Universidad Nacional Autónoma de México

Facultad de Ingeniería

KGraphFI: Grafo de conocimiento de propiedades para el análisis de la mortalidad por COVID-19 en la Ciudad de México

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Ingeniero en Computación

PRESENTA:

García Fernández Jesús Alejandro
Hernández Arrieta Carlos Alberto

DIRECTOR DE TESIS:
Guillermo Gilberto Molero Castillo

Ciudad de México, 2021

1. Datos del alumno

2. Datos del tutor

3. Datos del sinodal 1

4. Datos del sinodal 2

5. Datos del sinodal 3

6. Datos del sinodal 4

7. Datos del trabajo escrito
KGraphFI: Grafo de conocimiento de propiedades para el análisis de la mortalidad por COVID-19 en la Ciudad de México

2021

Resumen/Abstract

COVID-19 es una enfermedad originada en China, que se ha extendido alrededor del mundo hasta convertirse en una pandemia mundial. Esta enfermedad ha afectado en el mundo a millones de personas con contagio y muerte, siendo México uno de los países con mayores tasas de infección y muerte, esto es, cerca de 250 mil fallecimientos a julio de 2021. **Problema de investigación.** Dada la pandemia actual, la recolección y análisis de datos de COVID-19 ha llevado a un creciente interés en definir nuevos enfoques y métodos para extraer información útil a partir de los datos sobre esta enfermedad, que ha causado, en muchas partes del mundo, altas tasas de mortalidad. La Ciudad de México no es la excepción, donde existe una elevada mortalidad por esta enfermedad. Ante esto, los grafos de conocimiento han tomado importancia, puesto que permiten transformar los datos en forma de relaciones y patrones, que a su vez pueden ser útiles para tener un mayor entendimiento de la enfermedad. **Objetivo.** La presente tesis se enfoca en la implementación de un grafo de conocimiento de propiedades para el análisis de la mortalidad de mujeres embarazadas contagiadas de COVID-19 y que recibieron atención médica en hospitales de la Ciudad de México. Para cumplir con este propósito se utilizó la tecnología Neo4j y se analizó la relación existente entre los factores de riesgo y la mortalidad por esta enfermedad. **Motivación.** Hoy en día, hacer un análisis adecuado de datos es uno de los desafíos de la sociedad actual. Se decidió analizar la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México debido a que es una población vulnerable con alto riesgo de infección y rápido deterioro clínico, con síntomas variados de fiebre, tos, cefalea, diarrea y otras; siendo necesarios los tratamientos médicos inmediatos. **Método.** El método utilizado fue estructurado en cuatro etapas: a) adquisición de los datos, b) análisis y preparación de los datos, c) modelado del grafo de conocimiento, y d) creación del grafo en Neo4j. Las cuales fueron de tipo exploratoria, dado el hecho de que los grafos de conocimiento son una disciplina emergente, y aplicada debido a que estos se apoyan de conocimientos específicos. **Resultados.** A partir del grafo de conocimiento construido, se logró observar que existe un importante nivel de mortalidad en la población objeto de estudio, siendo tos, cefalea, fiebre y mialgias los principales síntomas presentados; así como la obesidad, como la principal enfermedad crónica padecida. Otro factor importante fue la edad en la que fallecieron las pacientes, esto es, entre 24 y 39 años. **Conclusión.** Si bien el riesgo general de enfermarse gravemente a causa de COVID-19 es alto, sigue siendo mayor para las embarazadas. Se ha identificado que tener ciertas afecciones, ocultas o no, y otros factores, incluida la edad, puede aumentar aún más el riesgo de contagio y muerte en las gestantes, a quienes, por su condición, a pesar del cuidado que puedan tener, no están exentas de mantener la distancia con personas que podrían estar enfermas, como médicos, enfermeras, familiares, vecinos y entorno en general.

Índice general

Capítulo 1

Introducción

1.1. Contexto de la investigación

En diciembre de 2019, surgió una serie de casos de neumonía de causa desconocida en Wuhan, Hubei, China, con presentaciones clínicas parecidas a la neumonía viral. Del análisis de las muestras del tracto respiratorio inferior indicaron un nuevo coronavirus (Huang y col., 2020), que se denominó coronavirus de 2019 (COVID–19). Esta enfermedad por COVID–19 es un padecimiento respiratorio, de reciente aparición, causado por el síndrome respiratorio agudo severo coronavirus 2 (SARS–CoV–2), el cual fue declarado por la Organización Mundial de la Salud (OMS) como pandemia mundial el 30 de enero de 2020 (X. Cao, 2020; World-Health-Organization, 2020c).

En la actualidad, la mayoría de los pacientes con COVID–19 presentan síntomas de leves a moderados, pero aproximadamente un 15 % pasan a tener una neumonía grave, y aproximadamente un 5 % desarrollan el síndrome de dificultad respiratoria aguda (SDRA), choque séptico y otros una insuficiencia orgánica múltiple (Huang y col., 2020). Aunque se están probando activamente varios medicamentos antivirales, ninguno ha sido aprobado específicamente para COVID–19. Además, el desarrollo de vacunas y nuevos tratamientos se ha convertido en un enfoque importante para bloquear el virus (X. Cao, 2020).

En el presente, se vive un brote de enfermedad por COVID-19, que ha provocado en el mundo entero más de 3.5 millones de muertes (World-Health-Organization, 2020f). Esta situación no es ajena al continente americano, donde se han registrado alrededor de 70 millones de casos confirmados y más de 1.7 millones de muertes, convirtiéndose así en el centro de la pandemia, siendo Estados Unidos, Brasil y México los países con más defunciones por esta enfermedad.

En México, el anuncio oficial del primer paciente infectado fue el 29 de febrero de 2020. A la fecha, 1 de junio de 2021, se informaron más de 232 mil fallecidos, alrededor de 2.5 millones de casos confirmados acumulados y más de 433 mil casos sospechosos. Además, México sigue entre los diez países con más contagios de COVID–19 (Gobierno-de-México, 2020a). De acuerdo con Méndez-Arriaga (2020), las características climáticas jugaron un papel importante en la infección local, siendo las regiones templadas, como Michoacán, Jalisco, Puebla, y otros, más vulnerables en comparación con las regiones secas, como Chihuahua, Durango o Zacatecas; o zonas tropicales, como Colima, Campeche, Morelos y otros.

En el caso específico de la Ciudad de México (CdMx), objeto de estudio de este trabajo de tesis, se han reportado también altos números de contagio en sus habitantes, lo que hace que se tenga también un alto porcentaje de mortalidad. A la fecha, 1 de junio de 2021, se reportaron más de 658 mil casos confirmados y 33 mil fallecidos (Gobierno-de-México, 2020b), existiendo como causalidad asociada diversos factores de riesgo, como enfermedades crónicas, hipertensión, diabetes, cáncer, entre otras, las cuales aumentan la posibilidad de presentar casos clínicos graves a causa de COVID-19.

Este alto número de casos infectados y fallecidos se debe, entre otras causas, al desconocimiento de los diferentes factores de riesgo por parte de la población y a las medidas de emergencia sanitaria que se implementaron a partir del 30 de marzo de 2020, que, junto con acciones de sana distancia, buscaron reducir la dispersión y transmisión comunitaria de COVID–19 en todo el territorio nacional (Gobierno-de-México, 2020a).

Por otro lado, debido a esta situación de pandemia, día a día se generan grandes cantidades de datos, lo que vuelve complejo el análisis e interpretación de éstos. Por lo que, es necesario estructurar los datos generados a partir de la transmisión del COVID-19 con el propósito de facilitar la generación de conocimiento a partir de grandes cantidades de información y así lograr una rápida interpretación de los datos. En consecuencia, en este trabajo de tesis se propone la implementación de un grafo de conocimiento de propiedades para estructurar la información y analizar la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México.

Se decidió analizar esta población vulnerable debido a que infección por COVID-19 en gestantes puede ser clasificada según la gravedad de la sintomatología respiratoria, esto es, leve, moderada y severa. La mayoría de casos durante la gestación presentan algún tipo de severidad de la infección, con síntomas frecuentes de fiebre y tos; y otras menos frecuentes como mialgias, disnea, cefalea y diarrea. Además, hay indicios de que en la gestación también hay un alto riesgo de infección por COVID-19, especialmente cuando se asocian factores de riesgo, como la edad, obesidad, hipertensión y diabetes. Otras comorbilidades son las enfermedades pulmonares y renales. Por lo que, se debe tener en cuenta que las embarazadas con infección por COVID-19 pueden presentar un rápido deterioro clínico, siendo necesarios los tratamientos médicos inmediatos.

1.2. Problema de investigación

En la sociedad actual se recopilan datos sobre un determinado contexto, momento y lugar (Aalst, 2016). Estos datos, información y conocimiento son considerados en la actualidad como un activo estratégico en las organizaciones (L. Cao, 2017). Prácticamente en todas las áreas de las organizaciones se recolectan datos sobre sus operaciones, flujos de trabajo y otros procesos de interés. Por ejemplo, en los centros de atención médica la recopilación y análisis de datos pueden ser útiles para identificar y mejorar los procesos de tratamientos médicos, reducir los tiempos de espera, reducir los costos de operación, mejorar las capacidades para satisfacer la demanda, mejorar la productividad de los recursos y aumentar la transparencia de los procesos. Por lo tanto, esta disponibilidad de datos ha llevado a un creciente interés en definir nuevos enfoques y métodos para extraer información útil y adquirir conocimiento a través del análisis de datos.

Así, a la luz de los desafíos que presenta COVID–19 en nuestra sociedad, se distinguen

diversas acciones y esfuerzos para hacer frente a esta enfermedad que ha causado, en muchas partes del mundo, altas tasas de mortalidad (Huang y col., 2020). La Ciudad de México no es la excepción, donde existe una elevada mortalidad por esta patología. Las medidas de emergencia sanitaria que se implementaron, a partir del 30 de marzo de 2020, junto con acciones de sana distancia, buscaron reducir la dispersión y transmisión comunitaria de COVID–19 en todo el territorio nacional (Gobierno-de-México, 2020a). Se buscó además disminuir la carga de la enfermedad, sus complicaciones y muerte en la población. Sin embargo, la propagación de la enfermedad, sujeta a la suspensión de actividades no esenciales para la contención de COVID–19, no fue suficiente y no logró ser efectiva.

En este sentido, a través de este trabajo de tesis se busca analizar los factores de riesgo y la mortalidad de embarazadas diagnosticadas con COVID-19 en la Ciudad de México. Esto con el propósito de detectar patrones de mortalidad en la población objeto de estudio. Para esto, se buscó explorar la asociación entre los casos locales diarios positivos confirmados y el número de muertes diarias reportadas mediante grafos de conocimiento (KG, por sus siglas en inglés), área de la Inteligencia Artificial, que en los últimos años ha tenido un importante desarrollo en aplicaciones sobre la integración de información heterogénea. Este tipo de grafos permiten crear y resaltar las relaciones semánticas entre diferentes entidades, donde las relaciones tienen un identificador, nombre y dirección. Por lo tanto, el grafo de conocimiento que se pretende desarrollar está relacionado con el motor de bases de datos con orientación a grafos, Neo4j. El cual es de código abierto y proporciona un lenguaje de consultas nativo, Cypher.

1.3. Pregunta de investigación

Se plantea la siguiente pregunta de investigación que surge de la problemática anterior y que se pretende responder:

- ¿Qué relación existe entre los factores de riesgo y la mortalidad de gestantes diagnosticadas con COVID-19 que fueron atendidas en la Ciudad de México?

1.4. Hipótesis

A partir de la problemática planteada y la pregunta de investigación, se establece la siguiente hipótesis:

- Existe relación entre los padecimientos por enfermedad, edad y colonia con la mortalidad de gestantes infectadas con COVID-19 que fueron atendidas en hospitales de la Ciudad de México.

Para probar la hipótesis se utilizó como caso de estudio la fuente de datos de la Base del Sistema Nacional de Vigilancia Epidemiológica (SINAVE) para el seguimiento de posibles casos de COVID-19 en la Ciudad de México. Estos datos fueron obtenidos a través del portal Web

de datos abiertos sobre salud pública, acciones sociales y gasto público en la Ciudad de México (Secretaría de Salud Gobierno de la Ciudad de México, 2021).

1.5. Objetivos general y específicos

1.5.1. Objetivo general

- Implementar un grafo de conocimiento de propiedades para el análisis de la mortalidad de gestantes diagnosticadas con COVID–19 en la Ciudad de México.

1.5.2. Objetivos específicos

- Diseñar y desarrollar el grafo de conocimiento de propiedades con base en la tecnología Neo4j.
- Determinar la relación que existe entre los factores de riesgo y la mortalidad de embarazadas diagnosticadas con COVID-19 en la Ciudad de México, según su edad, padecimientos de salud y alcaldía.

1.6. Justificación

Hoy en día, hacer un análisis adecuado de datos es uno de los desafíos de la sociedad actual. Esto debido a la creciente recolección de datos en diversos campos de aplicación, como el cuidado de la salud, seguridad pública, análisis de mercado, prácticas comerciales, procesos industriales, descubrimientos científicos, políticas públicas, entre otros temas de interés. Precisamente, una de las formas de extraer información a partir de los datos, de fuentes variadas, es a través de grafos de conocimiento, que técnicamente se trata de teoría de grafos (vértices, aristas y atributos), que tiene como objetivo organizar la información, de tal manera que ésta pueda extraerse de manera fácil a partir de entidades con atributos y sus relaciones con otras entidades, también con atributos.

En el contexto de la salud, el análisis de datos sobre COVID–19 es importante por varias razones: a) se necesitan analizar datos actualizados sobre la enfermedad para estudiar sus posibles causas y efectos; b) caracterizar los síntomas durante y después de la enfermedad, c) analizar los factores de riesgo, como la obesidad, diabetes, problemas renales y otros; d) investigar posibles grupos de contagio y sus causas; e) identificar mejores prácticas de tratamientos y vacunas, entre otros.

En México, uno de los efectos que debe estudiarse para distinguir el impacto de COVID–19, es la mortalidad, puesto que el número de contagios es amplio, a pesar de no haberse aplicado pruebas masivas a la población. Algunos estudios señalan la existencia de una importante selectividad de la mortalidad por edad, sexo, entidades federativas, por algunas condiciones

demográficas y socioeconómicas (Hernández, 2020). Además, no se debe eludir la controversia existente en torno a que el número de muertes por COVID–19 en México debería ser superior, dado que la magnitud de la pandemia está subestimada y una de las razones principales es el bajo número de pruebas de diagnóstico que se realizan (Forbes, 2020). Por lo que, es un tema de interés que debe analizarse a partir de la información disponible, puesto que es un fenómeno que sigue creciendo y que no ha concluido.

1.7. Organización del documento de tesis

El documento de tesis está dividido en cinco capítulos. En este primer capítulo se describe el contexto de la investigación, se incluye una descripción del problema, se plantea la pregunta de investigación y la hipótesis; y se especifican los objetivos y la justificación de la investigación. Dando lugar al desarrollo de este trabajo de investigación, cuyos resultados se presentan en los capítulos siguientes.

En el capítulo 2 se presenta los fundamentos de los grafos de conocimiento y los principios de su funcionamiento. Asimismo, se describen los modelos de grafos de conocimiento, las bases de datos orientadas a grafos y la minería de grafos. Además, se da a conocer las características de COVID-19, sus síntomas, transmisión y etapas. Enfermedad catalogada como pandemia a nivel mundial. Finalmente, se da a conocer los trabajos relacionados con esta investigación.

El capítulo 3 presenta la propuesta de solución de la construcción del grafo de conocimiento de propiedades para el análisis de la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México. Para esto, se definieron cuatro etapas de trabajo: adquisición de datos; análisis y preparación de datos; modelado del grafo de conocimiento; y creación del grafo en Neo4j. El resultado fue la creación de un grafo de conocimiento conformado por 1781 nodos y 9339 relaciones, que representan toda la complejidad de los datos elegidos para la presente investigación.

En capítulo 4 muestra los resultados obtenidos con base en la evaluación de la funcionalidad del grafo de conocimiento implementado. Con base en esto se hizo el análisis de la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México, cuyo periodo de evaluación fue desde el inicio de la pandemia en marzo de 2020 hasta el 24 de diciembre del año en mención. Se observaron patrones de datos significativos, resaltando que algunas pacientes fallecidas estaban en el rango de edad de 20 a 29 años, con 8 meses de gestación y que se dedicaban a las labores domésticas.

El Capítulo 5 presenta las conclusiones generales y particulares del trabajo de investigación realizado, y se establecen los trabajos futuros que se pretenden desarrollar con base en los resultados obtenidos.

Finalmente, se presentan una serie de anexos, donde se muestran información ampliada sobre las variables disponibles en la fuente de datos (Anexo A); y código fuente de las consultas en Cypher (Anexo B).

Capítulo 2

Marco Teórico y Estado del Arte

Gracias al avance de la tecnología es posible almacenar amplios volúmenes de datos de manera continua, ya sea de forma estructurada en bases de datos o en modo textual. Una de las ventajas de contar con estas fuentes de datos es la posibilidad de descubrir información de interés y adquirir conocimiento útil mediante el análisis de los datos. Sin embargo, el volumen de datos es una limitante para un análisis manual. Por lo que, se han desarrollado tecnologías especializadas que faciliten su manejo, como los grafos de conocimiento, que cuenta con fundamentos útiles para la estructuración de información a partir de los datos almacenados. Esto ha hecho que los grafos de conocimiento se conviertan en una herramienta útil, estableciendo en la actualidad un área activa de investigación y desarrollo.

En este capítulo se define los grafos de conocimiento y sus principales fundamentos. Un aspecto importante es el papel de los grafos de conocimiento como parte de la estructuración de información y el descubrimiento de conocimiento a partir de los datos. Asimismo, se revisan los modelos y técnicas comúnmente utilizados en los grafos de conocimiento. Se presenta además aspectos de interés relacionados con COVID-19, enfermedad objeto de estudio en este trabajo de investigación, que a la fecha ha ocasionado altas tasas de mortalidad a nivel mundial. Finalmente, se presentan los trabajos relacionados sobre grafos de conocimiento y analítica de datos asociados con CODIV-19.

2.1. Grafos

De acuerdo con Rosen (2011) un grafo $G = (V, E)$ es una estructura compuesta que se define como un conjunto no vacío de vértices V, conocidos también como nodos, y un conjunto de aristas E, donde las aristas representan la relación existente entre pares de vértices. Con base en la literatura, se identificaron dos tipos de grafos, dependiendo del tipo de aristas E:

- **Dirigidos**. Un grafo $G = (V, E)$ dirigido es aquel donde el conjunto de aristas E tienen una dirección, tal como se muestra, a modo de ejemplo, en la Figura 2.1.
- **No Dirigidos**. Un grafo $G = (V, E)$ no dirigido es aquel donde el conjunto de aristas E carecen de una dirección, esto es, son bidireccionales. La Figura 2.2 muestra un ejemplo de este tipo de grafos.

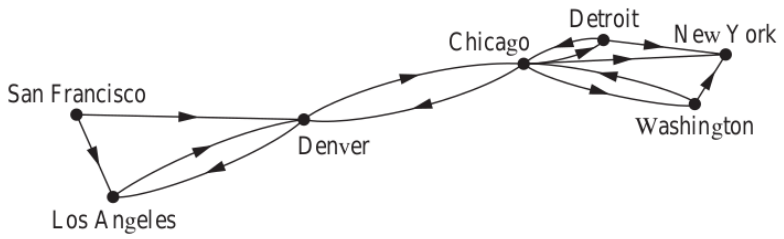

Fig. 2.1: Ejemplo de un grafo dirigido. Adaptado de Rosen (2011).

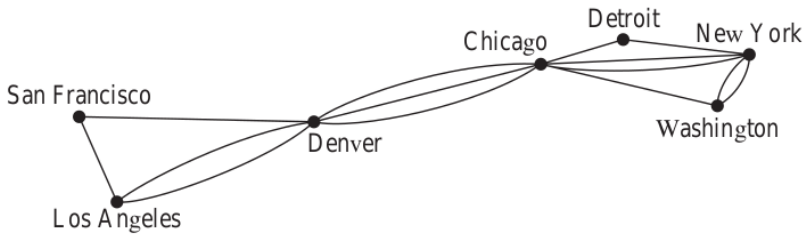

Fig. 2.2: Ejemplo de un grafo no dirigido. Adaptado de Rosen (2011).

2.2. Grafos de conocimiento

Un grafo de conocimiento (Knowledge Graph o KG, por sus siglas en inglés) es una estructura de datos basada en grafos que sirve para representar información con el propósito de organizarla y guardarla para su interpretación (Chen y col., 2020), donde los nodos (vértices) son las entidades y las aristas son las relaciones que existen entre esas entidades (Y. Cao y col., 2019).

Así, una entidad puede representar a un objeto tangible, como persona, lugar y organización; o concepto (intangible), como color, sentimiento, definición y otros. Estas entidades están conectadas por aristas que describen las relaciones entre éstas. Por lo que, analizar los datos del mundo real, de esta manera, ayuda a comprender el significado detrás de una consulta, lo que representa resultados más relevantes para el usuario. Un ejemplo de representación de información con un grafo de conocimiento es el cuadro que se ve en los resultados de la búsqueda de Steve Jobs en Google, tal como se muestra en la Figura 2.3.

Fig. 2.3: Información extraída por Google a partir de un grafo de conocimiento.

Por lo general, un grafo de conocimiento se construye sobre las bases de datos existentes, con el propósito de vincular los datos combinando información estructurada y no estructurada. Conectar conjuntos de datos, de manera relacional, ayuda a los usuarios a obtener conocimiento existente sobre una determinada situación o contexto. Por lo que, algunas empresas de tecnología como Amazon, Facebook y Google invirtieron millones de dólares para crear sus propios grafos de conocimiento. Por ejemplo, Google Knowledge Graph utiliza las relaciones entre palabras y conceptos para comprender el contexto de una consulta, y así asignar un significado específico a las intenciones del usuario.

2.3. Modelos de grafos de conocimiento

Para la construcción de un grafo de conocimiento se hace una abstracción de los datos con el propósito de modelar el grafo. Para esto existen modelos que permiten representar la información

de forma estructurada y concisa, entre los que destacan: *i*) Grafo dirigido con aristas etiquetadas (Directed edge labelled graphs) y *ii*) Grafo de propiedades (Property graphs).

2.3.1. Grafo dirigido con aristas etiquetadas

Este modelo tiene un conjunto de entidades y aristas dirigidas, cada una con una etiqueta, que representan las relaciones entre éstas. Por ejemplo, dado un conjunto definido "*Con*", este tipo de grafos dirigidos con aristas etiquetadas se representa a través de una tupla $G = (V, E, L)$, donde $V \subseteq Con$ es el conjunto de entidades, $L \subseteq Con$ es el conjunto de etiquetas, y $E \subseteq V \times L \times V$ es el conjunto de aristas (Hogan y col., 2020). La Figura 2.4 es una representación de un grafo dirigido con aristas etiquetadas.

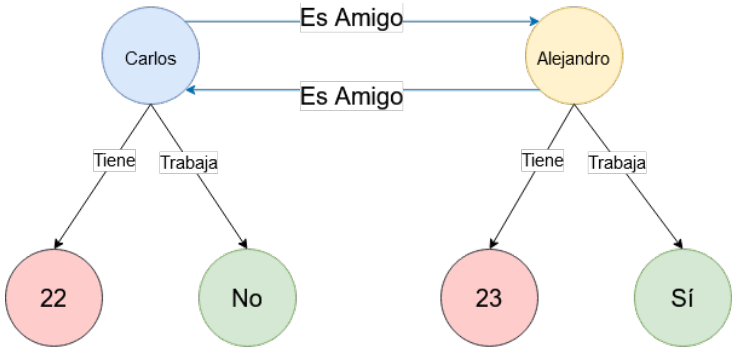

Fig. 2.4: Ejemplo de grafo dirigido con aristas etiquetadas.

Donde, $V = \{$Carlos, Alejandro, 22, 23, No, Sí$\}$ y $L = \{$Es amigo, Tiene, Trabaja$\}$. Por lo tanto, al hacer el producto Cartesiano $V \times L \times V$ se obtiene el conjunto de aristas $E = \{$(Carlos, Es amigo, Alejandro), (Alejandro, Es amigo, Carlos), (Carlos, Tiene, 22), (Carlos, Trabaja, No), (Alejandro, Tiene, 23), (Alejandro, Trabaja, Sí)$\}$. Así, con base en esta representación, a través del grafo, se puede dar respuesta a algunos cuestionamientos, como:

- ¿Qué edad tiene Carlos?
- ¿Qué edad tiene Alejandro?
- ¿Alejandro trabaja?

Para dar respuesta a estas interrogantes, es necesario ubicarse en cada nodo correspondiente y seguir la relación de las aristas etiquetadas. A manera de ejemplo, en el caso de la última pregunta el recorrido es: *Alejandro* → Trabaja → Si

2.3.2. Grafo de propiedades

Un grafo de este tipo está formado por nodos, relaciones (aristas) y propiedades. Las relaciones deben tener una etiqueta o nombre y una dirección, es decir, es un grafo dirigido. Así, todas las

relaciones tienen un nodo origen y un nodo destino (Pérez Solá y col., 2018). En este modelo, los nodos y las relaciones que los conectan pueden tener propiedades, que son pares de *clave-valor*. La *clave* es una cadena de caracteres que indica la semántica de la propiedad, esto es, representa una característica identificativa única. Mientras que el *valor* puede ser un conjunto de tipos de datos, como caracteres o números. Por lo tanto, el grafo de propiedades está formado por (Neo4j, 2019):

- Nodos. Son las entidades del grafo, que pueden o no contener algunos atributos, los cuales representan las propiedades, conocidas como pares de clave-valor. Estos nodos pueden identificarse por medio de etiquetas, que representan los roles en el dominio total del grafo.
- Relaciones. Proporcionan conexiones dirigidas entre dos nodos, las cuales son semánticamente relevantes. Una relación siempre tendrá una dirección, un tipo o etiqueta, un nodo inicial y un nodo final. Al igual que los nodos, las relaciones también pueden contener propiedades, como pesos, costos, distancias, calificaciones, intervalos de tiempo, entre otras. Un ejemplo de un grafo de propiedades se muestra en la Figura 2.5.

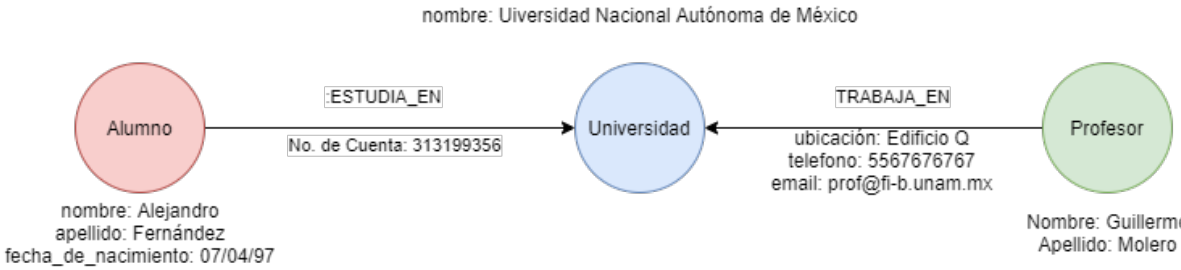

Fig. 2.5: Ejemplo de un grafo de propiedades.

El ejemplo muestra una forma de representación de información estructurada en tres entidades, con sus respectivos atributos y relaciones. A través de este grafo se puede dar respuesta a algunas interrogantes, como: *i*) ¿Cuál es el nombre del profesor?, *ii*) ¿Cuál es la fecha de nacimiento del alumno?, y *iii*) ¿Dónde trabaja el profesor y cuál es su ubicación?. Estas preguntas podrían ser respondidas accediendo a las propiedades de los nodos y siguiendo sus relaciones.

2.4. Bases de datos orientadas a grafos

Una base de datos orientada a grafos es un sistema de gestión de bases de datos no relacional que utiliza una estructura de grafos con nodos, aristas y propiedades para almacenar información (S.-T. Wang y col., 2015). Cuenta con cuatro propiedades básicas para el almacenamiento persistente de datos, esto es, CRUD por sus siglas en inglés (Pérez Solá y col., 2018): crear (create), leer (read), actualizar (update) y eliminar (delete). Además, este tipo de bases de datos está diseñada para tratar las relaciones y los datos con el mismo nivel de importancia (Robinson y col., 2015). Esto hace que se evite tener información aislada, dado que todos los datos están conectados de manera eficiente.

Por otro lado, mientras que las bases de datos convencionales calculan las relaciones al momento de las consultas (a través de JOIN), una base de datos orientada a grafos almacena las conexiones en el mismo modelo de datos, sin importar el tamaño de éstos. Además, este tipo de bases de datos destacan por la gestión de datos altamente conectados, con el propósito de explorar los datos vecinos, recolectando y añadiendo información de otros nodos y relaciones sin comprometer a los datos restantes (Robinson y col., 2015).

En cuanto a la estructura de una base de datos orientada a grafos existen dos conceptos: a) Labeled-Property Graph (grafo de propiedades etiquetadas) y b) Resource Description Framework (marco de descripción de recursos, RDF). En el primero, se asignan propiedades a los nodos y a las aristas. En el segundo, la estructura del grafo se regula mediante los elementos: sujeto-predicado-objeto. En general, en este tipo de bases de datos, se tienen algoritmos especiales para simplificar y acelerar las consultas a través de búsquedas en profundidad (nodo más profundo) y en anchura (de un nivel a otro). Además, estos algoritmos permiten encontrar patrones, conocidos como graph patterns, y nodos relacionados directa o indirectamente.

2.4.1. Características

Las bases de datos orientadas a grafos forman parte la familia NoSQL (Not only SQL), que están diseñadas para modelos de datos flexibles y crear aplicaciones modernas, las cuales han ganado importancia debido a sus características, funcionalidad, rendimiento, flexibilidad y representación de interconexión de datos del mundo real de manera sencilla y práctica para la consulta y análisis de éstos. Entre las características representativas de este tipo de bases de datos destacan (Robinson y col., 2015):

- *Rendimiento.* Una base de datos orientada a grafos tiene un rendimiento superior cuando se maneja información altamente conectada, comparada con una base de datos relacional, donde conforme la tabla crece de manera vertical, las operaciones JOIN se vuelven más costosas, a diferencia de las bases orientadas a grafos, que aunque el grafo se vuelve más grande, el costo de las operaciones no se incrementa.
- *Flexibilidad.* Las bases de datos orientadas a grafos ofrecen esquemas flexibles que permiten que el modelo vaya creciendo conforme nuevos datos surgen. Esto debido a que un grafo es una estructura de datos aditiva, por lo que, se pueden añadir nuevas entidades, relaciones o subgrafos sin alterar lo ya existente.
- *Escalabilidad.* En todas las aplicaciones, los requerimientos van cambiando y evolucionando. Gracias a la flexibilidad que ofrecen las bases de datos orientadas a grafos, se tiene también la posibilidad de ir evolucionando (escalar) el modelo de datos de manera sencilla, y así implementar nuevos requerimientos.

Por otra parte, dado que el propósito de una base de datos orientada a grafos es facilitar la creación y ejecución de aplicaciones que funcionan con conjuntos de datos altamente conectados, como redes sociales, motores de recomendación, detección de fraude y grafos de conocimiento, existen algunas herramientas importantes como Neo4j, Giraph, AllegroGraph, GraphDB y otras.

En este documento y dado el interés de explorar el potencial que ofrece Neo4j, para el manejo de propiedades, velocidad de procesamiento y los diversos métodos analíticos, se hace una descripción ampliada de esta tecnología.

2.4.2. NEO4J

Neo4j es una base de datos orientada a grafos con características de atomicidad, consistencia, aislamiento y durabilidad (ACID, por sus siglas en inglés). Este tipo de bases de datos es de código abierto con un potente motor para procesar grafos en forma de estructura (Lal, 2015). El almacenamiento se hace en forma de grafos y no en tablas, esto es, la información se almacena a través de relaciones, formando un grafo dirigido entre los nodos y sus relaciones.

Neo4j fue desarrollado en Java por Neo Technology Inc, el cual es un emprendimiento que inició en Suecia y Estados Unidos. La versión 1.0 de Neo4j fue lanzada en febrero de 2010. Esta base de datos está disponible para sistemas operativos Windows, Linux y OS X. En la actualidad, son diversas empresas las que utilizan esta tecnología, como Cisco, Walmart, Ebay, HP, entre otras. Además, cuenta con una amplia comunidad de desarrolladores, lo que permite que se agreguen y mejoren funcionalidades, creando un amplio ecosistema de herramientas alrededor de la base central. La Figura 2.6 muestra los componentes de la arquitectura de Neo4j.

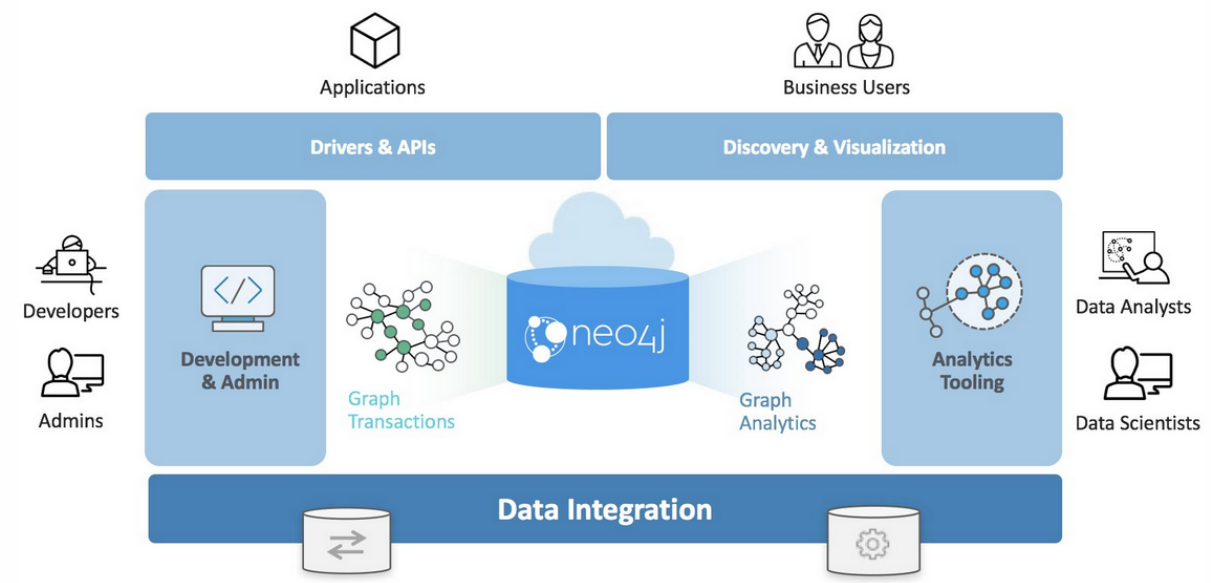

Fig. 2.6: Arquitectura de Neo4j. Fuente: *Graph Algorithms in Neo4j* (2018).

1. *Base de datos basada en grafos.* Es el corazón de la arquitectura, el cual soporta las operaciones transaccionales y la analítica de grafos.

2. *Desarrollo y administración.* Es donde los administradores y desarrolladores hacen uso de la base de datos para la integración de datos.

3. *Drivers y APIs.* Es lo que permite trabajar con lenguajes de programación para conectar y utilizar la base de datos de forma externa a la interfaz gráfica de Neo4j.

4. *Descubrimiento y visualización.* Son las aplicaciones que permiten explorar los grafos, así como también visualizarlos.

5. *Integración de datos.* Son las herramientas que permiten estructurar la información tabular y datos masivos en grafos.

6. *Herramientas de analítica.* Son las herramientas que permiten la extracción de patrones en los datos.

Las bases de datos basadas en grafos como Neo4j están llevando el almacenamiento de datos conectados a un nivel completamente diferente. Estas crecen constantemente en torno a técnicas de almacenamiento del mundo real y se cree que es uno de los logros complejos de la ingeniería. Por lo que, es importante conocer como Neo4j funciona internamente. La Figura 2.7 muestra los componentes internos de Neo4j.

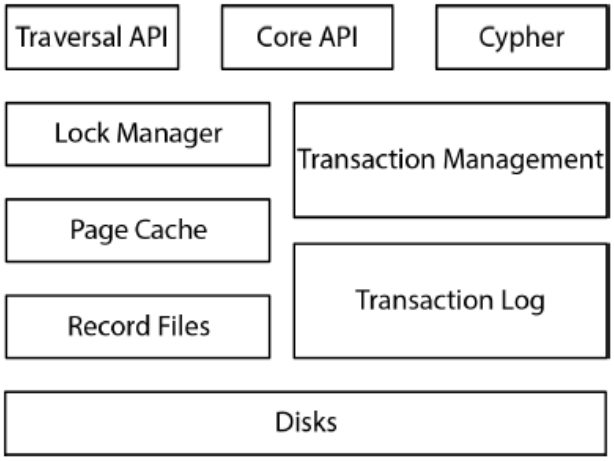

Fig. 2.7: Componentes internos de Neo4j. Fuente: Webber (2020).

1. *Cypher.* Es el lenguaje de consultas que permite guardar y extraer información de Neo4j. Su sintaxis en *ascii-art* consiste en poner nodos entre paréntesis y relaciones como flechas etiquetadas por corchetes.

2. *Core API.* Es una interfaz de programación de aplicaciones en Java (API) que expone las primitivas del grafo, como los nodos, relaciones, propiedades y etiquetas al usuario.

3. *Traversal API.* Es una interfaz de programación de aplicaciones en Java (API) que permite recorrer la estructura del grafo definido en la base de datos.

4. *Lock Manager.* Es el encargado de administrar la concurrencia de la base datos, decide qué tipo de bloqueo se aplica a diferentes transacciones y a qué recursos de la base de datos.

5. *Transaction Management.* Se encarga de procesar varias consultas de varios usuarios para que se pueda cumplir con las propiedades ACID.

6. *Transaction Log.* Es donde se guarda el historial de las transacciones realizadas.
7. *Page Cache.* Es donde se almacena temporalmente los resultados de las consultas realizadas con el objetivo de que puedan ser recuperadas y no se gasten recursos como CPU y memoria.
8. *Record Files.* Son archivos donde se almacenan una parte de los datos del grafo.
9. *Discos.* Son los discos físicos donde se almacena la información de la base de datos.

Modelado

El modelado de datos es el proceso mediante el cual un usuario describe cualquier dominio como un grafo compuesto de nodos conectados mediante relaciones con propiedades y etiquetas. Un modelo de datos de grafos de Neo4j es útil para responder preguntas en formas de consultas de Cypher (Lenguaje de consultas de Neo4j) y resolver problemas técnicos, sociales y comerciales mediante la organización de una estructura de datos.

Dominio

Lo primero que se realiza es la elección del tema o dominio sobre el cual se obtendrá el conjunto de datos. A manera de ejemplo, para entender el proceso de modelado, sobre esta tesis el dominio es la realización del trabajo de investigación, sobre el cual se recorre paso a paso la creación del modelo de grafos a partir del siguiente enunciado: "Dos alumnos, Carlos y Alejandro son autores de la tesis KGraphFI, siendo asesorados por el profesor Guillermo, quien les impartió la asignatura de Minería de Datos".

Se usa la información del fragmento anterior para generar el modelo identificando los componentes, como: a) nodos, b) relaciones, c) etiquetas, y d) propiedades.

a) Nodos. Las primeras entidades que se identifican en el dominio son los nodos. Estos nodos son unidades fundamentales que forman el modelo de los grafos de propiedades. Estos nodos pueden contener propiedades formadas por pares de datos de nombre-valor. Además, a los nodos se les pueden asignar roles o tipos usando una o más etiquetas.

Con base en el ejemplo anterior, las entidades identificadas fueron resaltadas en negritas: "Dos alumnos, **Carlos** y **Alejandro** son autores de la tesis **KGraphFI**, siendo asesorados por el profesor **Guillermo**, quien les impartió la asignatura de **Minería de Datos**". A partir de estas entidades se establecieron los siguientes nodos del modelo: Carlos, Alejandro, KGraphFI, Guillermo y Minería de Datos; mostrados en la Figura 2.8.

Fig. 2.8: Nodos identificados para el modelo del grafo.

b) Etiquetas. Son nombres que se utilizan para agrupar nodos en conjuntos. Todos los nodos etiquetados con el mismo nombre pertenecen a un mismo conjunto. Además, un nodo puede o no etiquetarse, lo que hace que éstas sean una adición opcional en el grafo.

Con base en el ejemplo anterior, se identificaron cuatro tipos de etiquetas: Alumno, Tesis, Profesor y Asignatura, las cuales fueron actualizadas en el modelo del grafo (Figura 2.9). Para Alejandro y Carlos se asigna el rol *Alumno*, para KGraphFi se asigna el rol *Tesis*, para Guillermo se asigna el rol *Profesor* y para Minería de Datos se asigna el rol *Asignatura*.

Fig. 2.9: Nodos y etiquetas identificados para la construcción del grafo.

c) Relaciones. Una vez identificadas las entidades y una forma de agruparlas, el siguiente paso es identificar las relaciones en el grafo. Una relación conecta dos nodos y permiten encontrar nodos de datos relacionados. Además, dado que una relación tiene un nodo inicial y un nodo final, no es posible eliminar un nodo sin eliminar también sus relaciones asociadas.

Continuando con el ejemplo planteado, las relaciones entre los nodos son (Figura 2.10):

- Alejandro **es autor de** KGraphFI
- Carlos **es autor de** KGraphFI
- Guillermo **imparte** Minería de Datos
- Guillermo **es asesor de** Alejandro
- Guillermo **es asesor de** Carlos

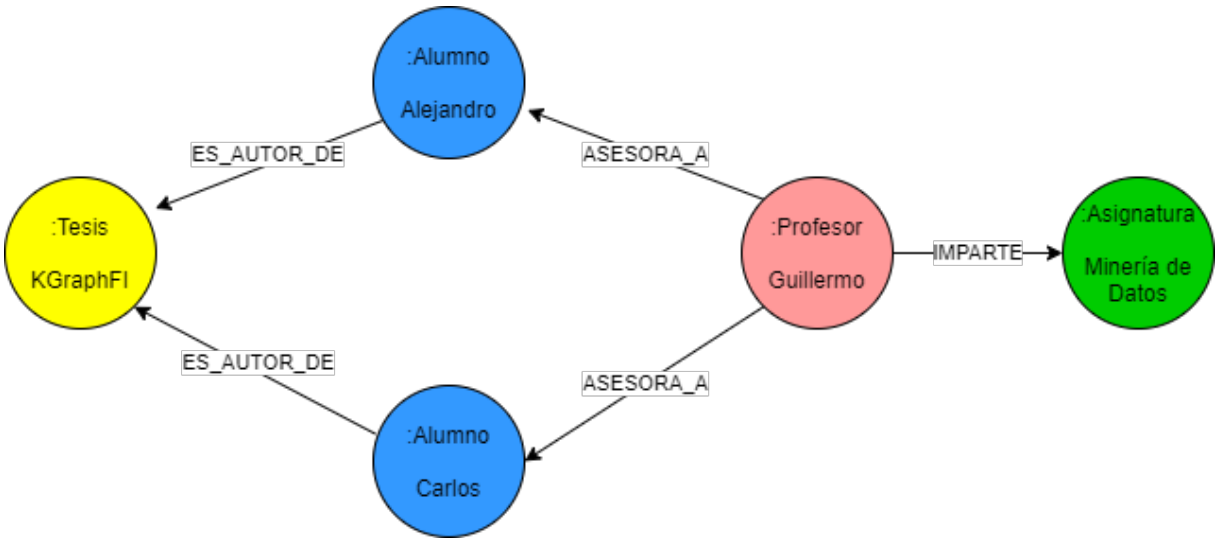

Fig. 2.10: Nodos, etiquetas y relaciones identificados para la construcción del grafo.

d) Propiedades. El último paso es definir, en el modelo de datos, las propiedades clave-valor. Estas propiedades permiten almacenar datos relevantes sobre el nodo o la relación con la entidad que describe. Las propiedades se pueden encontrar a través de preguntas que extraen información de un nodo en particular. Con respecto al ejemplo anterior, algunas preguntas de interés podrían ser:

- ¿Cuántas páginas tiene la tesis KGraphFi?
- ¿Qué edad tienen Carlos y Alejandro?
- ¿Desde cuándo Guillermo imparte la asignatura de Minería de Datos?
- ¿Qué clave tiene la asignatura de Minería de Datos?

A partir de estas y otras preguntas es posible identificar los atributos necesarios a agregar en las entidades del modelo de datos (Figura 2.11). El propósito es dar respuesta a estas preguntas.

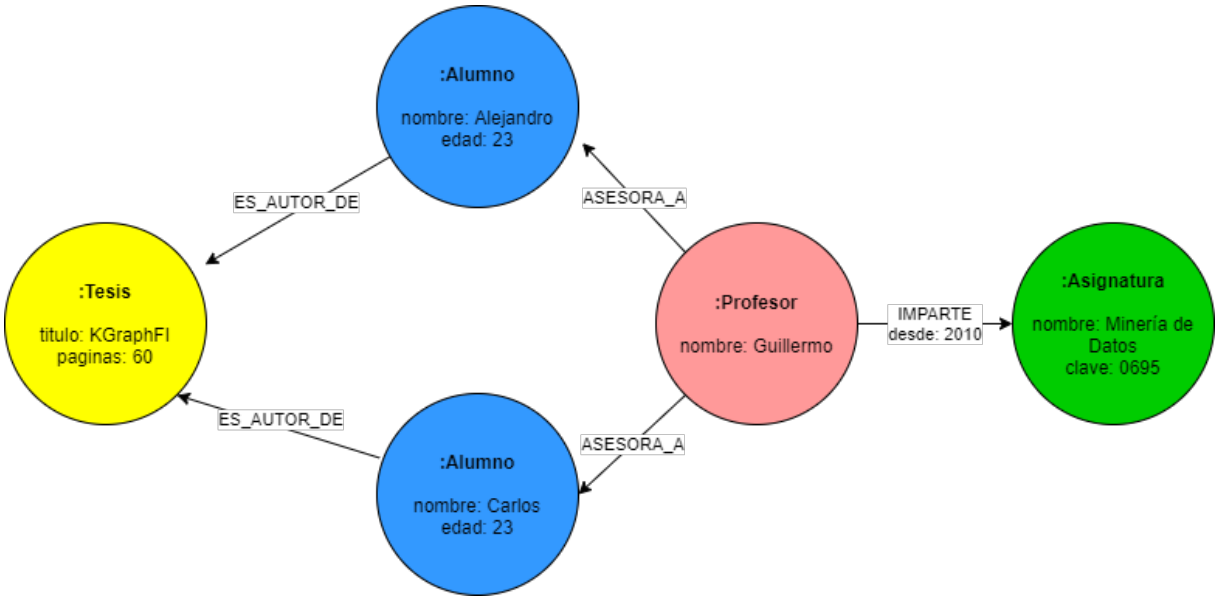

Fig. 2.11: Modelo de datos final obtenido orientado a grafos.

2.5. Minería de grafos

Los grafos de conocimiento, como se mencionó en la sección 2.2, permiten organizar la información de manera estructurada. A partir de esa información, es posible obtener conocimiento de interés a través de los nodos y siguiendo sus relaciones (aristas). Además, es posible obtener mayor conocimiento con base en la información no explícita en el grafo.

Por otra parte, de acuerdo con Han y col. (2012) la minería de datos es el proceso de descubrir patrones y conocimiento de largas cantidades de datos. En este sentido, tomando los datos (información) que se encuentran en el grafo de conocimiento, se puede identificar, a través de la minería de datos, patrones, tendencias e información de interés.

2.5.1. Conocimiento inductivo

El conocimiento inductivo (inductive knowledge) se refiere al hecho de realizar una generalización de patrones dado un conjunto de observaciones (Hogan y col., 2020). Se puede obtener este conocimiento utilizando dos aproximaciones:

1. *Métodos supervisados*. Este tipo de métodos aprenden una función o un modelo a partir de un conjunto de datos, para después mapear nuevos elementos a su correspondiente clasificación. Para este proceso se necesita que el conjunto de datos esté etiquetado.
2. *Métodos no supervisados*. Este tipo de métodos no requieren que los datos estén etiquetados previamente. En este caso el modelo se aplica a un conjunto de datos no etiquetados para así definir una posible etiqueta (categoría).

2.5.2. Analítica de grafos

Conforme a Hogan y col. (2020), la analítica es el proceso de descubrir, interpretar y comunicar patrones valiosos propios de las colecciones de datos. Por lo que, la analítica de grafos (Graph Analytics) es el uso de la analítica sobre un conjunto datos en grafos. Los algoritmos de grafos son un conjunto de herramientas utilizadas para llevar a cabo este proceso de descubrimiento y entendimiento de patrones de datos. De acuerdo con Needham y Hodler (2019), los algoritmos sobre grafos se pueden definir en tres categorías: i) algoritmos de exploración o búsqueda, ii) algoritmos de centralidad, iii) y algoritmos de detección de comunidades.

Algoritmos de exploración o búsqueda

Los algoritmos de búsqueda permiten explorar el grafo y recorrer las relaciones existentes entre los nodos. Algunos algoritmos dentro de esta categoría destacan:

1. *Búsqueda primero en amplitud* (Breadth-First Search o BFS, por sus siglas en inglés). Teniendo un nodo de partida, este algoritmo permite recorrer todos los nodos a los que pueden ser accedidos desde ese nodo. El algoritmo empieza en el nodo de partida y recorre todos los nodos vecinos y los marca como explorados, después va visitando cada uno de los nodos vecinos que aún no han sido explorados, así hasta que se hayan recorrido todos los nodos (Sedgewick y Wayne, 2011). Por ejemplo, sea el grafo mostrado en la Figura 2.12, el algoritmo de búsqueda primero en amplitud sería:

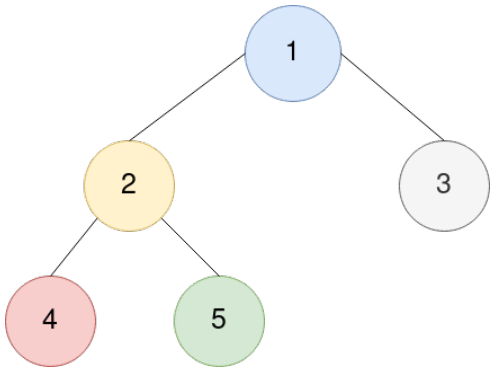

Fig. 2.12: Ejemplo de Breath First Search y Depth First Search.

   a) El nodo inicial es 1.
   b) Se visitan sus vecinos, que son los nodos 2 y 3.
   c) Se visitan los vecinos del nodo 2, que son los nodos 4 y 5.
   d) Como el nodo 3 no tiene vecinos, se pasa al siguiente nodo.
   e) Como el nodo 4 no tiene vecinos, se pasa al siguiente nodo.
   f) Como el nodo 5 no tiene vecinos, se pasa al siguiente nodo.
   g) Finaliza el recorrido.

De lo anterior, se deduce que se explora el grafo por niveles, el primer nivel siempre tiene un único nodo (nodo 1), después se exploran los nodos del nivel 2 (nodos 2 y 3), y así sucesivamente hasta explorar todos los niveles del grafo.

2. *Búsqueda primero en profundidad* (Depth First Search o DFS, por sus siglas en inglés). Teniendo un nodo de partida, este algoritmo al igual que el anterior permite encontrar o recorrer todos los nodos que pueden ser accedidos desde ese nodo. El algoritmo empieza en el nodo de partida y explora los nodos en profundidad (Sedgewick y Wayne, 2011). Por ejemplo, con base en la Figura 2.12, el algoritmo sería el siguiente:

   a) El nodo inicial es 1.
   b) Se visita al primer vecino (nodo 2) del nodo 1.
   c) Se visita al primer vecino (nodo 4) del nodo 2.
   d) Este nodo no tiene vecinos, entonces se regresa al nodo 2
   e) Se visita al siguiente vecino (nodo 5) del nodo 2.
   f) Este nodo no tiene vecinos, entonces se regresa al nodo 2.
   g) Este nodo no tiene más vecinos que visitar, se regresa al nodo 1.
   h) Se visita al siguiente vecino (nodo 3) del nodo 1.
   i) Este nodo no tiene vecinos, entonces se regresa al nodo 1.
   j) Finaliza el recorrido.

   De *a* a *c*, el recorrido queda como *1-2-4*, es decir, se va explorando el grafo hacia abajo (en profundidad). Una vez que no haya más nodos que explorar en profundidad se regresa al nodo anterior, como se observa en el punto *d*. A esta idea de regresar se le conoce como *backtracking* (Sedgewick y Wayne, 2011).

3. *Camino aleatorio* (Random walk). Este algoritmo regresa un camino random sobre un grafo (Needham y Hodler, 2019). El algoritmo empieza en un nodo y sigue una de las relaciones con sus vecinos de manera aleatoria, así sucesivamente con todos los nodos que se recorren hasta alcanzar una longitud deseada del camino (número de aristas recorridas). El término random se refiere al número de relaciones de un nodo y sus vecinos, lo cual influirá si pasa o no por un nodo. Se usa este algoritmo principalmente para el entrenamiento de modelos de aprendizaje automático.

Algoritmos de centralidad

Los algoritmos de centralidad son útiles para identificar los nodos más importantes en el grafo. Además, permiten identificar la dinámica de todo el grafo, por ejemplo, la accesibilidad y velocidad con la que se propaga la información, a través los nodos. Algunos algoritmos de centralidad conocidos son:

1. *Centralidad de grado* (Degree centrality). Este algoritmo cuenta con un número de relaciones de entradas y salidas de un nodo. El algoritmo se utiliza para buscar los nodos más populares en el grafo. El grado de un nodo es el número de relaciones directas que tiene, y se divide en grados de entrada y salida. El grado medio de un grafo es el número total de relaciones dividido por el número total de nodos. Sin embargo, esto puede estar sesgado por nodos de alto grado. Así, la distribución de grados es la probabilidad de que un nodo seleccionado al azar tenga un cierto número de relaciones (Needham y Hodler, 2019).

2. *Centralidad de proximidad* (Closeness centrality). Es una forma de detectar nodos que pueden difundir información de manera eficiente a través de un subgrafo, es decir, son los nodos que pueden alcanzar a los demás nodos de manera inmediata. La medida de la centralidad de un nodo es su distancia promedio (distancia inversa) a todos los demás nodos (Needham y Hodler, 2019). Los nodos con una puntuación de proximidad alta tienen las distancias más cortas de todos los demás nodos. Algunos algoritmos de centralidad de cercanía son: Closeness centrality, Wasserman-Faust, y centralidad armónica (Harmonic Centrality.

3. *Centralidad de intermediación* (Betweenness Centrality). Algunas veces, la parte más importante de un sistema no es la más fuerte o el estado más alto, sino son los intermediarios que conectan a los grupos o los que tienen control sobre los recursos o el flujo de información. El algoritmo de centralidad de intermediación calcula la influencia que tiene un nodo sobre el flujo de información o recursos del grafo (Needham y Hodler, 2019). Por lo general, se usa para encontrar nodos que sirvan como puente en una parte del grafo.

4. *PageRank*. Mide la influencia direccional de los nodos. Se calcula distribuyendo iterativamente el rango de un nodo entre sus vecinos o atravesando aleatoriamente el grafo y contando la frecuencia con la que se toca cada nodo. Este algoritmo mide el número y la calidad de las relaciones entrantes a un nodo para determinar una estimación de la importancia de ese nodo. Los nodos con más influencia sobre una red tienen más relaciones entrantes de otros nodos influyentes (Needham y Hodler, 2019).

Para ejemplificar lo mencionado, en la Figura 2.13 se muestra los algoritmos de centralidad, donde se toman los nodos con más conexiones en el grafo. La centralidad de proximidad toma el nodo que alcanza a los demás de manera inmediata. La centralidad de intermediación identifica los puentes en el grafo y PageRank es útil para conocer qué nodos son los más importantes en el grafo.

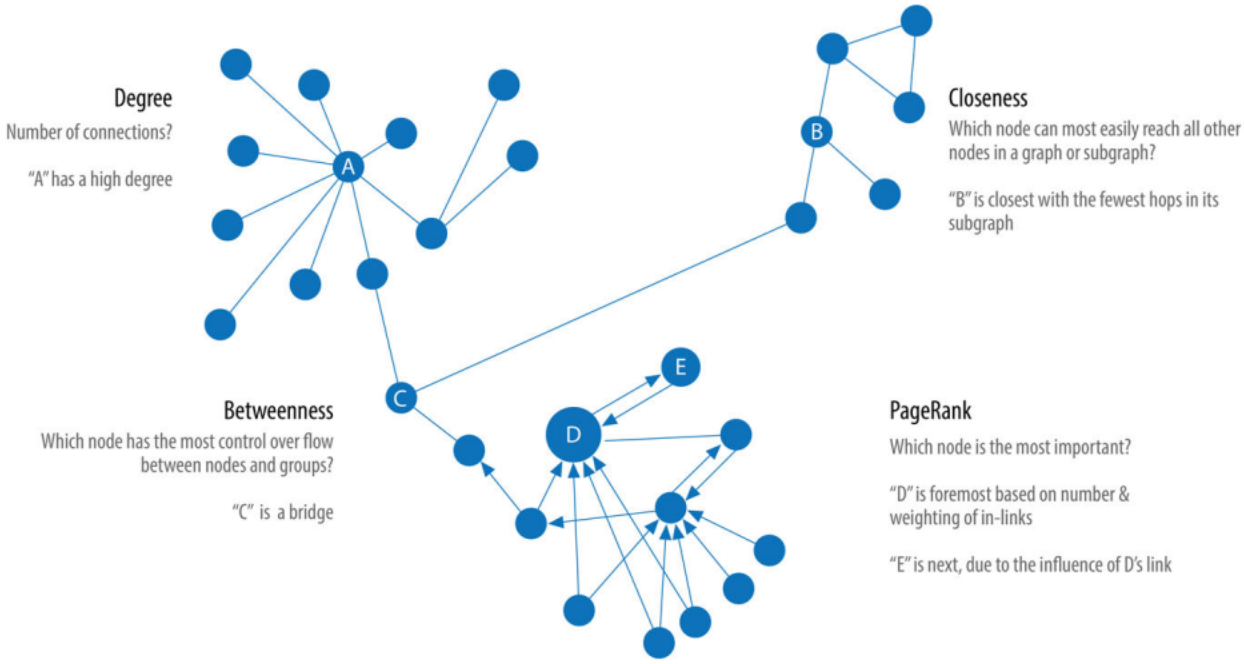

Fig. 2.13: Esquema de los algoritmos de centralidad. Fuente: Needham y Hodler (2019).

Algoritmos de detección de comunidades

Este tipo de algoritmos sirven para encontrar comunidades, donde los miembros significativos tienen más relaciones con otros nodos dentro del grupo, lo cual permite analizarlos y ver que características tienen en común. Algunos algoritmos de detección de comunidades son:

1. *Cuenta de triángulos y coeficiente de agrupamiento* (Triangle count and clustering coefficient). Un triángulo es un conjunto de 3 nodos, donde cada nodo tiene una relación con los otros dos nodos. Por lo que, el conteo de triángulos determina el número que pasan por cada nodo. Por otro lado, el coeficiente de agrupamiento calcula que tan agrupados están los nodos, comparado con los que podrían estar. Estos dos algoritmos tienen relación, puesto que el coeficiente de agrupamiento utiliza la cuenta de triángulos para ser calculado, y permiten saber que tan agrupados pueden estar los nodos vecinos (Needham y Hodler, 2019). Para el cálculo del coeficiente de agrupamiento se tienen dos tipos de coeficientes.

   a) *Coeficiente de agrupamiento local*. Se calcula el coeficiente de agrupamiento de un nodo y se define la probabilidad de que sus vecinos también estén conectados. Este es calculado usando la siguiente fórmula:

   $$CC(u) = \frac{2R_u}{k_u(k_u-1)}$$

   donde: u es el nodo del que se desea calcular el coeficiente de agrupamiento, R_u el número de triángulos que pasan por u y k_u es el grado de u.

La Figura 2.14 muestra un ejemplo de distintos coeficientes de agrupamiento para distintos nodos.

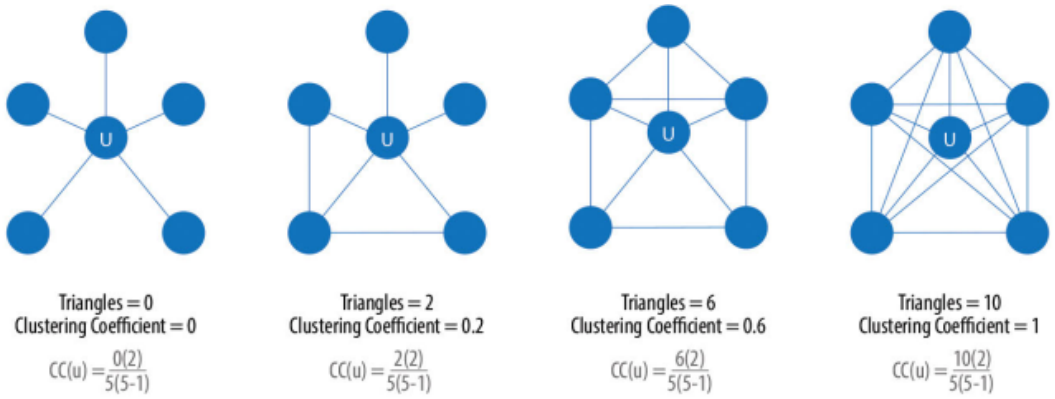

Fig. 2.14: Agrupamiento para distintos nodos. Fuente: Needham y Hodler (2019).

b) *Coeficiente de agrupamiento global.* Es la suma normalizada de los coeficientes de agrupamiento de los nodos:

$$\overline{CC} = \frac{1}{n}\sum_{i=1}^{n} CC(u_i)$$

donde: n es el número de nodos en el grafo.

Este algoritmo sirve para detectar comunidades en redes sociales y detectar que tan agrupadas están ciertas redes (grafos).

2. *Componentes altamente conectadas* (Strongly connected components). Un componente altamente conectado es un conjunto de nodos, tal que para cada par de vértices, estos se pueden alcanzar, es decir, para todo par de nodos $\{u, v\}$, se tiene un camino de aristas para llegar de u a v y otro para llegar de v a u (Cormen y col., 2009). A manera de ejemplo, en la Figura 2.15 se observa que existen dos componentes, una conformada por los nodos $\{A, B, C\}$ y otro con $\{D, E\}$. Para llegar del nodo b a c, el camino es $b \to a \to c$, y de c a b es sólo $c \to b$, por lo que $\{b, c\}$ son alcanzables entre estos.

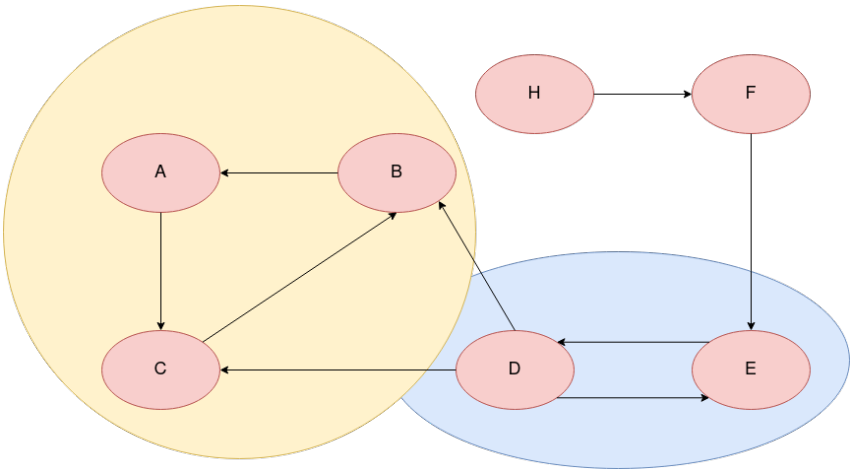

Fig. 2.15: Ejemplo de componentes altamente conectados.

Este algoritmo tiene como aplicación encontrar grupos altamente conectados, por ejemplo en sistemas de recomendaciones para encontrar grupos de personas con gustos similares.

3. *Label propagation* (Propagación de etiquetas). Es un algoritmo utilizado para encontrar comunidades en un grafo. Cada grupo de un nodo es seleccionado basado en sus vecinos directos. Se basa en que una etiqueta puede ser propagada en un conjunto de nodos densamente conectados, pero por otro lado tendrá problemas en ser propagada en un grupo de nodos dispersos. Este algoritmo es de tipo semi-supervisado, puesto que se ponen etiquetas preliminares (seed labels) en algunos nodos y los otros se quedan sin etiquetar para que el algoritmo pueda aprender (Needham y Hodler, 2019). La Figura 2.16 muestra un ejemplo de propagación de etiquetas.

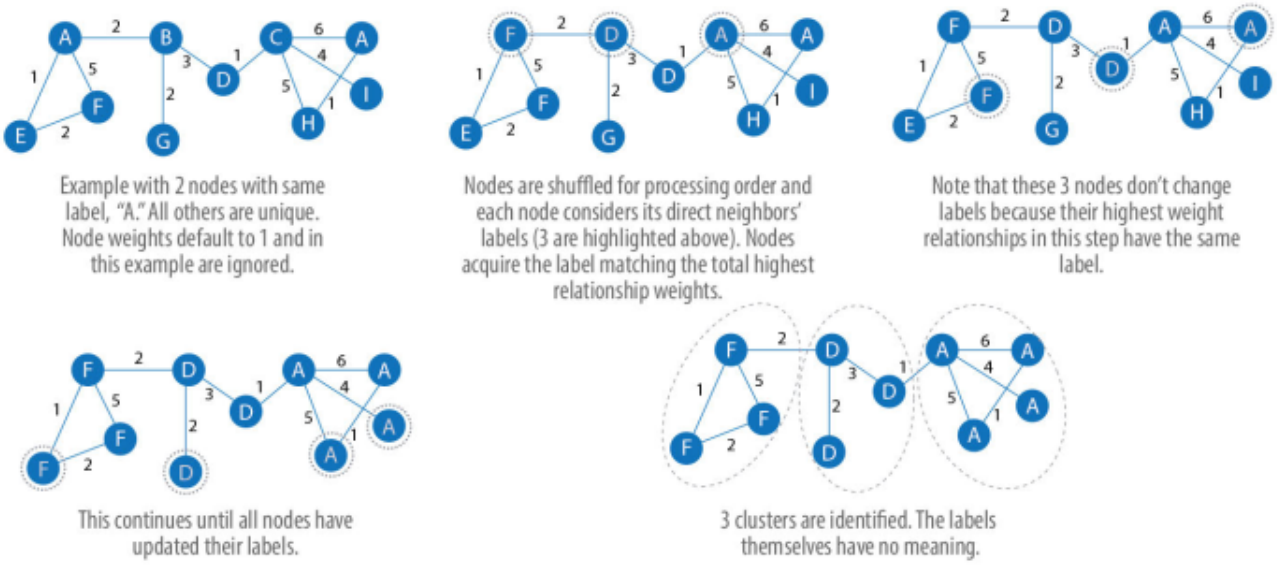

Fig. 2.16: Ejemplo de Label Propagation. Fuente: Needham y Hodler (2019).

El algoritmo es el siguiente:

a) Cada nodo es inicializado con un identificador único.

b) Cada etiqueta es propagada a través de la red.

c) En cada iteración, cada nodo actualiza su etiqueta para tener la misma que el nodo vecino con la arista con mayor peso.

d) El algoritmo termina cuando todos los nodos han sido etiquetados (basado en las etiquetas de sus vecinos).

2.5.3. Aplicaciones actuales

Existen numerosas áreas donde los grafos de conocimiento se puede aplicar, prácticamente en todas las actividades humanas en las que se generan datos.

Redes sociales

Un desafío en las redes sociales es buscar la representación y análisis, de manera eficiente, de los datos que cambian constantemente. Se busca aprovechar los registros de las actividades de los usuarios y sus interacciones a través de los diversos dispositivos de comunicación. Se busca además realizar consultas eficientes sobre la presencia de las interacciones de nodos individuales y los metadatos de los nodos, para esto se utilizan datos recopilados, variables en el tiempo, con el propósito de modelarlos en forma de grafos y analizar diferentes aspectos, como (Cattuto y col., 2013): popularidad, recomendaciones de amistad, intereses, entre otros.

Detección de fraude

Es especialmente útil para analizar fraude bancario, fraude de seguros, fraude de comercio electrónico o cualquier otro tipo de fraude; para esto se abordan dos retos esenciales (Sdowski y Rathle, 2020):

- Detectar el fraude lo más rápido posible para que los delincuentes no tengan el tiempo suficiente de seguir provocando un impacto negativo. A medida de que los procesos se vuelven más rápidos y automatizados, los margenes de tiempo para detectar el fraude se vuelven cada vez más pequeños, aumentando la necesidad de soluciones en tiempo real.
- Darle el valor y asegurar los datos conectados, dado que los delincuentes sofisticados han aprendido a atacar sistemas, donde la estructura de datos conectados es débil.

Las tecnologías tradicionales si bien siguen siendo adecuadas y necesarias, sin embargo, para algunos tipos de prevención no están diseñadas para detectar ciertos círculos o patrones de fraude. Por lo que, las bases de datos orientadas a grafos son la herramienta ideal para soluciones de detección de fraude eficientes y manejables. Desde redes de fraude y grupos organizados hasta

delincuentes que operan por su cuenta. Estas bases de datos ofrecen la capacidad para descubrir una variedad de patrones de fraude en tiempo real. Las colusiones que antes estaban ocultas se vuelven obvias cuando se analiza los datos conectados, utilizando consultas gráficas en tiempo real (Sdowski y Rathle, 2020).

Gestión de datos personales

Los requisitos generales de protección de datos (GDPR, por sus siglas en inglés) es la regulación de la Unión Europea sobre el almacenamiento y administración de los datos personales de sus residentes. Bajo estas reglas, todas las empresas deben comprender, controlar y administrar estrictamente la posesión de los datos personales o, en caso del no cumplimiento, enfrentar procesos judiciales y fuertes multas. Por lo que, las bases de datos orientadas a grafos son una solución para la regulación GDPR, que conecta datos personales en todos sus sistemas:

- La ubicación de la información privada.
- Qué sistemas y aplicaciones utilizan los datos.
- Cómo y cuándo se utilizan los datos personales.
- Quién mira y usa los datos.
- Qué permisos tiene para utilizar los datos, y cuándo y cómo se obtuvieron.
- Dónde y cómo se mueven los datos personales.

Esta tecnología ayuda a visualizar los datos privados y controlar su uso, mientras que los miembros del personal interno brindan respuestas rápidas a las consultas de privacidad. También permite a los administradores de privacidad rastrear los flujos de datos, investigar posibles infracciones y demostrar el cumplimiento a los reguladores. Sin la tecnología de grafos sería imposible comprender el ciclo de vida de los datos personales (*GDPR Compliance*, 2020).

Sistemas de recomendación

El enunciado “Esto te puede interesar...” es una frase engañosa que encapsula una nueva era en la gestión de relaciones con los clientes (Webber, 2020). Se puede ofrecer sugerencias personalizadas que maximicen el valor de las recomendaciones de productos en tiempo real. Esta capacidad de realizar ofertas atractivas requiere de la tecnología de las bases de datos orientadas a grafos. Esta tecnología captura el historial de compras del cliente y también analiza opciones actuales.

Así, la tecnología de grafos supera ampliamente a las bases de datos relacionales y las NoSQL, cuando se trata de conectar grandes cantidades de datos de compradores y productos para obtener información sobre las necesidades de los clientes y las tendencias de los productos. Las bases de datos gráficas son una plataforma tecnológica central de gigantes de Internet como Google, Facebook y LinkedIn.

Riesgos financieros

Los datos relacionados con el riesgo financiero están intrínsecamente conectados. Los bancos de hoy deben tener una comprensión completa y sistemática de los cálculos de riesgo y los datos asociados, desde su origen hasta cómo fluyen a través de todos los sistemas empresariales, para comprender su verdadero riesgo financiero.

Por lo que, las tecnologías de bases de datos orientadas a grafos y los datos conectados han transformado la presentación de los informes de riesgos en los bancos modernos para ayudar a prevenir pérdidas importantes y cumplir con las demandas de cumplimiento de informes o reportes (Mathur, 2019).

Campo sanitario

En la actualidad, detener la propagación de una infección o virus es la principal preocupación para los hospitales. Existen algunos avances, como en el hospital de Presbyterian ubicado en Nueva York, donde se propuso una forma de rastrear la propagación de una infección, buscando identificar puntos de conexión y contagio de un paciente infectado. Para esto la tecnología orientada a grafos ofrece una forma flexible de conectar datos de qué, cuándo y dónde sucedió. Posteriormente, el algoritmo agrupa los eventos en diversas especialidades como oncología y pediatría, validando el trabajo del equipo (Zelenetz, 2020).

Precisamente, a partir de este último campo de aplicación, se busca a través de este trabajo de tesis proponer una aproximación de grafos de conocimiento de propiedades para analizar la mortalidad de embarazadas diagnosticadas con COVID-19 atendidas en la Ciudad de México.

2.6. COVID-19

En diciembre de 2019, la ciudad de Wuhan, capital de la provincia de Hubei en China, se convirtió en el centro de un brote de neumonías atípicas causadas por un virus, llamado coronavirus tipo 2 del síndrome respiratorio agudo grave (SARS-CoV-2, por sus siglas en inglés). Según las autoridades de Salud de China, entre el 12 y 19 de diciembre se presentaron varios casos similares, siendo reportados a la Organización Mundial de la Salud (World-Health-Organization, 2020a).

Un evento en común entre los pacientes iniciales fue la asistencia al mercado mayorista de mariscos de Huanan, en Wuhan, donde también se vendían otro tipo de animales. Como consecuencia, el 1 de enero de 2020, las autoridades de China decidieron cerrar el mercado al descubrir que los animales vendidos podrían ser la fuente del virus desconocido. Posteriormente, se decidió aislar a las personas contagiadas y analizar las moléculas infectadas, las cuales no correspondían a neumonías provocadas por enfermedades conocidas. Por lo que, se inició una investigación en retrospectiva sobre el brote del virus el 5 de enero de 2020, y para el 7 de enero las autoridades de China confirmaron que identificaron el virus como un nuevo coronavirus (Cable-News-Network, 2020).

2.6.1. El nuevo coronavirus

De acuerdo con Sun y col. (2020), el coronavirus es una extensa clase de virus genéticos encontrados en animales, así como también en humanos. La primera aparición de coronavirus en el ser humano se registró entre 2002 y 2003, cuando se reportó una neumonía atípica en la provincia de Guangdong, al Sureste de China. Este padecimiento fue nombrado como Síndrome Respiratorio Agudo Severo (Severe Acute Respiratory Syndrome), el cual es una enfermedad respiratoria viral causada por coronavirus (SARS-CoV, por sus siglas en inglés). La enfermedad se propagó en varios países de Norteamérica, Suramérica, Europa y Asia antes de que pudiera ser contenida el brote global.

Posteriormente, otro coronavirus surgió en 2012 en Arabia Saudita, esta vez conocido como Síndrome Respiratorio del Este Medio (Middle East Respiratory Syndrome), MERS-CoV, el cual es una enfermedad respiratoria grave que involucra principalmente al tracto respiratorio superior, que causa fiebre, tos y dificultad para respirar (World-Health-Organization, 2020c). El MERS-CoV tiene como característica principal la baja transmisión entre humanos, pero con un alto índice de mortalidad, aproximadamente el 30 % de las personas que han contraído la enfermedad han muerto. Según reportes de World-Health-Organization (2020c), se han presentado desde 2012 hasta la fecha un total de 2494 casos a nivel mundial.

Estas enfermedades de SARS-Cov y MERS-CoV plantearon amenazas para la salud mundial debido a las altas tasas de mortalidad, esto es, de 9.6 y 34.4 %, respectivamente. Sin embargo, en la actualidad, se convive con una variante de coronavirus, SARS-Cov-2, que de acuerdo con World-Health-Organization (2020d) es un tipo cuyo contagio mundial ha provocado la pandemia de 2020, inicialmente llamado 2019-nCoV (nuevo coronavirus de 2019). A la fecha, uno de los mecanismos de transmisión, que puede producir el contagio de una persona a otra, es mediante las gotas de saliva expulsadas a través de la tos y el estornudo o al espirar, que puede provocar enfermedad respiratoria aguda y neumonía grave (Sun y col., 2020).

2.6.2. Síntomas

Los síntomas más frecuentes por COVID-19 son fiebre, tos seca, agotamiento. Algunos síntomas de menor presencia son la congestión nasal, dolor de cabeza, conjuntivitis, dolor de garganta, diarrea, pérdida del gusto y olfato, entre otros (Sun y col., 2020). Además, estos síntomas pueden comenzar de manera leve y agravarse con el tiempo. El 80 % de las personas se recuperan sin necesidad de un tratamiento hospitalario, mientras que el otro 20 % presentan un cuadro grave y experimentan dificultad para respirar.

Por otro lado, las personas de la tercera edad y las que padecen alguna enfermedad crónica, como diabetes, hipertensión, problemas cardiacos, problemas pulmonares o cáncer, tienen más probabilidad de presentar casos graves (Rodríguez y col., 2020a), con síntomas variados, como fiebre, disnea, tos, mialgia, astenia, delirio; y otros síntomas atípicos, como el deterioro del estado funcional y la existencia de delirio hiper e hipoactivo (Bianchetti y col., 2020; Godaert y Proye, 2020). Es importante mencionar que también se presentan casos asintomáticos, lo que hace más compleja la detección del virus en el momento oportuno.

2.6.3. Transmisión

El virus de COVID-19 se transmite principalmente entre personas a través del contacto y de gotículas respiratorias. Por lo que, cualquier persona que esté en contacto con otra infectada puede también enfermarse. El virus se propaga de manera rápida, de persona en persona, por medio de los residuos naturales producidos por la nariz o la boca al hablar, toser o estornudar; siendo esta la razón por la cual es importante mantenerse a una distancia mínima de 1.5 metros con respecto a otras personas (World-Health-Organization, 2020c). Las gotículas respiratorias suelen tener un diámetro de 5 a 10 micrómetros (μm), y otros núcleos goticulares tienen un diámetro aún menor de 5 μm (World-Health-Organization, 2020e).

Debido al riesgo de contagio por contacto directo cercano (a menos de un metro) en el entorno inmediato de una persona infectada, la recomendación ha sido evitar exponer las mucosas de la boca y nariz usando cubrebocas, así como la conjuntiva (ojos) mediante caretas faciales. Pero también el contagio puede ser de forma indirecta, esto es, por el contacto con superficies u objetos que hayan sido utilizados por personas infectadas, por ejemplo, un estetoscopio o termómetro, entre otros (Trilla, 2020).

Se han presentado también otras formas de transmisión, como la aérea, donde el virus puede estar latente en lugares específicos, por ejemplo, donde se efectúan procedimientos o tratamientos por intubación endotraqueal, broncoscopia, aspiración abierta, administración de un fármaco por nebulización, ventilación manual antes de la intubación, desconexión del paciente de un ventilador, ventilación no invasiva con presión positiva, traqueostomía y reanimación cardiopulmonar (Trilla, 2020). En este sentido, la transmisión por gotículas es distinta de la transmisión aérea, esta última tiene lugar a través de núcleos goticulares que contienen microbios. Los núcleos goticulares, que tienen un diámetro inferior a 5 μm, pueden permanecer en el aire durante periodos prolongados y llegar hasta las personas que se encuentren a más de un metro de distancia.

2.6.4. Etapas de COVID-19

Según Gobierno-de-México (2020c) la expansión del virus COVID-19 se divide en tres etapas:

1. *Importación de casos*. El número de infectados es reducido dónde los contagios han sido fuera del país. Esta fase carece de medidas de restricción de movilidad, solamente se toman las medidas de higiene adecuadas para evitar un contagio.
2. *Transmisión comunitaria*. Empiezan los contagios dentro del país que no hayan tenido necesariamente contacto con pacientes expuestos a la infección en el extranjero. Durante esta fase empiezan medidas de restricción de movilidad como suspensión de clases, trabajo a distancia y cancelación de eventos masivos debido a que suelen aumentar rápidamente los casos registrados.
3. *Etapa epidemiológica*. Es cuando hay miles de personas contagiadas. Para detener el avance del virus se ponen restricciones de movilidad más agresivas como cuarentena generalizada.

Comparado con el SARS-CoV y MERS-CoV, el COVID-19 es mucho más contagioso,

esto reflejado por el número de casos confirmados. El primero reportó un total de 8096 casos (World-Health-Organization, 2015) y el segundo 2494 (World-Health-Organization, 2020b) en comparación de los mas 81 millones de casos de COVID-19 en el mundo al 31 de diciembre de 2020 (World-Health-Organization, 2020f).

2.7. Trabajos relacionados

En Parra-Bracamonte y col. (2020) se analizaron datos de las características clínicas y los factores de riesgo de mortalidad de pacientes con COVID-19 en México. El método utilizado fue un modelo de regresión logística multivariante y las curvas de supervivencia de Kaplan–Meier para estudiar las probabilidades de muerte y comorbilidad en los pacientes con COVID-19. Los resultados arrojaron que la edad, el sexo y las comorbilidades más frecuentes como obesidad e hipertensión se asociaron significativamente al riesgo de muerte. Mientras que las comorbilidades menos frecuentes, como la enfermedad pulmonar obstructiva crónica y la enfermedad renal crónica, también mostraron un significativo riesgo de muerte. Así, se mostró que un paciente positivo vulnerable, con más riesgo de muerte, es representado por una persona de sexo masculino, mayor de 41 años, si presenta las comorbilidades de obesidad e hipertensión.

En Hernández (2020) se dieron a conocer las estadísticas de mortalidad por COVID-19 en México hasta el 27 de mayo de 2020. Para el análisis se consideraron los datos de los certificados de defunción que se asentaron como causa principal la palabra cov o covid. La investigación fue descriptiva y los resultados arrojaron que del 27 de mayo al 10 de junio de 2020 (fecha de corte del análisis), el número de muertos por COVID-19 pasó —según la información oficial— de 8597 a 15357 casos, es decir 78 % de incremento en trece días, lo que implicó un ritmo de duplicación cada 16 días, con respecto de la cifra base del análisis. Por lo tanto, llegaron a la conclusión de que aun cuando los resultados presentados son de carácter preliminar, apuntan a una tendencia, en términos del perfil demográfico y social de los mexicanos que sufren y sufrirán la consecuencia extrema de este padecimiento.

De manera similar, Rotmensch y col. (2017) estudiaron un proceso automatizado para aprender bases de conocimiento de alta calidad que vinculan enfermedades y síntomas directamente de los registros médicos electrónicos. Se extrajeron conceptos médicos de 273,174 registros de pacientes anonimizados y se utilizó la estimación de máxima verosimilitud de tres modelos probabilísticos para construir automáticamente los grafos de conocimiento: regresión logística, clasificación de Bayes y una red Bayesiana utilizando compuertas noisy OR. Se obtuvo un grafo de las relaciones enfermedad-síntoma a partir de los parámetros aprendidos y los grafos de conocimiento construidos se evaluaron y validaron, contra el grafo de conocimiento construido manualmente por Google y contra las opiniones de médicos expertos. El estudio mostró que la construcción automatizada de grafos de conocimiento de alta calidad, a partir de registros médicos, es factible. El modelo a partir de compuertas noisy OR produjo un grafo de conocimiento de alta calidad que alcanza una precisión de 0.85 (85 %).

Por otro lado, en Rodríguez y col. (2020b) se analizaron variables asociadas con la mortalidad en una población de pacientes mayores de 80 años y con algún grado de dependencia funcional, hospitalizados por COVID-19 en un servicio de geriatría. El propósito del estudio

fue describir las características de los pacientes, determinar la tasa de mortalidad e identificar factores asociados. Para el análisis se recopilaron variables sociodemográficas, clínicas, funcionales, mentales, analíticas, radiológicas, terapéuticas y asistenciales. Los resultados arrojaron altas tasas de mortalidad en pacientes mayores hospitalizados por COVID-19, con mayor riesgo de fallecer en aquellos con dependencia funcional severa o deterioro cognitivo. Estos hallazgos refuerzan la importancia de la valoración médica para elaborar estrategias que permitan adecuar la toma de decisiones diagnósticas y terapéuticas, y optimizar la atención al paciente ante un nuevo brote epidémico.

2.8. Síntesis

La complejidad de analizar variables asociadas con la mortalidad de pacientes con COVID-19, en los habitantes de la Ciudad de México, requiere el uso de tecnologías especializadas, como los grafos de conocimiento, mediante los cuales se puede hacer una estructuración y representación de los datos a través de nodos, relaciones y propiedades. Así, una base de datos orientada a grafos cuenta con cuatro propiedades básicas para el almacenamiento persistente de datos, esto es, crear, leer, actualizar y eliminar. Además, este tipo de bases de datos, como Neo4j, está diseñada para tratar las relaciones y los datos con el mismo nivel de importancia. Esto hace que se evite tener información aislada, dado que todos los datos están conectados de manera eficiente. Ante esto, surge interés natural de desarrollar un grafo de conocimiento de propiedades para el análisis de la mortalidad de la población gestante infectadas con COVID-19 que fueron atendidas en la Ciudad de México.

Capítulo 3

Método de solución

En el capítulo anterior se presentaron los fundamentos de los grafos de conocimiento, los cuales permiten que la información sea estructurada de forma que facilite su organización para el posterior análisis de los datos. Estos datos, con base en un análisis adecuado, permiten descubrir información de interés, como tendencias y patrones internos dentro de la estructura de los datos. Se destacó las características de los grafos de propiedades, formado por nodos, relaciones y sus propiedades; siendo las bases de datos orientadas a grafos, como Neo4j, la tecnología que soporta la gestión de los datos, utilizando una estructura de grafos para la creación, lectura, actualización y eliminación de la información. Por otra parte, se abordaron también las principales características de COVID-19, como síntomas, tratamiento y etapas de la enfermedad. Por último, se describieron los trabajos relacionados, de los cuales se identificaron sus fortalezas y debilidades, dando lugar a esta propuesta de investigación.

En este capítulo se presenta el método utilizado para la construcción del grafo de conocimiento de propiedades para el análisis de la mortalidad por COVID–19 en embarazadas atendidas en la Ciudad de México. Como parte del método de investigación se definieron cuatro etapas de trabajo: a) adquisición de los datos, b) análisis y preparación de los datos, c) modelado del grafo de conocimiento, y d) creación del grafo en Neo4j. Las cuales fueron de tipo exploratoria dado el hecho de que los grafos de conocimiento son una disciplina emergente, y aplicada debido a que éstos se apoyan de conocimientos específicos.

3.1. Adquisición de los datos

La fuente de datos a partir de la cual se realizó la construcción del grafo de conocimiento de propiedades, para el análisis de la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México, fueron datos provenientes de la base de datos del Sistema Nacional de Vigilancia Epidemiológica (SINAVE), de la Dirección General de Epidemiología de la Secretaría de Salud (Figura 3.1). Estos datos fueron adquiridos a través del portal Web de datos abiertos sobre salud pública, acciones sociales y gasto público en la Ciudad de México (Secretaría de Salud Gobierno de la Ciudad de México, 2021). El SINAVE es un conjunto de estrategias y acciones epidemiológicas que permiten la elaboración de información epidemiológica útil para la salud pública. El sistema integra información proveniente de todo México y de todas las instituciones del Sistema Nacional

de Salud (SNS). La información producida por el SINAVE es con base en lineamientos colegiados en tres niveles administrativos: a) Comité Nacional de Vigilancia Epidemiológica (CONAVE), integrado por representantes de instituciones del Sistema Nacional de Salud; b) Comités Estatales de Vigilancia Epidemiológica (CEVE) con la participación de representantes institucionales del sector en cada entidad federativa; y c) Comités Jurisdiccionales para la Vigilancia Epidemiológica (COJUVES), donde participan representantes institucionales en cada jurisdicción.

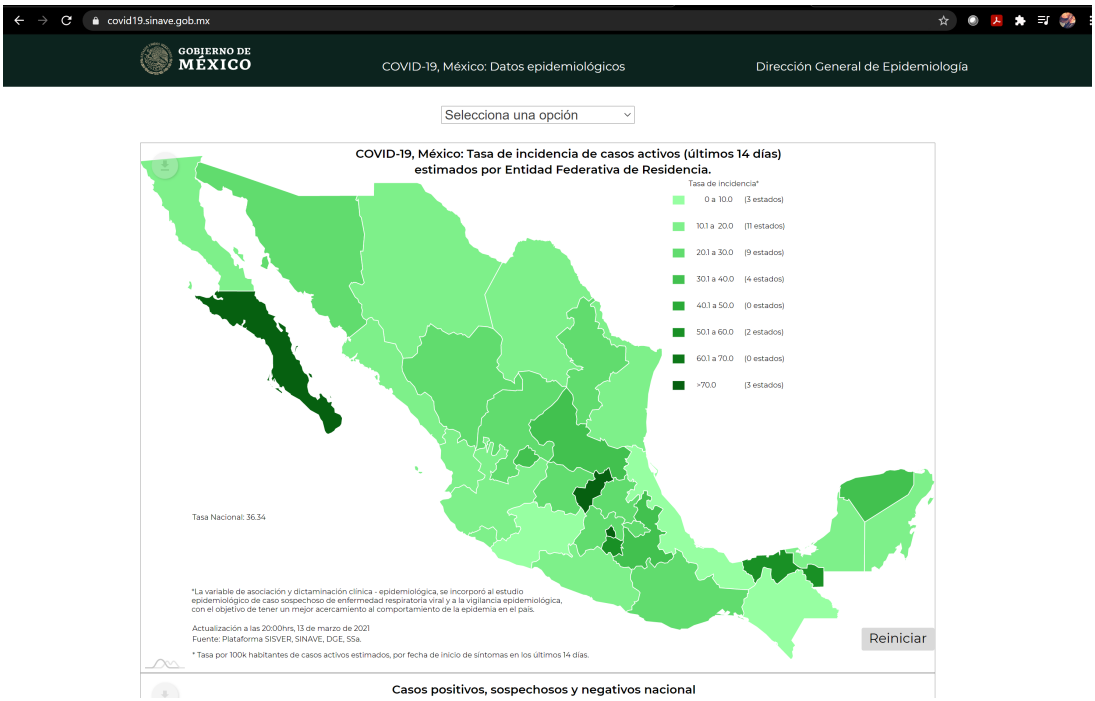

Fig. 3.1: Sitio web de datos abiertos sobre la incidencia de COVID-19 en México.

El método para la adquisición de datos fue retrospectivo debido a que se tomaron en cuenta casos registrados de mujeres embarazadas con confirmación de COVID-19, quienes fueron atendidas en hospitales ubicados en la Ciudad de México, correspondiente al periodo 2020. Se eligió este periodo de análisis debido a la disponibilidad de registros de datos de los pacientes gestantes con COVID-19 desde el inicio de la pandemia, marzo de 2020, hasta el 24 de diciembre de 2020, fecha de corte del análisis. Se determinó utilizar este grupo de población vulnerable debido a que el efecto de COVID-19 en el embarazo no se conoce por completo debido a la ausencia de datos fiables.

Por otra parte, los datos científicos parecen indicar actualmente que las personas de mayor edad, con sobrepeso y con afecciones preexistentes son más vulnerables a una forma grave de COVID-19. Sin embargo, de acuerdo con la OMS ("COVID-19: The Risk for Pregnant Women And Their Babies", 2020), nuevos estudios señalan que las embarazadas con COVID-19 tienen menos probabilidades de presentar síntomas que las no embarazadas con esta enfermedad, pero son más susceptibles de necesitar cuidados intensivos en caso de enfermedad grave.

En el Anexo A se muestra el total de variables obtenidas de la base de datos del Sistema Nacional de Vigilancia Epidemiológica (SINAVE), de la Dirección General de Epidemiología de

la Secretaría de Salud. Se observó que el total de variables recuperadas fue 90, las cuales son descritas en la Tabla A.1, indicando nombre, descripción, tipo de dato y los valores que éstas pueden tomar.

3.2. Análisis y preparación de datos

El análisis de datos es una de las actividades fundamentales en el proceso de la generación de grafos de conocimiento, mediante el cual se establece el contacto directo con el problema a resolver. El análisis de las fuentes de datos disponibles se realizó en dos etapas. La primera consistió en una revisión y análisis preliminar del total de variables listadas en la base de datos, esto con el fin de establecer aquellas características relevantes en función de sus registros. En la segunda etapa se hizo un análisis exploratorio de datos, previo al diseño y construcción del grafo de conocimiento de propiedades.

La base de datos obtenida considera todos aquellos casos sospechosos por COVID-19, que pueden o no tener la enfermedad. De la fuente de datos recuperada, al 24 de diciembre de 2020, se detectó un total de 935203 registros a nivel federal. A partir de esta fuente de datos se delimitó el objeto de estudio a las gestantes diagnosticadas con COVID-19 y que fueron atendidas en la Ciudad de México. Para esto se hizo una selección de registros bajo los siguientes criterios:

- A partir de la variable **resultado_definitivo**, que indica si el resultado de la prueba fue por laboratorio, se hizo el primer filtro de todos aquellos registros con valor SARS-CoV-2. Con esto se tomaron en cuenta solo los casos positivos por COVID-19.
- A partir de la variable **esta_embarazada** se hizo un segundo filtro, tomando los casos que se encontraban en estado de gestación al momento de ser diagnosticada con COVID-19.

Con base en lo anterior, del total de registros iniciales, 935203 a nivel federal, quedaron 1106 casos de pacientes embarazadas contagiadas de COVID-19 que recibieron atención en la Ciudad de México. Por lo que, en función del objetivo de esta tesis, se hizo el análisis de datos a partir de esta población seleccionada.

3.2.1. Variables significativas

Posterior a la selección de los registros, otro paso importante fue la selección de características. El cual es clave y se debe tener en cuenta para definir qué datos de entrada va a recibir el algoritmo y con qué variables se modelará el grafo de conocimiento. Esta selección de características fue un proceso que consistió en obtener un grupo de variables seleccionadas por su relevancia, discriminando aquellas que no guardan relación con el objeto de estudio y otras que no fueron necesarias. Como resultado del análisis, se observó que 48 de las 90 variables cumplieron con los criterios establecidos, las cuales se describen en la Tabla 3.1.

Tabla 3.1: Variables relevantes para la construcción del grafo de conocimiento.

Nombre de la variable	**Descripción**
sector	Designa si la USMI pertenece a alguna institución de salud del sector público o privado.
unidad_medica	Nombre de la unidad médica
entidad_residencia	Indica la entidad de residencia del paciente.
municipio_residencia	Indica el municipio de residencia del paciente.
localidad_residencia	Nombre de la localidad de residencia del paciente.
tipo_paciente	Indica el tipo de paciente, según si recibió atención médica ambulatoria (es decir, que recibió atención médica y de diagnóstico, pero no tuvo que pasar la noche en la unidad de salud), o si requirió hospitalización.
evolucion	Indica la evolución del paciente a la fecha de corte en la base de datos.
fecha_defuncion	Indica la fecha de defunción del paciente.
intubado	Indica si el paciente requirió intubación o no.
diagnostico_clinico_neumonia	Indica si cuando el paciente llegó a la unidad médica presentó un diagnóstico clínico de neumonía.
edad	Indica la edad del paciente.
meses_embarazo	Indica el número de meses de embarazo de la paciente.
ocupacion	Indica la ocupación del paciente.
servicio_ingreso	Indica el servicio al que ingresó el paciente para ser atendido.
fecha_inicio_sintomas	Indica la fecha en que el paciente comenzó con síntomas.
fiebre	Indica si el paciente presentó fiebre como síntoma.
tos	Indica si el paciente presentó tos como síntoma.
odinofagia	Indica si el paciente presentó odinofagía como síntoma.
disnea	Indica si el paciente presentó disnea (dificultad para respirar) como síntoma.
irritabilidad	Indica si el paciente presentó irritabilidad como síntoma.

Nombre de la variable	**Descripción**
diarrea	Indica si el paciente presentó diarrea como síntoma.
dolor_ toracico	Indica si el paciente presentó dolor torácico.
calofrios	Indica si el paciente presentó calofríos como síntoma.
cefalea	Indica si el paciente presentó cefalea como síntoma.
mialgias	Indica si el paciente presentó mialgias como síntoma.
artralgias	Indica si el paciente presentó artralgias.
ataque_ al_ estado _general	Indica si el paciente presentó un ataque al estado general.
rinorrea	Indica si el paciente presentó rinorrea como síntoma.
polipnea	Indica si el paciente presentó polipnea como síntoma.
vomito	Indica si el paciente presentó vómito como síntoma.
dolor_ abdominal	Indica si el paciente presentó dolor abdominal como síntoma.
conjuntivitis	Indica si el paciente presentó conjuntivitis como síntoma.
cianosis	Indica si el paciente presentó cianosis como síntoma.
inicio_ subito _sintomas	Indica los casos en los que los síntomas no fueron paulatinos, sino que iniciaron súbitamente.
diabetes	Indica si el paciente padecía diabetes al momento del diagnóstico.
epoc	Indica si el paciente padecía EPOC al momento del diagnóstico.
asma	Indica si el paciente padecía asma al momento del diagnóstico.
inmunosupresivo	Indica si el paciente es inmunosupresivo al momento del diagnóstico.
hipertension	Indica si el paciente padecía hipertensión al momento del diagnóstico.
VIH_ SIDA	Indica si el paciente padecía VIH/SIDA al momento del diagnóstico.

Nombre de la variable	**Descripción**
otra_ condicion	Indica si el paciente padecía alguna otra condición de salud (comorbilidad) al momento del diagnóstico.
enfermedad_ cardiaca	Indica si el paciente padecía alguna enfermedad cardiaca al momento del diagnóstico.
obesidad	Indica si el paciente padecía obesidad al momento del diagnóstico.
insuficiencia_ renal_ cronica	Indica si el paciente padecía insuficiencia renal crónica al momento del diagnóstico.
tabaquismo	Indica si el paciente padecía tabaquismo al momento del diagnóstico.
antiviral	Indica si el paciente tomó algún antiviral y cuál.
vacunado	Indica si el paciente recibió la vacuna contra la influenza en el último año.

Por otro lado, al aplicar las consideraciones anteriores, se discriminaron las siguientes 42 variables:

1. *origen*, debido a que no contiene información relevante para analizar las características de un paciente contagiado por COVID-19.
2. *cve_ entidad_ unidad_ medica*, debido a que todas las unidades dentro de la fuente de datos pertenecen a la Ciudad de México, este valor no cambia.
3. *entidad_ medica*, debido a que no cambia el valor al tratarse de la Ciudad de México.
4. *delegacion_ unidad_ medica*, debido al caso anterior y sus valores no son objeto de estudio.
5. *fecha_ registro*, es la fecha de registro en el sistema y no aporta información para el estudio.
6. *sexo*, debido a que la población en la que se enfoca el estudio corresponde al sexo femenino.
7. *cve_ entidad_ residencia*, debido a que proporciona información similar a entidad_residencia.
8. *clave_ municipio_ residencia*, proporciona información similar a municipio_residencia.
9. *clave_ localidad_ residencia*, debido a que proporciona información similar a 'localidad_residencia'.
10. *semana_ defuncion*, indica la semana de la fecha de defunción del paciente.
11. *nacionalidad*, debido a un número no significativo diferente a nacionalidad Mexicana.
12. *esta_ embarazada*, el valor no cambia debido a que el estudio se limita a mujeres embarazadas.
13. *es_ indigena*, debido a que un número no significativo es positivo.
14. *habla_ lengua_ indigena*, se presenta el mismo caso que la variable anterior.

15. *fecha_ ingreso*, hay otras variables, como fecha de inicio de síntomas, que son relevantes.
16. *diagnostico_ probable*, debido a que el objetivo son casos de embarazadas por COVID-19.
17. *recibio_ tratamiento*, la información que proporciona no es de interés para el estudio.
18. *recibio_ tratamiento_ antiviral*, debido a que hay otras variables, como *antiviral*, que proporcionan mayor información del antiviral que recibió la paciente.
19. *recibio_ tratamiento_ antibiotico*, debido al mismo caso de la variable anterior.
20. *fecha_ inicio_ antiviral*, debido a que la información no es de interés para el estudio.
21. *contacto_ infeccion _ viral*, la información que proporciona no es de interés para el estudio.
22. *contacto_ aves*, la información tampoco es de interés para el estudio.
23. *contacto_ cerdos*, la información que proporciona no es de interés para el estudio.
24. *contacto_ animales*, la información que proporciona no es de interés para el estudio.
25. *fecha_ estimada_ vacunación*, la información que proporciona no es de interés para el estudio.
26. *toma_ muestra*, la información que proporciona no es relevante para este estudio.
27. *laboratorio*, se presenta el mismo caso que la variable anterior.
28. *resultado_ definitivo*, al ser todos positivos a COVID-19 esta variable no aporta valor.
29. *es_ migrante*, debido a que aporta un número no significativo de casos.
30. *pais_ nacionalidad*, debido a que aporta un número no significativo de casos.
31. *pais_ origen*, se presenta el mismo caso que la variable anterior.
32. *fecha_ ingreso_ pais*, esta variable ya no aplica para el estudio.
33. *puerperio*, la información que proporciona no es de interés para el estudio.
34. *dias_ puerperio*, esta variable no aplica para el estudio.
35. *antipiréticos*, la información que proporciona no es de interés para el estudio.
36. *UCI*, existe otra variable que indica si el paciente es un caso grave.
37. *influenza_ tipo_ b*, la información que proporciona no es de interés para el estudio.
38. *viaje_ 1*, la información que proporciona no es de interés para el estudio.
39. *viaje_ 2*, la información que proporciona no es de interés para el estudio.
40. *viaje_ 3*, la información que proporciona no es de interés para el estudio.
41. *viaje_ 4*, la información que proporciona no es de interés para el estudio.
42. *viaje_ 5*, la información que proporciona no es de interés para el estudio.

3.2.2. Exploración de datos

Una buena práctica, antes de modelar el grafo de conocimiento, es hacer un análisis de éstos para resumir sus principales características. A esta práctica se conoce como análisis exploratorio de datos, que implica conocer los datos a través de métodos visuales. El propósito es tener una idea de la estructura del conjunto de datos, identificar variables objetivo y posibles técnicas de modelado.

En este sentido, se analizaron las variables seleccionadas en la etapa anterior. A modo de ejemplo se presentan en este apartado los resultados de algunas variables, como: a) sector, que hace referencia a alguna institución de salud del sector público o privado del país; b) enfermedades, que agrupa a un conjunto de condiciones que indica los padecimientos más frecuentes en las embarazadas con COVID-19; y c) síntomas, que agrupa una serie de síntomas que presentaron las pacientes objeto de estudio.

a) Sector

La Figura 3.2 muestra un resumen del conteo, valores únicos, el valor más frecuente y el número de apariciones de éste en el global de datos para la variable *sector*. Se observó que fueron ocho los valores únicos y el valor con mayor número de apariciones fue SSA (Secretaría de Salud) con un total de 726 veces.

	sector
count	1106
unique	8
top	SSA
freq	726

Fig. 3.2: Recuento estadístico para la variable *sector*.

A partir de lo anterior, la Figura 3.3 muestra un histograma sobre los sectores de salud en los que mayormente fueron atendidas las embarazadas con COVID-19. Se observó que en su mayoría fueron atendidas en el sector público a través de la Secretaría de Salud (SSA), seguido del Instituto Mexicano del Seguro Social (IMSS) y en menor medida a través de la Secretaría de Marina (SEMAR), Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE), Petróleos Mexicanos (PEMEX) y Secretaría de Medio Ambiente (SEDEMA). Existe también una porción menor de personas que fueron atendidas por el sector privado.

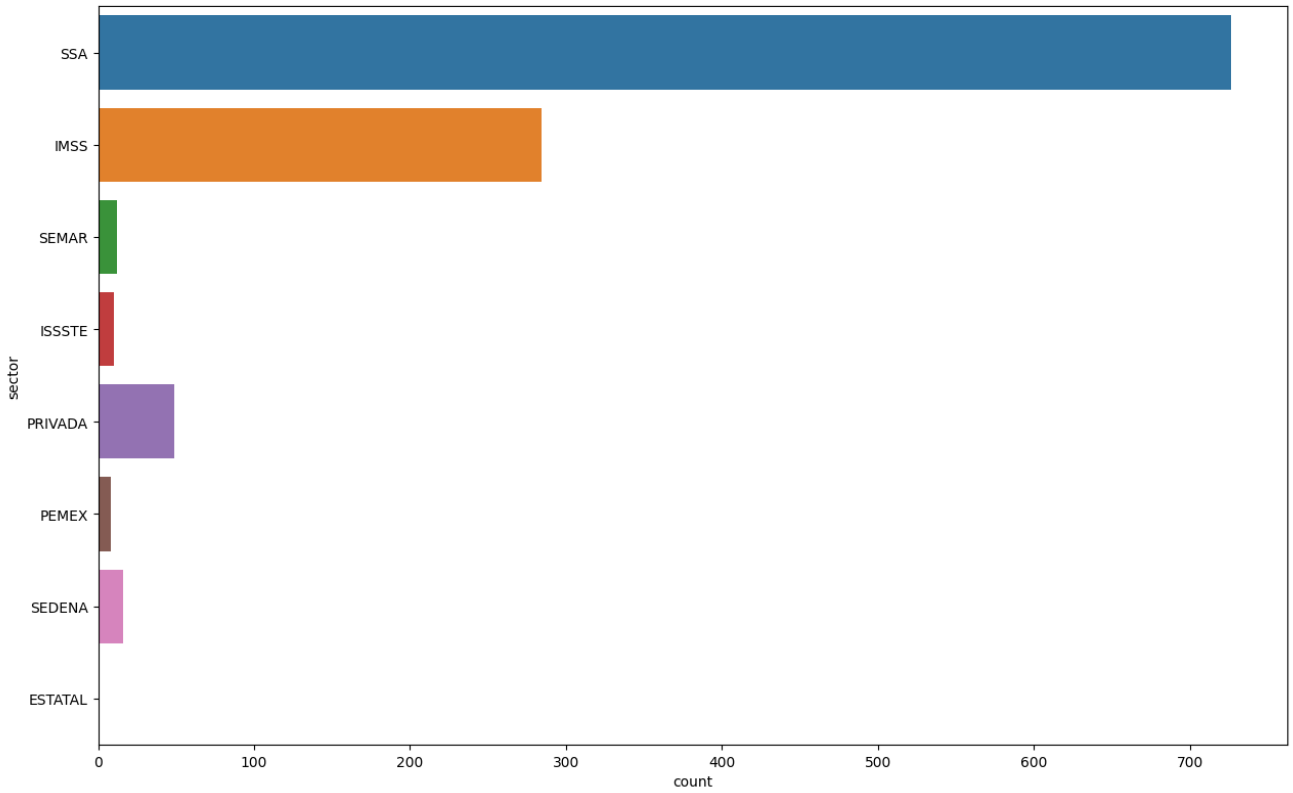

Fig. 3.3: Sectores hospitalarios en los que fueron atendidas las embarazadas con COVID-19.

b) Enfermedades

Otra de las condiciones analizadas fueron las enfermedades con las cuales las gestantes contagiadas de COVID-19 fueron diagnosticadas. Se observó, a través de la Figura 3.4, que la obesidad fue la enfermedad que mayormente presentaron las pacientes, seguido de la diabetes y tabaquismo. Se observó también otro grupo de enfermedades que en menor medida afectaron la condición de las pacientes, como: asma, hipertensión, inmunosupresivo, VIH y la enfermedad pulmonar obstructiva crónica (epoc). Esta información puede ser potencialmente útil para encontrar dentro del grafo patrones que relacionen estas enfermedades con el estado o evolución de las pacientes.

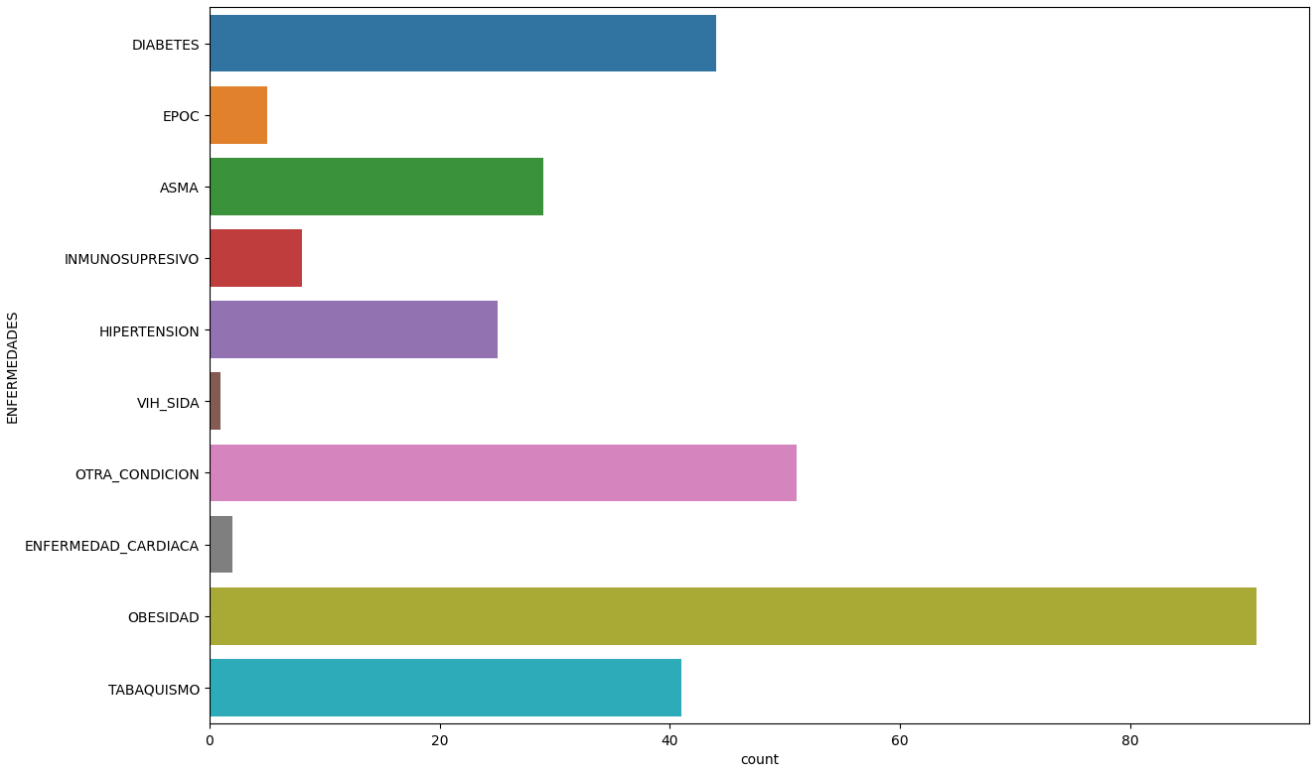

Fig. 3.4: Enfermedades diagnosticadas en embarazadas con COVID-19.

c) Síntomas

Por otra parte, los síntomas fue otra de las condiciones analizadas. De la cual se observó, a través de la Figura 3.5, que la tos y cefalea fueron los síntomas más frecuentes. Otros padecimientos, o la combinación de estos, fueron fiebre, odinofagia, disnea, irritabilidad, diarrea, dolor torácico, calofríos, mialgias, artralgias, ataque al estado general, rinorrea, polipnea, vómito, dolor abdominal, conjuntivitis y otros. Estos datos también son útiles para relacionar los síntomas y las diferentes enfermedades con las que fueron diagnosticadas, su evolución y posible identificación de patrones a través del grafo de conocimiento.

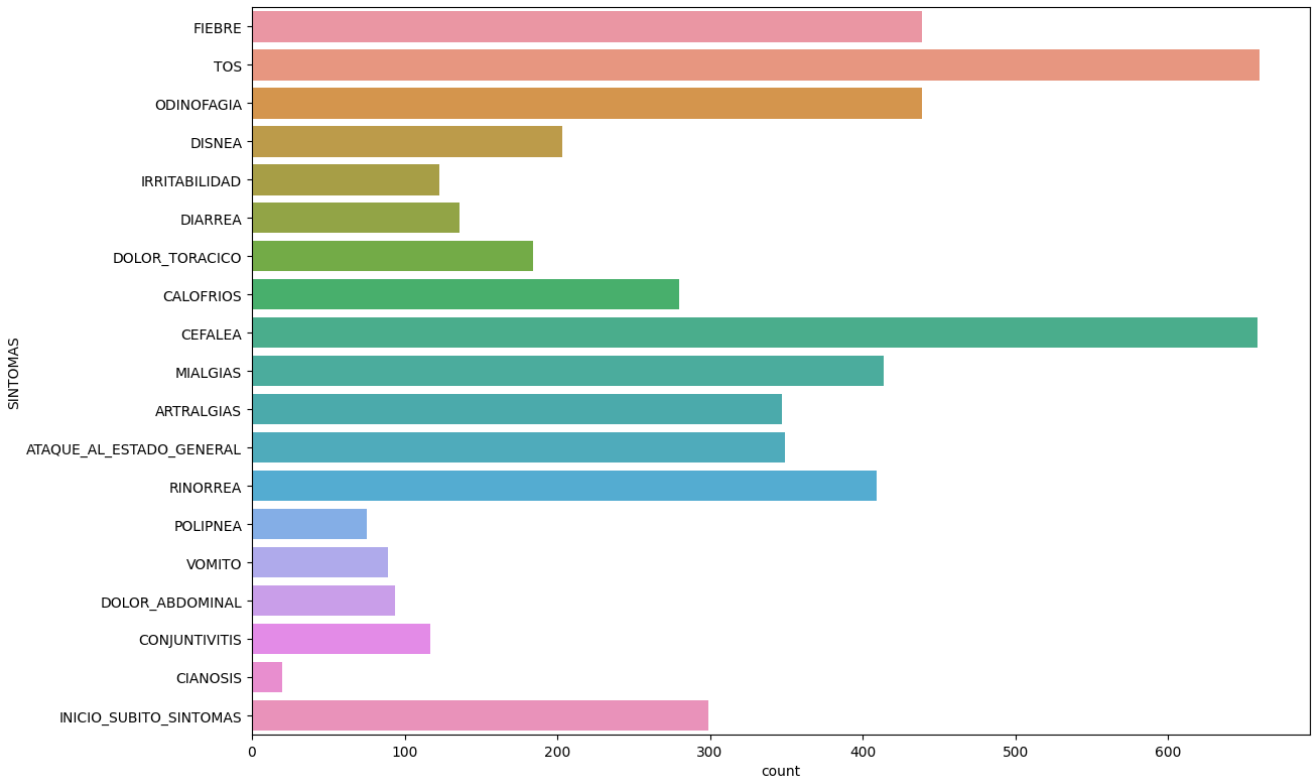

Fig. 3.5: Histograma con los síntomas de las embarazadas con COVID-19.

3.3. Modelado del grafo de conocimiento

Con base en lo descrito en el capítulo anterior, sección 2.3, para la construcción de un grafo de conocimiento se hace una abstracción de los datos con el propósito de modelar el grafo. Para esto existen modelos que permiten representar la información de manera estructurada, entre los que destacan: i) grafo dirigido con aristas etiquetadas, y ii) grafo de propiedades.

Para este trabajo de tesis, como se mencionó con anterioridad, para el modelado del grafo de conocimiento se empleó un grafo de propiedades, el cual está formado por nodos, relaciones y propiedades. En este tipo de grafos, las relaciones deben tener una etiqueta (nombre) y una dirección. Además, se utilizó la tecnología Neo4j, la cual es de código abierto, como enfoque de base de datos orientados a grafos. Entre sus propiedades básicas para el almacenamiento persistente de datos destacan: crear (*create*), leer (*read*), actualizar (*update*) y eliminar (*delete*). Otra característica es que a través de Neo4j se puede evitar tener información aislada, dado que todos los datos son conectados de manera eficiente en forma de grafos y no en tablas.

De esta forma, para el modelado de datos en el grafo, se emplearon las variables seleccionadas en la sección anterior, compuesto de nodos conectados mediante relaciones con propiedades y etiquetas.

3.3.1. Nodos

Se utilizaron nodos para representar las entidades conceptuales únicas, las cuales corresponden a grupos de variables relevantes asociadas con la enfermedad COVID-19 en embarazadas atendidas en hospitales de la Ciudad de México:

1. *Paciente*. Los registros de datos de un grupo de variables fueron utilizadas para estructurar información sobre el paciente y COVID-19, como: a) identificador, b) edad, c) vacunado, d) ocupación, e) estatus, y f) meses de embarazo.
2. *SectorSalud*. Este nodo fue utilizado para indicar en qué sector del área de salud se atendieron las pacientes. Se vincula con la variable *sector*.
3. *UnidadMédica*. Nodo designado para identificar el nombre de los hospitales en la Ciudad de México, donde fueron atendidas las pacientes embarazadas con COVID-19. Se vincula con la variable *unidad_ medica*.
4. *EntidadResidencia*. Nodo utilizado para hacer referencia a las entidades de residencia de las pacientes atendidas en la Ciudad de México. Se vincula con la variable *entidad_ residencia*.
5. *MunicipioResidencia*. Nodo utilizado para hacer referencia a los municipios de residencia de las pacientes que fueron atendidas en la Ciudad de México. Se vincula con la variable *municipio_ residencia*.
6. *LocalidadResidencia*. Nodo utilizado para hacer referencia a las localidades de residencia de las pacientes que fueron atendidas en la Ciudad de México. Se vincula con la variable *municipio_ residencia*.
7. *Enfermedad*. Nodo establecido para hacer referencia a otros padecimientos (enfermedades) que presentaron las pacientes al momento de ser atendidas por COVID-19. Se vincula con un grupo de variables, como: *diabetes, asma, hipertensión, VIH_ SIDA, enfermedad_ cardiaca, obesidad, insuficiencia_ renal_ cronica, tabaquismo, entre otros*.
8. *FechaSíntomas*. Nodo designado para hacer referencia a la fecha en la que las pacientes comenzaron a presentar los síntomas de COVID-19. Se vincula con la variable *fecha_ inicio_ sintomas*.
9. *FechaDefunción*. Nodo designado para hacer referencia a la fecha en la que las pacientes desafortunadamente fallecieron. Se vincula con la variable *fecha_ defuncion*.
10. *Síntomas*. Nodo que agrupa una serie de variables que proveen información sobre los síntomas de COVID-19 que presentaron las pacientes. Se vincula con un grupo de variables, como: *fiebre, tos, odinofagia, disnea, irritabilidad, diarrea, dolor_ toracico, calofrios, cefalea, vomito, conjuntivitis, entre otros*.

3.3.2. Relaciones

Se definieron relaciones para representar la interacción de diferentes entidades y conjunto de entidades. Las relaciones definidas fueron:

1. *ATENDIDO_EN*. Esta relación, *Paciente* → *UnidadMédica*, indica en qué hospital fue atendida una paciente. Tiene las siguientes propiedades:

   a) *TipoAtención*. Indica si la paciente fue atendida de forma ambulatoria o tuvo que ser hospitalizada.

   b) *FechaIngreso*. Indica cuándo inició la atención de la paciente en la unidad médica.

   c) *ServicioIngreso*. Indica el servicio que se le dio a la paciente al ingresar al hospital.

   d) *UCI*. Indica si el paciente necesitó de cuidados intensivos.

   e) *Intubado*. Indica si la paciente fue intubada o no.

   f) *Tratamiento*. Indica si la paciente recibió o no prescripción médica.

   g) *Antiviral*. Indica, si es el caso, qué antiviral se le prescribió a la paciente.

2. *VIVIO_EN*. Esta relación, *Paciente* → *Localidad*, indica el lugar de residencia de la paciente.

3. *DIAGNOSTICADO_CON*. Esta relación, *Paciente* → *Enfermedad*, indica qué otras enfermedades sufría la paciente al momento de ser diagnosticada con COVID-19.

4. *PRESENTO*. Esta relación, *Paciente* → *Síntomas*, indica cuales fueron los síntomas de COVID-19 que presentaron las pacientes.

5. *PERTENECE_A*. Esta relación, *Localidad* → *Municipio y Municipio* → *Estado*, hace referencia a qué municipio o entidad federativa pertenece la localidad de residencia de la paciente.

6. *FALLECIO_EL*. Esta relación, *Paciente* → *FechaDefunción*, indica la fecha de fallecimiento de la paciente.

7. *INICIO_EL*. Esta relación, *Paciente* → *FechaSíntomas*, indica la fecha en la que la paciente empezó a presentar los síntomas de COVID-19.

A partir de los nodos y relaciones mencionados, se estructuró, como parte del diseño, el modelo de datos en forma de grafo, el cual se muestra en la Figura 3.6.

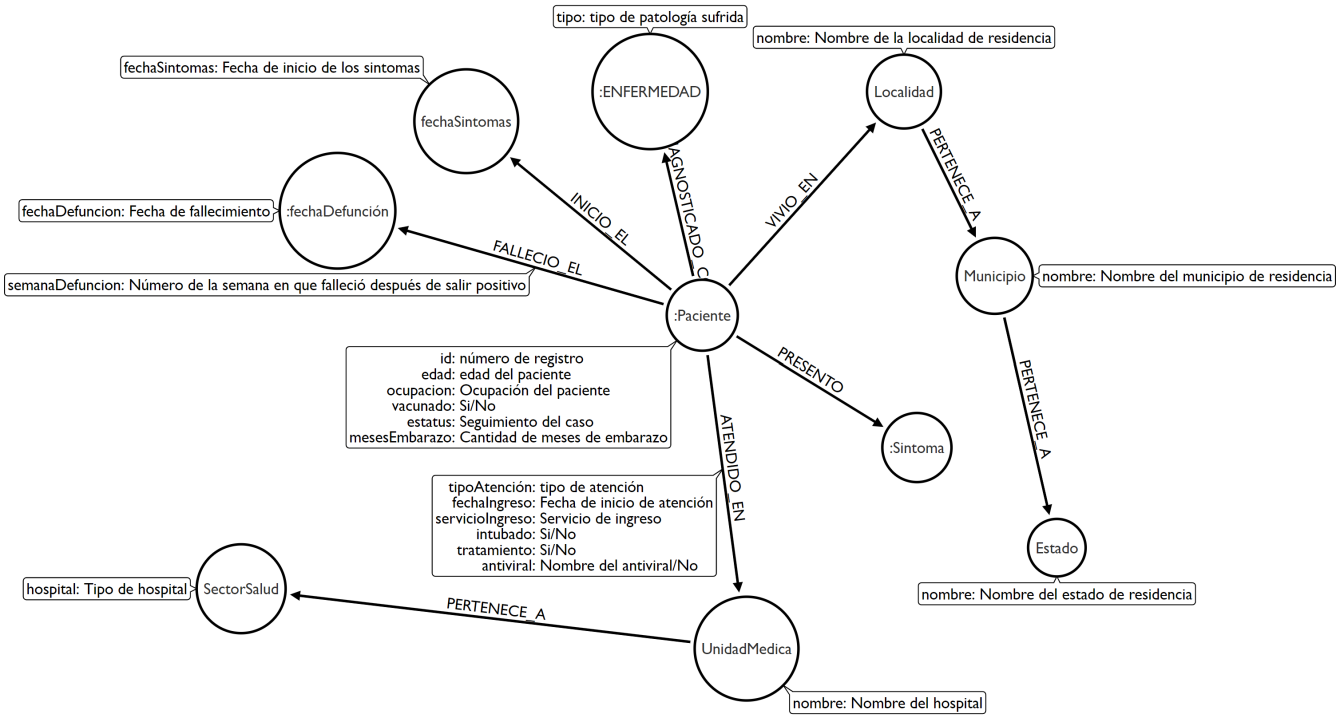

Fig. 3.6: Modelo de datos orientado a grafos.

3.4. Creación del grafo en Neo4j

Con base en el diseño del modelo mostrado en la sección anterior, se hizo la construcción del grafo en Neo4j. Para esta construcción se definieron las siguientes acciones:

1. Los datos seleccionados en la etapa de análisis y preparación de datos fueron almacenados en un archivo con extensión CSV (comma-separated values), para la lectura de los datos a través del lenguaje de consultas Cypher.

2. Posteriormente, se crearon los nodos de referencia a partir de las variables (columnas) descritas en la sección anterior. Para lograr esto se desarrolló y ejecutó el siguiente script:

```
LOAD CSV FROM 'file:///paciente.csv' AS line
with line[0] as id, line[12] as edad, line[14] as ocup, line[13] as emb,
    line[51] as vac
CREATE(P:Paciente{id:toInteger(id), edad:toInteger(edad), ocupacion:ocup,
    embarazada:emb,vacunado:vac})
```

3. A través del script se ejecutó la carga del archivo y se tomaron las columnas necesarias, por ejemplo, con *line[12] as edad* se tomó la variable que concentra la información sobre la edad de las pacientes. La base carga el archivo y línea por línea ejecuta *CREATE* para crear un nodo dentro de la base de datos con la etiqueta *Paciente*, y dentro de las llaves se incluyeron todas las propiedades para ese nodo. En la Figura 3.7 se muestra, a modo de ejemplo, el nodo creado (*Paciente*).

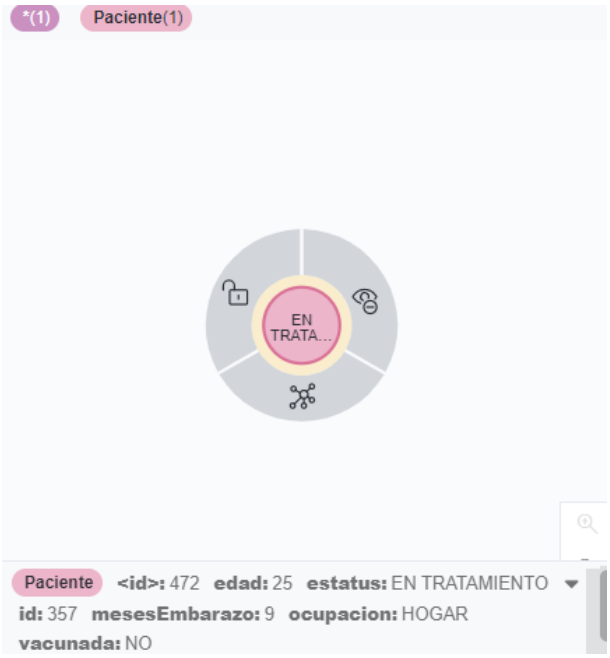

Fig. 3.7: Nodo Paciente.

4. Para la creación de los otros nodos restantes, como: *Localidad, Municipio, Entidad, Síntomas, UnidadMédica, FechaSíntomas, FechaDefunción y SectorSalud*, se utilizó un segundo script para recorrer el archivo línea por linea y ejecutar la instrucción condicional *FOREACH* para cada nodo.

```
1
2  LOAD CSV FROM 'file:///diccionario.csv' AS row
3  with row[0] as id, row[1] as nom
4  FOREACH ( ignoreMe in CASE WHEN id="sector" THEN [1] ELSE [] END | MERGE(S:
       SectorSalud{nombre:nom}))
5  FOREACH ( ignoreMe in CASE WHEN id="Hosp" THEN [1] ELSE [] END | MERGE(U:
       UnidadMedica{nombre:nom}))
6  FOREACH ( ignoreMe in CASE WHEN id="Est" THEN [1] ELSE [] END | MERGE(E:
       Estado{nombre:nom}))
7  FOREACH ( ignoreMe in CASE WHEN id="Mun" THEN [1] ELSE [] END | MERGE(M:
       Municipio{nombre:nom}))
8  FOREACH ( ignoreMe in CASE WHEN id="Loc" THEN [1] ELSE [] END | MERGE(L:
       Localidad{nombre:nom}))
9  FOREACH ( ignoreMe in CASE WHEN id="def" THEN [1] ELSE [] END | MERGE(fd:
       FechaDefuncion{fecha:nom}))
10 FOREACH ( ignoreMe in CASE WHEN id="sin" THEN [1] ELSE [] END | MERGE(fs:
       FechaSintomas{fecha:nom}))
11 FOREACH ( ignoreMe in CASE WHEN id="sint" THEN [1] ELSE [] END | MERGE(Si:
       Sintoma{tipo:nom}))
12 FOREACH ( ignoreMe in CASE WHEN id="enf" THEN [1] ELSE [] END | MERGE(en:
       Enfermedad{tipo:nom}))
13 RETURN *
```

5. A través de la instrucción MERGE, se busca el nodo para verificar su existencia. Si existe, recupera el nodo; sino crea el nodo con las propiedades especificadas. Con base a lo anterior,

se crearon todos los nodos dentro de la base de datos orientada a grafos, pero aún sin relaciones.

6. Para crear las relaciones dentro de la base de datos orientada a grafos, *FALLECIO_EL* (entre Paciente y FechaDefunción), *INICIO_EL* (entre Paciente y FechaSíntomas), *VIVIO_EN* (entre Paciente y Localidad), *PERTENECE_A* (entre Localidad y Municipio), *PERTENECE_A* (entre Municipio y Entidad), se utilizó un script en las instrucciones MATCH y MERGE, con las cuales se definieron dichas relaciones. En el Anexo B se muestra el código en extenso utilizado para la creación de las relaciones restantes.

```
1
2  LOAD CSV FROM 'file:///paciente.csv' AS line
3  with line[0] as id, line[8] as def, line[9] as semdef, line[17] as inicio,
       line[5] as loc, line[4] as mun,line[3] as est
4  MATCH(P:Paciente{id:id})
5  MATCH(F:FechaDefuncion{fecha:def})
6  MATCH(FS:FechaSintomas{fecha:inicio})
7  MATCH(L:Localidad{nombre:loc})
8  MATCH(M:Municipio{nombre:mun})
9  MATCH(E:Estado{nombre:est})
10 MERGE (P)-[R:FALLECIO_EL{semana:semdef}]->(F)
11 MERGE (P)-[R:INICIO_EL]->(FS)
12 MERGE (P)-[R:VIVIO_EN]->(L)
13 MERGE (L)-[R:PERTENECE_A]->(M)
14 MERGE (M)-[R:PERTENECE_A]->(E)
```

A manera de ejemplo, la Figura 3.8 muestra el nodo resultante de la construcción del nodo Entidad, específicamente para la Ciudad de México (Rojo), que a su vez tiene relación con el nodo *Municipio* (Café), y este tiene relación con *PERTENECE_A*, proveniente del nodo *Localidad* (Verde). Para lograr esta visualización se realizó la siguiente consulta:

```
1 MATCH (n:Estado) <-[r]- (m:Municipio) <-[r1]- (l:Localidad) WHERE n.nombre = "
      CIUDAD DE MEXICO" return n,m,l
```

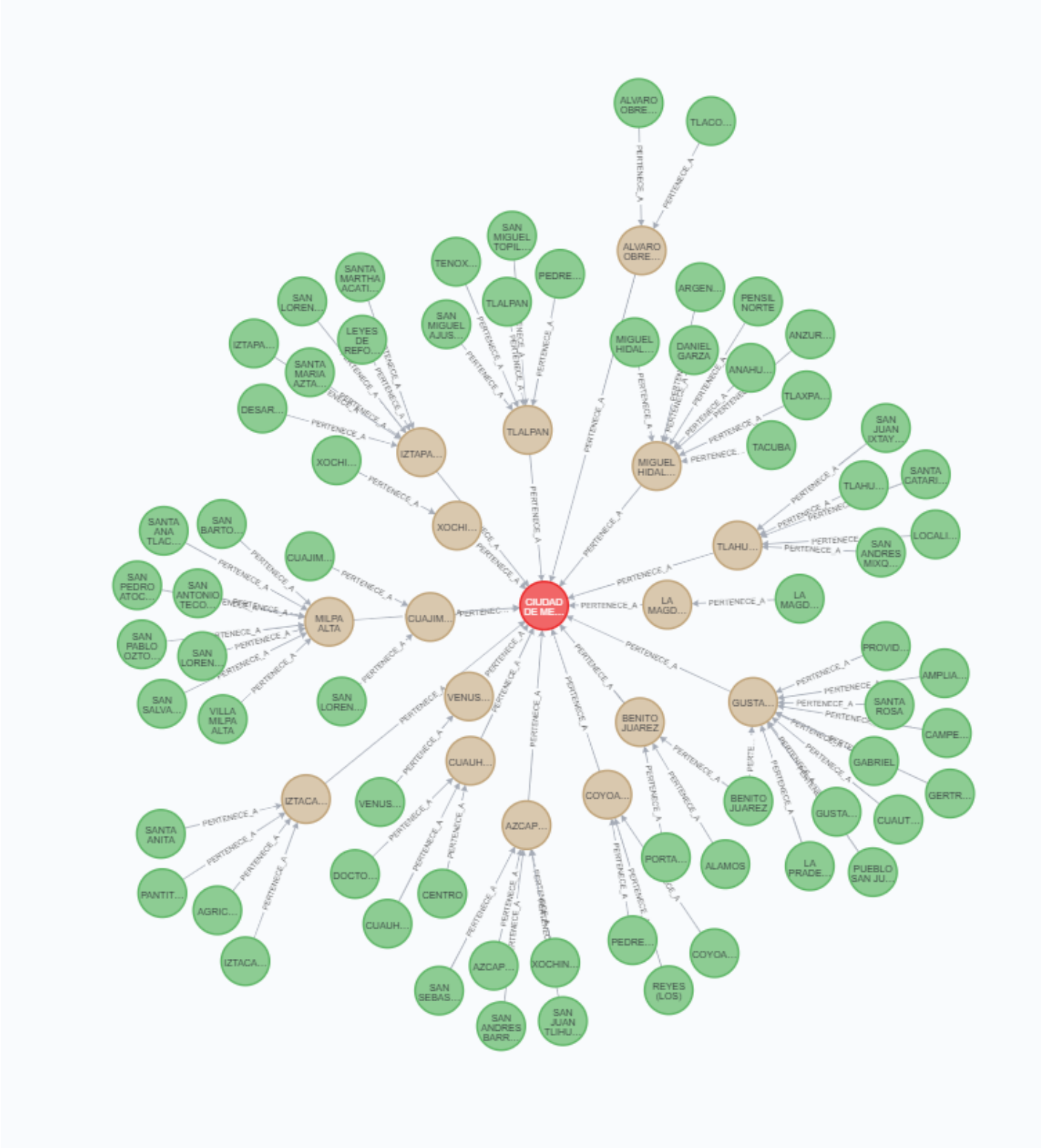

Fig. 3.8: Visualización del nodo Entidad (Ciudad de México) y sus relaciones con Municipio (Café) y Localidad (Verde).

También a manera de ejemplo, la Figura 3.9 muestra la construcción del nodo IMSS (Naranja) que tiene como etiqueta *SectorSalud*, y que a su vez tiene relación *PERTENECE_A* con del nodo *UnidadMédica*. Para lograr esta visualización se utilizó la siguiente consulta:

```
MATCH (n:SectorSalud) <-[r1]- (u:UnidadMedica) WHERE n.nombre = "IMSS" return n, u
```

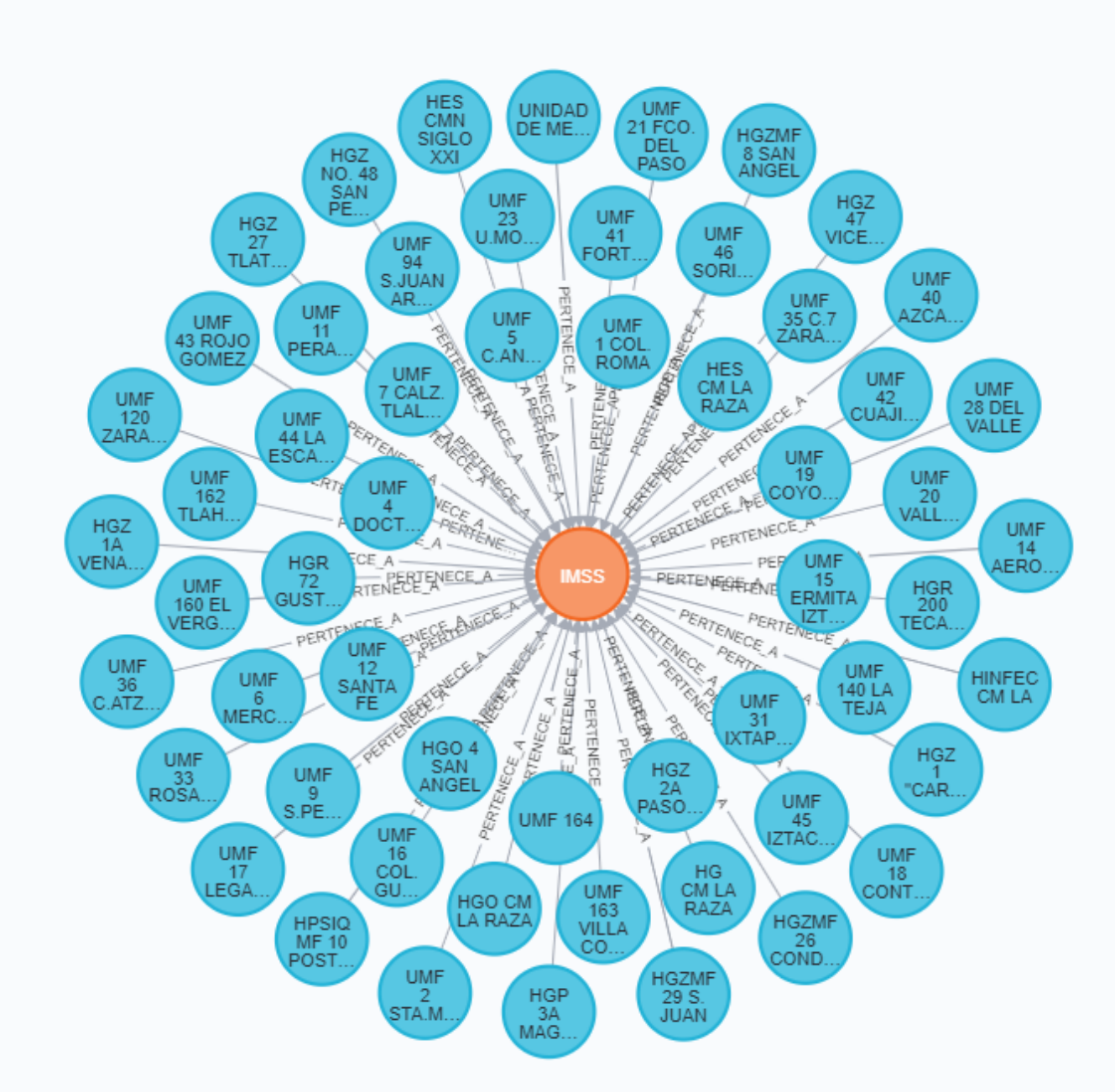

Fig. 3.9: Visualización del nodo SectorSalud (IMSS) con sus unidades de salud asociadas.

Por otro lado, la Figura 3.10 muestra el nodo Paciente y sus relaciones con otros nodos,

como ASMA siguiendo la relación *DIAGNOSTICADO_ CON*. En este caso la paciente inició con síntomas de COVID-19 el '13/11/2020' y fue atendida en el Hospital *C.S.T-III DR. JOSE ZOZAYA*, el cual pertenece al sector salud *(SSA)*. Otros datos de interés son: la paciente tiene 25 años, con 9 meses de embarazo, vive en la localidad Iztacalco, Municipio de Iztacalco, Ciudad de México; tiene como ocupación *HOGAR*, no fue vacunada y está en tratamiento. Además, la paciente ingresó mediante consulta externa y tuvo un tipo de atención ambulatoria. Para lograr esta visualización se utilizó la siguiente consulta:

```
MATCH (p:Paciente)
WHERE p.id="357"
WITH p
MATCH (p)-[r1:VIVIO_EN]->(l:Localidad)-[r2]->(m:Municipio)-[r3]->(e:Estado)
MATCH (p)-[r4]->(u:UnidadMedica)-[r5]->(s:SectorSalud)
MATCH (p)-[r]->(b)
RETURN *
```

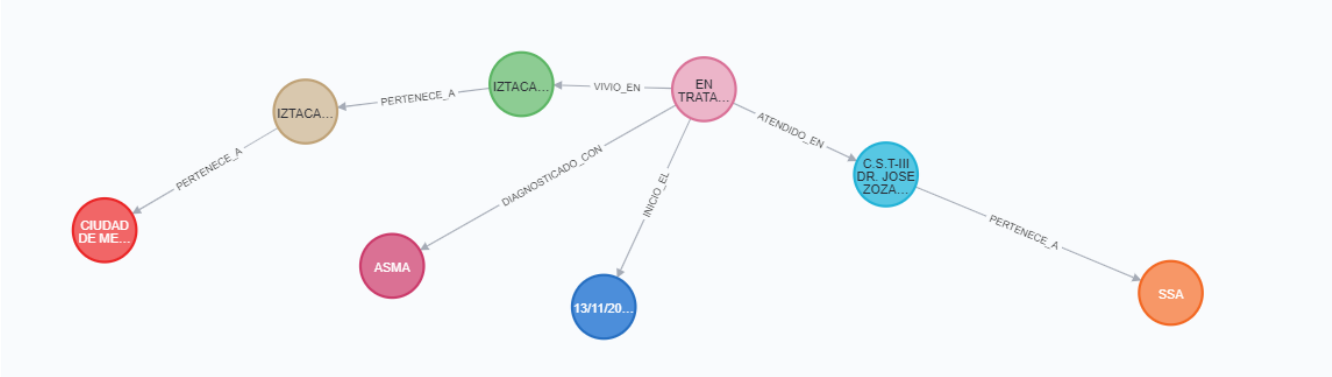

Fig. 3.10: Visualización del nodo Paciente y sus relaciones con los demás nodos.

El paso posterior es la discusión de resultados, los cuales se presentan en el capítulo siguiente. Estos resultados son con base en un análisis descriptivo de los patrones más significativos que se pueden descubrir a través del grafo de conocimiento, como: i) comorbilidad asociada con COVID-19, ii) tiempo promedio entre el inicio de los síntomas y el fallecimiento por esta enfermedad, iii) sectores de salud con más casos atendidos y más fallecimientos, iv) relación de la entidad de residencia y la atención recibida en la Ciudad de México, v) periodo con más fallecimientos, vi) síntomas en cada municipio, vii) enfermedades en cada localidad, y viii) casos graves.

3.5. Síntesis

El grafo de conocimiento formado fue el resultado de un amplio proceso de análisis. La primera fase, y una de las más importantes, fue la abstracción de la información disponible y modelarla, formando así las entidades que fueron representadas por medio de nodos y etiquetas. Posteriormente, se identificaron la forma en que estas interactúan, y así, obtener las relaciones que conforman el grafo. Una vez definido el modelo lógico, se conformó el grafo de conocimiento y se realizó la construcción de éste a través del motor de bases de datos orientado a grafos Neo4j. Para lograr lo anterior, fue necesario trabajar con Cypher, nombre del lenguaje de consultas que utiliza Neo4j,

para la creación y manipulación de las bases de datos. Finalmente, el resultado fue la creación de un grafo de conocimiento conformado por 1781 nodos y 9339 relaciones que representan toda la complejidad de los datos elegidos para el presente estudio, de los cuales se obtuvo información en forma de patrones de datos, que se describe en el siguiente capítulo.

Capítulo 4

Resultados

Como se mostró en el capítulo anterior, se realizó la construcción del grafo de conocimiento de propiedades para el análisis de la mortalidad por COVID–19 en embarazadas atendidas en hospitales de la Ciudad de México. Como parte del método utilizado se definieron cuatro etapas de trabajo: a) adquisición de datos, b) análisis y preparación de datos, c) modelado del grafo de conocimiento, y d) creación del grafo en Neo4j.

Con el objetivo de evaluar la funcionalidad del grafo de conocimiento obtenido, en este capítulo se presentan los resultados relacionados con el caso de estudio, esto es, análisis de la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México, cuyo periodo de evaluación comprende el 2020, hasta el 24 de diciembre del año en mención, fecha de corte del análisis. Los datos analizados fueron todos aquellos casos positivos de SARS-CoV-2.

4.1. Estructura del grafo de conocimiento

El grafo principal quedó conformado por 1761 nodos (Tabla 4.1) y 9209 relaciones (Tabla 4.2). Este grafo tiene un tamaño 6.3 Megabytes en disco y quedó estructurado por estos nodos y relaciones, tal como se muestra en la Figura 4.1.

Tabla 4.1: Total de nodos en el grafo de conocimiento.

Tipo de nodo	**Número de nodos**
FechaSintomas	226
UnidadMedica	220
Sintoma	19
Paciente	1106
Municipio	46
Estado	10
FechaDefuncion	11
SectorSalud	8
Enfermedad	11
Localidad	104

Tabla 4.2: Total de relaciones en el grafo de conocimiento.

Nombre de la relación	Total de relaciones
INICIO_EL	1992
ATENDIDO_EN	2212
PERTENECE_A	754
PRESENTO	10632
VIVIO_EN	2212
DIAGNOSTICADO_CON	594
FALLECIO_EL	22

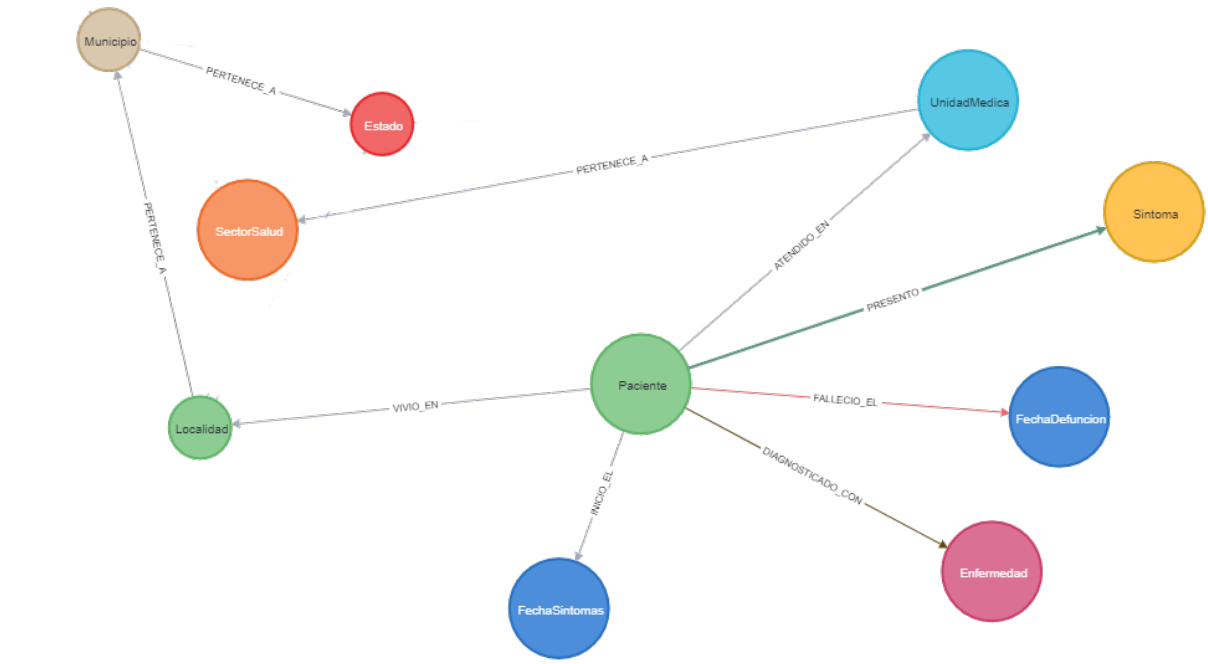

Fig. 4.1: Visualización del modelo de grafo de conocimiento.

4.2. Gestantes diagnosticadas con COVID-19

En total fueron 1106 casos positivos de embarazadas diagnosticadas con SARS-CoV-2, atendidas en la Ciudad de México. A partir de esta población, se identificaron algunos patrones de interés, derivados de las propiedades que fueron definidas en los nodos del grafo de conocimiento principal, por ejemplo, *Paciente*. Para esto se utilizó la consulta mostrada en la Lista 4.1.

```
-- edad
MATCH (n:Paciente)
    WHERE toInteger(n.edad) > -1 AND toInteger(n.edad) < 10
    return "0 a 9" as edad , count(n) as frecuencia
UNION
```

```
MATCH (n:Paciente)
    WHERE toInteger(n.edad) > 9 AND toInteger(n.edad) < 20
    return "10 a 19" as edad , count(n) as frecuencia
UNION
MATCH (n:Paciente)
    WHERE toInteger(n.edad) > 19 AND toInteger(n.edad) < 30
    return "20 a 29" as edad , count(n) as frecuencia
UNION
MATCH (n:Paciente)
    WHERE toInteger(n.edad) > 29 AND toInteger(n.edad) < 40
    return "30 a 39" as edad , count(n) as frecuencia
UNION
MATCH (n:Paciente)
    WHERE toInteger(n.edad) > 39 AND toInteger(n.edad) < 50
    return "40 a 49" as edad , count(n) as frecuencia
UNION
MATCH (n:Paciente)
    WHERE toInteger(n.edad) > 49 AND toInteger(n.edad) < 100
    return "50 o más" as edad , count(n) as frecuencia;
-- meses de embarazo
MATCH (n:Paciente)
    return toInteger(n.mesesEmbarazo) as Mes,
    count(n.mesesEmbarazo) ORDER BY Mes
-- ocupacion
MATCH (n:Paciente)
    return n.ocupacion, count(n.ocupacion) as ocupacion ORDER BY ocupacion DESC
-- Vacunadas
MATCH (n:Paciente)
    return n.vacunada, count(n.vacunada)
--Estatus
MATCH (p:Paciente) return  p.estatus, COUNT(p.estatus) as frecuencia ORDER BY
   frecuencia DESC
```

Listing 4.1: Consulta a partir del nodo Paciente.

A partir del script se obtuvieron patrones de interés, derivados del nodo *Paciente*, los cuales fueron agrupados en las siguientes tablas: a) acumulación de casos por rango de edades (Tabla 4.3); b) número de casos por meses de embarazo en las mujeres diagnosticadas con COVID-19 (Tabla 4.4); c) ocupación de las mujeres embarazadas con COVID-19 (Tabla 4.5); d) número de casos de vacunación contra la influenza de gestantes con COVID-19 (Tabla 4.6); y e) situación de las mujeres embarazadas con COVID-19 (Tabla 4.7).

Tabla 4.3: Acumulación de casos por rango de edades.

Edad	**Casos**
10 a 19	78
20 a 29	529
30 a 39	431
40 a 49	63
50 o más	5

Tabla 4.4: Casos por meses de embarazo en mujeres diagnosticadas con COVID-19.

Mes de embarazo	**Casos**
0	2
1	48
2	94
3	85
4	96
5	118
6	103
7	149
8	209
9	162
10	40

Tabla 4.5: Ocupación de las mujeres embarazadas con COVID-19.

Ocupación	**Casos**
Hogar	511
Empleadas	258
Estudiantes	51
Enfermeras	37
Comerciantes	31
Médicos	16
Maestras	14
Desempleadas	10
Dentistas	4
Obreras	3
Laboratoristas	2
Otras trabajadoras de la salud	17
Otras profesionistas	30
Otros	122

Tabla 4.6: Casos de vacunación contra la influenza.

Vacuna contra la influenza	**Casos**
No	803
Si	299
Sin información	4

Tabla 4.7: Estatus de mujeres embarazadas con COVID-19

Situación	**Casos**
Seguimiento terminado	414
En tratamiento	288
Seguimiento domiciliario	233
Alta - Mejoría	97
Caso no grave	20
Caso grave	17
Defunción	11
Referencia	9
Alta - Voluntaria	8
Caso grave - Traslado	5
Alta - Traslado	2
Alta - Curación	2

Con respecto al *grupo de edades*, se observó que la mayoría de las mujeres embarazadas y diagnosticadas con COVID-19, se presentaron en los rangos de edades de 20 a 29 años (529 casos) y de 30 a 39 años (431 casos). Existe también un grupo significativo de casos positivos en el rango de edad de 10 a 19 años (78 casos). Lo inusual se presenta en los casos diagnosticados para el grupo de 50 o más años, debido a que se tienen pacientes con edades de 59, 65, 66, 71 y 75 años, lo que podría deberse a valores atípicos por un error al momento registrar a estas personas.

Para el caso del *mes de embarazo* en el que las pacientes se contagiaron de COVID-19, se observó que la mayor cantidad de casos positivos se presentaron a partir del quinto mes del embarazo, con 118, 103, 149, 209 y 162, respectivamente. Mientras que en los primeros meses de embarazo, el número de contagios fue menor, en el orden de menos de 100 casos. De estas pacientes diagnosticadas con el virus SARS-CoV-2, en su mayoría, casi la mitad de los casos, tienen como *ocupación* las labores del hogar (511). Otro grupo importante son las que tienen un empleo (258), mientras que en otras actividades se registraron 122 casos. Estos niveles de contagio pudieron ser variados, esto es, ya sea a través de un familiar, los desplazamientos por ir a trabajar, o simplemente por las rutinas del día a día para la realización de compras, trámites, revisiones y demás.

De los casos analizados, 803 no fueron *vacunadas* contra la influenza (72.6 %), mientras que 299 si se vacunaron y existen otros 4 casos de los que se desconoce está información. Otro aspecto que llama la atención es que, a pesar de la alta cantidad de contagio, la mayoría de los casos se *recuperaron* de la enfermedad (414) de manera favorable, que corresponde al 40.32 %, otros 288 casos han seguido en tratamiento y otros 233 han tenido seguimiento domiciliario. Lamentablemente, existieron también 11 decesos y persistieron algunos casos graves (17) y traslados de casos graves (5). Por lo que, a pesar de la recuperación de gran parte de las mujeres afectadas por COVID-19, éstas sigue siendo una población vulnerable debido a la mortalidad de algunos casos identificados. Además, las embarazadas que tienen enfermedades subyacentes, como diabetes, hipertensión y otras, pueden tener un alto riesgo debido a COVID-19. Esto es importante, por el impacto que tiene el cuidado prenatal y de las gestantes en general.

4.3. Situación de la mortalidad por COVID-19

Como se mencionó en la sección anterior, del total de casos analizados, desafortunadamente 11 mujeres embarazadas fallecieron, quienes fueron atendidas en hospitales de la Ciudad de México. A través de consultas en el grafo de conocimiento, mostradas en las Listas 4.2 y 4.3, se obtuvieron detalles de los decesos, organizados a través de tablas, como: a) edad de las mujeres fallecidas (Tabla 4.8); b) meses de embazado al momento del deceso (Tabla 4.9); c) ocupación (Tabla 4.10); d) vacunación contra la influenza (Tabla 4.11); y e) periodo del deceso desde iniciado los síntomas (Tabla 4.12). Para esto se utilizaron los nodos *Paciente*, *FechaSintomas* y *FechaDefuncion*, relacionados a través de *FALLECIDO_EL*.

```
-- edad
MATCH (f:FechaDefuncion)<-[r2:FALLECIO_EL]-(n:Paciente)
    return n.edad, count(n.edad) ORDER BY n.edad
-- meses de embarazo
MATCH (f:FechaDefuncion)<-[r2:FALLECIO_EL]-(n:Paciente)
    return n.mesesEmbarazo, count(n.mesesEmbarazo) ORDER BY n.mesesEmbarazo
-- ocupacion
MATCH (f:FechaDefuncion)<-[r2:FALLECIO_EL]-(n:Paciente)
    return n.ocupacion, count(n.ocupacion)
-- Vacunadas
MATCH (f:FechaDefuncion)<-[r2:FALLECIO_EL]-(n:Paciente)
    return n.vacunada, count(n.vacunada)
```

Listing 4.2: Particularidades de los decesos por COVID-19.

```
MATCH (f:FechaDefuncion)<-[r2]-(n:Paciente)-[r1:INICIO_EL]->(s:FechaSintomas)
WITH apoc.date.parse(s.fecha, "ms", "dd/MM/yyyy") AS ms1,
    apoc.date.parse(f.fecha, "ms", "dd/MM/yyyy") AS ms2, n
return date(datetime({epochmillis: ms1})) as fechaSintomas,
    date(datetime({epochmillis: ms2})) as fechaDefuncion,
     apoc.date.convert(apoc.date.add(ms2 - ms1, "ms", 1, "day"), "ms", "d") as
    InicioDefuncion,
    n.ocupacion
```

Listing 4.3: Periodo de los decesos por COVID-19.

Tabla 4.8: Edad de mujeres embarazadas fallecidas por COVID-19.

Edad	Casos
24	1
25	1
28	1
29	2
31	2
32	1
36	1
38	1
39	1

Tabla 4.9: Meses de embarazo de mujeres embarazadas fallecidas por COVID-19.

Mes de embarazo	Casos
5	1
6	1
7	1
8	5
9	3

Tabla 4.10: Ocupación de mujeres embarazadas fallecidas por COVID-19.

Ocupación	Casos
Hogar	8
Enfermeras	2
Otras profesionistas	1

Tabla 4.11: Vacunación contra la influenza.

Vacuna contra la influenza	Casos
No	8
Si	3

Tabla 4.12: Periodo de inicio de síntomas y fecha de defunción.

Caso	Inicio de síntomas	Defunción	Total de días
1	2020-04-01	2020-04-15	15
2	2020-04-10	2020-04-23	14
3	2020-05-01	2020-05-24	24
4	2020-05-09	2020-05-17	9
5	2020-05-25	2020-06-12	19
6	2020-05-25	2020-06-02	9
7	2020-06-01	2020-06-14	14
8	2020-06-15	2020-07-07	23
9	2020-08-20	2020-09-15	27
10	2020-10-25	2020-11-01	8
11	2020-12-01	2020-12-14	14

Con relación a la edad en el que fallecieron las personas contagiadas con el virus SARS-CoV-2, se observó que el rango de edad del deceso oscila entre 24 a 39 años de edad, siendo 29 y 31 las edades donde se tuvieron dos casos. Esto evidencia que el riesgo de contraer la infección y tener un desafortunado desenlace, por lo que, se debe evitar tener, en lo posible, el contacto con cualquier persona que esté enferma o tenga los síntomas de COVID-19. Además, es importante el lavado de manos con frecuencia con agua y jabón y hacer uso de cubrebocas o mascarillas faciales en lugares públicos y en el trabajo.

Otro aspecto que llama la atención es la incidencia de muerte en etapas avanzadas del embarazo, siendo el octavo y noveno mes en el que se dieron ocho fallecimientos. Por otra parte,

de los 11 casos de muerte confirmados, 8 mujeres se dedicaban a las labores del hogar, 2 fueron enfermeras y otra persona con alguna otra profesión. Además, de estos casos de mortalidad, se pudo identificar también que la mayoría de mujeres embarazadas no estaban vacunadas contra la influenza (8 casos); y el total de días entre la fecha de inicio de los síntomas y la fecha de defunción de las gestantes fue variado, algunos casos transcurrieron en promedio 15 días o menos y en otros rebasaron los 20 días, después de ser diagnosticadas con COVID-19.

4.3.1. Síntomas identificados

La Figura 4.2 muestra el grafo de conocimiento sobre los síntomas que presentaron las pacientes fallecidas. El grafo fue obtenido a través de la siguiente consulta:

```
MATCH (f:FechaDefuncion)<-[r2]-(n:Paciente)-[r1:PRESENTO]->(s:Sintoma)
return s.tipo, COUNT(r1) as degreeCount ORDER BY degreeCount DESC
```

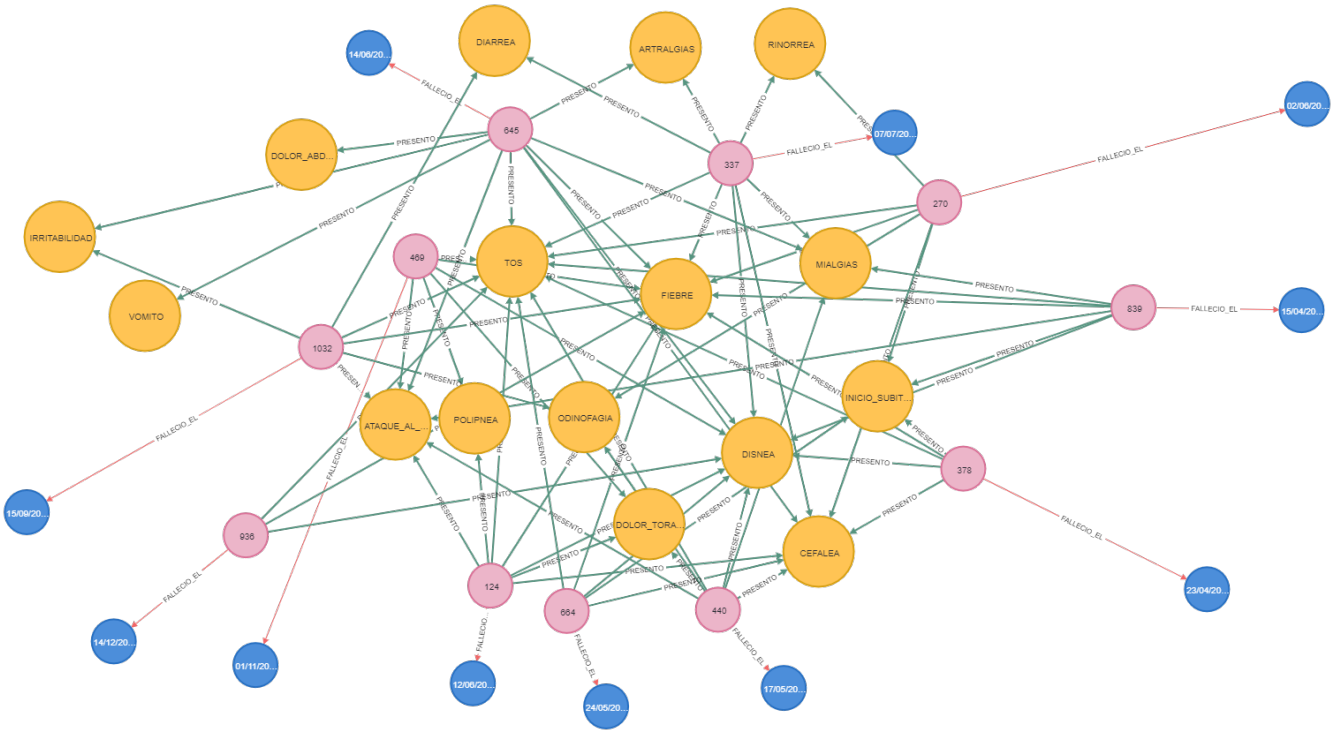

Fig. 4.2: Grafo sobre los síntomas que presentaron las pacientes fallecidas.

Entre los síntomas identificados, mostrados en la Tabla 4.13, en su totalidad, presentaron tos (11 casos), seguido de fiebre (10 casos). Otros síntomas significativos que presentaron fueron disnea (9 casos), cefalea (7 casos), ataque al estado cerebral (6 casos), mialgias (4 casos). Otros síntomas que en menor medida presentaron fueron dolor torácico, odinofagia, polipnea, rinorrea, artralgias, diarrea, irritabilidad, dolor abdominal y vómito.

Tabla 4.13: Síntomas de mujeres embarazadas fallecidas por COVID-19

Síntomas	**Casos**
Tos	11
Fiebre	10
Disnea	9
Cefalea	7
Ataque al estado cerebral	6
Inicio súbito de síntomas	4
Mialgias	4
Dolor torácico	3
Odinofagia	3
Polipnea	2
Rinorrea	2
Artralgias	2
Diarrea	2
Irritabilidad	2
Dolor abdominal	1
Vómito	1

Si bien los síntomas identificados fueron variados, existe un patrón de tos y fiebre de manera predominante. Por lo que, en caso de presentar este tipo de padecimientos, o se ha estado expuesto a alguien que tenga COVID-19, es importante acudir de inmediato con el profesional de la salud. Se recomienda además hacerse la prueba para detectar el virus que causa la enfermedad COVID-19. En caso de tener la enfermedad y se está embarazada, entonces el tratamiento se centrará en aliviar los síntomas, y puede incluir tomar líquido y descansar, así como tomar medicación para reducir la fiebre, aliviar el dolor, o reducir la tos. Existen casos graves que requieren de tratamientos especializados en los hospitales.

4.3.2. Enfermedades identificadas

Con respecto a las enfermedades identificadas en los casos de mortalidad, mostrados en la Tabla 4.14, estas fueron relacionadas mediante un grafo mostrado en la Figura 4.3, donde se muestran los padecimientos comunes relacionados con COVID-19. Este grafo fue obtenido a través de la siguiente consulta:

```
MATCH (f:FechaDefuncion)<-[r2]-(n:Paciente)-[r1:DIAGNOSTICADO_CON]->(e:
    Enfermedad)
return e.tipo, COUNT(r1) as degreeCount ORDER BY degreeCount DESC
```

Tabla 4.14: Enfermedades identificadas en mujeres embarazadas fallecidas por COVID-19.

Enfermedad	**Casos**
Obesidad	5
Otra condición	2
Tabaquismo	1
Diabetes	1
Inmunosupresivo	1

Una de las enfermedades con mayor afectación, que pudo conducir a la muerte, fue la obesidad, con 5 casos. Otros padecimientos en menor medida, pero que también fueron determinantes para producir el deceso, fueron el tabaquismo, diabetes, inmunosupresivo y otras condiciones de la salud.

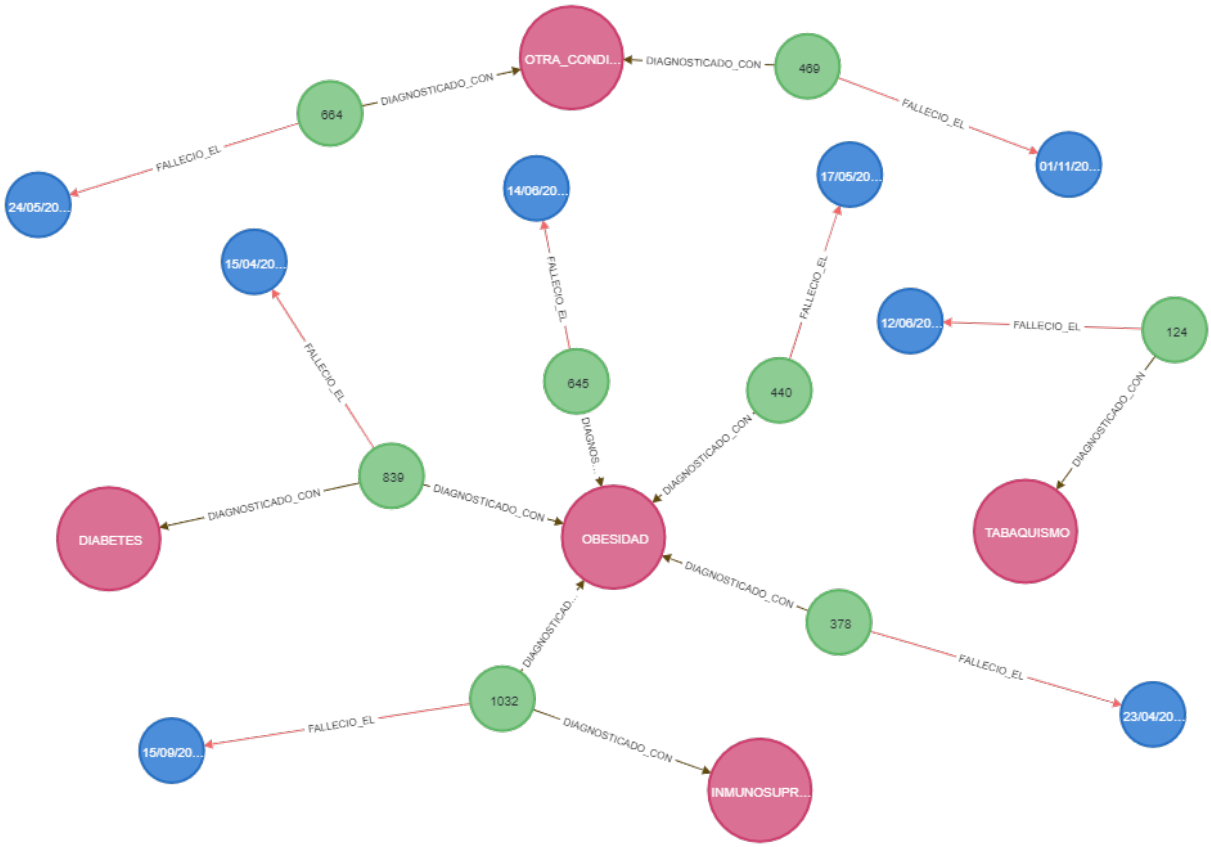

Fig. 4.3: Grafo sobre las enfermedades identificadas.

Se presentó también un grupo importante de gestantes intubadas, en total 15 casos mostrados en la Tabla 4.15, con riego de vida, de los cuales 6 sufren de obesidad, 2 de tabaquismo, una es inmunosupresiva, otra con hipertensión y 5 presentaron otra condición no especificada. Estos datos derivan directamente del grafo, mostrado en la Figura 4.4, obtenidos a través de la siguiente consulta:

```
MATCH (e:Enfermedad)<-[r2:DIAGNOSTICADO_CON]-(p:Paciente)-[r1:ATENDIDO_EN]->(u:
    UnidadMedica)
WHERE r1.intubado = "SI"
return e.tipo, COUNT(e.tipo) as cuenta ORDER BY cuenta DESC
```

Tabla 4.15: Enfermedades detectadas en gestantes intubadas con COVID-19.

Enfermedad	Casos
Obesidad	6
Otra condición	5
Tabaquismo	2
Inmunosupresivo	1
Hipertensión	1

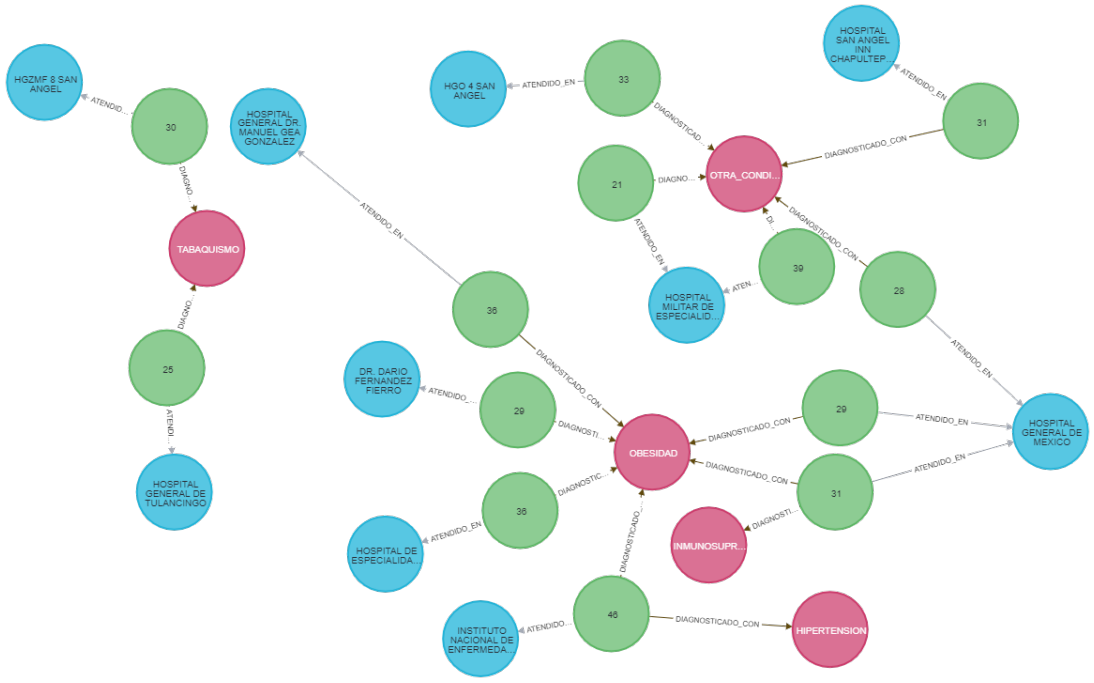

Fig. 4.4: Grafo sobre las enfermedades identificadas en gestantes intubadas.

Las gestantes no parecen tener una mayor susceptibilidad para infectarse por coronavirus, sin embargo, el tener ciertas enfermedades hacen que se presenten complicaciones graves, como la obesidad. Se evidencia que hay un porcentaje de las embarazadas, que se infectaron con SARS CoV-2 y presentaron complicaciones que les condujo a la muerte. Además, el que las gestantes sean intubadas aumenta la probabilidad de muerte de manera considerable. Por este motivo, es importante diagnosticarlas y tratarlas de forma oportuna e inmediata.

4.3.3. Identificación del sector salud

Otra de las características analizadas fue el sector salud en el que se atendieron las pacientes embarazadas, diagnosticadas con COVID-19, y que finalmente fallecieron. La Tabla 4.16 muestra un resumen de los hospitales que pertenecen al sector salud en el que fueron atendidas. Esta tabla fue obtenida a partir del grafo mostrado en la Figura 4.5, el cual fue obtenido a través de la siguiente consulta:

```
MATCH (f:FechaDefuncion)<-[r2]-(n:Paciente)-[r1:ATENDIDO_EN]->(u:UnidadMedica)-[
    r3:PERTENECE_A]->(s:SectorSalud)
return s.nombre, COUNT(r1) as degreeCount ORDER BY degreeCount DESC
```

Tabla 4.16: Sector Salud en el que fueron atendidas las gestantes fallecidas por COVID-19.

Sector salud	Casos
SSA	6
IMSS	2
ISSSTE	1
PRIVADA	1
SEDENA	1

Se observó que hay un patrón frecuente de seis casos que fueron atendidas en los hospitales del sector salud público, quienes a su vez se dedicaban a las labores del hogar. Por lo que, se infiere que fueron pacientes de escasos recursos económicos. Otros dos casos fueron atendidos en el IMSS y otros en el ISSSTE, SEDENA y alguna institución privada, respectivamente.

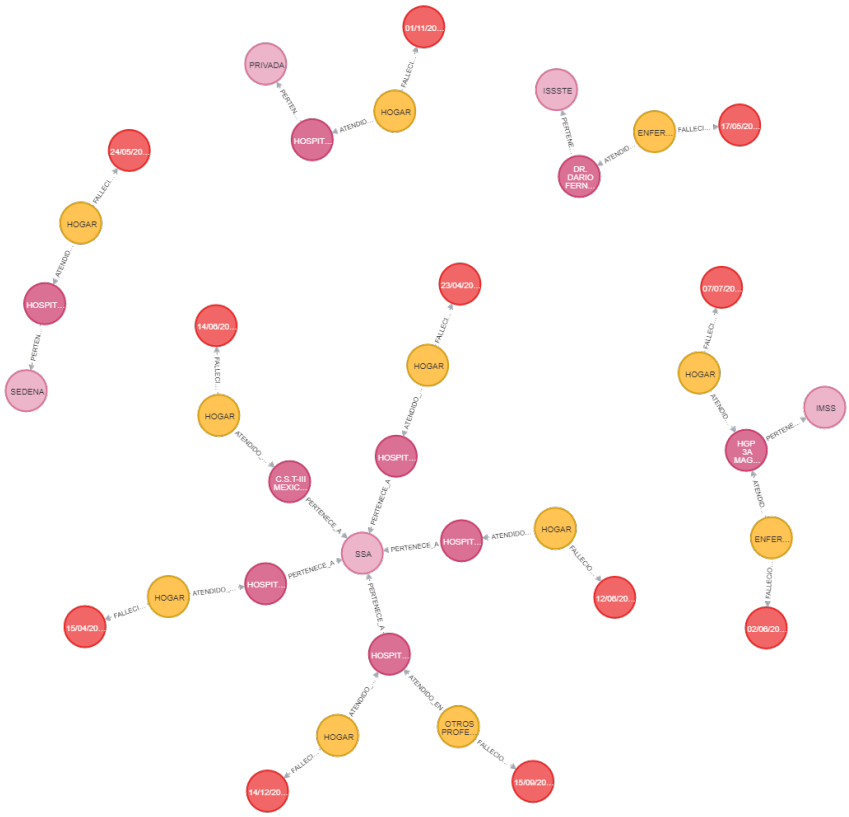

Fig. 4.5: Grafo sobre del sector salud en el que fueron atendidas las pacientes fallecidas.

4.3.4. Municipios de decesos y casos graves

Con respecto a los municipios de decesos de las gestantes que fallecieron se observó, Tabla 4.17, variados lugares de residencia, teniendo tres casos en Naucalpan de Juárez, dos en Gustavo A. Madero, y otros de un solo caso en Venustiano Carranza, Iztapalapa, Tlalpan, Cuauhtémoc, Coyoacán y Ecatepec de Morelos. Estos datos fueron obtenidos a partir del grafo mostrado en la Figura 4.6, a través de la siguiente consulta:

```
MATCH (f:FechaDefuncion)<-[r2]-(n:Paciente)-[r1:VIVIO_EN]->(l:Localidad)-[r3:
    PERTENECE_A]->(m:Municipio)
return m.nombre, COUNT(m) as degree ORDER BY degree DESC
```

Tabla 4.17: Municipio de residencia de mujeres embarazadas fallecidas por COVID-19.

Municipio	Casos
Naucalpan de Juárez	3
Gustavo A. Madero	2
Venustiano Carranza	1
Iztapalapa	1
Tlalpan	1
Cuauhtemoc	1
Coyoacán	1
Ecatepec de Morelos	1

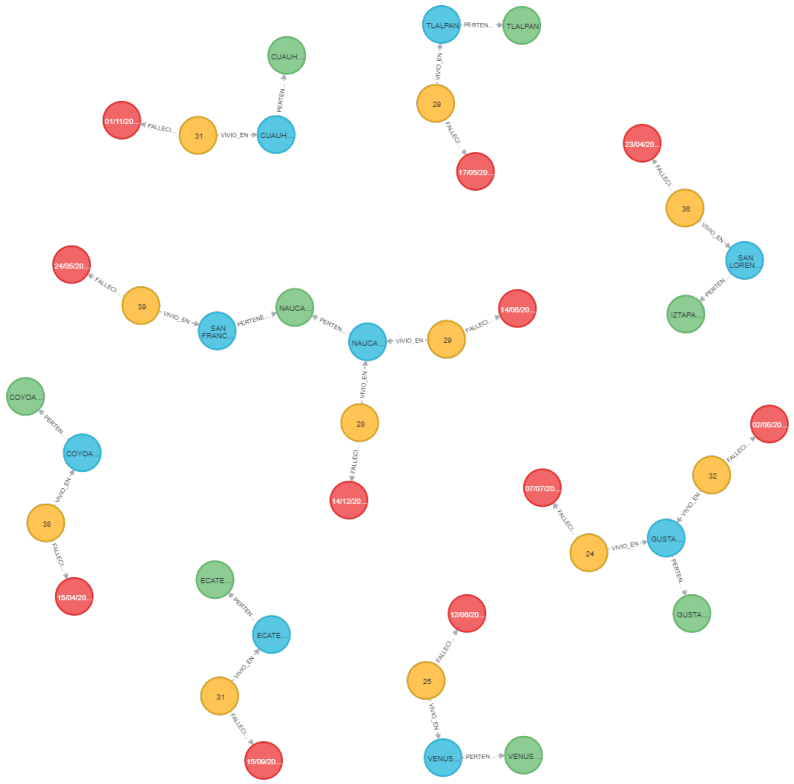

Fig. 4.6: Grafo sobre los municipios con los decesos detectados.

Caso similar ocurre con los casos graves detectados, que en total fueron 22, mostrados en la Tabla 4.18, de estos se observó que 4 pertenecen a Iztapalapa, 3 a Cuauhtemoc, 3 a Xochimilco, 2 a Gustavo A. Madero y otros con un caso en Venustiano Carranza, Azcapotzalco, Tlahuac, Alvaro Obregón, Milpa Alta, Naucalpan de Juárez, Hueypoxtla, Cuautitlan Izcalli, Tultepec y Atizapan de Zaragoza. Estos datos fueron obtenidos a partir del grafo mostrado en la Figura 4.7, a través de la siguiente consulta:

```
MATCH (p:Paciente)-[r1:VIVIO_EN]->(l:Localidad)-[r2:PERTENECE_A]->(m:Municipio)
WHERE p.estatus in ["CASO GRAVE -", "CASO GRAVE - TRASLADO"]
return m.nombre, COUNT(m.nombre) as cuenta ORDER BY cuenta DESC
```

Tabla 4.18: Municipio de casos graves de mujeres embarazadas con COVID-19.

Municipio	**Casos**
Iztapalapa	4
Cuauhtemoc	3
Xochimilco	3
Gustavo A. Madero	2
Venustiano Carranza	1
Azcapotzalco	1
Tlahuac	1
Alvaro Obregón	1
Milpa Alta	1
Naucalpan de Juárez	1
Hueypoxtla	1
Cuautitlan Izcalli	1
Tultepec	1
Atizapan de Zaragoza	1

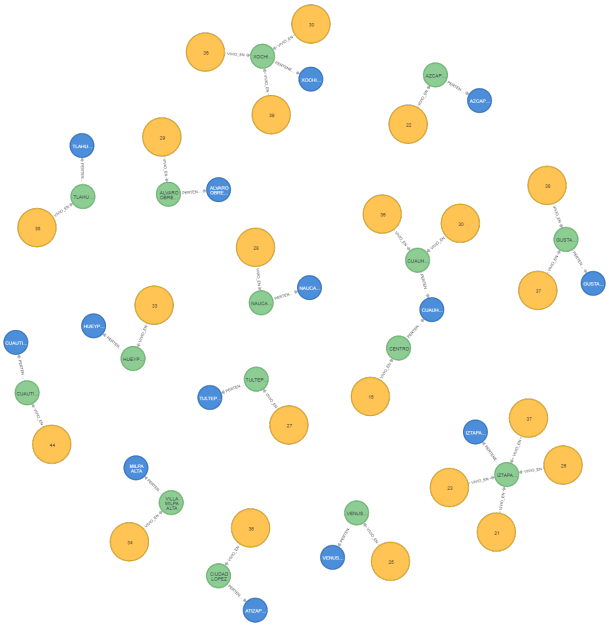

Fig. 4.7: Grafo sobre los municipios con casos graves detectados.

4.4. Patrones en el grafo de conocimiento

Con el objetivo de encontrar otros patrones de interés en el grafo de conocimiento, se utilizaron determinados algoritmos, soportados a través de Graph Data Science Playground, la cual es una extensión de Neo4j. Los algoritmos utilizados, descritos en la sección 2.5.2 (Analítica de grafos) fueron: a) Centralidad de grado (Degree centrality), b) PageRank, y c) Propagación de etiqueta (Label propagation).

a) Centralidad de grado

Como se mencionó previamente, este algoritmo se utiliza para identificar los nodos más populares en el grafo. Así, con base en la ejecución, la Figura 4.8, muestra el resultado obtenido, donde los nodos más importantes en el grafo fueron los que representan a los síntomas de la enfermedad (extremo izquierdo), seguido de sector salud donde las gestantes fueron atendidas, luego los nodos que representan las localidades de residencia, las enfermedades, y finalmente las unidades médicas.

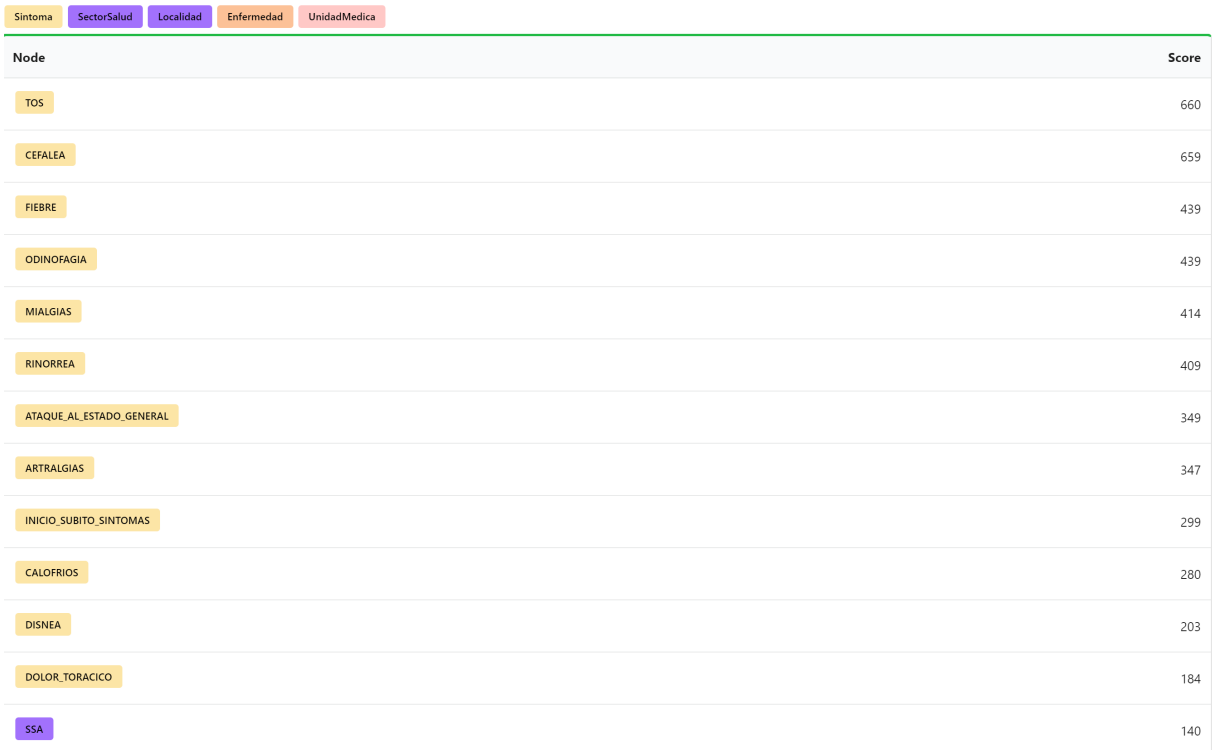

Fig. 4.8: Nodos significativos obtenidos a través del algoritmo centralidad de grado.

Se observó que los nodos más populares de los síntomas asociados con COVID-19 fueron: tos, seguido de cefalea, fiebre, odinofagia, mialgias, rinorrea, ataque al estado en general, artralgias, calofríos, disnea y dolor torácico. Caso similar se observó con los nodos más significativos de enfermedades que presentaron las gestantes diagnosticadas con coronavirus (Figura 4.9), entre los que destacan: diarrea, irritabilidad, conjuntivitis, dolor abdominal, obesidad, vómito, polipnea, entre otras.

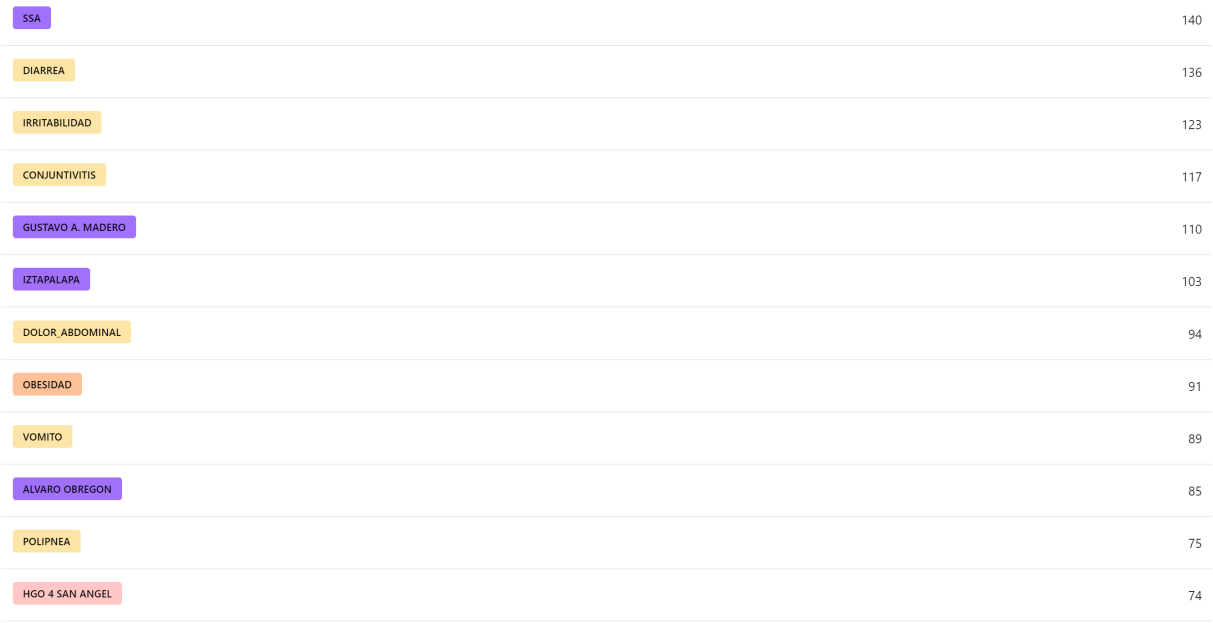

Fig. 4.9: Continuación de los nodos significativos obtenidos a través de centralidad de grado.

Por lo anterior, el patrón frecuente de las gestantes contagiadas con COVID-19 fueron tos y dolor de cabeza, quienes presentaron obesidad y que algunas de éstas acudieron al sector de salud público en la colonia Gustavo A. Madero, particularmente a la unidad médica HGO 4 de San Ángel.

b) PageRank

A través de este algoritmo se midieron los nodos que tuvieron mayor influencia en el grafo. De manera particular, se midieron las conexiones entre los nodos más influyentes del grafo de conocimiento. La Figura 4.10 muestra el resultado obtenido, donde los nodos más influyentes en el grafo fueron: a) sector salud (SSA), donde las pacientes fueron atendidas; b) seguido de la entidad federativa, en este caso Ciudad de México; y c) los síntomas más frecuentes en las gestantes, como tos y cefalea. Estos resultados contrastan también los obtenidos a través del algoritmo anterior (centralidad de grado).

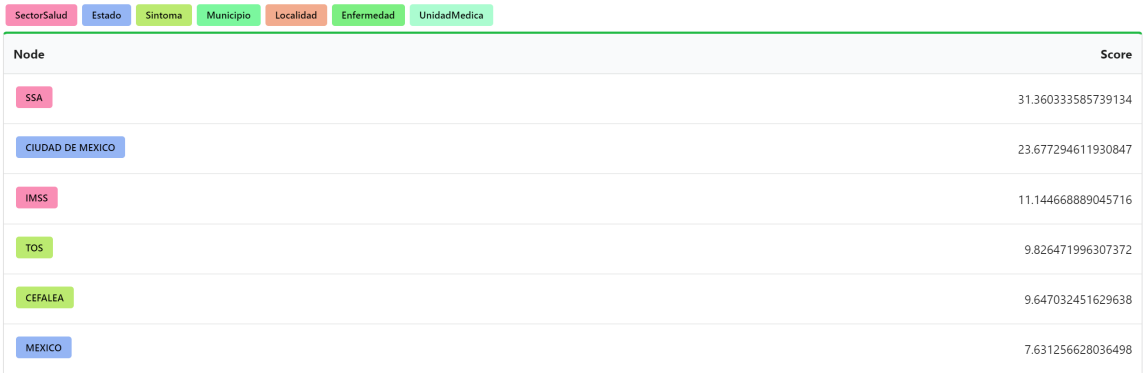

Fig. 4.10: Nodos con mayor influencia obtenidos a través del algoritmo PageRank.

Con base en lo anterior, el patrón más influyente de gestantes contagiadas de COVID-19 es congruente con lo obtenido anteriormente, siendo pacientes atendidas en el sector público, dentro de la Ciudad de México, con síntomas de tos y dolor de cabeza, principalmente.

c) Propagación de etiqueta

Para el caso de la detección de comunidades, se utilizó el algoritmo de propagación de etiquetas, mediante el cual en cada iteración se actualizaron las etiquetas de cada nodo hasta que un número importante de nodos vecinos resultaron con una misma etiqueta.

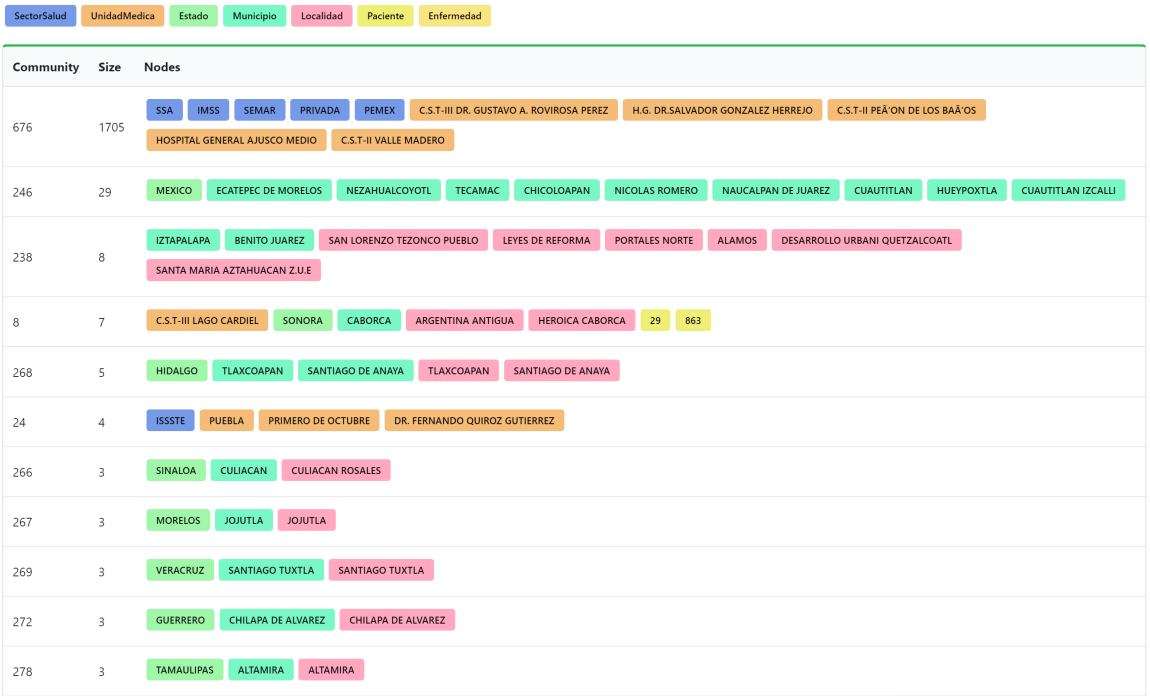

Fig. 4.11: Propagación de etiquetas en el grafo.

De los resultados obtenidos, se observó, Figura 4.11, la formación de comunidades representativas a partir de los nodos y sus etiquetas. La comunidad más representativa fue el Sector Salud y las unidades médicas en donde fueron atendidas las pacientes. Otra de las comunidades representativas fueron los municipios, donde fueron atendidas o residen las gestantes diagnosticadas con SARS-CoV-2.

4.5. Síntesis

Se observaron patrones significativos a través del análisis de los registros de datos de gestantes contagiadas con SARS-COV-2, de los cuales se pudo observar que algunas pacientes fallecidas estaban en el rango de edad de 20 a 29 años, con 8 meses de gestación y que se dedicaban a las labores domésticas. Otro dato importante fue que el 70 % de las gestantes fallecidas no se encontraban vacunadas contra la influenza, lo que permite llegar a la conclusión de que los anticuerpos generados por esta vacuna si pueden tener algún efecto positivo en el combate por reducir los contagios de SARS-COV-2. Finalmente, y de forma afortunada, resultó que el estado más frecuente en el seguimiento de las pacientes fueron aquellos casos que tuvieron el alta médico, pero también existe un importante nivel de mortalidad en esta población vulnerable, cuyos síntomas principales fueron tos, cefalea, fiebre y mialgias, principalmente; y la enfermedad de mayor afectación, que pudo conducir a la muerte, fue la obesidad. Otros padecimientos que también fueron determinantes para producir el deceso, fueron el tabaquismo, diabetes, inmunosupresivo y otras condiciones de la salud.

Capítulo 5

Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones del trabajo realizado y se establecen las futuras líneas de investigación como trabajo posterior, cuyas bases se sustentan de acuerdo a los resultados obtenidos.

5.1. Conclusiones

El impacto de la pandemia por COVID-19 en el mundo y en México ha provocado llegar al límite de las capacidades hospitalarias por varios días y semanas consecutivas, y con esto se ha presentado un alto número de fallecimientos y exceso de mortalidad. Como consecuencia de lo anterior, diversos estudios se han centrado en encontrar patrones de interés sobre los sectores más afectados por esta letal enfermedad, particularmente en grupos vulnerables, como las personas embarazadas.

Si bien el riesgo general de enfermarse gravemente a causa del virus SARS-COV-2 es alto, sigue siendo mayor para las personas embarazadas en relación con las personas que no. Se ha identificado que tener ciertas afecciones ocultas o no, y otros factores, incluida la edad, puede aumentar aún más el riesgo de enfermarse gravemente en personas embarazadas, a quienes, por su condición, se les ha dificultado cumplir el aislamiento preventivo.

Por otro lado, hay factores que pueden aumentar el riesgo de una persona embarazada de enfermarse gravemente a causa de contraer COVID-19, por ejemplo, las condiciones de los lugares en que viven, aprenden, trabajan, se entretienen y descansan también pueden incidir en los riesgos y consecuencias para la salud. Esto hace que las gestantes, a pesar del cuidado que puedan tener, no están exentas de mantener la distancia con personas que podrían estar enfermas, como médicos, enfermeras, familiares, vecinos y entorno en general. Por otra parte, las desigualdades sociales y la falta de acceso a la salud colocan a las embarazadas como un grupo de mayor riesgo de contraer COVID-19.

Para analizar a este grupo vulnerable de gestantes se utilizó el conjunto de datos abiertos del portal de la Ciudad de México, a partir de la cual se realizó la construcción del grafo de conocimiento para el análisis de la mortalidad por COVID–19 en embarazadas atendidas en la Ciudad de México. Los datos fueron adquiridos a través del Sistema Nacional de Vigilancia

Epidemiológica de la Dirección General de Epidemiología de la Secretaría de Salud. El método para la adquisición de datos fue retrospectivo debido a que se tomaron casos confirmados que fueron registrados desde marzo a diciembre de 2020, fecha de corte del análisis.

Dada la necesidad de analizar a esta población vulnerable, debido a la importante infección y muerte por COVID-19 en gestantes atendidas en la Ciudad de México, y ante la existencia de una amplia variedad de variables que registran información de las pacientes, se logró, mediante el diseño e implementación del grafo de conocimiento, identificar patrones de datos de interés sobre la mortalidad de gestantes diagnosticadas con COVID–19. Por lo tanto, como conclusiones particulares destacan:

- Se logró diseñar e implementar el grafo de conocimiento de propiedades con base en la tecnología Neo4j. Como resultado del grafo se analizó la mortalidad por COVID–19 en gestantes atendidas en la Ciudad de México. Es importante destacar que Neo4j, como base de datos orientada a grafos, está diseñada para desarrollar modelos de datos flexibles y crear aplicaciones modernas, las cuales siguen ganando importancia en los últimos años debido a sus características, funcionalidad, rendimiento, flexibilidad y representación de interconexión de datos del mundo real.
- Para el diseño y modelado del grafo de conocimiento de propiedades se tomaron en cuenta variables representativas asociadas con la enfermedad COVID-19 en embarazadas que fueron atendidas en la Ciudad de México. Este grupo de variables fueron estructuradas a su vez en nodos, como: paciente, sector salud, unidad médica, entidad de residencia, municipio de residencia, localidad de residencia, enfermedad, fecha de síntomas, fecha de defunción y síntomas.
- Con base en el grafo de conocimiento propuesto, en cuyos nodos y relaciones se estructuró el modelo de datos, se analizaron los patrones más significativos recuperados a través de consultas en el grafo, como: comorbilidad asociada; tiempo promedio entre el inicio de los síntomas y el fallecimiento por esta enfermedad; sectores de salud con más casos atendidos y más fallecimientos; relación de la entidad de residencia y la atención recibida en la Ciudad de México; periodo con más fallecimientos; síntomas en cada municipio, enfermedades en cada localidad; y casos graves.
- Los resultados alcanzados mostraron con éxito la conformación del grafo de conocimiento con 1761 nodos y 9209 relaciones. De los cuales se obtuvieron subgrafos y tablas a través de consultas para entender el comportamiento de la grave enfermedad por COVID-19 que puede conducir a la muerta a la población en general, y de manera particular en las personas gestantes. El total de casos analizados fue 1106 de gestantes diagnosticadas con SARS-CoV-2, de los cuales se identificaron patrones de interés, derivados de las propiedades que fueron definidas en los nodos del grafo de conocimiento.
- Se puede afirmar que el rango de edad en la que fallecieron las gestantes contagiadas con SARS-CoV-2 oscila entre 24 a 39 años de edad. Esto evidencia el alto riesgo de contraer la infección y tener un desafortunado desenlace, por lo que, se debe evitar tener, en lo posible, el contacto con cualquier persona que esté enferma o tenga los síntomas de COVID-19. Además, otro aspecto que llama la atención es la incidencia de muerte en etapas avanzadas del embarazo, siendo el octavo y noveno mes en el que se dieron la mayor cantidad de

decesos. Se pudo identificar también que la mayoría de embarazadas no estaban vacunadas contra la influenza.

- Por otra parte, se observó que en las gestantes existe susceptibilidad para infectarse por coronavirus, pero también existen enfermedades que hacen que se presenten complicaciones graves, como es el caso de la obesidad, que pueden conducir a la muerte. Otras enfermedades fueron el tabaquismo e hipertensión. Por este motivo, es importante diagnosticarlas y tratarlas de forma oportuna e inmediata.

- Adicionalmente, no hay manera de evitar por completo el riesgo de infección de COVID-19. Por lo que, es importante protegerse tanto como sea posible y considerar la situación y riesgo que puede llegar a tener una embarazada y su familia a la hora de interactuar con otras personas. Así, se deben tener en cuenta diversas acciones para reducir la propagación de COVID-19, como: i) considerar vacunarse contra COVID-19; ii) limitar las interacciones presenciales con otras personas que pudieron haber estado expuestas a COVID-19, o que podrían tener la infección; iii) usar mascarilla y evitar a otras personas que no la usen; iv) mantener distancia con otras personas; v) evitar las multitudes y espacios con mala ventilación; y vi) lavarse las manos con frecuencia y usar frecuentemente desinfectante de manos.

- Finalmente, se logró exitosamente cumplir con los objetivos de la investigación por medio de la presentación de los datos en una orientación emergente como es la orientación a grafos, superando los retos de conocer las tecnologías que implementan ésta y que facilitarían la representación deseada. Además se logró obtener el conocimiento suficiente en el lenguaje de consultas de la tecnología elegida que nos permitió navegar en el grafo resultante y encontrar los resultados presentados. Por lo anterior, se puede asegurar que la presentación de datos orientada a grafos en casos de uso particulares en los cuáles los datos se expresan de manera natural en forma de grafo resulta una mejor opción que la clásica orientación relacional con tablas debido a que los datos se pueden explorar y analizar más fácilmente como es en el caso de las representaciones gráficas mostradas durante esta tesis.

5.2. Trabajo futuro

Si bien los resultados obtenidos fueron favorables, el avance tecnológico deja abierta futuras líneas de investigación, sobre todo por la gran cantidad de datos e información almacenada sobre otros grupos vulnerables con COVID-19 que aqueja a la sociedad en general. Entre los trabajos futuros destacan:

- Repetir el análisis utilizando otros periodos de corte, así como otras variables disponibles en las diversas fuentes de datos abiertas, de la Ciudad de México y el país que se pueda encontrar. El propósito es a modo de corroborar los hallazgos encontrados hasta la fecha.

- El desarrollo del caso de estudio demostró una amplia aplicabilidad de los grafos de conocimiento de propiedades en el campo de la salud, por lo que se deja abierta esta posibilidad para ampliar el espectro de aplicación a otros grupos de poblaciones vulnerables, como los estudiantes, personal de la salud, adultos mayores, entre otros.

- Extender el trabajo con la implementación de otros algoritmos, como los de aprendizaje automático, con el propósito de encontrar patrones de datos a partir de aprendizaje supervisado, no supervisado y profundo.

Anexo A

Variables de la fuente de datos

En este apartado se presenta la fuente de datos recuperado de la base de datos del Sistema Nacional de Vigilancia Epidemiológica (SINAVE), de la Dirección General de Epidemiología de la Secretaría de Salud. La Tabla A.1 muestra el nombre de la variable, descripción, tipo de dato y los valores que éstas pueden tomar.

Tabla A.1: Variables disponibles en la fuente de datos.

Nombre de la variable	**Descripción**	**Tipo de dato**	**Valores que puede tomar**
origen	Registra si el paciente fue diagnosticado dentro de las Unidades de Salud Monitoras de Influenza (USMI) o fuera de éstas.	Booleano	Fuera de USMI o USMI
sector	Designa si la USMI pertenece a alguna institución de salud del ámbito público o privado.	Nominal	SSA; Privada; IMSS; ISSSTE; Cruz Roja; Estatal; IMSS-Oportunidades; PEMEX; SEDENA; SEMAR; Universitario
unidad_medica	Nombre de la unidad médica	Nominal	-
cve_entidad _unidad_medica	Designa la clave numérica de la entidad federativa en la que se localiza la Unidad Médica, según el Marco Geostadístico del INEGI.	Numérico	1-32
entidad_medica	Designa el nombre de la entidad federativa en la que se localiza la Unidad Médica.	Nominal	Algún estado de México

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
delegacion_ unidad _ medica	Las Delegaciones son unidades operativas que funcionan de manera autónoma en los estados, brindando servicios institucionales a la población local.	Nominal	-
unidad_ medica	Designa el nombre de la Unidad Médica de atención.	Nominal	-
fecha_ registro	Fecha de registro en el sistema.	Fecha	-
sexo	Indica el sexo del paciente	Booleano	Femenino; Masculino
entidad_ residencia	Indica la entidad de residencia del paciente.	Nominal	Algún estado de México
cve_ entidad _ residencia	Indica la clave de la entidad de residencia del paciente.	Numérico	1-32
municipio_ residencia	Indica el municipio de residencia del paciente.	Nominal	-
clave_ municipio _ residencia	Indica la clave del municipio de residencia del paciente.	Numérico	1-525
localidad_ residencia	Nombre de la localidad de residencia del paciente.	Nominal	-
clave_ localidad _ residencia	Indica la clave de la localidad de residencia del paciente, según el Marco Geostadístico del INEGI.	Numérico	1- 3128; 9999

Nombre de la Variable	**Descripción**	**Tipo de Dato**	**Valores que puede tomar**
tipo_paciente	Indica el tipo de paciente según si recibió atención médica ambulatoria (es decir, que recibió atención médica y de diagnóstico pero que no tuvo que pasar la noche en la unidad de salud), o si requirió hospitalización.	Booleano	Ambulatorio; Hospitalizado
evolucion	Indica la evolución del paciente a la fecha de corte de la base de datos.	Nominal	Alta - curación; Alta - mejoría; Alta - traslado; Alta - voluntaria; Caso grave; Caso grave-traslado; Caso no grave; Defunción; En tratamiento; Referencia; Seguimiento domiciliario; Seguimiento terminado
fecha_defuncion	Indica la fecha de defunción del paciente.	Fecha	-
semana_defuncion	Indica la semana de la fecha de defunción del paciente, contando a partir de la penúltima semana de diciembre de 2019.	Numérico	1-53
intubado	Indica si el paciente requirió intubación o no.	Booleano	Si; No
diagnostico_clinico_neumonia	Indica si cuando el paciente llega a la unidad médica presenta un diagnóstico clínico de neumonía.	Booleano	Si; No
edad	Indica la edad del paciente.	Numérico	-

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
nacionalidad	Indica la nacionalidad del paciente.	Booleano	Mexicana; Extranjera
esta_embarazada	Indica si la paciente se encontraba embarazada al momento del diagnóstico.	Booleano	Si; No
meses_embarazo	Indica el número de meses de embarazo de la paciente.	Numérico	0-10
es_indigena	Indica si el paciente se autoadscribe como persona indígena.	Booleano	Sí; No
habla_lengua_indigena	Indica si el paciente habla alguna lengua indígena.	Booleano	Sí; No
ocupacion	Indica la ocupación del paciente.	Nominal	-
servicio_ingreso	Indica el servicio al que ingresó el paciente para ser atendido.	Nominal	Infectología; Medicina interna; Neumología; UCI; UCIN; Urgencias adultos; Urgencias Cirugía; Urgencias Pediatría; UTIP
fecha_ingreso	Indica la Fecha de ingreso del paciente al servicio.	Fecha	-
fecha_inicio_sintomas	Indica la Fecha en que el paciente comenzó con síntomas.	Fecha	
diagnostico_probable	Indica el primer diagnóstico probable.	Nominal	Enfermedad Tipo Influenza (ETI); Infección Respiratoria Aguda Grave (IRAG)
fiebre	Indica si el paciente presentó fiebre como síntoma.	Nominal	Si; No; Se ignora
tos	Indica si el paciente presentó tos como síntoma.	Nominal	Si; No; Se ignora

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
odinofagia	Indica si el paciente presentó odinofagía como síntoma.	Nominal	Si; No; Se ignora
disnea	Indica si el paciente presentó disnea (dificultad para respirar) como síntoma.	Nominal	Si; No; Se ignora
irritabilidad	Indica si el paciente presentó irritabilidad como síntoma.	Nominal	Si; No; Se ignora
diarrea	Indica si el paciente presentó diarrea como síntoma.	Nominal	Si; No; Se ignora
dolor_toracico	Indica si el paciente presentó Dolor torácico.	Nominal	Si; No; Se ignora
calofrios	Indica si el paciente presentó calofríos como síntoma.	Nominal	Si; No; Se ignora
cefalia	Indica si el paciente presentó cefalea como síntoma.	Nominal	Si; No; Se ignora
mialgias	Indica si el paciente presentó mialgias como síntoma.	Nominal	Si; No; Se ignora
artralgias	Indica si el paciente presenta artralgias.	Nominal	Si; No; Se ignora
ataque_al_estado_general	Indica si el paciente presenta un ataque al estado general.	Nominal	Si; No; Se ignora
rinorrea	Indica si el paciente presentó rinorrea como síntoma.	Nominal	Si; No; Se ignora
polipnea	Indica si el paciente presentó polipnea como síntoma.	Nominal	Si; No; Se ignora
vomito	Indica si el paciente presentó vómito como síntoma.	Nominal	Si; No; Se ignora

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
dolor_ abdominal	Indica si el paciente presentó dolor abdominal como síntoma.	Nominal	Si; No; Se ignora
conjuntivitis	Indica si el paciente presentó conjuntivitis como síntoma.	Nominal	Si; No; Se ignora
cianosis	Indica si el paciente presentó cianosis como síntoma.	Nominal	Si; No; Se ignora
inicio_ subito _ sintomas	Indica los casos en los que los síntomas no fueron paulatinos, sino que iniciaron súbitamente.	Nominal	Si; No; Se ignora
diabetes	Indica si el paciente padecía diabetes al momento del diagnóstico.	Nominal	Si; No; Se ignora
epoc	Indica si el paciente padecía EPOC al momento del diagnóstico.	Nominal	Si; No; Se ignora
asma	Indica si el paciente padecía asma al momento del diagnóstico.	Nominal	Si; No; Se ignora
inmunosupresivo	Indica si el paciente es inmunosupresivo al momento del diagnóstico.	Nominal	Si; No; Se ignora
hipertension	Indica si el paciente padecía hipertensión al momento del diagnóstico.	Nominal	Si; No; Se ignora
VIH_ SIDA	Indica si el paciente padecía VIH/SIDA al momento del diagnóstico.	Nominal	Si; No; Se ignora

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
otra_ condicion	Indica si el paciente padecía alguna otra condición de salud o cormobilidad al momento del diagnóstico.	Nominal	Si; No; Se ignora
enfermedad_ cardiaca	Indica si el paciente padecía alguna enfermedad cardiáca al momento del diagnósitco.	Nominal	Si; No; Se ignora
obesidad	Indica si el paciente padecía obesidad al momento del diagnóstico.	Nominal	Si; No; Se ignora
insuficiencia_ renal _ cronica	Indica si el paciente padecía insuficiencia renal crónica al momento del diagnóstico.	Nominal	Si; No; Se ignora
tabaquismo	Indica si el paciente padecía tabaquismo al momento del diagnóstico.	Nominal	Si; No; Se ignora
recibio_ tratamiento	Recibió cualquier tipo de tratamiento antes de ser registrado.	Booleano	Si; No
recibio_ tratamiento _ antibiotico	Recibió tratamiento con antibiótico antes de ser registrado.	Booleano	Si; No
recibio_ tratamiento _ antiviral	Recibió tratamiento con antivirales antes de ser registrado.	Booleano	Si; No
antiviral	Indica si el paciente tomó algún antiviral y cuál.	Nominal	Aciclovir; Amantadina; Iopinavir Ritonavir; Kaletra; Lopinavir; Oseltamivir; Rimntadina; Sin; Zavamivir; No especifica

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
fecha_ inicio _ antiviral	Indica la fecha en que el paciente inició el tratamiento antiviral.	Fecha	-
contacto_ infeccion _ viral	Contacto con un caso de infección respiratoria viral en los últimos 7 días.	Booleano	Si; No
contacto_ aves	Pregunta si el paciente tuvo contacto con aves en los últimos 7 días.	Booleano	Si; No
contacto_ cerdos	Pregunta si el paciente tuvo contacto con cerdos en los últimos 7 días.	Booleano	Si; No
contacto_ animales	Indica si el paciente convive con animales y qué tipo de animales.	Nominal	-
vacunado	Si recibió la vacuna contra la influenza en el último año.	Nominal	-
fecha_ estimada _ vacunación	Fecha estimada de vacuna estacional.	Fecha	-
toma_ muestra	Indica si al paciente se le tomó la muestra para detectar COVID19.	Booleano	Si; No
laboratorio	Indica el nombre del laboratorio que tomó la muestra.	Nominal	-

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
resultado_definitivo	Indica el resultado definitivo de la muestra de laboratorio	Nominal	AH3; B; CORONA HKU1; CORONA HNL63; ENTERO-V//RHINOVIRUS; INF 1; INF AH1N1 PMD; NEGATIVO; NO ADECUADO; NO RECIBIDA; NO SUBTIPIFICADO; PARAINFLUENZA 1; PARAINFLUENZA 2; RECHAZADA; SARS-CoV-2; VSR
es_migrante	Pregunta si el paciente es migrante.	Booleano	Si; No
pais_nacionalidad	Pregunta la nacionalidad del paciente.	Nominal	-
pais_origen	Pregunta el país de origen, en caso de que sea extranjero.	Nominal	-
fecha_ingreso_mexico	Fecha de ingreso al país.	Fecha	-
puerperio	Indica si la paciente se encontraba en periodo puerperal al momento del diagnóstico.	Booleano	Sí; No
dias_puerperio	Indica los días en puerperio que llevaba la paciente al momento del diagnóstico.	Numérico	-
antipireticos	Indica si el paciente tomó antipiréticos antes del diagnóstico.	Booleano	SÍ; No

Nombre de la Variable	Descripción	Tipo de Dato	Valores que puede tomar
UCI	Si el paciente entra a una unidad de cuidados intensivos. Los pacientes ambulatorios por default no llenan esta variable.	Booleano	Sí; No
linaje_influenza_b	Sólo para los casos de Influenza-b.	Nominal	Victoria; Yamagata
viaje_1	Si el paciente realizó viajes en los últimos 14 días y a dónde lo realizó.	Nominal	-
viaje_2	Si el paciente realizó viajes en los últimos 14 días y a dónde lo realizó.	Nominal	-
viaje_3	Si el paciente realizó viajes en los últimos 14 días y a dónde lo realizó.	Nominal	-
viaje_4	Si el paciente realizó viajes en los últimos 14 días y a dónde lo realizó.	Nominal	-
viaje_5	Si el paciente realizó viajes en los últimos 14 días y a dónde lo realizó.	Nominal	-

Anexo B

Código en Cypher

```
1
2 LOAD CSV FROM 'file:///diccionario.csv' AS row
3 with row[0] as id, row[1] as nom
4 FOREACH ( ignoreMe in CASE WHEN id="sector" THEN [1] ELSE [] END | MERGE(S:
    SectorSalud{nombre:nom}))
5 FOREACH ( ignoreMe in CASE WHEN id="Hosp" THEN [1] ELSE [] END | MERGE(U:
    UnidadMedica{nombre:nom}))
6 FOREACH ( ignoreMe in CASE WHEN id="Est" THEN [1] ELSE [] END | MERGE(E:Estado{
    nombre:nom}))
7 FOREACH ( ignoreMe in CASE WHEN id="Mun" THEN [1] ELSE [] END | MERGE(M:
    Municipio{nombre:nom}))
8 FOREACH ( ignoreMe in CASE WHEN id="Loc" THEN [1] ELSE [] END | MERGE(L:
    Localidad{nombre:nom}))
9 FOREACH ( ignoreMe in CASE WHEN id="def" THEN [1] ELSE [] END | MERGE(fd:
    FechaDefuncion{fecha:nom}))
10 FOREACH ( ignoreMe in CASE WHEN id="sin" THEN [1] ELSE [] END | MERGE(fs:
    FechaSintomas{fecha:nom}))
11 FOREACH ( ignoreMe in CASE WHEN id="sint" THEN [1] ELSE [] END | MERGE(Si:
    Sintoma{tipo:nom}))
12 FOREACH ( ignoreMe in CASE WHEN id="enf" THEN [1] ELSE [] END | MERGE(en:
    Enfermedad{tipo:nom}))
13 RETURN *
```

```
1
2 LOAD CSV FROM 'file:///paciente.csv' AS line
3 with line[0] as id, line[8] as def, line[9] as semdef, line[17] as inicio, line
    [5] as loc, line[4] as mun,line[3] as est
4 MATCH(P:Paciente{id:id})
5 MATCH(F:FechaDefuncion{fecha:def})
6 MATCH(FS:FechaSintomas{fecha:inicio})
7 MATCH(L:Localidad{nombre:loc})
8 MATCH(M:Municipio{nombre:mun})
9 MATCH(E:Estado{nombre:est})
10 MERGE (P)-[R:FALLECIO_EL{semana:semdef}]->(F)
11 MERGE (P)-[R:INICIO_EL]->(FS)
12 MERGE (P)-[R:VIVIO_EN]->(L)
13 MERGE (L)-[R:PERTENECE_A]->(M)
14 MERGE (M)-[R:PERTENECE_A]->(E)
```

```
1 LOAD CSV FROM 'file:///paciente.csv' AS line
2 with toInteger(line[0]) as id, line[1] as sec, line[2] as hosp, line[6] as tipo,
    line[10] as intu, line[15] as serv,line[16] as ing,line[48] as trata, line
```

```
    [49] as vir,
MATCH(P:Paciente{id:id})
MATCH(S:SectorSalud{nombre:sec})
MATCH(U:UnidadMedica{nombre:hosp})
MERGE (P)-[R1:ATENDIDO_EN{tipoAtencion:tipo,fechaIngreso:ing,servicioIngreso:
    serv,intubado:intu,tratamiento:trata,antibiotico:anti,antiviral:vir}]->(U)
MERGE (U)-[R2:PERTENECE_A]->(S)
```

```
LOAD CSV FROM 'file:///paciente.csv' AS line
with line[0] as id, line[18] as fie, line[19] as tos, line[20] as odi, line[21]
    as dis, line[22] as irri,line[23] as di, line[24] as dol, line[25] as calo,
    line[26] as cefa, line[27] as mia,line[28] as art, line[29] as ataq, line
    [30] as rino, line[31] as poli, line[32] as vom,line[33] as abdo, line[34]
    as conju, line[35] as cian, line[36] as inicio, line[37] as diabe, line[38]
    as epoc, line[39] as asma, line[40] as inmuno, line[41] as hiper, line[42]
    as vih, line[43] as otra, line[44] as cardi, line[45] as obe, line[46] as
    insu, line[47] as taba
MATCH(P:Paciente{id:id})
MATCH(A:Sintoma{tipo:"FIEBRE"})
MATCH(A:Sintoma{tipo:"TOS"})
MATCH(A:Sintoma{tipo:odi})
MATCH(A:Sintoma{tipo:dis})
MATCH(A:Sintoma{tipo:irri})
MATCH(A:Sintoma{tipo:di})
MATCH(A:Sintoma{tipo:dol})
MATCH(A:Sintoma{tipo:calo})
MATCH(A:Sintoma{tipo:cefa})
MATCH(A:Sintoma{tipo:mia})
MATCH(A:Sintoma{tipo:art})
MATCH(A:Sintoma{tipo:ataq})
MATCH(A:Sintoma{tipo:rino})
MATCH(A:Sintoma{tipo:poli})
MATCH(A:Sintoma{tipo:vom})
MATCH(A:Sintoma{tipo:abdo})
MATCH(A:Sintoma{tipo:conju})
MATCH(A:Sintoma{tipo:cian})
MATCH(A:Sintoma{tipo:inicio})
MATCH(A:Enfermedad{tipo:diabe})
MATCH(A:Enfermedad{tipo:epoc})
MATCH(A:Enfermedad{tipo:asma})
MATCH(A:Enfermedad{tipo:inmuno})
MATCH(A:Enfermedad{tipo:hiper})
MATCH(A:Enfermedad{tipo:vih})
MATCH(A:Enfermedad{tipo:otra})
MATCH(A:Enfermedad{tipo:cardi})
MATCH(A:Enfermedad{tipo:obe})
MATCH(A:Enfermedad{tipo:insu})
MATCH(A:Enfermedad{tipo:taba})
FOREACH ( ignoreMe in CASE WHEN fie="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN tos="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN odi="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN dis="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN irri="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
```

```
FOREACH ( ignoreMe in CASE WHEN di="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN dol="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN calo="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN cefa="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN mia="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN art="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN ataq="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN rino="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN poli="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN vom="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN abdo="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN conju="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN cian="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN inicio="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    PRESENTO]->(A))
FOREACH ( ignoreMe in CASE WHEN diabe="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN epoc="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN asma="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN inmuno="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN hiper="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN vih="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN otra="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN cardi="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN obe="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN insu="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
FOREACH ( ignoreMe in CASE WHEN taba="SI" THEN [1] ELSE [] END | MERGE (P)-[R:
    DIAGNOSTICADO_CON]->(A))
```

Bibliografía

Aalst, W. v. d. (2016). *Process Mining Data Science in Action*. Berlin, Heidelberg, Springer. (Citado en la pág. 2).

Bianchetti, A., Rozzini, R., Guerini, F., Boffelli, S., Ranieri, P., Minelli, G., Bianchetti, L. & Trabucchi, M. (2020). Clinical Presentation of COVID19 in Dementia Patients. *The journal of nutrition, health & aging*, **24**(6), 560-562. https://doi.org/10.1007/s12603-020-1389-1 (citado en la pág. 28)

Cable-News-Network. (2020). *Cronología del coronavirus: así empezó y se ha extendido por el mundo el mortal virus pandémico*. https://cnnespanol.cnn.com/2020/05/14/cronologia-del-coronavirus-asi-empezo-y-se-ha-extendido-por-el-mundo-el-mortal-virus-pandemico/. (Citado en la pág. 27)

Cao, L. (2017). Data Science: Challenges and Directions. *Commun. ACM*, **60**(8), 59-68. https://doi.org/10.1145/3015456 (citado en la pág. 2)

Cao, X. (2020). COVID-19: immunopathology and its implications for therapy. *Nature reviews immunology*, **20**(5), 269-270 (citado en la pág. 1).

Cao, Y., Wang, X., He, X., Hu, Z. & Chua, T.-S. (2019). Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. *IW3C2 (International World Wide Web Conference Committee)*, 151-161 (citado en la pág. 8).

Cattuto, C., Quaggiotto, M., Panisson, A. & Averbuch, A. (2013). Time-Varying Social Networks in a Graph Database: A Neo4j Use Case. https://doi.org/10.1145/2484425.2484442 (citado en la pág. 25)

Chen, C., Akef Ebeid, I., Bu, Y. & Ding, Y. (2020). Coronavirus Knowledge Graph: A Case Study. *KDD2020: ACM Knowledge Discovery inDatabases*, 1-8 (citado en la pág. 8).

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2009). *Introduction to Algorithms* (3ra). Cambridge, Massachusetts, The MIT Press. (Citado en la pág. 23).

COVID-19: The Risk for Pregnant Women And Their Babies. (2020). *Journal of Pediatrics Neonatal Biology*, **5**(3). https://doi.org/10.33140/jpnb.05.03.03 (citado en la pág. 34)

Forbes. (2020). La OMS alerta sobre la magnitud de la pandemia en México: 'está subestimada' • Forbes México. https://www.forbes.com.mx/noticias-la-oms-alerta-sobre-la-magnitud-de-la-pandemia-en-mexico-esta-subestimada. (Citado en la pág. 5)

GDPR Compliance. (2020). https://neo4j.com/use-cases/gdpr-compliance/?ref=web-solutions-privacy-risk-compliance. (Citado en la pág. 26)

Gobierno-de-México. (2020a). Consejo de Salubridad General declara emergencia sanitaria nacional a epidemia por coronavirus COVID-19 (citado en las págs. 1-3).

Gobierno-de-México. (2020b). *COVID-19 Tablero México*. https://coronavirus.gob.mx/datos/. (Citado en la pág. 2)

Gobierno-de-México. (2020c). *Fases o escenarios de contingencia y nivel de propagación del COVID-19*. http://educacionensalud.imss.gob.mx/es/system/files/Fases-COVID19.pdf. (Citado en la pág. 29)

Godaert, L. & Proye, E. (2020). *Clinical characteristics of older patients: the experience of a geriatric short-stay unit dedicated to patients with COVID-19 in France*. http://dx.doi.org/10.1016/j.jinf.2020.04.009. (Citado en la pág. 28)

Graph Algorithms in Neo4j. (2018). https://neo4j.com/blog/graph-algorithms-in-neo4j-neo4j-graph-analytics/. (Citado en la pág. 13)

Han, J., Kamber, M. & Pei, J. (2012). *Data mining concepts and techniques* (3ra). Waltham, Massachusetts, Morgan Kaufmann Publishers. (Citado en la pág. 18).

Hernández, H. (2020). Mortalidad por covid-19 en México. Notas preliminares para un perfil sociodemográfico. *Notas de coyuntura del CRIM No.36*, 1-7 (citado en las págs. 5, 30).

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J. E. L., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.-C. N., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S. & Zimmermann, A. (2020). *Knowledge Graphs*. https://arxiv.org/abs/2003.02320. (Citado en las págs. 10, 18, 19)

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. Y col. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, **395**(10223), 497-506 (citado en las págs. 1, 3).

Lal, M. (2015). *Neo4j graph data modeling*. Packt Publishing Ltd. (Citado en la pág. 13).

Mathur, N. (2019). *Financial Services Compliance*. https://neo4j.com/whitepapers/financial-risk-reporting/?ref=web-solutions-privacy-risk-compliance. (Citado en la pág. 27)

Méndez-Arriaga, F. (2020). The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1. *Science of The Total Environment*, **735**, 139560. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.139560 (citado en la pág. 1)

Needham, M. & Hodler, A. (2019). *Graph Algorithms: Practical Examples in Apache Spark and Neo4j* (2da). Sebastopol, California, O'Reilly Media, Incorporated. (Citado en las págs. 19-24).

Neo4j. (2019). *What is a Graph Database?* https://neo4j.com/developer/graph-database/. (Citado en la pág. 11)

Parra-Bracamonte, G., Lopez-Villalobos, N. & Parra-Bracamonte, F. (2020). Clinical characteristics and risk factors for mortality of patients with COVID-19 in a large data set from Mexico. *Annals of Epidemiology*. https://www.sciencedirect.com/science/article/pii/S1047279720302866 (citado en la pág. 30)

Pérez Solá, C., Rodríguez González, M. E., Conesa Caralt, J. Y col. (2018). Bases de datos noSQL (citado en la pág. 11).

Robinson, I., Webber, J. & Eifrem, E. (2015). *Graph Databases*. Beijing, China, OReilly. (Citado en las págs. 11, 12).

Rodríguez, J. G., Muñoz, J. M., Muela, F. J., García-Prendes, C. G., Rivera, M. M. & Armas, L. G. (2020a). Variables asociadas con mortalidad en una población de pacientes mayores de 80 años y con algún grado de dependencia funcional, hospitalizados por COVID-19 en un Servicio de Geriatría. *Revista Española de Geriatría y Gerontología* (citado en la pág. 28).

Rodríguez, J. G., Muñoz, J. M., Muela, F. J., García-Prendes, C. G., Rivera, M. M. & Armas, L. G. (2020b). Variables asociadas con mortalidad en una población de pacientes mayores de 80 años y con algún grado de dependencia funcional, hospitalizados por COVID-19 en un Servicio de Geriatría. *Revista Española de Geriatría y Gerontología*. https://www.sciencedirect.com/science/article/pii/S0211139X20301098?via=ihub (citado en la pág. 30)

Rosen, K. (2011). *Discrete Mathematics and Its Applications* (7ma). New York, Estados Unidos, McGraw-Hill Education. (Citado en las págs. 7, 8).

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S. & Sontag, D. (2017). *Learning a Health Knowledge Graph from Electronic Medical Records*. https://www.nature.com/articles/s41598-017-05778-z. (Citado en la pág. 30)

Sdowski, G. & Rathle, P. (2020). *Graph Database Use Cases: Stop Fraudsters with Neo4j*. https://neo4j.com/use-cases/fraud-detection/. (Citado en las págs. 25, 26)

Secretaría de Salud Gobierno de la Ciudad de México. (2021). *Portal de Datos Abiertos de la CDMX*. https://datos.cdmx.gob.mx/dataset/base-covid-sinave. (Citado en las págs. 4, 33)

Sedgewick, R. & Wayne, K. (2011). *Algorithms* (4ta). Boston, Massachusetts, Addison-Wesley. (Citado en las págs. 19, 20).

Sun, J., He, W.-T., Wang, L., Lai, A., Ji, X., Zhai, X., Li, G., Suchard, M., Tian, J., Zhou, J. Y col. (2020). COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *CellPress REVIEWS, Trends in Molecular Medicine*, 483-495 (citado en la pág. 28).

Trilla, A. (2020). Un mundo, una salud: la epidemia por el nuevo coronavirus COVID-19. *Medicina Clınica*, **154**(5), 175 (citado en la pág. 29).

Wang, S.-T., Jin, J., Rivett, P. & Kitazawa, A. (2015). Technical Survey Graph Databases and Applications. *International Journal of Semantic Computing*, **09**(04), 523-545. https://doi.org/10.1142/S1793351X15500129 (citado en la pág. 11)

Webber, J. (2020). *Graph Database Use Cases: Optimize Real-Time Recommendations*. https://neo4j.com/use-cases/real-time-recommendation-engine/. (Citado en las págs. 14, 26)

World-Health-Organization. (2020a). *Cronología de la respuesta de la OMS a la COVID-19*. https://www.who.int/es/news-room/detail/29-06-2020-covidtimeline. (Citado en la pág. 27)

World-Health-Organization. (2020b). *Middle East respiratory syndrome coronavirus (MERS-CoV)*. https://www.who.int/emergencies/mers-cov/en/. (Citado en la pág. 30)

World-Health-Organization. (2020c). *Nuevo coronavirus 2019*. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019. (Citado en las págs. 1, 28, 29)

World-Health-Organization. (2020d). *Preguntas y respuestas sobre la enfermedad por coronavirus (covid-19)*. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses?gclid=CjwKCAjwyo36BRAXEiwA24CwGWf-UEWHC5Zf4hmtq-P9KIRNR2k-3VvPbK9SeqhYIjNH9fS3EgeUIBoCw2AQAvD_BwE. (Citado en la pág. 28)

World-Health-Organization. (2015). *Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003*. https://www.who.int/csr/sars/country/table2004_04_21/en/. (Citado en la pág. 30)

World-Health-Organization. (2020e). *Vías de transmisión del virus de la COVID-19: repercusiones para las recomendaciones relativas a las precauciones en materia de prevención y control de las infecciones*. https://www.who.int/es/news-room/commentaries/detail/

modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations. (Citado en la pág. 29)

World-Health-Organization. (2020f). *WHO Coronavirus Disease (COVID-19) Dashboard*. https://covid19.who.int/. (Citado en las págs. 1, 30)

Zelenetz, M. (2020). *Neo4j NewYork–Presbyterian Hospital Case Study*. https://neo4j.com/case-studies/newyork-presbyterian-hospital/?ref=web-solutions-life-sciences. (Citado en la pág. 27)