Universidad Tecnológica de Chihuahua

Tecnologías de la Información y Comunicación

Área: Desarrollo de Software

Extracción de Conocimiento en Bases de Datos

Enrique Mascote

Reporte de solución de caso de estudio en la que presente objetivo, alcance, justificación de la metodología y planeación de las etapas para el análisis de datos.

IDGS91N

PRESENTA:
SEBASTIÁN ACOSTA ORTIZ
IAN CARLOS CHÁVEZ ROJO
REGINA CHÁVEZ TAMAYO
IVÁN EDUARDO MARTÍNEZ MARTÍNEZ
ERICK FABIÁN TERRAZAS HERNÁNDEZ

Chihuahua, Chih., Mex.
Fecha de realización de la práctica: 27/09/2025
Fecha de entrega el reporte: 27/09/2025

índice

Introducción

El incremento constante en la demanda de los servicios de urgencias hospitalarias ha planteado grandes retos en la administración eficaz del tiempo y de los recursos médicos. Frente a esta situación, tecnologías emergentes como la inteligencia artificial, el aprendizaje automático (machine learning), la minería de datos (data mining) y el análisis de grandes volúmenes de información (big data) se convierten en herramientas estratégicas para optimizar los procesos de decisión tanto clínicos como operativos.

El presente informe expone la propuesta elaborada por el equipo de trabajo para enfrentar la problemática relacionada con los largos tiempos de espera en el área de urgencias del Hospital Universitario Salud Total. La iniciativa consiste en el diseño de un modelo de regresión predictiva que permita calcular de forma confiable los tiempos de atención a los pacientes, empleando para ello información histórica, operativa y contextual.

El documento desarrolla de manera detallada el propósito del proyecto, su alcance, las fuentes de datos empleadas, el tipo de análisis realizado y la justificación de la metodología seleccionada. De igual modo, se presenta la planificación por fases, considerando tanto los aspectos técnicos como los éticos, con la intención de entregar una solución factible, escalable y acorde con las necesidades reales del entorno hospitalario

Objetivo del proyecto

El Hospital Universitario Salud Total enfrenta como principal desafío la saturación en el área de urgencias, lo que ocasiona tiempos de espera prolongados, insatisfacción en los pacientes y una administración poco eficiente de los recursos médicos y espacios físicos. Si bien actualmente se recopilan datos sobre cada paciente (edad, síntomas, nivel de dolor, tiempos de llegada y atención), estos no se utilizan de manera sistemática para anticipar la demanda y mejorar la gestión del servicio.

El objetivo del proyecto es desarrollar un sistema predictivo que estime los tiempos de espera en urgencias mediante un modelo de regresión entrenado con datos históricos y variables contextuales (como hora y día de la llegada, brotes epidémicos locales o carga operativa). Este sistema incorporará un esquema híbrido de procesamiento: batch para el entrenamiento diario del modelo y streaming para la actualización de predicciones en tiempo casi real. De esta forma, se busca optimizar la asignación de personal y salas, reducir la saturación en urgencias y elevar la calidad de la atención brindada.

En términos específicos, se pretende:

- Anticipar el tiempo estimado que un paciente esperará desde su ingreso hasta recibir atención médica.
- Incorporar variables clínicas (triage: edad, síntomas, presión arterial, nivel de dolor) y contextuales (hora del día, día de la semana, brotes epidémicos).
- Generar información en tiempo casi real que apoye la toma de decisiones operativas del personal médico y administrativo.

Alcance

El proyecto se enfocará en el diseño y validación de un modelo predictivo de regresión para estimar tiempos de espera en urgencias, incluyendo el desarrollo de un pipeline de datos con procesos de ingestión, limpieza, preparación y actualización continua.

Fuentes de datos contempladas:

- Internas: Registros históricos de urgencias (datos de llegada, triage, tiempos de atención, diagnóstico).
- Contextuales: Hora y día de la semana, brotes epidemiológicos locales, disponibilidad de salas y personal médico por turno.

Tipo de análisis a implementar:

- Modelado de regresión supervisada (ej.: Random Forest, XGBoost) para predecir tiempos continuos en minutos u horas.
- Procesamiento en batch para entrenamientos diarios del modelo con registros históricos.
- Procesamiento en streaming para actualizar predicciones cada hora en función de los ingresos recientes.

Productos esperados:

- Modelo predictivo entrenado y evaluado.
- Pipeline de datos con ingestión, limpieza, integración y actualización.
- Reportes de tiempos de espera accesibles desde el centro de control hospitalario.

Limitaciones del proyecto:

- No se incluirá integración directa con expedientes clínicos electrónicos de otras áreas del hospital.
- No se considerarán variables externas como tráfico vehicular o disponibilidad de ambulancias.
- La validación se limitará a datos internos del Hospital Universitario Salud Total, sin extenderse a otras instituciones.
- El uso de datos sensibles requerirá procesos de anonimización y aprobación del comité de ética.

Justificación de la metodología

Para este proyecto se implementará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), un marco reconocido y ampliamente aplicado en proyectos de ciencia de datos por su enfoque ordenado y adaptable. CRISP-DM se compone de seis fases cíclicas (comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue), lo cual facilita la adaptación a ajustes y mejoras constantes durante la construcción del modelo. Esta metodología resulta especialmente útil porque vincula desde el inicio los objetivos organizacionales con los requerimientos técnicos, previniendo desajustes habituales en proyectos analíticos. Asimismo, su orientación hacia la documentación y la evaluación rigurosa asegura la trazabilidad y el perfeccionamiento continuo del modelo, factores cruciales en entornos dinámicos como la gestión hospitalaria.

¿Por qué esa metodología?

CRISP-DM resulta apropiado para la predicción de tiempos de espera en urgencias por diversas razones:

Adaptación al entorno hospitalario: La etapa de comprensión del negocio posibilita establecer métricas clave (como disminuir en un 20% los errores de predicción) y priorizar variables clínicas y operativas (ej.: triage, afluencia de pacientes). Esto garantiza que el modelo responda a necesidades concretas del hospital, como optimizar personal y recursos físicos.

Carácter iterativo: La saturación en urgencias constituye un fenómeno cambiante, afectado por elementos como brotes epidémicos o picos de horario. CRISP-DM facilita la retroalimentación entre fases (ej.: reentrenar modelos con datos en tiempo real) para ajustarse a variaciones repentinas.

Énfasis en la calidad de los datos: La fase de preparación es esencial para tratar registros faltantes (ej.: síntomas omitidos) y variables contextuales (como festivos), un reto frecuente en la práctica clínica.

Como antecedente, investigaciones recientes en hospitales evidencian que CRISP-DM incrementa la precisión de los modelos predictivos en urgencias al combinar información histórica y variables operativas en procesos iterativos. Además, su flexibilidad ante técnicas modernas (como la quantile regression para estimar percentiles de espera) lo convierte en una alternativa más efectiva que marcos rígidos como SEMMA, cuyo enfoque se orienta principalmente al análisis estadístico sin priorizar la alineación con el negocio.

Fuentes clave:

Guía oficial de CRISP-DM.
Estudios de aplicación en urgencias, como (PubMed) y (Journal of Archives in Military Medicine).
Casos de éxito en salud con CRISP-DM.

Planeación de etapas

En esta parte se describen las fases, actividades y productos del proyecto, en concordancia con la metodología CRISP-DM. El plan de trabajo se organiza en 6 etapas iterativas, diseñadas para asegurar una implementación eficaz del modelo predictivo, priorizando la calidad de los datos, la evaluación constante y la flexibilidad frente a los requerimientos del hospital. A continuación, se expone un desglose en formato tabular, que incluye las actividades principales y un calendario preliminar de 10 semanas (ajustable de acuerdo con la disponibilidad de recursos o posibles contingencias).

Tabla de las fases del proyecto

Fase	Actividades clave	Entregables	Semanas
1. Comprensión del negocio	Sesiones con los actores clave (médicos, personal administrativo), establecimiento de indicadores clave de desempeño (ej.: margen máximo de error en minutos).	Documento con los requerimientos y criterios de éxito.	1–2
2. Comprensión de datos	Exploración inicial (EDA), identificación de fuentes internas/externas y variables relevantes (ej.: horario pico).	Entregable: informe EDA + diccionario de datos.	3

3. Preparación de datos	Depuración (valores faltantes, atípicos), unificación de datasets (histórico + streaming), generación de variables.	Entregable: dataset listo y pipeline de preprocesamiento.	4-5
4. Modelado	Elección de algoritmos (Random Forest, XGBoost), entrenamiento mediante validación cruzada.	Entregable: prototipo del modelo + código de entrenamiento.	6-7
5. Evaluación	Medición con métricas (MAE, RMSE), comparación de modelos y ajuste de parámetros.	Entregable: reporte de desempeño con recomendaciones.	8
6. Despliegue	Implementación de pipeline batch/streaming (ej.: Apache Airflow), documentación para el personal médico.	Entregable: scripts de inferencia + manual de uso.	9-10

Notas adicionales:

- Iteratividad: Las fases 3 a 5 (preparación, modelado y evaluación) pueden repetirse en caso de requerirse la integración de nuevos datos o ajustes de requisitos.

- Riesgos: Posibles demoras en la obtención de información (ej.: autorizaciones éticas) podrían impactar el periodo de las semanas 3–5. Se sugiere considerar un margen adicional de 1–2 semanas.

Conclusión

La implementación del sistema predictivo para tiempos de espera en el servicio de urgencias del Hospital Universitario Salud Total ha seguido un marco metodológico robusto basado en CRISP-DM, estableciendo un proceso estructurado y flexible que va desde el análisis inicial del contexto hasta la puesta en marcha del modelo operativo. Esta aproximación metodológica facilita la convergencia entre las demandas institucionales del centro médico —tales como el desahogo de la congestión y la gestión eficiente de recursos— y las potencialidades técnicas del sistema de predicción, verificando que cada etapa, desde el procesamiento de información hasta la valoración de rendimiento, cuente con la validación de los actores estratégicos involucrados.

El esquema de trabajo diseñado, distribuido en 6 etapas durante un período de 10 semanas, no solamente satisface las metas establecidas inicialmente (anticipar duraciones mediante información histórica y situacional), sino que también contempla obstáculos operativos como la integridad de la información o las autorizaciones institucionales, aprovechando su naturaleza cíclica de mejora continua. La etapa de acondicionamiento de datos incorpora mecanismos para gestionar información fragmentada, mientras que la fase de implementación enfatiza la capacidad de expansión del modelo para incorporar actualizaciones constantes sin comprometer la estabilidad tecnológica hospitalaria.

Sin embargo, más allá de los aspectos técnicos, metodológicos y las herramientas empleadas, esta iniciativa nos ha proporcionado una perspectiva más profunda sobre cómo la innovación tecnológica puede convertirse en un puente hacia la solución de problemáticas tangibles que impactan el bienestar cotidiano de las

personas. Abordar una circunstancia compleja como la saturación en servicios de emergencia nos llevó a reflexionar sobre la relevancia de desarrollar propuestas que generen transformaciones concretas en la experiencia humana, trascendiendo el ámbito puramente conceptual o investigativo.

Esta experiencia nos permitió fortalecer nuestras competencias colaborativas, desarrollar criterios de decisión fundamentados en la responsabilidad social, y cultivar un pensamiento analítico que va más allá de las soluciones convencionales, reconociendo que los verdaderos desafíos requieren enfoques integrales y comprometidos con el impacto real en la comunidad.

Glosario de Conceptos Clave de la Tabla

1. Stakeholders: Personas clave que participan en el proyecto (ej.: médicos, directivos del hospital).
2. KPIs (Indicadores Clave de Desempeño): Medidas para evaluar el progreso del proyecto (ej.: "El modelo no debe fallar más de 10 minutos en sus estimaciones").
3. EDA (Exploratory Data Analysis): Análisis inicial de los datos para identificar tendencias (ej.: "¿Los viernes aumenta la cantidad de pacientes?").
4. Dataset: Conjunto de datos con información de los pacientes (edad, síntomas, tiempos de atención, entre otros).
5. Depuración de datos: Corrección de inconsistencias (ej.: eliminar registros incompletos o repetidos).
6. Feature engineering: Generación de nuevas variables relevantes para el modelo (ej.: "¿Corresponde a fin de semana? Sí/No").
7. Algoritmos (Random Forest, XGBoost): Métodos computacionales que permiten estimar los tiempos de espera.
8. Entrenamiento del modelo: Proceso de enseñanza a la máquina utilizando información histórica para aprender a predecir.
9. Validación cruzada: Evaluación del modelo con diferentes subconjuntos de datos para confirmar su fiabilidad.

10. Métricas (MAE, RMSE): Indicadores para cuantificar los errores (ej.: "El modelo presenta un promedio de 5 minutos de diferencia").
11. Pipeline (batch/streaming):
    a. Batch: Procesamiento de datos en lotes (ej.: durante la noche).
    b. Streaming: Procesamiento de datos en tiempo real (ej.: cada hora).
12. Scripts de inferencia: Programas que aplican el modelo entrenado para generar predicciones nuevas.

Referencias

Author, G. (2021, 18 mayo). Qué son los stakeholders, qué tipos existen y de qué manera impactan a una empresa. Rock Content - ES. https://rockcontent.com/es/blog/que-es-un-stakeholder/

Twin, A. (2025, 24 enero). KPIs: What Are Key Performance Indicators? Types and Examples. Investopedia. https://www-investopedia-com.translate.goog/terms/k/kpi.asp?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto

=tc

Análisis exploratorio de datos. (s. f.). https://www.jmp.com/es/statistics-knowledge-portal/exploratory-data-analysis

Lab, R. I. (2025, 2 junio). Qué es un dataset. The Information Lab. https://www.theinformationlab.es/blog/que-es-un-dataset/

Guide to Data Cleaning: Definition, benefits, components, and how to clean your data. (s. f.). Tableau. https://www.tableau.com/learn/articles/what-is-data-cleaning

¿En qué consiste la ingeniería de características? - Explicación sobre la ingeniería de características - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/feature-engineering/#:~:text=La%20ingenier%C3%ADa%20de%20caracter%C3%ADsticas% 20implica,el%20entrenamiento%20y%20la%20predicci%C3%B3n.

Zúñiga, J. J. E. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería Investigación y Tecnología, 21(3), 1-16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

¡Cómo entrenar un modelo de machine learning paso a paso! (s. f.). Tokio School. https://www.tokioschool.com/noticias/como-entrenar-modelo-machine-learning/

3.1. Cross-validation: evaluating estimator performance.(s. f.). Scikit-learn. https://scikit--learn-org.translate.goog/stable/modules/cross_validation.html?_x_tr_sl=en&_x_tr_tl=es&_ x_tr_hl=es&_x_tr_pto=tc

Pandelu, A. P. (2024, 13 noviembre). Day 10: Evaluation Metrics for Regression — MSE, MAE, RMSE, R2 Score.Medium. https://medium-com.translate.goog/@bhatadithya54764118/day-10-evaluation-metrics-for-regression-mse-mae-rmse-r%C2%B2-score-0ffb39e3ea26?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc