Universidad Nacional Autónoma de México

Facultad de Ciencias

Análisis del estado ocupacional de las mujeres a través de una red bayesiana

T E S I S

que para obtener el título de:

Matemática Aplicada

presenta:

Julia Trinidad Reyes

Tutora:

Laura Clementina Eslava Fernández

Ciudad Universitaria, CDMX 2023

Índice general

Índice de figuras

Índice de tablas

"Dame la perseverancia de las olas del mar, que hacen de cada retroceso un punto de partida para un nuevo avance" - Gabriela Mistral

Agradecimientos

Este trabajo fue posible gracias al valioso y constante apoyo de la Dra. Laura Eslava, quien me motivó, guió e inspiró a lo largo de este tiempo con paciencia y comprensión.

Gracias a mi mamá Ana, que ha sido mi mayor apoyo, motivación, mi amiga incondicional y mi más preciado regalo de vida; nada de esto sería posible sin ella. A mi ángel, mamá Maxi, que desde que tengo memoria me ha apoyado incondicionalmente en todo lo humanamente posible; que con su gran amor me ayuda, levanta y guía siempre. Ambas son una gran inspiración para mí.

Gracias a mi hermanito Dani, que me hace recordar las cosas importantes en la vida y es fuente de amor inagotable. Gracias a mis tías y primas, por sus porras infinitas, su apoyo, cariño y su desmesurada confianza en mí.

A mis redes de apoyo elegidas. Kevin, Gema y Brenda, que hicieron de la licenciatura un cúmulo de experiencias inolvidables y aprendizajes más allá de la carrera; además de apoyarme y alentarme siempre. A Daniela, Vianey y Paco que me han demostrado su cariño y apoyo durante más de la mitad de mi vida. A Abril y Carlos que siempre me motivan a ir por más y están cuando les necesito. A Dany, que está en casi todas las piezas de mi vida y es un gran apoyo y fuente de cariño para mí.

Las gracias más especiales son a mi papá Fidel, quien me apoyó desde mi comienzo en este mundo y hasta que la vida nos separó; y creyó en mí más que yo misma hasta su último aliento. Gracias por su cariño infinito e incondicional; por estar y por la fortuna de ser su hija.

Finalmente, gracias al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM, con folio 212721 y número TA100820 cuya beca hizo posible este trabajo.

Capítulo 1

Introducción

La brecha laboral de género es un tema de alcance mundial; según la Organización Internacional del Trabajo (OIT), las mujeres son quienes más problemas tienen para conseguir empleos y cuando trabajan son las más propensas a tener puestos con baja categoría y realizar actividades en condiciones de vulnerabilidad, (22).

Durante 2018, en México las tasas de desempleo para ambos sexos fueron muy parecidas (3.6 % para mujeres y 3.5 % para hombres), mientras que los índices de participación para mujeres y hombres correspondían al 44.1 % y 79 % respectivamente, lo cual se traduce en una brecha de 34.9 puntos porcentuales (21). Asimismo, en 2019 la tasa de desocupación de personas identificadas como hombres fue de 3.98 % mientras que la de las mujeres fue 4.54 %; una brecha de 0.56 puntos porcentuales.

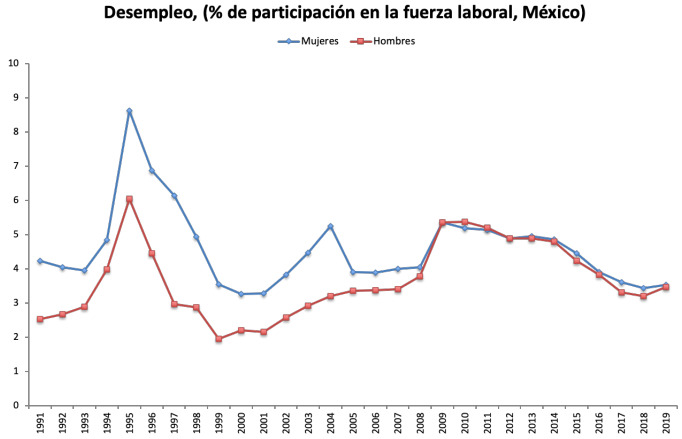


Figura 1.1: Cifras históricas del Banco Mundial desde 1991 a 2019, de la proporción de la población femenina y masculina mexicana activa que no tiene trabajo pero lo busca y está disponible para realizarlo (20).

Aunque la diferencia en la tasa de desempleo de hombres y mujeres ha disminuido a través de los años, como se dijo anteriormente, existe una brecha salarial entre hombres y mujeres, además de diferencia en las condiciones de trabajo. Una cuestión interesante es cuáles son las características sociales y económicas de las mujeres empleadas y desempleadas: ¿Hay similitudes?, ¿Hay factores puntuales que las diferencíen?, ¿Cómo se

relacionan dichos factores?

1.1. Estudios de la ocupación laboral con perspectiva de género

En 2004 se realizaron modelos de regresión y modelos dinámicos para analizar el problema en países de la OCDE, Azmat et al. (3). Se encontró que la brecha laboral de género tendía a ser más grande para las mujeres jóvenes, casadas o con hijos pequeños. Se observó además que si bien las mujeres pasan del empleo al desempleo con una tasa mayor que la de los hombres, la variable "responsabilidades domésticas" no influía de manera significativa en esta transición.

Los resultados de Livanos et al. (19) sugieren que atributos como el estado civil, el lugar en la familia, la educación, región de residencia y edad son factores que influyen significativamente en estar desempleado tanto en Grecia como en Reino Unido, aunque de manera distinta. Se encontró también que la brecha de desempleo no pudo ser explicada por las características distintas de hombres y mujeres, es decir, una explicación para la situación desfavorable de las mujeres es simplemente la discriminación del mercado laboral y no las características o habilidades que éstas poseen.

Landivar (17) encontró que el desempleo en Estados Unidos durante la recesión económica del 2008 no afectó por igual a hombres y mujeres debido a la "división" existente entre el campo laboral para hombres y mujeres; de esta manera los sectores usualmente ocupados por mujeres, por tratarse mayormente de educación y salud, no sufrieron grandes daños, por lo que concluye que en ambientes de incertidumbre económico las mujeres son quienes menos despidos sufren. Encontró también relación entre la probabilidad de que una mujer se emplee y el sector económico para el que su pareja trabaja; específicamente las esposas de hombres que se dedican a la agricultura o construcción sufren más desempleo que quienes son parejas de hombres dedicados a la salud, ayuda social o educación. También encontró una relación inversa entre el desempleo de las mujeres y las ganancias de su pareja masculina.

Baussola et al. (5) señalan que mientras en Italia la brecha de género de desempleo es significativa y constante, en los países anglosajones y del norte de Europa éste no parece ser un problema importante. Concluye también que durante la recesión económica de 2008 aunque la brecha laboral de Italia disminuyó, no fue porque la situación de las mujeres haya mejorado. Para Reino Unido también destaca que abandonar el estado de inactividad resulta más difícil para las mujeres que para los hombres. Belloc and Tilli (6) por su parte, muestran que la brecha laboral en Italia se reduce o persiste con el tiempo dependiendo de la región que se tome en cuenta.

En 2015 Koutentakis (16) realizó un análisis para 9 economías avanzadas europeas y Estados Unidos, obteniendo que en la mayoría de los países las oportunidades de empleo no eran muy distintas por género, sin embargo, las mujeres se suman a la población desempleada a una tasa mayor que los hombres.

Las referencias anteriores dan idea de que los factores sociales y económicos son importantes para el estado ocupacional de las personas. En específico para las mujeres de esas distintas regiones se puede decir que los factores importantes para determinar el estado ocupacional fueron el número de hijos, estado civil, edad y educación.

1.2. Población objetivo y herramientas

El interés de este proyecto es encontrar el conjunto de características sociales y económicas principales de las mujeres mexiquenses empleadas y desempleadas, así como definir en qué medida estos factores impactan en dicha situación. Se utilizan variables socio-económicas, debido a su relación con los roles de género asociados a las mujeres mexicanas; tales como ser cuidadoras, amas de casa, subordinadas a su pareja, etc. (12). Lo anterior bajo la hipótesis de que dichos roles y estereotipos sociales colaboran a la brecha de empleabilidad de las mujeres mexicanas, dado que en los estudios previos señalados en la Sección 1.1 dichas variables fueron de importancia para determinar el estado ocupacional de las personas.

Buscamos en literatura de los últimos 5 años (con respecto al 2020) y no encontramos algún estudio reciente con tal objetivo, pero los datos que el Instituto Nacional de Estadística y Geografía (INEGI) proporciona son suficientes para realizarlo. Para este trabajo se utiliza la base de datos de la *Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más edad* del INEGI del primer trimestre correspondiente al 2019; se eligió dicho conjunto de datos porque se requiere datos recientes, que no tengan influencia de la pandemia de SARS-CoV-2, ya que naturalmente la ocupación y empleo se han transformado y, a la fecha de inicio del proyecto, las consecuencias de este fenómeno y su estabilidad no son claras. Se trabaja con el primer trimestre, ya que es este cuando se utiliza el cuestionario de ocupación y empleo ampliado, que proporciona más datos que el cuestionario básico; específicamente este cuestionario contiene variables asociadas al tiempo que las personas dedican a actividades del hogar.

La población a estudiar son las mujeres con 18 años o más pertenecientes a la población del Estado de México, se elige el mínimo de edad ya que es a partir de los 18 años que es legal trabajar para cualquier persona en México, sin restricciones de acuerdo a la edad (18). Es preciso mencionar que el INEGI, hasta la encuesta de 2019 no provee información sobre la identidad de género de las personas, y que tampoco ofrece más opciones que clasificar a las personas que con el sexo *hombre* o *mujer*. Por lo anterior, en este trabajo se hace referencia únicamente a esos dos términos, pero contar con una encuesta que aborde la identidad de género sería mucho más interesante y completo.

Algunos análisis de índole económica utilizan modelos de regresión para analizar la importancia de las características socio-económicas en un hecho o ente. En este trabajo se experimenta un análisis utilizando un modelo gráfico probabilístico, en particular, una red bayesiana. Se ha elegido dicho modelo, ya que como se verá más adelante, las redes bayesianas permiten visualizar relaciones entre características, calcular algunas probabilidades asociadas e incluso inferir consultas complejas.

Los modelos gráficos probabilísticos son herramientas estadísticas frecuentemente utilizadas en estudios o biológicos; definen una familia de distribuciones de probabilidad y permiten la representación de las mismas en una gráfica (2). En este trabajo se intenta llevar esta herramienta al campo económico para encontrar relaciones entre las características de las mexiquenses mayores de edad y su estado ocupacional, así como inferir algunas probabilidades.

Lo anterior, resulta en la conclusión de que el tratamiento de los datos juega un papel clave en el proyecto; el número de hijos y estado conyugal de las mujeres son los factores más influyentes en el perfil socio-económico de los mujeres; el estado ocupacional está directamente relacionado con la asistencia o no a la escuela de la población objetivo, así como las horas dedicadas al hogar y su necesidad de trabajar.

1.3. Objetivos del proyecto

El objetivo general de este proyecto es encontrar relaciones entre el empleo/desempleo y distintas variables socio-económicas con el fin de proporcionar información adecuada para el mejoramiento de oportunidades laborales para las mujeres del Estado de México.

Los objetivos específicos de esta tesis son:

- La investigación y entendimiento del modelo de Redes Bayesianas; así como las bases probabilísticas, estadísticas y algorítmicas de este.
- Adaptar las herramientas necesarias para la construcción de la red bayesiana a los datos de la Encuesta Nacional de Ocupación y Empleo.
- Obtener relaciones relevantes entre variables sociales y económicas de las mexiquenses e identificar posibles factores de influencia en su condición de ocupación.

Para alcanzar los objetivos antes mencionados se modelará el fenómeno socio-económico del desempleo de mujeres en el Estado de México con una Red Bayesiana. Lo anterior utilizando la Encuesta Nacional de Ocupación y Empleo (ENOE) del INEGI. Con lo que se espera obtener un grafo matemático direccionado y acíclico, que relacione las diferentes variables económicas y se esboce algunas probabilidades entre ellas.

Capítulo 2

Redes Bayesianas

El Aprendizaje Estadístico es un campo cuyo objetivo es modelar y analizar conjuntos de datos de diferentes disciplinas, tales como Medicina, Finanzas, Economía, Astronomía, Industria, etc. Este campo ha tenido un gran auge durante los últimos años y su naturaleza no sólo es estadística, también es computacional.

Los problemas de aprendizaje pueden dividirse en dos categorías: *aprendizaje supervisado* y *aprendizaje no supervisado*. En el primero de éstos, el problema se compone de variables que explican el problema (*input*) y una variable de salida o variable respuesta (*output*), que es el objeto de estudio y frecuentemente el fin es predecirla con tanta exactitud como se pueda. En el segundo, el problema sólo considera un conjunto de variables o características (*input*) y lo que se busca es describir la asociación o patrones entre éstas (James et al. (13)).

Existen diversas herramientas para la realización de tales tareas, desde métodos ya bien conocidos como la regresión lineal hasta los más recientes como lo son las redes neuronales. Este trabajo aborda una herramienta generalmente utilizada en la búsqueda de relaciones de dependencia y causalidad en un modelo con varias variables; así como de fácil interpretación por su naturaleza gráfica: las redes bayesianas, que son un tipo de modelo gráfico probabilístico.

2.1. Modelos Gráficos Probabilísticos

Los modelos gráficos probabilísticos son una rama del aprendizaje de máquina, cuyo objetivo es modelar fenómenos de diferente índole a través de distribuciones de probabilidad, para que la incertidumbre que pueda existir en dicho fenómeno, sea tomada en cuenta.

Un problema de interés puede ser identificar si un subconjunto de variables aleatorias es independiente de otro subconjunto, más aún, identificar si un subconjunto de variables aleatorias es independiente de otro dado que se conoce a un tercer subconjunto. Estas preguntas pueden contestarse si se cuenta con la función de probabilidad conjunta del

vector de variables aleatorias $(X_1, X_2, ..., X_n)$:

$$P(X_1 = x_1, ..., X_n = x_n) \tag{2.1}$$

En el caso de las variables aleatorias discretas, que son las de interés en este trabajo, se puede obtener una representación de su densidad conjunta como una tabla *n-dimensional*. Si cada variable X_i puede tomar r valores, esto quiere decir que se tendrían que calcular r^n probabilidades; en el mejor de los casos, si $r = 2$, se necesitarían 2^n números, lo cual es muy costoso de obtener y representa un problema para representar funciones de probabilidad conjuntas de fenómenos reales; así, una manera de manipular dichas probabilidades es simplificando su representación con suposiciones sobre los datos. La simplificación sugerida en el enunciado anterior es el fundamento de los modelos gráficos probabilísticos.

Recientemente, los modelos gráficos probabilísticos (MGP) se han convertido en una herramienta del Aprendizaje Estadístico muy popular para el análisis de datos. Se trata de modelos que representan de manera compacta e intuitiva la distribución conjunta de variables aleatorias y las relaciones entre éstas mediante grafos.

Un MGP define una familia de distribuciones de probabilidad que pueden ser representadas en términos de una gráfica. Los nodos son variables aleatorias y la estructura se traduce en dependencias estadísticas entre las variables; esto puede conducir al cálculo de probabilidades marginales, condicionales o conjuntas que sean de interés de una manera menos costosa que la antes mencionada.

Las variables aleatorias pueden ser usadas para expresar la variabilidad de una cantidad observada o bien, para especificar factores que, aunque no se pueden observar, afectan al resultado. Las aristas del grafo especifican cuáles factores afectan a otros, mientras que los parámetros correspondientes a las distribuciones de las variables aleatorias completan la información, (2).

Un modelo gráfico probabilístico (MGP) puede caracterizarse utilizando dos componentes, (24):

- Una gráfica $G(V, E)$, que define la estructura del modelo. Donde V y E son un conjunto de vértices y un conjunto de aristas, respectivamente.
- Un conjunto de funciones $f(X_i)$ que definen los parámetros del modelo.

Estos dos elementos son la base de los dos principales pilares en la construcción de los MGP's: Inferencia y Aprendizaje. El primero se refiere a la obtención de información dado que se cuenta con un MPG, en otras palabras, realizar consultas sobre probabilidades marginales o condicionales de eventos o variables de interés. El segundo pilar es la estimación de parámetros y estructura de la red dado que se conoce un conjunto de datos; que es ajustar un MPG a los datos.

Este tipo de modelos resalta por su fácil interpretación, eficiencia y por su versatilidad en la construcción de la misma, puesto que puede construirse con ayuda de datos o de algún experto en el área de estudio correspondiente.

Existen distintos tipos de modelos gráficos probabilísticos dependiendo de su fin y lo que se pretenda modelar (dirigidos o no dirigidos; estáticos o dinámicos; probabilísticos o de decisión). En este trabajo se abordarán las Redes Bayesianas estáticas y se aplicarán a la problemática de desempleo de mujeres en el Estado de México.

2.2. Teoría de Probabilidad

La probabilidad es, en pocas palabras, el estudio de la incertidumbre. En esta sección se proveen los conceptos y propiedades concernientes a la Teoría de Probabilidad, que serán útiles en la definición y desarrollo de las Redes Bayesianas.

2.2.1. Conceptos básicos

Debido a que es del interés de esta tesis trabajar con más de una variable aleatoria, se recuerda que dado un conjunto de variables aleatorias $X_1, X_2, ...X_n$ con dominios $\underline{X_1}, ..., \underline{X_n}$, respectivamente, podemos denotar su **función de distribución conjunta** como $P(X_1, ..., X_n)$ y cumple las siguientes propiedades:

- $P(X_1, .., X_n) \geq 0$
- $\sum_i \sum_{\underline{X_i}} P(X_1, .., X_n) = 1$

más aún, se define la **función de distribución acumulada** $F(X_1, ..., X_n)$ como:

$$F(X_1, ..., X_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$

asimismo puede definirse la **función de probabilidad condicional** de la variable X_1 dado que se conocen los valores de las variables $X_2, ..., X_n$ $(X_1|X_2, ..., X_n)$, como sigue:

$$P(X_1 = x_1|X_2 = x_2, ..., X_n = x_n) = \frac{P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)}{P(X_2 = x_2, ..., X_n = x_n)}$$

del mismo modo, es importante recordar que un conjunto de variables aleatorias $X_1, ...X_n$ son **independientes** $(X_1 \perp X_2 ... X_n)$ si y sólo si:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2), ..., P(X_n = x_n)$$

para todos los valores posibles de $X_1, ..., X_n$, a saber: $\underline{X_1}, ..., \underline{X_n}$, respectivamente. Más aún, la independencia de dos o más variables aleatorias puede verse también en términos de la probabilidad condicional, pues si $X_1 \perp \{X_2, ..., X_n\}$, entonces:

$$P(X_1 = x_1|X_2 = x_2, ..., X_n = x_n) = P(X_1 = x_1)$$

cuya interpretación es que cuando X_1 es independiente de $\{X_2, ..., X_n\}$, el conocer qué valor toman estas últimas variables, no nos brinda información acerca de X_1.

Para finalizar este apartado, se enuncia una propiedad crucial en la definición y desarrollo del modelo de Redes Bayesianas, a saber, **la regla de la cadena**, resultado de la definición de probabilidad condicional:

$$\begin{aligned} P(X_1, ...X_n) &= P(X_n|X_1, ..., X_{n-1})P(X_1, ..., X_{n-1}) \\ &= P(X_n|X_1, ..., X_{n-1})P(X_{n-1}|X_1, ..., X_{n-2})P(X_1, ..., X_{n-2}) \\ &\vdots \\ &= P(X_1)\prod_{i=2}^{n} P(X_i|X_1, ..., X_{i-1}) \end{aligned}$$

Este resultado es aplicado en la interpretación de las redes bayesianas, de hecho, más adelante se verá que debido a las propiedades de éstas, no es necesario calcular todo el producto de las probabilidades condicionales, sino unos cuantos factores.

2.2.2. Independencia condicional

El concepto matemático de Independencia Condicional juega un papel muy importante en la definición y uso de una Red Bayesiana, pues es gracias a esta que se puede extraer información de interés de ciertos datos sin tener que conocer todos los valores posibles de las variables a estudiar, como más a delante se verá.

Definición 2.1 (Independencia condicional (versión 1))**.** *Sea U un conjunto discreto de valores, sea $P(\cdot)$ una función de probabilidad conjunta sobre las variables en U, y sean X, Y y Z cualesquiera tres subconjuntos de variables en U. Se dice que X es condicionalmente independiente de Y dado Z si*

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z) \tag{2.2}$$

con $P(Y = y|Z = z) > 0$.

La intuición detrás de este concepto es que dado que se sabe que un evento Z ocurre, el conocimiento adicional de que Y también ha ocurrido no cambia la probabilidad del evento X, es decir, Y no aporta más información de la que ya se tenía conociendo Z. Un ejemplo sería: si X es el número de palabras que un niño conoce, Y es la estatura del niño y Z es la edad del niño, entonces note que si Y es grande, muy probablemente X lo es también, esa es la información que podría obtener de X dado que conozco Y. Por otro lado, si se conoce también a Z, aunque las tres variables estén relacionadas, el conocer la edad del niño hace irrelevante su estatura para saber el número de palabras que el niño conoce; Y no aporta a X información extra (y viceversa, como más adelante se verá), Y ya no cambia la probabilidad de X porque el factor que "unía" o creaba dependencia entre estas dos variables era la edad, entonces X e Y son condicionalmente independientes dado Z.

Otra definición de independencia condicional es la que sigue.

Definición 2.2 (Independencia condicional (versión 2))**.** *Sea U un conjunto discreto de valores, sea $P(\cdot)$ una función de probabilidad conjunta sobre las variables en U, y sean X, Y y Z cualesquiera tres subconjuntos de variables en U. Se dice que X es condicionalmente independiente de Y dado Z si*

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z) \tag{2.3}$$

para cualquier $P(Z = z) > 0$.

Esta segunda definición puede resultar más familiar y fácil de recordar, debido a su similitud con la definición de independencia simple.

Proposición 1. *Las definiciones de Independencia condicional versión 1 y versión 2 son equivalentes.*

Demostración. Versión 1 implica versión 2.

$$\begin{aligned} P(X, Y|Z) &= \frac{P(X, Y, Z)}{P(Z)} \\ &= \frac{P(X|Y, Z)P(Y, Z)}{P(Z)} \\ &= P(X|Z)P(Y|Z) \end{aligned}$$

La última igualdad se cumple dada la primera definición de independencia condicional. Falta ver que la versión 2 implica la versión 1.

$$\begin{aligned} P(X|Y, Z) &= \frac{P(X, Y, Z)}{P(Y, Z)} \\ &= \frac{P(X, Y|Z)P(Z)}{P(Y, Z)} \\ &= \frac{P(X|Z)P(Y|Z)P(Z)}{P(Y|Z)P(Z)} \\ &= P(X|Z) \end{aligned}$$

□

En lo que sigue se usará la notación:

$$(X \perp\!\!\!\perp Y|Z) \Leftrightarrow P(X = x|Y = y, Z = z) = P(X = x|Z = z) \tag{2.4}$$

para decir que la variable X es independiente de Y dado que se conoce Z.

La relación central entre la estructura de la red y este tipo de independencia, es que la estructura de la red bayesiana debe corresponder con las relaciones de independencia condicional dadas por la distribución de probabilidad conjunta y viceversa. Este resultado se discutirá más adelante.

Si P es un modelo probabilístico y se cumple que "X es independiente de Y dado que conocemos Z", la relación de independencia condicional (I) satisface algunas propiedades. Esto se enuncia de manera formal en el siguiente teorema.

Teorema 2.1 (Pearl 1988)**.** *Sean X, Y y Z tres subconjuntos de variables disjuntos de un conjunto U. Si se cumple $(X \perp\!\!\!\perp Y|Z)$ "X es independiente de Y dado Z" en algún modelo probabilístico P, entonces la relación I debe satisfacer las siguientes condiciones independientes.*

- *Simetría*

  $(X \perp\!\!\!\perp Y|Z) \Leftrightarrow (Y \perp\!\!\!\perp X|Z)$

- *Descomposición*

  $(X \perp\!\!\!\perp Y \cup W|Z) \Rightarrow (X \perp\!\!\!\perp Y|Z) \& (X \perp\!\!\!\perp W|Z)$

- *Unión débil*

  $(X \perp\!\!\!\perp Y \cup W|Z) \Rightarrow (X \perp\!\!\!\perp Y|Z \cup W)$

- *Contracción*

  $(X \perp\!\!\!\perp Y|Z) \& (X \perp\!\!\!\perp W|Z \cup Y) \Rightarrow (X \perp\!\!\!\perp Y \cup W|Z)$

Si P es estrictamente positiva, entonces se satisface también:

- *Intersección*

  $(X \perp\!\!\!\perp Y|Z \cup W) \& (X \perp\!\!\!\perp W|Z \cup Y) \Rightarrow (X \perp\!\!\!\perp Y \cup W|Z)$

La interpretación de la propiedad de Simetría es que si conociendo un evento Z, el evento Y no es relevante para otro X, tampoco X será relevante para Y si se conoce a Z. La Descomposición asegura que si dos eventos W, Y en combinación son irrelevantes para otro X, dado un evento Z; serán también irrelevantes para X por separado. La Unión débil y la Contracción afirman que la información irrelevante para un evento no cambia la relevancia de otros eventos. La propiedad de Intersección establece que cuando una variable Y no afecta a X cuando se conoce W y al mismo tiempo W no afecta a X cuando se conoce Y, entonces ni Y, ni W, ni su combinación pueden afectar X.

Las propiedades anteriores son llamadas *Axiomas de grafoides* (23). La razón por la que se usa Independencia condicional en el desarrollo de las redes bayesianas es la siguiente conjetura:

Teorema 2.2 (Conjetura de completitud (Pearl y Paz, 1985)). *El conjunto de las primeras cuatro propiedades listadas en el Teorema 2.1 es un sistema de axiomas completo cuando I se interpreta como la relación de independencia condicional. En otras palabras, para toda relación que satisfaga estas propiedades, existe un modelo de probabilidad P tal que:*

$$P(X|Y,Z) = P(X|Z) \Leftrightarrow (X \perp\!\!\!\perp Y|Z);$$

Además, si se cumple la propiedad de intersección, decimos que P es positiva.

Esta conjetura nos dice que la única manera en la que se cumplen las propiedades antes expuestas es que la relación en juego sea la de independencia condicional.

2.3. Teoría de Gráficas

Como ya se ha mencionado antes, en los modelos de Redes Bayesianas (y en general en los modelos gráficos proabilísticos), los nodos de una gráfica representan variables aleatorias y las aristas, cierta relación entre ellas. Por esta razón, es importante recordar conceptos básicos de gráficas, que juegan un papel importante en la concepción de redes bayesianas.

Sea $G = (V, E)$ una gráfica no vacía, donde V es el conjunto de vértices (o nodos) y E es el conjunto de aristas. Si las aristas de G son dirigidas, entonces se trata de una **gráfica dirigida**, también llamada **digráfica**; más aún a dichas aristas pueden nombrarse **arcos**.

Dentro de un grafo, pueden encontrarse distintos tipos de secuencias de aristas o vértices. Un **camino** en una gráfica es una secuencia de aristas tal que cada arista comienza con el vértice "final"de la arista anterior. Si además se tiene que el camino respeta la dirección de cada arista (en el caso de digráficas), entonces es un **camino dirigido**. A un camino dirigido cerrado se le llama **ciclo dirigido**, y si una gráfica dirigida G no cuenta con ningún ciclo dirigido, entonces es **acíclica**. Conociendo lo anterior, a continuación define el tipo de gráfica con el que se trabaja en las Redes Bayesianas.

Definición 2.3 (Gráfica Acíclica Dirigida). *Gráfica cuyas aristas son dirigidas y no contiene ciclos dirigidos.*

Adicionalmente se define para cada nodo su conjunto de padres y se hace visible la relación padre-hijo entre nodos que se conectan directamente.

Definición 2.4 (Padre/Hijo). *En una gráfica $G = (V, E)$, si existe un arco dirigido del nodo A al B ($A \longrightarrow B$), A es padre de B y B es hijo de A.*

Al conjunto de nodos padre de un nodo A, se le denota de aquí en adelante como $PA(A)$. Siguiendo la lógica anterior también se definen los ancestros y descendientes de un nodo.

Definición 2.5 (Ancestro/Descendiente)**.** *Decimos que el nodo A es ancestro de B en una gráfica $G = (V, E)$ y B es descendiente de A si existe un camino dirigido $A_1, ..., A_k$ tal que $A_1 = A$ y $A_k = B$.*

La relevancia de estos conceptos recae en el rol que las gráficas tienen a la hora de proveer una representación de las variables que son relevantes para otro conjunto de ellas.

Hasta ahora, se han definido las gráficas acíclicas no dirigidas y se ha hablado de la independencia condicional y porqué es esta la relación usada en las redes bayesianas. Es momento de introducir el puente de unión entre la topología de una gráfica y la relación de independencia condicional.

2.4. Representación

En esta sección se realiza la conexión entre la Teoría de Probabilidad y la Teoría de Gráficas, pues las redes bayesianas son modelos de gráficas acíclicas dirigidas (DAG, por sus siglas en inglés), que como se verá más adelante, representan la función de probabilidad conjunta de ciertas variables aleatorias. Varios autores afirman que una red bayesiana queda definida (al igual que otros modelos probabilísticos) por su estructura (S) y sus parámetros (P), de tal modo que usan la notación $RB(S, P)$ para definir una red bayesiana.

La importancia de estos modelos recae en que para representar una función de distribución conjunta para un modelo con variables discretas es necesario construir una tabla de probabilidades que tendrá n dimensiones; suponiendo que cada variable es binaria, la complejidad computacional de obtener dicha información será $O(2^n)$, lo cual es muy costoso o hasta infactible dependiendo del número de variables. Las redes bayesianas ofrecen una solución simple a esta problemática basada en restringir las relaciones entre las variables y de dicha manera, reducir los cálculos; de tal manera que permiten representar y expresar relaciones de independencia entre los datos.

Así, en una red bayesiana, dada una DAG las variables del modelo son representadas como los nodos de la gráfica, mientras que los arcos denotan la existencia de influencia entre estas. De este modo, si existe un arco del nodo A al nodo B en la gráfica, esto se traduciría en una influencia causal o una dependencia directa (la interpretación varía dependiendo del problema) de la variable representada por el nodo A hacia la variable representada por el nodo B. La intensidad de esta influencia es expresada mediante un conjunto de probabilidades condicionales asociadas a la red.

En la Figura 2.1 se muestra un ejemplo de red bayesiana con cinco variables, recuperada del libro de Sucar (24). La interpretación sería que *Fiebre* depende de *Tifoidea* y *Gripe*, pero no existe dependencia directa entre *Fiebre* e ingerir *Alimentos contamina-*

dos dado que ya se tiene gripe o tifoidea. Por su parte, *Tifoidea* y *Gripe* podrían estar relacionados dependiendo de si se conoce o no el valor de *Fiebre*.

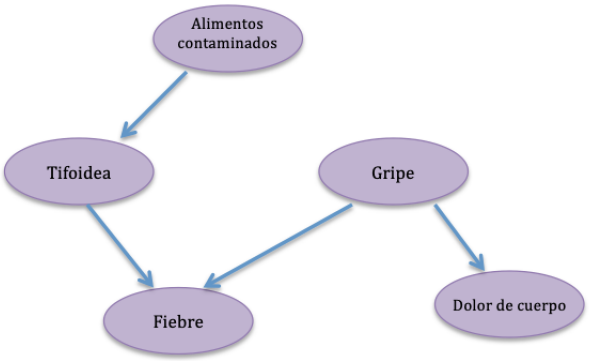


Figura 2.1: Ejemplo de red bayesiana.

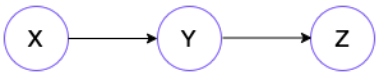
Para enlazar las gráficas y la probabilidad en términos del razonamiento bayesiano, es necesaria la invención de un concepto de "independencia" en las gráficas, a saber: el criterio de *d-separación*.

2.4.1. D-separación

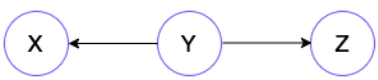
Las redes bayesianas tienen dos componentes fundamentales: la estructura de la red y las relaciones de probabilidad ligadas a esta. Anteriormente se mencionaba que la relación de independencia condicional será la base para calcular probabilidades, pero ¿De qué manera se puede relacionar esto con una red? El criterio de *d-separación* (por *directed separation*), permite determinar por inspección cuales conjuntos de variables son consideradas independientes dado un tercer conjunto.

Antes de definir dicho criterio, es necesario notar que dados dos arcos y tres variables a modo de nodos en una red bayesiana, se pueden definir tres estructuras básicas para el nodo Y:

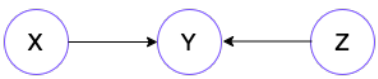
- Secuencial:



- Divergente:



- Convergente:



Aclarado esto, el criterio de *d-separación* establece que dos subconjuntos de nodos están *d-separados* si dado un tercer conjunto, para toda trayectoria en los dos principales sucede que: si un nodo es convergente, ni éste ni sus descendientes pertenecen al subconjunto sobre el que se condiciona; o no es convergente pero sí pertenece al conjunto sobre el que se condiciona. Formalmente:

Definición 2.6 (Criterio d-separación)**.** *Si X, Y y Z son subconjuntos de nodos disjuntos en una DAG, entonces Z* ***d-separa*** *a X de Y si a través de cualquier trayectoria entre un nodo en X y otro en Y , existe un nodo w que satisface alguna de las dos siguientes condiciones:*

- *w es convergente y ni w ni sus descendientes están en Z.*
- *w no es convergente y w está en Z.*

Ejemplo. En la Figura 2.1 el nodo **1** *d-separa* a los nodos **2** y **3**, pues $\mathbf{Z = \{1\}}$ no es convergente y es claro que $\mathbf{1} \in Z$. Por otro lado, **2** y **3** no están *d-separados* por $\mathbf{Z = \{4\}}$ debido a que éste último es un nodo convergente y claramente pertenece al conjunto Z.

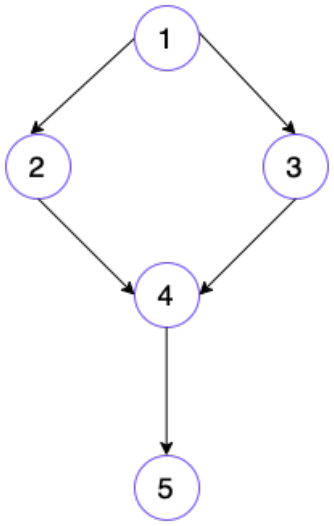


Figura 2.2: Ejemplo de gráfica acíclica dirigida

El criterio de *d-separación* es a una DAG, lo que la Independencia Condicional a un conjunto de variables aleatorias. En otras palabras, hay equivalencia entre las relaciones de independencia condicional entre las variables del modelo y la *d-separabilidad* dada la estructura de la red.

Para ilustrar esta última idea, nos auxiliaremos del **Algoritmo de la bola de Bayes**, que es una manera simple de aplicar la *d-separación* a una gráfica.

Si se considera una gráfica G dirigida y conexa, con nodos X, Y y Z, tal como se muestra a continuación. Y se considera una bolita que puede moverse entre los nodos de

G, entonces las reglas del algoritmo son las siguientes:

Reglas del Algoritmo de la bola de Bayes

1. Si Y es secuencial o divergente y no está sombreado, la bola puede pasar.
2. Si Y es secuencial o divergente y está sombreado, la bola no puede pasar (está bloqueada).
3. Si Y es convergente y no está sombreado, la bola se bloquea.
4. Si Y es convergente y está sombreado, la bola pasa.

Hay que mencionar que entenderemos a los nodos X, Y y Z como variables aleatorias. Si un nodo es sombreado, es porque la variable asociada a él es conocida y si la bola puede pasar de un nodo a otro, estos nodos son condicionalmente dependientes. De este modo, se tienen tres casos básicos en el análisis de independencia condicional de tres variables.

Caso 1. (Y es secuencial) La bola va de X a Y, pero si Y está sombreado, no pasa a Z, lo que nos diría que $X \perp\!\!\!\perp Z|Y$. Similarmente si la bola va de Z a Y, no puede pasar a X. Por el contrario, si Y no está sombreado, la bola sí puede pasar libremente y concluiríamos que $X \not\perp\!\!\!\perp Z|Y$.

Figura 2.3: Algoritmo cuando Y es secuencial

Como un ejemplo concreto para este caso podemos tomar X como el pasado, Y el presente y Z el futuro; por lo que la interpretación sería que dado que conocemos el presente, el futuro no depende del pasado (la propiedad de Markov).

Caso 2. (Y es convergente) Si Y está sombreado, la bola no pasa de X a Z. Si no está sombreada, sí puede pasar.

Figura 2.4: Algoritmo cuando Y es convergente.

En este caso, tomemos a X como la medida de zapato de un individuo, Z la cantidad de canas y Y la edad. Sabemos que X y Z dependen fuertemente una de otra, pero cuando se conoce la edad Y, la medida de zapato X ya no aporta nueva información a la cantidad de canas Z que el individuo pueda tener, y viceversa; por lo que se vuelven independientes condicionalmente.

Caso 3. (Y divergente) Siguiendo las reglas, si Y está sombreada y considerando que es un vértice divergente, la bola está bloqueada. Si Y no está sombreada, entonces la bolita pasa.

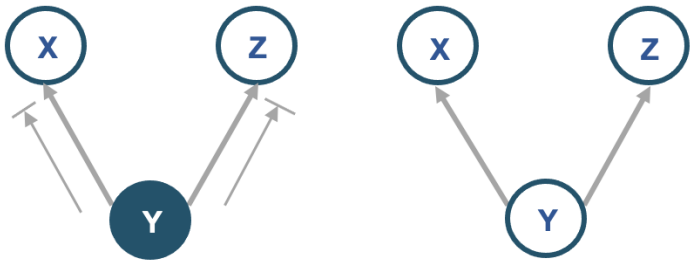


Figura 2.5: Algoritmo cuando Y es divergente

Este caso es especialmente interesante, pues nos presenta un escenario en el que dos variables X y Z son independientes, pero cuando conocemos la información de Y, se vuelven dependientes.

Si por ejemplo, un sujeto *Kevin* quedó de reunirse con *Gema* a las 12:00pm, llamemos **"tarde = sí"** al evento de que Gema llegue tarde a la reunión. Una explicación al evento de que Gema llegue tarde podría ser que Gema fue secuestrada por aliens miestras venía. Utilizando el teorema de Bayes, Kevin descubrió que la probabilidad de que los aliens hayan secuestrado a Gema aumenta dado que Gema no ha llegado, $P(aliens = si) < P(aliens = si|tarde = si)$. Sea **"reloj = no"** el evento de que Kevin olvidó retrasar su reloj por el cambio de horario de verano. Entonces Kevin ahora calcula $P(aliens = si|tarde = si, reloj = no)$ y encontró además que $P(aliens = si|tarde = si) > P(aliens = si|tarde = si, reloj = no)$. Como estas dos últimas probabilidades son diferentes, podemos concluir que **aliens** y **reloj** son depen-

dientes dado el hecho de que Gema llegue o no tarde.

El último caso da en el clavo a un tipo de eventos de dependencia condicional que aunque no son tan intuitivos gráficamente, pueden llegar a presentarse dentro de un modelo; eventos que por sí solos son independientes, pero cuando se conoce el resultado de un tercer evento, esta nueva información los hace dependientes.

2.4.2. Relación DAG-Modelo

Teniendo el problema de expresar la función conjunta de n variables discretas, así como poder obtener funciones marginales, en esta sección se relaciona al modelo probabilístico con una estructura de grafo. El propósito es ligar una DAG con una distribución de probabilidad, que además contiene suposiciones sobre independencia entre las variables.

Dada una función de probabilidad P definida sobre n variables y suponiendo un orden σ en éstas; $X_1, X_2, ..., X_n$. Sabemos que la regla de la cadena para probabilidad nos permite expresar la distribución conjunta de la siguiente manera:

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_j P(X_j = x_j | X_1 = x_1, ..., X_{j-1} = x_{j-1})$$

Si además se tiene que para una variable X_j, existe un subconjunto PA_j de $X = \{X_1, ..., X_n\}$, tal que una vez que los valores de PA_j son conocidos, cualquier otro antecesor ya no tiene influencia en X_j, i.e. PA_j hace condicionalmente independiente a X_j de las otras variables; entonces, podemos escribir:

$$P(X_j | X_1, ..., X_{j-1}) = P(X_j | PA(X_j))$$

Esta igualdad simplifica el cálculo de la probabilidad condicional requerida a través de la regla de la cadena, pues al tener independencias condicionales entre las variables, el producto a calcular se reduce. Al conjunto $PA(X_j) = PA_j$ se le conoce como los padres markovianos de X_j, dicho conjunto también puede ser denotado como π_j.

La existencia de un conjunto PA_j de variables que hagan que X_j sea independiente de otros ancestros en el orden σ de X_j puede ser ambigua, en el sentido de que puede agregarse más variables y seguir formando un conjunto que hace condicionalmente independiente a X_j de otros ancestros; en incluso puede ser tan grande como el conjunto original de variables, X. Por dicha razón se define al conjunto minimal de variables como *Padres Markovianos.*

Definición 2.7 (Padres Markovianos)**.** *Sea $X = \{X_1, ..., X_n\}$ un conjunto de variables con un orden σ y sea $P(X)$ la probabilidad conjunta de estas variables. Se dice que un conjunto PA_j es el conjunto de padres markovianos de X_j si PA_j es el conjunto minimal de predecesores tal que hace que X_j sea independiente de sus otros predecesores. En otras palabras, PA_j cumple que:*

$$P(X_j|PA(X_j)) = P(X_j|X_1, ..., X_{j-1})$$

y no existe subconjunto propio de PA_j que también cumpla esta condición.

el conjunto de padres markovianos de una variable X_j nos permite no considerar a los antecesores que no aportan nueva información en el cálculo de la probabilidad de X_j.

Este concepto se generaliza, de manera que permite no sólo *separar* los ancestros de X_j que no aporten información extra a la variable en cuestión, sino también considerar a los descendientes de dicha variable, y así, definir un conjunto de variables que si son conocidas, permitan independizar condicionalmente a X_j de las restantes. Esta generalización del conjunto se llama *Manta de Markov.*

Definición 2.8 (Manta de Markov)**.** *Sea G una DAG con nodos $X_1, ..., X_n$, la manta de Markov de un nodo X_j, $MB(X_j)$, es el conjunto de nodos en G tales que hacen X_j es independiente de todos los otros nodos en G. Esto es:*

$$P(X_j|X_1, ..., X_{j-1}, X_{j+1}, ...X_n) = P(X_j|MB(X_j))$$

La *Manta de Markov* de un nodo X_j, para una red bayesiana es:

Yendo algunos pasos adelante, dada una distribución P de X y su representación gráfica G, existe una relación de correspondencia entre la independencia condicional en P y en G. La siguiente definición marca la relación entre el criterio de *d-separación* en una DAG y la independencia condicional en un modelo.

Definición 2.9 (I-mapa)**.** *Una gráfica acíclica dirigida, D, es un I-mapa de un modelo M si toda d-separación en D corresponde a una relación de independencia condicional válida en M. Una DAG es un I-mapa minimal si al borrar alguno de sus arcos, ya no es $I - mapa$.*

En un *I-mapa* todas las relaciones de independencia condional en la gráfica G son ciertas para la distribución de probabilidad P (pero no al revés). De este modo queda bien definido el conjunto de independencias asociadas a una distribución P y con ello se define formalmente a una red bayesiana.

Definición 2.10 (Red bayesiana)**.** *Dada una distribución de probabilidad P sobre un conjunto de variables U una DAG $D = (U, \overrightarrow{E})$, donde $\overrightarrow{E}$ es un conjunto de arcos, es una red bayesiana de P si y sólo si es un $I - mapa$ minimal de P.*

Cualquier nodo X_j de la red es condicionalmente independiente de todos los nodos en la gráfica G que no se encuentren en la manta de Markov de X_j, más aún, es condicionalmente independiente de todos aquellos nodos de los que se pueda *d-separar*. Aunque

no siempre es posible obtener un "mapeo" perfecto de las relaciones de independencia entre la gráfica G y la distribución P, basta con encontrar un *I-mapa* minimal.

Pearl (23) expone resultados importantes para la construcción de una red bayesiana, algunos se enuncian aquí.

Proposición 2. *Sea una distribución de probabilidad $P(X_1, ..., X_n)$ y σ cualquier ordenamiento de las variables. La DAG obtenida de asignar como nodos padres de X_i a algún conjunto minimal Π_{X_i} de predecesores que satisfacen:*

$$P(x_i|\Pi_{X_i}) = P(X_i|X_1, ..., X_{i-1}), \Pi_{X_i} \subseteq X_1, ..., X_{i-1} \tag{2.5}$$

es una red Bayesiana de P. Si P es estrictamente positiva entonces todos los conjuntos de padres son únicos y la red bayesiana es única dado el ordenamiento σ.

Esta proposición ofrece una caracterización de las redes bayesianas a través de la construcción de padres markovianos de las variables y su representación como nodos padres de los nodos correspondientes en la red.

Por último, cabe mencionar que el criterio de *d-separación* determina la Manta de Markov para cualquier nodo X de una red bayesiana, pues ésta queda completamente definida con los descendientes y padres directos de X, así como los padres de sus descendientes directos.

En (23) y (15) se desarrollan a profundidad las bases matemáticas necesarias para la construcción y uso de las redes bayesianas.

2.5. Aprendizaje de la estructura de la red bayesiana

Las redes bayesianas se componen de su estructura y sus parámetros, que serán denotados como B_S y B_P. Algo característico de estos modelos es que la estructura de la red puede ser diseñada por expertos en el campo del problema que se quiere modelar; pero algunas veces realizar este diseño no es asequible (por la complejidad del problema, porque no hay una persona experta que pueda ayudar, etc.); en ese caso, la manera de definir la estructura de la red es aprendiéndola mediante los datos.

La tarea de generar la estructura a través del conjunto de datos es un problema *NP-difícil* [1] (24). Existen algoritmos con este objetivo, siendo uno de los más populares el Algoritmo K2, introducido por Gregory E. Cooper y Edward Herskovits en 1992 que se detalla en la sección siguiente.

[1] Un problema *NP-difícil* se distingue por ser al menos tan difícil como cualquier problema del conjunto *NP* o más difícil.

2.5.1. Algoritmo K2

El objetivo del Algoritmo K2 es encontrar la estructura de la gráfica cuya probabilidad sea máxima, en otras palabras, encontrar la estructura que se ajusta mejor a los datos.

Sea D un conjunto de datos y B_{S_i} la red bayesiana con la estructura S_i, se interpreta $P(B_{S_i}|D)$ como la probabilidad de que la red tenga la estructura B_{S_i} dada la información del conjunto de datos D. Así, se busca que $P(B_{S_i}|D)$ se maximice para alguna estructura S_i. Cabe mencionar que B_{S_i} contiene sólo a las variables de la base de datos D. Cooper and Herskovits (7) proponen una metodología para obtener $P(B_S, D)$ tras suponer lo siguiente:

a) Las variables de D son discretas.

b) Las observaciones en D son independientes dado el modelo de red bayesiana.

c) No hay datos faltantes en D.

d) Todas las estructuras S_i son igualmente probables si no se conocen los datos D.

Bajo las premisas anteriores, se calcula $P(B_S, D)$:

$$\begin{aligned} P(B_S, D) &= \int_{B_P} P(B_S, D, B_P) dP \\ &= \int_{B_P} P(D|B_S, B_P) P(B_S, B_P) dP \\ &= \int_{B_P} P(D|B_S, B_P) f(B_P|B_S) P(B_S) dP \end{aligned} \tag{2.6}$$

donde B_P es un vector que representa los valores condicionales asociados a la estructura B_S. Además, dada la suposición a), se asume que $P(D|B_S, B_P)$ es una función de masa.

De la segunda suposición, donde se especifica que ya que se conoce la red bayesiana, las observaciones son independientes (independencia condicional entre las observaciones, dada la red); se obtiene la siguiente expresión:

$$P(D|B_S, B_P) = \prod_{h=1}^{m} P(C_h|B_S, B_P) \tag{2.7}$$

donde C_h es la observación o caso *h-ésimo* del conjunto de datos D; y m el número de observaciones.La expresión de $P(B_S, D)$ se modifica de la siguiente forma:

$$P(B_S, D) = \int_{B_P} [\prod_{h=1}^{m} P(C_h|B_S, B_P)] f(B_P|B_S) P(B_S) dP. \tag{2.8}$$

El desarrollo posterior de la ecuación (2.6) se lleva a cabo en el Apéndice de Cooper and Herskovits (7); llegando a la expresión:

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2.9}$$

Donde r_i es el número de valores que la variable X_i puede tomar, N_{ijk} es el número de observaciones en la base de datos tales que la variable X_i toma el valor específico v_{ik} y el conjunto de padres está instanciado como w_{ij}; considerando que X_i puede tomar los valores $v_{i1}, ..., v_{ir_i}$; q_i es el número de instancias posibles para el conjunto de padres de X_i y N_{ij} es el número de casos en la base de datos donde $PA(X_i = j)$.

Como consecuencia de asumir que las estructuras B_S son equiprobables, es decir $P(B_S) = c$, entonces puede expresarse $P(B_S, D)$ como:

$$P(B_S, D) = c \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2.10}$$

Se hace hincapié en el hecho de que el "parámetro"que se puede optimizar es el número de padres en cada nodo, por lo que basta que para maximizar $P(B_S, D)$ se maximice el segundo producto, es decir:

$$\max_{B_S} P(B_S, D) = c \prod_{i=1}^{n} \max_{\pi_i} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2.11}$$

recordando que π_i es una forma de denotar a los nodos padre de la variable X_i.

Con el desarrollo anterior y los supuestos antes especificados, se construye el algoritmo *K2*, cuyo procedimiento general es recibir el conjunto de datos con un cierto orden, y encontrar la red bayesiana más probable para este conjunto. Se comienza asumiendo que ningún nodo tiene padres y éstos se van añadiendo gradualmente, eligiendo aquellos con los que la red resultante alcance su mayor probabilidad. El pseudocódigo del algoritmo se muestra a continuación.

Dado que este algoritmo da por hecho que se tiene un ordenamiento preestablecido de los nodos, $Pred(X_i)$ se define como el conjunto de nodos predecesores (o variables predecesoras) de la variable X_i en dicho ordenamiento.

Pseudocódigo del Algoritmo K2, Cooper 1992

Input: Un conjunto de n nodos ordenados, un límite superior, u, es el número de padres que un nodo puede tener. Una base de datos con m registros.
Output: El conjunto de padres para cada nodo.

```
1  Se inicia i = 1
2  for i = 1 a n do
3  |  π_i = {}
4  |  P_anterior = g(i, π_i)
5  |  bandera = True
6  |  while bandera y |π_i| < u do
7  |  |  sea z un nodo en Pred(X_i) − π_i que maximiza  g(i, π_i ∪ z)
8  |  |  P_nueva = g(i, π_i ∪ z)
9  |  |  if P_nueva > P_anterior then
10 |  |  |  P_anterior = P_nueva
11 |  |  |  π_i = π_i ∪ z
12 |  |  else
13 |  |  |  bandera = False
14 |  end
15 |  asignar π_i  al nodo x_i
16 end
```

En este procedimiento la función $g()$ se define en Cooper and Herskovits (7) como:

$$g(i, \pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2.12}$$

Para el caso específico de este trabajo, es necesario realizar ligeras modificaciones al algoritmo, para tomar en cuenta al Factor de Expansión de la ENOE (FAC), se cual se abordará a detalle en el Capítulo 3. Se definen

$$M = \{\text{Observaciones donde } X_i = k \text{ y } PA(X_i) = j\}$$

y

$$\Lambda = \{\text{Observaciones en el conjunto de datos, donde } PA(X_i) = j\}.$$

De modo que el cálculo de N_{ijk} se realiza:

$$N_{ijk} = \sum_{\mu \in M} FAC_\mu$$

Y de forma similar:

$$N_{ij} = \sum_{\lambda \in \Lambda} FAC_\lambda$$

Para concluir esta sección, queda decir que el algoritmo K2 es sumamente dependiente y sensible al orden de los nodos que recibe; es decir, no es robusto ante el orden

de las variables. El orden (σ) puede ser dado por alguna persona especializada en el tema; pero cuando se tiene un gran número de variables esto no es asequible y se utilizan algoritmos computacionales.

2.5.2. Orden de variables a través de la Entropía

Como se menciona en la sección anterior, el algoritmo de aprendizaje de la red K2, depende del orden en el que sean introducidos los nodos; ya que con cada ordenamiento distinto, la topología de la red será distinta también.

Aghdam et al. (1) experimentan con algunos de estos algoritmos, llegando a la conclusión empírica de que los mejores fueron los basados en la entropía de las variables. A continuación se exponen brevemente los conceptos de Entropía y Entropía Condicional.

El concepto de entropía, que inició como un desarrollo de la termodinámica, se lleva al campo de la probabilidad y estadística gracias a E.C. Shannon, caracterizada como la cantidad de información procesada en una fuente estocástica de datos.

Definición 2.11 (Entropía)**.** *La entropía de una variable aleatoria discreta X y con soporte $\underline{X}$, se define como:*

$$H(X) = -\sum_{x \in \underline{X}} p(x) \log p(x) \tag{2.13}$$

lo que es equivalente a:

$$H(X) = E[\log \frac{1}{p(x)}] = -E[\log(p(x))] \tag{2.14}$$

y refleja la incertidumbre, en promedio, de que la variable X tome uno u otro valor de su soporte $\underline{X}$.

La demostración de la ecuación (2.14) se sigue del siguiente teorema de probabilidad, informalmente nombrado en muchas ocasiones como el Teorema del estadístico inconsciente:

Teorema 2.3. *Sea X una variable aleatoria discreta y sea $f(X)$ una variable aleatoria con esperanza finita, entonces:*

$$E(f(x)) = \sum_{x \in \underline{X}} p(x) f(x) \tag{2.15}$$

Si se toma en cuenta que $f(x) = \log(p(x))$, entonces se comprueba que:

$$E[f(x)] = \sum_{x \in \underline{X}} p(x) \log(p(x)) = -H(X)$$

Ejemplo.

Si X es una variable aleatoria tal que:

$$X = \begin{cases} 1 & \text{con probabilidad} \quad p \\ 0 & \text{con probabilidad} \quad 1-p \end{cases}$$

entonces la entropía se calcula de la forma:

$$H(X) = -(p \log p + (1-p)\log(1-p))$$

Hay que hacer notar que la entropía depende únicamente de la distribución de probabilidad de la variable X, y no de los valores que toma. Además es no negativa, pues $p(x) \in [0,1]$ para cualquier valor que tome x; por lo anterior la desigualdad $\log(p(x)) \leq 0$ se cumple siempre.

El concepto anterior se extiende a dos variables aleatorias.

Definición 2.12 (Entropía conjunta)**.** *la entropía conjunta de dos variables X, Y con soportes $\underline{X}$ e $\underline{Y}$, respectivamente, se define como:*

$$H(X,Y) = -\sum_{x \in \underline{X}} \sum_{y \in \underline{Y}} p(x,y) \log(p(x,y)) \tag{2.16}$$

que puede verse también como:

$$H(X,Y) = -E[p(x,y)\log(p(x,y))] \tag{2.17}$$

y representa la incertidumbre de las variables X e Y de manera conjunta.

Finalmente, se define la entropía condicional de una variable dada otra de la siguiente forma.

Definición 2.13 (Entropía condicional)**.** *Para dos variables aleatorias X e Y con soportes $\underline{X}$ e $\underline{Y}$, respectivamente, la entropía condicional es:*

$$H(X|Y) = -\sum_{x \in \underline{X}} \sum_{y \in \underline{Y}} p(x|y) \log p(x|y), \tag{2.18}$$

que es equivalente a:

$$H(X|Y) = -E[\log(p(x|y))];$$

esto puede demostrarse con un procedimiento similar al de la demostración para una variable. Más aún, la entropía condicional cumple también la siguiente equivalencia:

$$H(X|Y) = H(X,Y) - H(X) \tag{2.19}$$

que se puede generalizar a la siguiente proposición:

Proposición 3. *Sean $X_1, ..., X_n$ variables aleatorias, se define X^{-i} como $\{X_j\}_{i\in\{1,...,n\}-\{i\}}$ y se denota a $\{X_1, ..., X_n\}$ como X^*; entonces:*

$$H(X_i|X^{-i}) = H(X^*) - H(X^{-i})$$

Demostración:

$$\begin{aligned} H(X_i|X^{-i}) &= \sum_{\underline{X_n}} \sum_{\underline{X_{n-1}}} \cdots \sum_{\underline{X_1}} P(X^*) \log\big(P(X_i|X^{-i})\big) \\ &= \sum_{\underline{X_n}} \cdots \sum_{\underline{X_1}} P(X^*) \log \frac{P(X^*)}{P(X^{-i})} \\ &= \sum_{\underline{X_n}} \cdots \sum_{\underline{X_1}} P(X^*)[\log P(X^*) - \log P(X^{-i})] \\ &= -\left[\sum_{\underline{X_n}} \cdots \sum_{\underline{X_1}} P(X^*) \log P(X^*) - \sum_{\underline{X_n}} \cdots \sum_{\underline{X_1}} P(X^*) \log P(X^{-i})\right] \end{aligned} \tag{2.20}$$

en la segunda igualdad se utiliza la definición de probabilidad condicional, mientras que la tercera igualdad resulta de aplicar una propiedad del logaritmo; la última igualdad se justifica con la propiedad distributiva de los números reales.

Por otro lado, desarrollando el segundo sumando:

$$\begin{aligned} -\sum_{\underline{X_n}} \cdots \sum_{\underline{X_1}} P(X^*) \log P(X^{-i}) &= -\sum_{\underline{X_n}} \cdots \sum_{\underline{X_{i+1}}} \sum_{\underline{X_{i-1}}} \cdots \sum_{\underline{X_1}} \log\big(P(X^{-i})\big) \sum_{\underline{X_i}} P(X^*) \\ &= -\sum_{\underline{X_n}} \cdots \sum_{\underline{X_{i+1}}} \sum_{\underline{X_{i-1}}} \cdots \sum_{\underline{X_1}} \log\big(P(X^{-i})\big) P(X^{-i}) \\ &= H(X^{-i}) \end{aligned}$$

Dado lo anterior, y de acuerdo con la última igualdad de (2.20), se obtiene que:

$$H(X_i|X^{-i}) = H(X^*) - H(X^{-i}).$$

La entropía condicional mide cuánta incertidumbre hay sobre la variable X una vez que ya se conoce el valor de la variable Y.

Aghdam et al. (1) hacen uso de diferentes métricas para obtener un orden de variables para el algoritmo *K2*, uno de los más efectivos para sus experimentos fue la entropía condicional.

Se calcula $H(X_i)$ para cada $i \in \{1, ..., n\}$ y el punto inicial será la i para la cual la entropía sea menor, supongamos $H(X_j)$. En cada paso posterior se condiciona sobre las variables elegidas anteriormente (en conjunto). Y se va eligiendo siempre la variable cuya entropía condicional sea menor. El orden en el que se vayan seleccionando las variables, será el ordenamiento deseado que se introducirá a *K2*.

Cabe mencionar que en la programación del algoritmo de ordenamiento, se utilizó (2.19); además, el logaritmo fue tomado en base 2. Por último, el Factor de Expansión de

la ENOE fue considerado en los cálculos de las entropías marginal, condicional y conjunta. Con lo anterior, se describe el algoritmo de este trabajo en el siguiente pseudocódigo.

Pseudocódigo de ordenamiento de variables usando entropía condicional

Input: Conjunto de datos D con n variables.
Output: Lista de variables l ordenadas.

```
ordenamiento = [],
entropias = [],
vars = {1, ..., n}
for i in vars do
    entropias = entropias.add(H(X_i))
end
ind_min = index(min(entropias))
ordenamiento = [ind_min]
vars = vars − {ind_min}
while |vars| > 1 do
    entropias = []
    for j in vars do
        H_conjunta = H(ordenamiento, X_j)
        entropias = entropias.add(H_conjunta − H(X_j))
    end
    ind_min = indice(min(entropias))
    ordenamiento = ordenamiento.add(ind_min)
    vars = vars − {ind_min}
end
return ordenamiento
```

Una vez obtenido el orden σ de las variables $\{X_1, ..., X_n\}$ puede llevarse a cabo el algoritmo K2.

2.6. Aprendizaje de parámetros en la red bayesiana

Recordando la idea de que una red bayesiana se define a través de su estructura y sus parámetros, y dado que en la sección se abordó cómo construir la estructura de la red, en esta sección se plantean las formas más comunes de aprender los parámetros de la red bayesiana y se ahonda en el uso de la máximo-verosimilitud.

El aprendizaje de los parámetros de la red se refiere a la obtención de la tabla de probabilidades condicionales de cada variable de la red (nodo) dado que se conocen sus padres, a esta tabla comúnmente se le llama **CPT** *(Conditional Probability Table)* y guarda la información de las relaciones entre las variables, siempre que estas sean discretas.

Zhiwei, et al (14) abordan los métodos de aprendizaje de redes desde dos perspectivas: los métodos para datos completos y para datos incompletos. Para los datos com-

pletos algunas técnicas conocidas son la estimación con Máximo-Verosimilitud (MLE) y el Método Bayesiano; mientras que para los datos incompletos se tienen la estimación con método Monte-Carlo, EM (Expectation-maximization), entre otros (14).

En este trabajo son de interés los métodos para datos completos, debido a que el tema de datos faltantes se resuelve con la imputación de datos MICE; mencionada en el Apéndice B.

2.6.1. Aprendizaje de parámetros con Máximo-Verosimilitud

Suponer que se tienen datos completos significa dar por hecho que no falta información en el conjunto de datos que se está utilizando y que no existen variables latentes. A continuación se presenta el método para aprender los parámetros de una red bayesiana basado en la Máximo-Verosimilitud.

La función de verosimilitud es la probabilidad de algún evento aleatorio dado un parámetro θ. Dada una muestra que se asume proveniente de un evento aleatorio; y suponiendo que dicha muestra sigue alguna distribución, se buscan los parámetros con los cuales la distribución hipótesis se ajusta mejor a los datos, o dicho de otra manera; se maximice la probabilidad de que la muestra tome una forma como la de hipótesis. Por lo anterior, el objetivo al ajustar una distribución de probabilidad es maximizar la verosimilitud.

El uso de la verosimilitud es llevado al estudio y construcción de redes bayesianas, además de ser uno de los métodos más comunes y simples de ejecutar.

En el entendido de que se tiene el conjunto de nodos $X = \{X_1, X_2, ..., X_n\}$ y se tiene al conjunto $D = \{d_1, d_2, ..., d_m\}$ de observaciones, donde $d_i = \{x_{1i}, x_{2i}, ..., x_{ni}\}$ indica la i-ésima muestra de las variables; la verosimilitud será denotada como:

$$L(\theta; \mathbf{D}) = P(\mathbf{D}|\theta) := P(D)$$

La siguiente proposición establece que en realidad, la Máxima-Verosimilitud para los parámetros de una Red Bayesiana se alcanza utilizando como parámetros (probabilidades) las frecuencias relativas de los datos.

Proposición 4. *P(D) es máxima si y sólo si:*

$$P_X(x|y) = \frac{|\{d \in D | d_X = x, d_{PA(X)} = y\}|}{|\{d \in D | d_{PA(X)} = y\}|}; \tag{2.21}$$

suponiendo que existe $d \in D$ *tal que* $d_{PA(X)} = y$. *En otro caso, puede elegirse* $P_X(x|y)$ *arbitrariamente, ya que* $P(D)$ *no depende de ello.*

Demostración. Se sigue de la independencia de las observaciones en D. Primero,

$$P(D) = \prod_{d \in D} P(d);$$

por la propiedad de factorización de redes bayesianas:

$$P(D) = \prod_{d \in D} \prod_{X \in \underline{X}} P_X(d|_{\{d, PA(d)\}}) = \prod_{X \in \underline{X}} \prod_{d \in D_{\{X, PA(X)\}}} P_X(d).$$

De manera que el conjunto en el que puede estar d se restringe a X y sus padres. La última igualdad se maximiza si el producto $\prod_{X \in \underline{X}} \prod_{d \in D_{\{X,PA(X)\}}} P_X(d)$ se maximiza para toda $X \in \underline{X}$. Entonces, se puede ver al producto a maximizar como:

$$P_X(X) := \prod_{X \in \underline{X}} \prod_{d \in D_{\{X,PA(X)\}}} P_X(d) = \prod_{x \in dom(X)} \prod_{y \in dom(PA(X))} P_X(X = x|PA(X) = y)^{n_D(x,y)}$$

donde $dom(X)$ y $dom(PA(X))$ son los conjuntos de valores que X y los padres de X pueden tomar, respectivamente. Además se define:

$$n_D(x, y) := |\{d \in D | d|_{X=x}, d_{PA(X)} = y\}|.$$

Dado que los conjuntos $dom(X)$ y $dom(PA(X))$ son finitos:

$$P_X(D) = \prod_{y \in dom(PA(X))} \prod_{x \in dom(X)} P_X(X = x|PA(X) = y)^{n_D(x,y)}.$$

Así, se puede maximizar $P_X(D)$ maximizando $\prod_{x \in dom(X)} P_X(X = x|PA(X) = y)^{n_D(x,y)}$, para toda $y \in dom(PA(X))$. Por cuestiones prácticas se denota:

$$P_X(X = x|PA(X) = y) =: P_X(x|y)$$

Por otro lado, se tienen las restricciones:

- $P_X(x|y) \in [0, 1]$
- $\sum_{x \in dom(X)} P_X(x|y) = 1$

Utilizando la segunda restricción, se puede descomponer:

$$\prod_{x \in dom(X)} P_X(x|y) = \prod_{\substack{x \in dom(X) \\ x \neq x_0}} P_X(x|y)^{n_D(x,y)} * [1 - \sum_{\substack{x \in dom(X) \\ x \neq x_0}} P_X(x|y)]^{n_D(x_0,y)}$$

Para algún valor x_0 arbitrario. Calculando la función logaritmo en ambos lados de la ecuación anterior y aplicando las propiedades de exponente y descomposición de producto de logaritmo, se obtiene la *log-verosimilitud*:

$$\log\left(\prod_{x \in dom(X)} P_v(x|y)\right) = \sum_{\substack{x \in dom(X) \\ x \neq x_0}} n_D(x,y) \log[P_X(x|y)] + n_D(x_0,y) \log[1 - \sum_{\substack{x \in dom(X) \\ x \neq x_0}} P_X(x|y)]$$

Sea $dom(X) := \{x_0, x_1, ...x_n\}$ y denotamos $p_i = p_{x_i}(x_i|y)$ y $n_i = n_D(x_i, y)$, entonces se puede escribir:

$$L(p) = \sum_{i=1}^{n} n_i \log(p_i) + n_0 \log\left(1 - \sum_{i=1}^{n} p_i\right)$$

Para obtener el máximo de $L(p)$, se deriva con respecto al parámetro p_j

$$\frac{\partial L}{\partial p_j} = n_j \frac{1}{p_j} - n_0 \frac{1}{1 - \sum_{i=1}^{n} p_i}$$

Igualando a cero y despejando p_j:

$$p_j = \frac{n_j}{n_0}(1 - \sum_{i=1}^{n} p_i) \tag{2.22}$$

Sumando sobre $i = \{1, 2, ..., n\}$ y desarrollando la ecuación:

$$\begin{aligned} \sum_{i=1}^{n} p_i &= \sum_{i=1}^{n} (\frac{n_j}{n_0})(1 - \sum_{i=1}^{n} p_i) \\ &= \frac{\sum_{i=1}^{n} n_i}{n_0 + \sum_{i=1}^{n} n_i} \end{aligned}$$

Así, sustituyendo en (2.22), se obtiene que:

$$p_j = \frac{n_j}{n_0}(1 - \sum_{i=1}^{n} p_i) = \frac{n_j}{n_0 + \sum_{i=1}^{n} n_i}.$$

□

Así, para estimar la probabilidad de que una variable tome un valor determinado, dado que sus padres tienen un conjunto de valores específicos, puede calcularse la frecuencia relativa de los eventos y esa es la forma más verosímil.

Un punto a remarcar específico de este trabajo es el Factor de Expansión de la ENOE en la estimación de probabilidades; pues como anteriormente se señaló, dicho factor fue tomado en cuenta en el cálculo de las frecuencias relativas más adelante expuestas.

Capítulo 3

Datos: ENOE, análisis y tratamiento

En este apartado se habla sobre la composición de la base de datos de la ENOE, en la que se basa este trabajo: detalles de su realización, las tablas que la componen, etc. Se avanza hacia la justificación de la selección de variables, los diferentes tipos de datos faltantes, así como su tratamiento. Finalmente se realiza un análisis descriptivo de los datos.

En adelante, en ocasiones se usará el término *semana de referencia* para denotar a la semana anterior en la que se entrevista a la persona muestra en la ENOE.

3.1. Encuesta Nacional de Ocupación y Empleo (ENOE)

La Encuesta Nacional de Ocupación y Empleo (ENOE) es el referente más importante sobre el mercado laboral en México, proporciona datos mensuales y trimestrales de la fuerza de trabajo, la ocupación, la informalidad laboral, la subocupación y la desocupación (INEGI). Uno de sus objetivos principales es el de proporcionar información estadística sociodemográfica que permita complementar y profundizar el análisis de las características ocupacionales de la población mexicana (11); en este trabajo, nos enfocaremos en la población femenina mayor de edad que habita en el Estado de México. Nuestro objetivo es encontrar las posibles relaciones entre las variables socio-económicas que estén influyendo en el estado ocupacional de este sector de la población.

3.1.1. Segmentación de la población

La encuesta divide a la población en dos sectores principales: la Población Económicamente Activa (PEA) y la Población No Económicamente Activa (PNEA); y se considera variedad de factores para clasificar a las personas dentro de estas categorías, que además cuentan con subcategorías, como veremos a continuación.

De acuerdo al INEGI (11), la **Población Económicamente Activa** se compone de las personas que ofrecen servicios laborales; sea que alguien demande sus servicios actualmente, o que estén ejerciendo presión en el mercado laboral para conseguir un trabajo. Es así que esta población se subdivide en otras dos categorías: Población Ocupada y Población Desocupada.

La **Población Ocupada** son los y las trabajadoras, ya sea independientes o asalariadas. Es decir, toda persona que se involucre en procesos de transacciones bilaterales; en la semana de referencia realizaron alguna actividad económica durante al menos una hora. Incluye a las personas ocupadas que tenían trabajo, pero no lo desempeñaron temporalmente por alguna razón, sin que por ello perdieran el vínculo laboral con este; así como a quienes ayudaron en alguna actividad económica sin recibir un sueldo o salario. Por otro lado, la **Población Desocupada** incluye a personas que no cuentan con un trabajo, pero se encuentran realizando acciones concretas de búsqueda para participar en el ámbito de las transacciones.

La **Población No Económicamente Activa** se refiere a las personas cuya subsistencia en la semana de referencia se basa en la transferencia de ingresos monetarios o no monetarios realizada por un familiar o terceras partes; y que además no están buscando involucrarse en el mercado laboral ni realizando actividades económicas. Esta categoría también tiene dos subdivisiones: la **Población Disponible** y la **Población No Disponible**. La primera hace referencia al grupo de personas que aunque al momento de la encuesta no ha ejercido acciones para conseguir trabajo, sí está interesada en trabajar y no hay impedimentos físicos o sociales que le detengan. La segunda incluye a las personas que no tienen interés en trabajar y/o están impedidas de manera permanente para hacerlo.

En el diagrama de la Figura 3.1 se muestran las categorías antes señaladas.

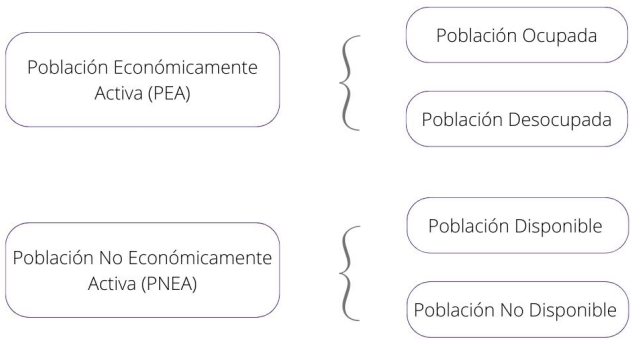


Figura 3.1: Segmentación de la población en categorías de ocupación de acuerdo a la ENOE 2019.

Dicho lo anterior, es importante notar que en la base de datos utilizada para la creación de la red bayesiana se incluyen mujeres provenientes de todas las clasificaciones laborales señaladas anteriormente, debido a que se considera importante analizar diferencias y similitudes de las características económicas y sociales de la población femenina respecto a su estatus laboral, sin importar a cuál pertenecen.

3.1.2. Factor de Expansión

Los datos de la encuesta incluyen un Factor de Expansión para cada observación, que de acuerdo al INEGI, es un coeficiente que otorga determinado peso en cada elemento de la muestra en función de su representatividad de otros tantos casos similares a él, tomando en cuenta el estrato socio-económico y lugar de residencia (10). En otras palabras, el Factor de Expansión es un número que indica cuántas viviendas son representadas por la vivienda que aparece seleccionada en la muestra, y todos los datos que proporcionan los habitantes dentro de esta vivienda se multiplican por ese número. Los detalles sobre cómo se construye este ponderador pueden consultarse directamente en el manual *"Cómo se hace la ENOE. Métodos y procedimientos"* (11).

Por último, cabe resaltar que este número es considerado en todos los análisis, algoritmos y modelos aquí presentados; desde la programación hasta la interpretación.

3.1.3. Cobertura Geográfica de la ENOE

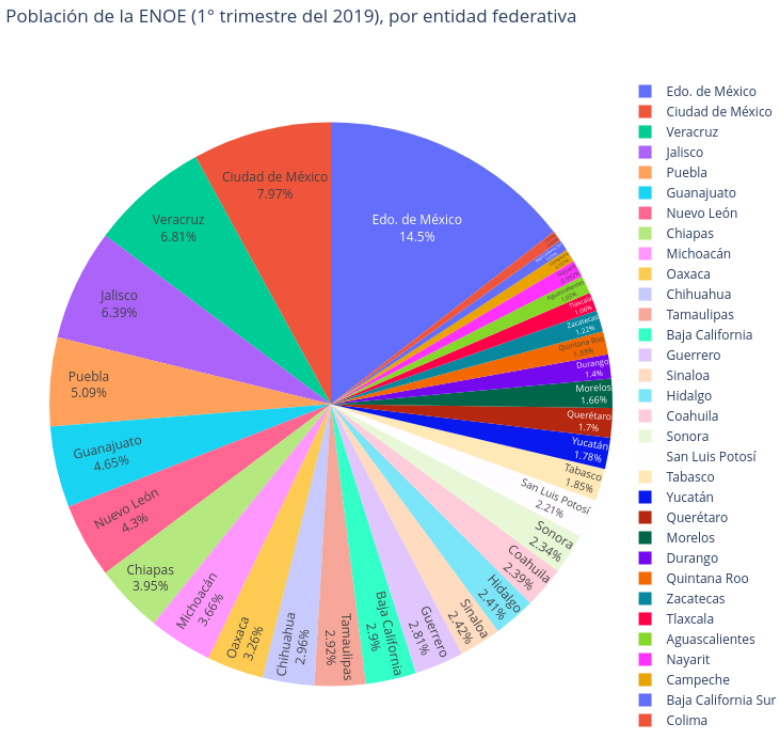


Figura 3.2: Gráfica de pastel de división de la población de la encuesta.

La ENOE tiene 3 dominios de estudio: 39 ciudades autorrepresentadas [1] (ciudades elegidas según su importancia en la entidad donde se localizan; según el INEGI,

[1]Para formar parte de la ENOE en cada una de las 32 entidades federativas se ha elegido una o más ciudades según su importancia, cada una de estas ciudades se considera **autorrepresentada** por

representan el 45.7 % de la población del país), el complemento urbano de alta densidad (formado por ciudades que no son autorrepresentadas y localidades de 2,500 a 99,999 habitantes; son el 31.2 % de la población) y el dominio rural (localidades de 2499 o menos habitantes, representando al 23.1 % de la población).

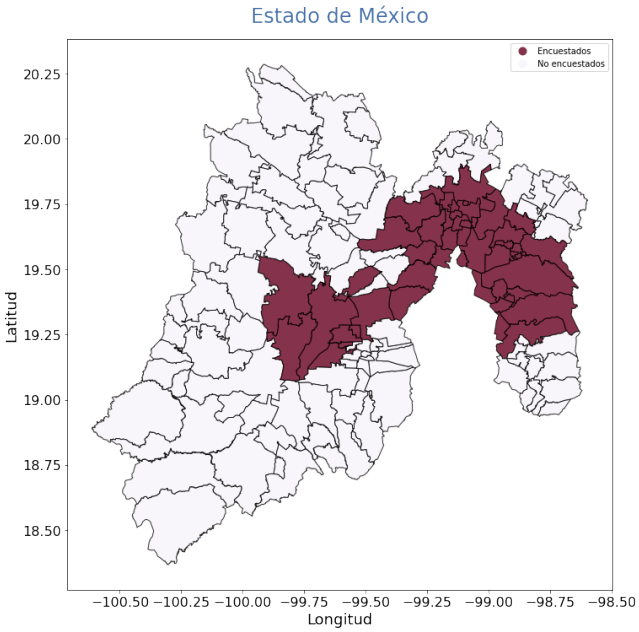


Figura 3.3: Distribución geográfica de los municipios del Estado de México contemplados en la ENOE.

Los criterios utilizados para clasificar las ciudades y localidades en cada una de las categorías anteriores están dados por la ENOE y son especificados con mayor detalle en el manual *"Cómo se hace la ENOE. Métodos y procedimientos"* (11).

En este proyecto se trabaja con la muestra proveniente del Estado de México, que representa el 14.5 % de la población (mujeres de 18 años o más) nacional de la encuesta. Asimismo, cuenta con 2 ciudades autorrepresentadas (Ciudad de México y Toluca); así como 31 % del complemento urbano y 3 % de las viviendas rurales a nivel nacional (11).

La distribución de las observaciones por entidad federativa se muestra en la Figura 3.2. Y en la Figura 3.3 se señalan los municipios contemplados en la encuesta, de acuerdo a la guía de *Métodos y Procedimientos de la ENOE* (11). Los nombres de los municipios correspondientes se encuentran en el Apéndice A.

lo que se le calcula un tamaño de muestra mínimo para poder dar resultados de manera independiente. Los criterios para elegir una ciudad como autorrepresentada en la ENOE son el político, el de desarrollo económico, su tasa de crecimiento medio anual y su desarrollo urbano, además de la clasificación según el Sistema Urbano Nacional (11).

3.2. Elección de variables y limpieza de datos

En este trabajo se utilizan los datos de la ENOE correspondientes al primer trimestre del 2019; esto porque se busca que la base de datos sea reciente y que el estudio se lleve a cabo bajo condiciones de salud pública estables, dada la actual pandemia por SARS-CoV-2.

En el primer trimestre de cada año se utiliza el Cuestionario de Ocupación y Empleo Ampliado, que proporciona más datos de Ocupación que el Cuestionario Básico (realizado el resto de los trimestres), específicamente es en el primer trimestre que se tienen más preguntas sobre las horas que dedican las personas a las obligaciones y tareas del hogar.

A su vez, la ENOE se compone de 5 tablas de datos, a saber: Sociodemográfico, Hogar, Vivienda y dos Cuestionarios de Ocupación y Empleo. En este trabajo únicamente se consideran las tablas Sociodemográfico y las dos de Ocupación y Empleo.

Aunque la población objetivo de la ENOE son hombres y mujeres de 15 años o más, la muestra con la que se trabaja de aquí en adelante son las mujeres de 18 años o más. No se considera en este proyecto a las mujeres menores de 18 años; porque aunque la Ley Federal del Trabajo les permite unirse al campo laboral, esto puede suceder únicamente bajo ciertas condiciones que restringen las oportunidades de esta población (18); por lo que los resultados obtenidos de estos datos podrían verse afectados.

Cabe recordar que se contemplan mujeres pertenecientes tanto a la Población Económicamente Activa, como a la Población No Económicamente Activa.

Para la selección se consideraron primero las variables que describieran mejor la situación social, económica y de vivienda de cada mujer: la edad, desarrollo urbano de su localidad, nivel de estudios, número de hijos, clasificación por actividad e inactividad, asistencia actual a la escuela, estado conyugal y antecedentes laborales.

Se añadieron también algunas de las variables de ocupación con enfoque de género de la ENOE (9), que consisten en el conteo de horas dedicadas a diferentes actividades durante la semana anterior a la de referencia (cuando la persona es encuestada). Las que se toman en cuenta son: tiempo dedicado a cuidar a niños y/o ancianos, trámites y seguridad del hogar, llevar a algún miembro del hogar al médico o escuela, así como el tiempo destinado a realizar quehaceres del hogar.

Es bueno recordar que las variables socio-económicas mencionadas anteriormente se eligieron como parte de la base de datos debido a la hipótesis de que se relacionan con el estado ocupacional de las personas; tal como se mencionó en la introducción.

Finalmente, se obtuvo una base de datos con 13 variables categóricas y 6,613 registros. Las variables son listadas en las Tablas 3.1, 3.2, 3.3 y 3.4.

Como información adicional, para definir mejor la variable T_LOC, ésta comprende a

cuatro grandes áreas de población: Áreas más urbanizadas: de 100,000 y más habitantes. Urbano medio: de 15,000 a 99,999 habitantes. Urbano bajo: de 2,500 a 14,999 habitantes. Rural: de menos de 2,500 habitantes (INEGI).

Variables Socio-económicas				
NOMBRE	DESCRIPCIÓN	RANGO	VALORES	
CLASE2	Condición de actividad de segunda categoría	(1,4)	1	Población ocupada
			2	Población desocupada
			3	Disponibles
			4	No disponibles
EDA7C	Intervalo de edad	(1,7)	1	De 18 a 19 años [1]
			2	De 20 a 29 años
			3	De 30 a 39 años
			4	De 40 a 49 años
			5	De 50 a 59 años
			6	De 60 años y más
			7	Edad no especificado
T_LOC	Tamaño de localidad	(1,4)	1	Urbano alto
			2	Urbano medio
			3	Urbano bajo
			4	Rural
HIJ5C	Clasificación de la población femenina de 15 años y más por número de hijos	(1,5)	1	Sin hijos
			2	De 1 a 2 hijos
			3	De 3 a 5 hijos
			4	De 6 hijos y más
			5	No especificado

[1] Este campo originalmente comprende el rango de edad 15 a 19 años, pero dada las restricciones de este trabajo, se modifica al valor de la tabla.

Tabla 3.1: Variables seleccionadas de la ENOE (Cuestionario ampliado) para el modelo.

Variables Socio-económicas				
NOMBRE	DESCRIPCIÓN	RANGO	VALORES	
CS_P13_ 1	¿Hasta qué grado aprobó en la escuela?	(0,9), 99	0	Ninguno
			1	Preescolar
			2	Primaria
			3	Secundaria
			4	Preparatoria o bachillerato
			5	Normal
			6	Carrera técnica
			7	Profesional
			8	Maestría
			9	Doctorado
			99	No sabe
CS_P17	¿Asiste a la escuela actualmente?	(1,2), 9	1	Sí
			2	No
			9	No sabe
E_CON	Estado conyugal	(1,6), 9	1	Vive con su pareja en unión libre
			2	Está separado(a)
			3	Está divorciado(a)
			4	Está viudo(a)
			5	Está casado(a)
			6	Está soltero(a)
			9	No sabe
D_CEXP_EST	Clasificación de los antecedentes laborales	(1,4)	0	No aplica
			1	Perdió o termino su empleo anterior
			2	Insatisfacción con el empleo anterior
			3	Dejo o cerro un negocio propio
			4	Otro

Tabla 3.2: Variables seleccionadas de la ENOE (Cuestionario ampliado) para el modelo.

Variables de Ocupación y Empleo				
NOMBRE	DESCRIPCIÓN	RANGO	VALORES	
P2F	Tiene necesidad de trabajar	(1,3), 9	0	No aplica
			1	Sí tiene necesidad de trabajar
			2	Sólo tiene deseos de trabajar
			3	No tiene ni necesidad ni deseos de trabajar
			9	No sabe
P11_H2	Horas que dedicó a cuidar o atender sin pago, de manera exclusiva a niños, ancianos, enfermos o discapacitados, durante la semana anterior a la de referencia	00-97, 98, 99, blanco	00-97	Número de horas dedicadas
			98	Realizó la actividad pero no sabe cuánto tiempo
			99	No sabe si realizó la actividad
			Blanco	No aplica
P11_H3	Horas que dedicó a realizar compras, llevar cuentas o realizar trámites para el hogar o encargarse de la seguridad, durante la semana anterior a la de referencia	00-97, 98, 99, blanco	00-97	Número de horas dedicadas
			98	Realizó la actividad pero no sabe cuánto tiempo
			99	No sabe si realizó la actividad
			Blanco	No aplica
P11_H4	Horas que dedicó a llevar a algún miembro del hogar a la escuela, cita médica, u otra actividad, durante la semana anterior a la de referencia	00-97, 98, 99, blanco	00-97	Número de horas dedicadas
			98	Realizó actividad, pero no sabe cuánto tiempo
			99	No sabe si realizó la actividad
			Blanco	No aplica

Tabla 3.3: Variables seleccionadas de la ENOE (Cuestionario ampliado) para el modelo.

Variables de Ocupación y Empleo				
NOMBRE	DESCRIPCIÓN	RANGO	VALORES	
P11_H7	Horas que dedicó a realizar los quehaceres de su hogar (lavar, planchar, preparar y servir alimentos, barrer), durante la semana anterior a la de referencia	00-97, 98, 99, blanco	00-97	Número de horas dedicadas
			98	Realizó actividad, pero no sabe cuánto tiempo
			99	No sabe si realizó la actividad
			Blanco	No aplica

Tabla 3.4: Variables seleccionadas de la ENOE (Cuestionario ampliado) para el modelo.

3.2.1. Variables P11

De aquí en adelante se denominará como "variables P11" al grupo de variables de ocupación y empleo cuya codificación en la Tablas 3.3 y 3.4 comienza con "P11". Dichas variables indican el número de horas dedicadas a alguna actividad específica.

Se hizo una modificación importante a este conjunto de variables. Originalmente, en la base de datos de la ENOE se tienen dos columnas para la característica de tiempo: "Número de horas que la persona dedicó a la actividad..." y "Número de minutos que la persona dedicó a la actividad..."; para comprimir esta información en una sola variable, se sumaron los minutos a la variable de horas. El resultado de lo anterior es que las variables *P11* indican el tiempo total destinado a una actividad concreta, tomando las horas como unidad, pero sin redondear (es decir, se consideran también los minutos).

También hay que remarcar que posterior al proceso de imputación (que se describe en el Apéndice B), las variables en cuestión fueron agrupadas para facilitar el análisis estadístico de los datos, concretamente para agilizar los algoritmos utilizados en la construcción de la red bayesiana; pues, como se verá más adelante, el tiempo de ejecución de dichos métodos dependen del número de variables y la cantidad de categorías que cada una de éstas posee.

Los grupos para la variable *P11_H2* son de 15 horas comenzando en 0 y terminando en "60 en adelante". Para la variable *P11_H7* son de 15 horas, comenzando en 0 y terminando en "45 en adelante". Finalmente, para las variables *P11_H3* y *P11_H4*, los grupos son de 4 horas comenzando en cero y terminan ambas en "12 en adelante", tal como se muestra en la Tabla 3.5.

Variable	Valor asignado	Intervalo de tiempo (min)
P11_H2	-1	No aplica
	1	[0-15)
	2	[15-30)
	3	[30-45)
	4	[45-60)
	5	60 o más
P11_H3	-1	No aplica
	1	[0-4)
	2	[4-8)
	3	[8-12)
	4	12 o más
P11_H4	-1	No aplica
	1	[0-4)
	2	[4-8)
	3	[8-12)
	4	12 o más
P11_H7	-1	No aplica
	1	[0-15)
	2	[15-30)
	3	[30-45)
	4	45 o más

Tabla 3.5: Discretización de variables de tiempo.

3.2.2. Datos faltantes e imputación

Dentro de las categorías que conforman las respuestas de la encuesta, existen opciones que dan lugar a datos faltantes o menos informativos que las categorías restantes de acuerdo a la información que se requiere en este trabajo. En las siguientes líneas se explica y justifica el trato que se le da a cada una de los tres casos de respuestas a tratar.

- **Respuestas no aplicables**

  La primera categoría es la correspondiente a la opción *No Aplica* de la encuesta, que está presente en las variables de Ocupación y Empleo de género (horas que se dedicaron a ciertas actividades, codificadas comenzando por *P11*), así como en las

variables *D_CEXP_EST* y *P2F*. De acuerdo al INEGI, los *No aplica*, se asignan cuando la pregunta del cuestionario no debe aplicarse, ya sea por la edad de la persona en cuestión o alguna otra circunstancia.

Estos valores se trataron como una categoría más de las variables que la tienen, puesto que aunque no dan una clasificación particular, sí brindan información sobre la observación, ya que están indicando que existe alguna razón por la que la realización de la actividad por parte de la mujer en cuestión no es aplicable; es decir, no es factible o lógico preguntarle.

De hecho, para algunas variables de empleo (las variables *P11*) estas respuestas representan un gran porcentaje de las observaciones; llegando a ser hasta el 84 % de éstas, por ejemplo para la variable *P11_H4* (horas dedicadas a estudiar o tomar cursos de capacitación). Por otro lado, tampoco es información faltante, ya que no es que la persona entrevistada no haya sabido la respuesta, sino que ni siquiera se le preguntó debido a ciertas condiciones de la persona, por lo tanto no se consideran datos faltantes.

De forma similar, la pregunta correspondiente a la variable *D_CEXP_EST* (clasificación de antecedentes laborales) sólo fue aplicada a las mujeres Desocupadas, así que todas las observaciones correspondientes a mujeres Ocupadas, Disponibles y No Disponibles tienen un *No aplica* en esa variable. Para estos últimos registros, *No aplica* no significa nada más que "la mujer en cuestión no pertenece a la Población Desocupada". En otras palabras, el *No Aplica* fuera del grupo de mujeres desocupadas no sucede bajo condiciones que no sean que la mujer en cuestión no pertenece a esa selección de la población, y dado que no hay una razón más allá de eso para utilizar "No aplica", no hay más interpretación que esa.

La situación anterior (D_CEXP_EST) difiere de las otras variables que contienen "No Aplica" en que para éstas últimas no se sabe la razón particular por la cual este valor fue asignado; puede deberse a cualquier situación social, económica, física o cualquier otro motivo que haga de la pregunta algo incongruente o no factible para la persona en cuestión.

- **Datos faltantes**

  La segunda categoría de datos a imputar son los valores que sirven para señalar que no se tiene información sobre la respuesta de la pregunta en cuestión, éstos son adjudicados a respuestas en la encuesta como "No sabe" o "No especificado"; por ejemplo, en la variable *P11_H7* (horas dedicadas a realizar quehaceres del hogar) la categoría 99 corresponde a las respuesta: "No sabe si realizó la actividad". También ocurre esto para las variables socio-económicas, como es el caso de *CS_P17* (¿Asiste a la escuela actualmente?), cuyo valor 9 se traduce en la respuesta "No sabe"por parte de la persona entrevistada.

  Este tipo de respuestas se encuentran tanto en variables socio-económicas como de ocupación y empleo; lo importante es señalar que fueron tratados como datos faltantes, y fue así porque aunque la pregunta correspondiente a la variable sí se hizo

a la persona entrevistada, la persona no conoce la respuesta; pero eso no quiere decir que la pregunta no sea aplicable, sólo no hay tal información.

- **Datos faltantes informativos**

  El tercer tipo de respuesta que se trata en este análisis sólo aplica para las preguntas de ocupación y empleo, específicamente las variables *P11*, y son las respuestas del tipo "Realizó la actividad, pero no sabe cuánto tiempo le dedicó".

  Dichas categorías son datos faltantes, pero no son iguales a los explicados en el primer punto, debido a que éstos brindan un poco de información al indicar que efectivamente se realizó la actividad de referencia, de este modo lo que se desconoce es únicamente el número de horas que se destinaron.

Los datos faltantes fueron imputados con ayuda de la paquetería MICE en R, este método y el procedimiento general de preparación de los datos se describen en el Apéndice B.1.

Los datos faltantes informativos también fueron imputados con la librería MICE de R, pero se creó una variable auxiliar para cada variable del tipo *P11* que diferencía entre los datos faltantes y los datos faltantes informativos. De esta manera, cada variable auxiliar contiene 4 valores, cada uno de ellos se describen en la Tabla 3.6.

Las variables auxiliares introducidas al modelo, se añaden para que MICE contemple la información que dan los datos faltantes del tercer tipo; lo anterior bajo la premisa de que MICE es un método que hace uso de otras variables para imputar una variable específica del conjunto de datos. Nuevamente, este procedimiento se explica con más detalle en el Apéndice B.1.

P11_HX auxiliar		
-3	No aplica	Primera categoría
-2	No sabe si realizó la actividad	Segunda categoría
-1	Realizó la actividad, pero no sabe cuánto tiempo	Tercera categoría
0	Respondió con el tiempo dedicado a la actividad	-

Tabla 3.6: Codificación de las categorías de las variables auxiliares de tipo"*P11_HX*"

Resumiendo lo anterior, los valores "No aplica" fueron tratados como otra categoría; los valores "No sabe" o "No especificado" se imputaron de manera normal; y los valores "Realizó la actividad, pero no sabe cuánto tiempo", se imputaron utilizando variables auxiliares que reflejan la información de que efectivamente se hizo la actividad que este tipo de dato brinda. Por último, es importante mencionar que para la imputación no se tomaron en cuenta los factores de expansión de las observaciones, debido a que la

paquetería de software utilizada no considera este caso; aunado a que trabajar con los datos desagregados no es factible debido al tamaño de la base que se genera y la memoria del equipo de cómputo utilizado.

En la Tabla 3.7 se desglosa el número de observaciones con el que contaba cada variable de cada uno de los tipos de dato faltante detallados anteriormente. Vale la pena mencionar que el número de observaciones finales; después de filtrar por Sexo, Edad y Entidad Federativa, tal cómo se mencionó al principio de la sección; es $6,613$ sin tomar en cuenta el Factor de Expansión (FAC). Tomando en cuenta el FAC, la muestra representa a $6,691,900$ mujeres en el Estado de México.

	Nueva categoría	Imputados	
Variable	**No aplica**	**No sabe/No especificado**	**Realizó la actividad, pero no sabe cuánto tiempo dedicó**
EDA7C	0	3	0
CS_P13_1	0	3	0
E_CON	0	1	0
D_CEXP_EST	6,507	0	0
P2F	2,988	0	0
P11_H2	4,880	0	3
P11_H3	1,171	0	0
P11_H4	5,554	0	4
P11_H7	232	0	9

Tabla 3.7: Número de observaciones de datos faltantes por tipo para cada variable.

Las variables no mostradas en la Tabla 3.7 no tienen observaciones con los valores señalados. Es decir, todos los renglones son cero.

3.3. Análisis descriptivo de los datos

Como ya se dijo anteriormente, se trabaja con un conjunto de datos de 6,613 observaciones y 13 variables categóricas. La proporción de mujeres según su estado de ocupación se indica en la gráfica de la Figura 3.4. Vale la pena mencionar que los porcentajes presentados en la gráfica consideran el Factor de Expansión de la ENOE, y en todos las gráficas y análisis estadísticos siguientes así será, a menos que se indique lo contrario; además, este análisis descriptivo se realizó a los datos filtrados y sin imputar, por lo que las conclusiones que de aquí salgan serán aplicables sólo a esta muestra.

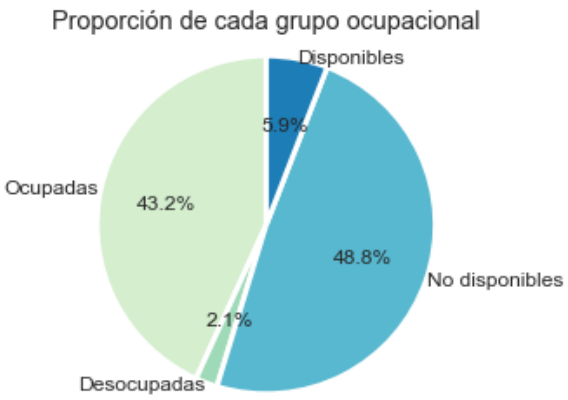


Figura 3.4: Proporción de mujeres según su estado de ocupación. Se toma en cuenta el Factor de Expansión.

Puede verse que la mayor parte de la muestra está condensada en las mujeres Ocupadas y No disponibles. Las mujeres Desocupadas son la población más pequeña de la muestra.

Hay que mencionar que los grupos de edad no están equitativamente repartidos, pues la categoría de 15 a 19 años sólo cuenta a las personas de 18 a 19 años (ya que como se mencionó en la Sección 3.2, sólo se toman en cuenta las mujeres mayores de edad) mientras que las demás categorías son de un rango de 10 años. Lo anterior se refleja en el histograma de la Figura 3.5(b); pues es la categoría en cuestión la que cuenta con menos registros. Asimismo, la mayor parte de la muestra es proveniente de localidades consideradas como Áreas más urbanizadas, seguidas por localidades del tipo Urbano bajo; tal como se muestra en la Figura 3.5(a).

En el histograma de la Figura 3.5(c) se puede notar que las mujeres de la muestra tienen más comúnmente de 1 a 2 hijos. Aunado a lo anterior, acorde con la Figura 3.6(b) se puede decir que las mujeres casadas encabezan a la mayor parte del estado civil de las encuestadas.

Por otro lado, de acuerdo a la Figura 3.6(a), la mayoría de las mujeres encuestadas no asisten a la escuela actualmente, más aún el nivel máximo de estudios más común entre la muestra es la Escuela Secundaria (Figura 3.5(d)).

Del histograma de la Figura 3.6(c) debe recordarse que esta variable sólo es para las mujeres *Desocupadas*, es así que puede concluirse del gráfico que la mayoría de las mujeres *Desocupadas* dejó su trabajo anterior porque se encontraba insatisfecha con él.

De la gráfica en la Figura 3.6(d) puede verse que las mujeres Ocupadas y Desocupadas son las que tienen necesidad de trabajar; y ocurre de la misma manera, con que el hecho de que tengan deseos de trabajar.

Las gráficas de pastel de las variables tipo *P11_HX* se realizaron tomando en cuenta las observaciones no aplicables, mientras que las gráficas de barras correspondientes, no

toman en cuenta las observaciones con valores “No aplica”.

En el caso de los gráficos en las Figuras 3.6(e) y 3.6(f) ocurre que la mayoría de ellos son valores *No aplica*, pero los que no lo son se agrupan mayormente en la clase de 0 a 15 horas a la semana dedicadas a cuidar o atender sin pago de manera exclusiva a niños, ancianos, enfermos o discapacitados.

Las gráficas 3.7(a) y 3.7(b) reflejan que la pregunta de si las mujeres realizan cuentas o trámites del hogar, así como encargarse de la seguridad; sí es aplicable a la mayoría de ellas, y más conmúnmente dedican de 4 a 8 horas a la semana a ello.

De acuerdo a las Figuras 3.7(c) y 3.7(d); la pregunta sobre el tiempo que las mujeres dedican a llevar algún miembro del hogar a la escuela, cita médica,etc. No es aplicable para la mayoría de ellas, y para quienes aplica, la mayoría no dedica más de 4 horas semanales a dichas actividades.

Por último, puede verse en las Figuras 3.7(e) y 3.7(f) que con mucha diferencia a las preguntas anteriores (de tipo P11_HX), a la mayoría de las mujeres les fue aplicable la pregunta sobre las horas dedicadas a realizar los quehaceres del hogar. Más aún, que la mayoría de ellas dedican de 15 a 30 horas semanales a esta actividad.

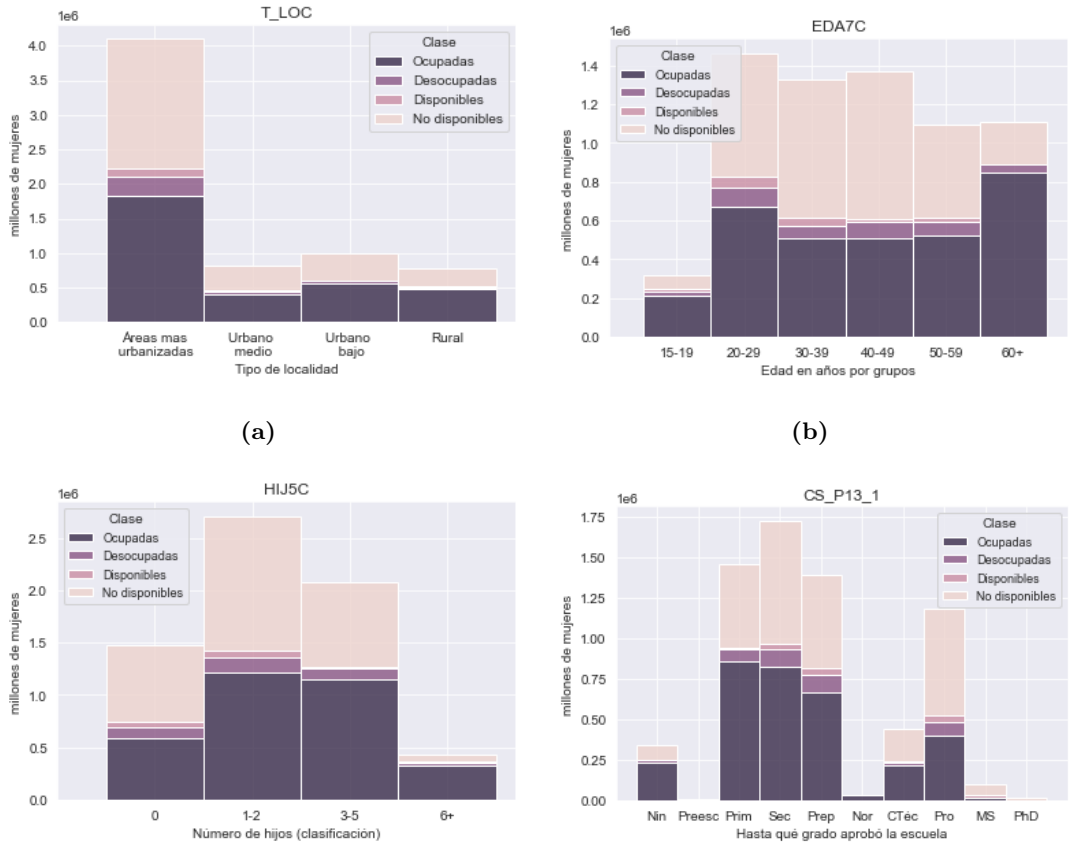


Figura 3.5: Distribución de la muestra para variables de localidad (T_LOC), edad (EDA7C), número de hijos (HIJ5C) y nivel escolar máximo (CS_P13_1); divididas por estado de ocupación.

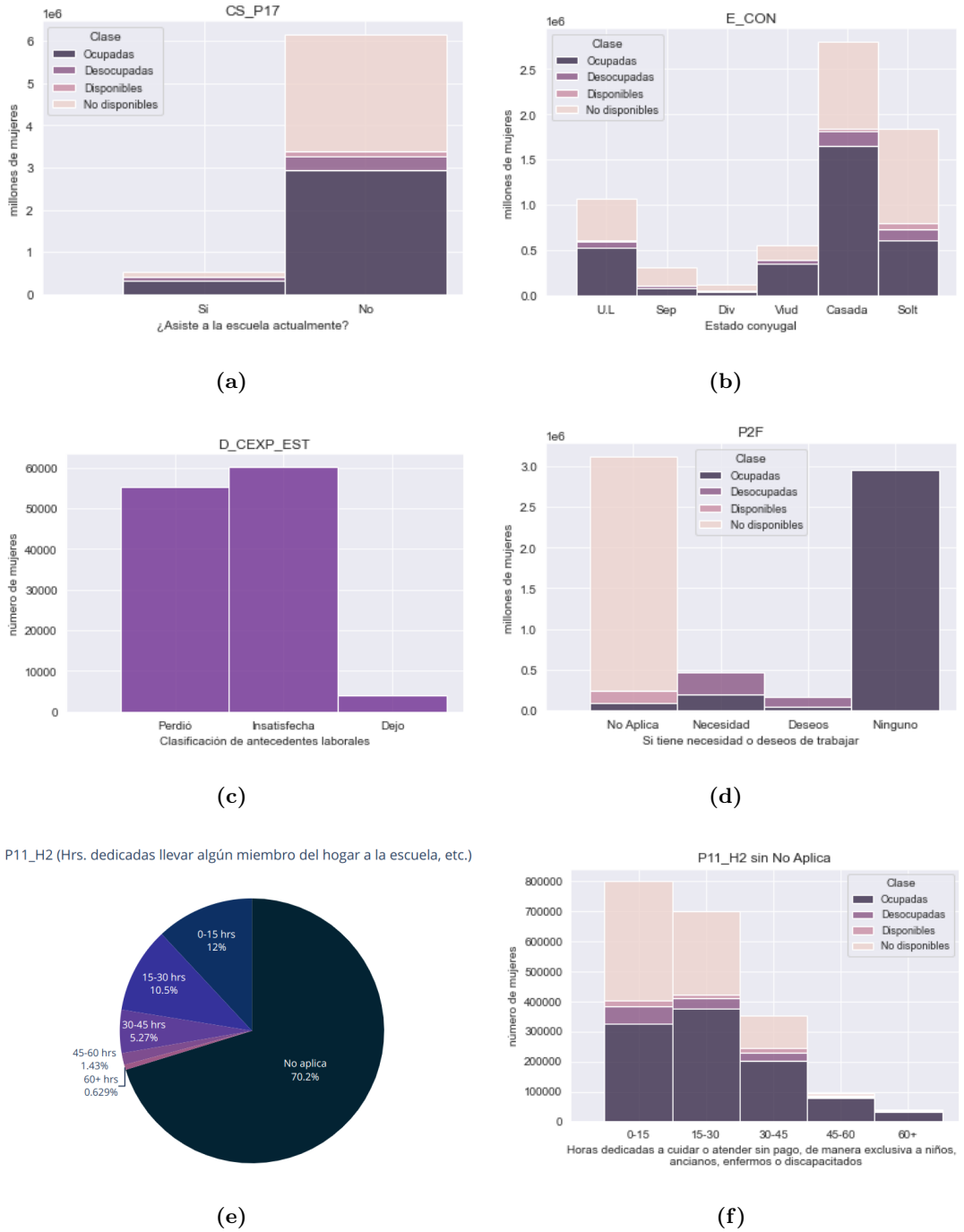
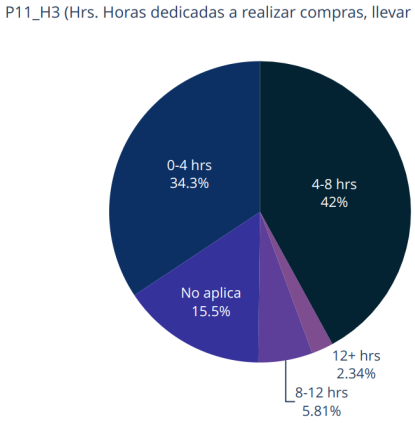
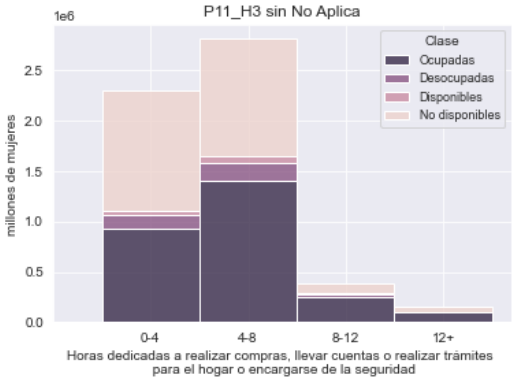


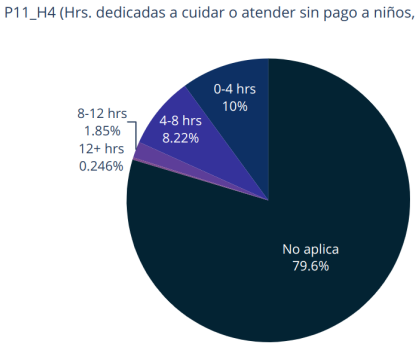
Figura 3.6: Distribución de la muestra para variables de asistencia a la escuela (CS_P17), estado conyugal (E_CON), antecedentes laborales (D_CEXP_EST), deseos de trabajar (P2F) y horas dedicadas a trasladar a miembros del hogar (P11_H2); divididas por estado de ocupación.



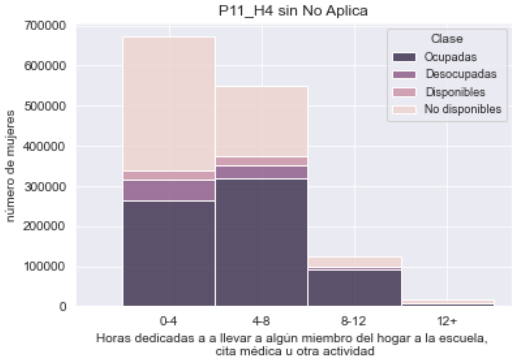
(a)



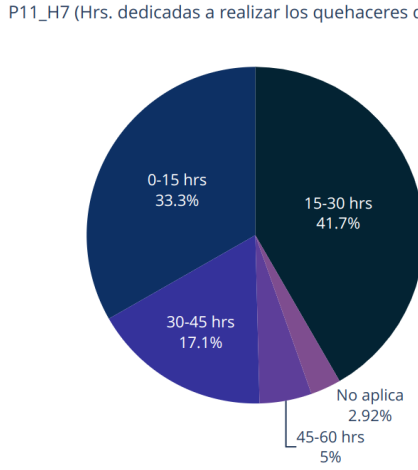
(b)



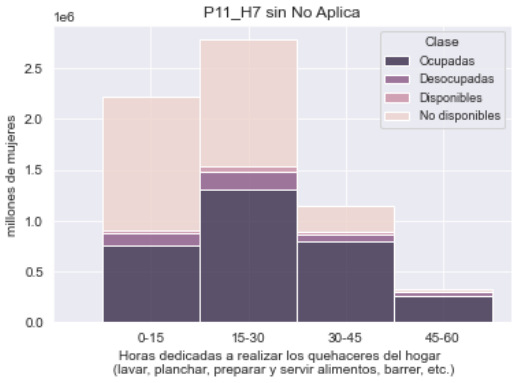
(c)



(d)



(e)



(f)

Figura 3.7: Distribución de la muestra para variables de número de horas dedicadas a: realizar compras y cuentas (P11_H3); cuidar o atender a otros (P11_H4); y realizar quehaceres del hogar (P11_H7); divididas por estado de ocupación.

Capítulo 4

Resultados y análisis

Con el conjunto de datos generado a partir de la ENOE del primer trimestre del 2019 y utilizando los conceptos y algoritmos presentados en el Capítulo 2, a continuación se presentan los resultados obtenidos.

Cabe recordar que debido al método utilizado para la imputación de los datos, MICE, expuesto en el Apéndice B.1; todos los algoritmos fueron aplicados a cinco conjuntos de datos. Más aún, ya que los resultados fueron idénticos para los cinco conjuntos de datos no fue necesario hacer un *pool* de los modelos.

4.1. Ordenamiento con entropía condicional

El ordenamiento de variables que se obtuvo del algoritmo de entropía condicional, y que posteriormente fue utilizado en el algoritmo K2, es el siguiente:

1. `D_CEXP_EST` (DC): Clasificación de los antecedentes laborales.
2. `CS_P17` (CS): Si asiste actualmente a la escuela.
3. `P11_H4` ($P4$): Horas que dedicó a llevar a algún miembro del hogar a la escuela, cita médica, u otra actividad, durante la semana anterior a la de referencia.
4. `CLASE2` (CL): Condición de actividad de segunda categoría.
5. `P2F` (PF): Si tiene necesidad de trabajar.
6. `P11_H2` ($P2$): Horas que dedicó a atender sin pago, de manera exclusiva a niños, ancianos, enfermos o discapacitados, durante la semana anterior a la de referencia.
7. `HIJ5C` (H): Clasificación de la población femenina por número de hijos.
8. `T_LOC` (L): Tamaño de localidad.
9. `E_CON` (CN) : Estado conyugal.
10. `P11_H7` ($P7$): Horas que dedicó a realizar los quehaceres de su hogar (lavar, planchar, preparar y servir alimentos, barrer) en la semana anterior a la de referencia.

11. `P11_H3` ($P3$): Horas que dedicó a realizar compras, llevar cuentas o realizar trámites para el hogar, o encargarse de la seguridad, durante la semana anterior a la de referencia.

12. `EDA7C` (E): Intervalo de edad.

13. `CS_P13_1` ($C13$): Grado hasta el que aprobó la escuela.

Teniendo en cuenta lo que significa la entropía condicional, se puede decir que cada variable en el orden de la lista anterior presentó la máxima certidumbre bajo la condición de que se conocen las variables anteriores. Así, `D_CEXP_EST` es la variable con más certidumbre del conjunto de datos; `CS_P17` presentó la mayor certidumbre condicionando al hecho de que se conoce la clasificación de antecedentes laborales; así sucesivamente.

4.2. Estructura de la red

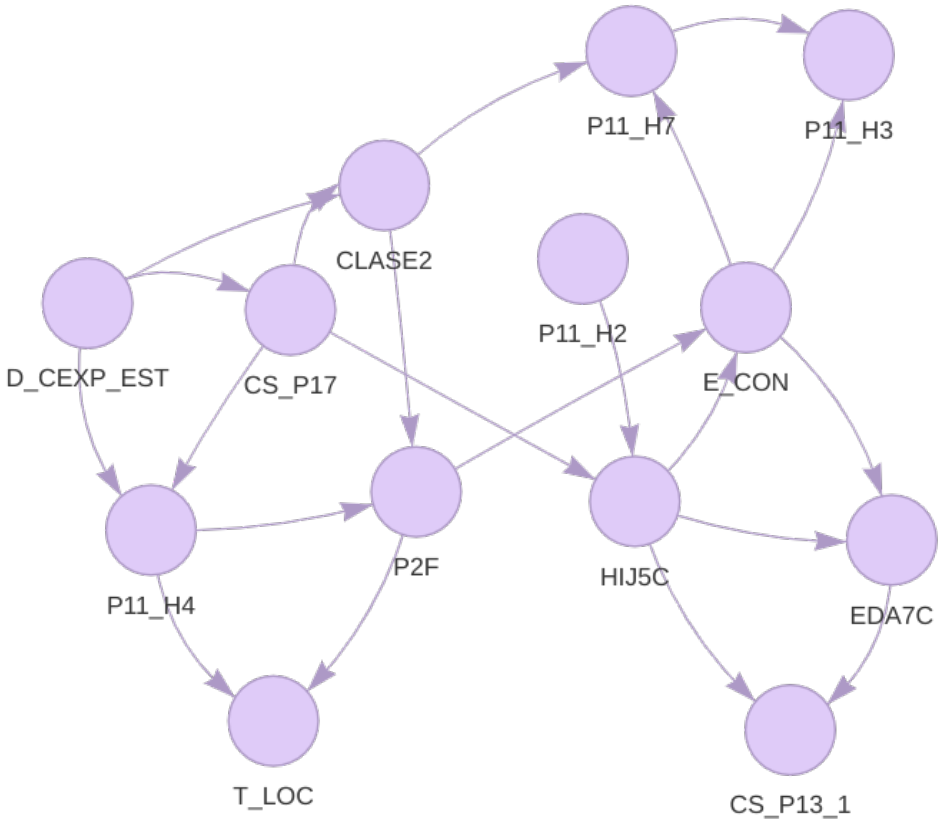


Figura 4.1: Red bayesiana obtenida del primer trimestre de la ENOE 2019.

La estructura de la red fue construida siguiendo el algoritmo K2, usando el ordenamiento antes señalado. Se limitó el número de padres de cada nodo a máximo 2. La red

obtenida se muestra en la Figura 4.1.

El análisis se centrará en la variable CLASE2, ya que es de interés saber cómo se relaciona con las otras variables. El conjunto de padres de CLASE2 se conforma:

$$PA_{CLASE2} = \{CS\_P17, D\_CEXP\_EST\}$$

recordando que el número de padres de cada variable se limitó a 2 en el algoritmo K2. De modo similar, los descendientes directos de $CLASE2$ fueron definidos por los algoritmos como:

$$DESC_{CLASE2} = \{P11\_H7, P2F\}$$

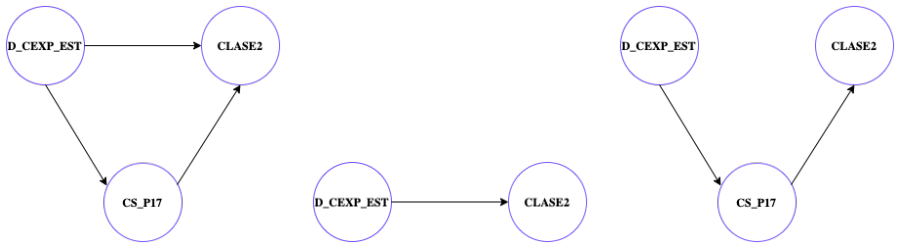
Ahora, veamos las independencias reflejadas por la estructura de la red, concernientes a la variable de interés, para lo cual se usa el criterio de *d-separación*.

4.2.1. Independencias reflejadas por la estructura

A continuación se exponen algunas de las relaciones de dependencia encontradas en la red de la Figura 4.1.

1. *Clasificación de los antecedentes laborales (**D_CEXP_EST**).*

Se tienen dos relaciones a simple vista. La primera corresponde a una relación padre e hijo y la segunda es que D_CEXP_EST es ancestro de CLASE2 (aunque no directo, como en el primer caso). Por el primer caso puede concluirse que estas dos variables son dependientes, incluso si se conoce el valor de CS_P17 (asistencia ala escuela), pues las variables están conectadas de manera directa.



(a) **(b)** **(c)**

Figura 4.2: Principales relaciones de D_CEXP_EST con CLASE2

*2. Si asiste actualmente a la escuela (**CS_P17**).*

Como en el caso anterior, dada la existencia de una relación de ancestro/descendiente directos, hay dependencia directa entre las variables CS_P17 y CLASE2 (el estado ocupacional), independientemente de si se condiciona sobre un conjunto Z de variables o no.

Figura 4.3: Principales relaciones de CS_P17 con CLASE2

*3. Horas que dedicó a llevar a algún miembro del hogar a la escuela, cita médica, u otra actividad, durante la semana anterior a la de referencia (**P11_H4**), Figura 4.4.*

En este caso son visibles dos relaciones. La primera es a través de un camino de causa común; en donde la causa común es CS_P17, por lo cual P11_H4 puede influenciar CLASE2 (y viceversa) únicamente si CS_P17 no es una variable observada. La segunda relación es mediante un camino de efecto común respecto a la variable P2F (si se tiene necesidad de trabajar), así, CLASE2 y P11_H4 son condicionalmente dependientes si se conoce el valor de P2F; influenciándose en el caso en que P2F o alguno de sus ancestros es observado; además también dependen condicionalmente de sí, cuando CS_P17 no se conoce.

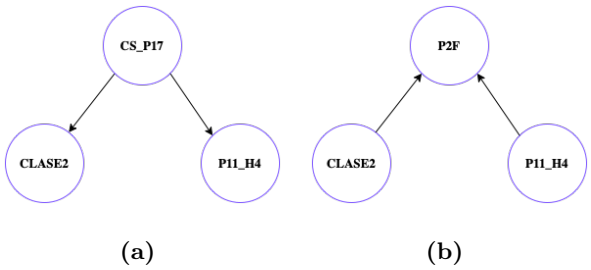


Figura 4.4: Principales relaciones de P11_H4 con CLASE2

*4. Si tiene necesidad de trabajar(**P2F**).*

Como en los casos anteriores, como los nodos están conectadas directamente, no existe *d-separación entre ellos* y dichas variables se influencían mutuamente.

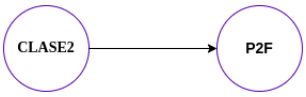


Figura 4.5: Principales relaciones de P2F con CLASE2

*5. Horas que dedicó a atender sin pago, de manera exclusiva a niños, ancianos, enfermos o discapacitados, durante la semana anterior a la de referencia (**P11_H2**).*

De acuerdo al Algoritmo de Bayes, y como los nodos P11_H2 (horas semanales dedicadas a atender a terceros), HIJ5C (número de hijos) y CS_P17 forman una estructura de V, P11_H2 y CS_P17 no están *d-separados*. A su vez, CLASE2 depende directamente de CS_P17, por lo que podemos concluir de la Figura 4.6 que CLASE2 depende condicionalmente de P11_H2 dado $Z = \{HIJ5C, CS\_P17\}$.

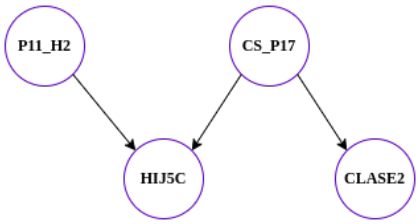
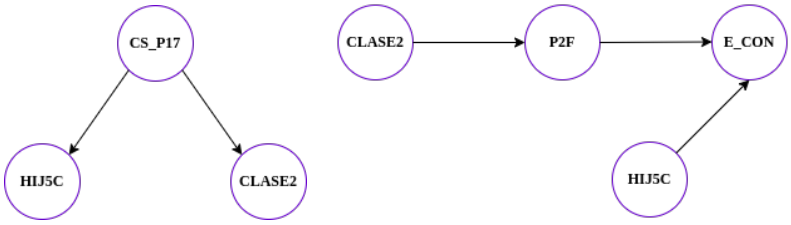


Figura 4.6: Principales relaciones de P11_H2 con CLASE2

*6. Clasificación de la población femenina por número de hijos (**HIJ5C**).*

Como notamos en la Figura 4.7(a), existe una relación de padre común entre HIJ5C y CLASE2 directa, por lo que son condicionalmente dependientes dada CS_P17. Además, es preciso notar que también existe una estructura (aunque no directa) de causa común, que involucra también a la variable P2F.



(a) **(b)**

Figura 4.7: Principales relaciones de HIJ5C con CLASE2

*7. Tamaño de localidad (**T_LOC**).*

En la Figura 4.8, se muestra una de las relaciones de CLASE2 con T_LOC, originada por una estructura secuencial con P2F, por lo que se puede decir que CLASE2 y T_LOC son condicionalmente independientes dado P2F.

Figura 4.8: Principales relaciones de T_LOC con CLASE2

*8. Estado conyugal (**E_CON**).*

Una de las relaciones más visibles es con respecto a la variable P11_H17 (horas semanales dedicadas a quehaceres del hogar), pues hay una estructura de efecto común, por lo que de acuerdo a la red obtenida, las variables CLASE2 y E_CON son condicionalmente dependientes dada P11_H7.

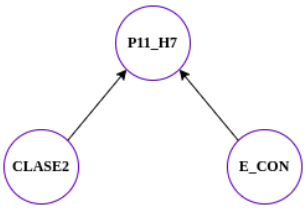


Figura 4.9: Principales relaciones de E_CON con CLASE2

*9. Horas que dedicó a realizar los quehaceres de su hogar (lavar, planchar, preparar y servir alimentos, barrer) en la semana anterior a la de referencia (**P11_H7**).*

Dado que se tiene una relación directa entre los nodos, Figura 4.10, la dependencia es entre dichas dos variables es verdadera sin importar sobre qué conjunto se condicione.

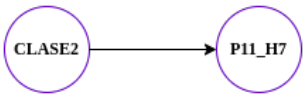


Figura 4.10: Principales relaciones de P11_H7 con CLASE2

*10. Horas que dedicó a realizar compras, llevar cuentas o realizar trámites para el hogar, o encargarse de la seguridad, durante la semana anterior a la de referencia (**P11_H3**).*

De acuerdo a las Figuras 4.11(a) y 4.11(b), se tienen relaciones secuenciales que involucran a las variables P11_H7, P2F (necesidad de trabajar) y E_CON (estado conyugal). De este modo se afirma que la variable de interés y P11_H3 dependen condicionalmente de P11_H7 y el conjunto de variables $Z = \{P2F, E\_CON\}$.

(a)

(b)

Figura 4.11: Principales relaciones de P11_H3 con CLASE2

*11. Intervalo de edad (**EDA7C**).*

Para la variable de edad la relación de dependencia condicional se da a través de las variables escuela e hijos, CS_P17 (si asiste actualmente a la escuela) y HIJ5C (número de hijos), respectivamente a partir de una estructura de causa común, Figura 4.12(a). Por otro lado, se tiene una estructura secuencial dada por las variables P2F y E_CON, Figura 4.12(b).

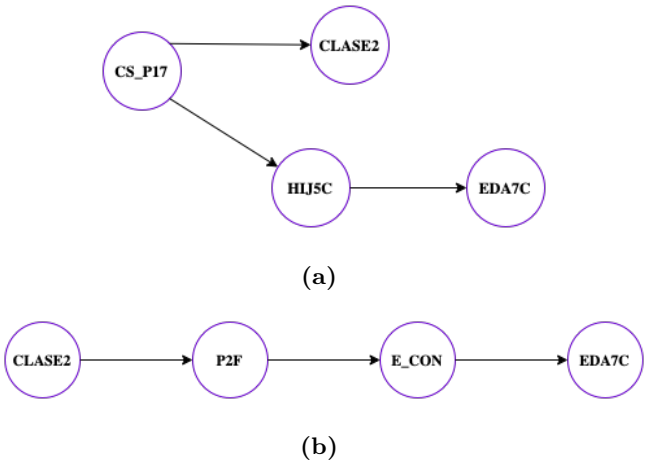


Figura 4.12: Principales relaciones de EDA7C con CLASE2

*12. Grado hasta el que aprobó la escuela (**CS_P13_1**).*

Se tiene una estructura de causa común, mediante las variables HIJ5C y CS_P17, Figura 4.13, por lo que se afirma la dependencia condicional de CS_P13_1 y CLASE2.

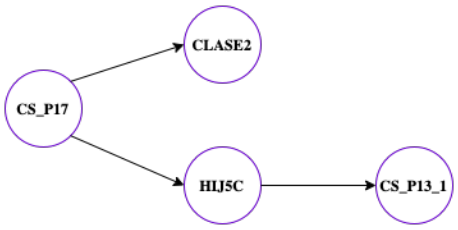


Figura 4.13: Principales relaciones de CS_P13_1 con CLASE2

De manera general, se encontraron dependencias condicionales desde la variable CLASE2 (estado de ocupación) con todas las restantes; más aún las dependencias directas fueron dadas por los nodos correspondientes a la asistencia actual a la escuela, antecedentes laborales, la necesidad de trabajar y las horas dedicadas a los quehaceres del hogar, es decir, los conjuntos de padres e hijos PA_{CL} y $DESC_{CL}$.

Hasta el momento, las principales relaciones encontradas han sido las dependencias directas mencionadas anteriormente, además de que las variables correspondientes al número de hijos y al estado conyugal de las encuestadas son las que más se relacionaron con otras variables de la red; tal como puede verse en 4.1 cada una con 5 relaciones de padre o hijo con las demás.

Otro descubrimiento hasta el momento, es que relación directa de D_CEXP_EST con CLASE2 se da a pesar de la restricción de que esta variable sólo es aplicable a la población desocupada. De forma similar, la variable CS_P17 forma parte de los padres de CLASE2, pese al desbalance en las clases que la conforman.

4.3. Parámetros de la red

En esta sección se calculan algunas de las probabilidades ligadas a la Red Bayesiana obtenida de los datos. Se consideran casos específicos con el fin de resaltar las relaciones más relevantes a los fines de esta tesis y mostrar el procedimiento general de la estimación; aunque se recuerda a la persona lectora que puede seguirse este análisis hasta la exhaustividad.

Primeramente se calculan probabilidades condicionales para la variable de interés CLASE2; remarcando el hecho de que uno de sus nodos padres en la red es D_CEXP_EST (antecedentes laborales), variable no aplicable para cualquier observación que no pertenezca a la población desocupada.

4.3.1. Análisis a variable CLASE2

Se calcula la probabilidad condicional de algunos valores de clasificación de ocupación dado que se conoce al conjunto PA_{CLASE}, las instancias obtenidas se dividen en casos. Los primeros casos abordados se derivan de considerar D_CEXP_EST como no aplicable (valor cero), como veremos a continuación.

1. Probabilidad de que una mujer pertenezca a la población Ocupada, dado que asiste actualmente a la escuela y no le fue aplicable la pregunta sobre antecedentes laborales.

$$P(CLASE2 = 1 | D\_CEXP\_EST = 0, CS\_P17 = 1) = 0.23 \tag{4.1}$$

2. Probabilidad de que una mujer pertenezca a la población No disponible, dado que asiste actualmente a la escuela y no le fue aplicable la pregunta sobre antecedentes laborales.

$$P(CLASE2 = 4 | D\_CEXP\_EST = 0, CS\_P17 = 1) = 0.62 \tag{4.2}$$

3. Probabilidad de que una mujer pertenezca a la población Ocupada, dado que no asiste actualmente a la escuela y no le fue aplicable la pregunta sobre antecedentes laborales.

$$P(CLASE2 = 1 | D\_CEXP\_EST = 0, CS\_P17 = 2) = 0.46 \tag{4.3}$$

4. Probabilidad de que una mujer pertenezca a la población No disponible, dado que no asiste actualmente a la escuela y no le fue aplicable la pregunta sobre antecedentes laborales.

$$P(CLASE2 = 4 | D\_CEXP\_EST = 0, CS\_P17 = 2) = 0.49 \tag{4.4}$$

Las anteriores son algunos ejemplos en donde, de acuerdo a los datos, existe más de un valor posible para CLASE2 dadas las instancias de sus padres. También se dan situaciones en las que para una instancia de padres $(X_1 = x_1, X_2 = x_2)$ sólo hay un valor posible de la variable de interés, en este caso CLASE2. Por lo anterior, existen instancias de una variable y sus padres que tienen probabilidad 1. A saber,

$$P(CLASE2 = 2|D\_CEXP\_EST = 1, CS\_P17 = 1) = 1 \quad (1)$$
$$P(CLASE2 = 2|D\_CEXP\_EST = 1, CS\_P17 = 2) = 1 \quad (2)$$
$$P(CLASE2 = 2|D\_CEXP\_EST = 2, CS\_P17 = 1) = 1 \quad (3)$$
$$P(CLASE2 = 2|D\_CEXP\_EST = 2, CS\_P17 = 2) = 1 \quad (4)$$
$$P(CLASE2 = 2|D\_CEXP\_EST = 3, CS\_P17 = 2) = 1 \quad (5)$$

Las ecuaciones anteriores representan las probabilidades de que una mujer pertenezca a la población desocupada dado que: (1) perdió o terminó su empleo anterior y asiste a la escuela actualmente; (2) perdió o terminó su empleo anterior y no asiste actualmente a la escuela; (3) hubo insatisfacción con el empleo anterior y asiste actualmente a la escuela; (4) hubo insatisfacción con el empleo anterior y no asiste actualmente a la escuela; (5) dejó o cerró un negocio propio y no asiste actualmente a la escuela.

Con la información anterior, puede decirse que sabiendo que una mujer asiste a la escuela y no pertenece a población desocupada; es 2.7 veces más probable que pertenezca a la población no disponible que a la ocupada. Por otro lado, si la condición es que no asista actualmente a la escuela, la probabilidad de que esté no disponible, es apenas 0.06 veces mayor. Además, debido a que la variable D_CEXP_EST sólo toma valores dentro de la clase de mujeres ocupadas, no es posible obtener información acerca de las probabilidades cuando el cálculo se centra en esta población.

Figura 4.14: Mapa de calor de probabilidades de CS_P17 y CLASE2.

En la Figura 4.14 se muestra con colores el tamaño de la probabilidad asociada a las posibles combinaciones de variables CS_P17 y CLASE2, cuando se fija D_CEXP_EST al valor *No aplica*. Se puede inferir que generalmente es más probable que una mujer no asista a la escuela, en cualquiera de sus estados ocupacionales. Además, quienes más probabilidad tienen de hacerlo son las mujeres no disponibles ocupacionalmente. Lo anterior, relativo a que no apliquen para ella antecedentes laborales.

4.3.2. Análisis a variable P11_H7

Al igual que en el caso anterior, se calculan algunas probabilidades condicionales de la variable P11_H7 (horas semanalaes dedicadas a tareas del hogar), que se relaciona estrechamente con la variable de interés.

De este modo, se calcula:

1. Probabilidad de que una mujer gaste de 0 a 15 horas en los quehaceres del hogar, dado que está casada y pertenece a la población ocupada.

$$P(P11\_H7 = 1|E\_CON = 5, CLASE2 = 1) = 0.31 \tag{4.5}$$

2. Probabilidad de que una mujer gaste de 0 a 15 horas en los quehaceres del hogar, dado que está casada y se clasifica ocupacionalmente como No disponible.

$$P(P11\_H7 = 1|E\_CON = 5, CLASE2 = 3) = 0.06 \tag{4.6}$$

3. Probabilidad de que una mujer gaste de 0 a 15 horas en los quehaceres del hogar, dado que está soltera y pertenece a la población ocupada.

$$P(P11\_H7 = 1|E\_CON = 6, CLASE2 = 1) = 0.65 \tag{4.7}$$

4. Probabilidad de que una mujer gaste de 30 a 45 horas en los quehaceres del hogar, dado que está casada y pertenece a la población ocupada.

$$P(P11\_H7 = 3|E\_CON = 5, CLASE2 = 1) = 0.14 \tag{4.8}$$

5. Probabilidad de que una mujer gaste de 30 a 45 horas en los quehaceres del hogar, dado que está soltera y pertenece a la población ocupada.

$$P(P11\_H7 = 3|E\_CON = 6, CLASE2 = 1) = 0.04 \tag{4.9}$$

En primer lugar, las ecuaciones (4.5) y (4.6) muestran que la probabilidad de que una mujer gaste de 0 a 15 horas (que representa el menor tiempo posible, de acuerdo a las clasificaciones) a los quehaceres del hogar, aumenta hasta 5.2 veces cuando la persona en cuestión pertenece a la población ocupada, que cuando es no disponible; más aún este aumento se duplica cuando en vez de considerar el estado conyugal como casada se considera a la mujer como soltera (4.7).

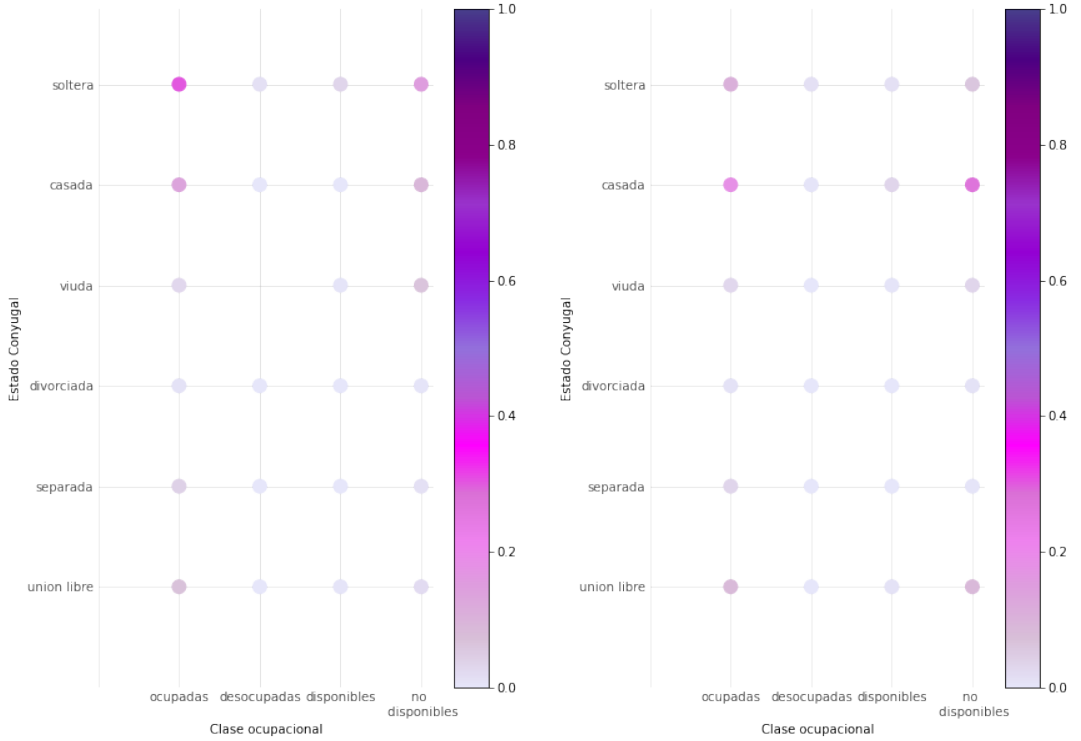
Aunado a lo anterior, las probabilidades calculadas en (4.8) y (4.9) dejan ver que si se considera a las mujeres ocupadas; la probabilidad de que dediquen de 30 a 45 horas a la semana en tareas del hogar, es 3 veces mayor cuando su estado conyugal es casada que cuando es soltera.

Se calcularon las probabilidades para mostrar gráficamente la distribución de las frecuencias de valores entre E_CON y P11_H7, fijando en cada caso a la variable CLASE2, Figura 4.15.

De acuerdo a la Figura 4.15, puede verse de manera general que las mujeres solteras, divorciadas y separadas dedican menos horas al hogar que las mujeres en unión libre o

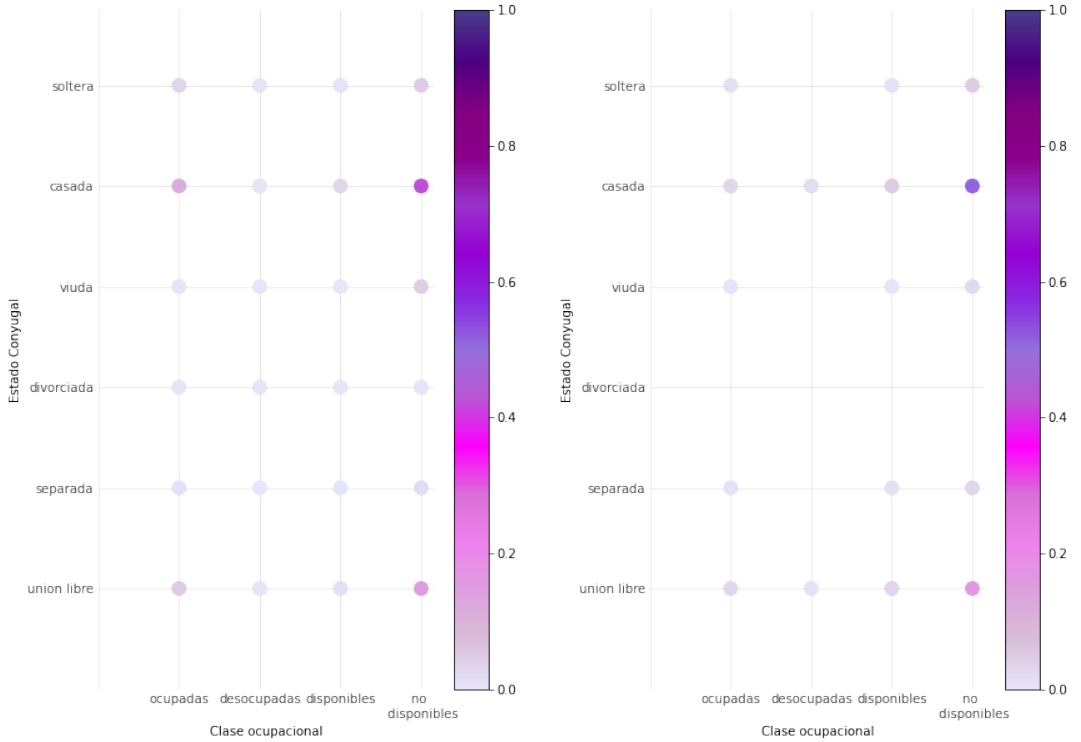
casadas. Estas últimas son las que más altas probabilidades tienen de dedicar más de 15 horas a los quehaceres del hogar, incluso perteneciendo a la clase Ocupada.

Asimismo, es más probable que las mujeres Ocupadas dediquen de 0 a 30 horas a los quehaceres del hogar; las Desocupadas pueden dedicar casi con igual probabilidad de 0 a 45 horas; las Disponibles pueden dedicar casi igualmente cualquier número de horas de acuerdo a los rangos, aunque las casadas Disponibles tienen ligeramente más probabilidad en dedicar 45 horas o más. Finalmente, las no disponibles tienen más probabilidad de 30 a 45 horas en la semana a los quehaceres del hogar.



(a) ***Mujeres que en una semana dedicaron de 0 a 15 horas a quehaceres del hogar***

(b) ***Mujeres que que en una semana dedicaron de 15 a 30 horas a quehaceres del hogar***



(c) ***Mujeres que en una semana dedicaron de 30 a 45 horas a quehaceres del hogar***

(d) ***Mujeres que en una semana dedicaron más de 45 horas a quehaceres del hogar***

Figura 4.15: Mapas de calor de probabilidades de CLASE2 y E_CON fijando P11_H7.

4.3.3. Análisis a variable P2F

Por último, se calculan algunos ejemplos de parámetros para la variable descendiente directa de CLASE2, P2F, que representa la necesidad o no necesidad de empleo de las encuestadas. En este caso, otra variable a tomar en cuenta es P11_H4 (horas dedicadas a otro miembro del hogar), pues pertenece al conjunto de padres de P2F.

1. Probabilidad de que una mujer tenga necesidad de trabajar, dado que dedicó de 0 a 4 horas (en la semana anterior a la de referencia) a llevar a algún miembro del hogar a la escuela, cita médica u otra actividad; y pertenece a la clase ocupacional disponible.

$$P(P2F = 1|P11\_H4 = 1, CLASE2 = 3) = 0.8 \tag{4.10}$$

2. Probabilidad de que una mujer tenga necesidad de trabajar, dado que dedicó de 8 a 12 horas (en la semana anterior a la de referencia) a llevar a algún miembro del hogar a la escuela, cita médica u otra actividad; y pertenece a la clase ocupacional Disponible.

$$P(P2F = 1|P11\_H4 = 3, CLASE2 = 3) = 0.67 \tag{4.11}$$

3. Probabilidad de que una mujer no tenga necesidad de trabajar, dado que dedicó de 0 a 4 horas (en la semana anterior a la de referencia) a llevar algún miembro del hogar a la escuela, cita médica u otra actividad; y pertenece a la clase ocupacional Disponible.

$$P(P2F = 2|P11\_H4 = 1, CLASE2 = 3) = 0.17 \tag{4.12}$$

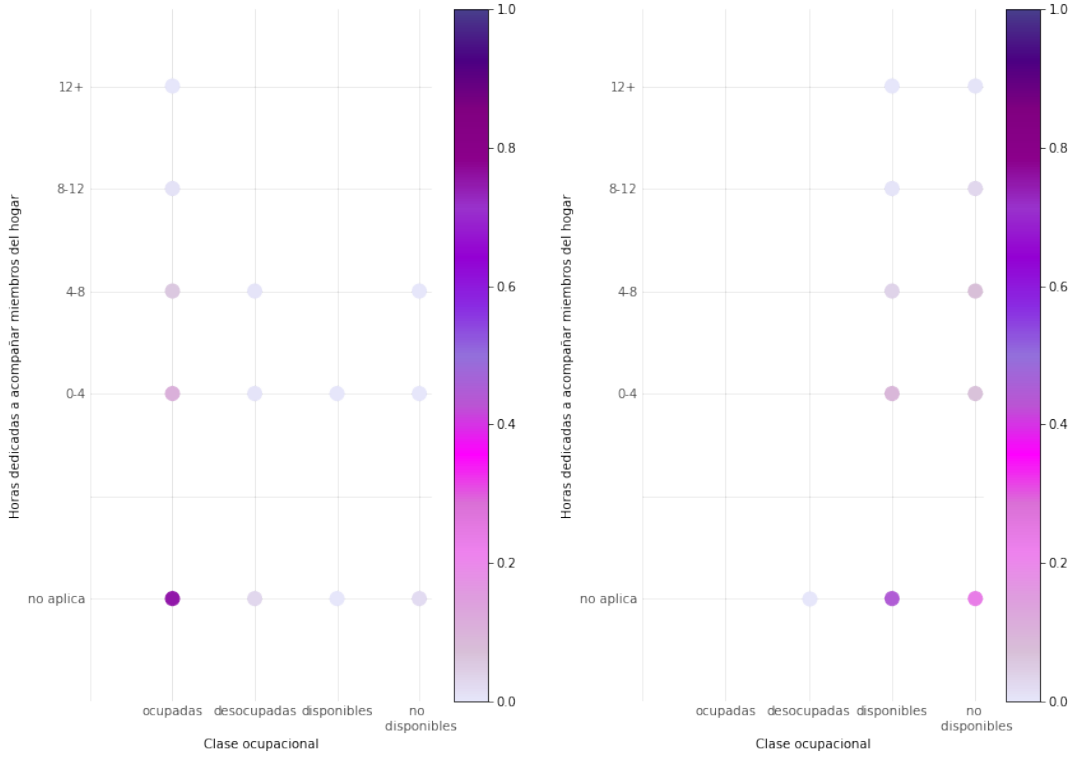
4. Probabilidad de que una mujer no tenga necesidad de trabajar, dado que dedicó de 8 a 12 horas (en la semana anterior a la de referencia) a llevar algún miembro del hogar a la escuela, cita médica u otra actividad; y pertenece a la clase ocupacional Disponible.

$$P(P2F = 2|P11\_H4 = 3, CLASE2 = 3) = 0.33 \tag{4.13}$$

De acuerdo a lo anterior, es más probable que una mujer tenga mayor probabilidad de tener necesidad de trabajar, dado que se encuentra disponible para trabajar, cuando dedica de 0-4 horas a algún miembro del hogar, que cuando dedica de 8 a 12 horas. La diferencia es del 19 %.

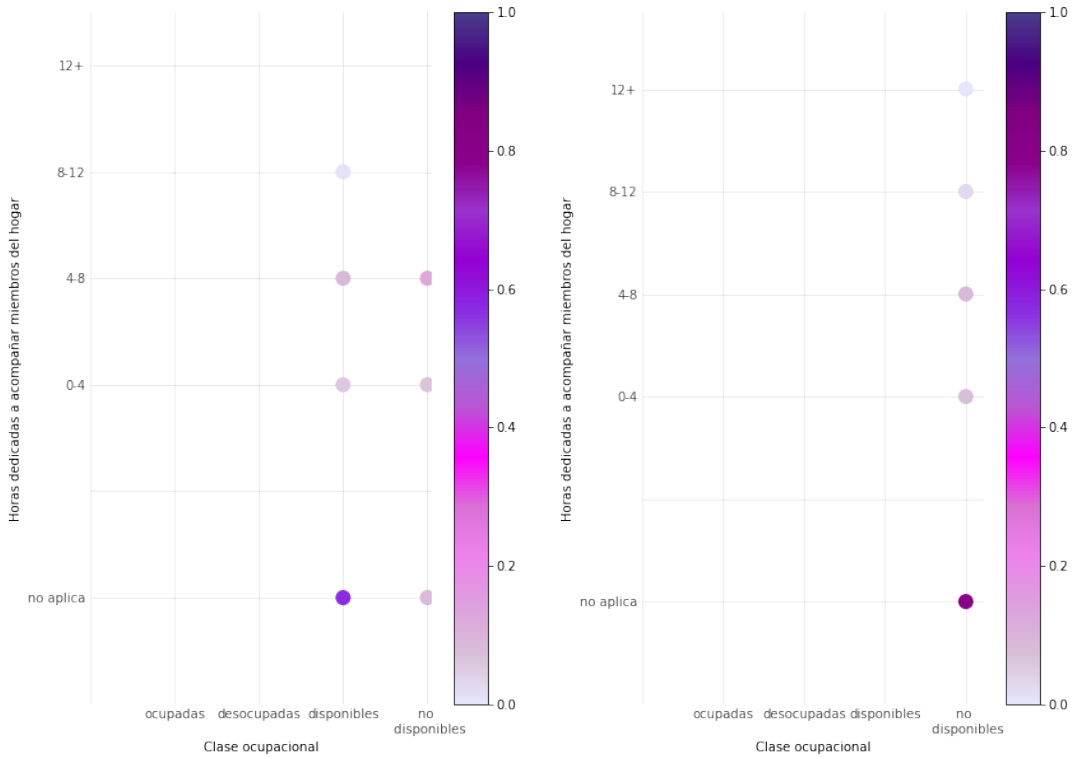
El caso anterior es al revés cuando el evento de interés es que la persona en cuestión no tenga necesidad de trabajar, como se esperaría después del resultado de arriba.

La Figura 4.16(a) muestra que la pregunta no es aplicable en mayor medida a las mujeres que ya tienen un trabajo, resultado esperado de la encuesta. A su vez, la Figura 4.16(b) muestra que es más probable que las mujeres que gastan menos horas en llevar a miembros del hogar a escuela, cita médica o cualquier otro lugar son las que más probablemente tienen necesidad de trabajar, junto con las que no se les aplica la pregunta de acompañar miembros del hogar. Además las mujeres disponibles y no disponibles tienen probabilidades parecidas de necesitar trabajar, excepto las muestras de *No aplica*.



(a) ***Mujeres para las que la pregunta sobre necesitar trabajo no fue aplicable***

(b) ***Mujeres con necesidad de trabajar***



(c) ***Mujeres sin necesidad de trabajar***

(d) ***Mujeres con deseos de trabajar***

Figura 4.16: Mapas de calor de probabilidades de CLASE2 y P11_H4 fijando P2F.

De acuerdo con la Figura 4.16(c). Las mujeres disponibles y no disponibles que dedican de 0 a 8 horas a acompañar a los miembros del hogar son las que mayor probabilidad tienen de no necesitar trabajar. Finalmente, vemos en la Figura 4.16(d) que las mujeres No disponibles son las más probables a tener deseos de trabajar, sin importar el tiempo que dediquen a los miembros de su hogar.

Los párrafos anteriores son ejemplos específicos de la estimación de parámetros y su interpretación *grosso modo*. El análisis puede ser extendido, pero no es el objetivo de este proyecto. Puede realizarse también un análisis usando diferentes consultas de probabilidad donde las variables no necesariamente tienen una relación padre-hijo, las ideas generales de ese tipo de inferencia son plasmadas en el Apéndice C, ya que dicho trabajo es un proyecto a futuro de esta tesis.

Capítulo 5

Discusión

Conociendo los resultados anteriores, a continuación se habla de algunos puntos importantes a discutir de este proyecto.

El tratamiento de los datos fue clave en el ordenamiento de las variables a ingresar en el algoritmo K2, y por tanto en la construcción de la red. Las primeras dos variables del ordenamiento fueron las de menor variabilidad de acuerdo a los valores que tomaron; pues la variable de Clasificacion de antecedentes laborales (D_CEXP_EST) solo toma los valores 1, 2 y 3 en la minoría de los casos (población desocupada) y "no aplica" en la mayoría de los datos (alrededor del 98 %).

Este ejemplo deja claro que el método de ordenamiento basándose en la entropía condicional es sensible a las muestras desbalanceadas; ya que D_CEXP_EST variable extremadamente desbalanceada logró tener la menor entropía gracias a que en la mayoría de los casos toma el mismo valor.

La variable de *Asistencia actualmente a la escuela* (CS_P17) es un caso similar al anterior, pues debido a que la mayoría de las mujeres de la muestra no asiste actualmente a la escuela, la variable en cuestión toma el valor "No" en el 91.3 % de la muestra. Por consiguiente, dicha característica carece de incertidumbre respecto a las demás, y es candidata a ser de las primeras en el ordenamiento con entropía condicional.

A su vez, como ya había sido señalado en (1), el ordenamiento jugó un papel crucial en la formación de la estructura de la red. En este caso, las variables con menos incertidumbre mencionadas arriba fueron los nodos padre de la variable del Estado Ocupacional, CLASE2; además de haberse convertido en ancestros de la mayoría de los nodos restantes.

El algoritmo K2 fue asequible para ser usado con las variables y los datos seleccionados; su ejecución es lenta cuando se utilizan variables con muchas categorías; así mismo, el ordenamiento de variables con entropía condicional es sensible a clases desbalanceadas.

Como mejora en el experimento, podría asignarse las variables con clases desbalanceadas al final de orden introducido en K2, o en un nivel más bajo de las variables de más interés; ya que al menos en este proyecto las variables desbalanceadas bloquearon en algunos casos la posibilidad de realizar un mejor análisis de la red, tener variables

balanceadas o sin la categoría “No aplica” hubiera aportado más información sobre lo que se quería investigar.

Capítulo 6

Conclusiones

En este trabajo se plantea la construcción de una red bayesiana con datos del primer trimestre de la ENOE 2019, con el objetivo de analizar algunas relaciones entre las variables socio-demográficas de las mujeres mexiquenses mayores de 18 años y sus estados ocupacionales. Obteniendo una red bayesiana de 13 nodos y 21 arcos; a partir de utilizar el algoritmo K2 para la construcción de la estructura de la red y el Estimador Máximo Verosímil para calcular algunos de sus parámetros.

Podemos afirmar que se alcanzó el objetivo de investigar y entender las Redes Bayesianas. Más aún, se logró adaptar dicho modelo a los datos de la ENOE dando posibles mejoras al procedimiento.

De los resultados de este proyecto, se concluye que las variable correspondientes al número de hijos y el estado conyugal de las encuestadas, fueron las que más relaciones directas tuvieron en toda la red. Cada una con 5 relaciones de padre o hijo con las demás. Lo anterior puede interpretarse como que dichos factores son los que más impactan o son impactadas (directamente) por las demás características económicas consideradas en este trabajo, tales como: la necesidad de trabajar; el número de horas dedicada a atender sin pago a niños, ancianos, enfermos, etc.; la edad; grado hasta el que aprobó en la escuela; las horas dedicadas a realizar compras o trámites para el hogar; asistencia actual a la escuela y las horas dedicadas a realizar quehaceres del hogar.

La variable de ocupación es directamente influenciada por los antecedentes laborales, la asistencia a la escuela, las horas dedicadas a los quehaceres del hogar y la necesidad de trabajar. En términos generales puede existir alguna relación de dependencia condicional entre la variable de ocupación y cualquier otra en el experimento; dependiendo de si algunas otras características son valores conocidos o no.

De manera más específica, se halló que para las mujeres mexiquenses mayores de 18 años de edad:

- Independientemente del tiempo dedicado a los miembros del hogar, la población no disponibles tiene alta probabilidad de querer trabajar.
- Las mujeres disponibles y no disponibles tienen probabilidad parecida de necesitar un trabajo.

- Las mujeres casadas y en unión libre dedican más tiempo a los quehaceres del hogar con más alta probabilidad que las solteras, divorciadas y viudas; incluso si son Ocupadas.

Como comentario final sobre los datos del 1° trimestre de la ENOE 2019, fue complicado realizar el análisis tomando en cuenta los valores *No aplica*; ya que a pesar de ser datos faltantes no aleatorios, no proporcionan información sobre los casoso en los que se presentan; en la documentación de la ENOE no está claro cuál es su significado o de qué manera son asignados.

A modo de cierre, se puede aseverar que se cumplió con el objetivo inicial de la tesis respecto a obtener relaciones importantes entre variables sociales y económicas de las mexiquenses, además de identificar algunos factores de influencia en su condición de ocupación.

Apéndice A

Municipios y Localidades muestra

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
002	Acolman	0001	Acolman de Nezahualcóyotl
		0005	San Bartolo
		0012	Santa Catarina
		0015	Tepexpan
011	Atenco	0001	San Salvador Atenco
		0002	San Cristóbal Nexquipayac
		0004	Santa Isabel Ixtapan
		0013	Nueva Santa Rosa
		0029	Granjas Ampliación Santa Rosa
013	Atizapán de Zaragoza	0001	Ciudad López Mateos
020	Coacalco de Berriozábal	0001	San Francisco Coacalco
023	Coyotepec	0001	Coyotepec
024	Cuautitlán	0001	Cuautitlán
		0088	San Mateo Ixtacalco
		0111	Galaxia Cuautitlán
024	Cuautitlán	0124	La Providencia
025	Chalco	0001	Chalco de Díaz Covarrubias
		0010	San Juan Tezompa
		0013	San Marcos Huixtoco

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
028	Chiautla	0001	Chiautla
		0004	Ocopulco
		0006	Santiago Chimalpa (Chimalpa)
029	Chicoloapan	0001	Chicoloapan de Juárez
030	Chiconcuac	0001	Chiconcuac de Juárez
031	Chimalhuacán	0001	Chimalhuacán
033	Ecatepec de Morelos	0001	Ecatepec de Morelos
037	Huixquilucan	0001	Huixquilucan de Degollado
		0005	Dos Ríos
		0009	Jesús del Monte
		0021	San Francisco Ayotuzco
		0023	San Juan Yautepec
		0025	Santiago Yancuitlalpan
		0026	Zacamulpa
		0071	Naucalpan de Juárez
039	Ixtapaluca	0001	Ixtapaluca
		0003	San Buenaventura
		0012	San Francisco Acuautla
		0064	Jorge Jiménez Cantú
044	Jaltenco	0001	Jaltenco
		0020	Alborada Jaltenco
053	Melchor Ocampo	0001	Melchor Ocampo
		0005	San Francisco Tenopalco
057	Naucalpan de Juárez	0001	Naucalpan de Juárez
		0267	Ejido de San Francisco Chimalpa
058	Nezahualcóyotl	0001	Ciudad Nezahualcóyotl
059	Nextlalpan	0001	Santa Ana Nextlalpan

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
060	Nicolás Romero	0001	Villa Nicolás Romero
		0016	Progreso Industrial
		0082	Veintidós de Febrero
069	Papalotla	0001	Papalotla
070	La Paz	0001	Los Reyes Acaquilpan
		0005	La Magdalena Atlicpac
		0008	San Sebastián Chimalpa
		0009	Tecamachalco
		0013	Emiliano Zapata
		0017	Profesor Carlos Hank González
		0019	El Pino
		0036	Arenal
		0037	Bosques de la Magdalena
		0038	Lomas de San Sebastián
		0039	Lomas de Altavista
		0040	San Isidro
		0041	San José las Palmas
		0042	Techachaltitla
		0043	Unidad Acaquilpan
081	Tecámac	0001	Tecámac de Felipe Villanueva
		0009	San Pablo Tecalco
		0019	Ojo de Agua
		0025	San Martín Azcatepec
		0098	Fracc. Social Progresivo San. Tomás Chiconautla
091	Teoloyucan	0001	Teoloyucan
		0010	San Bartolo

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
092	Teotihuacán	0001	Teotihuacán de Arista
		0002	Atlatongo
		0019	San Lorenzo Tlalmimilolpan
093	Tepetlaoxtoc	0001	Tepetlaoxtoc de Hidalgo
		0003	Concepción Jolalpan
095	Tepotzotlán	0001	Tepotzotlán
		0021	San Mateo Xoloc
		0026	Santiago Cuautlalpan
		0073	Santa Cruz del Monte
		0074	Ejido de Coyotepec
099	Texcoco	0001	Texcoco de Mora
		0012	Montecillo
		0016	La Purificación Tepetitla
		0020	San Bernardino
		0025	San Joaquín Coapango
		0029	San Miguel Coatlinchán
		0030	San Miguel Tlaixpán
		0042	Santiago Cuautlalpan
		0045	Santa María Tulantongo
		0048	Xocotlán
100	Tezoyuca	0001	Tezoyuca
		0002	Tequisistlán
		0007	Ejido de Tequisistlán Primero
104	Tlalnepantla de Baz	0001	Tlalnepantla
		0105	Puerto Escondido (Tepeolulco Puerto Escondido)

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
108	Tultepec	0001	Tultepec
		0014	Santiago Teyahualco
		0063	Fraccionamiento Paseos de Tultepec II
109	Tultitlán	0001	Tultitlán de Mariano Escobedo
		0003	Buenavista
		0025	San Pablo de las Salinas
		0068	Fuentes del Valle
		0069	Ampliación San Mateo (Colonia Solidaridad)
		0072	Colonia Lázaro Cárdenas (Los Hornos)
120	Zumpango	0001	Zumpango de Ocampo
		0054	San Sebastián
121	Cuautitlán Izcalli	0001	Cuautitlán Izcalli
		0020	Huilango
122	Valle de Chalco Solidaridad	0001	Xico
125	Tonanitla	0001	Santa María Tonanitla
005	Almoloya de Juárez	0032	San Francisco Tlalcilalcalpan
018	Calimaya	0006	San Lorenzo Cuauhtenco
		0008	Santa María Nativitas
051	Lerma	0001	Lerma de Villada
		0024	San Pedro Tultepec
		0086	Colonia los Cedros

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
054	Metepec	0001	Metepec
		0032	San Bartolomé Tlaltelulco
		0034	San Francisco Coaxusco
		0035	San Gaspar Tlahuelilpan
		0037	San Jerónimo Chicahualco
		0038	San Jorge Pueblo Nuevo
		0041	San Lorenzo Coacalco (San Lorenzo)
		0043	San Lucas Tunco (San Lucas)
		0047	San Miguel Totocuitlapilco
		0049	San Salvador Tizatlalli
		0052	San Sebastián
		0055	Santa María Magdalena Ocotitlán
055	Mexicaltzingo	0001	San Mateo Mexicaltzingo
067	Otzolotepec	0005	Colonia Guadalupe Victoria
		0050	Ejido de la Y Sección Siete A Revolución
076	San Mateo Atenco	0001	San Mateo Atenco
		0012	Santa María la Asunción
		0001	Toluca de Lerdo
		0043	Cacalomacán
		0044	Calixtlahuaca
		0050	El Cerrillo Vista Hermosa
		0051	La Constitución Toltepec
		0055	Jicaltepec Cuexcontitlán
		0062	San Andrés Cuexcontitlán
		0063	San Antonio Buenavista
		0068	San Diego de los Padres Cuexcontitlán

Municipio		Localidad	
Clave	**Nombre**	**Clave**	**Nombre**
106	Toluca	0070	San Felipe Tlalmimilolpan
		0072	San José Guadalupe Otzacatipan
		0077	San Marcos Yachihuacaltepec
		0079	San Mateo Otzacatipan
		0082	San Nicolás Tolentino
		0083	San Pablo Autopan
		0084	San Pedro Totoltepec
		0088	Santa Cruz Otzacatipan
		0112	San Miguel Totoltepec
		0127	Jicaltepec Autopan
		0155	San Diego los Padres Cuexcontitlán Sección 5B
		0194	Barrio Santa Cruz
118	Zinacantepec	0001	San Miguel Zinacantepec
		0053	San Antonio Acahualco
		0058	San Juan de las Huertas
		0059	Ejido San Lorenzo Cuauhtenco
		0064	Santa Cruz Cuauhtenco
		0075	Tejalpa

Apéndice B

Imputación de datos

B.1. Imputación Múltiple

La imputación de datos es el proceso de identificar datos faltantes y sustituirlos por algún valor. Desarrollada en los años 70's por Donald B. Rubin para resolver un problema de encuestas, la Imputación Múltiple es un método que crea varias imputaciones y refleja la incertidumbre de los datos faltantes.

El procedimiento general es crear $m > 1$ conjuntos de datos completos, utilizando métodos de simulación Monte Carlo. Vale la pena mencionar que este tipo de métodos asumen que los datos faltantes pueden ser MAR (*Missing at Random*: datos faltantes aleatorios que pueden tener diferencia con los valores observados, pero esta diferencia puede ser explicada completamente por los datos observados) o MCAR (*Missing Completely at Random*: datos faltantes completamente aleatorios; no hay diferencia con datos observados).

La imputación comienza con un conjunto de datos incompleto. Primero se generan m conjuntos de datos a partir del original pero con los datos faltantes imputados, estas imputaciones no necesariamente son iguales en cada uno. De esta forma, los m conjuntos de datos obtenidos se diferenciarán únicamente en las observaciones imputadas.

El siguiente paso es ajustar los parámetros de interés para cada uno de los *datasets* generados. Hay que tener en cuenta que los ajustes para cada uno serán diferentes, puesto que las observaciones imputadas lo son.

La última parte será presentar una sola estimación a partir de los m parámetros ajustados y calcular su varianza. Idealmente la estimación final será insesgada, van Buuren (25). Las etapas se muestran de manera gráfica para $m = 3$ en el diagrama siguiente.

Debido a que se crean varios conjuntos de datos imputados, con este método puede medirse la incertidumbre de los valores imputados. Esta es una de las mejores características de la Imputación Múltiple.

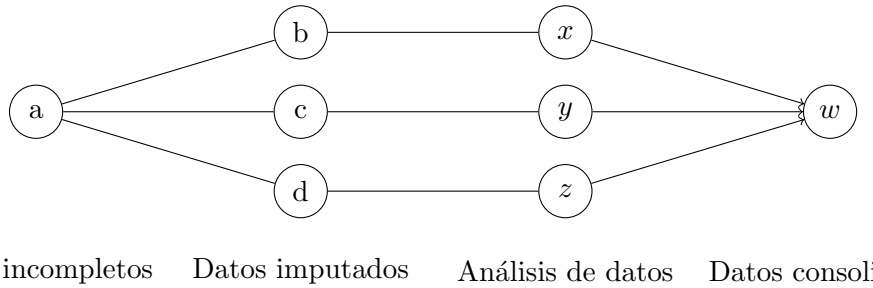


Figura B.1: Proceso de imputación de datos múltiple.

B.1.1. Método de imputación MICE

MICE (Multivariate Imputation by Chained Equation) es una forma particular de realizar el método de Imputación Múltiple; este procedimiento se basa en ecuaciones en cadena; y sirve para datos continuos, binarios y categóricos (tanto ordenados como desordenados).

Para cualquier modelo científico de interés, Q, si se tienen p variables incompletas, con $j = 1, ..., p$, denotamos a $X = (X_1, X_2, ..., X_p)$. Se denota como:

$$X^{obs} = (X_1^{obs}, X_2^{obs}, ..., X_p^{obs})$$

al conjunto de variables observadas y como

$$X^{mis} = (X_1^{mis}, X_2^{mis}, ..., X_p^{mis})$$

al conjunto de variables con datos faltantes. Finalmente, X_{-j} denota a todo el conjunto de variables X menos la variable j-ésima.

El modelo de *ecuaciones en cadena* permite que las columnas del *dataset* sean imputadas una a la vez y de forma separada. La hipótesis de este algoritmo es que X es una muestra de la distribución multivariada $P(X|\theta)$, que está completamente determinada por el vector de parámetros desconocidos θ.

El procedimiento MICE estima la distribución a posteriori de θ mediante el muestreo iterativo de distribuciones condicionales:

$$\begin{gathered} P(X_1|X_{-1}, \theta_1) \\ P(X_2|X_{-2}, \theta_2) \\ \vdots \\ P(X_p|X_{-p}, \theta_p) \end{gathered}$$

Cada uno de los parámetros θ_i es específico respecto a su densidad condicional. Empezando de una distribución marginal observada, la t-ésima iteración de las ecuaciones de cadena es un muestreo de Gibbs de la forma:

$$
\begin{aligned}
&\theta_1^{*(t)} \sim P(\theta_1 | X_1^{obs}, X_2^{t-1}, ..., X_p^{t-1}) \\
&X_1^{*(t)} \sim P(X_1 | X_1^{obs}, X_2^{t-1}, ..., X_p^{t-1}, \theta_1^{*(t)}) \\
&\vdots \\
&\theta_p^{*(t)} \sim P(\theta_p | X_p^{obs}, X_1^{t}, ..., X_p^{t}) \\
&X_p^{*(t)} \sim P(X_p | X_p^{obs}, X_1^{t}, ..., X_p^{t}, \theta_p^{*(t)})
\end{aligned}
$$

donde $X_j^{(t)} = (X_j^{obs}, X_j^{*(t)})$ es la j-ésima variable imputada en el tiempo t.

Azur et al. (4) explican el procedimiento general MICE en 6 pasos:

1. Se realiza una imputación simple a todos los datos faltantes del conjunto de datos (utilizando media, moda, etc.)
2. Para una variable X_i se eliminan los datos que fueron imputados en el paso anterior.
3. Se ajusta un modelo tomando como variable respuesta a X_i y usando únicamente las observaciones para las cuales X_i no tiene valores faltantes.
4. Las observaciones faltantes de X_i se rellenan con predicciones del modelo generado en el paso anterior. Cuando la variable X_i es usada posteriormente como predictora de otra, se utilizan los valores imputados como observados.
5. Los pasos 2-4 se repiten hasta que todas las variables con valores faltantes sean imputadas. Una vez cumplido este cometido, se dice que que se ha completado un **ciclo**.
6. Los pasos 2-4 se imputan por cierto número de ciclos actualizando los valores imputados en cada iteración.

A continuación se muestra el procedimiento MICE con un ejemplo simple. La tabla de datos a imputar, D, es de la siguiente forma:

X_1	X_2	X_3
1	2	
	3	1
2	3	1
1	1	3

Tabla B.1: Tabla de datos D

En el primer paso del algoritmo se imputan valores de acuerdo a una medida como la moda, por ejemplo:

X_1	X_2	X_3
1	2	1
1	3	1
2	3	1
1	1	3

Tabla B.2: Tabla de datos $\hat{D_1}$

utilizando el conjunto $\hat{D_1}$ sin el segundo renglón; ya que es ahí donde se encuentra la observación que se quiere imputar; se ajusta un modelo con parámetro estimado $\hat{\theta}$ (puede ser un vector). Con este modelo se predice el valor (o valores faltantes para X_1), supongamos que:

$$\hat{X_1} = 2$$

por lo que se tiene un segundo conjunto de datos, $\hat{D_2}$ de la forma:

X_1	X_2	X_3
2	2	1
1	3	1
2	3	1
1	1	3

Tabla B.3: Tabla de datos $\hat{D_2}$

posteriormente, se sigue el procedimiento para X_3. Se elimina el renglón de la observación imputada para esta variable, que en este caso es el primero. Quedando el conjunto de datos de la forma:

X_1	X_2	X_3
1	3	1
2	3	1
1	1	3

Tabla B.4: Tabla de datos $\hat{D_2}$ sin primera observación

Y nuevamente se ajusta un modelo, pero ahora la variable respuesta será X_3. Hecho lo anterior, se predice el valor faltante, supongamos que de acuerdo al modelo obtenido:

$$\hat{X_3} = 1$$

por lo que la tabla, se vería de la forma:

X_1	X_2	X_3
2	2	1
1	3	1
2	3	1
1	1	3

Tabla B.5: Tabla de datos $\hat{D_3}$

El procedimiento anterior fue un ciclo en la imputación, deben hacerse tantos como sean necesarios para que la cadena de valores estimados converja. Después del primer ciclo ya no se generan valores iniciales de las observaciones faltantes, pues se usa el valor estimado del ciclo anterior.

De acuerdo con observaciones empíricas de van Buuren and Groothuis-Oudshoorn (26), la convergencia puede darse desde la veinteava iteración. Además, los autores mencionan que MICE puede ser aplicado a datos del tipo MNAR (aunque no es el caso general de los métodos de imputación múltiple); pero en dicho caso se debe realizar un análisis de sensibilidad de la imputación.

B.1.2. MICE en los datos

Primeramente, todo el proceso de imputación se hace bajo la suposición de que los datos faltantes son del tipo MAR.

Se generaron cinco conjuntos de datos imputados, lo que quiere decir que se hicieron cinco cadenas de valores para cada variable imputada. En las Figuras B.2 y B.3 se muestran las trazas correspondientes a la imputación en cada ejercicio de cada variable. Cada color representa un ejercicio. Lo ideal es que las trazas sean aleatorias y sin tendencia.

Puede observarse en las Figuras B.2 y B.3 que exceptuando a la gráfica de la variable *E_CON*, las trazas tienen el comportamiento deseado; es decir, no siguen algún patrón definido o tendencia. Por lo que se concluye que se alcanzó una convergencia suficiente en las cadenas generadas. Respecto a las trazas de la variable *E_CON* (Estado conyugal de la persona), cuyos valores posibles son: 1, 2, 3, 4, 5 y 6; para las respuestas *Unión libre*, *Separada*, *Divorciada*, *Viuda*, *Casada* y *Soltera*; respectivamente; la desviación de los valores no tiene sentido, ya que se tiene sólo una observación. Se decidió no prestar atención en este suceso, ya que para esta variable solamente debe imputarse una observación.

Es importante mencionar el hecho de que la paquetería en R donde se implementa *MICE*, permite designar las variables del conjunto de datos que se usarán para imputar alguna otra variable del conjunto. Dado lo anterior, se añadieron variables auxiliares para *P11_H2*, *P11_H3* y *P11_H7* y dichas auxiliares se tomaron en cuenta únicamente para imputar a sus correspondientes; es decir, no se condicionó a otra variable diferente de *P11_HX* sobre *P11_HX_auxiliar*. Lo que se traduce en que estas variables auxiliares

no tienen efecto en la imputación de las que no sean la que le corresponde.

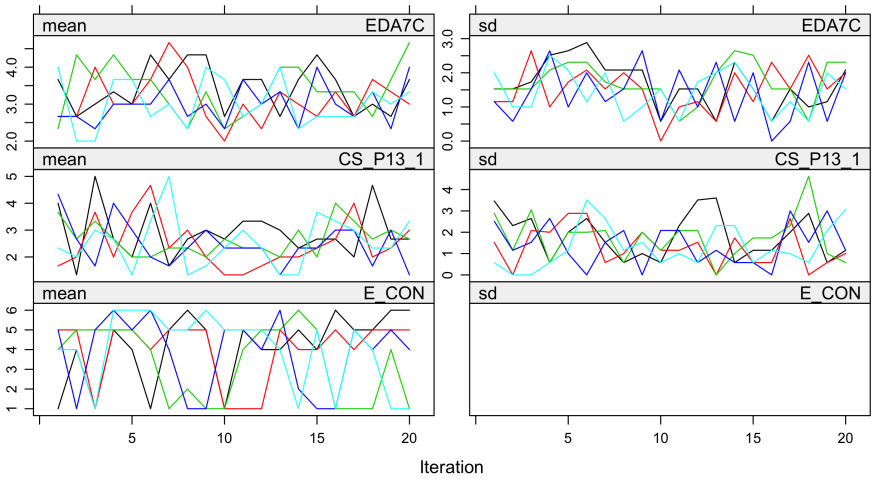


Figura B.2: Trazas de muestreos para las variables Edad, Clasificación por grados aprobados en la escuela y Estado conyugal.

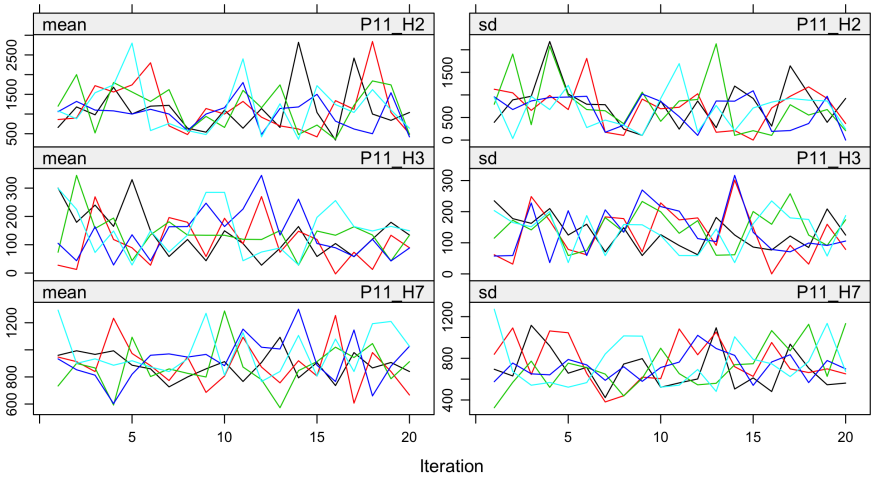


Figura B.3: Trazas de muestreos para las variables: Horas dedicadas a atender sin pago, de manera exclusiva a niños, ancianos, personas enfermas o con discapacitación; Horas que dedicó a realizar compras, llevar cuentas o realizar trámites para el hogar, o encargarse de la seguridad; así como Horas que dedicó a realizar los quehaceres del hogar.

Apéndice C

Inferencia Probabilística

En este apartado se introduce al concepto de inferencia en las redes bayesianas y se habla brevemente del problema principal que conlleva dicho procedimiento, así como algunas de sus soluciones.

Si se asume que se conoce la estructura de la red y suponiendo que $\mathbf{X}$ es un conjunto de variables de la red $\{X_1, X_2, ...X_k\}$ y a su vez $\mathbf{E}$ es un conjunto de variables $\{E_1, ..., E_j\}$ que toma los valores $\{e_1, ..., e_j\}$; uno puede focalizarse en realizar consultas del tipo: $P(\mathbf{X}|\mathbf{E} = \mathbf{e})$ a la red bayesiana, es decir, consultas sobre independencia condicional (15). En palabras de Sucar (24), la Inferencia Probabilística consiste en propagar los efectos de cierta evidencia en una red bayesiana para estimar su efecto en variables desconocidas.

Puede originarse dos casos de inferencia, dependiendo del conjunto $\mathbf{X}$. El caso en que $|\mathbf{X}| = 1$, es nombrado "una sola consulta"; mientras que en el caso en que $|\mathbf{X}| \geq 1$, se nombra muchas veces como "consulta de inferencia conjunta". En teoría puede usarse la misma red bayesiana para realizar las consultas antes mencionadas, pero en la práctica dicha tarea resulta en un problema que en el peor caso es *NP-difícil* (15), resultado de los cálculos para obtener las probabilidades conjuntas; e incluso con pocas variables, puede volverse un problema intratable (24).

Dado lo anterior, existen diversas formas de enfrentarse al problema de realizar consultas a la red, entre las que se destacan:

- Condicionamiento
- Eliminación de variables
- Simulación estocástica

El condicionamiento se basa en la idea de simplificar la consulta, a través de conocer los valores de ciertas variables en ella. La eliminación de variables se enfoca en calcular las distribuciones conjuntas de adentro hacia afuera, usando propiedades de los factores para evitar cálculos innecesarios. Por último, la simulación estocástica consiste en simular varias veces la red bayesiana, con el fin de obtener en cada simulación una muestra de los valores que las variables no observadas pueden tomar; tras repetir la simulación un cierto número de veces, se tendrá un estimado de la probabilidad posterior.

Bibliografía

[1] Aghdam, R., Rezaei Tabar, V., and Pezeshk, H. (2018). Some node ordering methods for the K2 algorithm. *Computational Intelligence*, 35(1):42–58. 25, 27, 66

[2] Airoldi, E. M. (2007). Getting started in probabilistic graphical models. *PLoS Computational Biology 3(12)*. 6, 8

[3] Azmat, G., Guel, M., and Manning, A. (2004). Gender gaps in unemployment rates. *Centre for Economic Performance. London School of Economics and Political Science*. 4

[4] Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49. 79

[5] Baussola, M., Mussida, C., Jenkins, J., and Penfold, M. (2015). Determinantes de la brecha de género de desempleo en italia y reino unido. un estudio comparado. *Revista Internacional del Trabajo*, 134(4):581–608. 4

[6] Belloc, M. and Tilli, R. (2013). Unemployment by gender and gender catching-up: Empirical evidence from the italian regions. *Papers in Regional Science*, 92(3):481–494. 4

[7] Cooper, G. F. and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9(4):309–347. 22, 24

[INEGI] INEGI. Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad. `www.inegi.org.mx/programas/enoe/15ymas/`. [Consultada en julio del 2021]. 33, 38

[9] INEGI (2005). *50 Preguntas y respuestas / Instituto Nacional de Estadística y Geografía*. INEGI. 37

[10] INEGI (2013). *Conociendo la base de datos de la ENOE. Datos ajustados a proyecciones de población 2010 / Instituto Nacional de Estadística y Geografía*. INEGI. 35

[11] INEGI (2019). *Encuesta Nacional de Ocupación y Empleo. Cómo se hace la ENOE : métodos y procedimientos. / Instituto Nacional de Estadística y Geografía. 2da. ed. México*. INEGI. 33, 35, 36

[12] INMUJERES (2007). El impacto de los estereotipos y los roles de género en méxico. 5

[13] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Inroduction to Statistical Learning with Applications in R*. Springer. 7

[14] Ji Z, X. Q. and G, M. (2015). A review of parameter learning methods in bayesian network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9227:3–12. 28, 29

[15] Koller, D. and Friedman, N. (2010). *Probabilistic Graphical Models*. The MIT Press. 21, 83

[16] Koutentakis, F. (2015). Gender unemployment dynamics: Evidence from ten advanced economies. *LABOUR*, 29(1):15–31. 4

[17] Landivar, L. C. (2012). The impact of the great recession on mothers' employment. *Economic Stress and the Family Contemporary Perspectives in Family Research*, 6:163–185. 4

[18] LFT (2021). *Ley Federal del Trabajo: artículos 22-23, 175-176, 178-180, reforma 2015*. Secretaría del Trabajo y Previsión Social, México. 5, 37

[19] Livanos, I., Yalkin, C., and Nuñez, I. (2009). Gender employment discrimination: Greece and the united kingdom. *International Journal of Manpower*, 30(8):815–834. 4

[20] Mundial, B. (2020). Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT). datos.bancomundial.org/indicador/SL.UEM.TOTL.FE.ZS? [Consultada en mayo del 2020]. III, 3

[21] OIT (2018). La brecha de género en el empleo: ¿qué frena el avance de la mujer? www.ilo.org/infostories/es-ES/Stories/Employment/barriers-women#global-gap. [Consultada en mayo del 2020]. 3

[22] Organization, I. L. (2019). Quick guide on interpreting the unemployment rate. *ILO Publications*, pages 6–7. 3

[23] Pearl, J. (1988). *PROBABILISTIC REASONING IN INTELLIGENT SYSTEMS*. MORGAN KAUFMANN PUBLISHERS, INC., San Francisco, CA. 12, 21

[24] Sucar, L. E. (2015). *Probabilistic Graphical Models*. Springer. 8, 14, 21, 83

[25] van Buuren, S. (2018). *Flexible Imputation with Missing Data, Second Edition*. CRC Press. 77

[26] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journalof Statistical Software*, 45(3):2–67. 81