

42

Fecha de presentación: octubre, 2018

Fecha de aceptación: diciembre, 2018

Fecha de publicación: febrero, 2019

NUEVO CLASIFICADOR BAYESIANO

SIMPLE PARA EL ANÁLISIS DE DATOS EDUCATIVOS

NEW BAYESIAN SIMPLE CLASSIFIER FOR EDUCATIONAL DATA ANALYSIS

Byron Oviedo Bayas¹

E-mail: boviedo@uteq.edu.ec

ORCID: <https://orcid.org/0000-0002-5366-5917>

Cristian Zambrano-Vega¹

E-mail: czambrano@uteq.edu.ec

¹ Universidad Estatal de Quevedo. Ecuador.

Cita sugerida (APA, sexta edición)

Oviedo Bayas, B., & Zambrano-Vega, C. (2019). Nuevo clasificador bayesiano simple para el análisis de datos educativos. *Universidad y Sociedad*, 11(2), 278-285. Recuperado de <http://rus.ucf.edu.cu/index.php/rus>

RESUMEN

En este artículo, se propone el uso de un nuevo clasificador bayesiano simple (SBND) que aprende rápidamente una frontera de Markov de la variable clase y una estructura de red que relaciona las variables de la clase y dicha frontera. Este modelo se compara con otros clasificadores bayesianos, para luego hacer uso de los modelos gráficos probabilísticos en el campo de la enseñanza para poder determinar el problema de deserción estudiantil en las universidades. Se han utilizado datos socioeconómicos de los estudiantes legalmente matriculados en la Facultad de Ciencias de la Ingeniería de la Universidad Técnica Estatal de Quevedo en Ecuador durante el periodo 2012-2013. Con esta base de datos se compararán los resultados que obtienen los algoritmos Naive Bayes, TAN, BAN, SBND y las combinaciones con diferentes métricas como K2, BIC, Akaike, BDEu. También se comparan los métodos RPDAG y C-RPDAG. El trabajo experimental fue realizado con la herramienta Weka que es de acceso libre y gratuito.

Palabras clave: Redes bayesianas, clasificador bayesiano, deserción estudiantil.

ABSTRACT

In this article, we propose the use of a new simple Bayesian classifier (SBND) that quickly learns a Markov boundary of the class variable and a red structure that relates the variables of the class and the boundary. This model is compared with other Bayesian classifiers, to then make use of probabilistic graphical models in the field of education in order to determine the problem of student desertion in universities. Socio-economic data of students legally enrolled in the Faculty of Engineering Sciences of the Technical State University of Quevedo in Ecuador during the 2012-2013 period have been used. This database compares the results obtained with the Naive Bayes, TAN, BAN, SBND and combinations with different metrics such as K2, BIC, Akaike, BDEu. The RPDAG and C-RPDAG methods are also compared. The experimental work was carried out with the Weka tool, which is free and has open access.

Keywords: Bayesian networks, Bayesian classifier, student desertion.

INTRODUCCIÓN

El problema de la deserción estudiantil viene siendo estudiado desde tiempo atrás por algunos investigadores. Magaña, Montesino & Hernández (2006), lo analizan agrupando a individuos u objetos de acuerdo a sus semejanzas, maximizando la similitud de los objetos dentro de los grupos a la vez que maximiza la heterogeneidad entre agregados. También se han utilizado técnicas de minería de datos para predecir la probabilidad de que un estudiante abandone la institución educativa; tal es el caso de Kuna, García Martínez & Villatoro (2010), quienes hicieron uso del conocimiento en reglas de asociación y en el enfoque TDIDT (Top Down Induction of Decision Trees) sobre la base de datos de la gestión académica del consorcio SIU de Argentina (que reúne 33 universidades de Argentina), permitiendo realizar un análisis para encontrar las reglas de comportamiento.

Lykourentzou, Giannoukos, Nikolopoulos, Mpardis & Loumos (2009), usan un método de predicción de deserción en los cursos de e-learning, basado en tres técnicas populares de aprendizaje automático: redes neuronales feedforward, máquinas de soporte vectorial y métodos de ARTMAP difuso simplificado. Por otro lado, Dekker, Pechenizkiy & Vleeshouwers (2009), haciendo uso de árboles de clasificación, naive Bayes, regresión logística y bosques de árboles comparan distintos modelos para predecir las tasas de abandono durante el primer semestre de los estudios de grado en la Universidad de Eindhoven obteniendo unas tasas de acierto entre el 75% y el 80%.

También se puede referenciar el trabajo de Porcel, Dapozo & López (2010), en el que se analiza la relación del rendimiento académico de los alumnos que ingresan a la Facultad de Ciencias Exactas y Naturales y Agrimensura de la Universidad Nacional del Nordeste (FACENA-UNNE) en Corrientes, Argentina, durante el primer año de carrera con las características socio-educativas de los mismos. Se ajustó un modelo de regresión logística binaria, el cual clasificó adecuadamente el 75% de los datos.

Kuna, et al. (2010), proponen un modelo basado en redes bayesianas para determinar el estilo de aprendizaje de cada estudiante tomando como base la información proporcionada por expertos (docentes), construyendo así la red con un alto grado de precisión.

Hay un trabajo enfocado al estudio de medidas de rendimiento colectivo, como es el caso del de Morales & Salmerón (2003), en el que se propone una metodología para el análisis de relevancia de indicadores de rendimiento basada en el uso de redes bayesianas.

Los modelos gráficos permiten obtener, de forma sencilla, las principales relaciones entre las variables a considerar en este tipo de estudios. La metodología propuesta se aplica a un caso práctico, mostrando que esta es una herramienta útil para ayudar en la toma de decisiones en la elaboración de políticas basadas en indicadores de rendimiento.

Esta tarea requiere el manejo de un alto número de variables de distintas naturalezas (cualitativas y cuantitativas), que pueden tener una compleja estructura de dependencias.

DESARROLLO

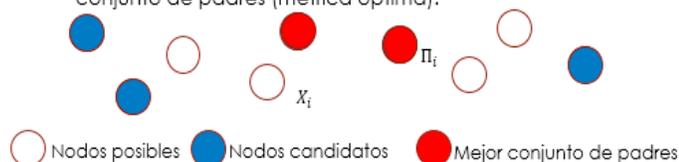
En este estudio empezamos introduciendo un nuevo clasificador que llamaremos *clasificador bayesiano simple SBND*, que será una red bayesiana genérica, pero aprendida con una técnica voraz. Luego se realiza el análisis de deserción de los estudiantes legalmente matriculados en la Facultad de Ciencias de la Ingeniería de la Universidad Técnica Estatal de Quevedo en el periodo 2012-2013 en base a los datos socio económicos (Oviedo, Moral & Puris., 2016). Posteriormente para realizar la clasificación se utiliza la herramienta Weka, usando clasificadores bayesianos, en este caso se utiliza Naive Bayes, BayesNet con K2 y un solo padre, BayesNet con K2 y un máximo de 5 padres, BayesNet con TAN, BayesNet con Hill Climber y un solo padre, BayesNet con Hill Climber y un máximo de 5 padres. Luego se usa los clasificadores de árboles, llegando a comparar los resultados obtenidos por los árboles J48 y Random Forest, a continuación, se analizan los datos con clasificadores que hacen uso de reglas de clasificación, para finalmente con todos estos resultados realizar un experimento en Weka.

Procedimiento inicial:

Para este clasificador antes que nada vamos a necesitar una función PADRES que dado una variable X_i y un conjunto de candidatos calcula el mejor conjunto de padres de X_i de entre ese conjunto de candidatos. El conjunto de padres se devuelve en π_i y al mismo tiempo devuelve un valor numérico que es el Score de esa variable dado ese conjunto de padres medido por un Score bayesiano.

La función PADRES es la que realiza una búsqueda heurística del mejor conjunto de padres dentro de un conjunto de candidatos y esto se da añadiendo y quitando padres mientras allá mejoras en el score (Oviedo, et al., 2016).

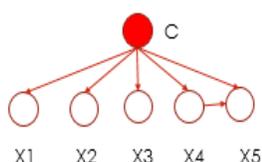
- Problema de clasificación supervisada.
 - Clase C y atributos $X = (X_1, \dots, X_n)$; \mathcal{D} =conjunto de observac.
 - Objetivo: Aprender una red bayesiana para $X' \cup \{C\}$, donde $X' \subseteq X$.
 - Idea: Introducir a C como nodo raíz en la RB \mathcal{R} .
 - Métrica a usar $Score(X_i, A|\mathcal{D})$
 - Procedimiento $PADRES(X_i, CANDIDATOS, \Pi_i)$ calcula mejor conjunto de padres de Π_i de X_i con la métrica seleccionada y devuelve un valor numérico igual al Score de X_i con su mejor conjunto de padres (métrica óptima).



Procedimiento de construcción del clasificador bayesiano

Este clasificador empezará con la variable clase y va añadiendo de forma interactiva distintos nodos al clasificador. Vamos a suponer que hemos añadido X' nodos en general, en ese momento es la clase (rojo) y se han añadido 5 nodos, de manera general nuestra red bayesiana va a ser la variable clase como padre de todos los nodos y una estructura genérica de red bayesiana y los nodos.

Para añadir un nodo para cada nodo candidato que no haya sido añadido previamente se calculan 2 valores. Primero uno con el mejor score de X_i cuando su padre se elige de entre todos los nodos previamente añadidos incluyendo la variable clase menos el score de X_i cuando los candidatos son los nodos previamente añadidos, pero sin incluir la variable clase. Este valor se supone que es de grado de dependencia condicional entre X_i y la clase dado los nodos previamente añadidos. Esto se considera como una medida de información que la variable X_i da sobre la clase condicionado a la variable ya introducida.



- Se calcula para cada variable $X_i \in X \setminus X'$ el valor:

$$Infor(X_i, C) = PADRES(X_i, X' \cup \{C\}, \Pi_i) - PADRES(X_i, X', \Pi_i)$$

$Infor(X_i, C) \geq 0$ aunque podría ser (-) ya que el mejor conjunto de padres se calcula aproximadamente

La variable que mayor información entregue siempre que esta sea positiva se añade a la red y se considera como su conjunto de padres, el mejor conjunto de padres que

nos ha devuelto esta función. Como la información es positiva hay que suponer que la variable clase está incluida en el conjunto de padres. El algoritmo termina si la función es ≤ 0 .

- Elegir la variable:

$$X_{max} = argmax_{X_i \in X \setminus X'} Infor(X_i, C)$$

- Si $Infor(X_{max}, C) > 0$ se inserta en la red y en X' . Su conjunto de padres es Π_i calculado con $PADRES(X_{max}, X' \cup \{C\}, \Pi_i)$.
- El proceso continúa de forma iterativa tratando de añadir una nueva variable y termina cuando en un paso $Infor(X_{max}, C) \leq 0$

Características para construir el clasificador

- Aprender una red bayesiana arbitraria con un subconjunto de las variables iniciales que influyen directamente en esta variable por lo tanto de manera implícita realiza una selección de variables.
- En la forma que hemos construido, la variable clase es siempre un nodo raíz.
- El orden de los atributos se basa en elegir de forma voraz primero los atributos que más informan sobre la clase, dados los atributos ya seleccionados.

Tradicionalmente se ha considerado que los algoritmos para aprendizaje sin restricciones de redes bayesianas, especialmente los basados en el paradigma de métrica+búsqueda, no son adecuados para la construcción competitiva de clasificadores basados en redes bayesianas (Acid & De Campos, 2005). Actualmente esto a cambiando debido al desarrollo de métodos de aprendizaje de redes genéricas que son muy competitivos (Acid & De Campos, 2005).

En esta parte experimental se han utilizado datos socioeconómicos de los estudiantes legalmente matriculados en la Facultad de Ciencias de la Ingeniería de la Universidad Técnica Estatal de Quevedo en Ecuador durante el periodo 2012-2013. Con esta base de dato se compararán los resultados que obtienen los algoritmos Naive Bayes, TAN, BAN, SBND y las combinaciones con diferentes métricas como K2, BIC, Akaike, BDEu. También se comparan los métodos RPDAG y C-RPDAG. Estos métodos construyen clasificadores que son redes bayesianas genéricas equivalentes en independencia y equivalentes en clasificación. El trabajo experimental fue realizado con la herramienta Weka (Garner, 1995). Las variables se ilustran en

la tabla 1. Los diferentes valores que asumen cada una de esas variables, se muestran en las tablas 2, 3, 4, 5.

Tabla 1. Variables y sus descripciones.

| Variable | Descripción |
|----------|--------------------------------|
| A | Carrera |
| B | Curso |
| D | Discapacidad |
| E | Costo de la educación |
| F | Vive separado de la familia |
| G | Tipo de vivienda de la familia |
| H | Propietario de la vivienda |
| I | Servicio de TV Cable |
| J | Servicio de tarjeta de crédito |
| K | Servicio de acceso a internet |
| L | Servicios básicos |
| M | Servicio de transporte privado |
| N | Servicio de plan celular |
| O | Servicio de carro propio |
| P | Viene en carro propio |
| Q | Trabaja actualmente |
| R | Aprobó |
| S | Desertó |

Tabla 2 Valores y descripción de la variable carrera.

| Valor | Descripción |
|-------|------------------------------------|
| FI024 | Ingeniería en Sistemas |
| FI025 | Ingeniería en Diseño Gráfico |
| FI026 | Ingeniería Mecánica |
| FI027 | Ingeniería Industrial |
| FI028 | Ingeniería en Telemática |
| FI029 | Ingeniería Eléctrica |
| FI030 | Ingeniería Agroindustrial |
| FI031 | Ingeniería en Seguridad Industrial |

Tabla 3. Valor y descripción de la variable curso.

| Valor | Descripción |
|-------|-------------|
| 1 | Primero |
| 2 | Segundo |
| 3 | Tercero |
| 4 | Cuarto |
| 5 | Quinto |

Tabla 4. Variable y descripción de valores.

| Variable | Descripción |
|----------|-------------|
| D | SI=1; NO=0 |
| F | SI=1; NO=0 |
| I | SI=1; NO=0 |
| J | SI=1; NO=0 |
| K | SI=1; NO=0 |
| L | SI=1; NO=0 |
| M | SI=1; NO=0 |
| N | SI=1; NO=0 |
| O | SI=1; NO=0 |

| | |
|---|------------|
| P | SI=1; NO=0 |
| Q | SI=1; NO=0 |
| R | SI=1; NO=0 |
| S | SI=1; NO=0 |

Tabla 5. Variable y descripción de valores.

| Variable | Descripción |
|----------|---|
| E | X<200=0; 200>X<800=1; X>800=2 |
| G | Media agua=0; Casa/Villa=1; Departamento=2; Cuarto de inquilinato=3; Otra=4; Rancho=5 |
| H | Padre y madre=0; Padre=1; Madre=2; Otro pariente=3; Otro=4 |

Weka permite trabajar con distintas herramientas de minería de datos como reglas de asociación, agrupación, clasificación y regresión. Podemos visualizar en la figura 1 los resultados obtenidos por cada uno de los atributos en referencia a la clase S (deserta). De forma visual, ya se puede observar que no hay variables que de forma individual den información significativa sobre la deserción.

De acuerdo a la figura 1 se ha construido las tablas 6, 7 y 8, en el que podemos encontrar el análisis descriptivo de las variables donde se puede visualizar el porcentaje de cada uno de sus valores.

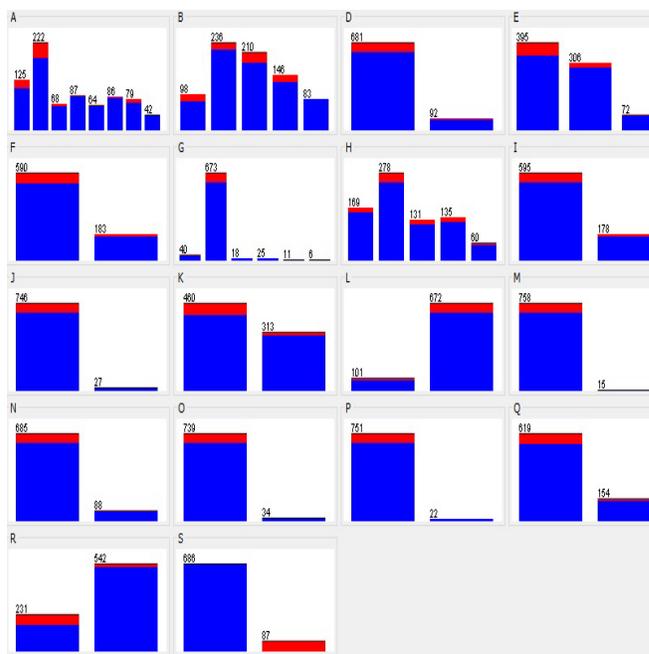


Figura 1. Resultados obtenidos por cada uno de los atributos en referencia a la clase.

Tabla 6. Variables y análisis descriptivo.

| Variable | Descripción | Cantidad | Porcentaje |
|----------|-------------|----------|------------|
| A | FCI024 | 125 | 16.17% |
| | FCI025 | 222 | 28.72% |
| | FCI026 | 68 | 8.80% |
| | FCI027 | 87 | 11.25% |
| | FCI028 | 64 | 8.28% |
| | FCI029 | 86 | 11.13% |
| | FCI030 | 79 | 10.22% |
| | FCI031 | 42 | 5.43% |
| B | 1 | 98 | 12.68% |
| | 2 | 236 | 30.53% |
| | 3 | 210 | 27.17% |
| | 4 | 146 | 18.89% |
| | 5 | 83 | 10.74% |
| D | No | 681 | 88.10% |
| | Si | 92 | 11.90% |

Tabla 7. Variables y análisis descriptivo.

| Variable | Descripción | Cantidad | Porcentaje |
|----------|--------------------|----------|------------|
| E | X<200 | 395 | 51.10% |
| | 200>X<800 | 306 | 39.59% |
| | X>800 | 72 | 9.31% |
| F | no | 590 | 76.33% |
| | si | 183 | 23.67% |
| G | Media agua | 40 | 5.17% |
| | Casa/villa | 673 | 87.06% |
| | Departamento | 18 | 2.33% |
| | Cuarto inquilinato | 25 | 3.23% |
| | Otro | 11 | 1.42% |
| | Rancho | 6 | 0.78% |
| H | Padre y madre | 169 | 21.87% |
| | Padre | 278 | 35.97% |
| | Madre | 131 | 16.95% |
| | Otro pariente | 135 | 17.46% |
| | Otro | 60 | 7.76% |
| I | no | 595 | 76.97% |
| | si | 178 | 23.03% |
| J | no | 746 | 96.51% |
| | si | 27 | 3.49% |
| K | no | 460 | 59.51% |
| | si | 313 | 40.49% |
| L | no | 101 | 13.07% |
| | si | 672 | 86.93% |
| M | no | 758 | 98.06% |
| | si | 15 | 1.94% |
| N | no | 685 | 88.62% |
| | si | 88 | 11.38% |
| O | no | 739 | 95.60% |
| | si | 34 | 4.40% |
| P | no | 751 | 97.15% |
| | si | 22 | 2.75% |

Tabla 8. Variables y análisis descriptivo.

| Variable | Descripción | Cantidad | Porcentaje |
|----------|-------------|----------|------------|
| Q | no | 619 | 80.08% |
| | si | 154 | 19.92% |
| R | no | 231 | 29.89% |
| | si | 542 | 70.11% |
| S | no | 686 | 88.75% |
| | si | 87 | 11.25% |

Se obtuvieron resultados usando como clasificadores Naive Bayes y BayesNet con diferentes alternativas como K2, TAN, Hill Climber con un padre y también con un máximo de 5 padres.

Tabla 9. Resultados obtenidos con diferentes clasificadores.

| Clasificador | Clasificados correctamente | Tasa TN | Tasa TP |
|-----------------------------|----------------------------|---------|---------|
| NaiveBayes | 88.0983 | 0.965 | 0.218 |
| BayesNet con K2-1 padre | 87.9690 | 0.964 | 0.218 |
| BayesNet con K2-5 padres | 91.0737 | 0.985 | 0.322 |
| BayesNet con TAN | 89.9094 | 0.974 | 0.318 |
| BayesNet con Hill Climber-1 | 88.6158 | 0.974 | 0.195 |
| BayesNet con Hill Climber-5 | 89.6507 | 0.975 | 0.276 |

Para obtener estos valores en la herramienta Weka hemos clasificado usando una validación cruzada de 10, con 773 casos teniendo como resultados que BayesNet con K2 y máximo 5 padres es el que mejor ha clasificado correctamente (91.0737%) y adicionalmente nos indica la tasa de verdaderos negativos (TN) o especificidad que corresponde a la probabilidad de que un estudiante que esté bien en su proceso académico tenga un resultado negativo en la prueba. En este caso solo se llega a detectar como falsos positivos un 1.5% (1- 0.985) y la tasa de verdaderos positivos (TP) o sensibilidad donde se considera que la clase positiva es los alumnos que desertan no es muy alta, pero se llega a detectar casi la tercera parte de los alumnos que van a desertar, pero también se debe indicar que el coste en función de falsos positivos es muy bajo 0,015

Se puede observar que hay un 32.20% de sensibilidad usando distintos clasificadores de Weka. Este es el porcentaje de estudiantes que se han clasificado correctamente entre aquellos que desertan. Éstos son los que deberían de recibir alguna atención y sobre los que habría que aplicar acciones especiales para disminuir este índice.

En la figura 2 podemos visualizar que todas las variables están relacionadas directamente con la clase deserción (S). La variable curso (B) depende también de la carrera (A) e influyen sobre el resultado académico (R). Por otro lado, se puede también considerar que la variable contar con servicio de tv cable (I) influye directamente sobre (J, K, M), servicio de plan celular (N), y éstas sobre servicio de vehículo propio (O) y trabajar actualmente (Q).

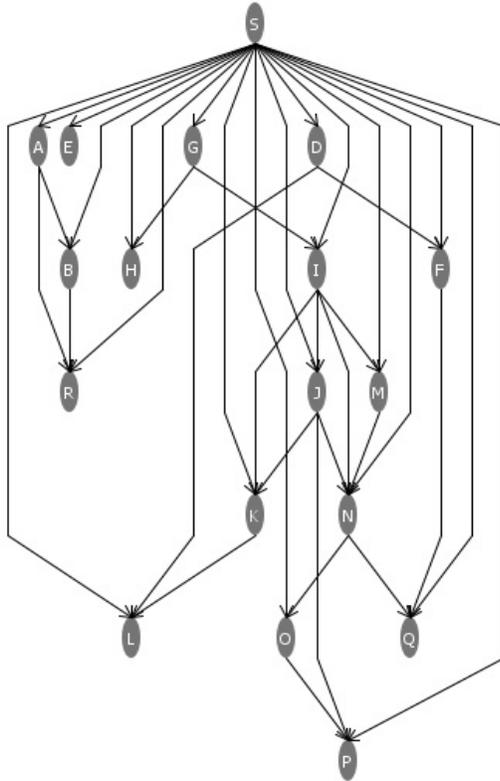


Figura 2. Red obtenida con clasificador BayesNet con K2 y un máximo de 5 padres.

En la figura 3 podemos visualizar la dependencia de las variables con la clase S; así como también que el curso (B) y la variable trabaja actualmente (Q) van a estar dependiendo directamente de la variable carrera (A).

El TAN considera sólo las relaciones más relevantes. Entre ellas la dependencia de la carrera para saber si se aprueba o no el curso (R), de igual manera que si el estudiante trabaja (Q) vivirá en un domicilio diferente al de la familia (F).

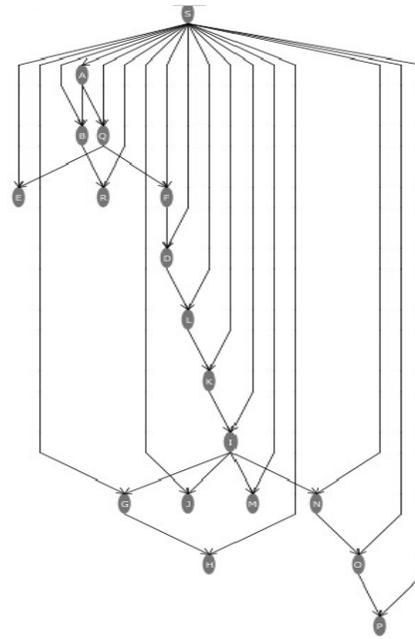


Figura 3. Red obtenida con clasificador BayesNet con TAN.

En la figura 4 podemos visualizar la red obtenida con con BayesNet y un solo padre. Hay influencia de 5 variables carrera (A), curso (B), costo de educación (E), servicio de internet (K) y aprueba (R) de manera directa con la clase deserta (S); así como también la de la variable estudiante trabaja actualmente (Q) con carrera (A). También se demuestra una dependencia fuerte de la variable tener servicio de internet (K) con las variables servicios básicos (L) y con servicio de tv cable (I).

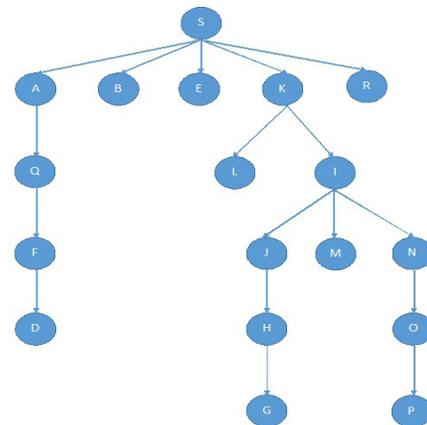


Figura 4. Red obtenida con clasificador BayesNet con HILL CLIMBER y un solo padre.

En la figura 5 podemos visualizar la dependencia de las variables carrera (A), curso (B), costo de educación (E), servicio de internet (K) y aprueba (R) con la clase deserta (S); se sigue manteniendo la dependencia tanto de servicios básicos (L) como servicio de plan celular (N) de servicio de acceso a internet (K) y está de servicio de tarjeta de crédito (J).

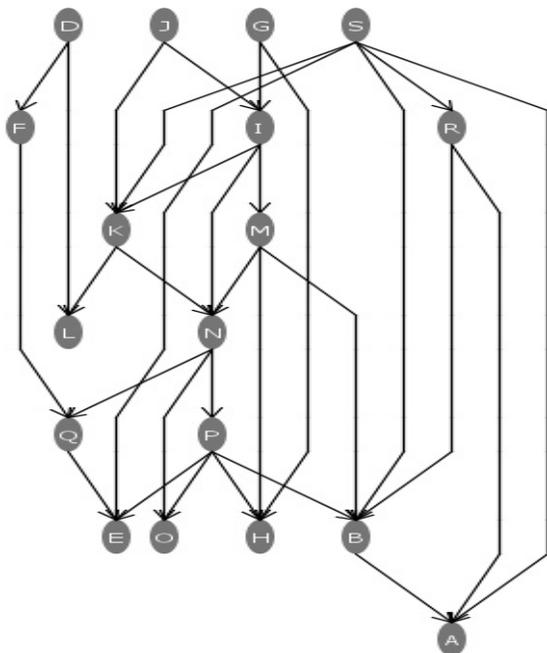


Figura 5. Red obtenida con clasificador BayesNet con HILL CLIMBER y un máx de 5 padres.

Los resultados que se obtuvieron usando como clasificadores de árboles J48 y Random Forest pueden ser visualizados en la tabla 10.

Tabla 10. Resultados obtenidos con clasificadores de árboles.

| Clasificador | Clasificados correctamente | Tasa TN | Tasa TP |
|--------------|----------------------------|---------|---------|
| J48 | 88.227 | 0.993 | 0.011 |
| RandomForest | 89.6507 | 0.978 | 0.253 |

Como se puede observar trabajando con un clasificador de árbol J48 los casos clasificados correctamente equivalen al 88.2277%. Además, se indica el porcentaje de sensibilidad y especificidad. Estos valores no mejoran en relación al del clasificador BayesNet con K2 y 5 padres. De igual manera se puede observar que trabajando con un clasificador de árbol Random Forest, los casos clasificados correctamente mejoran en referencia a J48. Se debe

indicar que es un bosque aleatorio de 100 árboles de los cuáles cada uno está construido con 5 características.

Los resultados usando como reglas de clasificación ZeroR y tablas de decisiones se pueden ver en la tabla 11.

Tabla 11. Resultados obtenidos con diferentes reglas de clasificación.

| Clasificador | Clasificados correctamente | Tasa TN | Tasa TP |
|---------------------|----------------------------|---------|---------|
| ZeroR | 88.7451 | 1 | 0 |
| Tabla de decisiones | 89.0039 | 0.981 | 0.172 |

Como se puede observar trabajando con reglas de clasificación ZeroR que corresponde a clasificadores triviales que siempre responden a la clase más frecuente ya sea negativa o positiva, los casos clasificados correctamente equivalen al 88.7451%. En el caso del trabajo con tablas de decisiones, los casos clasificados correctamente mejoran en referencia a ZeroR.

A continuación, se realizará una comparación de los resultados obtenidos con los diferentes algoritmos del estado del arte con la base de datos de variables socio-económicas de los estudiantes de la Facultad de Ciencias de la Ingeniería de la Universidad Técnica Estatal de Quevedo tal como se indica en la tabla 12.

Tabla 12. Resultados con base de datos de estudiantes de la UTEQ.

| Datos | SBND1 | SBND2 | SBND3 | SBND4 | BAN BDe | BAN BIC |
|----------------|--------|-----------|-----------|----------|---------|------------|
| Socioeconómico | 88.232 | 88.745 | 88.357 | 89.910 | 87.581 | 87.711 |
| Datos | BAN K2 | RPDag BDe | RPDag BIC | RPDag K2 | TAN | NaiveBayes |
| Socioeconómico | 87.723 | 87.584 | 87.972 | 89.657 | 88.743 | 87.456 |

Finalmente hemos aplicado los algoritmos de SBND y otros del estado del arte y como se puede observar el algoritmo que entrega mejores resultados es SBND con K2 con mucha diferencia de los otros con los que se han comparado y que los de peor resultado son con métrica BDe y BAN con la métrica BIC.

En virtud de que valor que se obtiene con el test de Friedman es mayor que 0.05, la hipótesis nula no es rechazada y se determina que no hay diferencias significativas entre las distribuciones y por lo tanto no es necesario seguir realizando pruebas. Estos resultados se originan en virtud de que se han comparado pocas bases de datos.

CONCLUSIONES

En la investigación se ha introducido un clasificador bayesiano que se basa en obtener de forma rápida una frontera de Markov que es fácil de aprender y muy competitivo.

El comportamiento del clasificador en determinados ejemplos es dependiente de la métrica que se use. BIC no da buen resultado, Akaike da buenos resultados en referencia a la media y K2 da buenos resultados en test no paramétricos.

En referencia al análisis de los datos de la UTEQ tenemos que decir que no es un problema fácil, pero aun así algunos clasificadores obtienen una sensibilidad cercana a la tercera parte con un coste en falsos positivos de sólo el 1,5%.

REFERENCIAS BIBLIOGRÁFICAS

- Acid, S., & De Campos, L. M. (2005). Learning Bayesian network classifiers: Searching in a space of partially directed acyclic graphs. *Machine Learning*, 59, 213-235. Recuperado de <https://link.springer.com/article/10.1007/s10994-005-0473-4>
- Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. *Educational Data Mining 2009*. Recuperado de <http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf>.
- Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. *Proceedings of the New Zealand Computer Science Research Students Conference*. Recuperado de <https://www.cs.waikato.ac.nz/ml/publications/1995/Garner95-WEKA.pdf>.
- Kuna, H., García Martínez, R., & Villatoro, F. (2010). Identificación de causales de abandono de estudios universitarios. Uso de procesos de explotación de información. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*. Recuperado de <http://sistemas.unla.edu.ar/sistemas/gisi/papers/TEYET-5-39-44-2010.pdf>.
- Lykourantzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3), 950-965. Recuperado de <https://eric.ed.gov/?id=EJ848773>
- Magaña, M.A., Montesino, O. A., & Hernández, C. M. (2006). Análisis de la evolución de los resultados obtenidos por los profesores en las evaluaciones ESDEPED y las realizadas por los estudiantes. *Revista de la Educación Superior*, 35(4), 29-48. Recuperado de http://publicaciones.anui.es/pdfs/revista/Revista140_S1A2ES.pdf
- Morales, M., & Salmerón, A. (2003). Análisis del alumnado de la Universidad de Almería mediante redes bayesianas. *27 Congreso Nacional de Estadística e Investigación Operativa*. Universitat de Lleida.
- Oviedo, B., Moral, S., & Puris, A. (2016). A hierarchical clustering method: Applications to educational data. *Intelligent Data Analysis*, 20(4), 933-951. Recuperado de <https://content.iospress.com/articles/intelligent-data-analysis/ida839>
- Porcel, E., Dapozo, G., & López, M. (2010). Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. *Revista Electrónica de Investigación Educativa*, 12(2), 1-21. Recuperado de <https://redie.uabc.mx/redie/article/view/264/730>