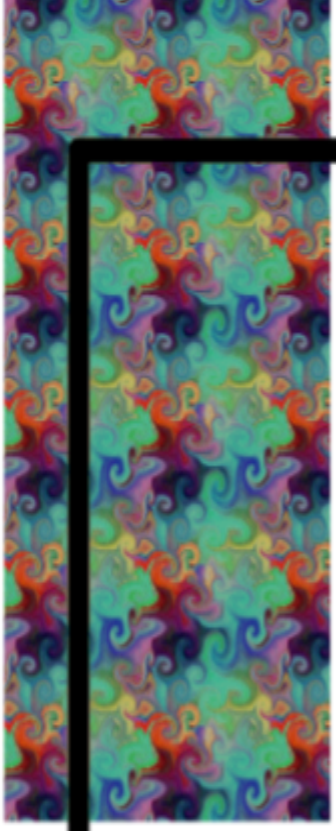

Actividad 2: Comparación de métodos

CE Inteligencia Artificial y Big Data

Sistemas de Big Data

2024/2025

Daniel Marín López
Carlos Chaves Hernández

Índice

1. Enunciado

Una empresa de servicios de salud quiere reducir las ausencias injustificadas de pacientes a sus citas médicas. Sospechan que hay patrones en los historiales de pacientes y el tipo de cita que podrían predecir si un paciente no asistirá. Te piden desarrollar un modelo para anticipar estas ausencias y tomar medidas preventivas (como enviar recordatorios o cambiar el tipo de cita).

2. Aplicación de métodos

Para ver cómo se ejecutarán los distintos métodos usaremos una tabla como se hizo anteriormente:

Método	Fase	Actividad
CRISP-DM	Comprensión del negocio	Reducir las ausencias injustificadas a las citas médicas mediante predicción proactiva, lo que tendría un impacto en la reducción de espera para una cita, mejora en la atención al paciente, Incremento de ingresos al reducir huecos no aprovechados.
	Comprensión de los datos	Demográficos: edad, sexo, ubicación, nivel socioeconómico. Historial: asistencia previa, número de citas canceladas/fallidas. Tipo de cita: especialidad, urgencia, presencial/virtual. Información contextual: día y hora, distancia a la clínica, recordatorios enviados, clima. Variables de comportamiento: llamadas previas, cambios de cita, comentarios administrativos.
	Preparación de los datos	Limpieza: imputar valores nulos, tratar outliers (e.g., edad > 120). Ingeniería de variables: • tasa_asistencia = nº_asistencias / nº_total_citas • día_semana, hora_pico (categorías de horario) • recibió_recordatorio (sí/no) • distancia_a_clínica (geocodificada si hay dirección) Codificación de variables categóricas y normalización de variables numéricas.
	Modelado	Clasificación binaria (asiste vs no asiste). Algoritmos candidatos pueden ser Árboles de decisión, Random Forest, XGBoost, Regresiones logísticas y Redes neuronales simples.

	Evaluación	Si se predicen con una precisión aceptable, por ejemplo ≥ 75% los pacientes que no asistirán, se puede usar para enviar recordatorios u otras medidas.
	Despliegue	Implementación: • API que reciba datos del paciente y devuelva probabilidad de ausencia. • Integración en el sistema de gestión de citas. Usuarios: • Personal administrativo para la planificación de llamadas. • Envío automático de recordatorios personalizados.
TDSP	Comprensión del problema empresarial	**Objetivo de negocio:** Reducir las ausencias injustificadas de pacientes a sus citas médicas. **Pregunta de ciencia de datos:** ¿Podemos predecir con antelación si un paciente no asistirá a su cita? **Tipo de problema:** Clasificación binaria (asiste = 0, no asiste = 1).
	Adquisición y comprensión de los datos	Podemos obtener los datos mediante sistemas de agendamiento, historiales médicos, CRM (para datos del paciente). Estos luego podemos procesarlos con programas como Spark y almacenarlos en HDFS o en Data Lakes. La calidad se puede sustentar mediante validaciones automáticas (formatos, valores nulos, duplicados), trazabilidad (mantener logs del pipeline y versionado de datasets) y verificación semántica (coherencia en fechas).
	Modelado	Se pueden probar varios modelos de clasificación como Regresión Logística, Random Forest o incluso Redes Neuronales. Podríamos revisar métricas como Recall y Precision para ver cómo de bueno es nuestro modelo. Usar validación cruzada para encontrar los mejores hiperparámetros, y que su interpretabilidad sea fácil para los médicos o administradores del hospital.
	Despliegue	Usando librerías como Flask se podría integrar una API Rest. El modelo se empaqueta y expone como endpoint para predicción en tiempo real o batch. Este se integraría con el sistema de agendamiento (EMR o CRM interno). Puede disparar alertas, recordatorios automáticos o reprogramaciones sugeridas.
	Aceptación del cliente	Los resultados pueden estar recogidos en un dashboard con las métricas de nuestro modelo. También se recogerán los casos reales acertados por el modelo y se puede mostrar qué variables son influyentes en el modelo. Se puede probar durante un tiempo corto y ver el feedback que el personal del hospital presenta ante las predicciones para ajustar el modelo.

ASUM-DM	Análisis	Necesidades del usuario: • Reducir citas vacías • Mejorar la planificación y comunicación con pacientes Entregables esperados: • Modelo predictivo listo para producción • Documentación y panel de control con alertas • Plan de mejora continua
	Diseño	Recursos técnicos: • Científicos de datos, ingenieros de datos, equipo TI • Infraestructura cloud (Azure, AWS o GCP) Entorno: • Desarrollo y pruebas en entorno cloud (Databricks o JupyterHub) • Producción en contenedor (Docker + Kubernetes si es escalable)
	Configuración y construcción	Ciclo 1: EDA + primer modelo básico Ciclo 2: optimización del modelo y métricas Ciclo 3: integración en sistemas reales
	Despliegue	Despliegue progresivo por centros. Supervisión paralela con sistema actual para evitar impactos. Posibilidad de revertir si hay fallos.
	Operación y optimización	Monitorización del rendimiento del modelo. Reentrenamiento mensual o si cambia el patrón de comportamiento. Logging de predicciones erróneas para revisión.

	Gestión del proyecto	Roles definidos: • Project Manager: coordinación y planificación. • Data Scientist: desarrollo del modelo. • Data Engineer: ETL y arquitectura. • DevOps: despliegue y monitorización. • Stakeholders clínicos: validación y feedback. Tareas: • Planificación de entregas iterativas (Scrum o Kanban). • Reuniones quincenales con stakeholders. • Documentación y checklist de producción.

3. Ventajas y desventajas

Una vez vistas todas las metodologías, podemos ver unas claras ventajas y desventajas de cada una de ellas:

Metodología	Ventajas	Desventajas
CRISP-DM	• Flexible y ampliamente adoptado: Ideal para proyectos exploratorios como el tuyo, sin requerimientos técnicos específicos. • Enfocado en el negocio: Comienza por comprender los objetivos del negocio, algo esencial cuando se quiere reducir ausencias médicas. • Iterativo: Permite refinar el modelo de forma continua conforme se aprende más del problema.	• No está orientado a la producción: No cubre detalles específicos sobre cómo desplegar modelos ni automatizarlos. • Poca formalización técnica: Requiere que el equipo complemente con prácticas de ingeniería para implementación moderna (APIs, DevOps, etc.).
TDSP	• Enfocado en implementación: Diseñado para proyectos que requieren desplegar modelos predictivos en producción. • Basado en herramientas modernas: Integra	• Más técnico: Puede ser excesivo si el equipo no tiene experiencia en MLOps o si el proyecto es exploratorio. • Ligero en exploración de negocio: Aunque contempla los objetivos, su foco está en

	fácilmente entornos colaborativos, control de versiones, pipelines y CI/CD. • Ideal para equipos técnicos: Estructura bien definida para equipos multidisciplinarios (científicos de datos, ingenieros, PMs).	el desarrollo técnico del modelo.
ASUM-DM	• Combina negocio y tecnología: Integra bien las fases de entendimiento del negocio, análisis de datos y planificación de despliegue. • Buen equilibrio entre agilidad y estructura: Es más moderno que CRISP-DM, con énfasis en gestión de proyectos y solución empresarial. • Soporta cambios: Diseñado para ambientes cambiantes y evolución de modelos.	• Menos conocido: Requiere más capacitación del equipo si no está familiarizado. • Más pesado que CRISP-DM: Puede implicar burocracia innecesaria si el equipo es pequeño o el caso es de baja complejidad.

Lo que se recomienda para cada caso es lo siguiente:

- Si es la primera vez que abordan un problema predictivo con fines de mejora operativa, se puede usar CRISP-DM que es un buen punto de partida.
- Si el objetivo es crear un sistema productivo con alarmas, recordatorios automáticos o integración con sistemas clínicos, TDSP será más adecuado.
- Si se desea una visión de negocio bien alineada con procesos analíticos y potencial de escalabilidad futura, usar ASUM-DM puede ser el mejor balance.