Algoritmo para predicción de éxito académico mediante árboles de decisión.

Cristian David Dávila García Universidad Eafit Colombia cddavi-lag@eafit.edu.c o	Andrés Rodríguez Barrientos Universidad Eafit Colombia arodri-gue2@eafit.edu.co	Miguel Correa Universi-dad Eafit Colombia ma-corream@eafit.edu.co	Mauricio Toro Uni-versidad Eafit Colom-bia mtorobe@ea-fit.edu.co

RESUMEN

A través del tiempo, la sociedad se ha desarrollado a un nivel en el que podemos y queremos hacer predicciones de nuestros resultados a lo largo de nuestra vida. Es por esto que hemos decidido comprobar cual será el nivel de éxito en las pruebas Saber Pro, teniendo en cuenta los resultados obtenidos de investigaciones en torno a los estudiantes que presentaron las pruebas Saber 11. Además, este problema nos ayuda a analizar si nuestros hábitos de vida son favorables para sacar buenos resultados en las pruebas Saber Pro. También, hablaremos de otros problemas más adelante.

El algoritmo propuesto para nuestro proyecto ha sido el árbol de decisión CART, sumando a este el uso de bosques aleatorios.

Palabras clave

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

1. INTRODUCCIÓN

Un día sentado en clase de lógica, hablando con mis compañeros, nos dimos cuenta que debíamos presentar unas pruebas finalizando la carrera universitaria, en ese momento nos preguntamos: "¿Me irá bien?". Y en ese momento se creó el problema, pero ahora, tenemos los recursos para resolverlo.

Entonces, con la predicción de los exámenes, podremos saber si debemos cambiar nuestro estilo de vida para lograr nuestro objetivo.

1.1. Problema

Al ser estudiantes, en los momentos previos de un examen nos preguntamos si nuestros métodos de estudio nos ayudarán en los futuros resultados.

Para esto se busca un algoritmo que, tomando diferentes datos del estudiante, pueda predecir su éxito académico. Esta predicción informa al estudiante de posibles cambios que pueda hacer en su manera de estudiar para lograr lo que quiere.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad.

Para la solución de este problema utilizamos el algoritmo CART, ya que sus resultados son fáciles de interpretar.

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

A continuación, veremos unos ejemplos relacionados con el tema

2.1 Arboles de decisión para predecir factores asociados al desempeño académico de los estudiantes de bachillerato en las pruebas saber 11

El problema era predecir patrones asociados al desempeño académico de los estudiantes colombianos que, encontrándose finalizando el grado undécimo de educación

media, presentaron las pruebas Saber 11° entre los años 2015 y 2016, a partir de la información socioeconómica, académica e institucional, almacenada en las bases de datos del ICFES, como solución se construyó un modelo de

clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta WEKA (*Waikato Environment for Knowledge Analysis*). Con esta herramienta, la precisión es proporcional al número de estudiantes, a mayor cantidad de estudiantes a evaluar, mayor es la precisión de la herramienta.

2.2 Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados

El objetivo de esta investigación era predecir la deserción de los estudiantes universitarios. Para esto, utilizaron un algoritmo llamado CBAD con parámetros optimizados utilizando el software RapidMiner. La precisión lograda fue del 87.27%, por lo que se concluyó que el uso de esta técnica optimizada, funcionó mejor que otras investigaciones.

2.3 Predicción del rendimiento académico aplicando tcnicas de minería de datos

El objetivo de esta investigación es predecir si los estudiantes de la UNALM en los semestres 2013 II y 2014 I van a tener una calificación final de aprobado o reprobado, esto, utilizando distintos métodos, tales como: Tecnicas de minería de datos de regresión logística, árboles de decisión, redes bayesianas y redes neuronales. Especialmente enfocados en la minería de datos. El método utlizado fue la matriz de confusión para comparar y evaluar la precisión de los clasificadores. Los resultados arrojaron un 71% de tasa de buena calificación

2.4 PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS

Se necesita tener el conocimiento previo de los estudiantes matriculados en una institución educativa para predecir su desempeño en futuras experiencias académicas. Los algoritmos usados fueron el ID3 y el C4.5. para un total de

182 estudiantes, el porcentaje promedio de precisión logrado en masa y Evaluaciones singulares es de aproximadamente 75.275%

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

Los árboles de decisión nos permiten resolver problemas de regresión o clasificación.

Los problemas de clasificación son todos aquellos en los que tenemos varios tipos de objetos y debemos separarlos o individualizarlos.

Los problemas de regresión se dan en su mayoría en el machine learning y la estadística, su mayor uso es para obtener relaciones entre 2 variables y escoger la mejor si es necesario.

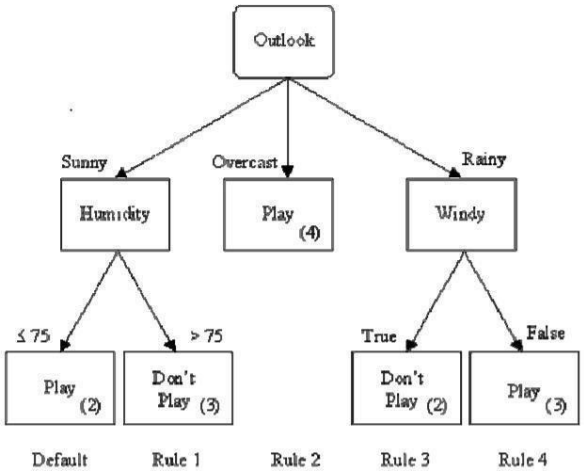
A continuación, hablaremos de unos tipos de algoritmos utilizados para resolver estos problemas.

3.2.1 ID3

"Para construir el árbol, el algoritmo utiliza el análisis de la entropía, la teoría de la información (basada en la entropía) y la ganancia de información." (IBM, s.f.)

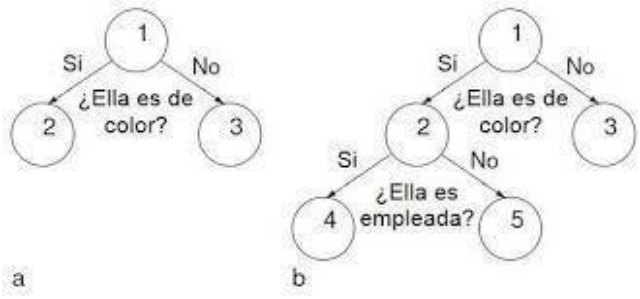
3.2.2 C4.5

El algoritmo C4.5 genera un árbol de decisión partiendo de divisiones ejecutadas recursivamente. Se realiza con el método de profundidad-primero (depth-first). Realiza diferentes pruebas posibles y selecciona la que al final le genere mayor ganancia al obtener información. Para cada atributo se realiza una prueba binaria realizando un número de pruebas igual al número de valores que puede tomar.



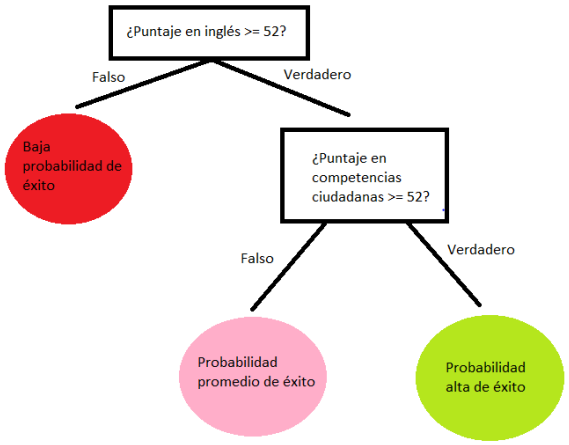
3.2.3 CART

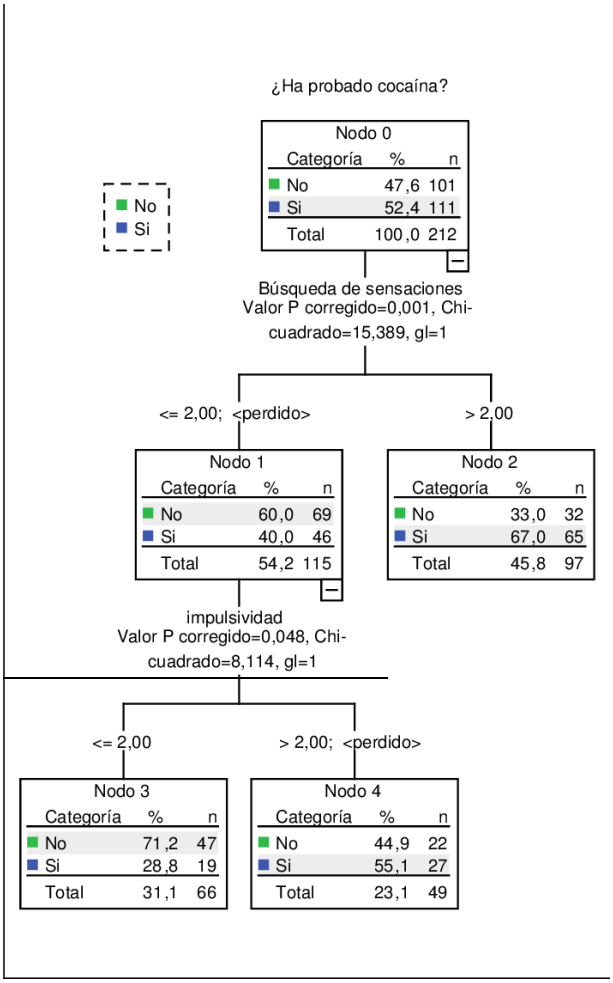
El método CART está basado en la impureza de Gini, funciona así: Comenzando por el nodo Principal se subdivide en 2 nodos hijos, se mide la impureza de estos resultados y el nodo hijo que tenga resultados más puros, será utilizado para continuar el árbol.



3.2.4 CHAID

Como mencionábamos antes, éste es un método de clasificación o selección que crea arboles mediante chi-cuadrado para encontrar divisiones o selecciones óptimas para el árbol. Tiene un nivel de dificultad bajo.





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

Un árbol de decisión binario es una estructura de datos que se basa en un proceso de decisiones que comienza desde la raíz dividida en dos nodos, repitiendo este proceso hasta que finaliza (en las hojas del árbol) con una operación tipo booleana que nos entrega la respuesta que buscamos.

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

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

4.2 Algoritmos

El algoritmo, toma los datos de los estudiantes que han realizado las pruebas SABER PRO anteriormente, y con estos, nos centramos en los que tuvieron éxito al realizar esta. Con esta información ya se tiene una base para seleccionar las variables que utilizaremos y poder predecir si los estudiantes que van a realizarla tendrán éxito.

4.2.1 Entrenamiento del modelo

Para entrenar nuestro modelo, utilizamos la mayor información posible y que el algoritmo inicie su ejecución tomando las diferentes decisiones. Haciendo que nuestro algoritmo entrene nuestro modelo. Al utilizar todos los datos, nuestro algoritmo se va a volver cada vez más preciso ya que estará calculando la impureza de cada una de las variables con las que contamos, y esto nos ayuda a encontrar el mejor camino posible y dar con la respuesta que buscamos.

Figura 2: En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

4.2.2 Algoritmo de prueba

Explique, brevemente, cómo probó el modelo: Esto equivale a explicar cómo su algoritmo clasifica los nuevos datos después de que se construya el árbol.

4.3 Análisis de la complejidad de los algoritmos

Explique en sus propias palabras el análisis para el peor caso usando la notación O. ¿Cómo calculó tales complejidades.

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

Tabla 2: Complejidad temporal de los algoritmos de entrenamiento y prueba. *N representa el número de filas. M representa el número de columnas.*

Algoritmo	Complejidad de memoria
Entrenar el árbol de decisión	O(N*M*2^M)
Validar el árbol de decisión	O(1)

Tabla 3: Complejidad de memoria de los algoritmos de entrenamiento y prueba. *N representa el número de filas. M representa el número de columnas.*

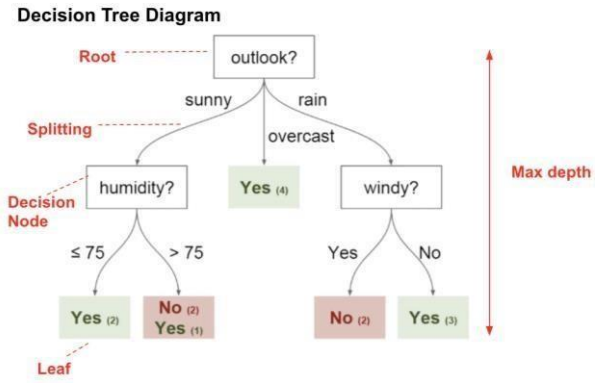
4.4 Criterios de diseño del algoritmo

Realizamos el Proyecto con el algoritmo CART con la impureza de Gini, porque este algoritmo nos ofrece una muy buena complejidad en tiempo, siendo el mejor si hablamos en términos de clasificación que el resto. Asimismo, los valores perdidos no influiran en el modelo, y en consecuencia será más fácil de entender.

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	0.79	0.75	0.78
Precisión	0.75	0.73	0.73
Sensibilidad	0.8	0.81	0.8

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	0.78	0.78	0.78
Precisión	0.75	0.72	0.73
Sensibilidad	0.79	0.81	0.81

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

5.2 Tiempos de ejecución

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Tiempo de entrenamiento	40 s	178 s	371 s
Tiempo de validación	9 s	63 s	122 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	66 MB	320 MB	447 MB

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

6. DISCUSIÓN DE LOS RESULTADOS

Observando las anteriores tablas, Podemos ver que en cuanto la precision, sensibilidad y exactitude está muy bien el algoritmo ya que se encuentra en los resultados que se esperaban al iniciar este proyecto. Aunque, el consume de memoria es algo que se tiene que mejorar, como vemos en la table 6, los resutados fueron muy altos. El tiempo de ejecución fue rápido considerando la cantidad de datos con los que se trabajaron.

6.1 Trabajos futuros

Nos gustaría optimizar el algoritmo, lo suficiente como para mantener el tiempo de ejecución igual a este y a la vez bajar el consumo de memoria.

AGRADECIMIENTOS

Nos gustaría agradecer a Simón Marín Giraldo y Daniel Alejandro Mesa Arango, estudiantes de Ingeniería de Sistemas y monitores de Estructura de Datos y Algoritmos I, quienes nos ayudaron y explicaron con todo lo relacionado al proyecto durante el desarrollo.