UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Ingeniería en Desarrollo y Gestión de Software

Extracción de Conocimiento en Bases de Datos

Evidencia 2: Solución del noveno caso de estudio de análisis de calidad de aire e identificación de focos críticos en zonas industriales.

IDGS91N

PRESENTA:

Luis Eduardo Aguilar Sarabia

Abraham Camacho Ríos

Giselle Cantú Chávez Karla

Alejandra de la Cruz Zea

Ricardo Hernández Martínez

NOMBRE DEL DOCENTE:

Ing. Luis Enrique Mascote Cano

Chihuahua, Chih., 27 de septiembre de 2025

ÍNDICE

ÍNDICE DE FIGURAS

OBJETIVO DEL PROYECTO

Como equipo, nos propusimos dar respuesta a un problema que sabemos es crítico en muchas ciudades: la calidad del aire en zonas industriales. Nuestro caso se centra en la empresa municipal *AirGuard*, que tiene la tarea de monitorear la contaminación en una zona metropolitana mediante una red de cincuenta sensores que registran partículas (PM2.5 y PM10) y gases contaminantes como CO y NO_2, además de datos meteorológicos (temperatura, humedad y viento) cada cinco minutos.

Nuestro objetivo principal es diseñar un sistema que permita identificar, en tiempo real, focos críticos de contaminación en las zonas más vulnerables, y que además cuente con un modelo de predicción a 24 horas que nos ayude a anticipar posibles brotes. Con ello buscamos que *AirGuard* pueda emitir alertas tempranas hacia la población, de modo que las autoridades y la ciudadanía puedan tomar decisiones oportunas para proteger la salud y el medio ambiente.

Para lograrlo debemos construir un pipeline en streaming que procese continuamente la información y que, en paralelo, trabaje con fases batch que aprovechen los registros históricos para entrenar el modelo predictivo. Nuestro compromiso como equipo es plantear una solución viable, escalable y útil para quienes gestionan la calidad del aire en la ciudad.

Según la Organización Mundial de la Salud (2024), la contaminación del aire está vinculada a millones de muertes prematuras cada año, y el Programa de las Naciones Unidas para el Medio Ambiente (2022) señala que la medición acompañada de análisis predictivo es esencial para diseñar políticas públicas efectivas. Queremos aportar con este proyecto una propuesta alineada a esas necesidades, integrando tanto el aspecto técnico como el impacto social.

ALCANCE DEL PROYECTO

Decidimos enfocar nuestra propuesta en lo que sí podemos cubrir de manera realista y en lo que, por razones de tiempo y recursos, queda fuera. Lo primero que trabajaremos son los datos de los cincuenta sensores desplegados en la zona metropolitana. Estos equipos registran contaminantes clave —PM2.5, PM10, monóxido de carbono (CO) y dióxido de nitrógeno (NO_2)— junto con variables meteorológicas como viento, temperatura y humedad cada cinco minutos. Ese flujo constante de información será la base de todo lo que desarrollemos.

Nuestro alcance incluye la construcción de un pipeline en streaming que ingiera y procese los datos en tiempo real. Allí mismo planeamos limpiar, transformar y organizar la información para que sea útil en la detección de focos críticos. A la par, vamos a implementar procesos batch para entrenar modelos con los históricos, aprovechando la riqueza de los registros acumulados. Dentro de los modelos, nos centraremos en redes LSTM, por su capacidad para trabajar con series temporales y anticipar los niveles de contaminación a 24 horas.

Queremos que el resultado final sea más que un prototipo funcional: buscamos dejar un esquema replicable, con datasets preparados, scripts de entrenamiento, métricas claras y un prototipo de sistema de alertas que refleje lo aprendido.

Reconocemos ciertas limitaciones, la calidad de los resultados dependerá de la confiabilidad de los sensores, de la infraestructura de cómputo disponible para entrenar modelos de deep learning y del tiempo que tenemos en este cuatrimestre. Aun así, creemos que la propuesta tiene un valor importante porque sienta las bases de un sistema escalable que podría evolucionar a un despliegue real.

JUSTIFICACIÓN DE LA METODOLOGÍA

Analizamos tres opciones principales: KDD, SEMMA y CRISP-DM. Cada una tiene su historia y utilidad, pero la comparación nos llevó a optar por CRISP-DM porque sentimos que es la que mejor refleja la naturaleza de nuestro proyecto y la urgencia de generar resultados claros en poco tiempo.

KDD (Knowledge Discovery in Databases) plantea un ciclo de descubrimiento de conocimiento que ha sido fundamental en el origen de la minería de datos. Sin embargo, lo percibimos más como una visión general que como una ruta práctica. SEMMA, por otro lado, está pensado para trabajar en entornos específicos como SAS. Esa dependencia lo hace potente en escenarios corporativos, pero poco flexible para un caso académico en el que usamos herramientas diversas (Python, frameworks de deep learning, bases de datos abiertas). Por eso, aunque reconocemos su valor histórico, descartamos SEMMA para este trabajo.

CRISP-DM, en contraste, nos da lo que buscábamos: claridad, estructura y flexibilidad. Está dividido en seis fases que encajan perfectamente con lo que nos piden en la consigna: comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación y despliegue. La fase de comprensión del negocio nos ayuda a no olvidar que el objetivo final es atender la necesidad de *AirGuard*: identificar focos críticos de contaminación y anticipar brotes para proteger a la población. La fase de comprensión de los datos nos orienta a estudiar a fondo las series de PM2.5, PM10, CO y NO_2, junto con las variables meteorológicas.

Nos gusta especialmente que CRISP-DM sea iterativo. En la práctica sabemos que entrenar un modelo LSTM no siempre resulta bien al primer intento: se puede sobrentrenar, los datos pueden estar incompletos, o las métricas no alcanzar los umbrales que buscamos. La metodología nos da permiso de volver atrás, ajustar lo que haga falta y mejorar continuamente. En un entorno como el análisis ambiental, donde los datos pueden ser ruidosos y muy variables, esa flexibilidad es indispensable (Álvarez, 2025).

Al revisar experiencias documentadas, confirmamos nuestra decisión. Espinosa-Zúñiga (2020) describe cómo CRISP-DM se aplicó en la segmentación geográfica, mostrando que puede adaptarse a datos complejos. De forma similar, el trabajo de la Universidad Simón Bolívar (s. f.) sobre la implementación de CRISP-DM como herramienta de apoyo muestra su utilidad en contextos académicos, justo como el nuestro. Y en la práctica profesional, MyTaskPanel (s. f.) y Blog MBA USPesalq (2024) resaltan que se ha convertido en un estándar de facto para proyectos de machine learning, al grado de ser llamado “la metodología del futuro” en ambientes de ciencia de datos.

Por otro lado, el uso de CRISP-DM en proyectos ambientales también tiene respaldo. Surmay et al. (2024) desarrollaron un modelo de monitoreo urbano en Chile aplicando metodologías similares y confirmaron que los enfoques iterativos ayudan a procesar datos ruidosos con éxito. Boudriki Semlali et al. (2024), en su propuesta de arquitectura para monitoreo de calidad del aire, muestran cómo el procesamiento en tiempo real se potencia al estructurar las fases bajo un marco organizado. Incluso en estudios más amplios como el de Muñoz et al. (2010) con el Observatorio Andino, se demuestra que contar con una metodología clara es lo que marca la diferencia entre un experimento aislado y un sistema sostenible de análisis ambiental.

La elección de CRISP-DM se conecta, aparte, directamente con la dimensión social del caso. La OPS (s. f.) y la OMS (2024) coinciden en que el monitoreo de la contaminación atmosférica es vital para prevenir riesgos de salud. No basta con medir; hay que procesar los datos, interpretarlos y actuar con rapidez. CRISP-DM nos da una hoja de ruta que asegura que no perdamos de vista ese objetivo. Como señala el PNUMA (2022), medir sin analizar y sin planear acciones es insuficiente. Justamente ahí es donde el marco metodológico aporta valor, nos obliga a traducir datos en conocimiento útil.

PLANEACIÓN DE ETAPAS

Para organizar nuestro trabajo elegimos seguir las seis fases de CRISP-DM. No queremos quedarnos solo en la teoría, así que pensamos en actividades concretas, entregables tangibles y un cronograma realista. La planeación busca que cada etapa tenga un inicio y un cierre definidos, pero también deja espacio para volver atrás cuando sea necesario, porque sabemos que en proyectos de ciencia de datos rara vez todo sale perfecto al primer intento.

La primera fase es la comprensión del negocio. Aquí nos enfocaremos en entender a fondo el problema de la calidad del aire en la zona metropolitana. Revisaremos regulaciones, umbrales de contaminantes establecidos por organismos internacionales y las necesidades reales de *AirGuard*. El entregable será un documento donde definamos indicadores clave de desempeño (KPIs), riesgos y objetivos específicos. Esta etapa será clave para no perder de vista que el propósito final no es solo entrenar un modelo, sino reducir riesgos de salud pública.

Después viene la comprensión de los datos. Recibiremos la información generada por los cincuenta sensores y la analizaremos para conocer su estructura, calidad y posibles problemas. Aquí identificaremos datos faltantes, valores atípicos y correlaciones entre contaminantes y variables meteorológicas. El producto será un informe descriptivo acompañado de gráficas exploratorias que nos permitan dimensionar el reto.

La tercera fase es la preparación de los datos, probablemente una de las más demandantes. Realizaremos limpieza, normalización, imputación de datos faltantes y generación de variables derivadas, como promedios horarios o índices compuestos. El entregable será un dataset limpio y documentado, listo para alimentar los modelos.

La cuarta fase corresponde al modelado. Aquí aplicaremos redes neuronales LSTM para pronosticar la contaminación a 24 horas. Probablemente hagamos pruebas con otros modelos de series de tiempo (ARIMA, Prophet) para comparar y reforzar los resultados. El entregable será un reporte con métricas de desempeño (precisión, RMSE, MAE), scripts de entrenamiento y un primer prototipo de sistema de alertas.

En la quinta fase, la evaluación, compararemos los modelos con base en las métricas definidas y validaremos si cumplen con los objetivos planteados. No se trata solo de ver qué modelo predice mejor, sino de verificar que realmente sirva para identificar brotes de contaminación a tiempo. El entregable será un reporte de evaluación con tablas comparativas y gráficas de desempeño.

Finalmente, la fase de despliegue busca dejar un prototipo funcional que pueda ser utilizado más allá de este trabajo académico. Prepararemos un pipeline en streaming que ingiera datos en tiempo real y ejecute el modelo entrenado. Aunque no construiremos una aplicación pública, dejaremos scripts listos y documentación clara para que alguien más pueda integrarlos a sistemas de mayor escala. El entregable será un manual técnico junto con el prototipo de alertas.

Tabla 1

Planeación de etapas.

Fase (CRISP-DM)	Actividades	Entregables	Semanas
Comprensión del negocio	Revisión de umbrales, definición de KPIs, análisis de requisitos de AirGuard	Documento de requisitos y objetivos	1 – 2
Comprensión de los datos	Exploración inicial, detección de valores atípicos, análisis de correlaciones	Informe exploratorio con gráficas	3 – 4
Preparación de los datos	Limpieza, imputación, normalización, creación de variables derivadas	Dataset limpio y documentado	5 – 6
Modelado	Implementación de LSTM, pruebas con modelos alternativos, ajuste de hiperparámetros	Scripts de entrenamiento, métricas de desempeño, prototipo inicial	7 – 8
Evaluación	Comparación de modelos, validación de resultados, verificación de KPIs	Reporte de evaluación con gráficas y tablas comparativas	9
Despliegue	Configuración de pipeline en streaming, documentación, entrega de prototipo	Manual técnico y prototipo funcional de alertas	10 – 11

Esta planeación refleja un avance progresivo, pero con posibilidad de retroalimentación en cada fase. La lógica es que, al terminar, encontremos un camino bien documentado que pueda servir de base para proyectos similares en el futuro. Surmay et al. (2024) ya demostraron en su investigación que dividir en fases claras acelera la adaptación de modelos ambientales. Lo mismo se observa en propuestas como la de Boudriki Semlali et al. (2024), donde la arquitectura de monitoreo en streaming se sostiene precisamente en la planeación metodológica.

CONCLUSIÓN

Al plantear este proyecto, la pregunta central fue si seríamos capaces de diseñar una solución que detectara focos críticos de contaminación y predijera brotes a 24 horas en un entorno industrial complejo. Después de todo el análisis, las comparaciones y la planeación, logramos estructurar una propuesta sólida que da respuesta directa al problema planteado, pensando tanto en la viabilidad técnica como en la utilidad social.

El resultado es un sistema completo que integra ingestión en streaming, análisis batch y un modelo predictivo basado en redes LSTM. Al organizar todo con la metodología CRISP-DM, cada fase nos permitió construir sobre lo anterior y llegar a un esquema final coherente, convirtiendo el objetivo inicial en un plan de acción bien definido.

La investigación nos llevó a reconocer que el verdadero reto está en lidiar con la calidad de los datos. Aprendimos a dar importancia al preprocesamiento, a detectar vacíos y anomalías, y a entender que sin una base limpia no existe modelo confiable, reflejando cómo funcionan los proyectos reales.

En cuanto al modelado, trabajar con LSTM nos exigió comprender cómo se comportan las series temporales ambientales. Nos dimos cuenta de que este tipo de modelos requieren ajustes, pruebas y validaciones constantes, pero que las técnicas de deep learning realmente pueden marcar una diferencia cuando se aplican con disciplina.

Reconocemos las limitaciones: los resultados dependen de factores fuera de nuestro control, como la calibración de los sensores o la infraestructura de cómputo, y no incluimos la parte legal o la integración de alertas en plataformas masivas. Aun así, lo que construimos cumple con lo esperado: un sistema técnico viable que puede crecer si se le destinan más recursos.

REFERENCIAS Y FUENTES CONSULTADAS

Álvarez, S. F. C. (2025). El CRISP – DM y la inteligencia artificial. *Revista Neuronum*. Recuperado el 27 de septiembre de 2025 de https://eduneuro.com/revista/index.php/revistaneuronum/article/view/579

Blog MBA USPesalq. (2024). Crisp-DM: las 6 etapas de la metodología del futuro. Recuperado el 27 de septiembre de 2025 de https://blog.mbauspesalq.com/es/crisp-dm-las-6-etapas-de-la-metodologia-del-futuro

Boudriki Semlali, B.-E., El Amrani, C., Ortiz, G., Boubeta-Puig, J. y García-de-Prado, A. (2024). SAT-CEP-monitor: An air quality monitoring software architecture combining complex event processing with satellite remote sensing. *arXiv*. https://arxiv.org/abs/2401.16339

CICIC. (2024). Aplicación de la metodología CRISP-DM en la detección de necesidades ciudadanas en base al análisis de tweets. *Proceedings of CICIC 2024*. Recuperado el 27 de septiembre de 2025 de https://www.iiis.org/CDs2024/CD2024Spring//papers/CB064WH.pdf

Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica. *Revista Mexicana de Investigación Educativa*. Recuperado el 27 de septiembre de 2025 de https://www.scielo.org.mx/scielo.php?pid=S1405-77432020000100008&script=sci_abstract

FMC. (2017). Efectos de la calidad del aire sobre la salud. *FMC*. Recuperado el 27 de septiembre de 2025 de https://www.fmc.es/es-efectos-calidad-del-aire-sobre-articulo-S1134207217301299

IQAir. (2024). ¿Cómo afecta la contaminación del aire al aire libre interior? *IQAir*. Recuperado el 27 de septiembre de 2025 de https://www.iqair.com/us-es/newsroom/como-afecta-la-contaminacion-del-aire-al-aire-libre-mi-calidad-del-aire-interior

La metodología CRISP-DM: desarrollo de modelos de machine learning. (s. f.). *MyTaskPanel*. Recuperado el 27 de septiembre de 2025 de https://www.mytaskpanel.com/la-metodologia-crisp-dm-desarrollo-de-modelos-de-machine-learning

Metodologías y estándares. (s. f.). *Universitat Oberta de Catalunya*. Recuperado el 27 de septiembre de 2025 de https://openaccess.uoc.edu/server/api/core/bitstreams/20b39332-4ac7-49e6-b9cc-c34c4401d380/content

Muñoz, Á. G., López, P., Velásquez, R., Monterrey, L., León, G., Ruiz, F., … Carrasco, G. (2010). An Environmental Watch System for the Andean countries: El Observatorio Andino. *arXiv*. https://arxiv.org/abs/1006.0926

Organización Mundial de la Salud. (2024, 24 de octubre). Contaminación del aire ambiente (exterior) y salud. Recuperado el 27 de septiembre de 2025 de https://www.who.int/es/news-room/fact-sheets/detail/ambient-%28outdoor%29-air-quality-and-health

Organización Panamericana de la Salud. (s. f.). Calidad del aire. Recuperado el 27 de septiembre de 2025 de https://www.paho.org/es/temas/calidad-aire

Ortiz, L., Sánchez-Salinas, E. y Castrejón-Godínez, M. L. (2015). La calidad del aire como instrumento para el desarrollo de gestión y política ambiental. En *Los indicadores ambientales como herramienta para la sustentabilidad* (pp. 171-209). Universidad Autónoma del Estado de Morelos. Recuperado el 27 de septiembre de 2025 de https://www.researchgate.net/publication/380360046_La_calidad_del_aire_como_instrumento_para_el_desarrollo_de_gestion_y_politica_ambiental

PNUMA. (2022). ¿Cómo se mide la calidad del aire? *Programa de las Naciones Unidas para el Medio Ambiente*. Recuperado el 27 de septiembre de 2025 de https://www.unep.org/es/noticias-y-reportajes/reportajes/como-se-mide-la-calidad-del-aire

Redalyc. (s. f.). Búsqueda de artículos (Calidad del aire). Recuperado el 27 de septiembre de 2025 de https://www.redalyc.org/busquedaArticuloFiltros.oa?q=Calidad+del+aire

Roig, J. G. (s. f.). Metodologías y estándares. *Universitat Oberta de Catalunya*. Recuperado el 27 de septiembre de 2025 de https://openaccess.uoc.edu/server/api/core/bitstreams/20b39332-4ac7-49e6-b9cc-c34c4401d380/content

Sáenz, R. (1999). Monitoreo de la calidad del aire en América Latina. *Organización Panamericana de la Salud*. Recuperado el 27 de septiembre de 2025 de https://iris.paho.org/bitstream/handle/10665.2/55453/monitoreocalidadal_spa.pdf

Surmay, R. M., Jara, J., Valdés, A., Pérez, M. y Díaz, C. (2024). Monitoreo de la calidad del aire urbano utilizando análisis funcional: caso Chile. *Revista Innovación e Ingeniería*. Recuperado el 27 de septiembre de 2025 de https://revistas.unisimon.edu.co/index.php/innovacioning/article/view/7067

Un mal día para el aire. (2017). *NIH Noticias de Salud*. Recuperado el 27 de septiembre de 2025 de https://salud.nih.gov/recursos-de-salud/nih-noticias-de-salud/un-mal-dia-para-el-aire

Universidad Simón Bolívar. (s. f.). Implementación de CRISP-DM como herramienta de apoyo. Recuperado el 27 de septiembre de 2025 de https://bibliotecadigital.usb.edu.co/entities/publication/50dc8656-5951-4127-a220-2993ca8681b0

Wang, K., Ling, C., Chen, Y. y Zhang, Z. (2023). Spatio-temporal Joint Modelling on Moderate and Extreme Air Pollution in Spain. *arXiv*. https://arxiv.org/abs/2302.06059

Yang, Y., Zheng, Z., Bian, K., Song, L. y Han, Z. (2017). Realtime Profiling of Fine-Grained Air Quality Index Distribution using UAV Sensing. *arXiv*. https://arxiv.org/abs/1711.02821