Selección de beneficiarios de apoyo en sector gubernamental basada en técnicas bayesianas. Caso de estudio: Comisión Nacional Forestal

RESUMEN: La Comisión Nacional Forestal (CONAFOR), como un organismo público descentralizado de la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) tiene como objetivo desarrollar, favorecer e impulsar las actividades productivas, de conservación y restauración en materia forestal, así como participar en la formulación de los planes, programas y en la aplicación de la política de desarrollo forestal sustentable bajo las Reglas de Operación del Programa de Apoyo para el Desarrollo Forestal Sustentable. Este artículo propone usar una metodología para modelar la selección de candidatos para recibir apoyo que ofrece la CONAFOR mediante un método bayesiano. El reconocimiento de diferencias en la conectividad de variables puede ser usada para clasificar patrones (a) normales. Este método bayesiano se usa para clasificar, analizar y evaluar gráficamente la existencia de (in)dependencia en la distribución espacial, obteniendo y comprobando una herramienta de apoyo a la toma de decisiones para facilitar la selección de los mejores candidatos para recibir el apoyo otorgado por la CONAFOR. Como resultado se obtiene la identificación de los criterios que se relacionan para ser catalogado como factible en el proceso de dictaminación, con una probabilidad de 72.3%. Perteneciendo a un municipio con nula y alta marginación y teniendo un proyecto técnicamente factible.

PALABRAS CLAVE: Árbol de decisión, Bayes, CONAFOR, Metodología, Modelo de selección, Toma de decisión.

Colaboración
Luis Armando Rodríguez Aguilar; Gabriel Grosskelwing Núñez; Roberto Ángel Meléndez Armenta; Jorge Cruz Salazar, Instituto Tecnológico Superior de Misantla

ABSTRACT: The Comisión Nacional Forestal (CONAFOR), as a decentralized public body of the Secretaría of Medio Ambiente and Recursos Naturales (SEMARNAT) aims to develop, promote and promote productive activities, conservation and restoration in forestry, as well as participate in the formulation of plans, programs and in the application of sustainable forest development policy under the Operating Rules of the Support Program for Sustainable Forest Development. This article proposes to use a methodology to model the selection of candidates to receive support offered by CONAFOR through a Bayesian method. The recognition of differences in the connectivity of variables can be used to classify normal (a) patterns. This Bayesian method is used to classify, analyze and graphically evaluate the existence of (in) dependence on spatial distribution, obtaining and verifying a decision support tool to facilitate the selection of the best candidates to receive the support granted by CONAFOR. As a result, the identification of the criteria that are related to be cataloged as feasible in the ruling process is obtained, with a probability of 72.3%. Belonging to a municipality with zero and high marginalization and having a technically feasible project.

KEYWORDS: Decision tree, Bayes, CONAFOR, Selection model, Decision making.

INTRODUCCIÓN

En la actualidad el planeta se encuentra inmerso en una gran cantidad de problemas ambientales, ante este contratiempo, México se encuentra en un proceso de construcción de estrategias y planes de manejo, la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) [1] se encarga de impulsar la protección, conservación y aprovechamiento sustentable de los ecosistemas y biodiversidad. Con ayuda de la Comisión Nacional Forestal (CONAFOR) [2] apoya a los dueños y poseedores de bosques, selvas, manglares, humedales y zonas áridas bajo las Reglas de Operación del

Programa de Apoyo para el Desarrollo Forestal Sustentable [3].

Sin embargo, detectar posibles beneficiarios de cientos de solicitantes es una tarea difícil, es en estas reglas de operación en donde CONAFOR ha detectado que un subproceso (la dictaminación), el análisis y selección de los candidatos lo realiza un número pequeño del personal con el que cuenta y los resultados se requieren en un periodo de tiempo muy corto.

Para una maximización de beneficios y que el Programa de Apoyo sea exitoso, es crucial identificar correctamente a los posibles beneficiarios, en este sentido, se propone un método bayesiano (red bayesiana), el cual fundamenta su funcionamiento en la teoría de probabilidad para estimar la posibilidad de que a un solicitante sea beneficiario por la comisión. Estos puntajes son necesarios para calcular la medida útil en un entorno de predicción para comprender por qué los solicitantes son o no catalogados como factibles y elaborar las estrategias correspondientes para ampliar los beneficios. Las reglas divididas de la red bayesiana se optimizan de acuerdo con la métrica de elección para el uso de un método de optimización de partición recursiva de búsqueda y es utilizado para dar respuesta a planteamientos que impliquen decisión multicriterio.

Este artículo ha sido organizado como sigue: un primer apartado en el cual se presenta un desarrollo teórico sobre los problemas de decisión y Teorema de Bayes. Seguido, se presentan los materiales y la metodología propuesta para resolver el problema de toma de decisión y finalmente unas conclusiones.

Hipótesis. Es posible modelar la selección de candidatos para recibir apoyo que ofrece la CONAFOR mediante un método bayesiano.

El presente artículo tiene como objetivo proponer un modelo para la selección de candidatos que reciban apoyo con los programas que ofrece la CONAFOR basado en un método bayesiano, el cual permite elegir entre varias alternativas, la que mejor responde a los múltiples criterios definidos para ello.

Teoría de la decisión.

González F.A. aclara que la teoría de la decisión se ocupa de analizar cómo elige una persona aquella acción que, de entre un conjunto de acciones posibles, le conduce al mejor resultado dadas sus preferencias [4].

La decisión puede ser paramétrica: si el contexto se considera dado, es decir, un parámetro o estrategia: si las decisiones de los actores son interdependientes. De forma que nuestra decisión dependa de lo que hagan los demás [5]. A este cuadro habría que añadirle la cantidad de información con que cuenta el individuo para decidirse por una opción u otra de su conjunto factible.

Si la información sobre los resultados de las distintas opciones es completa conocemos con toda seguridad las consecuencias de nuestras decisiones el decisor se hallará ante una situación de certidumbre; si, por el contrario, la información es incompleta desconocemos qué consecuencias tendrán nuestras acciones, la situación será de riesgo o bien de incertidumbre [6], el siguiente árbol recoge de manera resumida el panorama de la teoría de la decisión.

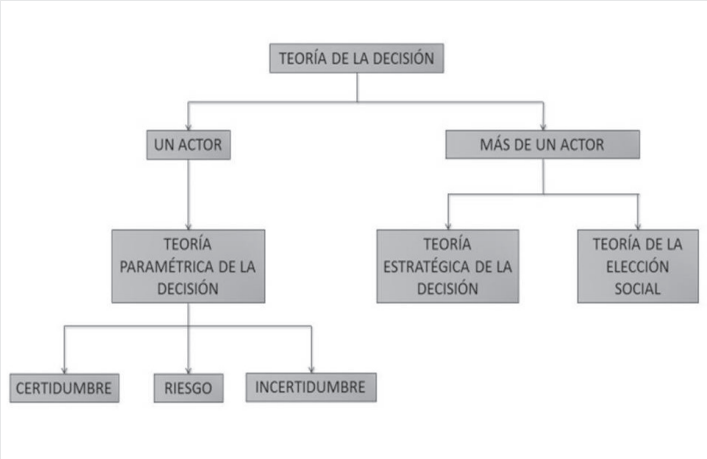

Figura 1. Teoría de la decisión. Tomado de [6].

Teorema de Bayes.

R. Aznar indica que Thomas Bayes estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados la influencia de las primeras investigaciones en que trataron el razonamiento condicional [7].

T. Bayes plantea de manera explícita que "Dado el número de veces que un suceso ha ocurrido y el de veces que no ha ocurrido, se requiere calcular la probabilidad de su ocurrencia en un solo experimento esté entre cualesquiera de los valores prefijados" (regla de Bayes) [8]. La regla de Bayes es una de las normas más importantes de la teoría de la probabilidad, ya que es el fundamento de la inferencia bayesiana. La idea principal de la metodología bayesiana proviene de la regla de Bayes y en los conceptos bayesianos los parámetros se consideran variables aleatorias [9].

El interés por el teorema de Bayes trasciende a la aplicación clásica, especialmente cuando se amplía a otro contexto en el que la probabilidad no se entiende exclusivamente como la frecuencia relativa de un suceso a largo plazo, sino como el grado de convicción personal acerca de que el suceso ocurra o pueda ocurrir.

Probabilidad a priori: Es la probabilidad incondicional asociada con la medición del grado de conocimiento inicial que se tiene de los parámetros en estudio. Si bien su influencia disminuye a medida que más información muestral es disponible, el uso de una u otra distribución a priori determinará ciertas diferencias en la distribu-

ción a posteriori [10]. Una vez que el decisor obtiene alguna evidencia referente a las variables aleatorias desconocidas que constituyen el dominio, las probabilidades priori ya no son aplicables.

Probabilidad a posteriori: Es la probabilidad condicional, utilizada en el criterio de valor esperado y suele estimarse a partir de datos históricos, puede mejorarse con experimentación adicional [11].

Árbol de decisión: Los árboles de decisión representan decisiones anidadas que sirven para clasificar los datos. Cuando se utiliza un árbol de decisión sobre los datos, se obtienen reglas que permiten clasificarlos. Un árbol se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver [12].

Red Bayesiana: La red bayesiana [13] [14], es una clase de modelo gráfico que permite una representación concisa a través de una distribución de probabilidad condicional entre un conjunto de atributos en un grafo dirigido acíclico. La dependencia entre dos atributos es descrita por la presencia de un arco entre ellos, y su influencia causal, por la dirección del arco. La independencia entre atributos se representa por la ausencia de un arco que conecte atributos particulares.

MATERIAL Y MÉTODOS

"Package party"

Una paquetería llamada "A Laboratory for Recursive Partytioning": Creada por Torsten Hothorn (versión 1.3-3), describe que es una caja de herramientas computacional para particiones recursivas. El núcleo del paquete es ctree (), una implementación de árboles de inferencia condicional que integran modelos de regresión estructurados en árbol en una teoría bien definida de procedimientos de inferencia condicional. Esta clase no paramétrica de árboles de regresión es aplicable a todo tipo de problemas de regresión, incluidas las variables de respuesta nominales, ordinales, numéricas, censuradas y multivariadas y escalas de medición arbitrarias de las covariables. Basado en árboles de inferencia condicional, cforest () proporciona una implementación de los bosques aleatorios de Breiman [15]. La función mob () implementa un algoritmo para particiones recursivas basado en modelos paramétricos, empleando pruebas de inestabilidad de parámetros para la selección dividida. La funcionalidad extensible para visualizar modelos de regresión estructurados en árbol está disponible [16].

Metodología para la base de datos.

Etapa 1 Recopilación.

Se cuenta con una base de datos relacional proporcionada por la CONAFOR (dictaminación del año 2018), donde se encuentran 697 solicitantes que fueron evaluados bajo los criterios de prelación establecidos en las Reglas de Operación del 2018 (atributos) [17] en un formato de una hoja de cálculo, cada caso contiene registro de los datos particulares de cada solicitante y es evaluado bajo 60 y 58 atributos numéricos y categóricos respectivamente.

Etapa 2 Limpieza y transformación.

Para efecto de confidencialidad de los datos se modificaron los nombres de los solicitantes y los folios de control, así como conversión de algunos atributos categóricos a numéricos, obteniendo un resumen de las características de los atributos.

Etapa 3 Construcción

Se evalúa la base de datos extrayéndola al programa RStudio (R versión 3.6, 2019-07-05, Copyright (C) 2019 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64, 64-bit) para ser corrida en la paquetería "Party" para comprobar que el modelo aplicado es funcional.

Etapa 4 Evaluación.

Una vez obtenida la red bayesiana se evalúa al modelo mediante una validación simple la técnica "bootstrapping" [11][18]: 1) Se divide el conjunto de datos (694 solicitantes) de manera aleatoria obteniendo dos disjuntos (conjunto de entrenamiento: 482 solicitantes y conjunto de test:212 solicitantes) en estos dos conjuntos construidos pueden contener datos repetidos. 2) Se aplica el algoritmo al conjunto de entrenamiento, 3) Se utiliza la matriz de confusión [19] [20] para mostrar el recuento de casos de las clases predichas y sus valores actuales para conocer mejor el tipo de error de nuestro modelo (Conocer la precisión del modelo).

Etapa 5 Interpretación.

Análisis e interpretación del grafo resultante para la descripción de su comportamiento (ver Figura 2)

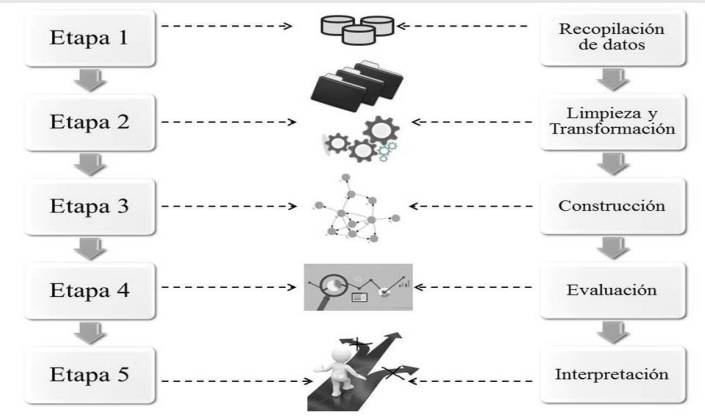

Figura 2. Metodología.

RESULTADOS

Como se aprecia en la Figura 3, se obtiene una red bayesiana con el siguiente contenido:

• Los nodos que definen el perfil de los solicitantes factibles son: Nodo 1-Nodo 5-Nodo 6, es decir, existe una mayor probabilidad de ser catalogado como factible si el solicitante cuenta con un proyecto Técnicamente Si Factible y perteneciendo a un Municipio con nula y alta marginación.

• Nodo 1: Es el factor (Y)= que describe a la variable dependiente que posteriormente ramifica en dos nodos: al Nodo2 (donde 3=Muy alta) y nodo 5 (donde 1=No y 2=Alta), tomando el criterio perteneciente a un municipio con un grado de marginación, indicando que ésta es la variable principal predictora

• Nodo 5: Lo precede el Nodo 1. Corresponde a la variable dependiente nombrada como Técnicamente no factible, ramificando en dos nodos: al Nodo 6 donde ("≤0"=No) y al Nodo 7 (donde y ">0"=Si)

• Nodo 6: Indica que de 505 solicitantes que caen en esta rama, existe una probabilidad de 27.7% y 72.3% de ser catalogado como No Factible y Si Factible respectivamente.

Se obtiene también la matriz de confusión (Tabla 1) que indica lo siguiente:

• NV es la cantidad de NO que fueron clasificados correctamente como NO.

• NF es la cantidad de NO que fueron clasificados incorrectamente como SI.

• SV es la cantidad de SI que fueron clasificados correctamente como SI.

• SF es la cantidad de SI que fueron clasificados incorrectamente como NO.

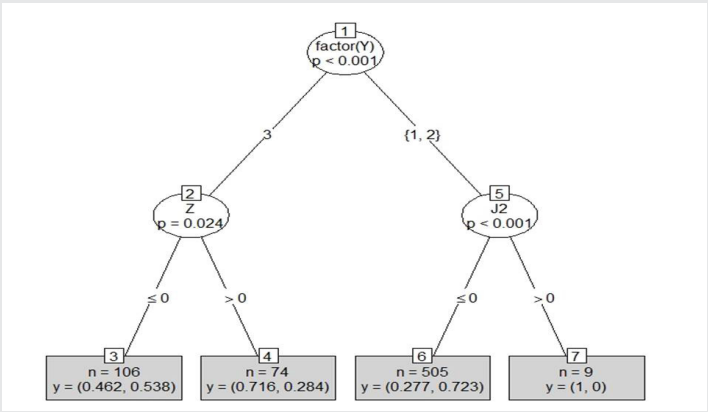

Figura 3. Red bayesiana.

Tabla 1. Matriz de confusión

Predicción/Real	Valor Real		
Valor Predicho	Predicción/Real	NO	SI
	NO	18 (NV)	8(NF)
	SI	66(SF)	120(SV)

Demostrando que cuenta con (ver Tabla 2):

• Sensibilidad: también se la llama recall o tasa de verdaderos positivos. Nos da la probabilidad de que, dada una observación realmente positiva.

• Especificidad: también llamado ratio de verdaderos negativos. Nos da la probabilidad de que, dada una observación realmente negativa.

• Precisión: también llamado valor de predicción positiva. Nos da la probabilidad de que, dada una predicción positiva, la realidad sea positiva también.

• Valor de predicción Negativa: Nos da la probabilidad de que, dada una predicción negativa, la realidad sea también negativa.

• Error de clasificación: Porcentaje de errores del modelo.

• Exactitud: Porcentaje total de los aciertos de nuestro modelo.

Prevalencia: La probabilidad de un positivo en el total de la muestra.

Tabla 2. Porcentaje de métricas

Métricas	Porcentaje
Sensibilidad	93.75%
Especificidad	21.43%
Precisión	64.52%
Valor de predicción negativa	69.23%
Error de clasificación	34.91%
Exactitud	65.09%
Prevalencia	60.38%

CONCLUSIONES

En este artículo presentamos un método bayesiano que describe de manera gráfica el proceso de dictaminación que lleva a cabo la CONAFOR, en él se demuestra la relación entre algunas de las variables (Y y J2), arrojando una mayor probabilidad de ser catalogado como factible perteneciendo a un Municipio con nula y alta marginación, teniendo un proyecto Técnicamente factible.

La precisión del modelo no garantiza que refleje la situación del problema actual, no obstante se debe contrastar el conocimiento que éste proporciona con el conocimiento previo que el experto decisor pudiera tener sobre el caso en particular y resolver los posibles conflictos.

Para trabajos futuros se recomienda usar la técnica validación cruzada para evaluar los resultados del análisis estadístico y garantizar (in)dependencia de la partición entre datos de entrenamiento y prueba.

El concepto de KDD se ha desarrollado, y continúa desarrollándose, desde la intersección de la investigación de áreas tales como bases de datos, aprendizaje automático, reconocimiento de patrones, estadística, teoría de la información, inteligencia artificial, razonamiento con incertidumbre, visualización de datos y Soft Computing, por estos motivos se recomienda para trabajos futuros incluir la minería de datos..

AGRADECIMIENTOS
La posibilidad de realizar el trabajo de investigación, ha sido gracias al apoyo económico y moral de muchas instituciones y personas, por ello mi agradecimiento:
Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada 719468 para realizar mis estudios de postgrado.

Agradezco a todos los investigadores, técnicos y personal administrativo y de apoyo del Instituto Tecnológico Superior de Misantla (ITSM) y al Tecnológico Nacional de México (TecNM); asimismo a todos mis compañeros de generación, por su compañía y apoyo.

De igual manera, agradezco el apoyo de la Comisión Nacional Forestal (CONAFOR) del estado de Veracruz, en especial a la oficina de Reforestación por su valiosa colaboración y aportación de información clave para este proyecto.

BIBLIOGRAFÍA

[1] SEMARNAT, (2017). Misión y Visión de la SEMARNAT, Obtenida el 12 de Noviembre del 2018, de la página electrónica: https://www.gob.mx/semarnat/acciones-y-programas/mision-y-vision-de-la-semarnat

[2] Estados Unidos Mexicanos. Cámara de Diputados del H. Congreso de la Unión. (2017). Ley orgánica de la administración pública federal. México

[3] Estados Unidos Mexicanos. Cámara de Diputados del H. Congreso de la Unión (2019). Quinta-Sexta Sección, Comisión Nacional Forestal. México.

[4] González, F. A. (2004). Teoría de la decisión e incertidumbre: modelos normativos y descriptivos. Empiria. Revista de metodología de ciencias sociales, (8), 139-160.

[5] Elster, J. (Ed.). (1986). Rational choice. NYU Press.

[6] Rapoport, A. (1983). Mathematical models in the social and behavioral sciences. John Wiley & Sons.

[7] Enrique R. Aznar, (2007). Biografías-Thomas Bayes, Obtenida el 24 de Octubre de 2018, de la página electrónica: https://www.ugr.es/□eaznar/bayes.htm

[8] Paredes-Cancino, C., & Cantoral, R. (2018). La noción de proporcionalidad en la construcción del teorema de Bayes. El caso del pensamiento estocástico.

[9] Walpole, R. E., Myers, R. H., & Myers, S. L. (1999). Probabilidad y estadística para ingenieros. Pearson Educación.

[10] Octavio Paredes Pérez (2013). Regresión Lineal por Medio del Análisis Bayesiano. Tesis Benemérita Universidad Autónoma de Puebla-Facultad de Ciencias Físico Matemáticas. México

[11] Russell, S. J., & Norvig, P. (2004). Inteligencia Artificial: un enfoque moderno (No. 04; Q335, R8y 2004.).

[12] Mestizo Gutiérrez, S. L. (2015). Árboles de decisión y redes bayesianas para el análisis de genes involucrados en la enfermedad de Alzheimer.

[13] Sahami, M. (1996). Aprendizaje clasificadores bayesianos de dependencia limitada. En KDD (Vol. 96, No. 1, pp. 335-338).

[14] Portugal, R., & Carrasco, M. (2007). Ensamble de Algoritmos Bayesianos con Árboles de decisión: una alternativa de clasificación. In XVII Congreso Chileno de Control Automático ACCA, Universidad de la Frontera, Chile.

[15] Breiman, L. (2001). Bosques al azar. Aprendizaje automático, 45 (1), 5-32.

[16] Torsten Hothorn, Kurt Hornik y Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15 (3), 651-674.

[17] Estados Unidos Mexicanos. Cámara de Diputados del H. Congreso de la Unión (2019). Décima-Décimoprimera Sección, Comisión Nacional Forestal. México.

[18] Hernández, O. José., Ramírez, Q. M., & Ferri, Ramírez. C. (2004). Introducción a la minería de datos. 3er. edición. Pearson, Prentice Hall. México.

[19] Zelada, C. (2017). Evaluación de modelos de clasificación, Obtenida el 25 de Enero del 2019, de la página https://rpubs.com/chzelada/275494.

[20] Pina, K. (2018). Matriz de confusión, Obtenida el 25 de Enero 2019 de la página https://koldopina.com/matriz-de-confusion/