Tecnológico de Monterrey

DECODE-EV: Framework de IA Multimodal para Ingeniería de Reconstrucción Funcional, Generación de Hipótesis de Lógica de Control y Correlación con Datos de Operación en Buses Eléctricos

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 21, 2025

IBM

Contenido

Pág.

Lista de figuras

Lista de tablas

1. Introducción

La industria automotriz se encuentra en plena transición hacia los vehículos definidos por software (SDV), donde la capacidad de desarrollar, mantener y evolucionar software embebido se convierte en un factor crítico de innovación. Según IBM (2025), se proyecta que para el año 2030 el 90% de los avances en el sector automotriz estarán impulsados por software, y de acuerdo con el estudio del IBM Institute for Business Value (2024), hacia 2035 la experiencia definida por software será el principal valor de marca para la mayoría de los fabricantes de vehículos.

En este escenario, Superpolo S.A.S., fabricante colombiano de carrocerías y chasis eléctricos, enfrenta el desafío estratégico de fortalecer su capacidad tecnológica en el desarrollo de software embebido propio. Para lograrlo, resulta esencial disponer de mecanismos que permitan correlacionar de manera integral diversas fuentes de información, tales como archivos BIN generados en los procesos de programación, definiciones de red contenidas en archivos DBC, tráfico de datos registrado en redes CAN y logs obtenidos en pruebas de campo.

DECODE-EV representa un framework de inteligencia artificial diseñado para correlacionar automáticamente logs de la red CAN (Controller Area Network) con secciones de código binario en sistemas embebidos de vehículos eléctricos, aprovechando estas fuentes de información para identificar patrones, generar representaciones semánticas y reconstruir modelos explicativos de la lógica de control del vehículo. Con ello, se busca habilitar un camino estratégico hacia la independencia tecnológica y la consolidación de una plataforma de software propia, robusta y sostenible en el tiempo.

2. Antecedentes

2.1. Información de la empresa

Superpolo S.A.S., filial de Marcopolo en Colombia, es una compañía con más de seis décadas de experiencia en la fabricación de carrocerías para buses urbanos e intermunicipales, y se ha consolidado como un actor estratégico en la transición hacia la movilidad sostenible en Latinoamérica. Su planta principal, ubicada en Cota, Cundinamarca, cuenta con certificaciones de calidad (ISO 9001:2015) y una capacidad instalada que le permite atender tanto el mercado nacional como exportaciones regionales (Marcopolo Superpolo, 2025).

En los últimos años, la empresa ha diversificado su portafolio incorporando plataformas eléctricas e híbridas, en línea con las políticas públicas colombianas de reducción de emisiones y con las tendencias globales de electrificación del transporte (Giraldo, 2024). En 2024, Superpolo presentó oficialmente dos prototipos ensamblados en el país: un bus eléctrico con autonomía de ~260 km y un bus de hidrógeno con autonomía de ~450 km, ambos equipados con motores Siemens y tanques de hidrógeno de Hexagon Purus (Giraldo, 2024). Estos proyectos marcaron un hito en el sector, posicionando a la empresa como pionera en la integración de tecnologías limpias y sofisticadas en Colombia.

El proceso de negocio directamente impactado por el presente proyecto corresponde al desarrollo de software embebido para chasis eléctricos, especialmente en lo relacionado con la integración y coordinación de subsistemas críticos: baterías CATL, sistemas de tracción Siemens/Meritor, frenos y suspensión WABCO, dirección BOSCH/WEG y pantallas ACTIA. La complejidad de esta integración exige a la compañía capacidades avanzadas en análisis de datos vehiculares, validación funcional y gestión del ciclo de vida del software, competencias que tradicionalmente dependían de proveedores externos.

En este escenario, la iniciativa DECODE-EV se inserta como un esfuerzo estratégico para construir capacidades propias en software definido por datos y algoritmos de inteligencia artificial, fortaleciendo la independencia tecnológica de Superpolo frente a terceros y habilitando nuevas oportunidades de innovación en productos y servicios de movilidad eléctrica.

2.2. Estado del arte y panorama tecnológico actual

La transición hacia vehículos definidos por software (SDV) redefine la industria automotriz, desplazando el foco de la innovación desde el hardware hacia el software. IBM (2025) estima que hacia 2030 el 90% de las innovaciones en el sector provendrán del software y que, para 2035, la experiencia digital será el principal diferenciador del valor de

marca. En consecuencia, los fabricantes buscan arquitecturas centralizadas y actualizables, capaces de evolucionar continuamente a través de código, en un modelo comparable al de los “smartphones sobre ruedas” (S&P Global Mobility, 2025).

Dentro de esta tendencia, la gestión de datos vehiculares es crítica. Los buses modernos operan múltiples redes Controller Area Network (CAN), donde circulan millones de mensajes diarios que coordinan motores, baterías, frenos y sistemas auxiliares. La interpretación de dichos mensajes es compleja debido a que los fabricantes codifican la semántica de señales de forma propietaria (Buscemi et al., 2023). Para abordar este reto, la literatura ha propuesto metodologías de análisis automatizado de CAN, destacándose herramientas como READ (Marchetti et al., 2018), LibreCAN (Pesé et al., 2019) y CANMatch (Buscemi et al., 2021), que aplican correlaciones con sensores externos y algoritmos de frame-matching para acelerar la identificación de señales. Estas soluciones han demostrado que es posible pasar de procesos manuales que requerían semanas a enfoques automatizados que producen resultados en horas (Buscemi et al., 2023).

De forma paralela, el análisis de código binario de ECU ha evolucionado hacia el uso de modelos de aprendizaje profundo que generan representaciones semánticas de funciones compiladas. Artuso (2024) identifica este como un campo de crecimiento exponencial, particularmente valioso para detectar vulnerabilidades, identificar bibliotecas reutilizadas y comprender la lógica funcional de sistemas críticos. Jiang et al. (2024) proponen BinaryAI, un modelo basado en transformers y aprendizaje contrastivo que alcanzó una precisión del 85,8% en la identificación de componentes en binarios, superando métodos tradicionales. Xia et al. (2023), por su parte, demostraron que el uso de n-gramas semánticos y grafos de flujo de control mejora la detección de similitudes funcionales entre binarios distintos.

El estado del arte converge hacia enfoques multimodales, en los que se integran datos de diferentes dominios para reconstruir una visión funcional completa del sistema. Ojima et al. (2024) muestran cómo los grafos de conocimiento combinados con modelos de lenguaje permiten capturar y consultar el conocimiento experto de ingenieros en análisis de fallas vehiculares. Este tipo de integración inspira el enfoque de DECODE-EV, que busca construir un framework capaz de correlacionar BIN, DBC y CAN, generando hipótesis explicables de lógica de control y validándolas en entornos físico-digitales como MATLAB/Simulink.

Bajo este marco, DECODE-EV se posiciona como un aporte de gran relevancia al aplicar modelos multimodales de IA a un caso real de bus eléctrico colombiano, con el doble propósito de generar conocimiento accionable para ingeniería y fortalecer la independencia tecnológica de la empresa.

3. Entendimiento del negocio

3.1. Formulación del problema

La empresa enfrenta el reto de interpretar y aprovechar de manera integral sus activos digitales vehiculares: archivos BIN de controladores, definiciones DBC parciales y millones de mensajes provenientes de las redes CAN registradas en pruebas de campo. Actualmente, la falta de un proceso estandarizado para correlacionar estas fuentes de información genera ineficiencias en diagnóstico, validación y desarrollo de nuevas funcionalidades.

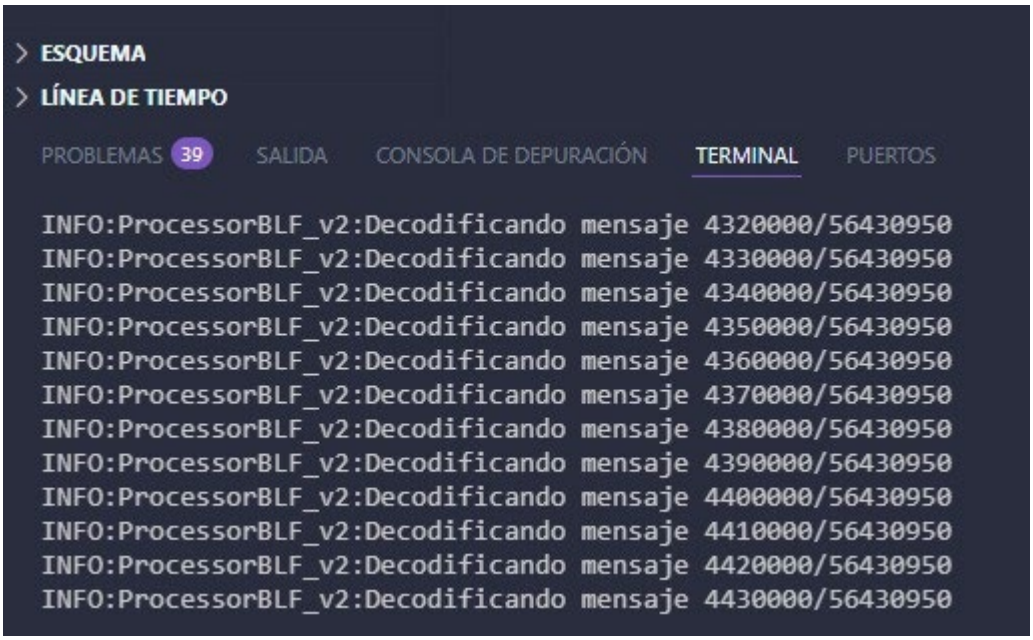
El problema central se define entonces como:

¿Cómo desarrollar un framework de inteligencia artificial multimodal que permita correlacionar firmware, definiciones de red y datos de operación vehicular, para reconstruir de forma funcional y explicable la lógica de control de un bus eléctrico?

3.2. Contexto e importancia

Figura 2-1

Flujo de trabajo propuesto para la reconstrucción de lógica de control



Nota: En la imagen se observa la cantidad de mensajes que se procesan como resultado de una prueba de autonomía del prototipo eléctrico, donde se decodifican los mensajes de los archivos .BLF con base en los .DBC disponibles. Fuente. (Superpolo,2025).

En el marco de la movilidad eléctrica, la innovación depende crecientemente del dominio del software embebido y de la capacidad de gestionar grandes volúmenes de datos

vehiculares (IBM Institute for Business Value, 2023). En pruebas de autonomía, un bus eléctrico de Superpolo puede generar más de 50 millones de mensajes en una sola jornada de 7 horas y 130 km recorridos, lo que hace inviable el análisis manual y resalta la necesidad de herramientas de IA para priorizar y correlacionar señales relevantes.

La relevancia de este proyecto radica en tres dimensiones:

1. Estratégica: fortalece la independencia tecnológica de Superpolo frente a proveedores externos, alineándose con las tendencias globales hacia vehículos definidos por software (S&P Global Mobility, 2025).
2. Operacional: habilita diagnósticos más rápidos y confiables, reduciendo tiempos de integración y posventa (Buscemi et al., 2023).
3. Académica: aporta un caso pionero en Latinoamérica de aplicación de IA multimodal a datos vehiculares, con potencial de transferencia tecnológica a otros sectores de transporte y movilidad.

3.3. Objetivos

3.3.1. General

Diseñar DECODE-EV, un framework de inteligencia artificial multimodal para reconstrucción funcional de lógica de control y correlación BIN–DBC–CAN en buses eléctricos, integrando hipótesis explicables, validación físico-digital y un asistente RAG para consulta técnica.

3.3.2. Específicos

- O1: Procesar logs CAN y generar representaciones vectoriales de señales y eventos relevantes.
- O2: Analizar firmware (BIN) y construir embeddings semánticos de funciones compiladas.
- O3: Correlacionar dominios (BIN ↔ CAN ↔ DBC) mediante técnicas de aprendizaje contrastivo y grafos de llamadas.
- O4: Desarrollar un asistente conversacional RAG para consultas en lenguaje natural, basado en Discovery y Watsonx Assistant.
- O5: Validar hipótesis de lógica de control mediante simulaciones en MATLAB/Simulink.

Nota: Los objetivos específicos están sujetos a modificaciones con base en el asesoramiento que se recibirá por parte del equipo de mentoría tanto de IBM como del TEC en el marco de la asignatura de proyecto integrador.

3.4. Preguntas Clave

1. ¿Qué mensajes y señales de la red CAN explican mayor proporción del comportamiento del bus eléctrico?
2. ¿Cómo relacionar funciones compiladas en el firmware con eventos observables en CAN, garantizando explicabilidad y coherencia física?
3. ¿Qué métricas permiten evaluar la validez de hipótesis de lógica de control sin acceso a documentación completa?
4. ¿Cómo integrar documentación técnica y manuales de proveedores en un proceso de consulta semántica?
5. ¿Qué arquitecturas de IA multimodal (transformers, aprendizaje contrastivo) ofrecen mejor desempeño en correlación de dominios heterogéneos?

3.5. Involucrados y roles

El proyecto se desarrolla en el área de Ingeniería de Innovación y Proyectos de Superpolo, con participación de varios niveles:

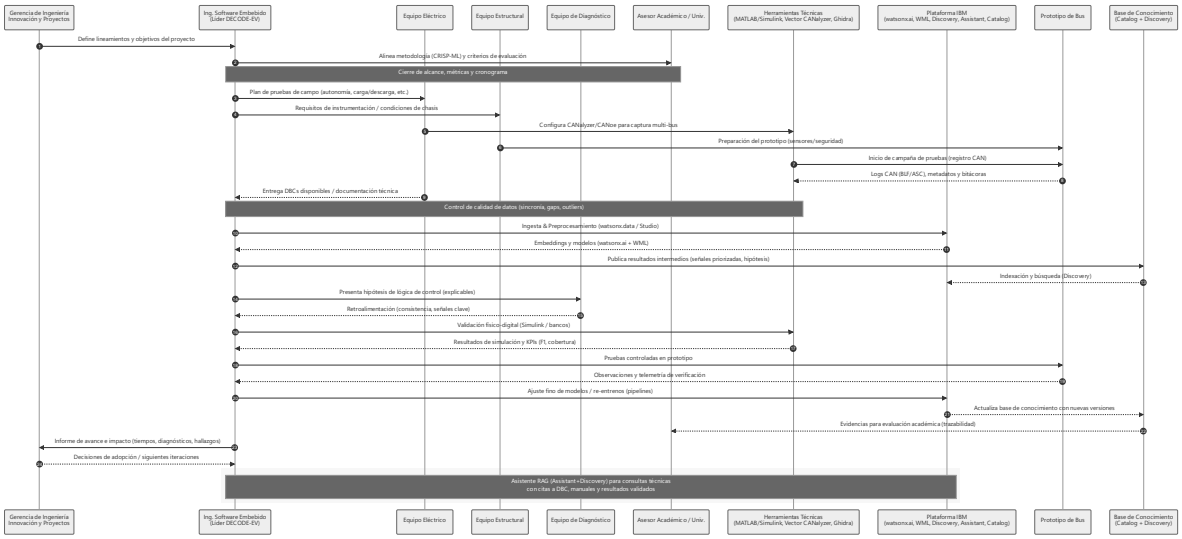
- Ingeniero de software embebido (líder del proyecto): responsable del diseño, implementación y validación del framework DECODE-EV. Lidera la integración académica y empresarial.
- Equipo eléctrico: aporta conocimiento de sistemas de baterías, cargadores y distribución de potencia.
- Equipo estructural: colabora en pruebas de chasis y en la integración física de sensores e interfaces.
- Equipo de diagnóstico: valida hipótesis de señales CAN y aporta experiencia en análisis de fallas.
- Gerencia de ingeniería – innovación y proyectos: actúa como sponsor, define lineamientos estratégicos y facilita recursos.
- Universidad (profesor y comité académico): garantiza rigor metodológico bajo CRISP-ML y supervisa cumplimiento de objetivos académicos.

Este ecosistema de involucrados asegura que el proyecto tenga un impacto práctico en la empresa y un aporte científico en el ámbito académico.

El diagrama de la Figura 3.2 representa la interacción entre los distintos actores, herramientas y el prototipo de bus eléctrico en el marco del proyecto DECODE-EV. A través de este esquema se visualiza cómo la gerencia de ingeniería define los objetivos estratégicos, el ingeniero de software embebido lidera la ejecución técnica, y los equipos eléctricos, estructurales y de diagnóstico colaboran en la preparación de pruebas y validación de resultados. Asimismo, se muestra la integración de plataformas IBM (watsonx), herramientas técnicas (MATLAB, Vector CANalyzer, Ghidra) y la base de conocimiento como elementos fundamentales para el procesamiento de datos, la generación de hipótesis de lógica de control y la retroalimentación continua en un ciclo de mejora validado académicamente bajo la metodología CRISP-ML.

Figura 3-2

Diagrama de secuencia de roles y flujo de trabajo en el proyecto DECODE-EV



Fuente. Elaboración Propia.

3.6. Sector Industrial al que pertenece

Según el Sistema de Clasificación Industrial de América del Norte (SCIAN) de INEGI, SUPERPOLO S.A.S. pertenece al:

- Sector 336 - Fabricación de Equipo de Transporte
  - Subsector 3362 - Fabricación de Carrocerías y Remolques

Este sector comprende unidades económicas dedicadas principalmente a la fabricación de equipo de transporte, como automóviles, camionetas y camiones; carrocerías

y remolques; partes para vehículos automotores; equipo aeroespacial, equipo ferroviario, embarcaciones y otro equipo de transporte.

La actividad específica de SUPERPOLO S.A.S. se clasifica bajo el código SCIAN 336210 - Fabricación de carrocerías y remolques, que incluye el ensamble de carrocerías sobre chasises comprados, así como la fabricación de remolques y semirremolques para diferentes tipos de vehículos.

3.7. Lugar de Aplicación

Tabla 3-1

Ubicación geográfica del proyecto

Concepto	Detalle
Domicilio principal	Vía Siberia Cota Km 1.6 Costado Oriental Hacienda Potrero Chico L
Municipio	Cota, Cundinamarca
País	Colombia

La planta industrial cuenta con más de 77,000 metros cuadrados de área total y 46,300 metros cuadrados de área construida. Adicionalmente, la empresa mantiene presencia nacional con oficinas y centros de servicio en Bogotá, Cali, Medellín y Barranquilla.

4. Entendimiento de los datos

El proyecto DECODE-EV cuenta con un conjunto heterogéneo de fuentes de datos vehiculares y técnicas que, al integrarse, permiten modelar y reconstruir la lógica de control de subsistemas clave del bus eléctrico. Estas fuentes son:

- Registros CAN (logs): capturados en pruebas de pista y banco, en formato BLF/ASC/CSV, con marcas de tiempo, identificadores (IDs) y payloads de hasta ocho bytes por trama. Una sola jornada de autonomía (~7 horas, 130 km) con solo dos de los cinco buses activos puede superar los 50 millones de mensajes, lo que evidencia la magnitud del problema de escalabilidad y la necesidad de técnicas de análisis automatizado (Buscemi et al., 2023).

- Archivos BIN: imágenes de firmware de la ECU principal, almacenadas en formato compilado (.bin, .hex). Estos archivos suelen tener entre 1 y 20 MB y contienen decenas de miles de instrucciones ensamblador, representando el comportamiento funcional de controladores críticos como el tren motriz. Estudios recientes muestran que es posible extraer características semánticas de estos binarios mediante embeddings aprendidos por modelos de deep learning (Artuso, 2024; Jiang et al., 2024).

- Archivos DBC y documentación técnica parcial: especificaciones parciales de red vehicular, manuales de proveedores (CATL, Siemens, Toyota) y diagramas eléctricos. Aunque fragmentarios, sirven como pistas de validación para verificar hipótesis generadas por los modelos.

- Metadatos de pruebas de campo: bitácoras de ingenieros, registros de OBD-II, grabaciones de video del tablero y etiquetas de eventos de operación (ej. frenado, regeneración, carga), que se utilizan como ground truth parcial para corroborar la correlación entre señales y funciones.

4.1. Características de volumen y complejidad

La red CAN opera típicamente a 500 kbps y puede transmitir entre 2000 y 10000 mensajes por segundo, dependiendo de la carga de la red (Buscemi et al., 2023). Esto se traduce en Big Data vehicular, donde los logs de una sola prueba exceden varios cientos de MB. El firmware, aunque pequeño en tamaño absoluto, es denso en complejidad: un archivo de 2 MB puede contener más de 20 000 funciones compiladas, difíciles de rastrear manualmente (Artuso, 2024). La documentación técnica es diversa en formato (tablas, PDF, esquemas) y en idioma (español, inglés, chino), lo que exige técnicas de procesamiento de texto y minería semántica para ser utilizable en consultas técnicas.

4.2. Variables de entrada y salida

El pipeline DECODE-EV contempla variables en tres dominios principales:

- **Entradas CAN:** IDs, payloads crudos, tasas de cambio por bit, autocorrelaciones entre señales, patrones de periodicidad.

- **Entradas BIN:** secuencias de instrucciones, grafos de llamadas, constantes numéricas y tablas, convertidas en **representaciones vectoriales** mediante modelos contrastivos (Jiang et al., 2024).
- **Entradas documentales:** fragmentos textuales normalizados (ej. “SOC en % cada 100 ms”), indexados para recuperación semántica (Ojima et al., 2024).
- **Salidas esperadas:** hipótesis de lógica de control expresadas en máquinas de estado, pseudocódigo explicable y mapas de correlación señal–evento; un archivo DBC generado con nuevas señales decodificadas; y un asistente RAG capaz de responder consultas técnicas en lenguaje natural con citas a la base de conocimiento.

4.3. Técnicas de aprendizaje automático previstas

El proyecto combina técnicas no supervisadas, supervisadas y profundas:

- **No supervisadas:** clustering de mensajes CAN para identificar grupos de señales relacionadas (Ezeobi, Keshav, & Sudbury, 2020); autoencoders para reducir dimensionalidad y detectar patrones ocultos.
- **Supervisadas:** clasificación de estados operativos (RUN, CHARGE, FAULT) cuando se dispone de etiquetas derivadas de pruebas instrumentadas.
- **Deep learning:** embeddings multimodales mediante transformers y aprendizaje contrastivo, permitiendo alinear representaciones de funciones BIN con señales CAN (Artuso, 2024; Jiang et al., 2024).
- **Razonamiento estructural:** grafos de llamadas y análisis de localidad para refinar coincidencias entre funciones y eventos (Xia et al., 2023).

4.4. Problemáticas de calidad y sincronización

Los datos presentan retos significativos:

- **Calidad de logs CAN:** posibles pérdidas de tramas, timestamps no monótonos y ruido de captura.
- **Sincronización temporal:** alineación de CAN con OBD-II, video y eventos externos requiere correlaciones ancla (ej. cambio brusco de velocidad).
- **Cobertura parcial:** ciertas condiciones (fallas, arranques en frío, ABS) pueden estar ausentes, generando sesgo en la inferencia.
- **Documentación heterogénea:** manuales escaneados con OCR defectuoso o en múltiples idiomas.

- **Explicabilidad:** necesidad de que las hipótesis generadas sean entendibles por ingenieros, no solo predicciones de caja negra.

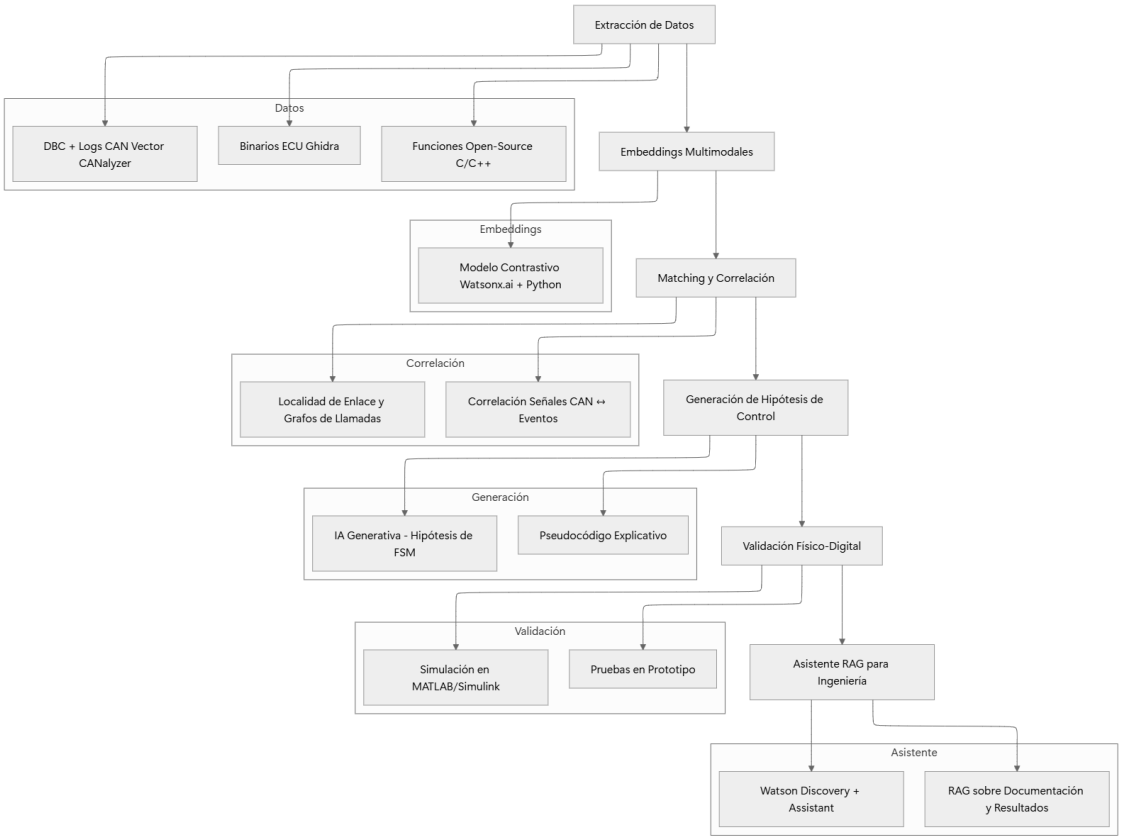
Estos retos hacen indispensable una estrategia de curación de datos, alineación temporal y validación.

5. Flujo general de trabajo propuesto

El flujo de trabajo propuesto inicia con la extracción de datos de binarios, funciones en C/C++ y registros CAN. Luego, se generan embeddings multimodales que permiten comparar y relacionar estas fuentes. A continuación, se realiza el matching y la correlación entre funciones y señales para inferir la lógica de control. Con estos resultados, se aplica IA generativa para proponer máquinas de estados y pseudocódigo explicativo, los cuales se validan mediante simulación físico-digital y pruebas en prototipo, culminando en un asistente RAG que facilita la consulta técnica en lenguaje natural.

Figura 5-1

Flujo de trabajo propuesto para la reconstrucción de lógica de control



Fuente. Elaboración propia.

6. Herramientas de desarrollo

A continuación, se expone una tabla unificada de herramientas que integra tanto las de IBM como las adicionales, mostrando su función y el uso concreto proyectado en el proyecto.

Tabla 6-1

Relación de herramientas para el proyecto

Herramienta	Función principal	Uso en el proyecto
watsonx.ai	Entrenamiento de modelos de IA	Generar embeddings multimodales para alinear binarios, funciones fuente y señales CAN.
Watson Studio	Entorno de ciencia de datos y ML	Gestión de datasets, desarrollo de notebooks colaborativos y orquestación de pipelines de entrenamiento.
Watson Machine Learning (WML)	Despliegue de modelos	Publicación de modelos de predicción y correlación como APIs consumibles en el framework.
watsonx.data	Repositorio estructurado	Almacenamiento de funciones, señales y resultados de correlación en formato tabular.
IBM Knowledge Catalog	Gobernanza de datos	Creación de un glosario de señales CAN, funciones de control y trazabilidad de resultados.
Watson Discovery	Indexación y búsqueda cognitiva	Carga de manuales, DBC y documentación técnica para consultas en lenguaje natural.
watsonx Assistant	Asistente conversacional con IA	Implementación del asistente RAG para consultas técnicas de ingeniería.
watsonx Orchestrate	Automatización de procesos	Automatizar flujos de ingestión de nuevos logs, reentrenamiento de modelos y publicación de resultados.

Speech-to-Text / Text-to-Speech	Procesamiento de voz	Interacción por comandos de voz en pruebas de pista y bancos de ensayo.
MATLAB/Simulink (Simscape Electrical, Driveline, Battery)	Modelado físico y simulación	Validación de máquinas de estados inferidas en un entorno de simulación físico-digital.
Python (NumPy, Pandas, Scikit-learn, PyTorch, NetworkX)	Análisis y ML	Desarrollo de pipelines de datos, entrenamiento de modelos de embeddings y clustering de señales.
Vector CANalyzer / CANoe	Registro y validación de CAN	Captura de logs CAN en pruebas de campo y validación en tiempo real de correlaciones.
Ghidra (plugin C166)	Ingeniería inversa de binarios	Descompilación y extracción de funciones de la ECU principal.
Tree-sitter	Parsing estructurado	Análisis sintáctico de pseudocódigo y funciones fuente en C/C++.
VS Code / Eclipse	IDEs de desarrollo embebido	Edición, compilación y depuración de software embebido.
GitHub/GitLab	Gestión de código	Versionamiento, colaboración y trazabilidad del desarrollo.
Docker/Kubernetes	Contenerización y despliegue	Empaquetado de pipelines y despliegue en la nube (IBM Cloud).

Fuente. Elaboración propia. Nota. Se requiere asesoría especializada para definir el alcance del proyecto en cuestión del tiempo disponible para su ejecución, por lo que estas herramientas solo dan un panorama general, hasta que se defina en conjunto con el asesor, el alcance real del proyecto.

7. Convenios (Ética, Confidencialidad y Uso Académico)

El desarrollo del proyecto DECODE-EV implica el acceso y análisis de fuentes de datos que pueden considerarse sensibles y estratégicas para la empresa y sus proveedores. Estos incluyen registros de red CAN de buses eléctricos, imágenes compiladas de firmware (archivos BIN), definiciones parciales de DBC y documentación técnica de subsistemas. Por ello, es imprescindible establecer lineamientos de uso responsable y convenios que garanticen tanto la validez académica del proyecto como la protección de la propiedad intelectual de la organización.

7.1. Uso académico de datos sensibles

El proyecto se adscribe a los principios de uso académico exclusivo de los datos. Esto significa que los registros y archivos proporcionados por Superpolo serán utilizados únicamente para fines de investigación, validación y documentación dentro del marco de la maestría. No se permitirá su divulgación pública ni su compartición con terceros no autorizados. Para publicaciones y repositorios personales, se emplearán datasets sintéticos o anonimizados, una práctica recomendada en proyectos industriales con componentes confidenciales (Ojima et al., 2024).

7.2. Confidencialidad y acuerdos de acceso

Dada la naturaleza de la información, se establece la necesidad de un acuerdo de confidencialidad (NDA) entre la empresa, la universidad y el estudiante responsable. Este convenio asegurará que:

- Los archivos BIN, registros CAN y documentos internos se manejen únicamente en entornos seguros.
- Cualquier resultado divulgado fuera de la empresa sea transformado en forma de abstracciones, hipótesis explicables o datos simulados.
- Se garantice la trazabilidad de quién accede a la información, siguiendo buenas prácticas de gobernanza de datos (IBM, 2023).

7.3. Consideraciones éticas

El proyecto se enmarca en un propósito legítimo: mejorar la interoperabilidad, la validación y el mantenimiento de buses eléctricos de Superpolo. Se evita explícitamente cualquier intención de clonar o replicar software de terceros, centrándose en la correlación funcional y en la generación de nuevo conocimiento. Esto se alinea con el principio de uso ético de técnicas de IA y minería de datos en contextos industriales (Artuso, 2024).

Un aspecto ético adicional corresponde a la seguridad vehicular. Durante el análisis de datos y firmware, es posible identificar vulnerabilidades o comportamientos no documentados. En tales casos, se aplicará un proceso de divulgación responsable, informando primero a los actores internos de Superpolo antes de considerar cualquier referencia en entregables académicos.

7.4. Privacidad

Si bien los datos vehiculares no contienen información personal directa, podrían inferirse patrones sensibles como rutas o condiciones de operación. Por ello, antes de ser utilizados en publicaciones o demostraciones, los registros serán anonimizados eliminando identificadores de tiempo absoluto, ubicaciones y cualquier información que pueda vincularse a personas o clientes específicos (Buscemi et al., 2023).

7.5. Garantía de integridad académica

El convenio incluye el compromiso de respetar la propiedad intelectual de terceros, citando correctamente bibliografía, software libre o componentes reutilizados. Además, se fomentará la transparencia mediante documentación completa del proceso, en línea con la metodología CRISP-ML, lo que garantiza reproducibilidad y confiabilidad de los resultados;

8. Resultados esperados

El proyecto DECODE-EV busca generar resultados tangibles que impacten tanto en el ámbito académico como en el industrial. Estos resultados se pueden organizar en dos dimensiones: logros técnicos e impacto en la empresa.

8.1. Logros técnicos

El principal resultado esperado es la implementación de un framework de IA multimodal capaz de integrar y correlacionar datos provenientes de registros CAN, archivos BIN y documentación técnica. Este framework deberá:

- Generar embeddings multimodales que representen señales CAN, funciones compiladas y fragmentos documentales en un espacio vectorial común, permitiendo medir similitudes y correlaciones (Artuso, 2024; Jiang et al., 2024).
- Construir hipótesis de lógica de control en forma de máquinas de estado, pseudocódigo explicable o diagramas de flujo que puedan ser interpretados por ingenieros sin necesidad de acceso al código fuente.
- Producir un archivo DBC extendido, en el que se documenten las señales decodificadas, su periodicidad y escalas, con un nivel de confianza basado en métricas de validación (Buscemi et al., 2023).

- Desarrollar un asistente RAG (basado en Watson Discovery y Watsonx Assistant) que permita consultas en lenguaje natural, con respuestas sustentadas en la base de conocimiento consolidada (Ojima et al., 2024).

8.2. Impacto en la empresa

En el plano organizacional, se espera que DECODE-EV:

- Reduzca significativamente los tiempos de análisis y diagnóstico. Actualmente, la interpretación de redes CAN desconocidas puede tardar semanas; con el framework, se busca obtener resultados iniciales en cuestión de horas, apoyando procesos de integración y pruebas de nuevos prototipos.
- Fortalezca la independencia tecnológica de Superpolo, disminuyendo la dependencia de proveedores externos para la comprensión de lógicas funcionales y protocolos de comunicación.
- Mejore la capacidad de soporte posventa, al ofrecer diagnósticos más rápidos y herramientas predictivas basadas en la correlación de datos históricos y en tiempo real.
- Posicione a la empresa como pionera en Latinoamérica en la adopción de enfoques de IA multimodal aplicados al desarrollo de buses eléctricos, alineándose con la tendencia global hacia los SDV (IBM Institute for Business Value, 2023).

8.3. Métricas de éxito previstas

Para evaluar la efectividad del framework, se plantean métricas tanto técnicas como de negocio:

- **Cobertura de señales decodificadas:** porcentaje de mensajes CAN que pudieron asociarse a una función o variable con confianza ≥70%.
- **Precisión en hipótesis de lógica de control:** concordancia entre las máquinas de estado inferidas y los resultados de validación físico-digital (F1 ≥ 0.80).
- **Latencia de consultas en el asistente RAG:** tiempo promedio de respuesta < 2 segundos en búsquedas dentro de la base de conocimiento.
- **Reducción de tiempos de diagnóstico:** disminución ≥50% en el tiempo requerido para identificar señales críticas en nuevos prototipos.
- **Satisfacción del equipo de ingeniería:** retroalimentación positiva de usuarios internos en pruebas piloto, medida mediante encuestas de utilidad y facilidad de uso.

9. Referencias Bibliográficas

Artuso, F. (2024). *Deep learning based binary code analysis* (Tesis doctoral). Università di Roma La Sapienza.

Buscemi, A., Turcanu, I., Castignani, G., Crunelle, R., & Engel, T. (2021). CANMatch: A fully automated tool for CAN bus reverse engineering based on frame matching. *IEEE Transactions on Vehicular Technology, 70*(12), 12358–12373. https://doi.org/10.1109/TVT.2021.3119401

Buscemi, A., Turcanu, I., Castignani, G., Panchenko, A., Engel, T., & Shin, K. G. (2023). A survey on Controller Area Network reverse engineering. *IEEE Communications Surveys & Tutorials, 25*(3), 1445–1481. https://doi.org/10.1109/COMST.2023.3267787

Ezeobi, S. A., Keshav, S., & Sudbury, J. (2020). A clustering-based approach to reverse engineer automotive CAN frames. *SAE Technical Paper 2020-01-1347*. SAE International. https://doi.org/10.4271/2020-01-1347

Giraldo, C. (2024, marzo 9). Marcopolo Superpolo presenta buses eléctricos y de hidrógeno, hechos en Colombia. *Portafolio*. https://www.portafolio.co/

IBM Institute for Business Value. (2023). *The software-defined vehicle: Redefining automotive value creation*. IBM. https://www.ibm.com/thought-leadership/institute-business-value

Jiang, L., An, J., Huang, H., Tang, Q., Nie, S., Wu, S., & Zhang, Y. (2024). BinaryAI: Binary software composition analysis via intelligent binary–source code matching. En *Proceedings of the 46th International Conference on Software Engineering (ICSE '24)*. ACM.

Marchetti, M., Stabili, D., Guido, A., & Colajanni, M. (2018). READ: Reverse engineering of automotive data frames. *IEEE Transactions on Information Forensics and Security, 14*(4), 1083–1097. https://doi.org/10.1109/TIFS.2018.2874645

Marcopolo Superpolo. (2025). *¿Quiénes somos?* https://superpolo.com.co

Ojima, Y., Shindo, R., Ichise, R., & Kajiwara, T. (2024). Knowledge management for automobile failure analysis using knowledge graph and large language model. *IEICE Transactions on Information and Systems, E107-D*(1), 11–20. https://doi.org/10.1587/transinf.2023KBP0003

Pesé, M. D., Stacer, T., Campos, C. A., Newberry, E., Chen, D., & Shin, K. G. (2019). LibreCAN: Automated CAN message translator. En *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2283–2300). ACM. https://doi.org/10.1145/3319535.3354258

S&P Global Mobility. (2025, septiembre 8). *The software-defined vehicle market is taking off*. S&P Global. https://www.spglobal.com/mobility

Xia, S., Ding, S., Gao, J., Yang, J., & Xu, Z. (2023). Binary code similarity analysis based on semantic n-grams of canonicalized binary code. *IEEE Transactions on Information Forensics and Security, 18*, 2370–2383. https://doi.org/10.1109/TIFS.2023.3260461