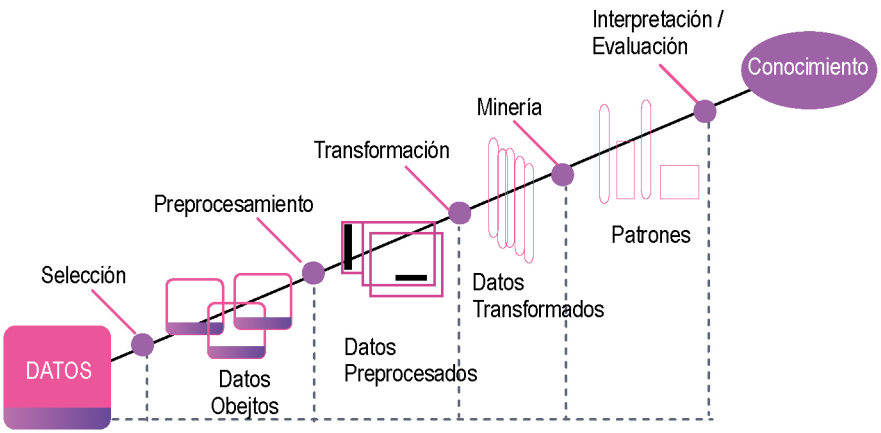
5. Acercamiento metodológico al problema de la cooperación en las IES

De acuerdo a la tesis o posicionamiento que se asumió en la definición del problema —el nivel de no cooperación entre las Instituciones de Educación Superior en México es mayor a su nivel de cooperación—, se establece como medio para obtener conocimiento, la Minería de Datos con la metodología denominada como Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases - KDD*). La pertinencia de emplear KDD es encontrar un modelo válido para explicar el fenómeno de la cooperación en las IES mexicanas. A continuación, se menciona el proceso de KDD de esta investigación:

- Abstracción del escenario, se establece el contexto del problema, así como las limitaciones, reglas y metas a conseguir.
- Selección de datos, se seleccionan las fuentes para obtener datos y seleccionar los datos convenientes a las metas.
- Limpieza y pre-procesamiento, se revisan los datos para garantizar su utilidad removiendo valores atípicos o generando datos faltantes, y se eliminan datos no útiles.
- Transformación de los datos, se transforman los datos para mejorar su calidad por medio de convertir valores numéricos a categóricos (discretización).
- Selección de la apropiada tarea de Minería de Datos, se elige el método para encontrar el modelo de explicación de acuerdo a las metas de la investigación.
- Elección del algoritmo, se aplica el método seleccionado a los datos las veces necesarias hasta obtener el óptimo resultado deseado..
- Evaluación, se revisan los patrones y rendimiento del modelo obtenido vía la técnica de validación cruzada, la cual consiste en partir la base de datos para realizar un entrenamiento y una prueba para observar el funcionamiento del algoritmo.
- Aplicación, se aplica el modelo obtenido y se observa su relación con el problema e hipótesis de investigación.

Imagen 1. Método Knowledge Discovery in Databases (KDD)



Fuente: tomado de http://traduccionesbigdata.blogspot.com/2017/07/el-proceso-kdd.html

5.1 Descubrimiento de conocimiento en bases de datos (KDD)

5.1.1 Abstracción del escenario

Se estudia el fenómeno de la cooperación entre las IES en México, tanto públicas como privadas. Se ha escogido el caso de la cooperación entre investigadores adscritos al Sistema Nacional de Investigadores (SNI) del Consejo Nacional para la Ciencia y Tecnología (Conacyt) de distintas universidades para publicar un artículo científico en revista indizada. Es decir, se consideran aquellas publicaciones en colaboraciones, pero que los autores pertenezcan a distintas IES mexicanas, en los últimos diez años. La meta es encontrar un modelo que nos brinde luz sobre cómo o cuál es el comportamiento de los individuos y/o IES en tanto la cooperación, para estudiar las probables causas y efectos de dicha conducta cooperativa o no cooperativa.

5.1.2 Selección de datos

Las fuentes de información para estructurar el estudio de caso son los datos de acceso libre del SNI del Conacyt y la base de datos del Explorador de Datos del Estudio Comparativo de las Universidades Mexicanas (Execum) de la Universidad Nacional Autónoma de México (UNAM). Los datos recolectados cubren un periodo de ocho años, iniciando en el 2010 hasta el 2017; pues no se cuenta con más datos a utilizar fuera de los periodos mencionados, ni en el Execum ni en el SNI del Conacyt. Ambas fuentes se combinarán para construir una única base de datos.

5.1.3 Limpieza y pre-procesamiento

De las fuentes de información empleadas en esta investigación, se recolectaron indicadores para estructurar la base de datos, cuyo objetivo es crear una perspectiva general de las instituciones públicas y privadas que se encuentran en el país, la cantidad de profesores con los que cuenta cada institución y la relación que dio origen a la correlación entre docentes y nivel de docentes adscritos con categoría SNI; para obtener el nivel de cooperación que existe entre las IES en México. A partir del tratamiento de los datos, se removieron dos IES por considerarse como datos atípicos (IPN y UNAM) ya que sus indicadores son notablemente superiores a las demás IES en México.

5.1.4 Transformación de datos

Basado en la información recabada, se estructuró un análisis de extracción de parámetros para determinar las variables: publicaciones por profesores, profesores, publicaciones, ISI, Scopus, SNI, colaboraciones, tipo

de Universidad; el conjunto de ellas nos posibilita analizar la cooperación. Para determinar los parámetros se realizó el promedio de las variables, es decir, el total de cada una de estas entre el número de resultados alcanzados. Las variables analizadas son determinantes para la presente investigación: denotan factores de orden primario de la cooperación entre las IES investigadas. Se calculó el valor promedio de cada variable, basado en los valores que influyen en el nivel de colaboración que existe entre las IES. De la presente base de datos se discretiza el nivel de cooperación que existe entre las IES mexicanas seleccionadas, marcando los factores de muy alto, alto, medio, bajo, muy bajo.

5.1.5 Selección de la apropiada tarea de minería de datos

La base de datos se analiza en el *software* de Weka a través de un clasificador de redes bayesianas, que funciona a partir de construir una clase y emplea un elemento de decisión. Emplea un algoritmo de refuerzo y, a la par, realiza una regresión basada en el error cuadrático medio y clasifica con base en la entropía. La falta de datos se considera como un valor separado.

5.1.6 Elección de algoritmo de minería de datos

El software que se utiliza es Samlam (*Sensitivity Analysis, Modeling, Inference and More*) con el algoritmo conocido como *Loopy Belief Propagation*.

5.1.7 Aplicación de algoritmo

Se carga en formato de archivo, el resultado de la red bayesiana obtenido en Weka, para procesarlo en Samlam con el algoritmo *Loopy Belief Propagation*.

5.1.8 Evaluación

Se aplicó una validación cruzada a la base de datos; se realiza una partición a la base de datos del 70% para entrenamiento y generar el modelo, posteriormente se prueba con el 30% para observar que trabaje bien de acuerdo al objetivo.

5.1.9 Resultado

Se obtiene nuevo conocimiento como resultado de la minería de datos, y se compara con la teoría para integrarlo al contexto.

5.2 Redes bayesianas, Samlam y NetLogo

Las redes bayesianas son modelos gráficos que representan información a nivel tanto cuantitativo como cualitativo. Estas se conforman por: una entrada que se compone de la base de datos a través de las redes bayesianas, agentes que utilizan la base de datos para realizar inferencias, y una salida, la cual produce datos artificiales.

Las redes bayesianas se componen de elementos cualitativos representados por los nodos, estos son, en realidad, las variables del estudio que se encuentran enlazadas por flechas para determinar la cadena causal o lógica. Por otro lado, los elementos cuantitativos son trascendentes en la construcción de la red bayesiana, pues la construcción de dicha red se basa en una distribución de probabilidad condicionada entre nodos. Cada parte de la red contiene la probabilidad condicional de las variables y, con base en esta característica, es posible conocer las probabilidades de los estados en cada variable.

El modelado del sistema se basa en herramientas del software Weka para considerar la red bayesiana, así mismo se emplean herramientas de Samlam para obtener las inferencias probabilísticas con base en las variables conocidas.

El modelado social se realiza con NetLogo, para reproducir el fenómeno de la cooperación entre las IES con el objetivo de observar las relaciones entre las variables, pero, en concreto, la conducta manifiesta en los agentes para cooperar o no cooperar. El beneficio específico de emplear NetLogo es obtener un modelo para compararlo con otras propuestas o, incluso, en posteriores investigaciones, adaptarlo a las condiciones de la realidad empírica.

Las redes bayesianas como modelo probabilístico para la simulación social. Se utiliza la herramienta denominada Samlam para representar y estudiar las inferencias; también se emplea un modelado basado en agentes conocido como NetLogo, el cual complementa la simulación social requerida para observar el fenómeno de la cooperación en el espacio de la actividad académica de las IES en México.

5.3 NetLogo y protocolo *Overview, Design and Details* (ODD)

La comprobación de la hipótesis de esta investigación se basa en un estudio sobre la cooperación entre IES, tanto privadas como públicas, en México. El estudio se concentra en el caso de la publicación de artículos científicos para la generación del conocimiento, de acuerdo a los estándares del Conacyt, donde un investigador realiza una petición a otro investigador, de una universidad distinta a la suya, para publicar en revistas indizadas; con el fin de mejorar su nivel académico y realizar una publicación para generar conocimiento. El protocolo *Overview, Design and Details* (ODD) es una descripción normalizada sobre modelos para la simulación basada en agentes, la cual ofrece una estructura estándar para la presentación del modelo desarrollado y hace posible compararlo con otros. A continuación, se describen los componentes de un protocolo ODD para cualquier modelo, así como una breve descripción de ellos:

1. Propósito, es o son los objetivos que brindan sentido al sistema.
2. Entidades, estados y escalas, se mencionan las entidades del sistema,

las variables de estado en sus características y representación; la temporalidad y espacio del modelo.

3. Visión general y planificación de procesos, descripción de la actividad de las entidades, sus tiempos y circunstancias.

4. Conceptos de diseño, son de utilidad para comprender los resultados y muestran las decisiones sobre el diseño del modelo.

5. Inicialización, se especifican las condiciones al inicio de arranque del modelo.

6. Datos de entrada, se menciona la base de datos utilizada por el modelo y cuándo se utiliza.

7. Submodelos, son los parámetros de los modelos que subyacen al sistema y los tiempos cuando son utilizados.

A continuación, se presenta el protocolo ODD para la programación de NetLogo propuesto en este trabajo:

1. Propósito.

El modelo tiene el propósito de estudiar las siguientes variables:

1. Publicaciones por profesor; miembros del sistema nacional de investigadores (SNI); 2. Publicaciones en el Instituto para la Información Científica (ISI); 3. Volaboraciones registradas en el Instituto para la Información Científica (ISI); 4. Publicaciones de los investigadores; 5. Publicaciones en la base de datos bibliográfica Scopus bajo las cuales se da la cooperación entre las universidades para generar conocimiento, entendiendo por cooperación la colaboración entre investigadores de distintas universidades para publicar artículos científicos en revistas indizadas.

2. Entidades, variables de estado y escalas

Las entidades son:

Individuales: investigador, y sus atributos son publicaciones (variable dinámica), universidad a la que pertenece (variable estática), nivel de SNI (dinámica).

Colectivas: universidades formadas por investigadore. Sus propiedades (variables) son: cantidad de profesores (variable aleatoria); publicaciones por

SNI (variable dinámica); colaboraciones (variable dinámica); proporción de SNI (variable dinámica) y publicaciones en Scopus (variable dinámica), proporción de cooperación (variable aleatoria).

La medición del nivel de cooperación entre universidades se realiza de manera cualitativa en las dimensiones de: muy bajo, bajo, medio, alto y muy alto.

Agentes:

1. Investigador
2. Universidades

Ambiente:

La probabilidad de aceptación para publicar en colaboración con otro investigador se inicia con la petición por parte de un agente a otro, con la opción de aceptación o rechazo.

No existe un ambiente específico, sin embargo, las universidades y los investigadores interactúan en un espacio donde se realiza un número determinado de invitaciones por parte de los investigadores, para colaborar con la meta de publicar en una entidad editorial que aceptará o rechazará los artículos. En el caso de aceptación de un artículo, por parte de la entidad editorial, este resultado actualizará las variables de los agentes y esto afectará a los agentes colectivos para observar cómo impacta en la cooperación entre universidades.

3. Proceso en general y programación

Un tic representa un año donde se actualizan las variables del agente colectivo.

Existen dos procesos en la simulación: el primero se refiere a la publicación de artículos donde el responsable es un investigador, el cual colabora o no, con el resultado de publicar un artículo o no, y con el impacto de elevar su nivel como académico; el segundo, donde se realiza una petición para colaborar, donde el investigador invita a otro investigador a escribir un artículo, el cual, eventualmente, se publica o no.

En resumen, los agentes individuales realizan peticiones a otros agentes individuales para colaborar con la meta de publicar un artículo científico. En ese momento de la toma de decisión sobre si colaborar o no, el agente

investigador consulta a su universidad para preguntar si colabora o no, en función de si el agente colectivo es cooperativo o no, será la respuesta para el investigador. En estos casos solo tendremos dos opciones: la universidad coopera; la universidad no coopera. Una vez que se obtiene respuesta por parte de los agentes, se actualizan sus variables.

A continuación, se explica la petición de los agentes individuales a los agentes colectivos sobre si colaborar o no:

Tabla 2. Relación entre agentes del modelo de simulación social en la cooperación entre IES mexicanas.

	Agente 1	Agente 2	Respuesta
Tipo de cooperación de la universidad	Coopera	Coopera	Coopera
Tipo de cooperación de la universidad	No coopera	No coopera	No coopera

Fuente: elaboración propia

Cuando se da la interacción entre una universidad que coopera y otra que no coopera, la decisión se realiza de forma aleatoria bajo el siguiente proceso: a partir de los datos obtenidos para elaborar el modelo de cooperación, se observa que las universidades etiquetadas como no cooperativas tienen un porcentaje bajo de cooperación, entonces, se decidió incluir una variable aleatoria que determine si coopera o no la universidad, esta variable aleatoria está en función de una proporción de cooperación definida inicialmente para cada universidad.

El etiquetado para este modelo de cooperación, en sus agentes colectivos, se realizó con base en un árbol de decisión construido a partir de la base de datos conformada por las fuentes de información de Conacyt y Execum de la UNAM. Esta base de datos contiene diez años de información sobre la actividad académica en el sector universitario de México, con las variables antes mencionadas en el apartado de variables de estado.

4. Conceptos de diseño

Este modelo gira alrededor del concepto de cooperación, el cual se define como la colaboración entre investigadores de distintas universidades para publicar artículos científicos en revistas indizadas; este concepto y las acciones que conlleva en los agentes, nos lleva a pensar en la emergencia, con base en las decisiones sobre si cooperar para publicar, o no cooperar y, al mismo tiempo, la adaptación de la conducta por parte de los investigadores a publicar de acuerdo a la universidad a la cual pertenecen.

La conducta de adaptación por parte de los investigadores se modela con base en la universidad a la cual pertenecen. Mientras que la conducta de las universidades se adapta cada vez que recolecta información de los agentes individuales, y esto lo hace en cada tic del programa.

Lo estocástico se usa para representar la decisión y elección del investigador para alcanzar la meta de publicar, para lo cual elige con cuál universidad y con qué investigador hacer la petición para publicar en colaboración.

5. Inicialización

El etiquetado para este modelo de cooperación en sus agentes colectivos se realizó con base en un árbol de decisión construido a partir de la base de datos construida con las fuentes de información de Conacyt y Execum de la UNAM.

Esta base de datos contiene diez años de información sobre la actividad académica en el sector universitario de México, con las variables antes mencionadas en el apartado de variables de estado.

6. Datos de entrada.

En este modelo no se asume un medio ambiente, por tal motivo dicho modelo no tiene datos de entrada.

7. Submodelos

El primer submodelo se llama red bayesiana con el algoritmo HillClimber y un estimador de 0.5.

El segundo submodelo es J48, es un árbol de decisión que regresa una serie de reglas para clasificar los datos.

El tercer submodelo es Samlam, para el análisis, modelado e inferencia en la representación de las relaciones y los pesos porcentuales de las variables independientes en la variable dependiente (cooperación).