Proyecto Avance 1: Creación de la Base de Datos Heterogénea Federada

Sector Salud (CIE-10 F10–F19)

Fernando Rodrigo Valenzuela García de León
fer_rodri-val@hotmail.com
Daniel Rojo Mata
danielrojomata@gmail.com

---

1. Introducción

Se presenta el proceso seguido para construir la primera versión de una **base de datos heterogénea federada en español en el sector salud**, cuyo propósito es integrar información proveniente de distintas fuentes y modalidades de datos: tablas relacionales, grafos y texto libre.

La integración se centra en diagnósticos de la **Clasificación Internacional de Enfermedades (CIE-10), códigos F10–F19**, que corresponden a trastornos mentales y del comportamiento debidos al consumo de sustancias psicoactivas.

2. Fuentes de datos utilizadas

2.1. Bases relacionales (CSV)

Se emplearon dos conjuntos de datos tabulares, previamente depurados y en formato CSV:

- **defunciones_uso_sustancias_clean.csv**: contiene registros de defunciones en México relacionadas con el consumo de sustancias.
  Columnas principales: anio_defuncion, entidad_defuncion, edad_quinquenal, sexo, F10–F19, cve_entidad, fecha, entidad_defuncion_etq.
  Los datos fueron obtenidos de la siguente página:
  https://datos.gob.mx/dataset/defunciones_relacionadas_consumo_sustancias_psicoactivas
- **urgencias_uso_sustancias_clean.csv**: contiene registros de atenciones de urgencias por consumo de sustancias.
  Los datos fueron obtenidos de la siguiente fuente:
  Columnas principales: anio, entidad, edad_quinquenal, sexo, F10–F19, fecha.
  https://datos.gob.mx/dataset/ingresos_urgencias_relacionados_consumo_sustancias_psicoactiva

2.2. Base en grafo (CIE-10)

La jerarquía de diagnósticos de la CIE-10 [1] se representó como un grafo dirigido a partir de:

- **cie10_f10_f19_nodes.csv**: nodos con descripciones de diagnósticos.
- **cie10_f10_f19_edges_enriched.csv**: aristas que representan relaciones jerárquicas y clínicas entre códigos.

Este grafo incluye nodos raíz (F10–F19) y sus subtipos (F10.0–F10.9, F11.0–F11.9, etc.), permitiendo consultas estructuradas sobre relaciones diagnósticas.

Un aspecto fundamental en la construcción del grafo fue incorporar relaciones de **comorbilidad** y **policonsumo** documentadas en la literatura científica y en reportes de salud pública recientes. Estas asociaciones motivaron la inclusión de aristas adicionales entre nodos CIE-10, más allá de la jerarquía taxonómica estándar, reflejando patrones epidemiológicos de co-ocurrencia en el uso de sustancias.

A continuación se presentan las combinaciones más relevantes:

- **Alcohol (F10) y tabaco (F17):** Existe una fuerte asociación entre ambos trastornos; entre el 80 % y el 95 % de las personas con alcoholismo también fuman cigarrillos [2]. El consumo conjunto potencia riesgos de cáncer oral y de faringe, entre otros [3].
- **Alcohol (F10) y cocaína (F14):** La prevalencia de consumo simultáneo de alcohol entre usuarios de cocaína es cercana al 74 % [4], y aproximadamente el 60 % de quienes presentan trastorno por cocaína también tienen trastorno por alcohol. Esta combinación produce cocaetileno, metabolito de alta toxicidad cardíaca y hepática [5].
- **Cannabis (F12) y estimulantes (F14/F15):** Estudios reportan que 17–18 % de usuarios mensuales de cannabis consumen también cocaína, y cerca del 24 % reportan uso de MDMA [6]. Además, más del 90 % de los usuarios de cocaína tienen antecedente de consumo de cannabis [6], sugiriendo trayectorias de policonsumo.
- **Opioides (F11) y depresores (F13/F10):** Más del 90 % de las personas con trastorno por opioides han usado al menos otras dos sustancias en el mismo año, y más del 25 % presentan dos o más diagnósticos adicionales [7]. La combinación de opioides con benzodiacepinas potencia la depresión respiratoria, aumentando el riesgo de muerte [8].
- **Opioides (F11) y estimulantes (F14/F15):** El consumo combinado de heroína o fentanilo con cocaína o metanfetaminas ("speedball") ha aumentado en la última década. En EE.UU., las muertes por sobredosis que involucraron opioides y cocaína crecieron un 450 % entre 2007 y 2019 [9], reflejando la llamada "cuarta ola" de la crisis de sobredosis.

2.3. Construcción del grafo

Con `NetworkX` se creó un grafo dirigido de 110 nodos y 107 aristas. Se añadieron atributos de descripción a cada nodo a partir del archivo de nodos, y las aristas se clasificaron en dos tipos:

1. **Aristas jerárquicas (taxonómicas):** Definidas a partir de la estructura propia de la CIE-10. Ejemplo: F10 → F10.0 (Trastorno mental y del comportamiento debido al consumo agudo de alcohol). Estas aristas reflejan la descomposición formal de los códigos principales (F10–F19) en sus subtipos.

2. **Aristas epidemiológicas (comorbilidad/policonsumo):** Incorporadas a partir de literatura médica y reportes de salud pública recientes. Estas aristas representan la co-ocurrencia frecuente de dos trastornos por uso de sustancias en la práctica clínica o en estudios poblacionales. A diferencia de la jerarquía estándar, aquí se modelan conexiones horizontales entre códigos distintos. Por ejemplo:

   - F10 (alcohol) ↔ F17 (tabaco).
   - F10 (alcohol) ↔ F14 (cocaína).
   - F12 (cannabis) ↔ F15 (anfetaminas).
   - F11 (opioides) ↔ F13 (benzodiacepinas).
   - F11 (opioides) ↔ F14/F15 (estimulantes, speedball).

   Estas conexiones no aparecen en la CIE-10 original, pero dan más información al grafo para reflejar patrones de policonsumo y riesgo documentados en los últimos años.

De esta manera, el grafo resultante no solo captura la estructura formal de la CIE-10, sino que también modela la realidad epidemiológica de las adicciones múltiples, permitiendo consultas sobre *subtipos diagnósticos* y sobre *combinaciones clínicas frecuentes*.

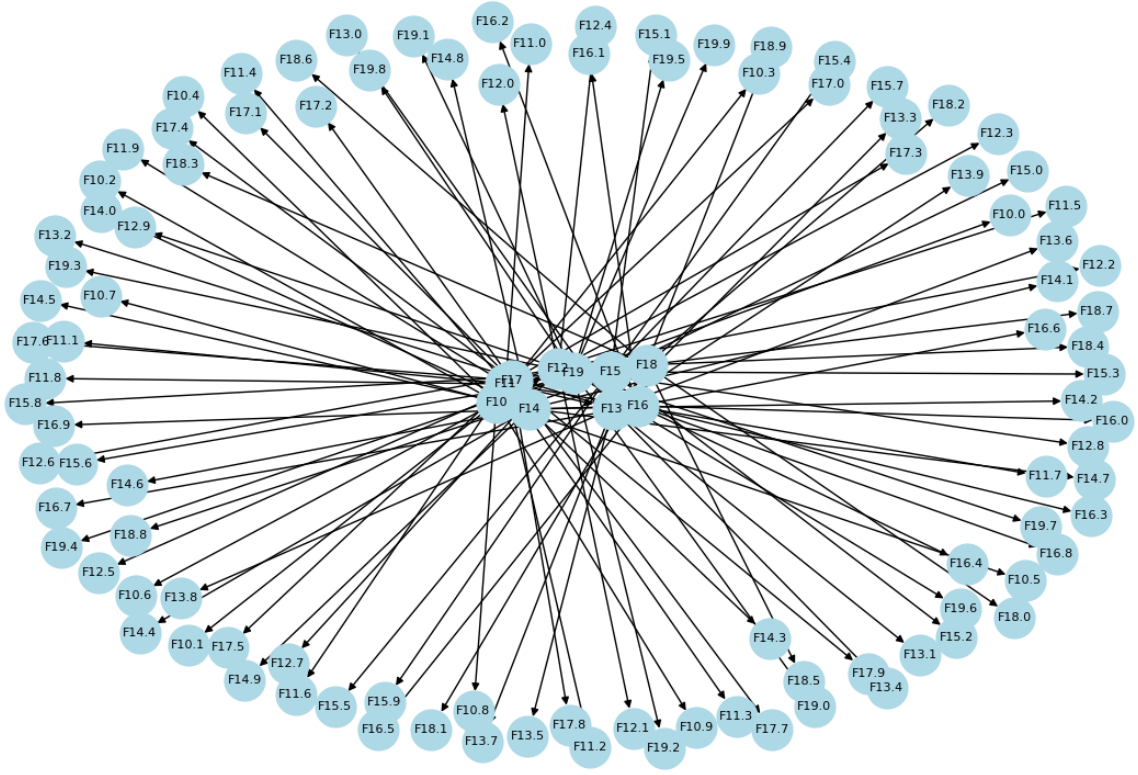

Figura 1: Red creada. Se muestran todos los nodos y las adyacencias, incluyendo jerarquía taxonómica y relaciones epidemiológicas.

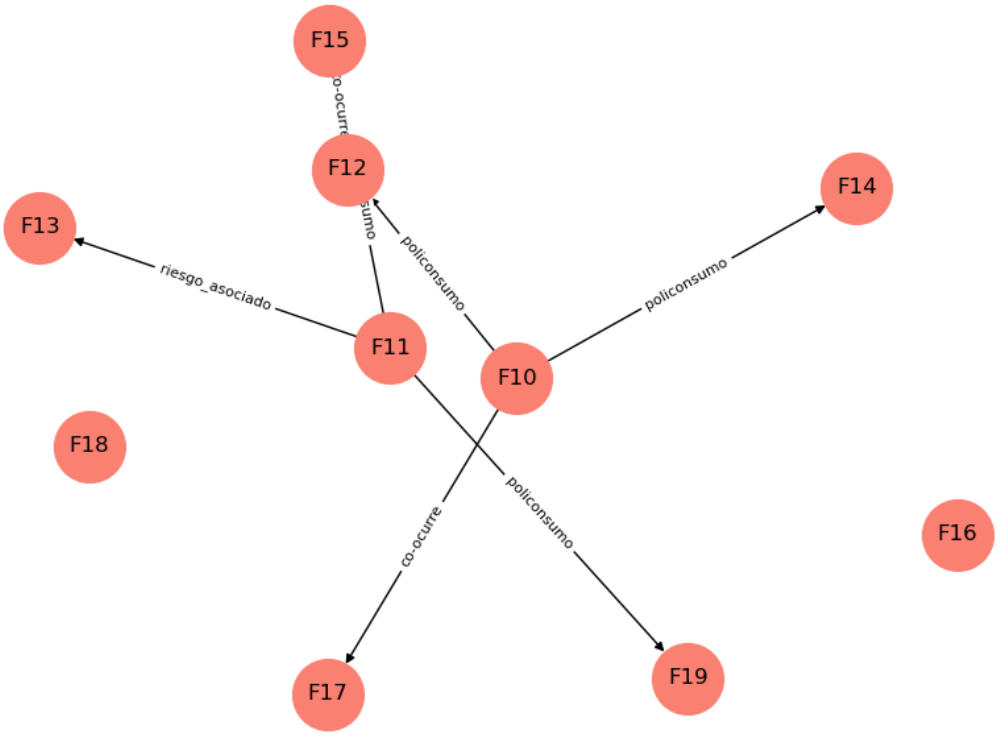

Figura 2: Subred de comorbilidad/policonsumo. Se visualizan las aristas producidas a partir de evidencia epidemiológica.

3. Base de texto (CoWeSe)

La tercera fuente de información considerada fue el corpus **CoWeSe.txt**, un conjunto extenso de textos en español. El objetivo fue extraer de este corpus menciones a sustancias psicoactivas relevantes (alcohol, tabaco, cocaína, cannabis, opioides, etc.) y mapearlas a los códigos CIE-10 de la categoría F10–F19.

Para ello se implementó un proceso de **preprocesamiento de lenguaje natural (NLP)** en varias etapas:

1. **Segmentación en oraciones:** El archivo de texto se dividió en unidades oracionales usando reglas simples basadas en signos de puntuación (`.`, `!`, `?`). El resultado fue un archivo `cowese_sentences.csv` con 50,000 oraciones en la primera corrida de prueba.

2. **Diccionario de palabras clave → CIE-10:** Se diseñó un diccionario inicial de términos en español asociados a sustancias, por ejemplo: *alcohol, etílico, bebidas alcohólicas* → F10; *cocaína, crack* → F14; *cannabis, marihuana* → F12; *benzodiacepinas, clonazepam, diazepam* → F13; *tabaco, nicotina, cigarrillo* → F17; entre otros. A partir de este mapeo se identificaron oraciones que contenían menciones directas a alguna sustancia.

3. **Extracción de coincidencias:** De las 50,000 oraciones procesadas, se detectaron 770 oraciones con al menos una mención asociada a un código CIE-10. Estas coincidencias se guardaron en `cowese_matches.csv`, con las siguientes columnas: `doc_id`, `sent_id`, `sentence`, `keyword`, `cie10`.

4. **Construcción de un índice semántico (TF-IDF):** Se entrenó un modelo **TF-IDF (Term Frequency – Inverse Document Frequency)** sobre el conjunto completo de oraciones. Este

índice generó un vocabulario de 117,389 términos y permite realizar búsquedas por similitud textual. Los modelos se almacenaron en `cowese_vectorizer.pkl` y `cowese_tfidf.pkl`.

5. **Consultas sobre el índice:** Con el índice es posible ejecutar consultas del tipo *"intoxicación por alcohol en jóvenes"*, recuperando las oraciones más relevantes aunque no contengan exactamente las mismas palabras clave. Este mecanismo complementa al diccionario de palabras clave, habilitando búsquedas más flexibles y semánticas.

Resultados preliminares:

- Oraciones procesadas: 50,000.
- Oraciones con mención explícita a sustancias F10–F19: 770.
- Vocabulario del índice TF-IDF: 117,389 términos.
- Códigos más frecuentes en texto: F10 (alcohol, 463 menciones), F17 (tabaco, 145 menciones), F11 (opioides, 73 menciones).

En conjunto, esta base de texto permite conectar el lenguaje natural con los diagnósticos médicos del CIE-10, vinculando oraciones reales con códigos diagnósticos y habilitando su integración con las otras dos fuentes (tablas y grafo).

4. Ejemplo de consulta federada

Consulta con la palabra clave **"cocaína"**:

- Códigos detectados en texto: F10, F11, F12, F14, F15, F17.
- Resultados en SQL (urgencias, 2018): F14 con 89 casos en Aguascalientes, 21 en Baja California, etc.
- Subtipos en grafo: F14.0–F14.9.
- Ejemplos de oraciones extraídas de CoWeSe.

5. Diagrama General

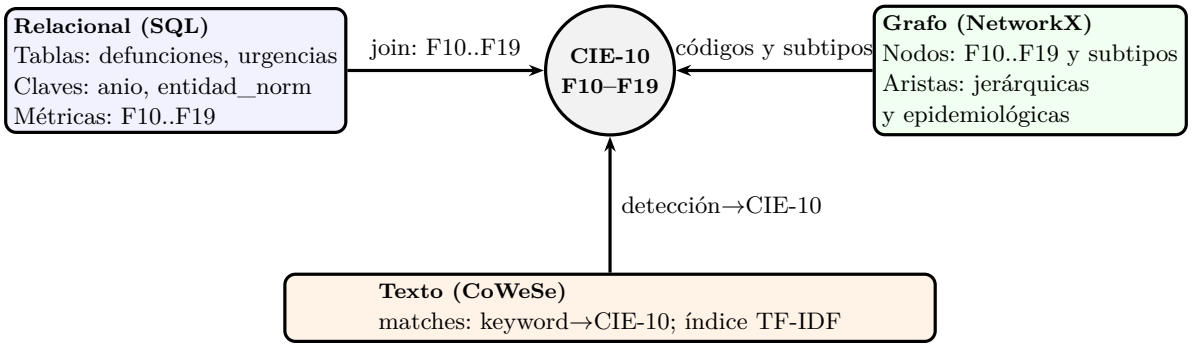

Figura 3: Integración federada: las tres fuentes se conectan vía CIE-10 F10–F19.

Referencias

[1] https://mediately.co/es/icd?chapterCode=F01-F99&setCode=F10-F19&classificationCode=F19#active

[2] ATTC Network. Epidemiology of Alcohol and Tobacco Co-use. Disponible en: https://attcnetwork.org

[3] National Cancer Institute. Alcohol and Tobacco: Risks for Cancers. Disponible en: https://cancercontrol.cancer.gov

[4] American Addiction Centers. Alcohol and Cocaine: A Dangerous Mix. Disponible en: https://americanaddictioncenters.org

[5] Organización Panamericana de la Salud. Alcohol and Cocaine Co-use and Cocaethylene Risks. Disponible en: https://paho.org

[6] PLOS Journals. Patterns of Cannabis and Stimulant Co-use in Epidemiological Studies. Disponible en: https://journals.plos.org

[7] Nature. Epidemiology of Opioid Use Disorders and Comorbidities. Disponible en: https://nature.com

[8] National Institute on Drug Abuse. Opioids and Benzodiazepines. Disponible en: https://nida.nih.gov

[9] PubMed Central. Trends in Opioid-Stimulant Overdose Deaths. Disponible en: https://pmc.ncbi.nlm.nih.gov