ÁRBOLES DE DECISIÓN PARA PRONOSTICAR EL ÉXITO EN LAS PRUEBAS SABER PRO

Juan David Correa Universidad Eafit Colombia jdcorread@eafit.edu.co	Stiven Ossa Universidad Eafit Colombia sossas@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

La finalidad de este informe es llevar a cabo una adecuada implementación de una estructura de datos y algoritmos, que permitan determinar todos aquellos factores que influyen en el desempeño de un estudiante en las pruebas de Educación Superior. Todo esto por medio de árboles de decisión y teniendo en cuenta múltiple información: académica, socioeconómica, institucional, etc. Gracias a esto, se aportará a los procesos de reforma educacionales, para así fortalecer la calidad de la educación en el país.

Palabras claves

Árboles de decisión, aprendizaje automático, éxito académico, predicción de los resultados de los exámenes

1. INTRODUCCIÓN

A medida que pasa el tiempo, el ser humano busca adaptarse a las circunstancias, encontrar soluciones y comodidad frente a los problemas que se enfrenta a diario, la educación no se queda atrás, ahora se usan los medios Y métodos tecnológicos, como los arboles de decisión para enfrentar brechas que a día de hoy se nos presenta, el manejo de datos y la tecnología nos ha ayudado a ver por qué los estudiantes desertan, como es que logran mayores éxitos en su vida académica, entre otras aspectos; para este caso queremos predecir el éxito que pueden tener en las pruebas de Educación Superior.

1.1. Problema

El problema enfrentado es, por medio del estudio de estructuras de datos y algoritmos, poner en práctica esto e implementar eficientemente un árbol de decisión capaz de predecir el éxito académico para las pruebas del estado de la Educación superior.

La solución a este problema sería un gran aporte a la educación, sobre todo a formación de los futuros profesionales del país.

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

2.1 Análisis del Desempeño Académico del Examen de Estado para el ingreso a la Educación Superior Aplicando Minería de Datos

Estudio de minería de datos basado en los resultado del Examen de Estado para el ingreso a la Educación Superior del año 2012. Se aplican técnicas de análisis de agrupamiento para construir un modelo que da entender de manera más clara la estructura de los datos. Se aplica el algoritmo K-means que permitió caracterizar los estudiantes que obtuvieron los diferentes niveles de desempeño en la prueba dadas sus condiciones. En cuanto a la precisión se obtuvieron datos consistentes.

2.2 Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia

Se analizan los resultado de las pruebas Saber Pro, usando metodología de extracción de conocimiento en bases de datos llamada KDD. El algoritmo aplicó la correcta correlación y pudo predecir de manera satisfactoria los puntajes de las pruebas, comparándolo con el resultado real del estudiante. Aunque en ciertas partes la predicción no fue exacta, pero igual los resultados fueron acordes.

2.3 Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO

Aplicando la metodología llamada CRISP-DM, se realizó un estudio de los resultados obtenidos en la prueba, seleccionando los factores que más influyen en el desempeño de la prueba, y así predecir el resultado en las pruebas basado en las variables seleccionadas. Se logró predecir el desempeño en las pruebas con una exactitud del 81%.

2.4 Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales

Lograron predecir los resultados que alcanzaría los estudiantes en las materias de Estructuras de Datos 1 y 2. Se implementó una aplicación que predecía los resultado teniendo en cuenta resultados académicos de semestres anteriores y datos socioeconómicos. Se realizaron varios entrenamiento y pruebas para buscar la mejor efectividad. Se logró una precisión mayor al 75%.

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 .

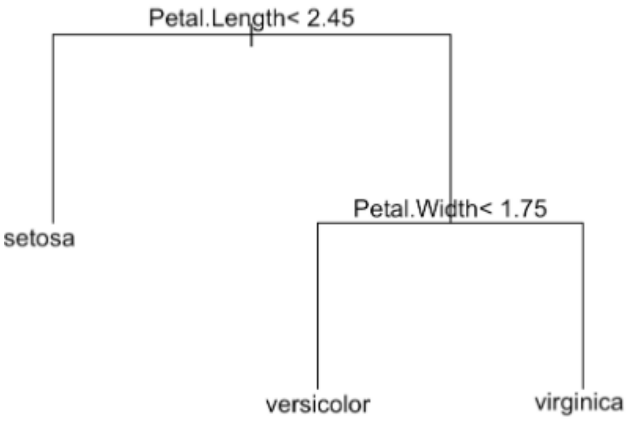
	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3	Conjunto de datos 4	Conjunto de datos 5
Entrenamiento	15,000	45,000	75,000	105,000	135,000
Validación	5,000	15,000	25,000	35,000	45,000

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

3.2.1 Árbol CART

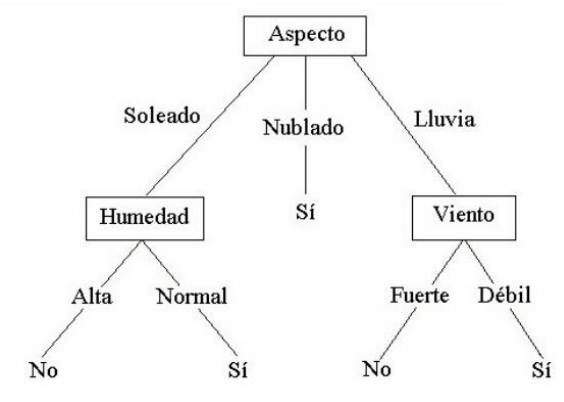
Diseñado por Breiman en 1984. Este algoritmo genera árboles de decision binaries (los nodos se dividen en dos). No solo se utiliza para la clasificación sino también para la regression. El algoritmo CART se basa, en lugar de Entropía y valores de ganancia de información, en la función índice de Gini, que simplifica la forma de definir la pureza de los nodos y hojas. Es fácil de usar debido a la facilidad con la que puede dar un buena interpretación de los datos, solo necesitando una Buena representación del problema.



3.2.2 Árbol ID3

El algortimo ID3 construye un árbol de decision de manera descendente, este algoritmo basado en un criterio estadístico se usa comúnmente en máquinas de aprendizaje que, a partir de un conjunto de datos, crea un árbol de decisiones; para ello es necesario conocer la entropía y la información de máxima ganancia donde primero se relaciona con la incertidumbre o probabilidad de un evento y el segundo con la diferencia entre la entropía y una de sus opciones. El algoritmo comienza en un nodo raíz que tiene el conjunto de datos, luego el algoritmo busca el mejor

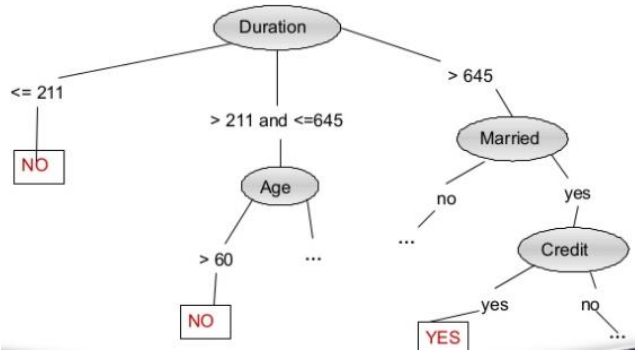
atributo de acuerdo con la información de ganancia que se está buscando, luego el conjunto se dividirá por esta decisión hasta obtener un nodo hoja que nos dé la clase o clasificación de esa información a través de una respuesta binaria (sí o no).



3.2.3 Árbol C4.5

Tiene la misma estructura del árbol del algoritmo ID3, con nodos, hojas y ramas, donde el curso de este es de acuerdo a la entropía y ganancia de información, comenzando a calcular el mejor atributo para dividir la información de entrenamiento, disminuye la entropía hasta obtener una decisión, porque mientras la iteración divide los datos en cada uno, será más fácil elegir la clase de información, ya que la ganancia de información es mayor en cuanto es menor la entropía; es decir, para cada atributo, calcula los datos potenciales mediante una prueba para obtener el mejor atributo para ramificar hasta el final del árbol.

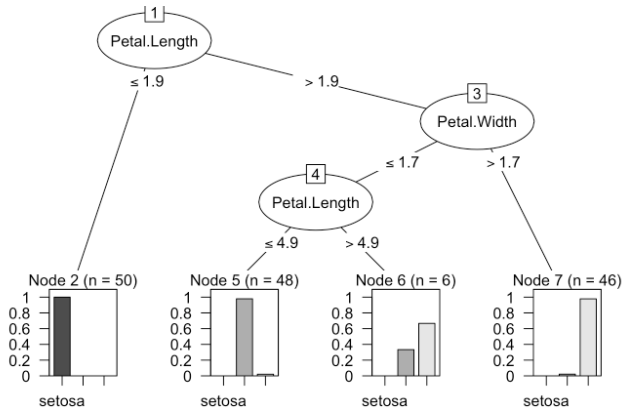
Una de las ventajas de este árbol es un proceso llamado poda que es útil para datos grandes y mejora la eficiencia, evitando un problema del algoritmo ID3 que es el sobreajuste.



3.2.4 Árbol C5.0

Su estructura es muy parecida a la de C4.5 y utiliza la misma información para construir el árbol, pero tiene una nueva implementación para mejorar la velocidad de procesamiento, la memoria y la eficiencia. Tiene una funcionalidad especial, la cual hace una preselección automática de los predictores más importantes y elabora un árbol sobre estos. También puede generar varios árboles y combinarlos para hacer una mejor predicción.

Es el estándar de la industria para la construcción de decisiones debido a su facilidad para resolver la mayoría de los problemas de inmediato y las facilidades que tiene para que los usuarios lo apliquen.



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

La estructura de datos usada es el árbol binario de decision, al realizer la entrada de los datos en este, va tomando un camino, por cada decisión opta entre dos caminos (de allí su nombre), sus partes son: raíz o nodo principal, ramificaciones, hojas-subhojas.

Figura 1: Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito.

[1]http://www.github.com/ ?????????/proyecto/

4.2 Algoritmos

El algoritmo empleado es el árbol CART, realizando un árbol de decisión, donde se clasificarán los datos cumpliendo con las condiciones relevantes, por medio del cálculo de la impureza de Gini e impureza ponderada de Gini, resaltando que entre menor sea esta impureza mejor será la precisión formado por estos conjuntos de datos.

4.2.1 Entrenamiento del modelo

Para llevar a cabo la preparación del árbol binario de decisión, se basa en tener en cuenta que tan cruzados se encuentran los datos determinando así la ganancia de la información, al emplear el cálculo de la impureza de Gini esto puede ser precisado. Al realizar el cálculo de esta impureza a los aspectos, se ira clasificando de acuerdo a su impureza, teniendo presente los aspectos que generen menor impureza y a la vez una menor impureza ponderada, para así realizar la mejor posible división de los datos cumpliendo con las principales condiciones y resultantes.

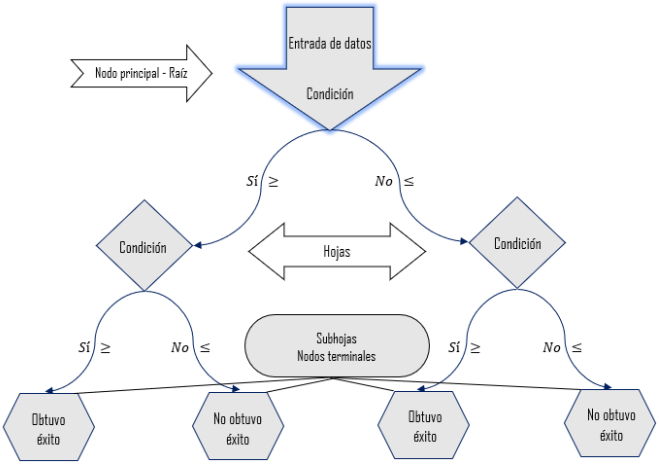


Figura 2: Entrenamiento y estructura fundamental de un árbol de decisión binario usando los datos de los resultados de los estudiantes en la pruebas saber 11 para predecir el éxito en las pruebas de educación superior.

4.2.2 Algoritmo de prueba

Se emplea el árbol para la realización de las predicciones hacía los estudiantes. El algoritmo de prueba es efectuado por el ingreso de los datos de más de 100000 estudiantes que realizaron las pruebas de Estado, comparando los resultados del modelo con los reales formalizando cuan preciso es el modelo delineado.

4.3 Análisis de la complejidad de los algoritmos

Para el cálculo de la complejidad en tiempo del entrenamiento se tiene en cuenta la organización que se tiene para la selección de las condiciones dependiendo el atributo, se realiza en O(NlogN), N es la cantidad de datos, esto se ejecuta según los aspectos que se tengan, y se repite para cada nodo del árbol, quedando así en $O(M^2*NlogN)$.

La validación del árbol los datos pasan por el árbol, son comparados para valida y tomar la decision, por cada nodo quedando así la complejidad en O(N*M)

Algoritmo	La complejidad del tiempo
Entrenar el árbol de decisión	$O(M^2*N*logN)$
Validar el árbol de decisión	O(N*M)

Tabla 1: Complejidad temporal de los algoritmos de entrenamiento y prueba.

Para calcular la complejidad de memoria de los algoritmos, se considera que los datos son almacenados en un marco de datos, de tamaño NxM, donde N son la cantidad de datos y M es la cantidad de aspectos resultantes, además el árbol al crearse ocupa ciertos nodos dependiendo de estos aspectos quedando de forma final para su entrenamiento como $O(N*M^2)$.

Algoritmo	Complejidad de memoria
Entrenar el árbol de decisión	$O(N*M^2)$
Validar el árbol de decisión	O(M)

Tabla 2: Complejidad de memoria de los algoritmos de entrenamiento y prueba.

4.4 Criterios de diseño del algoritmo

El algoritmo CART para el árbol de decision, fue empleado, debido los buenos resultados que trae, es muy eficiente en cuanto a velocidad al obtener los resultados y descartando puntos de poco influencia. Además, es muy cómodo para la implementación-aplicación de un árbol binario de decision.

5. RESULTADOS

5.1 Evaluación del modelo

En esta sección, presentamos algunas métricas para evaluar el modelo. La precisión es la relación entre el número de predicciones correctas y el número total de datos de entrada. Precisión. es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos identificados por el modelo. Por último, Sensibilidad es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos en el conjunto de datos.

5.1.1 Evaluación del modelo en entrenamiento

A continuación presentamos las métricas de evaluación de los conjuntos de datos de entrenamiento en la Tabla 3.

	Conjunto de datos 1	*Conjunto de datos 2*	*...Conjunto de datos n*
Exactitud	0.7	0.75	0.78
Precisión	0.7	0.75	0.79
Sensibilidad	0.7	0.73	0.75

Tabla 3. Evaluación del modelo con los conjuntos de datos de entrenamiento.

5.1.2 Evaluación de los conjuntos de datos de validación

A continuación presentamos las métricas de evaluación para los conjuntos de datos de validación en la Tabla 4.

	Conjunto de datos 1	*Conjunto de datos 2*	*...Conjunto de datos n*
Exactitud	0.67	0.7	0.73
Precisión	0.66	0.71	0.75
Sensibilidad	0.67	0.69	0.71

Tabla 4. Evaluación del modelo con los conjuntos de datos de validación.

5.2 Tiempos de ejecución

Calcular el tiempo de ejecución de cada conjunto de datos en Github. Medir el tiempo de ejecución 100 veces, para cada conjunto de datos, e informar del tiempo medio de ejecución para cada conjunto de datos.

	Conjunto de datos 1	*Conjunto de datos 2*	*...Conjunto de datos n*
Tiempo de entrenamiento	305.5 s	680.3 s	1002.4 s
Tiempo de validación	35.7 s	61.7 s	96.1 s

Tabla 5: Tiempo de ejecución del algoritmo *CART* para diferentes conjuntos de datos.

5.3 Consumo de memoria

Presentamos el consumo de memoria del árbol de decisión binario, para diferentes conjuntos de datos, en la Tabla 6.

	Conjunto de datos 1	*Conjunto de datos 2*	*...Conjunto de datos n*
Consumo de memoria	112 MB	432 MB	768 MB

Tabla 6: Consumo de memoria del árbol de decisión binario para diferentes conjuntos de datos.

6. DISCUSIÓN DE LOS RESULTADOS

Los resultados obtenidos fueron bastante buenos, se logró una muy buena precisión, exactitud y sensibilidad del algoritmo, no es necesario que el algoritmo fuera perfecto dado que las acciones que realizan los humanos son muy variables. En cuanto al consumo de tiempo y memoria establecemos que el apropiado eso es lo que ofrece dicho algoritmo. Por lo que se llega que este modelo se puede aplicar con certeza en la situación propuesta.

6.1 Trabajos futuros

La complejidad del algoritmo realizado no es de las más altas, pero este se podría mejorar, desarrollando más a fondo su estructura y funcionamiento, forma de trabajar al ingresar e interpretar los datos, podría ser aún mucho mejor implementando bosques aleatorios.

AGRADECIMIENTOS

Agradecemos a aquellos compañeros que nos dieron sugerencias para la mejora y solución de las situaciones que se iban presentando en el desarrollo de este proyecto.

REFERENCIAS

1. Bookdown.org. 6 Métodos de clasificación | Estadística y Machine Learning con R: https://bookdown.org/content/2274/metodos-de-clasificacion.html

2. José M. Sempere. Aprendizaje de árboles de decisión, from Universidad Politécnica de Valencia. http://www.academia.edu/download/43392762/decision.pdf

3. Blanco, V. Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos (Tesis de Maestría), from Universidad Nacional de Colombia. http://bdigital.unal.edu.co/51414/1/39004913.2015.pdf

4. García-González, J. R., Sánchez-Sánchez, P. A., Orozco, M., & Obredor, S. Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia https://scielo.conicyt.cl/scielo.php?pid=S0718-50062019000400055&script=sci_arttext&tlng=e

5. Carrascal, A. I. O., & Giraldo, J. J. Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO.

https://revistas.elpoli.edu.co/index.php/pol/article/view/1499

6. Álvarez Blanco, J., Lau Fernández, R., Pérez Lovelle, S., & Leyva Pérez, E. C. Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales. Ingeniare. https://scielo.conicyt.cl/scielo.php?pid=S0718-33052016000400015&script=sci_arttext&tlng=p