afectados: Herramientas descriptivas actuales (competencia funcional).

7. Propuesta de solución

Incluye capas de ingestión, procesamiento, modelos (RF, XGBoost), dashboard y API opcional.

8. Datos

Fuente principal (confirmada):

Madrid Traffic Dataset (MTD): ~554 sensores (completo) durante ~30 meses (jun-2022 a nov-2024) + subset de 300 sensores/17 meses, con clima, calendario, localización y matriz de adyacencia incluida.

Hay diferentes formatos como CSV/NPY/H5, con secuencias listas para DL. Licencia: pública para investigación (Mendeley Data).

Fuentes alternativas/complementarias (opcionales):

DGT — Punto de Acceso Nacional de Tráfico y Movilidad (incidencias, cámaras, DATEX2 v3.6, etc.). Nota: mayor complejidad técnica por formato y acceso.

Open Data Barcelona (movilidad/servicios; útil como inspiración o para comparativas, no tan específico en tráfico vehicular).

Kaggle (tráfico/congestión) para benchmarking de modelos/ideas.

Tipo de datos: Series temporales multivariantes (intensidad, clima), tabular, geoespacial (lat/lon).

Riesgos de acceso/licencias: bajos al ser datos públicos; revisar términos de uso si mezclamos con DGT/APIs en vivo.

9. Viabilidad técnica

Fácil/medio:

- Ingesta CSV del MTD, feature engineering temporal, baseline con RF/ XGBoost, dashboard web.

Complejo:

- Manejo de grandes volúmenes (optimizar memoria / muestrear sensores).
- Modelos secuenciales (C) o GNN (D) por complejidad matemática/infra.
- Integraciones en tiempo real con DGT (si se consideraran).

Stack previsto: Python, Pandas, NumPy, Scikit-learn, XGBoost, Streamlit.

Opcional: FastAPI.

Para opciones C/D: PyTorch/TensorFlow yo Geometric.

Incertidumbres: métricas alcanzables por sensor/zona; heterogeneidad entre sensores; *data leakage* temporal si no cuidamos splits.

10. Riesgos y Plan B

Riesgo	Impacto	Mitigación / Plan B
Dataset voluminoso → lentitud	Retrasos en EDA/entrenos	Usar subset MTD (300 sensores) o muestrear zonas/ventanas
Métricas flojas del modelo	Baja utilidad	Ajustar horizonte a 30 min, cambiar a clasificación de niveles, feature climática/calendario
Complejidad de modelos C/D	Sobrecarga de equipo/tiempo	Empezar por A y evolucionar a B, C/D como "fase 2"
wvivo (DGT)	Sobreesfuerzo en Fase 0/1	Mantener OUT para MVP, evaluar más adelante

11. Reparto de tareas

REPARTO DE TAREAS ORIENTATIVO.

- **Data (Ingesta/EDA/Features):** Preparar datasets por sensor/zona, validación de splits temporales. **DONOVAN, JOKIN**
- **ML (Modelado/Evaluación):** Baselines (A), clustering y modelos por zona (B) o secuencias (C). **JOKIN, FRANCISCO**
- **Platform/BI (Web & Dashboard):** Streamlit + despliegue (HF Spaces/Render), visualizaciones y KPIs. **DONOVAN, MARCOS**
- **MLOps/Calidad (opcional):** Scripts de training reproducibles, control de versiones (DVC opcional), métricas. **FRANCISCO, MARCOS**
- **PM/Docs:** Documentación, lectura de resultados, redacción y pitch final. **JOKIN**

12. Próximos pasos (7 días)

1. Explorar y organizar el dataset MTD (Día 1–2)
2. Preparar el entorno de trabajo y repositorio (Día 1)
3. Realizar un EDA inicial sobre los sensores seleccionados (Día 2–3)
4. Construir el primer baseline predictivo (Día 3–4)
5. Preparar el primer prototipo del dashboard web (Día 4–5)
6. Reunión de decisión técnica en base a resultados iniciales (Día 6–7)