ESTRUCTURAS DE DATOS PARA LA PREDICCIÓN DE RENDIMIENTO EN LAS PRUEBAS SABER 11

Juan Pablo Restrepo Universidad Eafit Colombia jprestrepe@eafit.edu.co	Juan José Sánchez Universidad Eafit Colombia jjsanchezc@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

Si bien se conoce que el objetivo principal de las Pruebas Saber 11 es el mejoramiento de la calidad de la educación colombiana se han hecho diferentes análisis, teniendo en cuenta los factores que inciden en el desempeño de los estudiantes. Buscamos definir la probabilidad respecto a puntajes totales que sean indicios de un buen rendimiento académico por medio de árboles de precisión y algoritmos. Los cuales sean capaces de determinar un nivel de educación media lineal a los estándares esperados según el promedio de cohorte.

Palabras clave

Árboles de decisión, aprendizaje automático, predicción de los resultados de los exámenes, rendimiento académico, algoritmos, pruebas saber 11, minería de datos, modelos predictivos.

1. INTRODUCCIÓN

Se tiene a la prueba saber 11 como el camino final de todo estudiante, por lo cual la mayoría de estos poseen una capacitación a lo largo de su educación media en búsqueda de un buen puntaje que pueda proveer mayores oportunidades de acceder a la educación superior en un futuro. Muchos estudiantes crean una dependencia de estas pruebas como su única oportunidad de estudio, esforzándose así obtener buenos puntajes resultantes de altos percentiles por encima de la media.

Evidenciando lo anterior son muchas las entidades las que buscan conocer todos estos puntajes, para poder así otorgar diferentes tipos de becas, teniendo como objetivo regular el buen desarrollo académico de la persona dentro de un ámbito social.

1.1. Problema

A través de árboles de decisión y algoritmos se buscará promediar el éxito académico de un estudiante en relación con las variables académicas y sociodemográficas. Dando a conocer datos más específicos y estadísticos sobre el nivel de estudio que presentan las personas que aplican a esta prueba.

En pocas palabras, explique el problema, el impacto que tiene en la sociedad y por qué es importante resolver el problema. *(En este semestre, el problema es predecir el éxito académico)*

1.3 Estructura del artículo

En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. Finalmente, en la sección 6, discutimos los resultados y proponemos algunas direcciones de trabajo futuras.

2. TRABAJOS RELACIONADOS

Explique cuatro (4) artículos relacionados con el problema descrito en la sección 1.1. Puede encontrar los problemas relacionados en las revistas científicas. Considere el Google Scholar para su búsqueda. *(En este semestre, el trabajo relacionado es la investigación de árboles de decisión para predecir los resultados de los exámenes de los estudiantes o el éxito académico)*

2.1 Propuesta para el uso de árboles de decisión en la predicción de la trayectoria de los estudiantes

En junio del 2019 la revista Iberoamericana de Ciencias propuso utilizar los árboles de decisión para la predicción de la trayectoria de los estudiantes a partir de un estudio de la Universidad Veracruzana en México. Su objetivo principal era encontrar un sistema por el cual se pudiera predecir el éxito o fracaso de los nuevos estudiantes pertenecientes al programa de Ingeniería de Tecnologías Computacionales tomando como datos fundamentales la información personal, familiar, socioeconómica y académica de estudiantes egresados y de aquellos que no lograron terminar la carrera. Determinando así por medio del algoritmo de J48 en Weka una precisión de 0.935 para el éxito y 1.000 para el fracaso.

2.2 Predicción del rendimiento académico aplicando técnicas de minería de datos

En enero del 2017 la Universidad Nacional Agraria La Molina en Perú publicó un articulo donde se plantea la predicción del rendimiento académico a partir de múltiples algoritmos para el funcionamiento de los arboles de decisión. El algoritmo más destacado es el M5P considerando variables únicas para los parámetros qué en conjunto con el algoritmo J48 es comparado con las redes bayesianas, los cuales dieran como resultado este sistema que sea capaz de predecir el buen rendimiento académico de los estudiantes de Ingeniería o en el caso contrario de un mal rendimiento poder así avisar a los docentes para hacer un correcto asesoramiento. Respecto a la precisión se obtuvo un 94% respecto a estudiantes aprobados y desaprobados.

2.3 Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°

En junio del 2019 la revista de investigación, desarrollo e innovación realizó en la Universidad UPTC en Duitama un modelo de clasificación que por medio de árboles de decisión diera a conocer todos los factores que determinaban el desempeño estudiantil de grado undécimo correspondientes al periodo 2015-2016. Utilizando el algoritmo J48 e implementando a su vez el C.45 teniendo como privilegio utilizar la variable confianza C para poder lograr así una capacidad de predicción estable. Respecto a la precisión lograron concluir su trabajo con instancias clasificadas correctamente en un 67 % e instancias clasificadas incorrectamente en un 33%.

2.4 Predicción del rendimiento académico en las nuevas titulaciones de grado de la EPS de la Universidad de Córdoba

En julio del 2012 en la Universidad de Córdoba en Madrid realizó una investigación para lograr la predicción del rendimiento académico respecto a las titulaciones de grado de las ingenierías. En esta investigación se utilizaron los algoritmos NaiveBayes, PART y J48 ya que eran previstos como poseedores de máxima exactitud. Pero se hizo énfasis en el algoritmo PART Ya que fue el que dio más resultados trabajando con menos reglas de If y Then para poder tener menos elementos afectantes de variables o factores. Teniendo variables como nivel de implicación de los profesores, Nivel de dificultad/exigencia, Asistencia, Nivel de apoyo, entre otros. Se logró la captura y procesado de la información que dio como resultado una precisión del 60% dando como razón de una precisión muy baja los múltiples factores que pueden afectar el rendimiento académico.

3. MATERIALES Y MÉTODOS

En esta sección se explica cómo se recopilaron y procesaron los datos y, después, cómo se consideraron diferentes alternativas de solución para elegir un algoritmo de árbol de decisión.

3.1 Recopilación y procesamiento de datos

Obtuvimos datos del *Instituto Colombiano de Fomento de la Educación Superior* (ICFES), que están disponibles en línea en ftp.icfes.gov.co. Estos datos incluyen resultados anonimizados de Saber 11 y Saber Pro. Se obtuvieron los resultados de Saber 11 de todos los gradudados de escuelas secundarias colombianas, de 2008 a 2014, y los resultados de Saber Pro de todos los graduados de pregrados colombianos, de 2012 a 2018. Hubo 864.000 registros para Saber 11 y 430.000 para Saber Pro. Tanto Saber 11 como Saber Pro, incluyeron, no sólo las puntuaciones sino también datos socioeconómicos de los estudiantes, recogidos por el ICFES, antes de la prueba.

En el siguiente paso, ambos conjuntos de datos se fusionaron usando el identificador único asignado a cada estudiante. Por lo tanto, se creó un nuevo conjunto de datos que incluía a los estudiantes que hicieron ambos exámenes estandarizados. El tamaño de este nuevo conjunto de datos es de 212.010 estudiantes. Después, la variable predictora binaria se definió de la siguiente manera: ¿El puntaje del estudiante en el Saber Pro es mayor que el promedio nacional del período en que presentó el examen?

Se descubrió que los conjuntos de datos no estaban equilibrados. Había 95.741 estudiantes por encima de la media y 101.332 por debajo de la media. Realizamos un submuestreo para equilibrar el conjunto de datos en una proporción de 50%-50%. Después del submuestreo, el conjunto final de datos tenía 191.412 estudiantes.

Por último, para analizar la eficiencia y las tasas de aprendizaje de nuestra implementación, creamos al azar subconjuntos del conjunto de datos principal,

como se muestra en la Tabla 1. Cada conjunto de datos se dividió en un 70% para entrenamiento y un 30% para validación. Los conjuntos de datos están disponibles en https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

	Conjunto de datos 1	**Conjunto de datos 2**	**Conjunto de datos 3**	**Conjunto de datos 4**	**Conjunto de datos 5**
Entrenamiento	15,000	45,000	75,000	105,000	135,000
Validación	5,000	15,000	25,000	35,000	45,000

Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

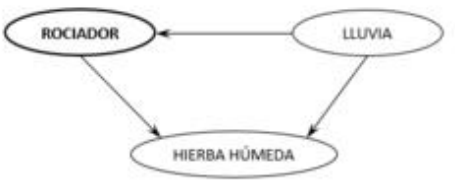
3.2 Alternativas de algoritmos de árbol de decisión

En lo que sigue, presentamos diferentes algoritmos usardos para construir automáticamente un árbol de decisión binario. *(En este semestre, ejemplos de tales algoritmos son ID3, C4.5 y CART).*

3.2.1 Red Bayesiana

También conocidas como redes probabilísticas funcionan en forma de que se combina la potencia del teorema de Bayes con la expresividad semántica de un grafo, creando así un conjunto de variables aleatorias y sus dependencias condicionales. Su objetivo es representar probabilidades condicionales por medio de relaciones causa – efecto entre los diferentes nodos que se presentan siendo así representada por un grafo acíclico dirigido.

LLUVIA	ROCIADOR T	ROCIADOR F
F	0.4	0.6
T	0.01	0.99



LLUVIA T	LLUVIA F
0.2	0.8

ROCIADOR	LLUVIA	HIERBA HÚMEDA T	HIERBA HÚMEDA F
F	F	0	1
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Un ejemplo muy interesante de su algoritmo podría ser el siguiente aplicado a la probabilidad de Aprobar o Desaprobar de un estudiante con precisión a posteriori.

$$P(C_k / X_0^k) = \prod_{j=1}^{p} P(Y^g) P(X_j^k / C_g); \quad g = 1,2,...,m$$

3.2.2 M5

El algoritmo M5 (Notado también como M5P) se destaca en la prueba de tests en este algoritmo se examinan todas las posibilidades, pero solo son seleccionadas aquellas posibilidades que tengan una reducción de esperada de fallo. Por ejemplo, el Algoritmo CART realiza esta misma función, pero en relación a cada variable o factor a tener en cuenta es de ahí que M5 es una variación de CART. Luego de estimar el error se construyen modelos lineales por medio de arboles de decisión que van a tener nodos intermedios como estructura base y su reducción de error estaría notada de la siguiente manera

$$\Delta\text{error} = sd(E) - \sum_i \frac{|E_i|}{|E|} \times sd(E_i)$$

Y figura vectorizada se puede representar de la siguiente manera:

```
CHMIN ≤ 7 : RM0
CHMIN > 7 :
|   MMAX > 24000 : RM1
|   MMAX ≤ 24000 :
|   |   CACH ≤ 48 : RM2
|   |   CACH > 48 : average 217.5
```

Model:	RM0	RM1	RM2
	11.5	-101.4	11.9
Cycle time			
Min mem		0.030	
Max mem	0.003		0.008
Cache size	0.902		
Min chans			
Max chans	0.518	4.686	

Figure 1: Model tree for CPU performance

	Original Attributes		Transformed Attributes	
	Correlation	Percentage Deviation	Correlation	Percentage Deviation
Ein-Dor (retrial)	–	–	.966	33.9%
IBP	–	35.0%	–	33.0%
M5	.921	34.9%	.956	34.0%
M5 (no smoothing)	.908	37.2%	.957	33.9%
M5 (no models)	.803	49.9%	.853	48.6%

Table 1: CPU performance data

3.2.3 PRISM

Notado como [CEN87] asume dos variables p y t donde los dos son cubiertos por la regla y su función es maximizar la condición de la relación donde específicamente para cada clase p hay que tener ejemplos de t . Siguiendo una orden de [decision list] el

algoritmo interpreta cada regla y lo registra hasta obtener la información requerida para eliminarlo al terminar su operación y por consiguiente busca una nueva regla que cumpla con su condición.

Su figura vectorizada se representa de la siguiente manera:

Regla 1. Clase "Sí".

Añadir a "If <vacío> Then Sí"	p/t
Vista = Soleado	2/5
Vista = Nublado	4/4
Vista = Lluvioso	3/5
Temperatura = Alta	2/4
Temperatura = Media	4/6
Temperatura = Baja	3/4
Humedad = Alta	3/7
Humedad = Normal	6/7
Viento = Sí	3/6
Viento = No	6/8

If Vista = Nublado Then Sí

Regla 2. Clase "Sí".

Añadir a "If <vacío> Then Sí"	p/t
Vista = Soleado	2/5
Vista = Lluvioso	3/5
Temperatura = Alta	0/2
Temperatura = Media	3/5
Temperatura = Baja	2/3
Humedad = Alta	1/5
Humedad = Normal	4/5
Viento = Sí	1/4
Viento = No	4/6

Añadir a "If Humedad=Normal Then Sí"	p/t
Vista = Soleado	2/2
Vista = Lluvioso	2/3
Temperatura = Alta	
Temperatura = Media	
Temperatura = Baja	
Viento = Sí	
Viento = No	

If Humedad=Normal **and** Viento = No **Then** S

Lista de Decisión Completa:

If Vista = Nublado **Then** Sí
If Humedad=Normal **and** Viento = No **Then** Sí
If Temperatura = Media **and** Humedad = Normal **Then** Sí
If Vista = Lluvioso **and** Viento = No **Then** Sí
If Vista = Soleado **and** Humedad = Alta **Then** No
If Vista = Lluvioso **and** Viento = Sí **Then** Sí

3.2.4 PART

Su entorno fue creado por los desarrolladores de Weka y es proporcionado por C4.5 [QUI93] tiene como función realizar dos fases. En su primera fase se generarán las reglas de clasificación para luego en una segunda fase refinarlas de forma de que la optimización global de las reglas se alcance con totalidad. La diferencia del algoritmo PART es que este de cierta forma optimiza esta globalización de forma en que sus reglas no necesitan ser mejoradas convirtiéndose así en PARTial decision trees. Entonces su funcionamiento pasa a ser la generación de reglas donde luego de eliminar sus ejemplares pasa en búsqueda de más reglas hasta que los ejemplos ya no se puedan clasificar.

Su figura vectorizada se representa de la siguiente manera

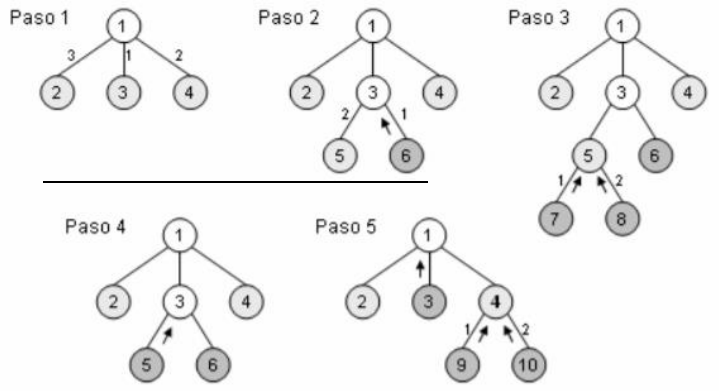


Figura 3.25: Ejemplo de generación de árbol parcial con PART.

4. DISEÑO DE LOS ALGORITMOS

En lo que sigue, explicamos la estructura de los datos y los algoritmos utilizados en este trabajo. La implementación del algoritmo y la estructura de datos se encuentra disponible en Github[1].

4.1 Estructura de los datos

Explique la estructura de datos utilizada para hacer la predicción y haga una figura que la explique. No utilice imágenes de Internet. *(En este semestre, la estructura de datos es un árbol de decisión binario)*

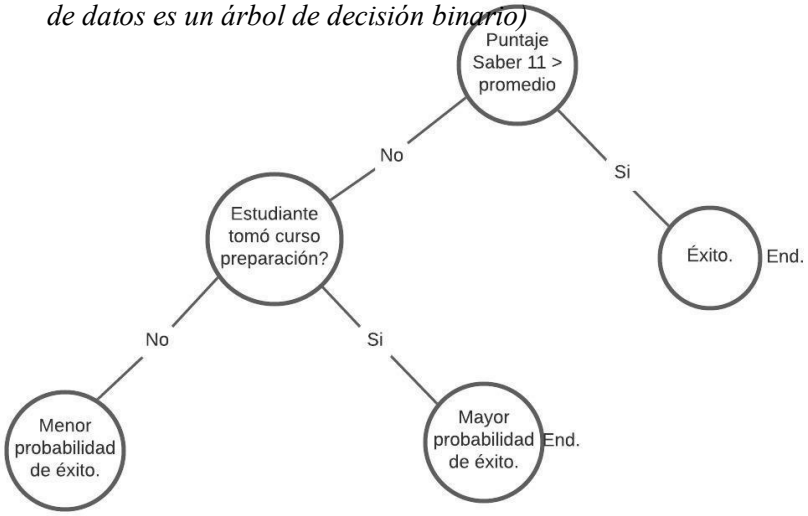


Figura 1: El algoritmo que escogimos fue CART. Con el cual podremos trabajar para determinar la probabilidad de éxito a partir de 2 variables como lo son el puntaje y el curso de preparación creando así producciones con más o menos probabilidad.

4.2.2 Algoritmo de prueba

Al crear cada nodo, se calculará la probabilidad por medio de la impureza de Gini definiendo así cada nodo según el porcentaje que entre más bajo, más alta será su probabilidad de éxito.

[1] http://www.github.com/jprestrepe/ST0245-002/tree/master/proyecto

Referencias

1. Pérez Cáceres, S., Rodriguez Flores, C., Morales Mendoza, E., Cruz Ramírez, H. and Carballo Franco, A., 2019. Propuesta Para El Uso De Árboles De Decisión En La Predicción De La Trayectoria De Los Estudiantes. [online] Reibci.org. Recuperado de: http://www.reibci.org/publicados/2019/jun/3400111.pdf.

2. Menacho Chiok, C., 2017. Predicción Del Rendimiento Académico Aplicando Técnicas De Minería De Datos. Dialnet.unirioja.es. Recuperado de: https://dialnet.unirioja.es/descarga/articulo/6171237.pdf.

3. Timarán Pereira, R., Caicedo Zambrano, J. and Hidalgo Troya, A., 2019. Árboles De Decisión Para Predecir Factores Asociados Al Desempeño Académico De Estudiantes De Bachillerato En Las Pruebas Saber 11°. Scielo.org.co. Recuperado de: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S2027-83062019000100363.

4. Romero, C., Zafra, A., Gibaja, E., Luque, M. and Ventura, S., 2012. *Predicción Del Rendimiento Académico En Las Nuevas Titulaciones De Grado De La EPS De La Universidad De Córdoba*. Upcommons.upc.edu. Recuperado de: https://upcommons.upc.edu/bitstream/handle/2099/15094/067.pdf.

5. Wikipedia. Red Bayesiana. Recuperado de: https://es.wikipedia.org/wiki/Red_bayesiana.

6. Quinlan, J., 2006. LEARNING WITH CONTINUOUS CLASSES. Sci2s.ugr.es. Recuperado de: https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Quinlan-AI.pdf.

7. Molina, J. and García, J., 2006. TÉCNICAS DE ANÁLISIS DE DATOS. Matema.ujaen.es. Recuperado de: http://matema.ujaen.es/jnavas/web_recursos/archivos/weka%20master%20recursos%20naturales/apuntesAD.pdf.