Universidad Veracruzana
Centro de Investigación en Inteligencia Artificial

Simulación social basada en redes bayesianas: Aproximación a un enfoque lógico-estadístico

Presenta:

Lic. Jean Christian Díaz Preciado

Director de tesis:
Dr. Alejandro Guerra-Hernández

Co-Director:
Dr. Nicandro Cruz Ramírez

Documento para obtener el titulo de
Maestro en inteligencia artificial, Diciembre 2016

Agradecimientos

Quiero expresar mi agradecimiento a todas las personas que me apoyaron de diferentes formas para poder concluir esta etapa de mi vida:

- A Larissa, por ser mi motivación y enseñarme a ver las cosas como si fuera la primera vez.

- A Patty, por todo el apoyo que me ha brindado y su compañía en todo momento.

- A mi mamá, por darme las fuerzas con su cariño, escucharme y alentarme siempre a buscar nuevas metas.

- A Alex, por todos los momentos que hemos compartido.

- Al Dr. Alejandro Guerra, por darme la oportunidad de desarrollar este trabajo y su invaluable guía.

- Al Dr. Nicandro Cruz, por todo el apoyo que me brindó.

- A todo el personal académico y administrativo del CIIA, por hacer esta experiencia inolvidable.

- A Conacyt, por otorgarme la beca 633473, para poder realizar mis estudios.

- Al Dr Efrén Mezura, por permitirme participar en el proyecto Conacyt No 220522, que me permitió concluir este documento.

- A mis sinodales: Dra. María del Carmen Mezura Godoy y Dr. Guillermo de Jesús Hoyos Rivera por sus invaluables observaciones y apoyo que ayudaron a mejorar este documento

Dedicación

A Patty y Larissa

Contenido

Lista de tablas

Lista de figuras

Capítulo 1

Introducción

La Inteligencia Artificial (IA) es un área de la ciencia de la computación que estudia el diseño y creación de herramientas que presenten conductas inteligentes y sean capaces de encontrar soluciones por sí mismas. En un principio la IA solo se había involucrado con el modelado de la cognición individual, pero en la década de los 80s hubo un creciente interés en la IA distribuida. Con el desarrollo de internet muchos investigadores se interesaron en el concepto de agente, que es un programa capaz de recibir o recolectar información de otros programas y decidir las acciones a tomar según la experiencia adquirida por el software [13]. Así con la IA distribuida y la tecnología de agentes, se dieron modelos de interacción entre agentes autónomos, con los que podían simular la interacción de sociedades humanas.

Los objetos de estudio en las ciencias sociales son siempre entidades dinámicas (personas, hogares o empresas) que cambian con el tiempo y reaccionan a los cambios de su entorno. En muchas ocasiones es complicado analizar la interacción que existe entre las entidades y los factores que influyen en los comportamientos de éstas, por lo que se ha recurrido a generar modelos que las representen en forma abstracta, es decir, simplificado, de manera más compacta y menos compleja.

Un modelo puede tomar diferentes formas, dependiendo la metodología que se utilice en la abstracción de lo que se quiere observar. Los fenómenos sociales han sido analizados desde perspectivas cuantitativas y cualitativas. La información cuantitativa se obtiene mediante la

recopilación de información de las entidades, por medio de censos o encuestas, la cual es estructurada con la finalidad de obtener estadísticas que muestren un panorama general del caso de estudio; mientras que la información cualitativa se obtiene de teorías de expertos en el área generadas mediante la observación, que tienen como finalidad explicar los procesos sociales desde la perspectiva del comportamiento de las entidades, es decir, cómo reaccionan a los cambios de su entorno o a las situaciones a las que se enfrentan.

Es por ello que los modelos en las ciencias sociales han tomado forma de ecuaciones matemáticas o de declaraciones lógicas. Sin embargo, independientemente de la forma que tome el modelo, es necesario examinar su comportamiento en el tiempo. La simulación social ha ganado terreno en las ciencias sociales por el dinamismo que requiere en sus modelos y por que permite observar el desarrollo de los mismos en el tiempo.

1.1 Antecedentes

Los primeros programas de simulación aparecieron en la década de los 60s, con la llegada de las primeras computadoras a las universidades, y se enfocaban en la observación de eventos discretos o sistemas dinámicos. Las simulaciones de eventos discretos consisten en hacer pasar unidades, que pueden representar personas, hogares o empresas, a través de colas o procesos para medir el rendimiento, por ejemplo el tiempo de espera de clientes en una fila. En cuanto a los sistemas dinámicos se hace uso de un gran número de ecuaciones diferenciales para trazar la trayectoria de las variables en el tiempo, por ejemplo los estudios del Club de Roma para tratar de predecir la economía mundial [10].

En la Figura 1.1, se muestra en el área gris, los trabajos desarrollados con modelos basados solo en ecuaciones diferenciales, y en el área blanca los modelos basados en objetos y eventos. También se observa como la IA aparece como una metodología basada en objetos y eventos que no tiene raíces en los modelos basados en ecuaciones diferenciales. En la década de los 90s llegan los modelos basados en agentes como un área de la IA, a los cuales se les agregó de la figura original una procedencia metodológica con los autómatas celulares y la teoría de juegos,

ya que proponen una simulación de individuos autónomos que interactúan entre ellos, como un autómata celular, implementando estrátegias de la teoría de juegos, por medio de sentencias lógicas que determinan sus acciones, con la finalidad de observar patrones de comportamiento emergente.

Como podemos observar las simulaciones de autómatas celulares (sCA) tienen la misma procedencia que los modelos multiagente, la diferencia entre estos es que un autómata se basa en una colección de celdas con dos estados posibles, los pasos de tiempo son discretos y sus reglas de comportamiento se basan en los estados de las celdas vecinas, mientras que en los modelos basados en agentes se pueden simular diferentes tipos ambientes con diversas arquitecturas de agentes y los pasos de tiempo pueden ser continuos o discretos.

A raíz de la diferencia en la metodología utilizada para modelar los procesos sociales, las simulaciones se pueden clasificar en dos grandes enfoques: el enfoque estadístico y el enfoque lógico [7]. Sin embargo, los modelos basados en agentes no pueden ser clasificados solo en uno de estos enfoques, ya que la abstracción de los procesos se expresa en formalismos lógicos y puede incluir información cuantitativa que regule mecanismos en el sistema. A continuación se describen éstas metodologías para modelar los procesos a simular.

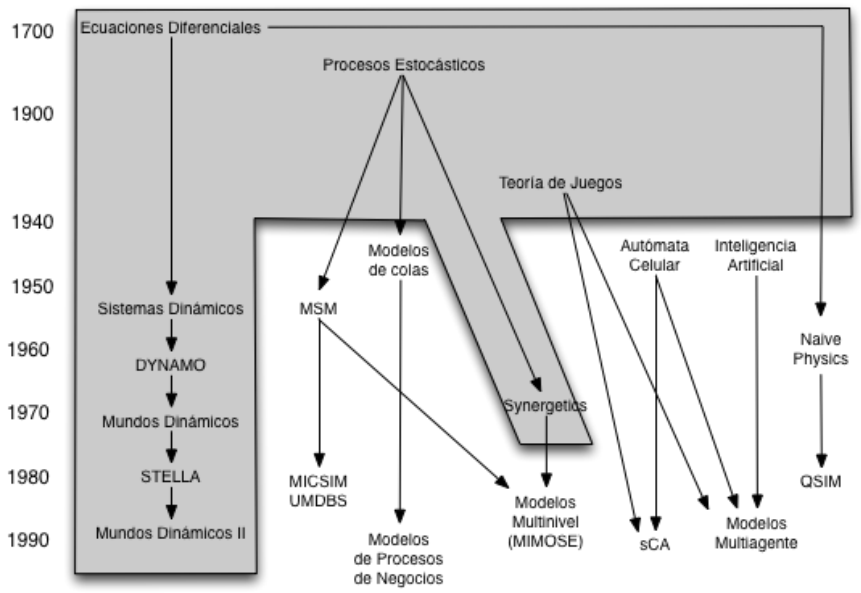


Figura 1.1: Desarrollo de enfoques de la simulación en las ciencias sociales (Adaptado de [7])

1.1.1 Enfoque estadístico

En la Figura 1.1, podemos identificar los modelos de simulación microanalítica (MSM), que se encuentran en el área blanca pero que tienen procedencia o implementan procesos estocásticos. De esta metodología de modelado nace la Microsimulación (MICSIM) en la década de los 80s e influye en los modelos de simulación multinivel (MIMOSE) a principios de la decáda de los 90s, en la que se simula la interacción de personas. Éstas metodologías por su naturaleza estocástica y la forma que toma la abstracción de los fenómeno podemos considerarlas bajo un enfoque estadístico.

En la Figura 1.2 podemos observar el procedimiento para desarrollar una simulación bajo un enfoque estadístico. Para elegir el proceso social a simular, es necesario contar con datos de éste. Los datos son recabados de las entidades que participan en el proceso, ya sea a través de censos o encuestas.

La abstracción del fenómeno se modela por medio de ecuaciones generadas de los datos, midiendo las variables que intervienen y la relación que existe entre ellas para representar los comportamientos que tienen en la realidad, y la simulación consiste en cambiar los parámetros de estimación de las entidades en el tiempo según las probabilidades de transición calculadas.

La salida de estos sistemas proporcionan datos predictivos de las distribuciones de las variables de las entidades, que contienen los cambios que pueden tener en el tiempo dado el modelo. La validación de los datos predictivos generados se hace a través de comparaciones estadísticas con datos reales recopilados en un periodo de tiempo igual al simulado. A pesar que trata cada entidad de forma individual no existe comunicación entre ellas y los cambios con respecto al tiempo se dan en respuesta de una tirada de dados representada por la generación de números aleatorios, tal como se hace en la Microsimulación [11].

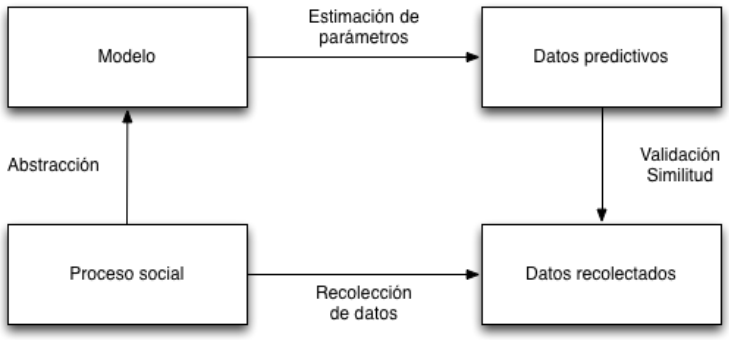


Figura 1.2: Modelado bajo un enfoque estadístico [7]

1.1.2 Enfoque lógico

En la Figura 1.3 podemos observar que en las simulaciones generadas bajo un enfoque lógico el proceso a observar es determinado por una teoría de comportamientos propuesta por un experto; la abstracción de la teoría se modela mediante formalismos lógicos, es decir, puede verse como un conjunto de proposiciones más un conjunto de reglas de inferencia que pueden usarse para deducir nuevas proposiciones, los modelos generados bajo este enfoque los podemos identificar en la Figura 1.1 en el área blanca y que no tienen características de procesos estócasticos ni de ecuaciones diferenciales, y que en algunos casos tienen descendencia de las estrátegias de la teoría de juego, como los autómatas celulares o los modelos basados en agentes.

La simulación consiste en dejar que las entidades desarrollen las acciones determinadas en el modelo. Los datos simulados que se obtienen, describen el comportamiento de las entidades al enfrentarse a ciertas situaciones denotadas por las reglas de inferencia, dichos comportamientos son comparados con la información del proceso real del que se generó la teoría. La información cualitativa puede ser recabada por entrevistas, cuestionarios, observación oa revisión de documentos del proceso social a estudiar. Este tipo de simulaciones han sido cuestionadas ya que la información de tipo cualitativa es considerada como subjetiva, sesgada, poco fiable y demasiado específica para cada contexto.

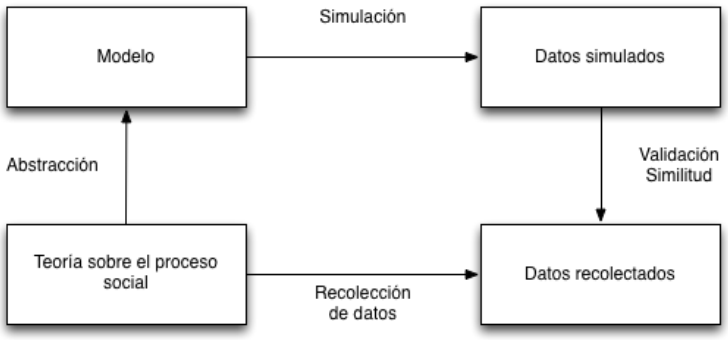


Figura 1.3: Modelado bajo un enfoque lógico [7]

1.1.3 Modelado basado en agentes

Las ventajas que ofrece este tipo de modelado las podemos observar partiendo de la definición de agente en la IA, en donde se consideran programas que tienen la capacidad de actuar de manera autónoma para satisfacer sus metas y objetivos, mientras se encuentran situados persistentemente en su ambiente [17, 14]. En la Figura 1.4, se muestra el proceso interno de un agente reactivo que percibe el estado actual de su ambiente con lo que determina que acción realizar según sus reglas de comportamiento, influyendo en el ambiente.

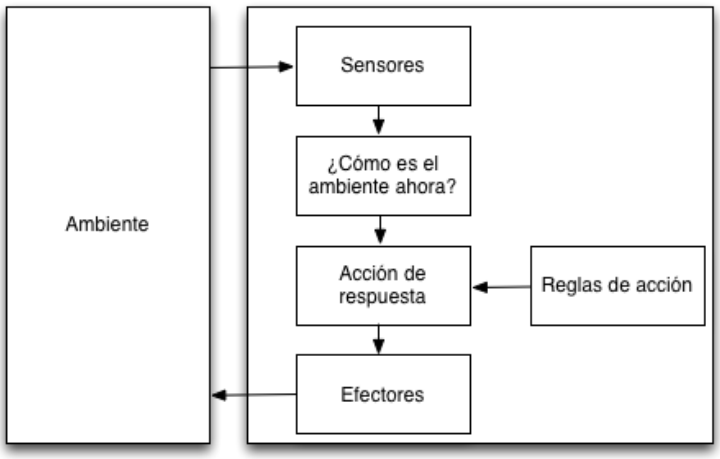


Figura 1.4: Abstracción de un agente reactivo

Esta arquitectura de agente es capaz de expandirse según el nivel de la capacidad del proceso lógico que se requiera:

- Agentes con estado. Tienen un estado interno en donde se lleva un registro de lo que ha percibido del ambiente
- Agentes lógicos. El modelo que rige el comportamiento de los agentes esta diseñado con fórmulas lógicas, así la toma de decisión de la acción a ejecutar es por medio de deducciones.
- Agentes basados en metas. La ejecución de las acciones depende de lo que cree el agente lo llevará a cumplir sus objetivos.
- Agentes basados en utilidad. La meta se define indirectamente, es decir, se determina una función de utilidad, y los agentes deciden que acciones realizar con la finalidad de maximizarla.

Con base en su comportamiento autónomo y flexible los agentes son reactivos, ya que responden a los cambios que perciben de su ambiente a través de acciones, su comportamiento puede estar determinado por sus metas y son capaces de socializar con otros agentes.

En un sistema multiagente se pueden tener diferentes tipos de agentes que interactúan en un mismo ambiente, dando oportunidad de simular los efectos que tienen sus comportamientos en otros agentes o en el mismo ambiente. Es por esto que en las ciencias sociales ha ganado mucho terreno como metodología para modelar procesos sociales, además de que ofrece diferentes arquitecturas de agentes que permiten diversificar el modelo de comportamiento ajustandose a lo que se desea analizar.

1.2 Validación de los modelos

La validación es la evaluación de la credibilidad de los modelos, es decir, que tan bien representa a la realidad. Considerando que las simulaciones sociales pueden implementar modelos generados con diferentes metodologías, no existe un criterio definitivo que garantice la validez de los resultados obtenidos. Por lo que validación depende del próposito del modelo a medir y la interpretación de los resultados debe hacerse según el contexto del fenómeno a estudiar [1].

En el resto de esta sección, se revisa el tema del propósito de validar las simulaciones considerando: el objetivo general de la simulación en la complejidad social, tres concepciones metodológicas básicas de tipos de validez y por último un conjunto de técnicas habituales aplicados en simulación social [7].

1.2.1 Objetivo general de la simulación

El objetivo de las simulaciones bajo un enfoque estadístico es predecir futuras distribuciones de las variables a partir de un modelo generado de un escenario pasado, por lo que la validación se hace por medio de comparaciones entre los datos obtenidos del sistema y los datos reales. La validación de los modelos bajo un enfoque multiagente evalua si el diseño de las acciones a nivel micro, generados a partir de teorías de expertos o de información cualitativa, pueden demostrar similitud con conductas sociales y las interacciones que se determinen entre ellos produzcan efectos a nivel macro que sean consistentes con datos cualitativo o teorías de comportamiento.

La similitud cualitativa a los datos reales se refiere a una comparación con el modelo en términos de los resultados, representados como características cualitativas, es decir, que interacciones y comportamientos sucedieron para llegar a los resultados. La validación de los modelos multiagente determina si son buenas o malas representaciones de la conducta social y la interacción.

1.2.2 Tipos generales de validación

Para evaluar si un modelo es capaz de reproducir las características esperadas de un fenómeno, existen tres metodologías generales: la predicción, retrodicción y semejanza estructural, que se describen a continuación.

Validación por predicción

El objetivo de la implementación de un modelo predictivo en una simulación social es suponer futuros estados del sistema, si las predicciones son satisfactorias en los periodos analizados, se

espera que dé buenos resultados al probarse en situaciones similares.

Validación por retrodicción

El objetivo es reproducir aspectos ya observados, dada la existencia de datos históricos, considerando que si el modelo es capaz de reproducir un registro histórico consistente y correctamente, puede ser de confianza para reproducir un registro futuro.

Validación por semejanza estructural

Considerando la implementación de acciones a nivel micro en la simulación, las clases de comportamiento en la escala macro se identifican en el modelo y se comparan con las clases de comportamiento identificados en el objetivo. Las validaciones evaluán el comportamientos de los agentes, las relaciones que existen entre ellos y la similitud de la evolución de la estructura social con el proceso social a estudiar.

1.2.3 Técnicas de validación

En esta sección se describen las técnicas de validación utilizadas en la simulación social. Por lo general se utiliza más de una técnica para validar un sistema ya que debe ser consistente desde diferentes perspectivas.

Validez aparente

Se implementa durante el desarrollo de la simulación y es utilizada para evaluar el modelo conceptual, sus componentes y su comportamiento. Esto se hace a través de la documentación, representando sus datos en gráficos o con animaciones de los modelos conforme se mueven en el tiempo.

Validez histórica

Es un tipo de retrodicción en donde los resultados del modelo se comparan con los resultados de los datos con los que se cuenta. Si solo se utiliza una parte de la información histórica disponible para diseñar el modelo, los datos generados a partir de este se utilizan para probar su capacidad predictiva.

Validez de eventos

Compara la ocurrencia de eventos en el modelo con la ocurrencia en los datos de origen. Esto se puede evaluar a nivel de trayectoria individual o de conjunto de agentes. Los eventos son situaciones a las que se enfrentan los agentes en el sistema.

Prueba de condiciones extremas

Se utiliza para verificar y validar los sistemas ingresando factores o combinaciones poco probables para probar que la simulación sigue teniendo sentido.

Análisis de sensibilidad

El análisis de sensibilidad es sinónimo de pruebas en las que los parámetros o incluso las interrelaciones de los componentes del modelo son variados sistemáticamente con el fin de determinar el efecto sobre el comportamiento del modelo.

Validez de sección transversal

Se refiere a la evaluación de la similitud de los datos sociales a los resultados que producen los modelos de simulación en un punto específico en el tiempo. Esto puede lograrse mediante la comparación de los datos, como por ejemplo una encuesta de la sección transversal, con la salida generada por un modelo en un solo periodo de tiempo.

Comparación con otros modelos

La comparación entre modelos es muy utilizada por la poca disponiblidad de datos con los que normalmente se cuenta o por que la abstraccón de los procesos tienen poca compatibilidad con los mismos. La validación se da en la comparación de resultados de modelos distintos que simulan el mismo proceso, con lo que se asume que los mecanismos de comportamiento de alguno no abstrae correctamente al fenómeno.

Validez de elementos cruzados

En esta validación se da cuando los modelos tienen una arquitectura de agentes similar pero mecanismos diferentes por lo que no se comparan los resultados del sistema no los modelos en general. Por ejemplo uno puede estudiar los efectos de la utilización de un modelo con agentes en un juego de negociación que funcionen con estrategias de aprendizaje o aprendizaje por refuerzo evolutivos, y evaluar que una de las estrategias produce resultados compatibles con el análisis teórico de la teoría de juegos.

Enfoques participativos

Se refieren a la participación de expertos sociales en la creación y validación del modelo, son muy utilizados en el desarrollo de estrátegias colaborativas e implementación de políticas.

1.3 Planteamiento del problema

Las simulaciones modeladas bajo enfoques estadísticos son mas fáciles de validar dado que se cuenta con información cuantitativa del proceso, pero sus fines son predictivos y los resultados no muestran una descripción del comportamiento del sistema. Mientras que las simulaciones basada en descripciones teóricas, que nos permiten observar el impacto de los supuestos comportamientos de las entidades a estudiar, son susceptibles a validaciones empíricas o basadas en enfoques participativos, que es muy cuestionado por la subjetividad que conlleva.

1.4 Estado del arte

El estado del arte que se muestra en esta sección esta orientado en dos partes: la primera enfocada al desarrollo de la investigación concerniente a la simulación social, especificamente en el trabajo de Silverman [15]; la segunda enfocada al análisis estadístico que se implementa en México del caso de estudio de la informalidad laboral, el cual es detallado en el siguiente capítulo.

1.4.1 Simulaciones sociales

En la actualidad se ha intentado desarrollar sistemas que consideren información estadística en modelos basados en agentes, con lo que surge una cuestión metodológica para decidir el grado en que los modelos deben ser diseñados bajo alguno de los enfoques previamente introducidos, el lógico y el estadístico. Existen trabajos que de alguna forma implementan ambos enfoques como los modelos de demografía computacional basado en agentes [3], en donde se han simulado fenómenos sociales como la migración [16], la movilidad residencial [2], entre otros. Silverman[15], propone una metodología de simulación social en donde el modelo combina el enfoque estadístico y el enfoque lógico. Como podemos observar en la Figura 1.5 se definen diferentes tipos de individuos determinados por las propiedades de la población a simular. Los modelos estadísticos se generan de datos obtenidos mediante censos o encuestas, con el objetivo de generar datos predictivos de las propiedades futuras de las personas. Los modelos basados en agentes de las personas simuladas están diseñados con reglas que determinan su comportamiento, con la finalidad de explicar los patrones observados en el sistema. Considerando que los modelos basados en agentes nos permiten probar diferentes comportamientos, se dá oportunidad de contestar a la pregunta ¿Qué pasaría si..?, y utilizar el modelo estadístico permite observar escenarios de cambios demográficos en periodos de tiempo más amplio que los que por lo general desarrollan los demógrafos. Por lo que la síntesis de los escenarios permite observar patrones de las distribuciones estadísticas dados los comportamientos complejos implementados en poblaciones artificiales que contienen información del mundo real.

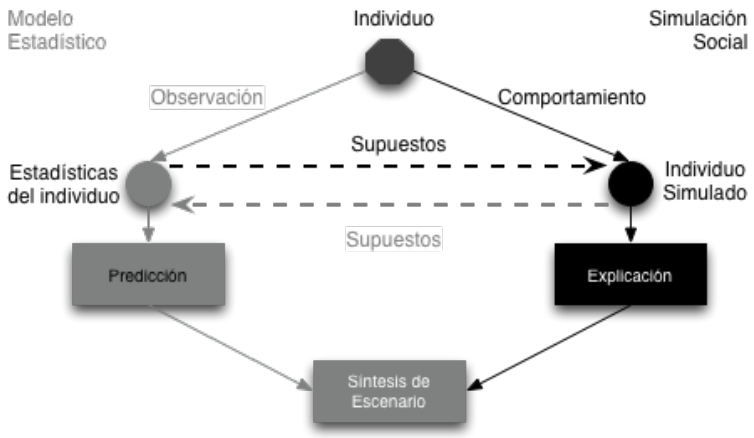


Figura 1.5: Propuesta de modelado combinando enfoques lógico y estadístico

1.4.2 Análisis de la informalidad laboral en México

En México el Instituto Nacional de Estadística y Geografía (INEGI), se encarga de recabar información referente a la situación de las personas, hogares y empresas del país, a través encuestas y censos. El objetivo del INEGI, es brindar información relevante y actualizada que sirva como herramienta de apoyo para la generación de políticas públicas que regulen las situaciones desfavorables que se observen en los diferentes análisis estadísticos que se realizan con dicha información.

En el año 2013 el INEGI, publica en su sitio web el documento titulado "México: Nuevas estadísticas de informalidad laboral", en donde presentan una matriz de Hussmanns, la cual se muestra en la Figura 1.6, la cual contiene tres clasificaciones generales de la población: tipo de unidad económica, la posición en la ocupación de las personas y su condición de formalidad o informalidad laboral. La tabla consiste en presentar un resumen del conteo de personas que inciden según sus características laborales; el conteo es a nivel nacional y la unidad de medida es millones de personas. Podemos observar que el mayor número de personas que laboran en condiciones de formalidad se concentran en la unidad económica correspondiente a empresas, gobierno e instituciones con un total 19.593, por otro lado la unidad económica empleadora denominada como sector informal concentra 14.177.

Debido a la dificultad que presenta la observación y el análisis del sector informal, y considerando los datos de esta tabla, se puede determinar que el área de oportunidad para im-

Clasificación según el tipo de la unidad económica empleadora	Clasificación según la posición en la ocupación y condición de informalidad										Totales por perspectiva de la unidad económica y/o laboral	
	Trabajadores subordinados remunerados				Empleadores		Trabajadores por cuenta propia		Trabajadores no remunerados			
	Asalariados		Con percepciones no salariales									
	Informal	Formal	Informal	Formal	Informal	Formal	Informal	Formal	Informal	Formal	Informal	Formal
Sector Informal	3.844		0.792		0.904		7.444		1.192		14.177	
Trabajo doméstico remunerado	2.128	0.059	0.020	0.000							2.148	0.059
Empresas, Gobierno e Instituciones	5.373	17.122	0.913	0.211		0.988		1.272	0.598		6.884	19.593
Ámbito agropecuario	2.169	0.289	0.219	0.017		0.313	2.553		1.129		6.070	0.618
Subtotal	13.514	17.470	1.944	0.227	0.904	1.301	9.997	1.272	2.920		29.279	20.270
Total	30.984		2.172		2.205		11.269		2.920		49.549	

Unidad de medida: millones de personas *Nota:* La suma de los componentes puede no coincidir con los totales debido al redondeo.

Figura 1.6: Matriz Hussmanns de Personas, segundo trimestre 2013

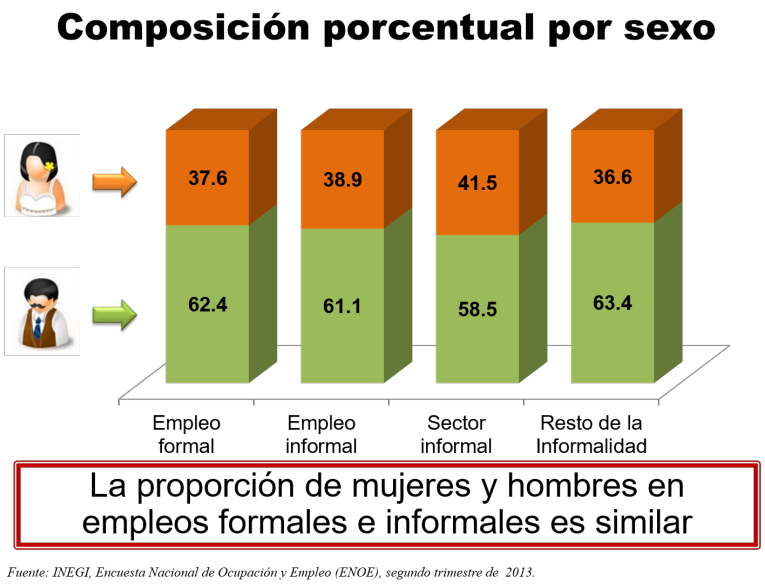


Figura 1.7: Composición porcentual por sexo

plementar una política pública que disminuya el número de personas que laboren en el sector informal se encuentra en los 6.884 millones de personas que laboran en empresas, gobierno e instituciones y lo hacen bajo un contexto informal, así como los 6.070 millones de personas que laboran en el ámbito agropecuario de igual manera bajo un contexto informal.

Como complemento al análisis realizado con la matriz de Hussmanns, el INEGI presenta de forma gráfica las condiciones laborales de las personas. En la Figura 1.7, presentan un análisis de la incidencia de las personas en la formalidad o informalidad según su sexo, en donde determinan que la proporción es similar.

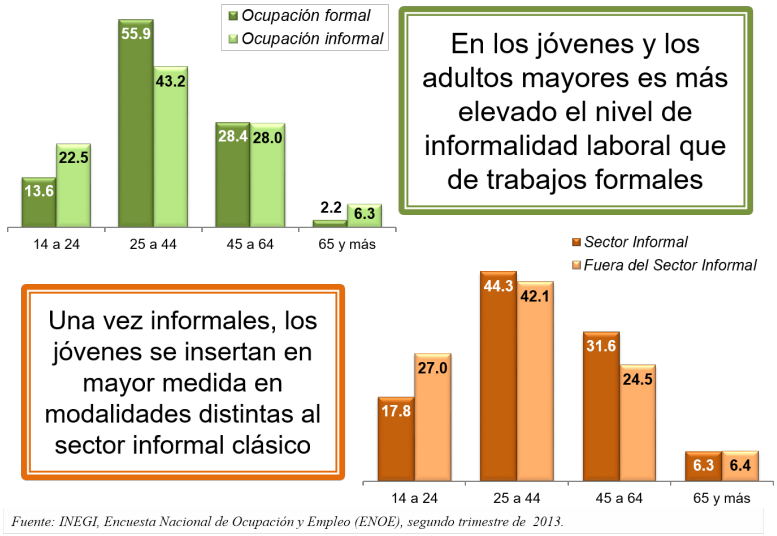


Figura 1.8: Grupos de edad

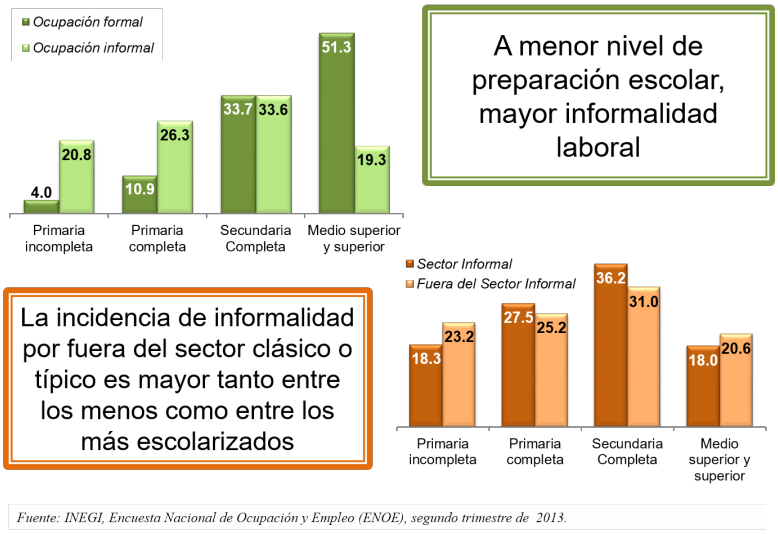


Figura 1.9: Nivel de instrucción

En la Figura 1.8, se hace el análisis según la edad de las personas, en donde determinan que existe más incidencia en la informalidad en adultos mayores y jóvenes; y por último la Figura 1.9 muestra la gráfica según el nivel de instrucción o escolaridad en donde se determina que a menor nivel de preparación escolar existe mayor incidencia en la informalidad laboral.

1.5 Propuesta

Desarrollar una simulación social basada en agentes que contemple modelos estadísticos bajo la forma de redes bayesianas. En la Figura 1.10, se muestra la inclusión del modelo de red

bayesiana en la metodología propuesta en [15]:

1. Se determina la base de datos que contiene la información del proceso social a simular.

2. Los agentes, que representan a las personas, obtienen sus propiedades directamente de la base de datos.

3. El modelo de red bayesiana se genera a partir de la base de datos, considerando las variables concernientes al proceso social y variables de las características de las personas como su sexo, edad, escolaridad, etc.

4. Los agentes hacen inferencias al modelo enviando como evidencia sus propiedades que definen sus características o contexto individual, obteniendo las probabilidades que tienen de obtener las propiedades del fenómeno.

5. Los cambios de estado de los agentes dependen de una tirada de dados considerando las probabilidades de transición obtenidas de las estadísticas demográficas representadas por índices o tasas.

6. Los agentes reportan todas sus propiedas en cada paso de tiempo, generando datos artificiales que pueden ser comparados estadísticamente con datos de encuestas de periodos posteriores.

7. En este trabajo no se incluyeron modelos de comportamiento, pero el objetivo es que las decisiones que tomen los agentes consideren su estado y las propiedades adquiridas del modelo estadístico.

8. Realizar un análisis estadístico basado en el modelo de red bayesiana generado para observar las oportunidades laborales de las personas según sus características de sexo, edad y escolaridad.

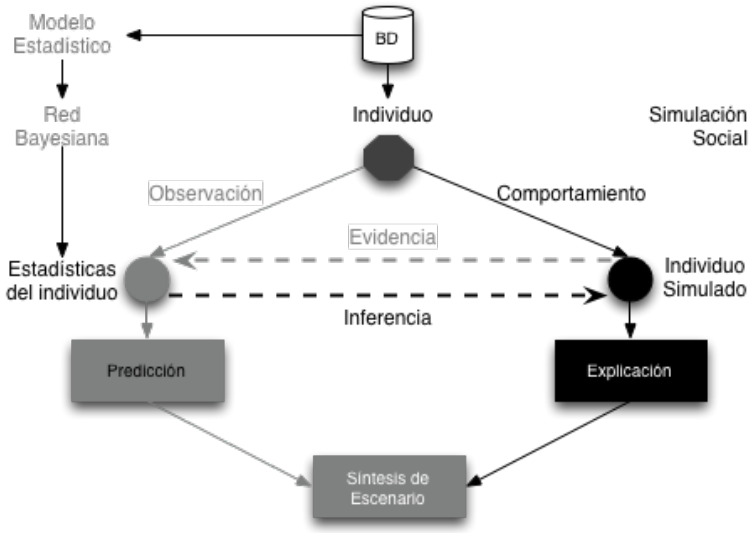


Figura 1.10: Implementación de redes bayesianas a la propuesta de Silverman

1.6 Hipótesis

Una simulación social en donde la abstracción del fenómeno tome una forma de red bayesiana para el manejo de la incertidumbre, percepción del ambiente o ambas, proporciona datos predictivos muy similares a la realidad.

1.7 Justificación

El desarrollo de una simulación social bajo un enfoque estadístico implementada en un sistema multiagente nos permite agregar modelos de comportamiento de las entidades a simular. Con esto podemos validar teorías de comportamiento en procesos sociales mediante la comparación estadística de los datos predictivos generados por el sistema con datos reales.

1.8 Objetivos

El objetivo de este trabajo es generar datos artificiales, mediante un sistema multiagente, que nos permitan predecir las características de las entidades a observar en periodos de tiempo

determinados. Validar los datos generados implementando comparaciones estadísticas con datos reales.

1.8.1 Objetivos específicos

- Obtener mediante procesos de minería de datos, un conjunto de casos representativos de las entidades a simular y determinar los atributos que describen mejor el fenómeno a observar. Considerando una base de dato generada por medio de censos y encuestas de los actores a simular de un periodo de tiempo determinado.
- Generar uno o más modelos de red bayesiana, que nos permitan entender la relación entre los atributos seleccionados.
- Crear un sistema multiagente, implementado en una plataforma Jason-CArtAgO [4], en donde se encapsulen en artefactos la base de datos y los modelos de red bayesiana generados de ésta, poniéndolos a disposición de los agentes.
- Generar datos artificiales que representen 4 años y medio, mediante el sistema multiagente.
- Determinar las pruebas estadísticas para validar los resultados obtenidos.
- Analizar el fenómeno a observar a partir de los modelos de red bayesiana generadas.

Capítulo 2

Metodología

Con la finalidad de probar la implementación de redes bayesianas como modelo probabilista en una simulación social, se desarrolló un sistema multiagente en una plataforma Jason-CArtAgO [4], en donde los modelos de red bayesiana y la base de datos son encapsulados en artefactos que son puestos a disposición de los agentes. Para construir el modelo se extendió JaCa-DDM [9], definiendo una nueva estrategia de aprendizaje basada en redes bayesianas y encapsulando un nuevo artefacto que nos permite realizar inferencias a los modelos utilizando herramientas de SamIam [5].

En la Figura 2.1, se plasma el diagrama general del sistema que tiene como entrada una base de datos y una red bayesiana que puede ser ingresada manualmente o generada por el sistema a partir de los datos; los agentes obtienen sus propiedades de la base de datos, y al paso del tiempo, el cambio de sus nuevas propiedades esta denotado por inferencias a la red bayesiana; el sistema proporciona los modelos de red generados y las bases de datos de cada periodo simulado.

2.1 Redes bayesianas

Se determinó que los modelos probabilistas tomasen la forma de una red bayesiana [12] que se define como un modelo gráfico que proporciona información a nivel cualitativo y cuantitativo.

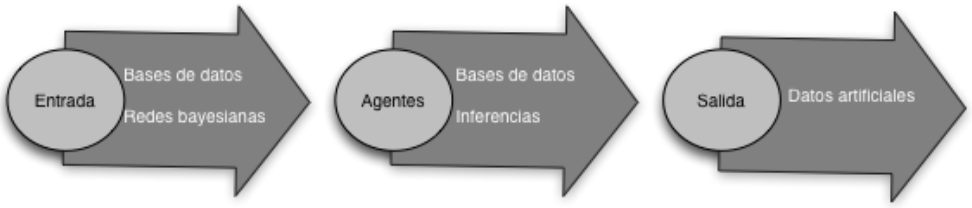


Figura 2.1: Diagrama general del sistema multiagente

La información a nivel cualitativo de una red bayesiana esta representada por un grafo acíclico dirigido compuesto por nodos, que representan las variables del caso de estudio, los cuales están unidos por flechas que determinan la causalidad ó influencia que existe entre ellas, cuando una flecha va de un nodo X a un nodo Y, se dice que X es un factor que influye en lo que pueda suceder en Y.

La información cuantitativa se observa en cada nodo, ya que tienen una distribución de probabilidad condicionada al efecto de sus nodos padre, es decir, la combinación de posibles valores. Estas distribuciones están contenidas en tablas en donde cada renglón contiene la probabilidad condicional de cada variable; esto nos permite conocer las probabilidades de los estados que pueden tener variables que no conocemos considerando que estamos observando o conocemos otras.

El sistema encapsula herramientas de Weka para la lectura de los datos y la generación de los modelos, y herramientas de SamIam que es una herramienta que permite realizar inferencias a los modelos de red bayesiana dadas ciertas variables conocidas o evidencia.

2.2 Protocolo ODD

Para probar la hipótesis se determinó como caso de estudio la informalidad laboral y subocupación en el Estado de Veracruz, la descripción de la simulación se muestra mediante el protocolo ODD [8] que estandariza la descripción de modelos basados en agentes. El protocolo ODD tiene como finalidad facilitar la revisión de los modelos con la finalidad de compararlos con otros. La descripción del modelo sigue la siguente estructura:

- **Propósito.** Se menciona el propósito, los objetivos y para que va a ser utilizado el sistema.
- **Entidades, estados y escalas.** Se describe el tipo de las entidades que conforman el sistema, las variables de estado que tienen y cómo se representan; la extensión espacial y temporal del modelo.
- **Visión general y planificación de procesos.** Se describe que hace cada entidad, en que momento y bajo que circunstancias.
- **Conceptos de diseño.** Se presentan para facilitar la interpretación de los resultados y tienen como fín mostrar las decisiones que se tomaron en el diseño del modelo.
- **Inicialización.** Es fundamental en este tipo de modelos conocer las condiciones con las que incian.
- **Datos de entrada.** Se indica si se considera una base de datos en el modelo y en que momentos se utilizan.
- **Submodelos.** Se describe a detalle los parámetros de los modelos que subyacen del sistema, y los momentos en los que se implementan.

Con la finalidad de poner en contexto al lector en cuanto al caso de estudio, la siguiente sección presenta una breve descripción de éste, y posteriormente se detalla como se incluyó en el sistema propuesto mediante el protocolo ODD.

2.3 Caso de estudio

El Instituto Nacional de Estadística y Geografía (INEGI), es un organismo autónomo del gobierno mexicano, encargado de captar, procesar y difundir información sobre el territorio, la economía y la población. El trabajo de recopilación se realiza a través de censos y encuestas cada determinado periodo de tiempo y con diferentes temáticas. El objetivo principal del INEGI es proporcionar a la sociedad y al Estado información estadística que muestre el

panorama nacional. Estas estadísticas sirven como parámetro para la toma de decisiones en la implementación de políticas públicas.

Para el desarrollo de esta simulación se trabajo con la Encuesta Nacional de Ocupación y Empleo (ENOE), que se realiza con una periodicidad trimestral recabando información de alrededor de 800,000 personas al año. La información estadística de la ENOE concerniente a la situación laboral de la población, es presentada por indicadores estrátegicos, siendo la tasa de desempleo (TD) el principal.

La tasa de desempleo es un indicador muy pólemico debido a los bajos niveles que reporta considerando que México es un país con mucha pobreza. Para tener un panorama más real de la situación en México es necesario analizar el comportamiento de las personas respecto a sus expectativas y cómo evalúan sus oportunidades ante las diferentes opciones laborales que se les presentan. Para ello se calculan indicadores complementarios como la Tasa de Ocupación en el Sector Informal (TOSI) y la Tasa de Subocupación (TS):

- TOSI. Se refiere a las personas que laboran en unidades económicas que no llevan un registro o contabilidad de sus actividades en la forma en que puede ser auditable por la autoridad fiscal

- TS. Engloba a todas las personas que tienen un empleo pero continuán en busca de otro

Estos indicadores complementarios se calculan como porcentaje de las personas con empleo, la subocupación es un fenómeno laboral que se da en México debido a que las personas no están preparadas para el desempleo ya que no se cuenta con un ahorro ni un seguro para ello, por esto al momento de quedar desempleadas fijan sus expectativas laborales en un nivel, pero si encuentran un empleo menor a sus expectativas lo toman para sobrevivir y continúan en busca del empleo que desean.

Es por esto que la informalidad laboral presenta niveles tan elevados, ya que muchas personas deciden laborar en este sector, ya que les permite continuar con su búsqueda de empleo y en ocasiones tener más de un trabajo.

2.3.1 Propósito

El propósito del sistema es observar las tasas de subocupación y ocupación en el sector informal en el Estado de Veracruz, en un periodo de cuatro años y medio, considerando las condiciones laborales que tienen las personas (prestaciones, jornada laboral, ingreso y sector en el que se desarrolla) dadas sus propiedades individuales como sexo, edad, escolaridad y si se encuentran estudiando.

El sistema genera datos artificales los cuales son válidados por comparaciones estadísticas con datos reales de los periodos simulados publicados por el INEGI. A partir de los datos generados se analiza el impacto que tiene la edad y el género de las personas en sus oportunidades laborales.

2.3.2 Entidades, estados y escalas

Las entidades en el sistema toman la forma de agentes que representan a personas. En la Figura 2.2, se muestra la clasificación que hace el INEGI de la población según su condición y ocupación. La población mayor de 15 años, edad mínima legal para laborar en México, puede ser económicamente activa (PEA) ó no económicamente activa (PNEA), dependiendo si ejercen presión en el mercado laboral o no.

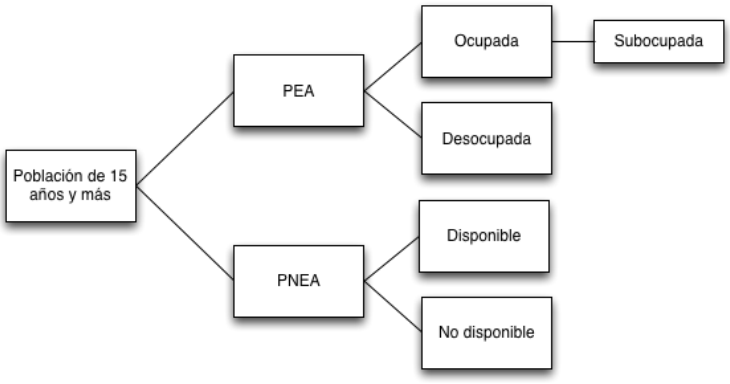


Figura 2.2: Clasificación de la población según su ocupación

Los estados de los agentes representan las condiciones de las personas según la clasificación del INEGI. En la tabla 2.1 se muestran los estados que pueden tener los agentes.

Estado	Descripción
Menor	Persona menor de 15 años
Ocupado	Persona que tiene empleo
Desocupado	Persona que no tiene empleo y está en busca de uno
Disponible	Persona que no tienen empleo y no busca uno
No disponible	Persona que se encuentra bajo un contexto que le impide laborar

Tabla 2.1: Estados de los agentes en el sistema

Los agentes tienen propiedades que determinan su contexto individual como sexo, edad, escolaridad y si se encuentan estudiando. Los agente ocupados conocen las propiedades de su empleo: jornada, ingreso, rama, posición en la ocupación, tipo de contrato, y si tienen acceso a seguridad social. La subocupación e informalidad laboral se representan a través de propiedades con valores binarios que solo tienen los agentes en estado de ocupados.

La escala temporal se determinó según la periodicidad de aplicación de las encuestas, por lo que cada periodo de tiempo en el sistema representa un trimestre. El primer periodo corresponde al primer trimestre del año 2010, y la simulación concluye en el periodo 19 que representa el tercer trimestre del año 2014.

2.3.3 Visión general y planificación de procesos

Los cambios de estado que pueden tener los agentes mayores de 15 años se muestran en la tabla 2.2. Dado que cada paso de tiempo representa un trimestre, todos los agentes aumentan su edad cada cuatro periodos. Con el cambio de edad algunos agentes menores cumplen 15 años con lo que su estado cambia.

t_n	t_{n+1}
Ocupado	Desocupado
Desocupado	Ocupado
Disponible	Ocupado
Disponible	Desocupado

t_n	t_{n+1}
Menor	Ocupado
	Desocupado
	Disponible
	No disponible

Tabla 2.2: Cambios de estado en los agentes

Dado que en esta simulación no se ingresaron datos de las empresas que nos permitieran modelar las condiciones bajo las que se desemplea a las personas se decidió que en cada período todos

los agentes desocupados encuentran empleo, y se desemplea a agentes ocupados según la tasa de desempleo publicada por el INEGI en el periodo simulado, como se muestra en la Tabla 2.3.

	t_n	t_{n+1}
Tasa de desempleo (TD)	3.7%	4.7%
Ocupados	96	95
Desocupados	4	5

Tabla 2.3: Probabilidad de cambio de estado según tasa de desocupación

Cuando un agente cambia de condición a ocupado adquiere las propiedades de su empleo. En la tabla 2.4, se ejemplifica las propiedades que puede adquirir un agente al cambiar de estado desocupado a ocupado.

	Propiedades	t_n	t_{n+1}
Propiedades generales	Sexo	Mujer	Mujer
	Edad	24	24
	Escolaridad	Bachillerato	Bachillerato
	¿Estudia?	No	No
Condición		**Desocupado**	**Ocupado**
Propiedades laborales	Jornada		15 a 34 horas
	Ingreso		$3,000
	Rama		Terciario
	Posición		Empleado
	Subocupación		Sí
	Seguridad social		No
	Clasificación empleo		Informal

Tabla 2.4: Cambio de condición: desocupados a ocupados

2.3.4 Concepto de diseño

Principios básicos

La simulación se basa en modelos estadísticos, de los que se rigen los comportamientos de los agentes. Se implementaron dos modelos generados de los datos de la ENOE como se muestra a continuación:

- **Red bayesiana de condiciones laborales.** Modela las oportunidades laborales que

tienen las personas, por lo que se construye a partir de los datos de la población ocupada considerando sus propiedades generales y laborales.

- **Red bayesiana de ocupación.** Modela la ocupación de las personas que cumplen 15 años en el periodo simulado, esta red se genera a partir de los datos correspondiente a las personas de 15 a 24 años de edad, considerando su escolaridad, sexo y condición.

Se utiliza la tasa de desocupación calculada de los datos como la probabilidad de transición que determina el cambio de estado de los agentes ocupados a desocupados, es decir, determina el número de agentes desempleados por período.

Los agentes adquieren sus propiedades de la base de datos y cada vez que cambian de estado a ocupados adquieren las propiedades de su empleo a través de inferencias a la red bayesiana laboral, enviando como evidencia sus propiedades generales. Los agentes menores que cumplen 15 años, conocen su nuevo estado por medio de inferencias a la red bayesiana de condición, enviando como evidencia su sexo, escolaridad y si se encuentran estudiando. Se crean datos artificiales que nos permiten generar estadísticas y calcular las tasas complementarias, con la finalidad de compararlas con datos reales.

Emergencia

Los comportamientos de los agentes están determinados por los modelos estadísticos, por lo que con las bases de datos generadas en el sistema podemos observar las oportunidades laborales de las personas según sus características y calcular las tasas de subocupación e informalidad laboral en cada periodo. También podemos observar qué condición adquieren las personas que cumplen 15 años, lo que nos permite analizar las oportunidades que pueden tener laboralmente dadas las estadísticas. Por la forma que toma el modelo estadístico se puede analizar a detalle las oportunidades laborales que van teniendo los agentes en el tiempo según sus propiedades, es decir, podemos observar los patrones a nivel sexo, edad, escolaridad, o desde la perspectiva de las propiedades laborales como que propiedades generales tienen mayor incidencia en ciertas jornadas laborales, rangos de salario, en empleos informales, etc.

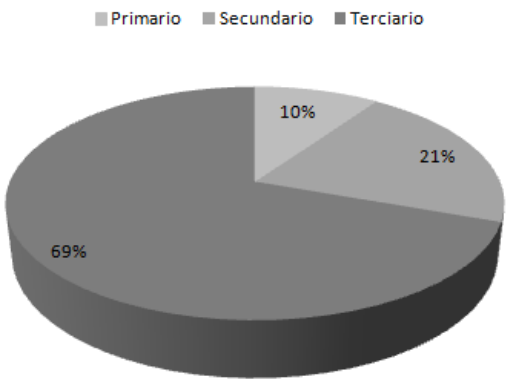


Figura 2.3: Selección por ruleta del sector en que se desarrolla el empleo

Predicción

El sistema predice la tasa de subocupación y la tasa de ocupación en el sector informal. La predicción del sistema se puede observar desde el conjunto de datos generados por periodo, ya que las tasas se calculan a partir de toda la población; no existe predicción a nivel agente ya que no tienen conocimiento de sus futuros cambios de condición.

Percepción

La percepción de las propiedades de los empleos se da a través de inferencias a la red bayesiana de condiciones laborales, que está encapsulada en un artefacto que la pone a disposición de los agentes.

Estocasticidad

Los agentes, al hacer una inferencia a un modelo de red bayesiana, obtienen la tabla de probabilidad condicional de cada propiedad que desconocen; la selección de la propiedad que adquiere se hace implementando, como acción interna, una ruleta propuesta por De Jong [6], en donde se sitúan las variables con mayor probabilidad tienen las porciones más grandes; la selección de la variable se determina generando un número aleatorio entre [0..1] y devolver la variable que se encuentre en esa posición, en la figura 2.3 se muestra un ejemplo de la ruleta que se simula para determinar el sector en el que se desarrolla el empleo.

Observación

En cada período se genera un archivo de texto, en donde los agentes reportan todas sus propiedades conocidas. Las tasas de subocupación y ocupación en la informalidad se calculan como porcentaje de la población economicamente activa, es decir, de los agentes que están en estado de ocupado y desocupado. Tener bases de datos de cada periodo nos permite observar la trayectoria de los agentes de forma individual o por subconjuntos delimitados por las propiedades que se deseen observar, asi como los cambios que presentan en el tiempo.

2.3.5 Inicialización

Es necesario ingresar la base de datos para generar los modelos de red bayesiana y ponerlos a disposición de los agentes. Los agentes que representan a las personas, al ser creados, se les asigna un número que corresponde al número de caso de la base de datos, y obtienen sus propiedades directamente de ésta. La población total es de 1,000 agentes, que son una muestra representativa de la base de datos original del Estado de Veracruz que tiene 13,295 casos. Por lo que las propiedades con las que incian los agentes dependen de la base de datos preprocesada que se ingrese.

2.3.6 Datos de entrada

El sistema recibe la base de datos preprocesada previamente, los valores de la tasa de desocupación de todos los periodos a simular calculadas de las bases de datos reales, y de forma opcional los modelos de red bayesiana, los cuales pueden ser generados por el sistema en caso de no ingresarse. Los valores de las propiedades que contienen la base de datos los podemos observar en el Tabla 2.5.

Atributos generales

Atributo	Variable	Descripción
Sexo	1	Hombre
	2	Mujer
Edad	1	15 a 24 años
	2	25 a 44 años
	3	45 a 64 años
	4	65 años y más
Escolaridad	1	Preescolar
	2	Primaria
	3	Secundaria
	4	Bachillerato
	5	Normal
	6	Carrera técnica
	7	Profesional
	8	Maestría
	9	Doctorado
¿Estudia?	1	Sí
	2	No

Atributos laborales

Atributo	Variable	Descripción
Jornada	1	Ausentes temporales
	2	Menos de 15 horas
	3	De 15 a 34 horas
	4	De 35 a 48 horas
	5	Más de 48 horas
Ingreso	1	Hasta un salario mínimo
	2	Más de 1 hasta 2
	3	Más de 2 hasta 3
	4	Más de 3 hasta 5
	5	Májs de 5
	6	No recibe
Posición en la ocupación	1	Subordinados y remunerados
	2	Empleadores
	3	Cuenta propia
	4	Sin pago

Atributo	Variable	Descripción
Rama	1	Primario
	2	Secundario
	3	Terciario
Seguridad social	1	Con acceso
	2	Sin acceso
Subocupación	0	No
	1	Sí
Número de empleos	1	1 empleo
	2	2 empleos
Formalidad/ Informalidad	1	Informal
	2	Formal

Tabla 2.5: Valores de las propiedades obtenidas de la base de datos

El vector que contiene las tasas de desempleo de cada periodo se muestra en la Tabla 2.6.

2010/1	**2010/2**	**2010/3**	**2010/4**	**2011/1**	**2011/2**	**2011/3**	**2011/4**	**2012/1**	**2012/2**	**2012/3**	**2012/4**
3.7%	4.7%	4.7%	4.6%	5.1%	4.3%	6.1%	5.3%	3.9%	3.8%	5.9%	4.3%

2013/1	**2013/2**	**2013/3**	**2013/4**	**2014/1**	**2014/2**	**2014/3**
3.3%	4.3%	4.1%	5.1%	5.3%	4.7%	4.6%

Tabla 2.6: Tasas de desempleo calculadas de los datos reales

En la Figura 2.4, se muestra el modelo de red bayesiana de condiciones laborales y en la Figura 2.5, el modelo de red bayesiana de ocupación.

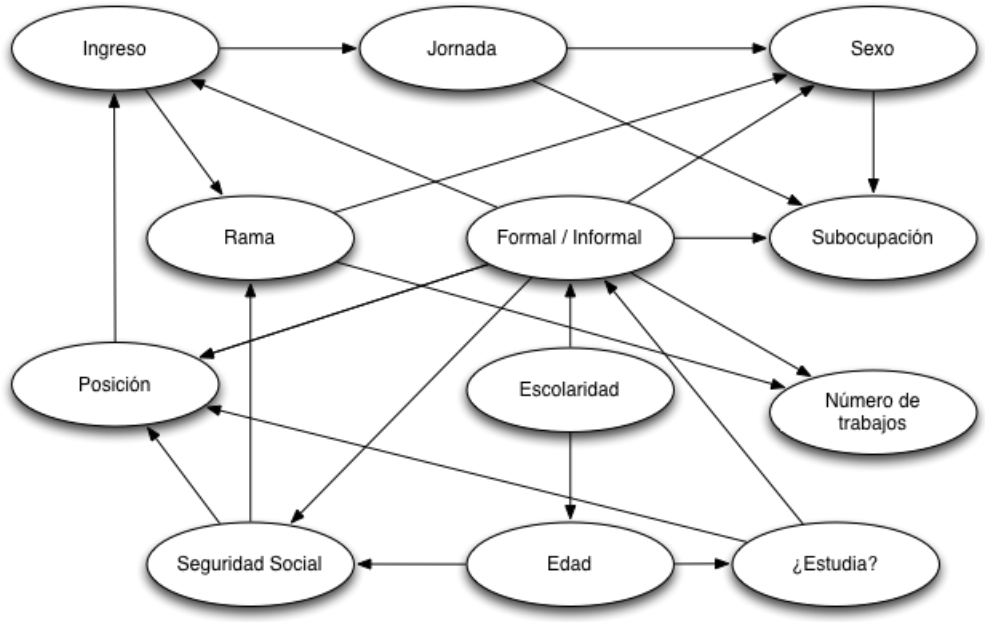


Figura 2.4: Red bayesiana de condiciones laborales

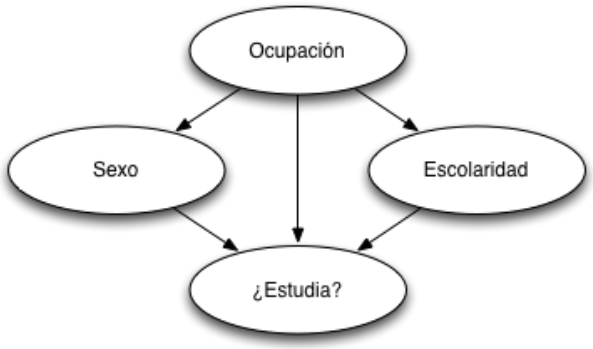


Figura 2.5: Red bayesiana de ocupación

2.3.7 Submodelos

En la figura 2.6, se muestra el proceso de un agente que representa a una persona, quien comienza en el tiempo cero (t_0) recibiendo sus propiedades directamente de la base de datos ENOE, cuando adquiere la condicion de ocupado en cualquier periodo (t_n), hace inferencia a la red bayesiana laboral, enviando como evidencia sus propiedades generales obteniendo las tablas de probabilidad condicional (TPC) de las propiedades laborales de su nuevo empleo, en cuanto conoce todas sus propiedades las reporta a la base de datos del periodo. Cuando un agente cumple años en un periodo, deja de ser menor y debe cambiar su condición. En la figura 2.7, se muestra la inferencia que realiza al modelo, enviando su sexo, escolaridad y si se encuentra estudiando, obteniendo la TPC de la condición. Si su condición cambia de menor a ocupado, en el mismo periodo realiza una inferencia al modelo de red bayesiana laboral.

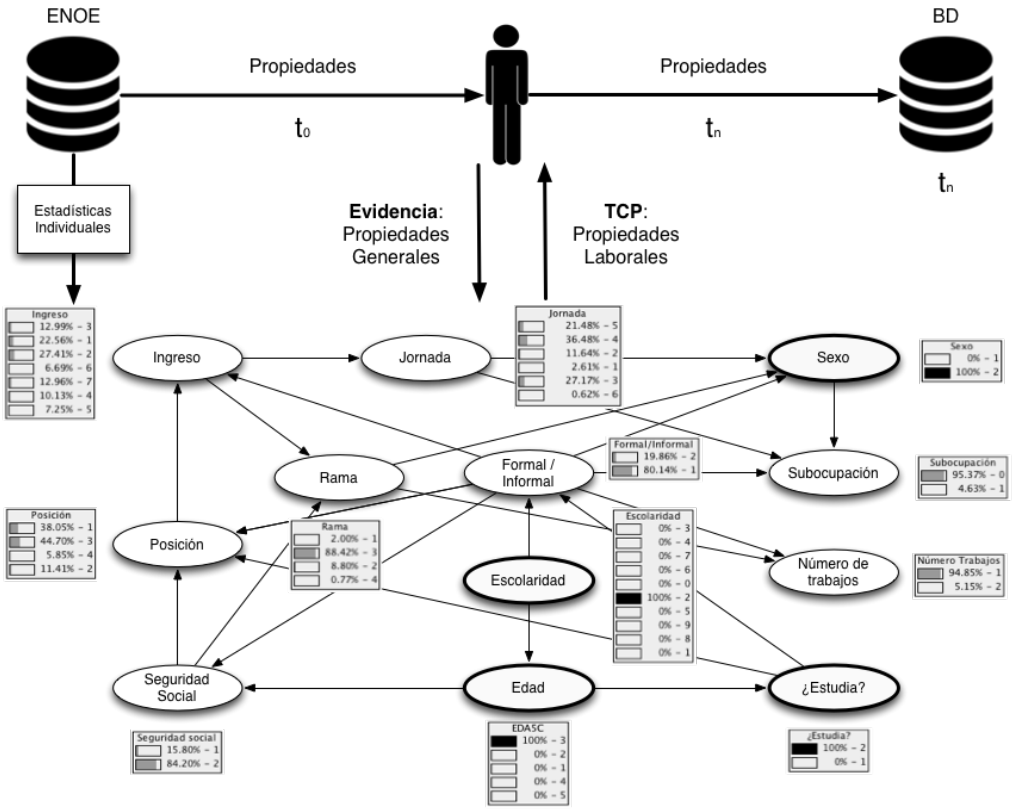


Figura 2.6: Proceso de agente persona con nuevo estado ocupado

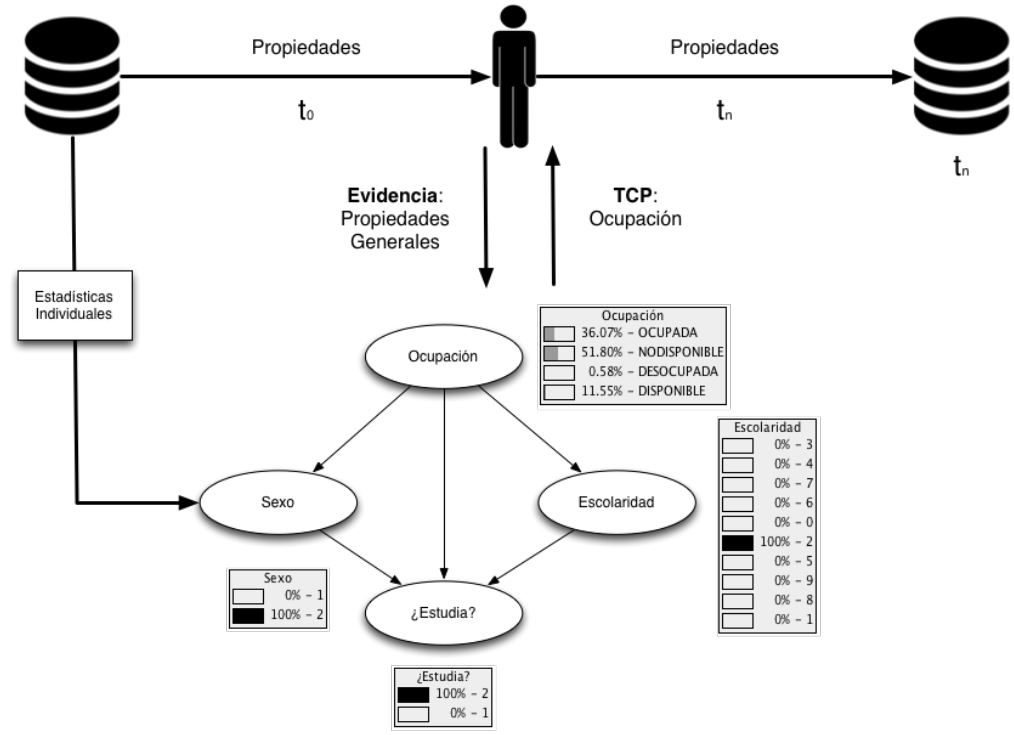


Figura 2.7: Proceso de agente persona que cumple 15 años

Capítulo 3

Resultados y discusión

Como se mencionó en el diagrama general del sistema (Figura 2.1), éste proporciona como salida los modelos de red bayesiana que se implentan en la simulación y las bases de datos de cada período simulado, por lo que en esta sección se presenta los resultados y el análisis que se realiza desde las siguientes perspectivas:

- Personas que ingresan formalmente, por mayoría de edad, al mercado laboral.
- Impacto del género de las personas en sus oportunidades laborales.

Los resultados obtenidos en este documento tienen la finalidad de mostrar el posible uso de la herramienta propuesta y no se consideran concluyentes en el caso de estudio ya que pueden existir otros factores que no han sido considerados.

3.1 Modelos generados

La Tabla 3.1, muestra los parámetros implementados para la generación de las redes bayesianas, los cuales se determinaron con base en la observación de las redes generadas en experimentos utilizando Weka. En estos experimentos se observó que los parámetros que modificaban significativamente al modelo fueron el iniciarlo con un modelo Naive y la implementación de la

manta de Markov, ya que sin ella, la variable que determina si el trabajador se encuentra estudiando quedaba fuera de la red, y para el caso de estudio esta variable es considerada relevante. Se llevó a cabo una validación cruzada del modelo con diez pliegues.

Estimador	Estimador Simple	A 0.5
Algoritmo de busqueda	HillClimber	
	initNaiveBayes	False
	MarkovBlanket	True
	mxNrOfParents	10,000
	scoreTYPE	MDL
	useArcReversal	True

Tabla 3.1: Parámetros para generar modelos

3.1.1 Clasificación de modelos

Para la determinación de los modelos que se implementan en el sistema se observó en los experimentos el porcentaje de clasificación y su desviación estándar. En la Tabla 3.2 se muestra que el modelo implementado de red bayesiana de condición laboral obtuvo un 94.11% de desempeño ó casos clasificados correctamente con respecto a la formalidad o informalidad de los empleos y una desviación estándar de 1.42, mientras que la red bayesiana de condición obtuvo un 65.73% con una desviación estándar de 0.74, considerando que, por la naturaleza de los datos, que se obtienen mediante encuestas los porcentajes de clasificación obtenidos son aceptables.

Red bayesiana ocupación	Desempeño	65.73%
	DesvEst	1.42
Red bayesiana condición laboral	Desempeño	94.11%
	DesvEst	0.74

Tabla 3.2: Porcentajes de clasificación y desviación estándar

3.2 Validación de datos artificiales

La validación de los resultados obtenidos en el sistema se llevo a cabo por retrodicción, dado que el objetivo es reproducir aspectos ya observados en la realidad y compararlos; si el modelo es capaz de reproducir datos similares se puede considerar para reproducir datos futuros. Se

implementó la técnica de validación de sección transversal, en donde se evalua la similitud de los datos sociales generados mediante encuestas con los datos artificiales genreados por el sistema. Para esto, se calcularon las tasas de desocupación (TD), la tasa de subocupación (TS) y la tasa de ocupación en el sector informal (TOSI) para compararlas estadísticamente con las tasas de los datos obtenidos del INEGI. En la Tabla 3.3, se muestran las tasas de los datos artificiales y los datos reales, en donde se resaltan los periodos en los que se presentaron las mayores diferencias.

Periodo	**TD**		**TS**		**TOSI**	
	Artificial	**Real**	**Artificial**	**Real**	**Artificial**	**Real**
1	0.036	0.036	0.043	0.043	0.583	0.577
2	0.038	0.035	0.043	0.048	0.581	0.580
3	0.047	0.043	0.048	0.049	0.567	0.567
4	0.047	0.041	0.047	0.044	0.565	0.574
5	0.051	0.046	**0.043**	**0.059**	0.575	0.582
6	0.042	0.042	0.049	0.046	0.573	0.581
7	0.051	0.054	0.055	0.050	0.563	0.588
8	0.043	0.043	0.054	0.055	0.575	0.593
9	0.039	0.037	0.059	0.058	**0.559**	**0.586**
10	0.036	0.037	**0.055**	**0.046**	0.565	0.586
11	0.049	0.046	**0.058**	**0.047**	0.569	0.590
12	**0.043**	**0.033**	**0.058**	**0.036**	0.594	0.575
13	0.033	0.032	**0.061**	**0.038**	0.588	0.579
14	0.042	0.040	**0.066**	**0.042**	0.578	0.594
15	0.039	0.042	0.063	0.059	0.582	0.595

Tabla 3.3: Comparación de tasas

Se calculó el error cuadrático medio (EMC) que podemos observar en la Tabla 3.4, en donde el valor más alto se obtuvo en la TOSI con un .0002432, lo que indica que se obtuvieron resultados muy similares a los datos reales.

	TD	**TS**	**TOSI**
EMC	0.000014866	0.000147335	0.000243216

Tabla 3.4: Error cuadrático medio

Se generaron las gráficas de las tasas presentadas en la Tabla 3.3 con la finalidad de observar las tendencias y las diferencias señaladas que existen entre ellas. En la tasa de desocupación la diferencia más grande se obtuvo en el periodo que representa el periodo 12 con una diferencia

de 1%, Figura 3.1.

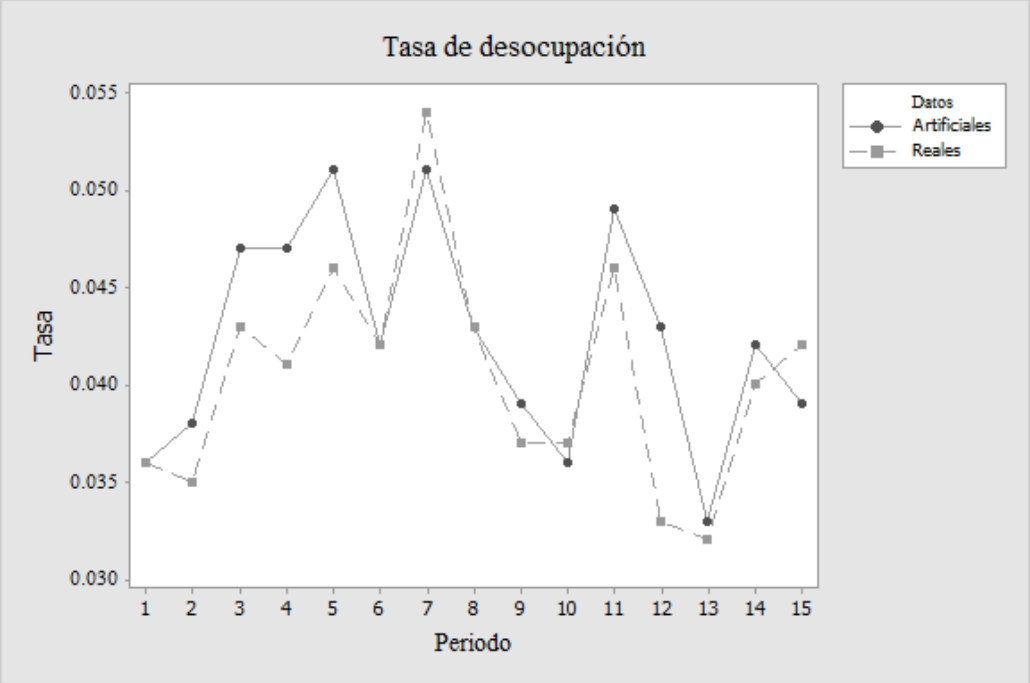


Figura 3.1: Comparación gráfica Tasas de Desocupación

La tasa de subocupación presenta sus mayores diferencias del periodo 10 al 14, sin embargo se puede observar que en estos periodos el INEGI reporta una caída de entre 1 y 2 %, sin tener cambios considerables en las otras tasas complementarias, Figura 3.2.

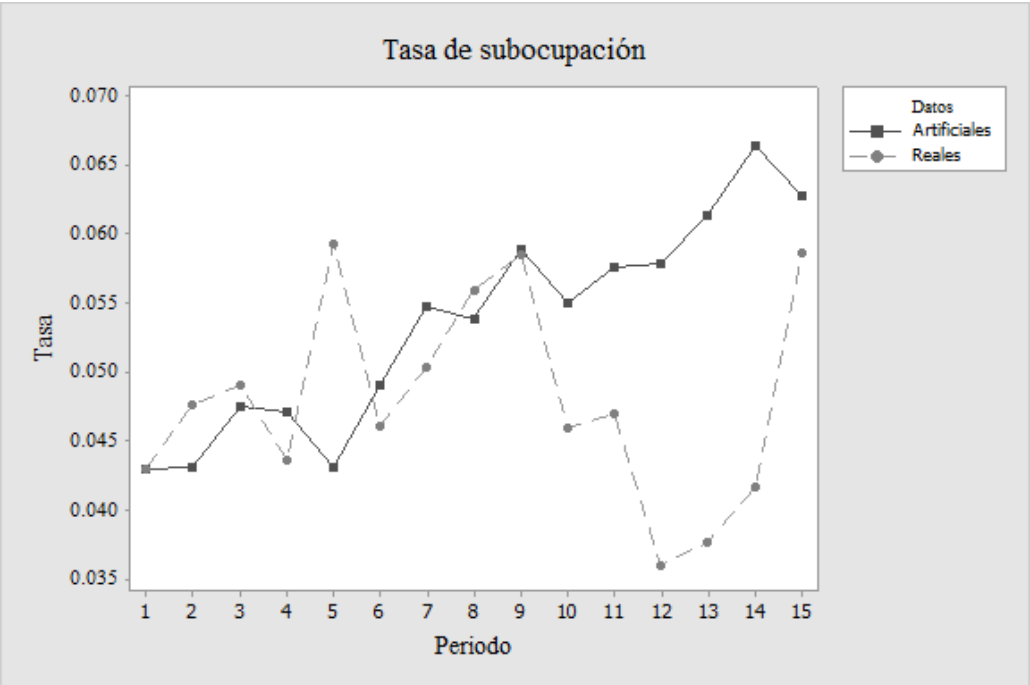


Figura 3.2: Comparación gráfica Tasas de Subocupación

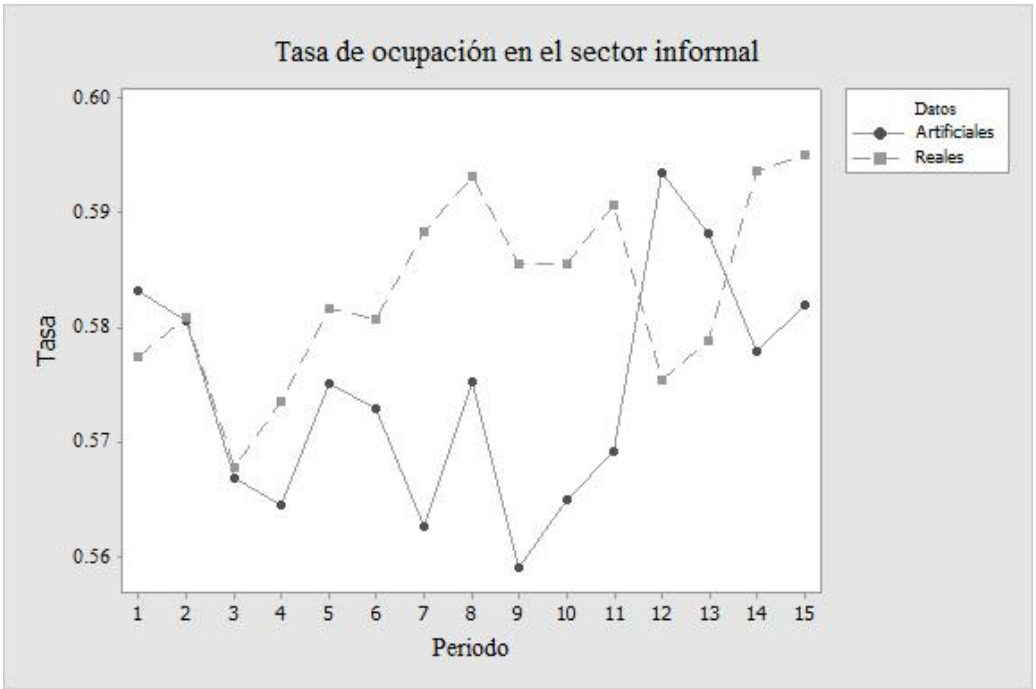


Figura 3.3: Comparación gráfica Tasas de Ocupación en Sector Informal

La tasa de ocupación en el sector informal presenta la mayor diferencia en el calculo de tasas en el periodo 9 con 2.6%, Figura 3.3

3.2.1 Prueba de normalidad

Se realizó la prueba de Komogorov-Smirnof para verificar la normalidad de los datos artificiales y reales para determinar el test estadístico a implementar en la verificación. En la Figura 3.4 observamos como todos los valores se ajustan a la normal, y en la Tabla 3.5 se presentan los valores del P-Value, en donde todos son mayores a 0.01 lo que se determina como evidencia de que los datos proceden de una distribución de tipo normal.

Tasa	Origen	P-Value
TD	Artificial	0.60
	Real	0.62
TS	Artificial	0.48
	Real	0.42
TOSI	Artificial	0.91
	Real	0.87

Tabla 3.5: P-value de prueba de normalidad

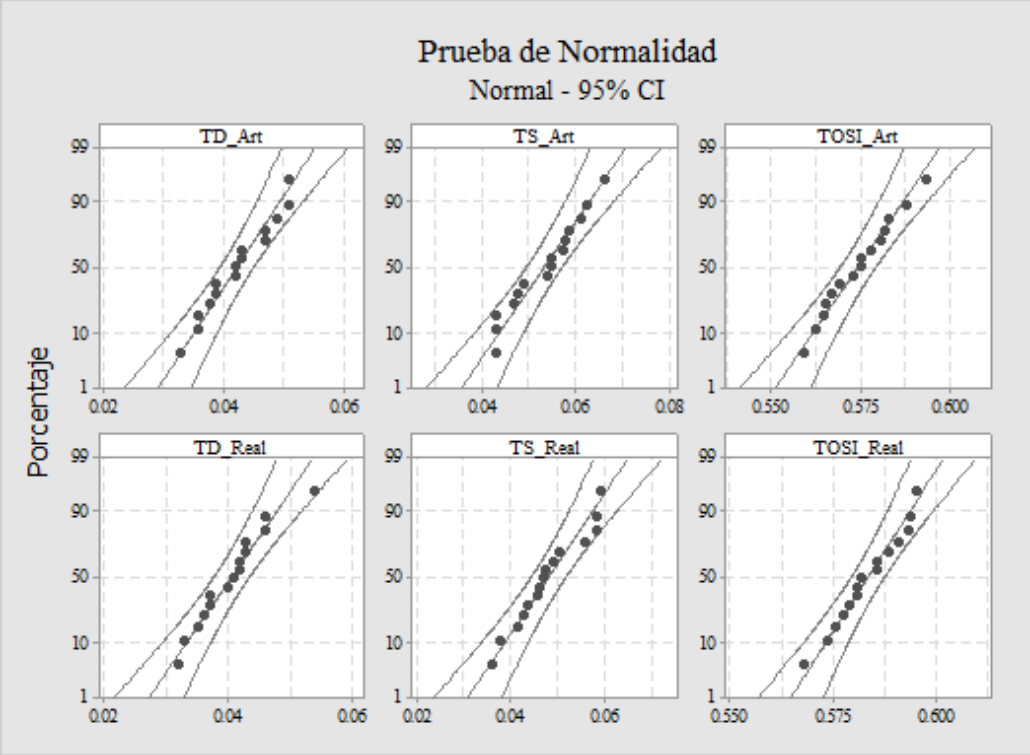


Figura 3.4: Prueba de normalidad de las tasas

3.2.2 Pruebas paramétricas de comparación

Dado que se comprobó que los datos provienen de una distribución normal se implementó la comparacion por T-Test, con un nivel de significación del 0.05, obteniendo los siguientes resultados:

- Las tasas de desocupación obtuvieron un p-value de .717666 que es mayor a 0.05, por lo que se rechaza la hipótesis nula, y se determina que las muestras no presentan diferencias significativas. En la Figura 3.5 se observa graficamente los valores calculados por cada muestra y la distancia que existe entre ellas.

- Las tasas de subocupación obtuvieron un p-value de .717666 que es mayor a 0.05, por lo que se rechaza la hipótesis nula y se determina que las muestran no presentan diferencias significativas. En la Figura 3.6 se observa graficamente los valores calculados por cada muestra y la distancia que existe entre ellas.

- Las tasas de ocupación en el sector informal obtuvieron un p-value de .380299 que es mayor a 0.05, por lo que se rechaza la hipótesis nula y se determina que las muestran no

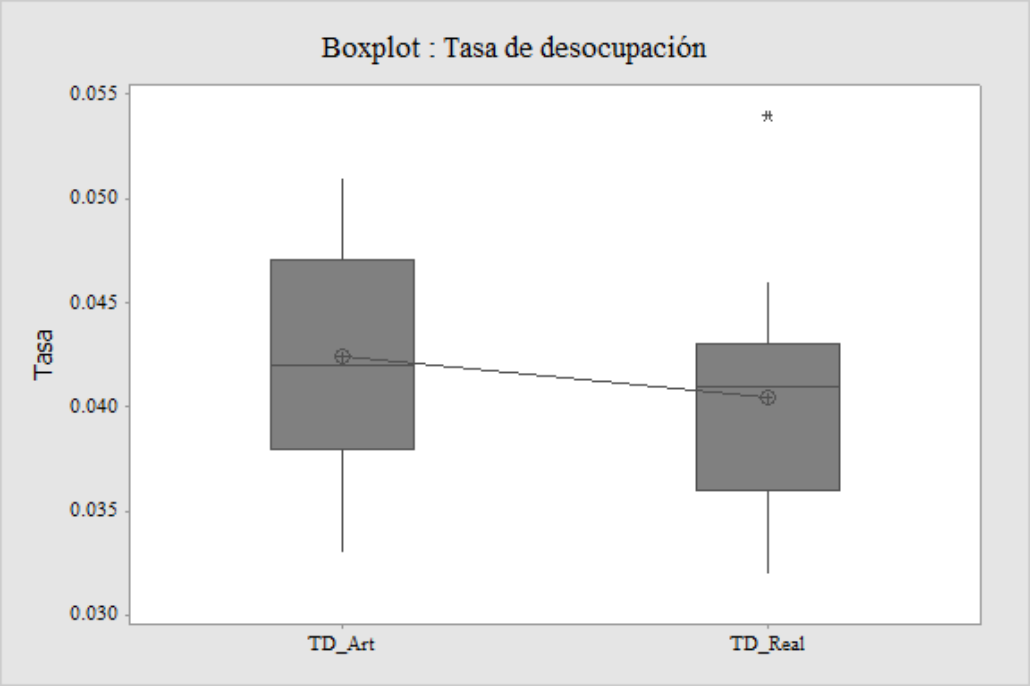


Figura 3.5: T-Test Tasa de desocupación

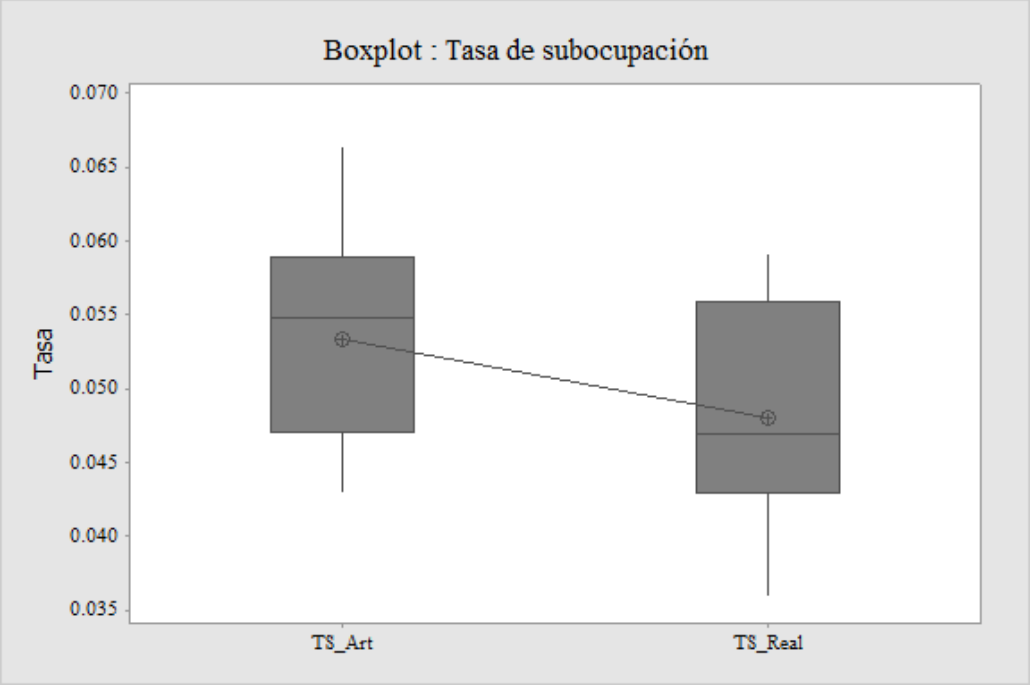


Figura 3.6: T-Test Tasa de subocupación

presentan diferencias significativas. En la Figura 3.7 se observa graficamente los valores calculados por cada muestra y la distancia que existe entre ellas.

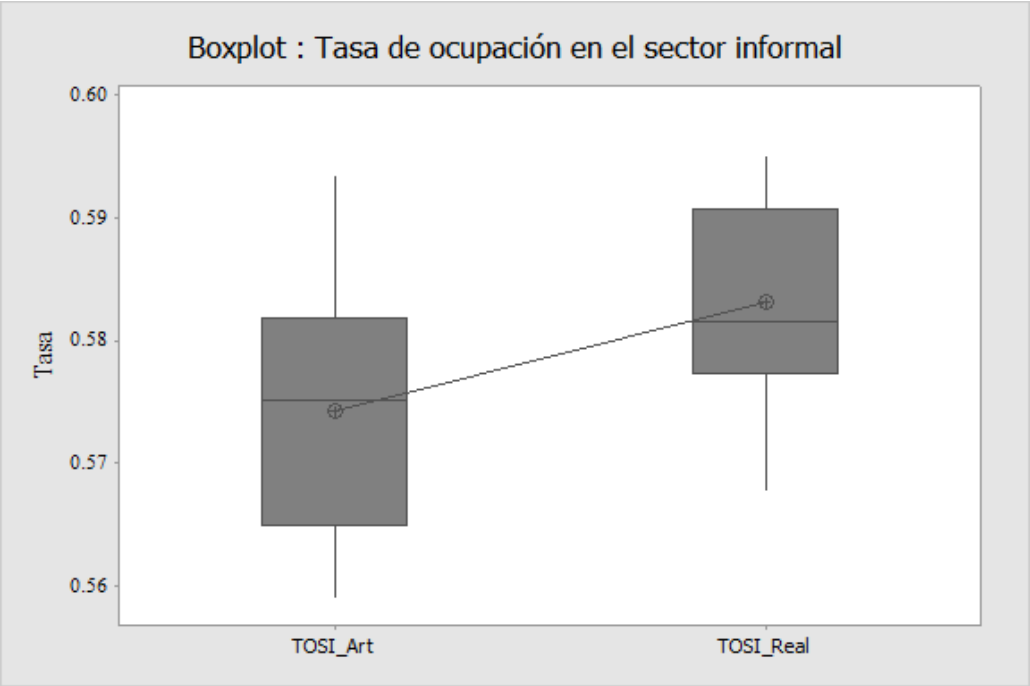


Figura 3.7: T-Test Tasa de ocupación en el sector informal

3.3 Análisis de personas que cumplen 15

Durante los periodos de la simulación 122 agentes cumplieron 15 años, por lo que se generó una base de datos con estos para analizar sus propiedades dada la ocupación que obtuvieron mediante inferencias a la red bayesiana correspondiente. En la Figura 3.8, se muestra la distribución de la ocupación siendo la condición no disponible la que tuvo mayor incidencia con un 71.3%, esto se debe a que muchos agentes se encontraban estudiando la secundaria, en la Figura 3.9 se muestra la distribución de las propiedades generales de los agentes con 15 años de edad. En la Figura 3.10 se presentan las tablas de probabilidad condicional del modelo de red bayesiana de ocupación. Para analizar la incidencia en la no diponibilidad laboral de los agentes se muestran las tablas de probabilidad condicional instanciando como evidencia que se encuentran estudiando el nivel de secundaria, como se puede observar en la Figura 3.11.

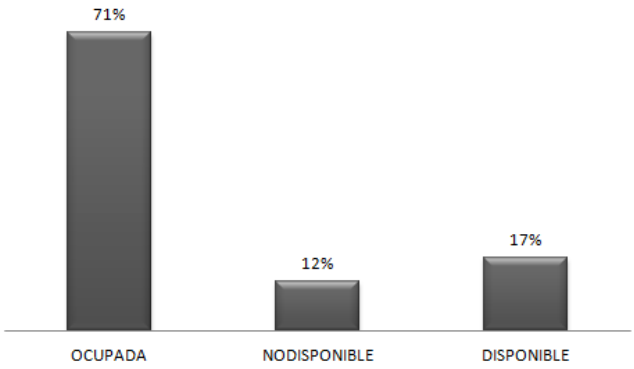


Figura 3.8: Distribución de ocupación agentes de 15 años

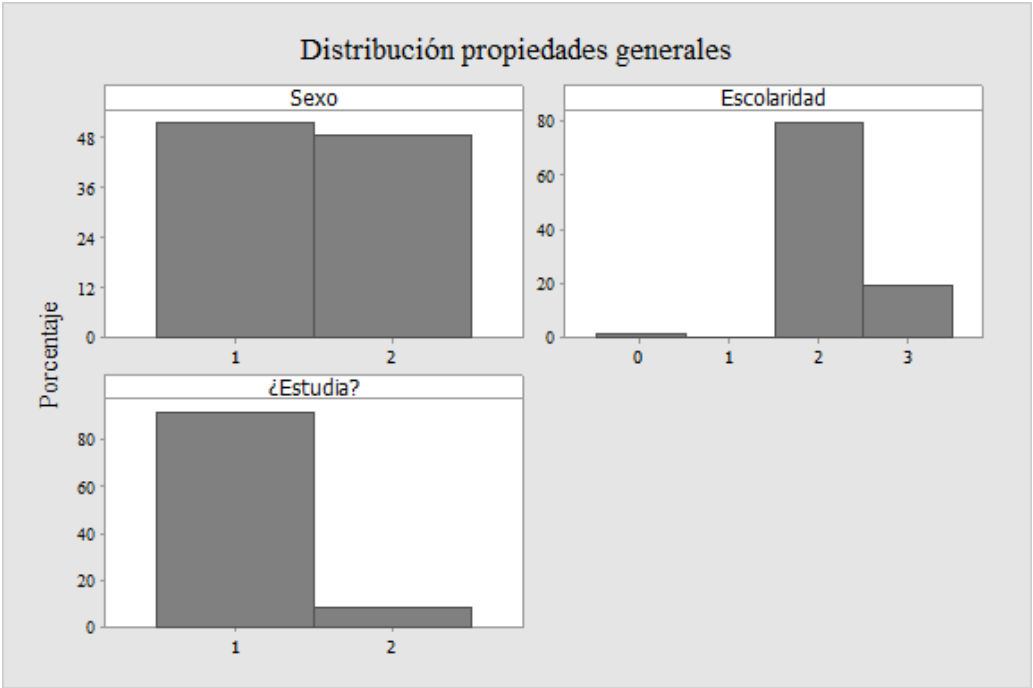


Figura 3.9: Propiedades generales

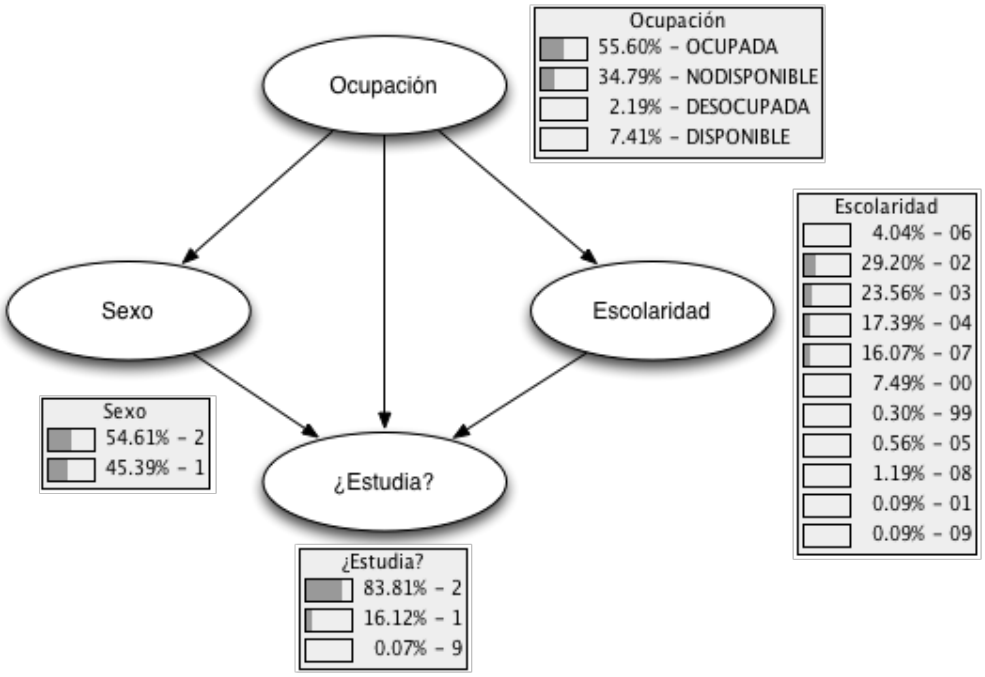


Figura 3.10: Tablas de probabilidad condicional

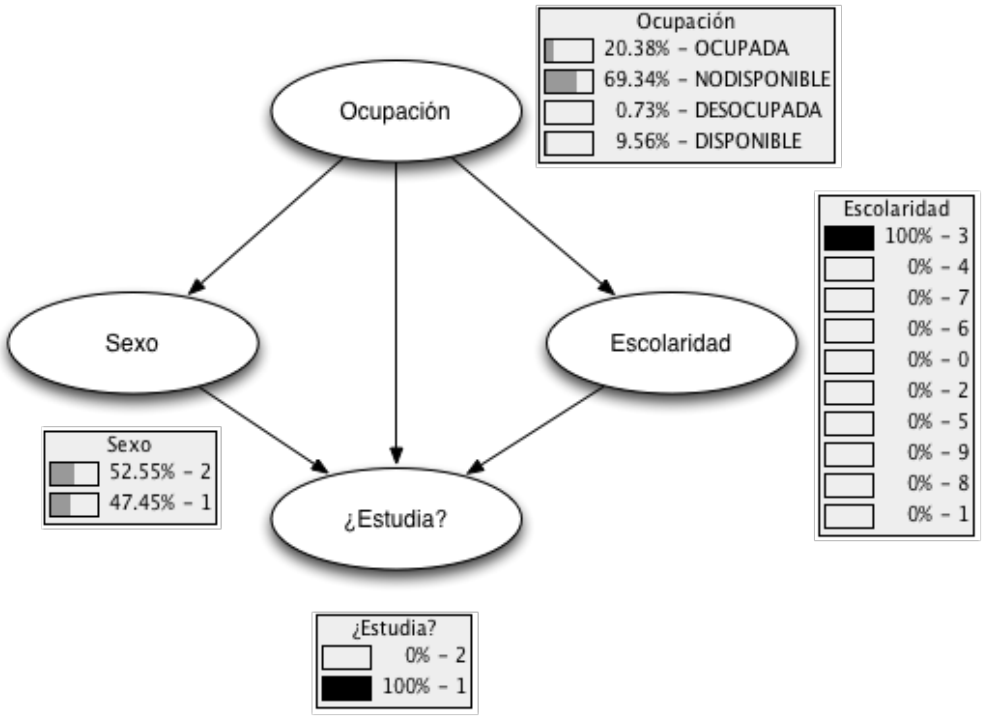


Figura 3.11: Inferencia de agentes que se encuentran estudiando

Por último se cambio la propiedad que determina si se encuentran estudiando y se observa en las tablas de probabilidad condicional que aumenta la posibilidad de convertirse en ocupados,

como se observa en la Figura 3.12.

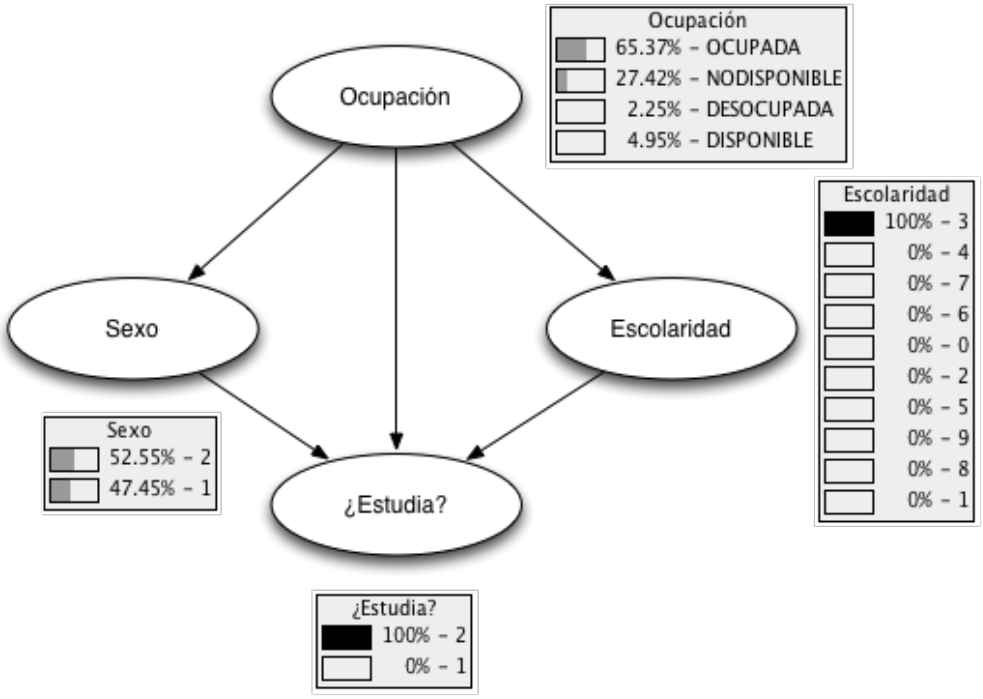


Figura 3.12: Inferencia de agentes que no se encuentran estudiando

3.3.1 Oportunidades laborales de individuos de 15 años

Se generó una base de datos de los individuos que cumplieron 15 años y obtuvieron un empleo con la finalidad de analizar las oportunidades laborales que tienen dado el modelo de red bayesiana laboral. En la figura 3.13, podemos observar las distribuciones de las propiedades laborales :

- **Posición en la ocupación.** Se observa una mayor incidencia en trabajos subordinados, trabajos por cuenta propia y trabajos sin salario sucesivamente.
- **Seguridad social.** La mayoría de estos agentes no cuentan con acceso a seguridad social.
- **Rama.** Obsevamos una gran incidencia en el sector secundario que corresponde a empresas dedicadas a la transformación de materias primas, como la minería o maquilas.
- **Nivel de ingreso.** Los niveles de ingreso en donde se encontró mayor incidencia son por debajo de los 3 salaríos mínimos

- **Jornada laboral.** El número de horas laboradas a la semana con más incidencia se dan en jornadas con menos de 15 o más de 48 horas.
- **Informalidad laboral.** La mayoría de éstos agentes obtuvieron trabajos en el sector informal.
- **Subocupación.** Los resultados nos muestran que la mayoría de estos agentes no se encuentran en busca de otro empleo.

.

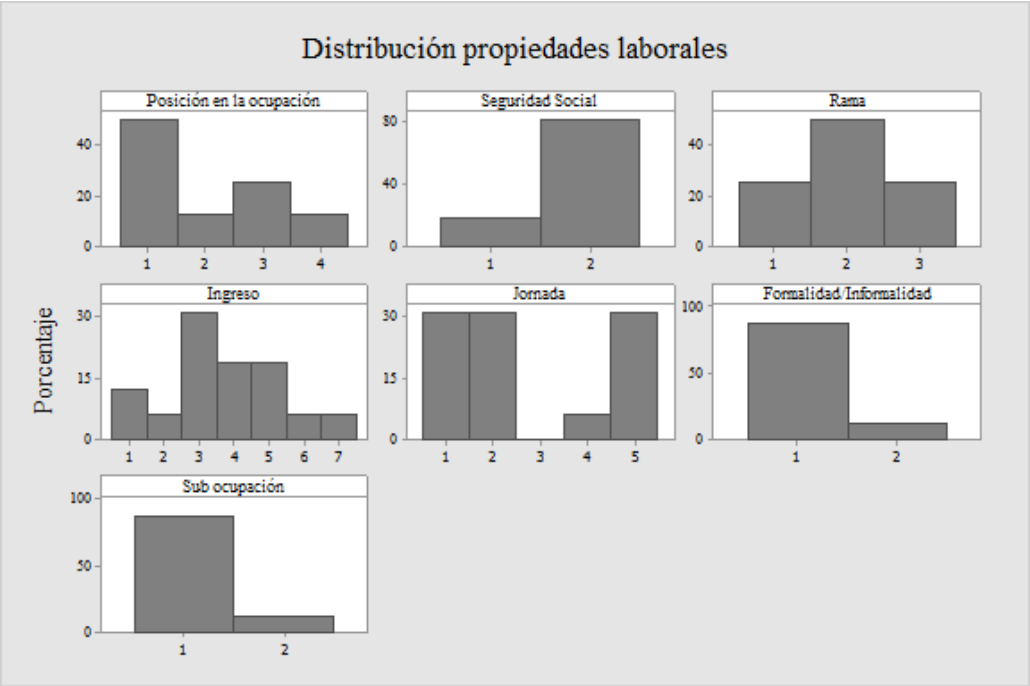


Figura 3.13: Propiedades laborales de los individuos de 15 años

3.4 Análisis de red bayesiana laboral

El modelo de red bayesiana, como se mencionó en el Capítulo 1, nos permite observar las relaciones causales de las propiedades que se determinaron más relevantes para la abtracción o representación del fenómeno, para este trabajo las condiciones laborales en el Estado de Veracruz, y al mismo tiempo nos permite observar las oportunidades laborales que tienen las personas dadas sus propiedades individuales mediante inferencias al modelo. En esta sección

se presenta el análisis del fenómeno a observar desde la perspectiva cualitativa y cuantitativa del modelo generado.

El fenómeno a observar son las condiciones laborales en Veracruz, por lo que se determinó generar el modelo considerando la informalidad o formalidad del empleo como la propiedad laboral a clasificar dadas las propiedades individuales de las personas y las condiciones de su empleo.

A continuación se hace un análisis por nodo o variable y las relaciones que existen entre ellas, según el modelo de red bayesiana y sus tablas de probabilidad condicional el cual podemos observar en la figura 3.14:

- **Escolaridad.** Es la única variable independiente del modelo. La escolaridad influye directamente en el tipo de empleo de las personas y en la edad. En la TCP podemos observar que los niveles de escolaridad se concentran en los valores 2 y 3, que corresponden a la primaria y secundaria, con un 28% y 24% respectivamente, seguidos por el nivel 7 y 4, nivel profesional y preparatoria con un 18.66% y 16.21%.

- **Edad.** En este modelo la edad tiene dependencia del nivel de escolaridad que tienen las personas. Se observa que el hecho que las personas estudien y si tienen acceso a seguridad social depende de la edad. La edad de las personas de la muestra con la que se construyó el modelo tiene mayor reincidencia en los rangos 2 y 3, que representan las edades de 25 a 44 y de 45 a 64 años respectivamente.

- **¿Estudia?** Solo un 5.70% de las personas que laboran continúan con sus estudios. Podemos observar que depende de la edad de las personas para determinar si estudian. El hecho que se encuentren estudiando influye en la formalidad o informalidad de su empleo y la posición que ocupan en él.

- **Sexo.** La muestra con la que se generó el modelo tiene más hombres con un 60.72% que mujeres con un 39.28%. Podemos observar como el sexo de las personas está relacionado con el número de horas que laboran, la rama y la formalidad o informalidad de su empleo,

por lo que se determinó hacer un análisis de las oportunidades laborales que tienen las personas según su género el cual se presenta en la siguiente sección.

- **Formal/Informal.** Un 56.82% de las personas que laboran, se desempeñan en la informalidad. La informalidad influye directamente en el ingreso de las personas, en que tengan o no acceso a la seguridad social, la posición que ocupan en su empleo y si se encuentran en constante busqueda de otro trabajo.

- **Seguridad social.** Un 35.71% de las personas que laboran cuentan con acceso a seguridad social, lo que depende de la formalidad o informalidad de su empleo. El acceso a la seguridad tiene influencia en la posición de la ocupación de las personas, en la rama en la que se desarrollan y en el número de horas que laboran a la semana.

- **Posición.** En la muestra con la que se generó el modelo, el 41.31% de las personas trabajan por cuenta propia y el 35.48% son subordinados. Solo el 16.49% son empleadores. La posición en la ocupación de la personas influye en el ingreso que tienen.

- **Ingreso.** El ingreso de las personas tiene mayor incidencia en los rangos que incluyen menor número de salaríos mínimos, con un 24.13% de uno a dos salaríos, 15.44% un salrio y 19.59% de dos a tres salarios, es decir, el 59.16% de las personas perciben menos de tres o hasta tres salarios mínimos. El ingreso depende de la posición en la ocupación y de la formalidad o informalidad del empleo, y de él depende la jornada laboral y la rama.

- **Jornada.** Los porcentajes más altos en la jornada laboral por semana se observan en los rangos de 35 a 38 a horas con 42.79% y más de 48 horas con 32.59%. La jornada laboral depende del ingreso y de si las personas tienen acceso a la seguridad social. Y de ella depende si las personas con empleo continuan en busca de otro y el sexo de las personas.

- **Subocupación.** Solo un 6.10% de las personas que laboran se encuentran en busca de otro empleo. El modelo nos dice que la jornada laboral, el sexo de las personas y la formalidad o informalidad del empleo son factores que influyen en que las personas continuen en busca de un empleo.

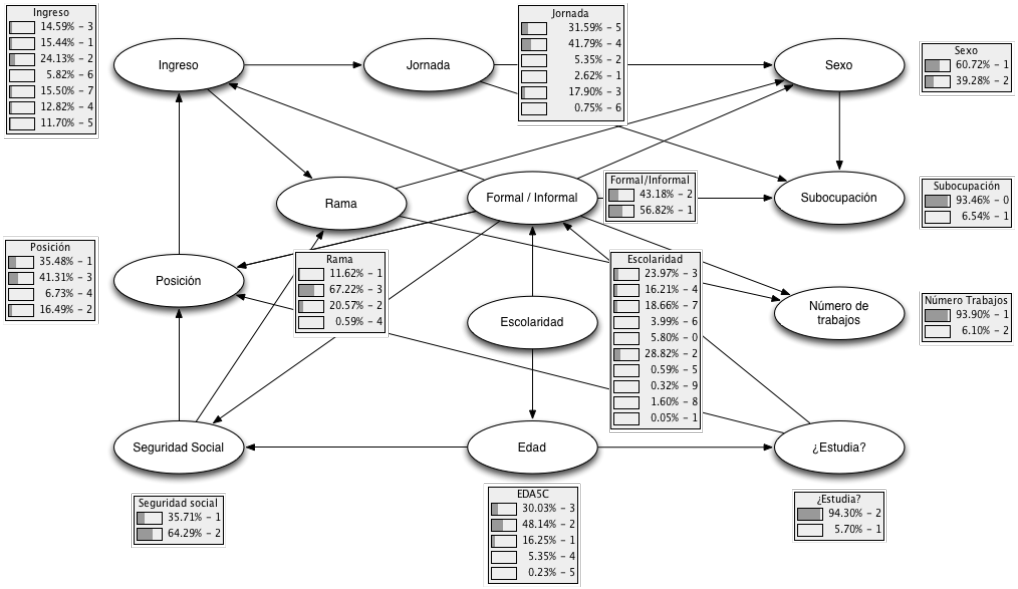


Figura 3.14: Red bayesiana con tablas de probabilidad condicional

3.4.1 Oportunidades laborales por sexo

Para analizar las oportunidades laborales por sexo se ingresó como evidencia en el modelo el sexo de las personas y se observó la influencia en las propiedades laborales. En la figura 3.15, se muestra el impacto de la propagación de las probabilidades de las variables con el sexo femenino y en la figura 3.16 con el sexo masculino.

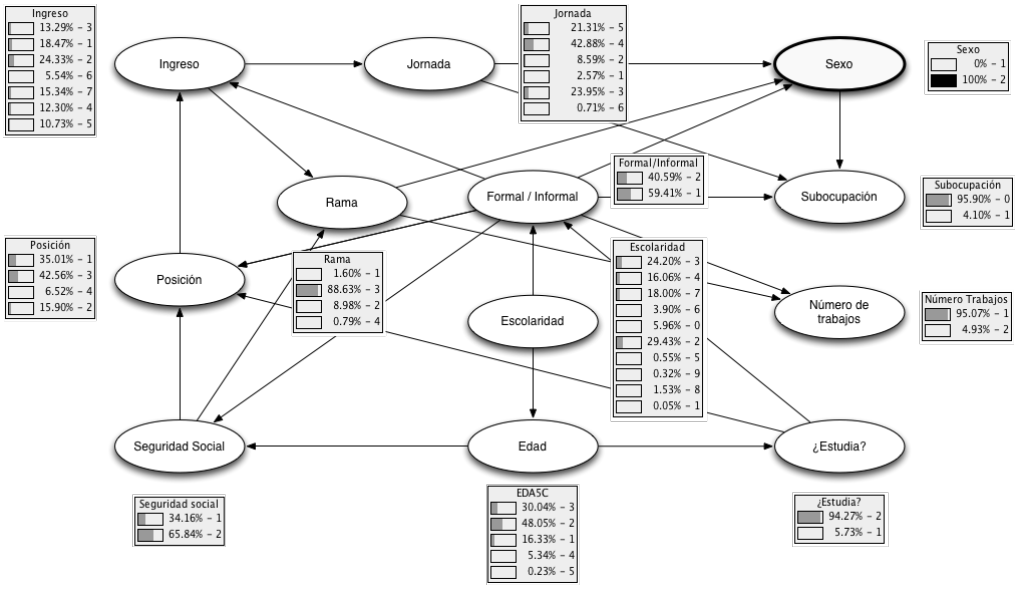


Figura 3.15: Propagación de probabilidades de las condiciones laborales del sexo femenino

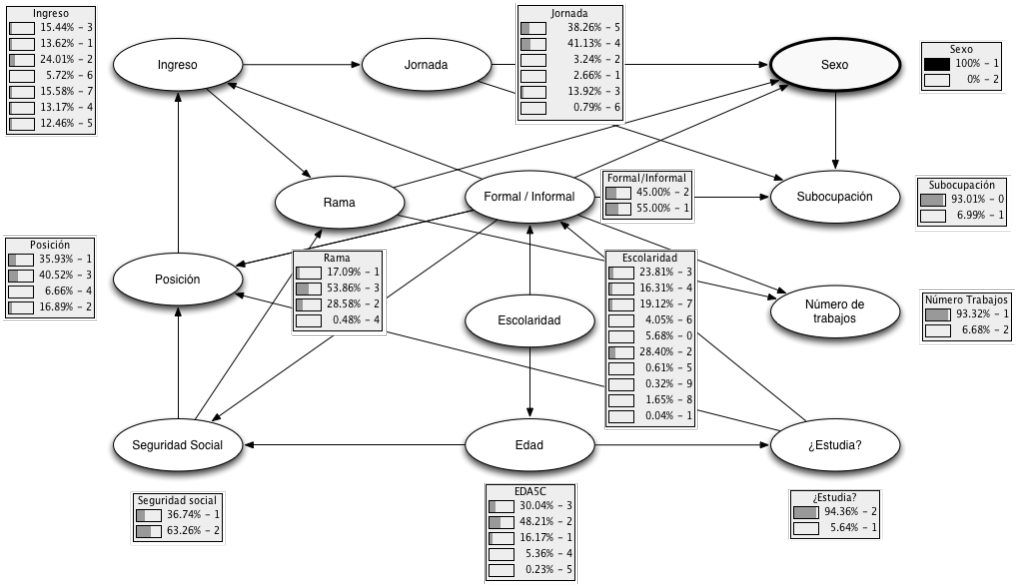


Figura 3.16: Propagación de probabilidades de las condiciones laborales del sexo masculino

Se determinó observar las variables de ingreso, jornada, rama, subocupación y formalidad o informalidad laboral, por lo que se generaron gráficas con la distribución de las tablas de probabilidad condicional obtenidas de las propagaciones con el sexo de las personas como evidencia. En las gráficas el eje Y representa la probabilidad que tienen las personas por sexo de tener las propiedades laborales determinadas en el eje X:

- **Ingreso** En la gráfica 3.17, podemos observar que las mujeres tienen probabilidades más altas solo en los rango uno y dos, en donde el ingreso es menor a dos salarios mínimos. A partir de un ingreso mayor a 2 salarios mínimos hasta 5 salarios mínimos los hombres tienen más probabilidades de obtener este ingreso, teniendo porcentajes de probabilidad muy similares en los rangos 6 y 7.

- **Jornada** En la gráfica 3.18, podemos observar que las mujeres tienen mayor probabilidad de encontrar un empleo con jornadas laborales de hasta 48 horas a la semana, y los hombres tienen mas probablidad de encontrar un empleo en donde laboren más de 48 horas a la semana. Los rangos en los que se presenta la mayor probabilidad de jornada laboral en la mujeres son 3 y 4, de 15 a 34 horas y de 35 a 48 horas, mientras que los hombre tienen más probabilidades de encontrar empleos en donde laboren de 35 a 48 horas ó más de 48 horas a la semana.

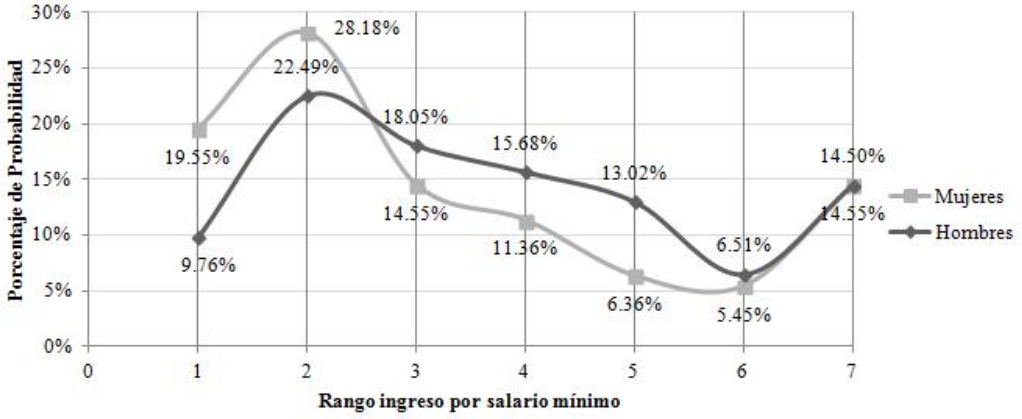


Figura 3.17: Comparación de ingreso por sexo

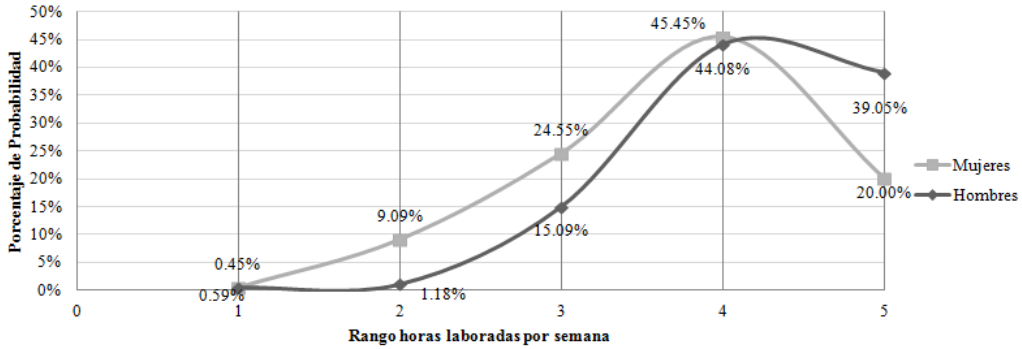


Figura 3.18: Comparación de jornada por sexo

- **Rama** En la gráfica 3.19, podemos observar que los hombres tienen más probabilidades de encontrar un empleo en los sectores primario y secundario, mientras que con las mujeres es más probable que encuentre un empleo en el sector terciario.

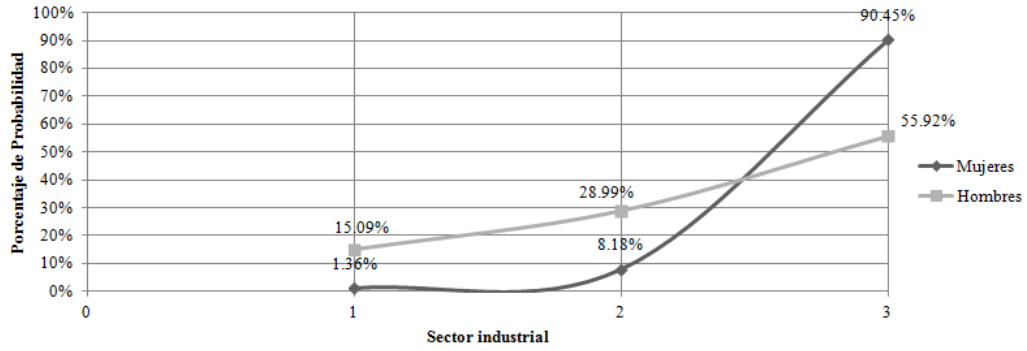


Figura 3.19: Comparación de rama por sexo

- **Subocupación** En la gráfica 3.20, podemos observar que los hombres tienden más a buscar otro empleo a pesar de ya tener uno.

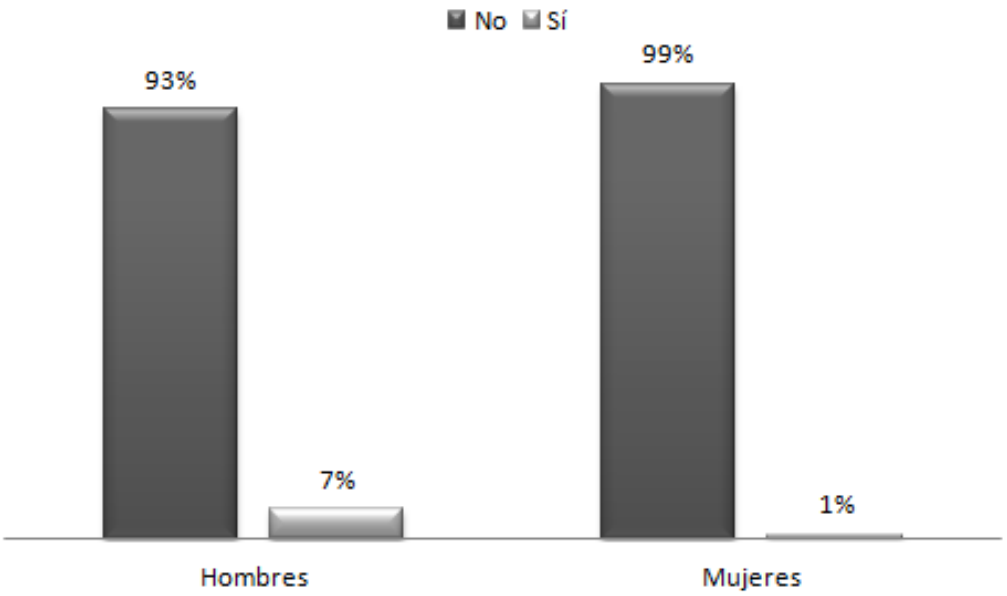


Figura 3.20: Comparación de subocupación por sexo

- **Formalida/Informalidad** En la gráfica 3.21, podemos observar que las mujeres tienen mas probalidad de encontrar un empleo informal que los hombres. También se muestra que los hombres tienen más posibilidades de encontrar un empleo formal.

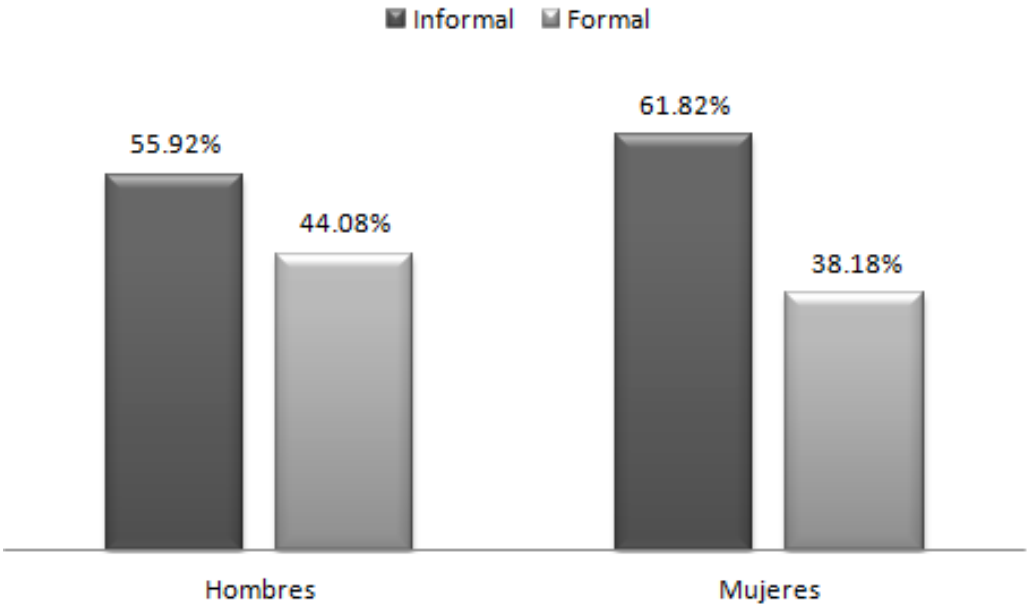


Figura 3.21: Comparación de formalidad o informalidad laboral por sexo

Capítulo 4

Conclusiones

Los datos generados en una simulación social en donde la abstracción del fenómeno a estudiar toma forma de modelos de redes bayesianas y las propiedades iniciales de los agentes son obtenidas directamente de la base de datos de la que se generan los modelos, nos permite obtener resultados que presentan una gran similitud a datos reales.

La generación de los datos artificiales, en donde los agentes reportan cada período de tiempo sus propiedades, nos permite comparar los resultados con bases de datos reales de los periodos simulados. Esto nos permite validar estadísticamente los resultados para poder determinar objetivamente la similitud que existe entre los datos reales y los datos artificiales.

Los modelos de red generados por el sistema nos permiten explicar el fenómeno, mediante la observación de la relación causal que existe entre las propiedades seleccionadas de la base de datos. Es importante resaltar que se generaron diferentes modelos para determinar los parámetros y las propiedades que generaron la red bayesiana que mejor se ajustaron a la explicación del fenómeno a observar, sin embargo esto no quiere decir que los modelos implementados sean los únicos que lo pueden hacer.

La implementación del sistema en una plataforma Jason-CArtAgO [4], nos facilitó el desarrollo de la generación de los modelos y el manejo de la base de datos. La simulación presentada en este trabajo se desarrolla bajo un enfoque estadístico, por lo que se puede decir que una

plataforma multiagente no es la herramienta natural para su implementación, sin embargo este trabajo es una primera aproximación a una simulación en donde se combinen los enfoques lógico y estadístico, por lo que tener el sistema en esta plataforma, nos permitirá agregar comportamientos individuales o colaborativos, especificados lógicamente.

A partir de las pruebas estadísticas que se muestran en el capítulo anterior, podemos determinar que los resultados obtenidos consistentes, ya que presentan gran similitud con los datos reales con los que se compararon. Y los modelos generados por el sistema nos permiten analizar el desarrollo del sistema a nivel individual y el impactó de los cambios de propiedades de los agentes a nivel global. Lo que nos permite observar y analizar las relaciones causales del modelo mediante el seguimiento de individuos con ciertas características, como son las oportunidades laborales de las personas en el Estado de Veracruz dadas su edad ó sexo.

4.1 Trabajo futuro

Este trabajo es una primera aproximación a una simulación social bajo un enfoque lógico-estadístico, por lo que en un futuro se implementarán los componentes lógicos al sistema implementado para el caso de estudio; para esto primero se desarrollarán las siguientes extensiones:

- Las bases de datos que generá el INEGI, nos permiten conocer a que hogar pertenece cada persona; con lo que se crearán artefactos que representen a los hogares, con la finalidad de que los agentes conozcan las propiedades de los integrantes de su familia.

- Se pretende impementar por medio de artefactos, una abstracción de políticas públicas, para observar de que forma afectan o influyen a los individuos.

Una vez que se tenga la representación de los hogares y la abstracción de políticas públicas, se implementarán los componentes lógicos, los cuales determinarán las acciones y la toma de decisiones de los agentes considerando lo siguiente:

- Las propiedades individuales que tiene cada agente, obtenidas mediante las inferencias a los modelos de red bayesiana.
- Las propiedades de los miembros de su hogar o familia.
- Las propiedades a nivel hogar, es decir, podremos observar el ingreso total por familia, el sexo del jefe de familia, el número de hijos, el número de personas que laboran y el número de personas que estudian.

La toma de decisiones de los agentes, consistirá en el conocimiento de esta información, pero se determinará el curso de acción a través de la implementación de diferentes estrategias colaborativas, con la finalidad de encontrar con cual se obtienen resultados más cercanos a los datos reales. La determinación de los comportamientos que se asemejen más a la realidad se hará por medio de comparaciones estadísticas, con lo que se evitará que la validación de los componentes lógicos del sistema sean subjetivas.

En base a la experiencia ganada con el caso de estudio, se pretende diseñar un sistema en el que se puedan simular diferentes fenómenos sociales, determinados por las bases de datos que se implementen y poniendo a disposición diferentes estrátegias que determinen los comportamientos de los agentes.

Bibliografía

[1] Frédéric Amblard, Pierre Bommel, and Juliette Rouchier. Assessment and validation of multi-agent models. *Phan, D., Amblard, F.(eds.)*, pages 93–114, 2007.

[2] Itzhak Benenson, Itzhak Orner, and Erez Hatna. *Agent-based modeling of householders' migration behavior and its consequences*. Springer, 2003.

[3] Francesco C Billari. *Agent-based computational modelling: applications in demography, social, economic and environmental sciences*. Taylor & Francis, 2006.

[4] Rafael H Bordini, Jomi Fred Hübner, and Michael Wooldridge. *Programming multi-agent systems in AgentSpeak using Jason*, volume 8. John Wiley & Sons, 2007.

[5] Adnan Darwiche. Samiam. *Software available from http://reasoning. cs. ucla. edu/samiam*.

[6] JONG De. Ka an analysis of the behavior of a class of genetic adaptative systems. *Ann Arbor, USA, Ph. D Thesis-Department of Computer and Comunication Sciences, University of Michigan*, 1975.

[7] Nigel Gilbert and Klaus Troitzsch. *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[8] Volker Grimm, Uta Berger, Donald L DeAngelis, J Gary Polhill, Jarl Giske, and Steven F Railsback. The odd protocol: a review and first update. *Ecological modelling*, 221(23):2760–2768, 2010.

[9] Xavier Limón, Alejandro Guerra-Hernández, Nicandro Cruz-Ramírez, and Francisco Grimaldo. An agents and artifacts approach to distributed data mining. In *Advances in Soft Computing and Its Applications*, pages 338–349. Springer, 2013.

[10] Dennis L Meadows, William W Behrens, Donella H Meadows, Roger F Naill, Jørgen Randers, and Erich Zahn. *Dynamics of growth in a finite world*. Wright-Allen Press Cambridge, MA, 1974.

[11] Guy H Orcutt, Joachim Merz, and Hermann Quinke. *Microanalytic simulation models to support social and financial policy*, volume 7. North Holland, 1986.

[12] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288, 1986.

[13] Simulating Social Phenomena. editors r. conte, r. hegselmann, 1997.

[14] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.

[15] Eric Silverman, Jakub Bijak, Jason Hilton, Viet Dung Cao, and Jason Noble. When demography met social simulation: a tale of two modelling approaches. *Journal of Artificial Societies and Social Simulation*, 16(4):9, 2013.

[16] F Willekens. Migration: A perspective from complexity science. In *Migration workshop of the Complexity Science for the Real World network, Chilworth, UK*, volume 16, 2012.

[17] Michael Wooldridge, Nicholas R Jennings, et al. Intelligent agents: Theory and practice. *Knowledge engineering review*, 10(2):115–152, 1995.

Apéndice A

Código de agentes

A.1 Agente machine_learning

```
1  // Agent machine_learning in project simulacionENOE
2
3  +!machineLearning : true
4      <-
5    !obtenDatos;    //InstancesBase
6    !generaRed;     //BayesNet
7    !preparaInferencia. //SamIam
8
9  //Herramientas de Weka
10
11 +!obtenDatos : true
12   <-
13   makeArtifact("instancesBase","empleo_informal.InstancesBase",[],InstancesBaseId);
14   focus(InstancesBaseId);
15   cargaARFF("Datos_Sim.arff"). // operacion del artefacto InstacesBase
16   makeArtifact("BayesNet","empleo_informal.BayesNetArtifact",[],BayNetId);
17     focus(BayNetId);
18     linkArtifacts(BayNetId, "portInstancesBase", InstancesBaseId).
19
20 //Herramientas de Weka
21
22 +!generaRed : true
23   <-
24   leeBaseDatos;  //Obtiene Base de Datos del Artefacto InstacesBase
```

```
25    enviaAlgBus; //Selecciona el Algoritmo de Busqueda
26    enviaParametros; //Determina los parametros del Algoritmo de Busqueda para generar la red
27    generaClasificador; //Genera modelo de red bayesiana
28    calculaTCP; //Calcula tablas de probabilidades condicionales
29    guardaBN.
30
31  // Herramientas SamIam
32  +!preparaInferencias : True
33    <-
34    makeArtifact("Red_Empleo","simulacionENOE.Red_empleo",[],RedEmpleoId);
35    makeArtifact("Red_Condicion","simulacionENOE.Red_Condicion",[],RedCondicionId);
36    focus(RedEmpleoId);
37    focus(RedCondicionId);
38    leerRedBayesiana; //Obtiene red bayesiana laboral
39    iniciaParametros; //Proporciona par\'ametros para consulta
40    obtenVar; //Obtiene valores de propiedades de artefacto InstancesBase
41    .send (controller,tell,comienza). //Avisa a controller que la red esta lista
```

A.2 Agente control

```
1   //Crea a los agentes que representan a las personas
2   +comienza[source(Ag)]
3     <-
4     !iniciaEscenario.
5
6   +!iniciaEscenario : numPersonas(Num) & porDesocu(PorDesocu) & porOcu(PorOcu) &
7                       porDispo(PorDispo) & porNodispo(PorNoDispo) & porMenor(PorMenor)
8     <-
9     focusWhenAvailable("baseDatos");
10    !generaPoblacion("DESOCUPADO",PorDesocu);
11    !generaPoblacion("OCUPADO",PorOcu);
12    !generaPoblacion("NODISPONIBLE",PorNoDispo);
13    !generaPoblacion("DISPONIBLE",PorDispo);
14    !generaPoblacion("MENOR",PorMenor).
15
16
17  +!generaPoblacion(Con,Por) : true
18    <-
19    for (.range(X,0,Por-1)){
20      obtenFil(X,Con,Fil);
21      idPer(Fil,Con,Id);
```

```
22      registro(Id,Con);
23      .create_agent(Id,"persona.asl",[agentArchClass("c4jason.CAgentArch")]);
24    }.
25
26  //Al recibir mensaje de todos los agentes creado incia cambios de periodos de tiempo
27  @notifyStart[atomic]
28  +started(_)
29  : not started &  numPersonas(Num) & .count(started(_), Num) & periodo(P)
30  <-  startTime[artifact_name("tiempo")];
31    -+periodo(P + 1);
32    !preparaCambioTiempo.
33
34  +!preparaCambioTiempo: periodo(T)
35    <-
36    obtenTD(T,NTD); //Obtiene la tasa de desocupacion del periodo simulado
37    tombola(NTD,PEA_F); //Selecciona aleatoriamente a los agentes a desemplear
38    nuevo_Per(T);
39    .broadcast(tell, nuevo_periodo(T)).
40
41  //Recibe mensaje de los agentes persona despues de reportar sus propiedades
42  +listo(_) : not listo & numPersonas(Num) & .count(listo(_), Num) & periodo(P) & pasoTiempo(N)
43    <-
44    if (N = P){
45      !stop;
46    }else{
47      .println("Termina Periodo : ",P);
48      .abolish(listo(_));
49      -+periodo(P + 1);
50      !preparaCambioTiempo;
51    }.
52
53  //Detiene la simulacion al concluir los periodos determinados
54  +!stop : true
55    <-
56    .println("FIN").
```

A.3 Agente persona

```
1  // Agent persona in project empleo_informal
2
3  /* Initial goals */
```

```
4
5   !start.
6
7   /* Plans */
8
9   +!start : true
10    <-
11    !obtenPropiedades;
12    .send(controller, tell, started(Me)).
13
14  //Obtiene sus propiedades generales de la base de datos
15  +!obtenPropiedades : true
16    <-
17    .my_name(Me);
18    .term2string(Me,N);
19    obtenAtt(N,1,Sex); //Sexo
20    +sexo(Sex);
21    obtenAtt(N,2,Eda); //Edad
22    +edad(Eda);
23    obtenAtt(N,3,GraEs); //Escolaridad
24    +gradoEsco(GraEs);
25    obtenAtt(N,4,Estu); //Estudia
26    +estudia(Estu);
27    if(Cond == "OCUPADA"){  // Si son personas con ocupadas obtiene propiedades laborales
28        obtenAtt(N,10,Pos_Ocu);
29        +posOcu(Pos_Ocu);
30        obtenAtt(N,11,Seg_Soc);
31        +segSoc(Seg_Soc);
32        obtenAtt(N,12,Rama);
33        +rama(Rama);
34        obtenAtt(N,13,Ingreso);
35        +ingreso(Ingreso);
36        obtenAtt(N,14,Jornada);
37        +jornada(Jornada);
38        obtenAtt(N,16,T_Tra);
39        +tTra(T_Tra);
40        obtenAtt(N,17,Emp_Ppal);
41        +emp_ppal(Emp_Ppal);
42        obtenAtt(N,18,Sub_O);
43        +sub_o(Sub_O);
44    }
45    logging(N,Propiedades).
46
47  //Si obtienen un nuevo empleo
```

```
48  +nuevo_periodo(T)[source(Ag)] : nuevoEmpleo(N)
49    <-
50    cumple_anio(Eda,T,N_Eda);
51    .concat("SEX",",",Sex,",","EDA5C",",",N_Ran,",","CS_P13_1",",",GraEs,",","CS_P17",",",Estu,
          Evi); //Evidencia
52    obtenMrPr("POS_OCU","SEG_SOC","RAMA_EST1","EMP_PPAL","ING7C","SUB_O","DUR_EST","T_TRA",Evi,
53                Pr_Pos,Pr_Seg,Pr_Rama,Pr_Em,Pr_Ing,Pr_Sub,Pr_Dur,Pr_Tt); //Obtiene TCP de SamIam
54    utils.ruleta(Pr_Pos,Pos_Ocut); //Accion interna de ruleta para determinar propiedades
          laborales
55    utils.ruleta(Pr_Seg,Seg_Soct);
56    utils.ruleta(Pr_Rama,Ramat);
57    utils.ruleta(Pr_Em,Emp_Ppalt);
58    utils.ruleta(Pr_Ing,Ingresot);
59    utils.ruleta(Pr_Sub,Sub_Ot);
60    utils.ruleta(Pr_Dur,Jornadat);
61    utils.ruleta(Pr_Tt,T_Trat);
62    logAttOcu(N,Cond1,Sex,N_Ran,GraEs,AnEs,Estu,Pos_Ocu,Seg_Soc,Rama,Ingreso,Jornada,Tip_Con,
          T_Tra,Emp_Ppal,Sub_O,Busqueda)[artifact_name("logging")];
63    .send(controller, tell, listo(Me)).
64
65  //Menores que cumplen 15 anios
66  +nuevo_periodo(T)[source(Ag)] : Cond == "MENOR"
67    <-
68    cumple_anio(Eda,T,N_Eda);
69    if(N_Eda == "15"){
70      .concat("SEX",",",Sex,",","CS_P17",",",Estu,",","CS_P13_1",",",GraEs,Evi);
71      obtenMrPr_Cond("CONDICION",Evi,Pr_Cond); //Obtiene TCP
72      utils.ruleta(Pr_Cond,N_Cond); //Accion interna que determina nueva condicion
73      if(N_Cond = 1){
74        N_Cond1 = "NODISPONIBLE";
75      }
76      if(N_Cond = 2){
77        N_Cond1 = "OCUPADA";
78        nuevoEmpleo; //Hace inferencias al modelo de condiciones laborales para obtener las
              popiedades
79          }
80      if(N_Cond = 3){
81        N_Cond1 = "DISPONIBLE";
82      }
83      if(N_Cond = 4){
84        N_Cond1 = "DESOCUPADA";
85      }
86    }
87    else{
```

```
88      logAttMenor(N,Cond,Sex,"0",GraEs,AnEs,Estu)[artifact_name("logging")];
89    }
90    .send(controller, tell, listo(Me)).
91
92  //Agentes con empleo
93  +nuevo_periodo(T)[source(Ag)] : Cond == "OCUPADA"
94      <-
95    cumple_anio(Eda,T,N_Eda);
96    obten_cambio_ocu(N,Kn); //Verifica si pierde empleo
97    if(Kn == "S"){ //Pierde empleo
98      Cond1 = "DESOCUPADA";
99    }else{  //Continua con su empleo
100     logAttOcu(N,Cond,Sex,N_Ran,GraEs,AnEs,Estu,Pos_Ocu,Seg_Soc,Rama,Ingreso,Jornada,Tip_Con,
            T_Tra,Emp_Ppal,Sub_O,Busqueda)[artifact_name("logging")];
101   }
102   .send(controller, tell, listo(Me)).
```

Apéndice B

Publicación

Se presentaron los primeros resultados de la implementación de los modelos de redes bayesianas en el Congreso Mexicano de Inteligencia Artificial, celebrado en el Instituto Nacional de Astrofísica, Óptica y Electrónica, en la ciudad de Tonanzintla, Puebla del 23 al 24 de Mayo del 2016. El trabajo presentado en este congreso será publicado en la revista *Reserach in Computing Science*, ISSN 1870-4069, indexada en DBLP, LatIndex y Periodica.

Un modelo de red bayesiana de la informalidad laboral en Veracruz orientado a una simulación social basada en agentes

Jean Christian Díaz-Preciado, Alejandro Guerra-Hernández, and Nicandro Cruz-Ramírez

Universidad Veracruzana, Centro de Investigación en Inteligencia Artificial
Sebastián Camacho No 5, Xalapa, Ver., México 91000
jean_christian12@msn.com, aguerra@uv.mx, ncruz@uv.mx

Resumen La informalidad en el empleo en México es un fenómeno social de interés, ya que cerca de un 60 % de los trabajadores se desempeña en este sector. En este trabajo se propone un análisis de la informalidad en el empleo con la creación de un modelo de red bayesiana a partir de la base de datos generada de la Encuesta Nacional de Ocupación y Empleo, obtenida mediante el Instituto Nacional de Estadística y Geografía, con el propósito de utilizarla posteriormente en una simulación social basada en agentes y artefactos, en la cual los agentes obtengan algunas de sus propiedades de la base de datos y otras por inferencias a la red bayesiana generando datos artificiales. Se hace una comparación estadística de los datos generados en el sistema con los datos reales, lo cual se utilizará en un futuro para la validación de la simulación social bajo un paradigma lógico-estadístico.

Keywords: base de datos, encuesta, redes bayesianas, agentes, validación estadística

A bayesian network model of labour informality in Veracruz oriented to agent-based social simulation

Abstract In Mexico, informal employment is a social phenomenon of interest, given that about 60 % of the workers are in this situation. This paper presents an analysis of informal employment based a bayesian network model obtained from the data from the National Survey of Occupation and Employment, obtained by the National Institute of Statistics and Geography. The model is intended to be used in an agent-based social simulation, where agents get some properties directly from the data base and some others through bayesian inference, generating in this way artificial data. A statistical comparison of the data generated in the system and the real data will be used in the future for validation of social simulation under a logical-statistical paradigm.

Keywords: data base, survey, bayesian network, agents, statistical validation

1. Introducción

En México, el Instituto Nacional de Estadística y Geografía (INEGI) es el encargado de recabar información de personas, hogares y empresas, mediante la aplicación de encuestas y censos en periodos determinados, que tienen como objetivo principal proveer información para la generación de estadistícas que sirven como parámetro para la toma de decisiones en la implementación de políticas públicas. Estas encuestas y censos tienen diferentes temáticas, en este trabajo nos enfocamos en la Encuesta Nacional de Ocupación y Empleo (ENOE), que proporciona información sobre la ocupación de las personas, es decir, si tienen empleo o no; y en el caso de los empleados, que características tienen sus empleos. Los resultados publicados por el INEGI acerca de la informalidad laboral son presentados mediante indicadores y estadísticas con una cobertura geográfica nacional y estatal, desglosadas por sexo, edad y sector de la actividad [6].

El objetivo de este trabajo es construir un modelo que nos permita clasificar la situación laboral de una persona como formal o informal, dadas las prestaciones que proporciona su empleo y el contexto individual del trabajador. Se decidió que el modelo tomase la forma de una red bayesiana y se adoptó una aproximación de minería de datos basada en agentes para su construcción. Aunque esto no es mandatorio, nuestro interés por usar el modelo más adelante, en una simulación social basada en agentes, justifica la decisión. Hemos adoptado una aproximación de minería de datos basada en Agentes y Artefactos [8], muy similar a la usada en la herramienta JaCa-DDM [7]: Se provee una serie de artefactos basados en Weka [10], para almacenar datos, generar el modelo y evaluarlo. Los agentes usan estos artefactos en el proceso de aprendizaje. Puesto que el modelo y los datos son accesibles a los agentes, vía estos artefactos, los agentes pueden obtener algunas de sus propiedades de la base de datos y otras mediante inferencias bayesianas a partir del modelo. Los agentes reportan sus actividades generando una base de datos artificial, la cual es comparada estadísticamente con la base de datos real. Eventualmente nos gustaría modelar como afecta la implementación de políticas públicas la decisión laboral de los agentes, es decir, si optan por una situación formal o informal.

El artículo está organizado de la siguiente manera: En el capítulo 2 se muestran las características de la base de datos ENOE y la descripción de las variables utilizadas en este trabajo. En el capítulo 3 se hace la descripción del sistema que genera el modelo y los datos artificiales. Posteriormente, el capítulo 4 describe el diseño experimental y el capítulo 5 los resultados obtenidos del mismo. Finalmente se presentan las conclusiones y trabajo futuro, en los capítulos 6 y 7, respectivamente.

2. Base de datos ENOE

La base de datos ENOE está construída a partir de entrevistas realizadas en 120,000 viviendas repartidas en todo México, recabando información de alrededor de 800,000 personas. La información de la ENOE, puede estudiarse a nivel

de vivienda, hogar y persona, la base de datos con la que se realizó el modelo fue obtenida de la tabla sociodemográfica (SDEMT110 [5]) del primer trimestre del año 2010. Los periodos posteriores podrán usarse para la validación de los datos artificiales generados.

El INEGI clasifica a la población en edad de trabajar legalmente, mayores de 15 años, en dos grandes grupos: Población Económicamente Activa (PEA), que es la que ejerce presión en el mercado laboral; y Población No Económicamente Activa (PNEA). Como podemos observar en la figura 1, la PEA se divide a su vez en Población Ocupada y Desocupada, según se tenga o no. Dentro de la población ocupada hay población sub ocupada refiriéndose a las personas que tienen un empleo pero continúan en busca de otro. La PNEA se divide en las personas disponibles, que aunque no están ocupadas ni buscando empleo al momento de la encuesta, bajo ciertas circunstancias podrían decidir incorporarse al mercado laboral; y las personas no disponibles, que son las que se encuentran bajo un contexto que les impide laborar.

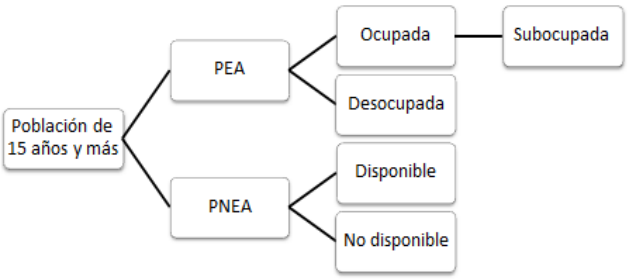


Figura 1. Clasificación de la población [4]

Con la idea de atender la problemática regional, en este trabajo se considera la población sub ocupada que reside en el estado de Veracruz.

2.1. Descripción de variables

El cuadro 1 describe las variables consideradas para este trabajo, las cuales dividimos en generales y laborales. Las variables generales incluyen sexo, edad, escolaridad y si las personas estudian al momento de la encuesta; como se mencionó, nos enfocamos en las personas sub ocupadas, por ello se incluyen las variables que nos indican si se encuentran en busqueda de un nuevo empleo, y el motivo de la busqueda. Las variables laborales corresponden a algunas de las características de los empleos, como es el nivel de ingreso, la duración de la jornada y la formalidad o informalidad del empleo. La variable de clasificación de empleo nos indica la infomalidad o formalidad del mismo, por ello es considerada como variable clase, ya que se quiere observar que tipo de empleo puede tener una persona dadas sus características.

Cuadro 1. Variables seleccionadas para la creación del modelo [5]

Atributos generales		
Atributo	**Variable**	**Descripción**
Sexo	1	Hombre
	2	Mujer
Edad	1	14 a 24 años
	2	25 a 44 años
	3	45 a 64 años
	4	65 años y mas
Escolaridad	1	Preescolar
	2	Primaria
	3	Secundaria
	4	Bachillerato
	5	Normal
	6	Técnica
	7	Profesional
	8	Maestría
	9	Doctorado
Estudia actualmente	1	Sí
	2	No
Busca otro empleo	1	Sí
	2	No
Motivo de busqueda	1	Para tener otro empleo
	2	Para cambiarse de empleo
	3	No buscó

Atributos laborales		
Atributo	**Variable**	**Descripción**
Ingreso	1	Hasta 1 salario mínimo
	2	De 1 a 2 salarios mínimos
	3	De 2 a 3 salarios mínimos
	4	Des 3 a 5 salarios mínimos
	5	Mas de 5 salarios mínimos
Jornada	1	Ausente temporales
	2	Menos de 15 horas
	3	De 15 a 34 horas
	4	De 35 a 48 horas
	5	Mas de 48 horas
Clasificación de empleos	1	Empleo informal
	2	Empleo formal

Los nombres de las variables de la fuente original fueron cambiados para dar mayor claridad. Es importante especificar la relación que existe entre las variables generales y su representación como propiedades de los agentes que definiremos en nuestro sistema como trabajadores; así como la relación entre las variables laborales y su representación como propiedades de los artefactos que definiremos en nuestro sistema como empresas. El siguiente capítulo presenta la descripción detallada del sistema de Agentes y Artefactos propuesto.

3. Sistema multiagente para crear el modelo

Los agentes del sistema están basados en redes probabilísticas para el manejo de la incertidumbre. Se implementó una red bayesiana, la cual se define como un modelo probabilísitico representado mediante un grafo acíclico dirigido (GAD), en el cual los nodos representan las variables del fenómeno y las dependencias probabilistas que existan entre ellas se encuentran en la estructura del grafo. Asociada a cada nodo de la red hay una distribución de probabilidad condicional (TPC), dependiente de los nodos padre [9].

La razón de utilizar redes bayesianas es facilitar la interpretación del modelo mediante el GAD y que nos permite observar la probabilidad que tiene una persona de tener un empleo formal o informal, y de las características del empleo, según su edad, sexo y escolaridad. Esto se realiza por medio de inferencias al modelo enviando como evidencia las variables generales [3].

Dado que el sistema está basado en Agentes y Artefactos, es deseable que la red bayesiana sea accesible a los agentes, ya sea como parte de ellos o como parte del algún artefacto. Para construir el modelo, hemos extendido JaCa-DMM [7], definiendo una nueva estrategia de aprendizaje basada en redes bayesianas y agregando un artefacto que encapsula SamIam [1], para realizar las inferencias. La figura 2 muestra el diagrama general del sistema, que tiene como entrada una base de datos y una red bayesiana que puede ser ingresada manualmente o generada por el sistema a partir de los datos. A continuación se describen a detalle las tareas realizadas por los artefactos y agentes en el sistema, en la figura 3 podemos observar el diagrama de los procesos realizados por estos.

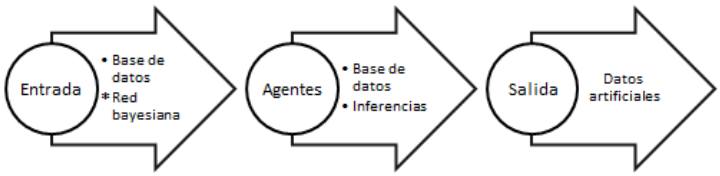


Figura 2. Diagrama general del sistema multiagente

3.1. Artefactos

El sistema cuenta con 3 artefactos que realizan tareas correspondientes a la minería de datos, los cuales se describen a continuación:

InstancesBase Este artefacto se adopta directamente de la herramienta JaCa-DDM, que a su vez la adopta de Weka. Se trata de un repositorio de ejemplos de entrenamiento que puede cargar archivos ARFF para su uso en clasificadores, evaluadores y demás herramientas Weka. Al tomar la forma de artefacto, los ejemplos y sus atributos son accesibles a los agentes del sistema y a otros artefactos.

BayesNet Este artefacto encapsula los métodos de construcción de redes bayesianas implementados en Weka. Su tarea principal es construir modelos a partir de datos y ponerlos a disposición del artefacto basado en SamIam. Dependiendo de las entradas del sistema realiza las siguientes tareas:

- Leer una red bayesiana generada previamente en formato XMLBIF
- Generar una red bayesiana a partir de los datos del artefacto InstancesBase y crear un modelo con los parámetros descritos en el capítulo 4 (Diseño experimental). Estos parámetros son fijos. Una vez generado el modelo, éste se guarda en formato XMLBIF.

SamIam Es el artefacto que los agentes usan para hacer inferencias basadas en una red bayesiana, la generada con el artefacto BayesNet. Por medio de las herramientas de SamIam, envía los parámetros para tener acceso a las tablas TPC; Recibe las evidencias de los agentes trabajadores para hacer la inferencia de las variables no conocidas, poniendo a disposición de los agentes las TPC obtenidas.

3.2. Agentes

El sistema cuenta con tres clases de agentes, dos se utilizan para crear agentes que controlan los experimentos y una para crear agentes que representan trabajadores. A continuación se describen las tareas que realiza cada agente:

Agente learning El sistema inicia su ejecución con este agente, el cual tiene como meta realizar las tareas correspondientes a la minería de datos por medio de los artefactos, para que los demás agentes tengan acceso a la base de datos y al modelo. Para poder concretar su meta, crea los artefactos y establece las ligas necesarias entre ellos. Su plan sigue esta secuencia:

- Leer base de datos (IntancesBase)
- Generar modelo (BayesNet)
- Preparar modelo para inferencias (SamIam)

El agente está diseñado para iniciar sus acciones ya sea recibiendo la base de datos y crear la red bayesiana; O recibir un modelo generado previamente. Al concluir sus tareas, envía un mensaje al agente control para que este comience sus actividades.

Agente control La meta de este agente es crear a los agentes que representan a los trabajadores a partir de los datos almacenados en el artefacto InstancesBase. Su única creencia es el número de personas que debe crear. Al momento de crear un nuevo agente le proporciona un nombre, que corresponde al número de caso de la base de datos, para que el nuevo agente obtenga sus propiedades generales de éste. También es el encargado de controlar el acceso al artefacto SamIam al momento de las inferencias de los agentes que representan los trabajadores.

Agente persona Esta clase de agente se usa para representar trabajadores, por lo que su meta principal es instanciar sus propiedades generales y laborales. Para las propiedades generales, recupera sus datos almacenados en el artefacto InstancesBase a partir de su nombre. Las propiedades laborales son inferidas en el artefacto SamIAm, con base en la evidencia que proveen las propiedades generales del agente y el modelo almacenado en el artefacto BayesNet. Puesto que las TPC generadas tienen 2 o más variables, se ejecuta una acción interna que representa la ruleta propuesta por DeJong [2] y según su probabilidad condicional, se determina la propiedad laboral del agente. Una vez obtenidas todas sus propiedades, estas son almacenadas en un archivo de texto.

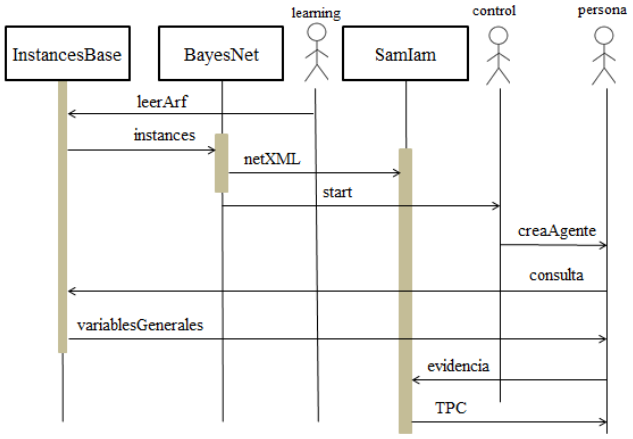


Figura 3. Diagrama de procesos del sistema multiagente

4. Diseño experimental

El sistema multi-agente propuesto, proporciona como salida una base de datos artificial y, si se desea, la red bayesiana generada de la base de datos. Como se mencionó, se determinó construir este modelo a partir de las personas sub ocupadas con residencia en el estado de Veracruz, las cuales en la ENOE del primer trimestre del 2010 constituyen un universo de 595 casos.

El cuadro 2, muestra los parámetros implementados para la generación de la red bayesiana, los cuales se determinaron en base a la observación de las redes generadas en experimentos utilizando Weka. En estos experimentos se observó que los parámetros que modificaban significativamente al modelo fueron el iniciarlo con un modelo Naive y la implementación de la manta de Markov, ya que sin ella, dos variables consideradas como relevantes, la edad y si el trabajador está estudiando, quedaban fuera de la red. Se llevó a cabo una validación cruzada del model con diez pliegues.

Cuadro 2. Parámetros para generar el modelo en Weka

Estimador	Simple Estimador	A 0.5
Algoritmo de Busqueda	HillClimber	
Parámetros del Algoritmo de busqueda	initNaiveBayes	False
	MarkovBlanket Classifier	True
	mxNrOfParents	10,000
	scoreTYPE	MDL
	useArcReversal	True

Dado que las variables generales son obtenidas directamente de la ENOE, la validación de los datos artificiales se realiza comparando la distribución de las

variables laborales. Posteriormente se generó una red bayesiana con la base de datos artificial la cual se comparó con la base de datos generada por el sistema para complementar la validación.

5. Resultados

El modelo generado por el sistema a partir de los datos reales, es muy similar al generado por Weka con los mismos datos. Con los datos artificiales se generó un modelo en Weka usando los mismo parámetros. El modelo obtenido con los datos de la ENOE se muestra en la figura 4(a), y el modelo generado por los datos artificiales en la figura 4(b). Los resultados estadísticos calculados por Weka se pueden observar en el cuadro 3.

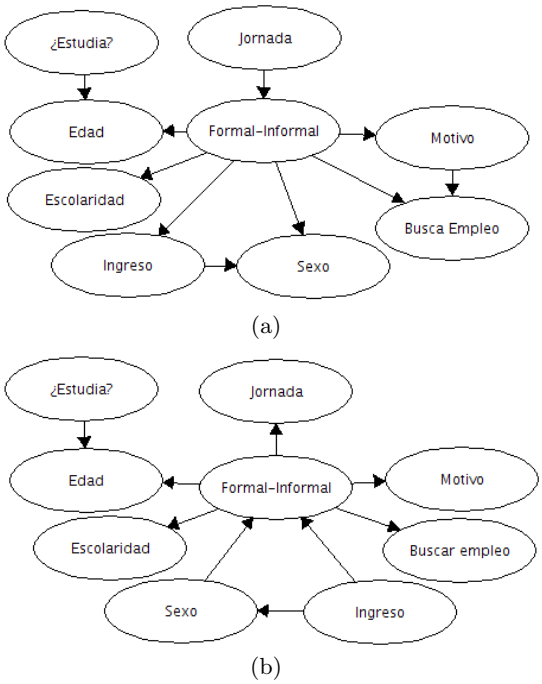
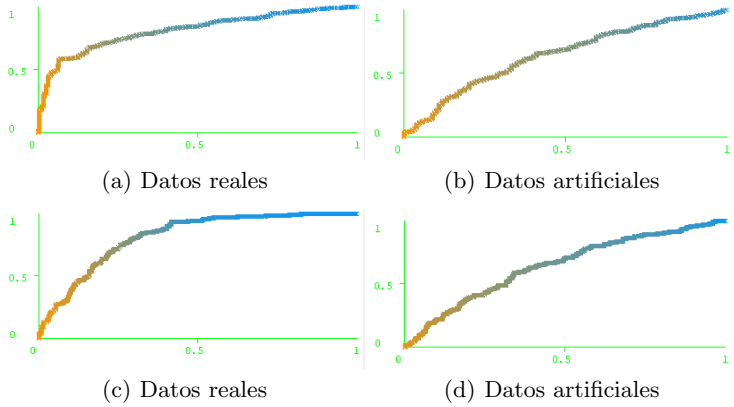


Figura 4. Modelos de red bayesiana obtenidos: (a) Datos reales y (b) Datos artificiales.

Se generaron las gráficas de curva ROC, con umbral de 0.5, para cada variable de la clase, con ambas bases de datos. Las gráficas ROC para la variable de informalidad en el empleo se muestran en las Figura 5(a) para los datos reales y en la Figura 5(b) para los artificiales, y para la variable de formalidad en el empleo en las Figuras 5(c) y 5(d).

Cuadro 3. Comparación de estadísticas de la generación del modelo

	Datos reales	Datos Artificiales
Porcentaje de clasificados correctamente	74.11 %	63.80 %
Desviación Estándar	5.46	6.13
Área bajo la curva ROC	0.8180	0.6383



(a) Datos reales (b) Datos artificiales

(c) Datos reales (d) Datos artificiales

Figura 5. Curvas ROC de los modelos generados

También se realizó una comparación de las distribuciones de las variables que se obtuvieron en las inferencias al modelo. Dado que se generó el mismo número de agentes que de casos de la base de datos, la distribución de las variables generales son las mismas. En el Cuadro 4 se muestran las estadísticas de las variables laborales.

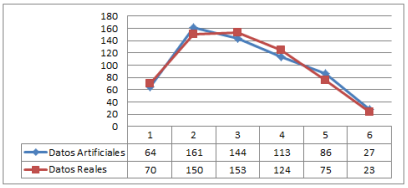
Cuadro 4. Comparación de estadísticas

Variable	Media		EE Media		DesvEst		Varianza		Asimetria		Curtosis	
	Real	Artificial	Real	Artificial	Real	Artificial	Real	Artificial	Real	Artificial	Real	Artificial
Ingreso	3.0891	3.1294	0.0547	0.0560	1.3335	1.3652	1.7782	1.8637	0.25	0.28	-0.71	-0.80
Jornada	3.5345	3.5244	0.0432	0.0434	1.0541	1.0576	1.1112	1.1185	-0.49	-0.48	-0.38	-0.39
Buscar	1.7916	1.7983	0.0167	0.165	0.4065	0.4616	0.1652	0.1613	-1.44	-1.49	0.07	0.22
Motivo	2.2168	3.2134	0.0314	0.0310	0.7665	0.7564	0.5876	0.5722	-0.39	-0.38	-1.21	-1.17
Formal/Informal	1.3849	1.3849	0.0200	0.0200	0.4870	0.4870	0.2371	0.2371	0.47	-1.78	-1.78	-1.78

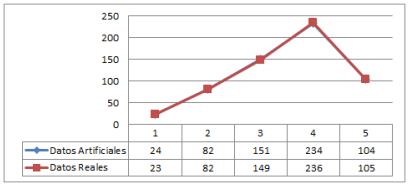
La Figura 6 muestra la comparación gráfica de la distribución de las variables. Los datos generados por medio de las inferencias y la aplicación de la ruleta, nos proporcionan resultados muy parecidos a los reales.

5.1. Informalidad en el Estado de Veracruz

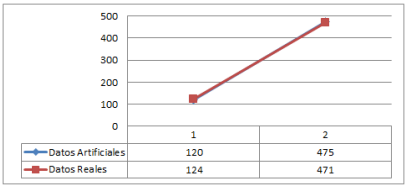
Como se mencionó, uno de los principales objetivos del trabajo del INEGI es dotar de estadísticas e información a los órganos encargados de la generación de



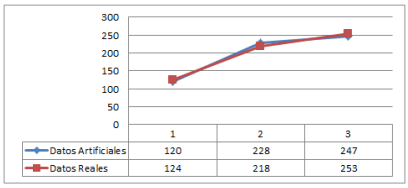
(a) Ingreso



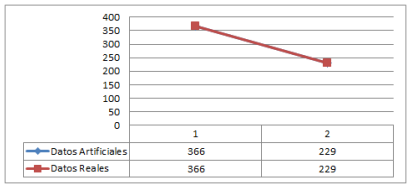
(b) Jornada



(c) Buscar Empleo



(d) Motivo



(e) Informalidad

Figura 6. Comparación de distribución de variables

políticas públicas. Estas estadísticas se presentan como gráficas con 2 variables a observar, por ejemplo la cantidad de personas por sexo que tienen un empleo formal o informal.

En este trabajo se realizó el análisis de los datos por medio de consultas al modelo, observando la propagación de las probabilidades dados ciertos valores de las variables que se observaron. Es importante mencionar que para este experimento, no se consideraron todas las variables que provee la base de datos ENOE y solo se trabajó con las personas subocupadas del Estado de Veracruz. A continuación se presentan tres consultas generadas:

- La variable clasificiación de empleo es dependiente del número de horas que trabaja una persona, se observó que solo en el rango de 35 a 48 horas laborales, hay más personas con empleos formales que informales, figura 7.
- Para los empleos informales se observó que las personas cuentan con un menor nivel escolar, los sueldos son menores y hay más personas en busca de un nuevo empleo para dejar el actual, Figura 8.
- Para los empleos formales se observó que las personas tienen un mayor nivel escolar, los sueldos se sitúan en rangos más altos y es menor el número de persona en busca de otro empleo, sin embargo las que están en busca de otro empleo es en su mayoría por que desean dejar su empleo actual, figura 9.

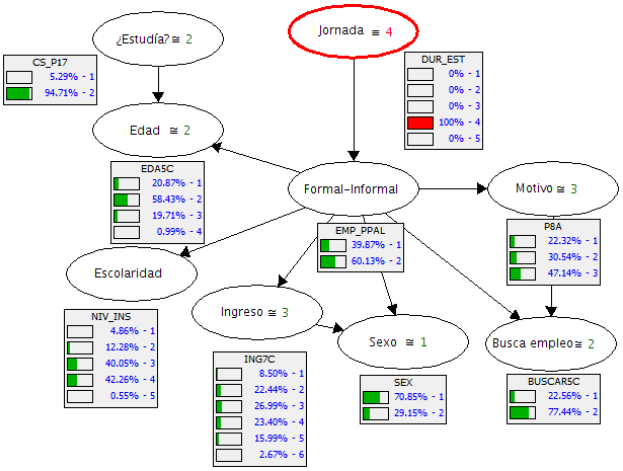


Figura 7. Propagación por Jornada Laboral [1]

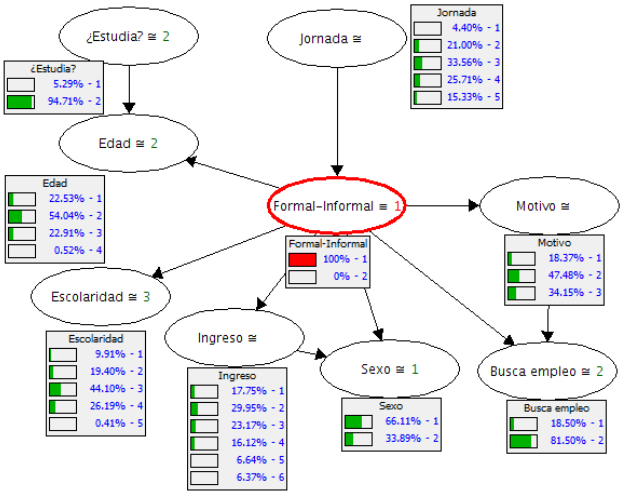


Figura 8. Propagación por Informalidad en el Empleo [1]

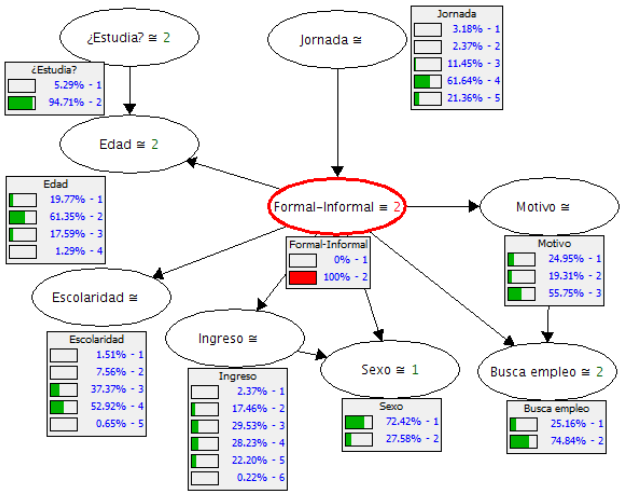


Figura 9. Propagación por Formalidad en el Empleo [1]

6. Conclusiones

Podemos observar como la red generada con los datos artificiales, es muy similar a la generada con los datos reales, considerando que el valor de las variables que adquieren los agentes como propiedades laborales, está determinado por una ruleta, se da oportunidad de adquirir valores que no tengan la probabilidad más grande. Las diferencias en los GAD de los modelos se describen en el cuadro 5.

Cuadro 5. Diferencia entre modelos

Datos reales	Datos artificiales
El nodo Jornada es independiente	El nodo Jornada es dependiente del nodo de clasificación del empleo
El nodo de clasificación del empleo es dependiente del nodo Jornada	El nodo de clasificación del empleo es dependiente de los nodos Ingreso y Sexo

A pesar que los datos con los que se genera el modelo son solo 595, y el número de nodos de la red son 9, la distribución de las variables que se obtienen por medio de inferencias tienen una gran similitud con las reales. Consideramos que el porcentaje de casos clasificados correctamente con el modelo generado es bueno teniendo 74.11 % con una desviación estándar de 5.46, y a pesar de que las distribuciones de las variables de los datos artificiales son similares, el porcentaje que se obtuvo es de 63.80 %, con una desviación estándar de 6.13, evidentemente el porcentaje disminuye más de un 10 %, sin embargo la desviación estándar no

es muy alta, por lo que se espera tener mejoras experimentando con bases de datos que contemplen un número mayor de casos.

Se generó el área bajo la curva ROC, para visualizar como los datos artificiales tienden a disminuir la generalidad del modelo, y con esto a cometer errores en la clasificación. Estos resultados nos proporcionan un panorama del número de agentes que debe tener nuestra simulación así como la complejidad de la red bayesiana que se implementará, para mantener la consistencia que nos proporciona el modelo en nuestros datos artificiales. Las comparaciones estadísticas que observamos en el cuadro 4 nos muestran las similitudes entre las bases de datos. Observamos que el atributo correspondiente al ingreso es el que tiene mayor diferencia en la distribución de los datos, esto se debe al número de variables que tiene y la distribución de las mismas, que observamos en la desviación estándar y la varianza, como complemento se calculó la curtosis que es de -0.71 y el valor de asimetria de 0.25; el atributo jornada también tiene un número mayor de variables, pero no se presentan cambios significativos en los datos ya que tiene una desviación estándar menor, y tiene una curtosis de -0.39 con un valor de asimetría de -0.48. Los atributos que solo tienen dos o tres variables presentaron las menores diferencias en la comparación, teniendo una distribución igual en la variable correspondiente a la formalidad o informalidad del empleo. El error cuádratico medio (EMC) calculado es de 0.81, con lo cual se determinó que los resultados son aceptables para el caso de estudio.

La implementación de la minería de datos en un entorno de agentes y artefactos, es una herramienta útil que nos facilita observar el flujo de trabajo que se realiza. Los artefactos nos proporcionan la distribución de las tareas que se requieran efectuar, ya sea desde la lectura de la base de datos para la generación del modelo o partiendo de un modelo ya generado. Los agentes se utilizaron para definir el orden en que se deben efectuar los procesos. Los resultados de la predicción mediante inferencias a redes bayesianas nos proporcionan datos aproximados a los datos con los que se genera el modelo. La determinación de analizar una base de datos obtenida de encuestas a personas con redes bayesianas, nos proporciona información sobre la causalidad del fenómeno a observar.

7. Trabajo futuro

En un futuro se prentede utilizar las inferencias que realizan los agentes al modelo en una simulación social, en donde el valor de los atributos que obtengan influyan en su toma de decisiones. Dicho esto se probará las inferencias con bases de datos con más casos, y con periodos de tiempo definidos, por ejemplo, observar el comportamiento de las variables en un periodo de un año equivalente a cuatro encuestas de la ENOE. Como se mencionó la ENOE puede observarse a nivel hogar, por lo que se pretende hacer experimentos a este nivel, representando los hogares por medio de artefactos, con el fin que los agentes que pertenezcan a un mismo hogar tengan acceso a la información de los integrantes de su familia. Se pretende implementar por medio de artefactos, una abstracción de políticas públicas, es decir, de qué forma afectan a un hogar o individuo, y

agregar comportamientos de los agentes con respecto a los cambios con los que se enfrentará.

Agradecimientos

El primer autor cuenta con el apoyo de la beca CONACyT número 633473.

Referencias

1. Darwiche, A.: Modeling and Reasoning with Bayesian Networks. Cambridge University Pres, New York (2009)
2. De Jong, K.A.: Analysis of the behavior of a class of genetic adaptive systems. Tech. Rep. 185, The University of Michigan (1975)
3. Glymour, C.: Discovering Causal Structure. Academic Press, Orlando, Florida (1987)
4. INEGI: Conociendo la base de datos de la ENOE. Datos ajustados a proyecciones de población 2010. INEGI (2010)
5. INEGI: ENOE. Descripción de Archivos. INEGI (2010)
6. INEGI: México: Nuevas estadísticas de informalidad laboral. INEGI (2013)
7. Limón, X., Guerra-Hernández, A., Cruz-Ramírez, N., Grimaldo, F.: An agents and artifacts approach to distributed data mining. In: Advances in Soft Computing and Its Applications, pp. 338–349. Springer (2013)
8. Omicini, A., Ricci, A., Viroli, M.: Artifacts in the A&A meta-model for multi-agent systems. Autonomous Agents and Multi-Agent Systems 17(3), 432–456 (2008)
9. Pearl, J.: Probabilistic Reasoning in Intelligence Systems. Morgan Kauffman, San Mateo, CA (1988)
10. Witten, I.H., Frank, E.: Data mining, Practical Machine Learning Tools and Techniques. Morgan Kauffman, San Francisco, CA (2011)