Aplicación de la Metodología CRISP-DM para el Análisis de Datos de Cáncer

Guevara - Hernández

UCV

2025

Introducción

La minería de datos se define como el proceso de descubrir conocimiento o patrones implícitos, desconocidos y potencialmente útiles a partir de los datos [1], [2]. En este proyecto, aplicamos este proceso a un conjunto de datos médicos para facilitar la toma de decisiones clínicas [3].

Justificación de CRISP-DM (I)

La elección de CRISP-DM como marco de trabajo se fundamenta en su posición como estándar de facto en la industria [3, 4].

- **Orientación al Negocio:** A diferencia de modelos puramente técnicos como SEMMA, CRISP-DM inicia con el entendimiento de los objetivos comerciales y criterios de éxito [3, 5].

Justificación de CRISP-DM (I)

La elección de CRISP-DM como marco de trabajo se fundamenta en su posición como estándar de facto en la industria [3, 4].

- **Orientación al Negocio:** A diferencia de modelos puramente técnicos como SEMMA, CRISP-DM inicia con el entendimiento de los objetivos comerciales y criterios de éxito [3, 5].
- **Reducción de Riesgos:** Sistematiza el desarrollo, lo que permite minimizar las probabilidades de reprocesos costosos durante el ciclo de vida del proyecto [6].

Justificación de CRISP-DM (II)

- **Ciclo Iterativo y Flexible:** Su naturaleza no es lineal; permite retroceder entre fases (ej. del Modelado a la Preparación de Datos) según los hallazgos técnicos [7, 8].

Justificación de CRISP-DM (II)

- **Ciclo Iterativo y Flexible:** Su naturaleza no es lineal; permite retroceder entre fases (ej. del Modelado a la Preparación de Datos) según los hallazgos técnicos [7, 8].
- **Visión de Despliegue:** Incluye explícitamente el mantenimiento y monitoreo, fases críticas para gestionar la "deuda técnica.en sistemas de aprendizaje automático [9, 10].

Justificación de CRISP-DM (II)

- **Ciclo Iterativo y Flexible:** Su naturaleza no es lineal; permite retroceder entre fases (ej. del Modelado a la Preparación de Datos) según los hallazgos técnicos [7, 8].
- **Visión de Despliegue:** Incluye explícitamente el mantenimiento y monitoreo, fases críticas para gestionar la "deuda técnica.en sistemas de aprendizaje automático [9, 10].
- **Interdisciplinariedad:** Facilita la colaboración entre expertos del dominio y analistas de datos [11, 12].

Etapas de CRISP-DM (I): Entendimiento

Las fases iniciales sientan las bases del conocimiento necesario para el éxito del proyecto [13].

1. **Entendimiento del Negocio:**
   - Definición de objetivos cuantificables.
   - Evaluación de recursos, riesgos y relación costo-beneficio [14].

Etapas de CRISP-DM (I): Entendimiento

Las fases iniciales sientan las bases del conocimiento necesario para el éxito del proyecto [13].

1. **Entendimiento del Negocio:**
   - Definición de objetivos cuantificables.
   - Evaluación de recursos, riesgos y relación costo-beneficio [14].
2. **Entendimiento de los Datos:**
   - Recolección inicial y descripción de atributos.
   - Análisis de calidad (identificación de valores nulos, anómalos o sesgos) [13, 15].

Etapas de CRISP-DM (II): Acción Técnica

En esta etapa se concentra el mayor esfuerzo computacional y analítico [16, 17].

3. **Preparación de los Datos:**
   - Preprocesamiento estructural (*Tidy Data*) y funcional (escalamiento, codificación) [18, 19].
   - Representa aproximadamente el 80 % del tiempo del proyecto [16, 20].

Etapas de CRISP-DM (II): Acción Técnica

En esta etapa se concentra el mayor esfuerzo computacional y analítico [16, 17].

3. **Preparación de los Datos:**
   - Preprocesamiento estructural (*Tidy Data*) y funcional (escalamiento, codificación) [18, 19].
   - Representa aproximadamente el 80 % del tiempo del proyecto [16, 20].
4. **Modelado:**
   - Selección de algoritmos (clasificación, regresión, agrupación).
   - Configuración de hiperparámetros y diseño de pruebas de robustez [21, 22].

Etapas de CRISP-DM (III): Cierre y Valor

Se verifica que el conocimiento extraído sea realmente útil y aplicable [5, 23].

5. **Evaluación del Modelo:**
   - Comparación de resultados frente a los objetivos de negocio.
   - Análisis de métricas técnicas (Precisión, F1-Score) y lógicas [5, 24].

Etapas de CRISP-DM (III): Cierre y Valor

Se verifica que el conocimiento extraído sea realmente útil y aplicable [5, 23].

5. **Evaluación del Modelo:**
   - Comparación de resultados frente a los objetivos de negocio.
   - Análisis de métricas técnicas (Precisión, F1-Score) y lógicas [5, 24].
6. **Despliegue:**
   - Plan de implementación (APIs, integración en sistemas).
   - Reporte final, documentación técnica y plan de mantenimiento [5, 24].

Estructura y Características del Dataset

Este conjunto de datos (Aquí el link) se compone de características visuales extraídas de muestras celulares para el diagnóstico clínico de cáncer de mama,.

- **Identificador (id):** Código único para cada paciente (no predictivo),.
- **Variable Objetivo (diagnosis):** Clasificación binaria categórica con estados:
  - **M** (Maligno - Cáncer presente).
  - **B** (Benigno - Ausencia de malignidad).
- **Características Visuales (Valores Medios):** Atributos como radio_media, textura_media, perímetro_media, área_media, suavidad_media, compacidad_media, concavidad_media y puntos cóncavos_media,.
- **Atributos Categóricos:** Incluye variables etiquetadas con valores numéricos para análisis estadístico,.

Utilidad del Dataset y Análisis Exploratorio (EDA)

El propósito fundamental es el entrenamiento y validación de algoritmos de diagnóstico médico asistido por computadora,.

- **Distribución de Rangos:** Cada característica se asigna a tablas de frecuencia que contienen la cantidad de valores en rangos específicos,.
- **Análisis Visual:** El uso de histogramas permite examinar la dispersión de las medias visuales del cáncer para detectar grupos naturales,.
- **Preprocesamiento Sugerido:** Debido a que las características numéricas tienen rangos dispares (ej. área vs suavidad), es necesario aplicar *escalamiento funcional* (Standardization) para garantizar la convergencia de los modelos,.

Fase 4: Casos de Uso de Algoritmos Predictivos

Basado en la naturaleza binaria y visual del dataset, se proponen los siguientes enfoques de modelado bajo supervisión,:

- **Regresión Logística:** Ideal para clasificación binaria (M vs B). Permite predecir el tipo de cáncer basándose en la relación lineal de las características visuales,.
- **K-Vecinos más cercanos (k-NN):** Clasifica muestras analizando la similitud entre pacientes cercanos. Se asume que características visuales similares tienden a diagnósticos iguales,.
- **Máquinas de Vectores de Soporte (SVM):** Algoritmo potente para la separación clara de clases en espacios de alta dimensionalidad, optimizando el margen de decisión médico,.

Fase 1: Entendimiento del Negocio y los Datos

Según el estándar CRISP-DM, antes de cualquier procesamiento técnico, se deben ejecutar las siguientes acciones estratégicas sobre el dataset de *Cancer Data* [1], [3]:

- **Establecimiento de Objetivos:** Definir el problema como una tarea de *clasificación binaria* para predecir la malignidad de un tumor basándose en atributos celulares [4], [1].
- **Criterios de Éxito:** Determinar métricas de rendimiento críticas. En este contexto médico, es vital minimizar los *falsos negativos* mediante el análisis de la Matriz de Confusión y el F1-Score [5], [6].
- **Reporte de Calidad de Datos:** Realizar una auditoría inicial para detectar:
  - Presencia de valores nulos o ausentes que requieran imputación [7], [8].
  - Consistencia en el formato de las variables numéricas (radio, textura, perímetro) [9], [10].
  - Identificación de *outliers* que puedan sesgar el modelo predictivo [11], [12].

Auditoría de Calidad Extendida (Fase 2)

Además de los nulos y outliers, se deben verificar los siguientes criterios estáticos para asegurar la integridad del análisis,:

- **Valores Duplicados:** Detectar y consolidar registros idénticos que puedan causar confusión y sesgar el rendimiento del modelo,.
- **Verificación de Tipos:** Garantizar que cada atributo (booleano, numérico, categórico) sea consistente con el significado documentado,.
- **Distribución Estadística:** Evaluar el sesgo (*skew*) y la curtosis de las variables numéricas; distribuciones extremas (sesgo > 1, curtosis > 7) sugieren el uso de algoritmos simbólicos en lugar de estadísticos,.
- **Vigencia de los Datos:** Confirmar que las observaciones coinciden con la ventana de tiempo que se desea analizar para evitar el uso de información caduca,.

Análisis de Interdependencia y Redundancia

Antes del modelado, es crítico evaluar la relación entre variables numéricas (como radio y perímetro),:

- **Colinealidad:** La alta correlación entre variables dificulta la atribución del comportamiento a un atributo específico y hace inestables a algunas técnicas de modelado,.
- **Variables No Informativas:** Identificar y eliminar variables con varianza cero o redundantes que consumen recursos computacionales innecesarios sin aportar valor predictivo,.
- **Uso de PCA:** En casos de alta dimensionalidad, considerar la reducción por proyección mediante Análisis de Componentes Principales para capturar la mayor varianza posible en un espacio más pequeño,.

Preprocesamiento Estructural (Tidy Data)

Para que el dataset sea procesable por librerías como *scikit-learn*, debe cumplir con los principios de *Tidy Data*,:

- **Encapsulamiento:** Garantizar que cada fenómeno (ej. medición celular) esté en una única tabla y cada observación en una sola fila,.
- **Normalización Estructural:** Cada variable debe estar documentada en una columna única, evitando que los descriptores de columna sean en realidad valores de una variable,.
- **Consistencia de Significado:** La representación física de los datos debe ser consistente con la semántica del dominio médico (ej. unidades de medida uniformes),.

Análisis Exploratorio de Datos (EDA)

El EDA es fundamental para comprender las dinámicas del dataset [14].

- **Distribuciones:** Uso de histogramas o diagramas de densidad para observar la dispersión de las variables cuantitativas [14], [15].

Análisis Exploratorio de Datos (EDA)

El EDA es fundamental para comprender las dinámicas del dataset [14].

- **Distribuciones:** Uso de histogramas o diagramas de densidad para observar la dispersión de las variables cuantitativas [14], [15].
- **Correlaciones:** Implementación de mapas de calor (Heatmaps) para detectar colinealidad entre atributos numéricos (ej. relación entre radio y perímetro) [16], [17].

Análisis Exploratorio de Datos (EDA)

El EDA es fundamental para comprender las dinámicas del dataset [14].

- **Distribuciones:** Uso de histogramas o diagramas de densidad para observar la dispersión de las variables cuantitativas [14], [15].
- **Correlaciones:** Implementación de mapas de calor (Heatmaps) para detectar colinealidad entre atributos numéricos (ej. relación entre radio y perímetro) [16], [17].
- **Discriminación:** Comparar las características medias entre el subgrupo de tumores benignos frente a malignos [18], [19].

Fase 3: Preparación de Datos (Preprocesamiento)

Es la etapa que más tiempo consume, asegurando que los datos sean estructural y funcionalmente correctos [20], [21].

- **Preprocesamiento Estructural:** Garantizar el formato *Tidy Data* (una variable por columna, una observación por fila) [22], [23].

Fase 3: Preparación de Datos (Preprocesamiento)

Es la etapa que más tiempo consume, asegurando que los datos sean estructural y funcionalmente correctos [20], [21].

- **Preprocesamiento Estructural:** Garantizar el formato *Tidy Data* (una variable por columna, una observación por fila) [22], [23].
- **Preprocesamiento Funcional:**
  - **Escalamiento:** Aplicar *StandardScaler* o *MinMaxScaler* para que las variables tengan rangos comparables [24], [25].
  - **Codificación:** Transformar la variable objetivo (Diagnosis) de categórica a numérica (0/1) para el modelado [24], [26].
  - **Reducción de Dimensionalidad:** Evaluar el uso de PCA para reducir el ruido si existen atributos altamente correlacionados [27], [28].

Fase 4: Modelado

Selección y aplicación de algoritmos especializados [6], [8].

- **Tarea:** Clasificación (Aprendizaje Supervisado) [7], [29].

Fase 4: Modelado

Selección y aplicación de algoritmos especializados [6], [8].

- **Tarea:** Clasificación (Aprendizaje Supervisado) [7], [29].
- **Algoritmos Candidatos:**
  - **k-NN:** Efectivo si los atributos están igualmente escalados [30], [31].
  - **Regresión Logística:** Proporciona una separación lineal robusta para variables binarias [30], [32].
  - **Árboles de Decisión (CART):** Ofrece alta interpretabilidad para el personal médico [30], [33].

Fase 5 y 6: Evaluación y Despliegue

- **Evaluación:** Uso de *Cross-Validation* para garantizar la fiabilidad de las métricas obtenidas [34], [35]. Análisis de la matriz de confusión para minimizar falsos negativos (casos de cáncer no detectados) [9].

Fase 5 y 6: Evaluación y Despliegue

- **Evaluación:** Uso de *Cross-Validation* para garantizar la fiabilidad de las métricas obtenidas [34], [35]. Análisis de la matriz de confusión para minimizar falsos negativos (casos de cáncer no detectados) [9].
- **Despliegue:** Planificar la integración del modelo mediante una API o reporte final para su uso en entornos productivos [36], [37].

Fase 5 y 6: Evaluación y Despliegue

- **Evaluación:** Uso de *Cross-Validation* para garantizar la fiabilidad de las métricas obtenidas [34], [35]. Análisis de la matriz de confusión para minimizar falsos negativos (casos de cáncer no detectados) [9].
- **Despliegue:** Planificar la integración del modelo mediante una API o reporte final para su uso en entornos productivos [36], [37].
- **Deuda Técnica:** Considerar factores de riesgo como la dependencia de datos inestables y la necesidad de monitoreo constante de la precisión [38], [39].

Bibliografía

- J. Han, M. Kamber y J. Pei, *Data mining concepts and techniques*, 3ra ed., 2012.
- R. Wirth y J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining", 2000.
- D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems", 2015.
- R. D. King et al., "STATLOG: Comparison of Classification Algorithms on Large Real-World Problems", 1995.
- W. González, "Proceso de Minería de Datos", *6213 - Facultad de Ciencias UCV*, 2025 [1, 16].
- S. Martins et al., "Propuesta de Artefactos para el Subproceso de Gestión de Proyectos de Explotación de Información", *SEDICI - UNLP*, 2016 [17].
- H. Wickham, "Tidy Data", *Journal of Statistical Software*, 2014.