Tecnológico de Monterrey

Análisis exploratorio de datos en el marco de desarrollo de DECODE-EV

Nombre del Autor	Código
Henry Junior Aranzales Lopez	A01794020
Jorge Arturo Hernández Morales	A01794908
Luis Alejandro González Castellanos	A01795481

Dra. Grettel Barceló Alonso y Dr. Luis Eduardo Falcón Morales

Instituto Tecnológico de Monterrey
Programa de IA Aplicada

TC5035: Proyecto integrador
Septiembre 28, 2025
IBM

Contenido

Pág.

1. Análisis exploratorio de datos

El presente documento proporciona una caracterización técnica exhaustiva del conjunto de datos utilizado en el proyecto DECODE-EV, el cual comprende registros operacionales de un bus eléctrico de fabricación colombiana operado por Superpolo S.A.S. El dataset incluye datos temporales de alta frecuencia del sistema de gestión de batería, así como especificaciones completas de la arquitectura de red Controller Area Network (CAN) del vehículo. Esta caracterización establece las bases para el desarrollo de algoritmos de inteligencia artificial orientados a casos puntuales como detección de anomalías y mantenimiento predictivo en sistemas de transporte eléctrico urbano.

2. Introducción y contexto tecnológico

Los sistemas de comunicación vehicular basados en Controller Area Network (CAN) constituyen la columna vertebral de la electrónica automotriz. En un bus eléctrico moderno, múltiples controladores electrónicos (ECUs) interactúan sobre diferentes redes CAN, cada una con responsabilidades específicas: propulsión, gestión de baterías, carrocería y sistemas auxiliares.

Los archivos DBC y BLF representan dos caras complementarias del mismo fenómeno: uno define la gramática de la red (DBC), mientras el otro captura la narrativa en tiempo real de la comunicación (BLF). A partir de ellos, es posible decodificar, validar y analizar la dinámica de la red vehicular.

2.1. Naturaleza de los DBC

Un archivo DBC funciona como un diccionario técnico. En él se establece cómo interpretar cada trama CAN: qué identificador corresponde a qué mensaje, qué campos dentro de la trama corresponden a señales físicas y cómo deben escalarse esos valores crudos para obtener magnitudes reales.

Tabla 2-1

Sintaxis esencial de los archivos DBC

Token	Descripción
BO_	Mensaje con ID, tamaño y transmisor
SG_	Señal: posición de bit, longitud, signo, endianness, factor y offset
CM_	Comentarios para documentación
VAL_	Tablas de valores (enumeraciones)
BA_DEF_ / BA_	Definición y asignación de atributos (ciclo, timeout, etc.)

Fuente: Elaboración propia

El DBC no es solo un archivo de configuración, sino un modelo formal de interpretación. Sin él, un registro BLF sería apenas un conjunto de bytes. Con él, esos bytes se convierten en temperaturas de celdas, niveles de carga, par motor solicitado, o estado de puertas.

Se deben contemplar los siguientes conceptos también:

- **Endianness**: la interpretación de bits puede cambiar totalmente el valor de una señal. En Motorola (big-endian) la numeración de bits difiere de Intel (little-endian).

- **Escalado**: mediante el factor y offset, una señal cruda como 0x0A puede traducirse en una temperatura de 25 °C o en un voltaje de 3.3 V.
- **Multiplexación**: un mismo mensaje puede contener distintas disposiciones de señales dependiendo de un campo multiplexor, optimizando ancho de banda.

2.2. Naturaleza de los BLF

El formato BLF captura la vida cotidiana de la red: cada interacción, cada mensaje, cada error. A diferencia de un CSV, es un formato binario compacto y eficiente, preservando no solo los datos, sino también la temporalidad precisa y la información del canal físico.

Tabla 2-2
Elementos típicos de un BLF

Objeto	Significado
CAN Message	Trama estándar de datos o remota
CAN Error	Registro de un error detectado en la red
Overload	Indica saturación temporal
Markers	Anotaciones o eventos del usuario
Objeto	Significado

Fuente: Elaboración propia

Los BLF suelen convertirse a CSV, MDF o Parquet para análisis masivo. Sin embargo, la decodificación en flujo aplicando el DBC sobre el BLF original garantiza la trazabilidad y la calidad de los datos. Por ejemplo, con respecto a un archivo .CSV, los BLF tienen las siguientes ventajas":

- **Eficiencia**: reducción de tamaño hasta 10 veces.
- **Precisión**: evita redondeo de tiempos.
- **Metadatos**: incluye información de hardware, canales y configuración de medición.

2.3. Comportamiento de la red CAN

La red CAN se comporta como un foro democrático, donde todos los nodos pueden hablar, pero la prioridad la define el identificador del mensaje.

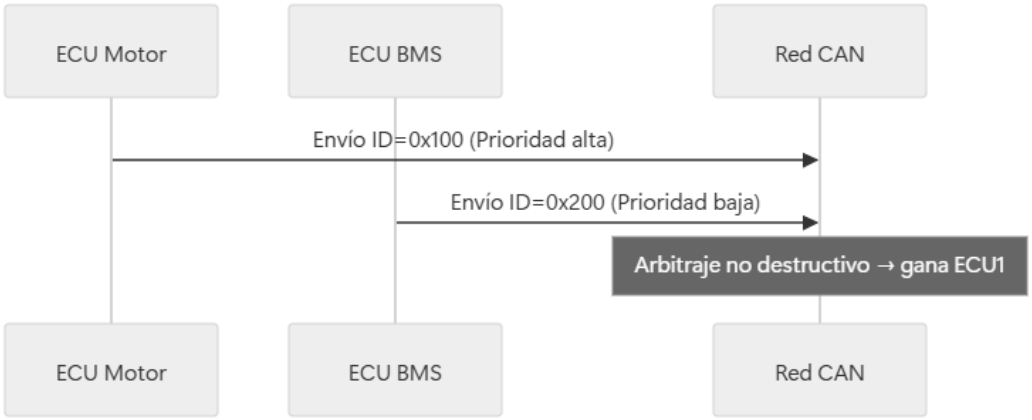
2.3.1. Dinámica del arbitraje

Los IDs más bajos tienen mayor prioridad. Cuando dos ECUs transmiten a la vez, gana la que transmite un bit dominante sobre un recesivo. Esto asegura que los mensajes críticos (ej. frenos, seguridad) nunca se retrasen.

La CAN funciona como un mecanismo de resolución de conflictos donde los bits dominantes prevalecen. Esto asegura que los mensajes más críticos siempre sean transmitidos primero.

Los mensajes se publican con ciclos definidos. La suma de estas publicaciones determina la carga del bus. Superar el 60% de carga puede generar retrasos y pérdida de determinismo.

Figura 2-1
Representación de la dinámica de arbitraje



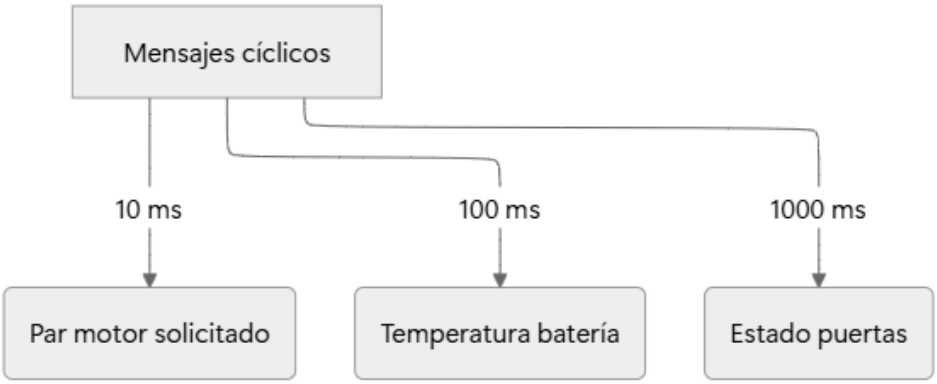
Fuente. Elaboración propia.

La carga del bus (Bus Load) se calcula como:

$$\boldsymbol{Carga}\ (\%)\ =\ (\boldsymbol{Tiempo\ total\ de\ transmisión}\ /\ \boldsymbol{Tiempo\ de\ observación})\ *\ \mathbf{100}$$

Valores mayores al 60% implican riesgo de latencia.

Figura 2-2
Estructura temporal y carga



Fuente. Elaboración propia.

2.4. J1939 y la semántica extendida

En buses eléctricos, la norma J1939 define mensajes de 29 bits que incluyen prioridad, PGN y direcciones de origen. Esto permite interoperabilidad entre distintos fabricantes de ECUs y facilita la integración con diagnóstico (DM1, DM2).

La norma establece los siguientes tres conceptos clave:

- **PGN**: identifica la función (ej. PGN 61444 para velocidad de motor).
- **SPN**: parámetros individuales (ej. SPN 190 = Engine Speed).
- **Diagnóstico**: mensajes DM1 y DM2 reportan DTCs activos e históricos.

A continuación, se expone la estructura general del pipeline detallado de análisis para los datos de la red CAN:

Figura 2-3
Estructura para el análisis de los datos de la red CAN

Fuente. Elaboración propia.

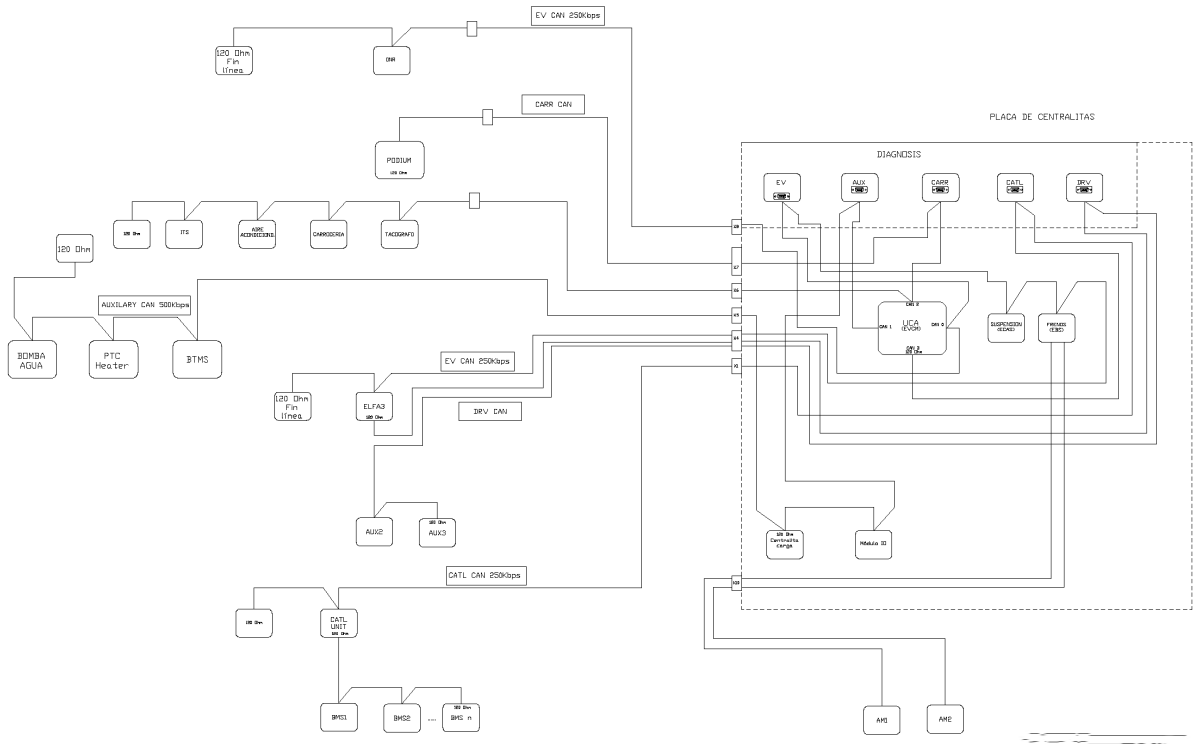
3. Configuración redes CAN en el Vehículo

En la siguiente figura, se muestra cómo se interconectan las distintas unidades electrónicas (ECUs) del chasis mediante líneas CAN, que son el sistema nervioso del autobús. El objetivo es asegurar la comunicación confiable y jerárquica entre módulos de control y sensores en tiempo real.

El sistema cuenta con 4 canales CAN principales, con diferentes velocidades y aplicaciones:

Figura 3-1

Esquemático Red de comunicaciones CAN (Control Area Network)



Fuente. (Superpolo,2025)

Tabla 3-1

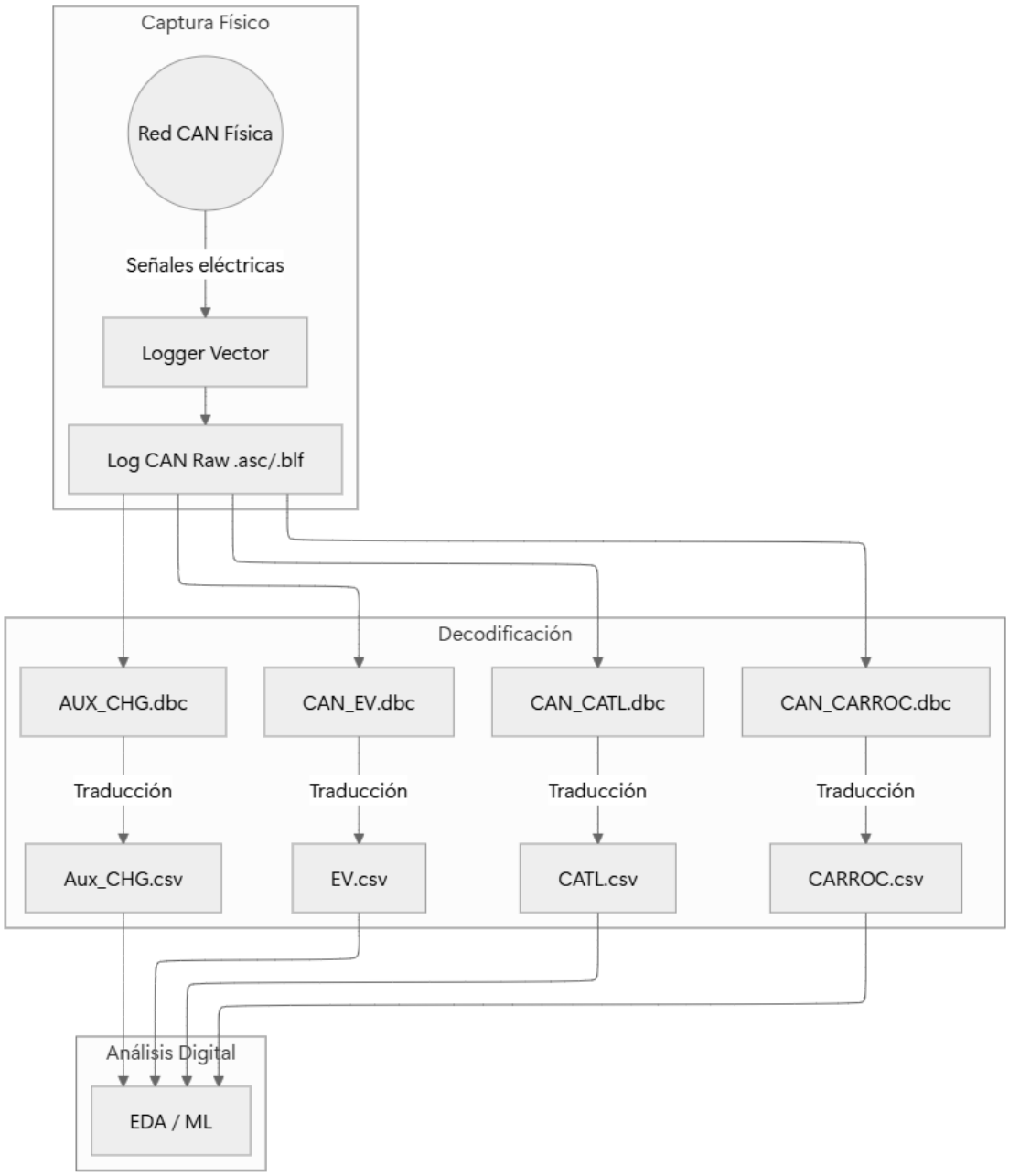
Atributos de las líneas CAN presentes en el vehículo

Línea CAN	***Nombre lógico***	***Función principal y conexiones***	***Velocidades***
CAN EV (CAN PRICIPAL)	*Electric Vehicle*	*ECU principal (TTControl), Inversor de tracción, conector de diagnóstico*	*250 kbps*
CAN DRIVE (Es privada no tenemos acceso)	*Drive / Conducción*	*Dirección electrónica, sensores de rueda, ECU EBS, sensor de freno, pedal acelerador*	*250 kbps*
CAN AUX	*Auxiliares*	*Refrigeracion, gestión de carga, ventiladores, sensores térmicos, actuadores auxiliares.*	*500 kbps*
CAN CATL	*Batería (BMS CATL)*	*BMS (CATL), ECU principal, sensores térmicos y de voltaje. PowerBOX y Alta Potencia.*	*250 kbps*
CAN CARR	*Carrocería*	*Pantallas, ITS, controlador de puertas, aire acondicionado, tacógrafo, PODIUM.*	*250 kbps*

Fuente: Elaboración propia

3.1. Estructura general del ecosistema de datos

Figura 3-2
Diagrama general del ecosistema de datos



Fuente. Elaboración propia

3.1.1. Captura Físico (Nivel Hardware):

El proceso comienza en la red física CAN del bus eléctrico, donde las señales eléctricas viajan entre las distintas ECUs y componentes. El hardware Logger Vector funciona como interfaz que captura estas señales eléctricas binarias y las convierte en archivos de registro (.asc/.blf) que preservan toda la información de los mensajes CAN sin interpretación, manteniendo timestamps, IDs y datos hexadecimales.

3.1.2. Decodificación (Nivel Interpretación):

Los archivos de registro raw son procesados simultáneamente con múltiples archivos DBC (Database CAN), cada uno especializado en un subsistema del vehículo. Esta fase es crítica porque transforma datos binarios incomprensibles en valores ingenieriles con significado. Los archivos DBC contienen diccionarios que definen cómo interpretar cada byte, aplicando factores de conversión, offset y unidades específicas para cada señal.Por ejemplo:

- CAN_EV.dbc decodifica los mensajes relacionados con el tren motriz eléctrico
- CAN_CATL.dbc interpreta los mensajes del sistema de gestión de baterías (BMS)
- CAN_CARROC.dbc traduce la información de la carrocería del vehículo
- AUX_CHG.dbc maneja los datos relacionados con sistemas de carga auxiliar

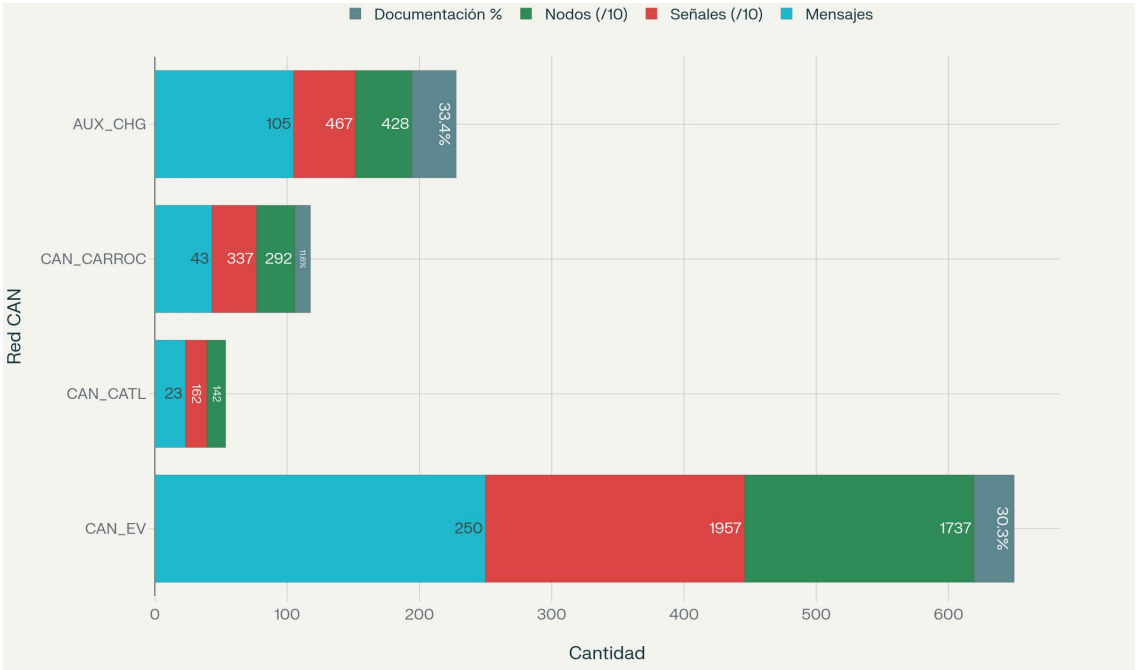
El resultado de esta decodificación son archivos CSV donde cada columna representa una señal específica con valores ya convertidos a unidades de ingeniería (voltios, amperios, grados, etc.), facilitando así su análisis posterior.

3.1.3. Análisis Digital (Nivel Procesamiento):

Esta fase final recibe todos los archivos CSV generados y los integra en un pipeline de procesamiento que incluirá análisis exploratorio de datos (EDA) y posiblemente algoritmos de machine learning. Aunque esta sección aparece truncada en el diagrama seleccionado, sugiere que el propósito final es realizar análisis avanzados sobre los datos del vehículo para optimización, diagnóstico o desarrollo.

Figura 3-3

Arquitectura CAN del BUS



Fuente. Elaboración propia

En la anterior imagen se muestra la distribución de complejidad de las redes CAN, mostrando que CAN_EV es la red más crítica con 1,957 señales, pero CAN_CATL carece completamente de documentación.

4. Resultados del análisis exploratorio

4.1. Descripción del Proyecto:

Se desarrolló un sistema comprehensivo de análisis exploratorio de datos (EDA) especializado en el procesamiento y análisis de información proveniente de sistemas de diagnóstico vehicular. El proyecto se centró en la creación de una herramienta robusta capaz de manejar datos de redes CAN (Controller Area Network) y archivos BLF (Binary Logging Format) para aplicaciones de mantenimiento predictivo y diagnóstico automotriz.

4.2. Componentes Técnicos Implementados:

El sistema incluye un notebook interactivo de Jupyter con capacidades modulares que abarca:

1. **Sistema de Configuración Personalizada**: Implementación de configuraciones flexibles para el procesamiento de archivos BLF específicos, incluyendo rutas personalizables, archivos DBC para decodificación de mensajes CAN, y filtros temporales avanzados.
2. **Carga Multi-Fuente de Datos**: Desarrollo de un sistema jerárquico de carga que prioriza configuraciones personalizadas BLF, implementa búsqueda automática de archivos, utiliza datos CSV de respaldo y genera datasets sintéticos para validación.
3. **Análisis Estadístico Avanzado**: Implementación de análisis descriptivo completo con evaluación de normalidad mediante tests Shapiro-Wilk y Kolmogorov-Smirnov, cálculo de medidas de asimetría y curtosis, y optimización computacional para datasets de gran volumen.
4. **Evaluación de Calidad de Datos**: Sistema de evaluación multidimensional que incluye análisis de completitud, verificación de consistencia, detección de duplicados y análisis de rangos válidos adaptados a las características específicas de datos vehiculares.
5. **Análisis Correlacional**: Implementación de matrices de correlación de Pearson con visualizaciones especializadas, identificación automática de correlaciones significativas y generación de scatter plots para relaciones bivariadas.
6. **Detección de Valores Atípicos**: Sistema multi-método que incorpora técnicas IQR, Z-Score y métodos de aislamiento estadístico, con visualizaciones mediante boxplots y análisis contextual para el dominio vehicular.
7. **Análisis Temporal**: Desarrollo de capacidades para identificar patrones cronológicos, análisis de periodicidad, detección de tendencias y evaluación de ciclos operacionales vehiculares.

4.3. Optimizaciones Implementadas:

Se implementaron múltiples optimizaciones computacionales para manejar datasets de gran escala típicos en aplicaciones vehiculares, incluyendo técnicas de muestreo estratificado, procesamiento por lotes y gestión eficiente de memoria para datasets que pueden exceder los 38 millones de registros.

4.4. Documentación Académica:

Se desarrolló documentación académica comprehensiva que incluye metodologías estadísticas empleadas, interpretaciones contextuales para el diagnóstico vehicular, consideraciones computacionales y recomendaciones para implementación práctica.

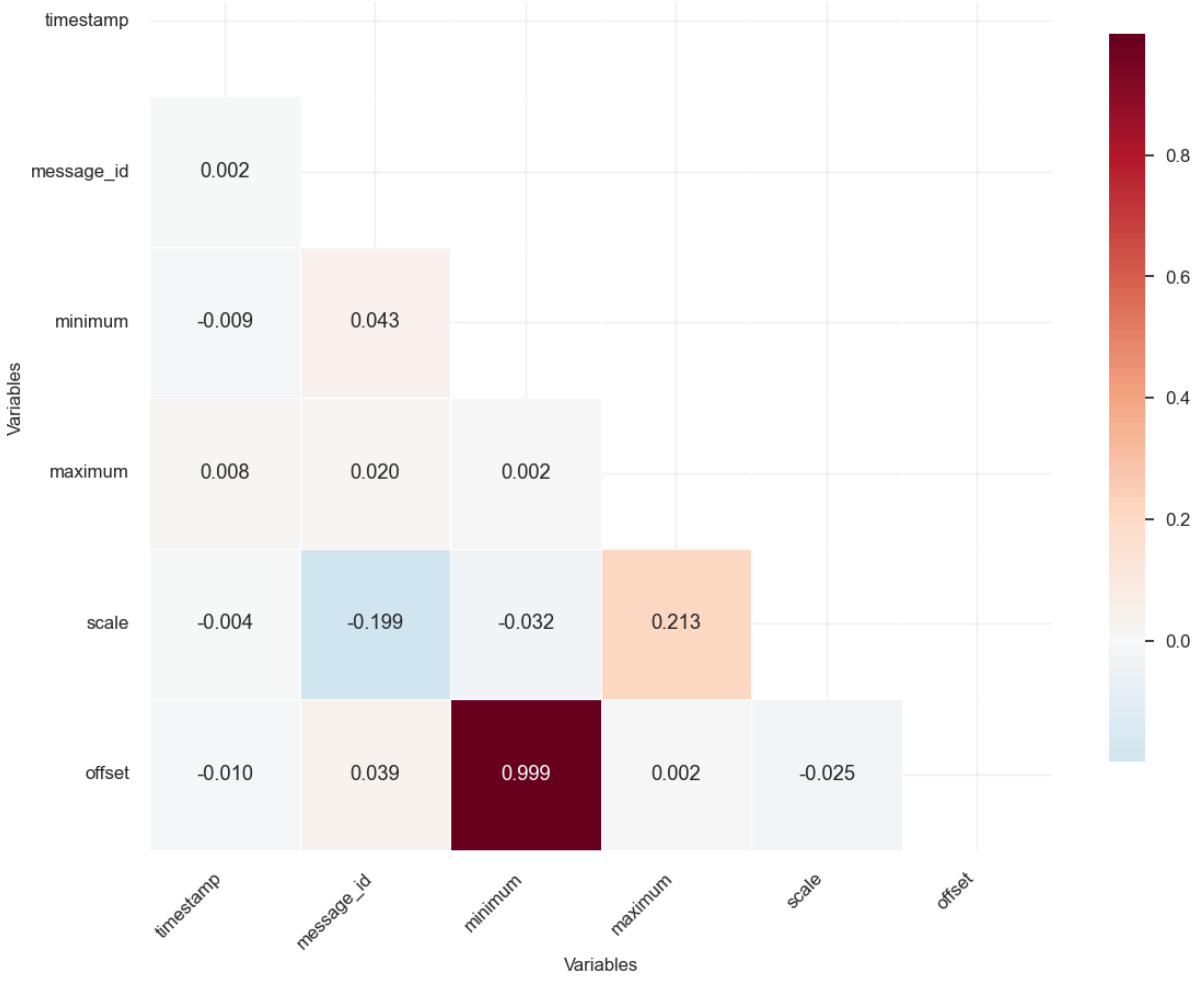
5. Resultados Técnicos

A continuación, se comparten los resultados del análisis de 85 minutos de grabación en tiempo real de los canales CAN EV y CATL, en el marco de una prueba que buscaba validar la autonomía real del vehículo eléctrico bajo condiciones de operación definidas, a fin de comparar su rendimiento con las especificaciones teóricas y determinar factores que afectan su desempeño energético.

5.1. Matriz de Correlación de Variables del Sistema Vehicular BLF

Figura 4-1

Matriz de correlación – Variables del Sistema en archivos BLF



Fuente. Elaboración propia

La gráfica presentada corresponde a una matriz de correlación entre variables asociadas a la decodificación de datos vehiculares almacenados en archivos BLF. Este tipo de análisis permite identificar dependencias lineales entre parámetros de definición de señales (ej. *mínimos, máximos, escalas, offsets*) y metadatos como el *message_id* o el *timestamp*.

La escala cromática de la matriz va del azul (correlación negativa) al rojo oscuro (correlación positiva fuerte), permitiendo visualizar rápidamente las relaciones más significativas.

Las siguientes son las observaciones principales:

1. **Correlación entre minimum y offset (0.999, rojo intenso)**
    - La gráfica muestra un cuadrado rojo oscuro en la intersección de estas variables, lo que refleja una correlación positiva casi perfecta.
    - Interpretación: el valor mínimo registrado y el offset definido en el DBC se comportan prácticamente como la misma variable. Esto implica redundancia y falta de independencia entre ellas.

2. **Relación entre scale y maximum (0.213, tono anaranjado claro)**
    - Se observa una correlación positiva débil.
    - Implicación: en algunos casos, un rango máximo mayor está ligeramente asociado a factores de escala más grandes, pero no es un patrón generalizable.

3. **Relación negativa entre scale y message_id (–0.199, tono azul claro)**
    - El cuadro en azul claro indica una correlación negativa débil.
    - Esto sugiere que ciertos IDs de mensajes tienden a estar asociados a escalas menores, y otros a escalas mayores, lo que refleja variabilidad entre subsistemas vehiculares.

4. **timestamp sin correlación significativa**
    - En la matriz se observan valores cercanos a cero para el *timestamp* frente al resto de variables.
    - Esto es deseable: el tiempo de registro no depende de cómo se definieron las señales, confirmando independencia metodológica.

5. **Otras correlaciones cercanas a cero**
    - Entre minimum y maximum (~0.002), o entre offset y maximum (~0.002), se evidencia prácticamente nula relación lineal.
    - Esto indica que cada variable aporta información distinta en la caracterización de las señales.

Lo anterior tiene las siguientes implicaciones practicas:

- **Redundancia en datos**: la correlación casi perfecta entre minimum y offset sugiere eliminar una de ellas en modelos de análisis para evitar multicolinealidad.
- **Validación de DBC**: esta relación también puede ser un indicador de cómo se definieron los DBC (ej. offset construido directamente a partir del mínimo).
- **Selección de variables**: conviene priorizar variables más independientes (como scale y maximum) en análisis predictivos o exploratorios.
- **Heterogeneidad del sistema**: la débil correlación general entre variables confirma que cada señal vehicular responde a características particulares de su ECU y subsistema.

La matriz de correlación revela un hallazgo clave: minimum y offset son prácticamente la misma variable, lo que plantea un riesgo de redundancia en análisis estadísticos y modelos de machine learning. Por otro lado, la ausencia de correlaciones significativas en la mayoría de los pares respalda la idea de que los parámetros del sistema vehicular son heterogéneos, reflejando la complejidad de la arquitectura CAN.

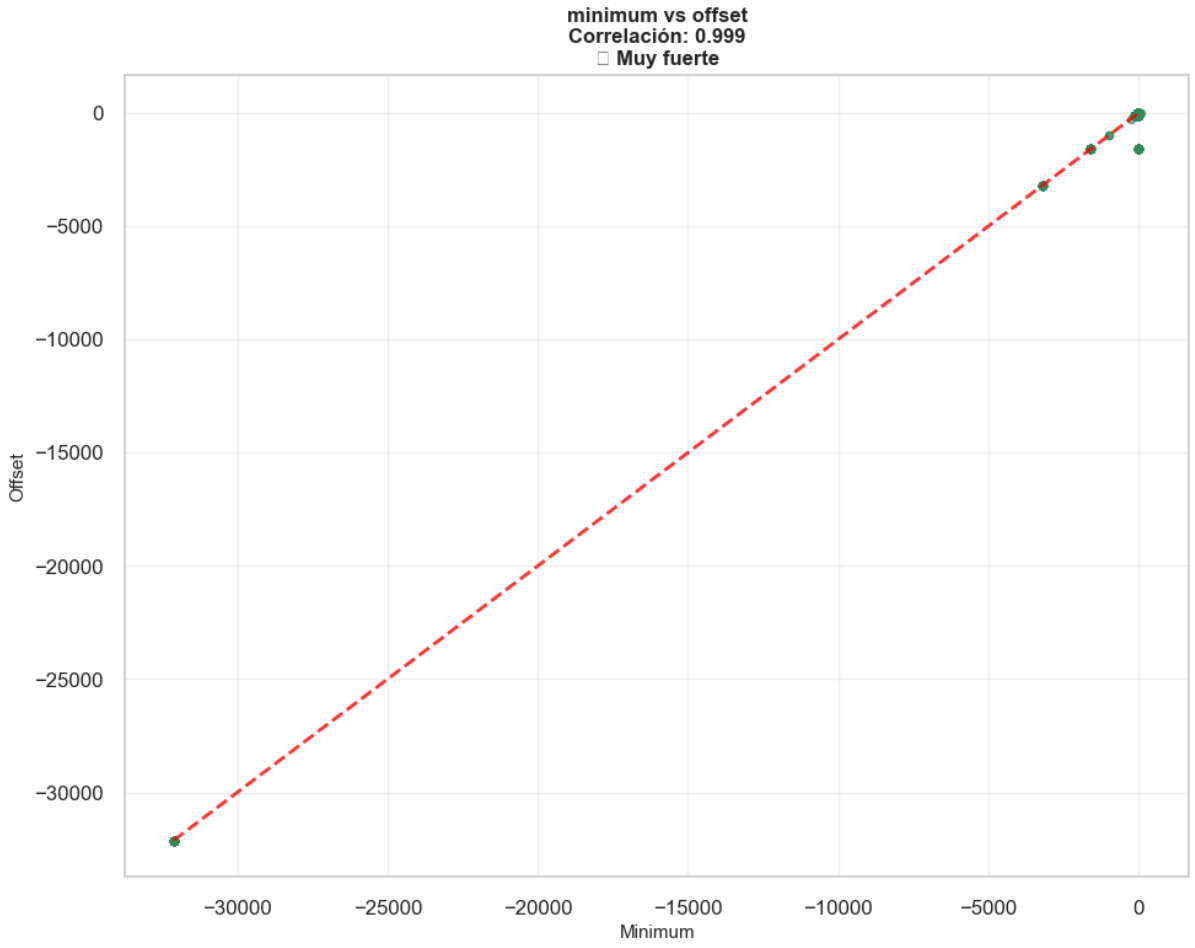
En resumen, este tipo de análisis es valioso para depuración de datasets, detección de redundancias y preparación de datos para inteligencia artificial aplicada a movilidad eléctrica.

5.2. Relación entre las variables de *minimum* y *offset* en el Sistema Vehicular

El gráfico presentado corresponde a un diagrama de dispersión entre los valores de minimum y offset extraídos de las definiciones de señales en un archivo BLF/DBC. La recta roja representa la tendencia lineal ajustada, mientras que los puntos verdes corresponden a las observaciones reales. El coeficiente de correlación mostrado en la gráfica es 0.999, lo que indica una relación casi perfecta.

Figura 4-2

Relación entre minimum y offset



Fuente. Elaboración propia

Se puede observar que:

- Correlación extremadamente alta (0.999): La nube de puntos se alinea de manera casi exacta sobre la recta roja. Esto confirma una relación lineal y directa entre ambas variables: a medida que el valor minimum disminuye, el offset disminuye en la misma proporción.

- Signo de la relación: La pendiente positiva de la línea de tendencia indica que minimum y offset crecen o decrecen en conjunto. Los valores negativos reflejan la forma en que las señales fueron definidas en el DBC (ej. rangos de tensiones o temperaturas con desplazamientos negativos para codificación binaria).

- Baja dispersión: La casi total ausencia de puntos alejados de la recta confirma que no existen valores atípicos significativos. Esto refuerza la hipótesis de que minimum y offset son, en esencia, la misma información duplicada bajo diferente etiqueta.

Dado que ambas variables contienen la misma información, incluirlas juntas en análisis estadísticos o modelos de machine learning provocaría problemas de multicolinealidad.

Es recomendable conservar solo una de las dos variables (ej. offset) para evitar duplicidades y mantener modelos más interpretables.

Esta correlación puede reflejar una práctica común en la generación de archivos DBC: derivar el offset directamente del valor mínimo permitido para la señal.

La gráfica evidencia que la relación entre minimum y offset no es casual, sino estructural. Su correlación de 0.999 indica que ambas variables son prácticamente equivalentes, lo cual debe tenerse en cuenta al momento de preparar datasets para análisis avanza

Este hallazgo contribuye a depurar la base de datos, eliminando redundancias y garantizando que los modelos construidos sobre estos datos reflejen únicamente relaciones genuinas y no artefactos derivados de definiciones duplicadas en el DBC.

6. Conclusiones

La dualidad DBC–BLF es la llave para descifrar las conversaciones invisibles entre ECUs. El primero establece la semántica, el segundo la narrativa temporal. Al profundizar en ambos, se obtiene una comprensión holística de la red CAN, esencial para: Ingeniería de integración, diagnóstico avanzado, modelos de IA para movilidad eléctrica, optimización de la seguridad y eficiencia.

En buses eléctricos, donde la confiabilidad y la gestión energética son críticas, este conocimiento se convierte en un recurso estratégico.

El análisis de datos CAN en vehículos eléctricos presenta desafíos únicos debido a la naturaleza distribuida y heterogénea de los sistemas embebidos. Según Marchetti et al. (2018), la ingeniería inversa de protocolos CAN requiere enfoques sistemáticos que combinen análisis estadístico con conocimiento del dominio automotriz. El presente estudio contribuye a esta área de investigación mediante la caracterización detallada de un dataset real de operación comercial.

La organización jerárquica de los archivos DBC, combinada con la rigurosa extracción controlada de logs CAN, permite que cualquier proceso de ingeniería inversa, análisis estadístico o inteligencia artificial en proyectos de vehículos eléctricos sea reproducible, trazable y científicamente robusto. Este pipeline garantiza que todo log registrado pueda mapearse fielmente a variables técnicas interpretables, siempre que se cuente con el DBC actualizado y correctamente referenciado durante todo el análisis.

Este enfoque estructurado es esencial para habilitar pipelines de IA, garantizar la integridad de los datos y convertir señales eléctricas abstractas en conocimiento accionable para diagnóstico avanzado y mantenimiento predictivo en flotas eléctricas.