Revista Americana de Educación Farmacéutica 2018; 82 (7) Artículo 6980.

COMENTARIO

La aplicación de los árboles de clasificación a las admisiones en las facultades de farmacia

Samuel C. Karpen, PhDa Steve C. Ellis, MSb

[a] Facultad de Medicina Veterinaria, Universidad de Georgia, Athens, Georgia

[B] Facultad de Farmacia Bill Gatton, Universidad Estatal del Este de Tennessee, Johnson City, Tennessee

Presentado el 26 de enero de 2018; aceptado el 8 de mayo de 2018; publicado en septiembre de 2018.

En los últimos años, la Asociación Estadounidense de Facultades de Farmacia (AACP) ha fomentado la aplicación de técnicas analíticas de big data a la educación farmacéutica. De hecho, el Informe del Comité de Asuntos Académicos de 2013-2014 incluyó una sección de "Analítica de aprendizaje en la educación farmacéutica" que revisó los beneficios potenciales de adoptar técnicas de macrodatos. aula, práctica y admisiones.2 Si bien ambos informes fueron exhaustivos, ninguno discutió técnicas analíticas específicas. En consecuencia, este comentario introducirá árboles de clasificación, con especial énfasis en su uso en la admisión. Con las solicitudes electrónicas, las facultades y facultades de farmacia ahora tienen acceso a registros detallados de los solicitantes que contienen miles de observaciones. Con la disminución de solicitudes en todo el país, el análisis de admisiones puede ser más importante que nunca.3 Palabras clave: análisis predictivo, admisiones, árbol de decisiones

INTRODUCCIÓN Los

árboles de clasificación intentan separar los datos en grupos homogéneos al máximo en términos de un resultado de interés. Por ejemplo, el restaurador representado en la Figura 1 quiere saber cuándo esperarán los clientes para sentarse en su restaurante, en lugar de irse para ir a otro lugar.4 En consecuencia, registró información sobre 12 clientes. El restaurador encontró que el número de clientes en el restaurante diferenciaba mejor a los clientes en términos de

espera. Cuando no había clientes en el restaurante, nadie esperaba; cuando había algunos clientes, todos esperaban; y cuando el restaurante estaba lleno, dos de cada seis clientes esperaban. El tipo de restaurante, sin embargo, no fue tan informativo como el tamaño de la multitud, porque tantas personas esperaron como no esperaron en cada tipo de restaurante: en los restaurantes francés e italiano, uno esperó y el otro no, mientras que en los restaurantes tailandés y de hamburguesas , dos esperaron y dos no. En otras palabras, no se pudo trazar una distinción clara entre esperar y no esperar. La figura 1 también introduce otro atributo: si los clientes tenían o no hambre. Ninguno de los clientes saciados del grupo “completo” esperó, pero sí la mitad de los

hambrientos. El siguiente paso sería determinar qué variable(s) divide(n) a los clientes hambrientos en grupos homogéneos de meseros y no meseros.

Autor para correspondencia: Samuel C. Karpen, Facultad de Medicina Veterinaria, Universidad de Georgia, 501 DW Brooks Dr., Athens, GA 30602. Tel: 423-439-6883. Fax: 706-542-5460. Correo electrónico: sckarpen@uga.edu

La ventaja más clara de los árboles de clasificación es su interpretabilidad. Al presentar el análisis como una serie de clasificaciones binarias, brindan un gráfico sencillo para quienes no son estadísticos, a diferencia de la regresión logística, que puede ser difícil de interpretar cuando se incluyen muchas variables.
Los árboles de clasificación, sin embargo, pueden capturar relaciones complejas sin la dificultad adicional de interpretación.
Además, cuando la relación entre los predictores de uno y la variable dependiente es marcadamente no lineal, los árboles de clasificación tienden a producir predicciones mucho más precisas que la regresión logística.5-7 A pesar de sus ventajas, los árboles de clasificación

tienden a sobreajustarse.5-7Eso es , los árboles de clasificación son aptos para modelar tanto la relación entre las variables de interés como las idiosincrasias de los datos. Por ejemplo, si un investigador está tratando de modelar la relación entre la información demográfica de los estudiantes y las calificaciones de primer año de PharmD, y los datos contienen cuatro estudiantes de Nueva York con GPA de P1 altos, un árbol de categorización puede indicar que vivir en Nueva York es un predictor importante de P1 GPA cuando es solo una peculiaridad de esos datos en particular. Como consecuencia, el algoritmo del árbol no hará predicciones precisas cuando se aplique a nuevos datos con nuevas idiosincrasias.

Debido a esta limitación, es importante construir el árbol de clasificación de uno en un conjunto de datos inicial (el conjunto de entrenamiento) y luego aplicarlo a un nuevo conjunto de datos (el conjunto de prueba) para determinar qué tan bien se generaliza antes de implementa
Tradicionalmente, el analista divide aleatoriamente los datos 80-20, utilizando el 80 % de los datos para el conjunto de entrenamiento y el 20 % para el conjunto de prueba.7 Los analistas también pueden optar por utilizar la validación cruzada k-fold

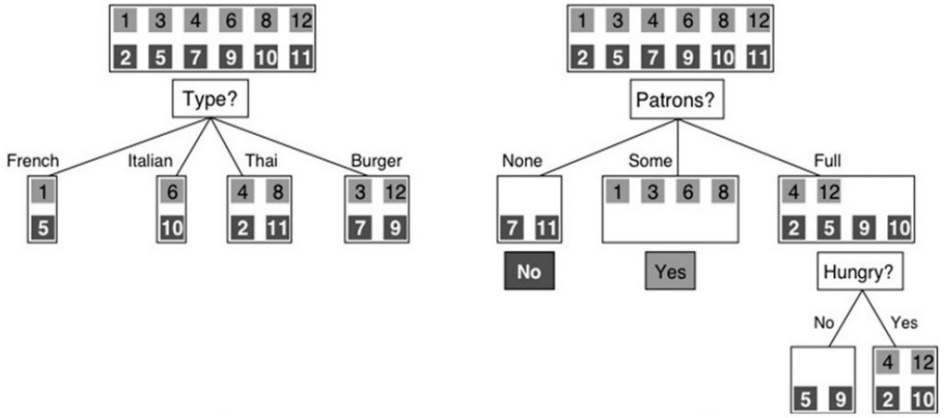

Figura 1. El número de clientes en un restaurante hace un mejor trabajo al dividir a los clientes potenciales en grupos de meseros y no meseros que el tipo de comida que se sirve.

para aumentar la generalización. Para mejorar aún más la generalización de un árbol, los investigadores pueden usar bosques aleatorios. Para generar un bosque aleatorio, el algoritmo selecciona aleatoriamente un subconjunto de variables de todas las variables y genera un árbol de clasificación basado en cada subconjunto. Por ejemplo, los programas de farmacia pueden estar interesados en predecir si un solicitante progresará normalmente. Cuando se utilizan variables comunes como la edad del solicitante, el GPA de ciencias previas a la farmacia, PCAT de biología y el título más alto del solicitante, el algoritmo puede seleccionar PCAT y GPA de ciencias previas a la farmacia, y generar un árbol basado solo en esas variables. Además de seleccionar variables al azar, el algoritmo también selecciona casos al azar, de modo que las variables seleccionadas al azar solo se usan para construir un árbol en un subconjunto de los casos. Las predicciones de los árboles generados aleatoriamente se combinan luego para producir una estimación más generalizable. La generación y combinación aleatoria de árboles mejora la generalización porque cada árbol solo produce estimaciones basadas en parte de los datos; por lo tanto, hay una gran variación en los árboles. Dado que es poco probable que los árboles dependan unos de otros (porque se generan a partir de un conjunto diferente de variables), algunos sobreestimarán el resultado y otros lo subestimarán; por lo tanto, su estimación agregada debe ser una representación precisa del resultado en la población.

En consecuencia, este árbol debería generalizarse bien. Los analistas, sin embargo, aún deben usar conjuntos de datos de entrenamiento y prueba o validación cruzada de k-fold con bosques aleatorios y regresión para el caso.

Los árboles de clasificación también adolecen de limitaciones no estadísticas, ya que requieren capacitación estadística avanzada y capacidad de programación. La mayoría del personal de evaluación no tiene este conjunto de habilidades. Además, los árboles de clasificación requieren conjuntos de datos muy grandes para un rendimiento óptimo, idealmente 30001 casos. La mayoría de las facultades de farmacia no tienen tantos registros de estudiantes. Si una universidad es grande, bien establecido y tiene la experiencia necesaria, sin embargo, los árboles de clasificación pueden ser herramientas valiosas.

Ejemplo El

árbol de clasificación de la Figura 2 muestra la relación entre el estado de inscripción y varias variables disponibles en las solicitudes de las escuelas de farmacia. El árbol se puede interpretar de la siguiente manera: los ceros y unos en la parte superior de cada cuadro muestran el resultado de la mayoría de los solicitantes del grupo: cada grupo en el que al menos el 50 % de los solicitantes inscritos está etiquetado con un uno. En el primer cuadro, que representa los datos antes de cualquier división, hay un cero en la posición superior porque la mayoría de los solicitantes a los que se les extendió una oferta no se inscribieron. La proporción indica el porcentaje de personas de cada grupo que se inscribieron. En el primer recuadro, el 0,48 significa que el 48% de los alumnos a los que se les prorrogó alguna oferta finalmente se matriculó. Finalmente, el porcentaje indica el porcentaje de todos los solicitantes que están en ese grupo.
Este número es 100 % en el primer cuadro porque representa todos los datos antes de dividirlos en grupos.

El cuadro en la esquina inferior izquierda del árbol representa la información en la que una escuela estaría más interesada (es decir, el tipo de solicitante que es más probable que se inscriba). Como se muestra, el 36 % de los solicitantes con al menos un título de licenciatura con un padre (ParentEd .5.5 es "sí"), que asistieron a una universidad de cuatro años (FourYear .5.05 es "sí") y obtuvieron un GPA anterior a la farmacia superior a 2.6 (PrePharmGPA.52.6 es "sí") inscrito. Colectivamente, este grupo representó el 50% de todos los solicitantes. Cabe señalar que si una variable se codifica como 1 y 0, como la educación de los padres (ParentEd), el software que generó el árbol denota divisiones por menos o más de 0,5. Menos de 0,5 es "el grupo codificado como 0" mientras que mayor de 0,5 es "el grupo codificado como 1". En el caso del árbol de clasificación en la Figura 2 FourYear .50.5 es "no" significa que el solicitante no asistió a una institución de cuatro años porque "institución de cuatro años" se codificó como 1,

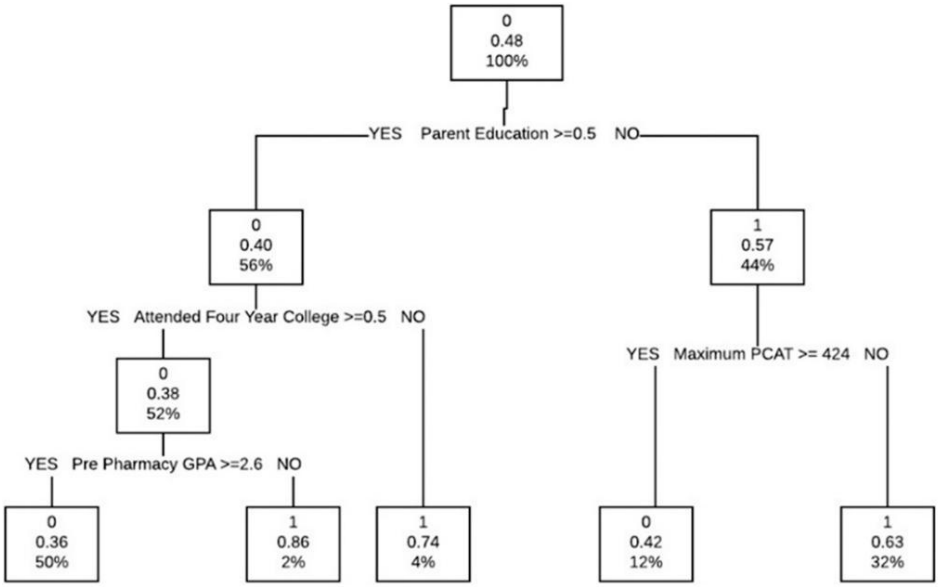

Figura 2. El árbol muestra una predicción general (05menos del 50 % del grupo inscrito, 15más del 50 % del grupo inscrito), la proporción de cada grupo que se inscribió, y el porcentaje de la muestra total constituida por un grupo dado. Por ejemplo, el 57% de solicitantes cuyos padres no tenían un título de cuatro años inscrito (ParentEd .5no). Este grupo representó el 44% de todos los solicitantes.

Cuando se aplicó al conjunto de prueba, el árbol de ejemplo predijo con precisión el 70,6 % de los casos. Para mejorar aún más precisión del árbol, los investigadores podrían incluir información adicional atributos (por ejemplo, si un estudiante tiene una oferta competidora) o usar un bosque aleatorio. En consecuencia, el bosque aleatorio en los datos de entrenamiento predijeron el 77% de los casos en la prueba conjunto de datos correctamente. Si el liderazgo (decanos, asociados/asistentes decanos, etc.) se siente cómodo con el 77%, entonces el árbol será listo para su uso en las decisiones de admisión.

CONCLUSIÓN

Los árboles de categorización se pueden utilizar para facilitar una variedad de decisiones en todo el espectro de la educación farmacéutica. Una El uso potencial es desarrollar un modelo predictivo para ayudar en decisión de admisión como lo describen Muratov y col.8 En este proyecto, los autores desarrollaron una farmacia modelo predictivo de desempeño escolar para identificar postulantes para entrevistas basadas en su probabilidad de éxito académico. Mientras que el modelo de Muratov y sus colegas se basa principalmente en factores cognitivos, tales modelos de admisión también puede incluir factores no cognitivos que pueden ampliar la utilidad del modelo para una institución. Por ejemplo, para ayudar en identificar a los solicitantes que se ajusten a la misión de la escuela o, como ilustrado en este documento, para identificar a los solicitantes con mayor probabilidad de inscribirse, que a su vez puede informar el proceso de contratación. A El uso potencialmente innovador de los árboles de categorización respalda la evaluación de una escuela de la preparación APPE de los estudiantes como abordado en el Estándar 24.3 de los Estándares ACPE 2016. De terminar la preparación APPE es un desafío porque la preparación incluye habilidades comúnmente conocidas como habilidades blandas; la competencia de los estudiantes en estas áreas (o la falta de ella) a menudo no se manifiesta hasta que han comenzado las APPE. Examinar datos de cursos y experiencias relevantes durante el años didácticos pueden permitir una identificación más temprana de estudiantes que pueden necesitar preparación adicional antes de ingresando a la porción APPE del plan de estudios. De hecho, los árboles de clasificación y técnicas similares pueden ser útiles para cualquier situación en la que se requiera predicción.

Los árboles de clasificación también son muy útiles para fines exploratorios. análisis Por ejemplo, si una institución desea identificar las características demográficas de los estudiantes con dificultades, una árbol de clasificación podría ser una herramienta valiosa. Para tal fin, los autores de este comentario construyeron un árbol de clasificación para visualizar la relación entre la información de la aplicación: Puntuaciones de PCAT, GPA de matemáticas/ciencias antes de la farmacia, ya sea el estudiante asistió a una institución de cuatro años, ya sea el estudiante era de nuestra región, edad y sexo, y si el estudiante reprobó al menos una clase durante su primer año. Si bien los resultados no fueron del todo sorprendentes en el sentido de que los estudiantes con promedios de calificaciones bajos en matemáticas/ciencias antes de la farmacia y bajos Los puntajes PCAT de biología tenían más probabilidades de reprobar un curso que aquellos con puntajes más altos, el árbol también indicó que los estudiantes fuera del rango de edad tradicional (menos de 22 o mayores de 32) eran propensos a luchar. Mientras este solitario tree no debe usarse para la predicción, descubrió un problema potencial que de otro modo podría haber pasado desapercibido.

Cada vez es más importante que el Doctor en Farmacia los programas desarrollan métodos para analizar datos programáticos en formas que faciliten la toma de decisiones. Los macrodatos y los análisis asociados ya se han mostrado prometedores en las universidades

que estén dispuestos a invertir en ellos. Después de utilizar el análisis de big data para identificar a los estudiantes en riesgo (y asignarles las intervenciones adecuadas), la Universidad Estatal de Georgia experimentó un aumento del 6 % en la tasa de graduación durante tres años, una disminución de medio semestre en la cantidad de tiempo necesario para graduarse y mejor desempeño en los cursos STEM por parte de los estudiantes de primera generación.9 Si bien fue la intervención de la escuela lo que finalmente aumentó el desempeño y la retención de los estudiantes, los modelos estadísticos mostraron que debían enfocar sus esfuerzos. De manera similar, los programas de farmacia pueden usar herramientas como el análisis de árboles de clasificación para obtener información sobre sus grupos de solicitantes y el desempeño de los estudiantes de maneras que van más allá de los análisis tradicionales.

AGRADECIMIENTOS Los autores agradecen a David Roane por sus comentarios relacionados con este trabajo.

REFERENCIAS

1. Cain J, Conway JM, DiVall MV, et al. Informe del Comité de Asuntos Académicos 2013-2014. Soy J Pharm Educ. 2014;78(10): Artículo S23.

2. Baldwin JN, Bootman JL, Carter RA, et al. Práctica farmacéutica, educación e investigación en la era de los grandes datos: informe de la Comisión Argus 2014-2015. Soy J Pharm Educ. 2015;79(10):Artículo S26.

3. Reunión del consejo administrativo del Consejo de Decanos de la AACP. Presentación oral en la reunión intermedia de la AACP. 25 de febrero de 2017, Río Grande, Puerto Rico.

4. Russell S, Norvig P. Inteligencia artificial: un enfoque moderno. Essex, Reino Unido: Pearson Education; 2014.

5. Breiman L. Bosques aleatorios. Aprendizaje automático. 2001;45(1): 5-32.

6. Hayes, T Usami S, Jacobucci R, McArdle. Uso de árboles de clasificación y regresión (CART) y bosques aleatorios para analizar el desgaste: resultados de dos simulaciones. Envejecimiento Psicológico. 2015;30(4): 911-929.

7. Strobl C, Malley J, Tutz G. Una introducción a la partición recursiva: justificación, aplicación y características de los árboles de clasificación y regresión, embolsado y bosques aleatorios. Métodos Psicológicos. 2009;14(4):323-348.

8. Muratov E, Lewis M, Fourches D, Tropsha A, Cox WC. Soporte de decisiones asistido por computadora para la admisión de estudiantes en función de su rendimiento académico previsto. Soy J Pharm Educ. 2017; 81(3): Artículo 46.

9. Kamenetz A. Cómo una universidad usó big data para impulsar las tasas de graduación. NPRED. 30 de octubre de 2016. http://www.npr.org/sections/ed/2016/10/30/499200614/how-one-university-used-big data-to-boost - graduation-rates.