📊 ¿Qué es CRISP-DM?

CRISP-DM (Cross-Industry Standard Process for Data Mining) es un modelo estándar utilizado para llevar a cabo proyectos de minería de datos de manera estructurada y eficiente. Sirve como guía metodológica para resolver problemas mediante el análisis de datos.

🔍 Características principales:

- Estructurado en 6 fases:
    1. Entendimiento del negocio
    2. Entendimiento de los datos
    3. Preparación de los datos
    4. Modelado
    5. Evaluación
    6. Despliegue
- Iterativo y flexible: Las fases no son estrictamente lineales; es común volver atrás y ajustar.
- Independiente de la industria: Puede aplicarse tanto en salud, finanzas, retail, etc.
- Enfoque en el negocio: Comienza entendiendo el problema desde una perspectiva no técnica, lo que asegura que los resultados aporten valor real.

📊 CRISP-DM y su aplicación en nuestro proyecto

CRISP-DM (Cross-Industry Standard Process for Data Mining) es un modelo estándar que guía el desarrollo de proyectos de minería de datos a través de seis fases estructuradas. Fue creado para asegurar que los proyectos no solo se enfoquen en lo técnico, sino que realmente resuelvan problemas del mundo real.

📊 Resumen por Fases – CRISP-DM en el Proyecto Neumonía

Fase	Acciones Clave	Herramientas / Enfoques
1. Entendimiento del Negocio	Definición del problema clínico, impacto esperado y restricciones éticas	Análisis contextual, revisión bibliográfica, entrevistas con usuarios potenciales
2. Comprensión de los Datos	Revisión de imágenes, eliminación de datos corruptos, análisis de distribución de clases	Exploración visual y estadística, verificación manual, detección de desbalance
3. Preparación de los Datos	Limpieza, normalización, redimensionamiento, codificación y aumento de datos	Técnicas de preprocesamiento, One-Hot Encoding, data augmentation (Keras, OpenCV)
4. Modelado	Diseño e implementación de CNN básica para clasificación multiclase	TensorFlow/Keras, optimizador Adam, función Categorical Crossentropy, métricas múltiples
5. Evaluación e Implantación	Evaluación con datos de prueba, métricas por clase, discusión de aplicabilidad clínica futura	Accuracy, Precision, Recall, AUC. Consideraciones éticas y de integración futura en hospitales

🔄 Fases del modelo CRISP-DM aplicadas al proyecto:

Entendimiento del negocio

→ Identificamos la problemática clínica: la necesidad de detectar neumonía de forma rápida y precisa a partir de radiografías, sobre todo en contextos con pocos especialistas.

Entendimiento de los datos

→ Analizamos la calidad, el origen y la distribución de las imágenes radiográficas. Identificamos clases desbalanceadas (por ejemplo, más casos normales que virales).

Preparación de los datos

→ Incluyó limpieza, normalización, redimensionamiento de imágenes y técnicas de aumento de datos para mejorar la generalización del modelo.

Modelado

→ Construimos una red neuronal convolucional (CNN) con capas convolucionales y densas para clasificar entre neumonía viral, bacteriana y normal.

Evaluación

→ Usamos métricas como precisión, recall y AUC para comprobar el rendimiento del modelo y entender cómo responde ante cada clase.

Despliegue (parcial)

→ Si bien no hicimos un despliegue clínico, el modelo quedó preparado para ser implementado como herramienta de apoyo diagnóstico en futuras iteraciones.

✅ ¿Por qué usamos CRISP-DM?

"Porque es un modelo flexible, iterativo y centrado en el problema real. Nos permitió alinear los pasos técnicos con los objetivos del negocio médico, asegurando que el trabajo tenga un impacto más allá del aula."

📝 Notas para CRISP-DM – Proyecto Neumonía

🔷 1. Entendimiento del Negocio

- Problema: Diagnóstico de neumonía depende de radiólogos, lo cual es lento y limitado en algunos contextos.
- Objetivo: Crear una herramienta de apoyo para agilizar y mejorar el diagnóstico.
- Impacto: Mayor eficiencia, reducción de costos y mejor atención médica.
- Rol de IA: Apoyo, no reemplazo del profesional.
- Restricciones: Calidad de datos, explicabilidad del modelo, validación clínica.

🔷 2. Comprensión de los Datos

- Origen de datos: Radiografías públicas anotadas (3 clases).
- Calidad revisada: Se eliminaron imágenes corruptas/mal etiquetadas.
- Problemas detectados: Desbalance de clases, variabilidad visual.

🔷 3. Preparación de los Datos

- Procesos aplicados:

    - Limpieza y redimensionamiento.
    - Normalización [0–1].
    - One-Hot Encoding.
    - División: 70/15/15 (entrenamiento/validación/prueba).
- Aumento de datos: Rotación, zoom, flips, contraste.

🔷 4. Modelado

- Modelo: CNN secuencial con capas convolucionales y densa.
- Optimización: Adam (lr=0.0001), Categorical Crossentropy.
- Métricas: Accuracy, Precisión, Recall, AUC.
- Prevención de sobreajuste:
    - EarlyStopping
    - ReduceLROnPlateau
    - Validación cruzada
    - Aumento de datos

🔷 5. Evaluación e Implantación

- Resultados: Buen rendimiento general, especialmente en clase bacteriana.
- Limitaciones: Recall bajo en clase *normal*.
- Implantación:
    - Aún experimental.
    - Requiere validación clínica.
    - Potencial uso en sistemas de salud públicos y rurales.

❓ Posibles preguntas

🧠 1. Entendimiento del negocio

¿Por qué eligieron utilizar el modelo CRISP-DM para este proyecto?

Respuesta sugerida:

Elegimos CRISP-DM porque es una metodología robusta, flexible y ampliamente aceptada en la comunidad de ciencia de datos. Nos permitió estructurar el proyecto desde una perspectiva clara y ordenada, comenzando por entender el problema clínico real que queríamos resolver, y asegurándonos de alinear cada etapa técnica con un objetivo práctico y aplicable. Además, al ser un modelo iterativo, nos permitió ir ajustando nuestro enfoque a medida que avanzábamos con el análisis y desarrollo.

¿Cómo definieron los objetivos de negocio en esta fase?

Respuesta sugerida:

Partimos del análisis de la situación clínica actual: la necesidad de agilizar y mejorar el diagnóstico de neumonía, especialmente en contextos con pocos recursos. A partir de eso, establecimos como objetivo general desarrollar una herramienta de apoyo basada en deep learning que pudiera identificar automáticamente casos sospechosos en imágenes de rayos X. Luego lo vinculamos con beneficios concretos para el sistema de salud: mayor eficiencia, reducción de costos por diagnósticos tardíos y mejora en la calidad del servicio médico.

¿Cuáles fueron las principales restricciones o desafíos detectados en esta fase?

Respuesta sugerida:

Detectamos varias limitaciones clave: primero, la disponibilidad y calidad de los datos médicos —ya que deben estar correctamente anotados y ser representativos—; segundo, la responsabilidad ética, ya que estamos trabajando en un área sensible como la salud, donde cualquier error puede tener consecuencias graves; y tercero, la necesidad de explicabilidad del modelo, para que los profesionales puedan confiar en sus resultados y entender cómo se toman las decisiones.

¿Qué impacto esperan que tenga el proyecto en el entorno clínico?

Respuesta sugerida:

Esperamos que esta herramienta pueda actuar como un soporte diagnóstico que permita priorizar casos críticos, reducir los tiempos de respuesta y extender el alcance del diagnóstico a lugares con menos recursos humanos especializados. A largo plazo, puede contribuir a una mejor toma de decisiones médicas y a una distribución más equitativa de los servicios de salud.

🧩 2. Comprensión de los Datos

¿Qué tipo de datos utilizaron y cómo validaron su calidad?

Respuesta sugerida:

Utilizamos un conjunto de imágenes de radiografías de tórax provenientes de bases de datos públicas, que incluían tres clases: neumonía viral, neumonía bacteriana y casos normales. Para validar la calidad de los datos, realizamos una inspección manual inicial para eliminar imágenes corruptas, duplicadas o mal anotadas. También analizamos la distribución de clases y las características visuales generales para asegurarnos de que los datos fueran adecuados para el entrenamiento de un modelo de clasificación.

¿Detectaron algún problema con los datos durante esta fase?

Respuesta sugerida:

Sí, uno de los principales desafíos fue el desbalance entre clases. Por ejemplo, había más imágenes de neumonía bacteriana que de viral. También notamos cierta variabilidad en el contraste y tamaño de las imágenes, lo cual influye directamente en el aprendizaje del modelo. Estas observaciones nos guiaron en las decisiones que tomamos durante la preparación de los datos.

🩹 3. Preparación de los Datos

¿Qué técnicas aplicaron durante la preparación de los datos?

Respuesta sugerida:

Realizamos varias tareas clave: limpieza de datos (eliminando imágenes no útiles), redimensionamiento a un tamaño uniforme (224x224 píxeles), normalización de valores de píxeles y codificación One-Hot de las etiquetas. Además, dividimos los datos en tres conjuntos: entrenamiento (70%), validación (15%) y prueba (15%) para asegurar una evaluación justa del modelo.

¿Utilizaron aumento de datos? ¿Por qué?

Respuesta sugerida:

Sí, aplicamos técnicas de aumento de datos como rotación, zoom, flips horizontales y ajustes de contraste. Estas técnicas nos permitieron generar más variaciones de las imágenes originales, lo que ayudó a reducir el sobreajuste y mejorar la capacidad del modelo para generalizar a nuevos datos.

🧠 4. Fase de Modelado

¿Qué tipo de modelo entrenaron y por qué eligieron esa arquitectura?

Respuesta sugerida:

Diseñamos una red neuronal convolucional (CNN) secuencial, ya que este tipo de arquitectura es particularmente efectiva para el procesamiento de imágenes médicas. Las CNN son capaces de extraer automáticamente características relevantes sin requerir ingeniería manual de atributos. Optamos por una arquitectura sencilla en esta etapa inicial para validar el enfoque, con planes de incorporar modelos preentrenados en futuras iteraciones.

¿Cómo configuraron el modelo y qué métricas utilizaron?

Respuesta sugerida:

Utilizamos el optimizador Adam con una tasa de aprendizaje de 0.0001 y la función de pérdida *Categorical Crossentropy*, ya que se trata de un problema de clasificación multiclase. Evaluamos el modelo utilizando métricas como *accuracy*, *precisión*, *recall* y *AUC*, para obtener una visión más completa del rendimiento en cada clase.

¿Qué medidas tomaron para evitar el sobreajuste?

Respuesta sugerida:

Implementamos técnicas como el *Early Stopping*, que detiene el entrenamiento cuando la mejora se estabiliza, y *ReduceLROnPlateau*, que ajusta dinámicamente la tasa de aprendizaje. También hicimos uso de validación cruzada y aumento de datos, lo que ayudó a evitar que el modelo memorice los datos de entrenamiento.

✅ 5. Evaluación e Implantación

¿Cómo evaluaron el rendimiento del modelo?

Respuesta sugerida:

Evaluamos el modelo usando el conjunto de prueba, que el modelo no había visto antes. Las métricas mostraron un rendimiento sólido en general: por ejemplo, en la clase *bacterial pneumonia* se obtuvo una precisión de 0.80 y un recall de 0.86. La clase *normal* tuvo alta precisión pero bajo recall, lo que nos indica que hay margen de mejora en la detección de falsos negativos.

¿Planean implementar este modelo en un entorno clínico real?

Respuesta sugerida:

Nuestro enfoque actual es exploratorio y experimental. Aun así, los resultados obtenidos demuestran que es factible avanzar hacia una aplicación clínica, siempre que se cumplan requisitos éticos y normativos. Para llegar a una implantación real, sería necesario realizar validaciones clínicas con profesionales de la salud, integrar explicabilidad en el modelo, y asegurar que el sistema pueda integrarse a los flujos de trabajo hospitalarios existentes.

¿Qué impacto esperan de su posible implementación?

Respuesta sugerida:

Esperamos que esta herramienta actúe como un sistema de apoyo a la toma de decisiones clínicas, permitiendo identificar rápidamente los casos sospechosos de neumonía. En términos prácticos, puede reducir tiempos de diagnóstico, optimizar recursos y mejorar la calidad del servicio, especialmente en hospitales con escasez de especialistas.

🎯 **Preguntas Desafiantes (y cómo responderlas)**

¿Por qué no usaron un modelo preentrenado como ResNet?

Porque el objetivo inicial era validar la viabilidad del enfoque desde cero. Sin embargo, está en nuestras recomendaciones futuras, ya que modelos como ResNet pueden mejorar la precisión y ahorrar tiempo de entrenamiento.

¿Cómo manejaron el desbalance de clases en el dataset?

Aplicamos técnicas de aumento de datos, y consideramos usar métricas como *recall* por clase para evaluar el rendimiento en lugar de solo accuracy general, que puede ocultar fallos en clases minoritarias.

¿Cómo se asegurarían de que su modelo sea clínicamente aplicable?

Requiere validación con especialistas en un entorno real, evaluación con datos clínicos de diversos orígenes y garantizar interpretabilidad de resultados (por ejemplo, usando mapas de calor tipo Grad-CAM).

¿Qué harían diferente si repitieran el proyecto?

Si repitiéramos el proyecto en un entorno laboral, incorporaríamos modelos preentrenados desde el inicio, usaríamos técnicas más avanzadas de balanceo de clases y trabajaríamos en paralelo con un profesional médico para validar los resultados durante el proceso.