# **Informe de Avance – Unidad 1**

**Proyecto:** Predicción de aptitud vehicular en San Isidro, Lima  
**Metodología:** CRISP-DM (Fases 1 a 3)  
**Integrantes:** [Nombres de integrantes]  
**Fecha del informe:** [dd/mm/aaaa]

***NOTA***: Puedes seguir esta plantilla o ir al cuaderno júpiter [MD04 GP\_GA\_Unidad\_1.ipynb](https://drive.google.com/file/d/1Qzyf73RY0TR7jdua6Qgs02VIhMdR66kx/view?usp=drive_link)

## **Fase 1 – Business Understanding**

### **🎯1. Objective**

Construir un modelo que prediga si un automóvil estará **APPROVED** (APPROVED = 1, DISAPPROVED = 0) para circular por las calles de **San Isidro, Lima** usando características del vehículo (sin usar sus emisiones)

### **Success criteria (SC)**

* **Técnico:** alcanzar un F1-score ≥ 0.80 en el conjunto de validación.
* **De negocio:** contribuir al monitoreo sistemático de los vehículos no aptos en circulación y generar evidencia que permita a la Municipalidad de San Isidro diseñar estrategias orientadas a reducir progresivamente su número.

## **Fase 2 – Data Understanding**

### **📥 2. Data collection**

El dataset proviene de registros de campañas de inspección vehicular en San Isidro (*CAMPANADESENSIBILIZACIONRESPIRAAIRELIMPIO*). Incluye variables como:

* Datos del vehículo: Vehicle type, Fuel type, Vehicle year, Vehicle age.
* Resultados técnicos: VTI CO percVol, VTI HC ppm, VTI CO+CO2 min perc.
* Límites normativos: MPL CO percVol, MPL HC ppm, MPL CO+CO2 min perc.
* Estado final: State (APPROVED / DISAPPROVED).

### **📊 3. Descriptive analysis**

### **3.1. Análisis descriptivo básico**

* Resumen estadístico de todas las variables (describe()).
* Tipos de variables (numéricas, categóricas, fechas).
* Conteo de valores únicos en categóricas (ej. Vehicle type, Fuel type).
* Distribución de variables numéricas (media, mediana, desviación estándar).

### **3.2. Análisis descriptivo gráfico**

* Histogramas y boxplots para variables numéricas (ej. Vehicle year, Vehicle age).
* Gráficos de barras para categóricas (ej. Fuel type, Vehicle type).
* Gráfico de conteo de la variable objetivo State (Approved vs Disapproved).

### **3.3. Análisis exploratorio con estadísticas**

* Correlaciones entre variables numéricas (heatmap de correlaciones).
* Cruces de variables categóricas con la variable objetivo (State) mediante tablas de contingencia o groupby.
* Pruebas estadísticas simples (Chi-cuadrado para categóricas vs State, t-test o ANOVA para numéricas vs State).

### **3.4. Diagnóstico de calidad inicial**

* **Valores faltantes**: identificar y cuantificar (df.isnull().sum()).
* **Posibles inconsistencias**:  
  + Años fuera de rango (ej. vehículos con Vehicle year futuro).
  + Edades negativas o mayores a 50 años.
* **Distribución de la variable objetivo (State)**: revisar balance entre APPROVED y DISAPPROVED.

### **3.5. Perfil comparativo (Approved vs Disapproved)**

* Comparar medias/medianas de variables numéricas entre aprobados y desaprobados.
* Comparar distribuciones de categóricas (ej. mayor tasa de desaprobación en vehículos a diésel).

### **3.6 Conclusión preliminar**

El dataset es adecuado para análisis, pero requiere procesos de limpieza y transformación.

## **Fase 3 – Data Preparation**

### **🧹 4. Data cleaning**

* **Faltantes:** imputados con mediana (numéricas) o categoría “Unknown” (categóricas).
* **Duplicados:** eliminados con drop\_duplicates.
* **Outliers:** tratados con winsorización o recorte de valores extremos.
* **Inconsistencias:** unificación de formatos de combustible (ej. “Diesel” vs. “DIESEL”).

### **🔧 5. Data transformation**

* **Codificación de categóricas:** Vehicle type y Fuel type convertidos mediante *One-Hot Encoding*.
* **Feature engineering:**
  + Cálculo de Vehicle\_age\_calc = 2025 – Vehicle year.
  + Creación del ratio CO\_CO2\_ratio para evaluar eficiencia de combustión.
* **Integración:** posibilidad de unir con datos municipales (ej. multas o historial de revisiones).
* **Reducción:** exploración con PCA para análisis exploratorio (opcional).

### **✅ 6. Data validation**

* Se realizó un **EDA post-limpieza** con histogramas y matriz de correlación.
* Se depuraron variables redundantes y con alta colinealidad.
* Se confirmó que las variables predictoras están listas para modelado.
* La variable objetivo State fue recodificada como binaria (State\_bin: 1=APPROVED, 0=DISAPPROVED).
* **Se obtuvo el entregable unidad 1 (v3):**Un **dataset pre-modelado** (vehiculos\_sanisidro\_v3.csv), limpio, transformado, validado y documentado.

## **Conclusión**

La **Unidad 1** permitió:

1. Definir claramente el **objetivo de negocio y analítico**.
2. Comprender el dataset inicial mediante **descripción y diagnóstico**.
3. Aplicar procesos de **limpieza, transformación y validación** de los datos.

El resultado es un dataset confiable y documentado, **listo para la fase de modelado (Unidad 2)**, en la cual se entrenarán algoritmos de clasificación para predecir la aptitud vehicular.

## **Anexo**:

**Introducción o descripción del caso**

**1. Contexto global** En las últimas décadas, la contaminación ambiental se ha consolidado como uno de los principales problemas que afectan la calidad de vida en las ciudades a nivel mundial. Organismos internacionales como la Organización Mundial de la Salud [OMS] y la Organización Panamericana de la Salud [OPS] han advertido que los niveles de material particulado y gases contaminantes en zonas urbanas superan ampliamente los límites recomendados, generando un impacto directo en la salud pública y en la sostenibilidad ambiental. Dentro de las principales fuentes de contaminación, el parque automotor se ubica como responsable de entre el 70 % y el 90 % de las emisiones en áreas metropolitanas, lo cual convierte al tránsito vehicular en un factor crítico a controlar (OMS, 2024; OPS, 2023).

**2. Contexto del problema** En este escenario, las ciudades se ven obligadas a buscar alternativas innovadoras para mejorar la calidad del aire y optimizar la movilidad. El transporte vehicular no solo contribuye a la contaminación atmosférica, sino también a la congestión y al deterioro de la salud de los habitantes. A nivel técnico, los sistemas tradicionales de fiscalización dependen de revisiones manuales y reactivas, que suelen ser costosas y poco eficientes. Frente a ello, la ciencia de datos y el aprendizaje automático han demostrado ser herramientas prometedoras al permitir anticipar riesgos, clasificar vehículos y apoyar políticas públicas de sostenibilidad urbana (World Bank, 2022; Liu et al., 2024).

**3.1 Estado del arte internacional** Diversos estudios internacionales han explorado soluciones desde la perspectiva de la ciencia de datos. Zhou et al. (2021) en Asia y Park et al. (2022) en Europa demostraron que los modelos de predicción aplicados al transporte permiten anticipar incumplimientos de normas de emisiones y optimizar procesos de inspección. En América Latina, Fernández et al. (2022) en México y Gómez y Ríos (2023) en Colombia reportaron resultados positivos en la aplicación de técnicas de minería de datos para el control vehicular, evidenciando la utilidad de algoritmos de clasificación en la gestión ambiental. Estos hallazgos refuerzan que el aprendizaje automático constituye una estrategia eficaz para abordar los desafíos de la fiscalización vehicular.

**3.2 Estado del arte nacional/local** En el Perú, la aplicación de enfoques predictivos aún es incipiente. Luna y Salazar (2023) han señalado la necesidad de modernizar los sistemas de fiscalización vehicular mediante enfoques basados en datos, aunque las experiencias locales son limitadas. Las autoridades se concentran en cumplir con la Ley General de Transporte y Tránsito Terrestre (Ley N.° 27181) y con los Estándares de Calidad Ambiental fijados por el Ministerio del Ambiente (MINAM, 2020). Sin embargo, no se han identificado estudios que apliquen modelos de minería de datos o aprendizaje automático en campañas municipales, lo que evidencia una brecha entre la normativa vigente y las herramientas tecnológicas disponibles.

**4. Vacío en el conocimiento** Sin embargo, pese a los avances internacionales en el uso de modelos de predicción vehicular y a las investigaciones nacionales centradas en la calidad del aire, en el Perú no se han desarrollado estudios que apliquen aprendizaje automático para anticipar, a partir de las características técnicas de los vehículos, su aprobación o desaprobación en la circulación urbana.

**5. Propósito / Objeto de estudio** En este marco, la presente investigación tuvo como propósito **construir un modelo predictivo que determine si un automóvil es aprobado o desaprobado para circular en el distrito de San Isidro**, utilizando como insumo sus características técnicas y normativas vigentes. El estudio se contextualiza en las campañas municipales de inspección vehicular realizadas en el año 2025, en cumplimiento de la Ley N.° 27181 y los estándares establecidos por el MINAM (2020). De esta manera, se busca ofrecer una herramienta basada en datos que complemente los procesos actuales, predominantemente manuales y reactivos, optimizando la fiscalización y favoreciendo la prevención temprana.

**6. Justificación** La investigación se justifica en cuatro dimensiones. En el plano **teórico**, aporta evidencia sobre la aplicación de minería de datos en el control vehicular y la calidad del aire, un campo poco explorado en el Perú. En el plano **práctico**, ofrece a la Municipalidad de San Isidro una herramienta predictiva que optimice la fiscalización y reduzca costos operativos. En el plano **metodológico**, aplica la metodología CRISP-DM y técnicas de aprendizaje automático, asegurando un proceso estructurado y replicable en otras municipalidades. Finalmente, en el plano **social**, contribuye a la protección de la salud de la población y al cumplimiento de los Objetivos de Desarrollo Sostenible relacionados con ciudades sostenibles y acción climática (ONU, 2023).

**7. Pregunta, objetivos e hipótesis** En consecuencia, la investigación respondió a la pregunta: ¿es posible construir un modelo predictivo confiable que, a partir de las características técnicas de los vehículos, determine si un automóvil estará aprobado para circular en San Isidro? En coherencia, la investigación tuvo como **objetivo general** construir un modelo predictivo que determine, a partir de las características técnicas de los vehículos, si un automóvil será aprobado o desaprobado para circular en el distrito de San Isidro, en el marco de las campañas municipales de inspección vehicular del año 2025. Para alcanzarlo, se plantearon como **objetivos específicos**: (1) analizar y comprender el problema de negocio y los datos disponibles en campañas de inspección, definiendo las variables relevantes de estudio; (2) preparar y transformar el conjunto de datos vehiculares, aplicando técnicas de limpieza, integración y selección de variables; (3) desarrollar y entrenar modelos predictivos de clasificación, seleccionando el de mejor desempeño según métricas de validación; y (4) evaluar y documentar el modelo final, verificando su precisión y utilidad práctica, así como desplegarlo en una aplicación web/móvil que facilite su uso por parte de los inspectores municipales en los procesos de fiscalización vehicular.

Esa es la **ruta lógica del embudo**:

1. **Tema / Contexto global** → “Qué está pasando en el mundo” (ej. OMS, OPS, OIT, ONU).
2. **Problema / Contexto específico** → “Cómo se manifiesta el problema en el área de estudio” (ej. transporte vehicular y contaminación en ciudades).
3. **Antecedentes** → Internacionales y nacionales/locales, con evidencia científica reciente.
4. **Vacío en el conocimiento** → Breve, concreto: “Lo que falta investigar”.
5. **Objeto de estudio / Propósito** → Tu caso específico (modelo predictivo de clasificación).
6. **Justificación** → Relevancia en planos teórico, metodológico, práctico, social y otros.
7. **Objetivos e hipótesis** → Pregunta de investigación, objetivo general, objetivos específicos e hipótesis según el caso.