Predicción del rendimiento de un estudiante por medio de un árbol de decisiones

Juan Felipe Ortiz Salgado	Tomas Duque Giraldo	Miguel Correa	Mauricio Toro
Universidad Eafit	Universidad Eafit	Universidad Eafit	Universidad Eafit
Colombia	Colombia	Colombia	Colombia
jfortizs@eafit.edu.co	tduqueg@eafit.edu.co	macorream@eafit.edu.co	mtorobe@eafit.edu.co

Para cada versión de este informe: 1. Detalle todo el texto en rojo. 2. Ajustar los espacios entre las palabras y los párrafos. 3. Cambiar el color de todos los textos a negro.

Texto rojo = Comentarios

Texto negro = Contribución de Miguel y Mauricio

Texto en verde = Completar para el 1er entregable

Texto en azul = Completar para el 2º entregable

Texto en violeta = Completar para el tercer entregable

RESUMEN

El objetivo de este informe es analizar y predecir el éxito académico de un estudiante al realizar las pruebas Saber Pro. Lo que se busca es diseñar un algoritmo basado en árboles de decisión para obtener un puntaje total superior al promedio al momento de realizar las pruebas. Existen varios problemas similares al que se plantea en este informe y aunque es poco lo que se ha realizado para predecir el éxito en educación superior, algunos de estos serán analizados con el propósito de hallar una solución más efectiva.

¿Cuál es el algoritmo propuesto? ¿Qué resultados obtuvieron? ¿Cuáles son las conclusiones de este trabajo? El resumen debe tener como máximo **200 palabras**. (*En este semestre, usted debe resumir aquí los tiempos de ejecución, el consumo de memoria, la exactitud, la precisión y la sensibilidad*)

Palabras clave

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

1. INTRODUCCIÓN

Actualmente el papel de la tecnología es un factor clave en el proceso de la transformación digital de la educación. En el pasado, se han estudiado problemas tales como las causas y motivaciones que influyen en la deserción académica, con la cual a partir de la tecnología se han realizado algoritmos con diferentes factores que influyen en que ocurra esta. Sin embargo, es poco lo que se ha logrado para predecir el éxito académico en educación superior.

1.1. Problema

El problema al cual nos enfrentamos se basa en crear a través de árboles de decisión, un algoritmo para predecir el éxito académico de las pruebas Saber pro.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad (¡falta *una* cita *para este argumento!*). Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad (¡Falta *una* cita para *este argumento!*).

Explique, brevemente, su solución al problema *(En este semestre, la solución es una implementación de un algoritmo de árbol de decisión para predecir el éxito académico. ¿Qué algoritmo elegiste? ¿Por qué?)*

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)*

3.1 Un algoritmo de árbol de decisiones relacionado con el análisis y la predicción del desempeño de los estudiantes.[1]

Se buscaba por medio un muy conocido algoritmo de minería de datos poder predecir la nota de los estudiantes de pregrado de ingeniería donde 4 arboles de decisiones fueron comparados donde después de hacer una comparación entre los cuatro se encontró que el árbol con mayor exactitud, para ser más específicos una exactitud del 80.15% fue el arbol de decisión J48, que es la versión de java del c4.5 lo que indica que este modelo es bueno para predecir las notas de los estudiantes por su gran porcentaje de exactitud.

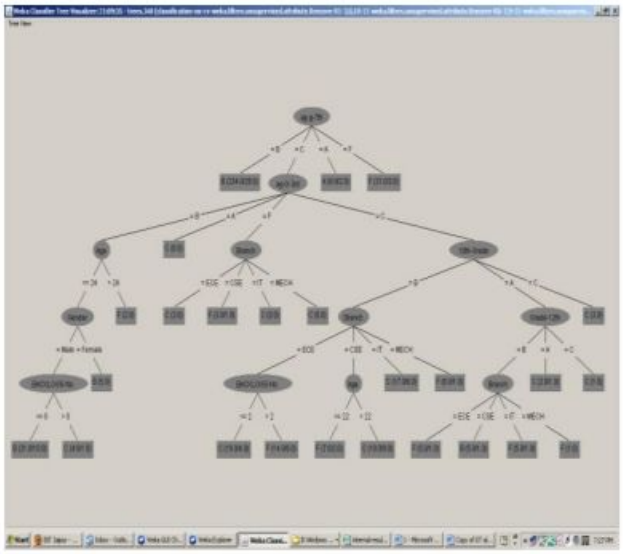
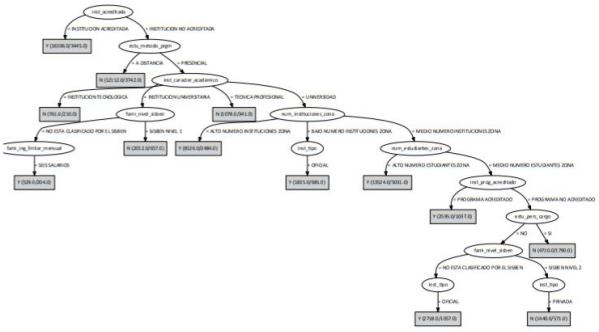


Fig 1: Decision Tree Construction

3.2 Un título para el segundo problema relacionado

En este caso se buscaba definir un patrón específico de desempeño académico en competencias genéricas de los estudiantes de programas profesionales. Únicamente aplicando las técnicas y algoritmos de minería de datos, que en el caso de esta investigación es clasificación por árboles de decisión.



https://repository.ucc.edu.co/handle/20.500.12494/1039

3.3 rendimiento académico.

en este se buscaron modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios. Para la formulación y ajustes de los modelos de predicción, se utilizaron alternativamente, técnicas de minería de datos clásicas y métodos simbólicos o inteligentes, evaluando su desempeño en la predicción del rendimiento académico de los alumnos.

http://sedici.unlp.edu.ar/handle/10915/19846

3.4 Algoritmo para predecir tensiones con técnicas de inteligencia artificial en una tibia humana.

en este problema se buscó crear un algoritmo que permita dar solución al problema de remodelación ósea de una tibia humana bajo diferentes valores de cargas mecánicas.se utilizó el Método de los Elementos Finitos. Se usó el software profesional ABAQUS/CAE para el cálculo de tensiones y deformaciones y una red neuronal para el procesamiento de los valores obtenidos. a partir del uso de las técnicas de inteligencia artificial y con el empleo del método de los elementos finitos, fue posible obtener un modelo que pronosticò las magnitudes de tensiones.

https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=63994

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

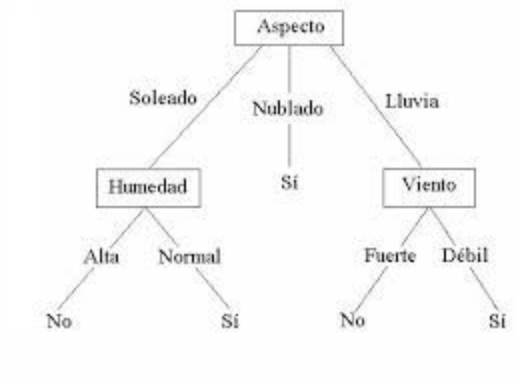
Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

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).*

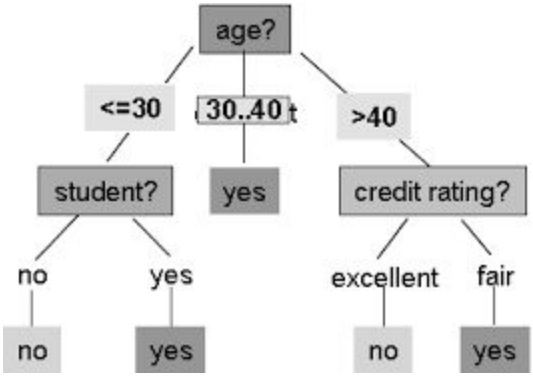
3.2.1 Algoritmo ID3

Es un algoritmo inventado por Ross Quinlan que se utiliza para generar un árbol de decisiones a partir de un conjunto de datos. ID3 es el precursor del algoritmo C4.5 y normalmente se utiliza en los dominios de aprendizaje automático y procesamiento del lenguaje natural .ID3 es más difícil de usar en datos continuos que en datos factorizados.



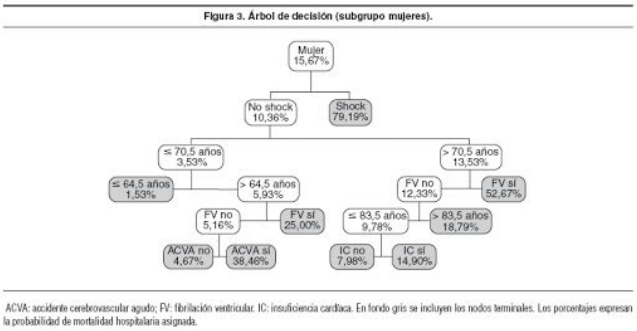
3.2.2 Algoritmo c4.5

Desarrollado a su vez por Ross es una extención del ID3 El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero. Entre sus principales diferencias con ID3 se encuentran: Evitar Sobreajuste de los datos , determinar que tan profundo debe crecer el árbol de decisión y reducir errores.



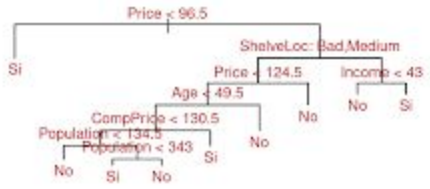
3.2.3 Algoritmo Cart

El algoritmo CART es el acrónimo de Classification And Regression Trees este modelo admite variables de entrada y de salida nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión.



3.2.4 Algoritmo c5.0

C5.0 es el algoritmo sucesor de C4.5, creado por la misma persona. Entre sus características se encuentran la capacidad para generar árboles de predicción simples, modelos basados en reglas y asignación de distintos pesos a los errores.



4. DISEÑO DE LOS ALGORITMOS

En lo que sigue, explicaremos el diseño del algoritmo que realizamos

4.1 Estructura de los datos

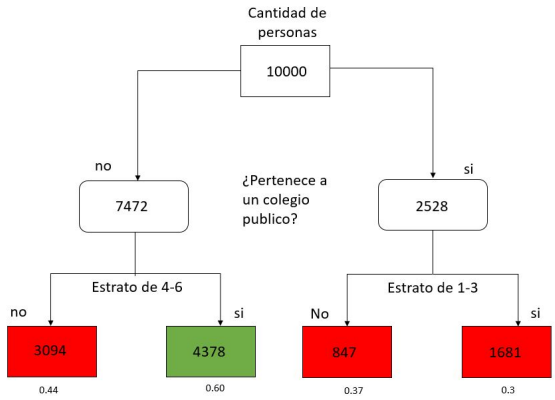
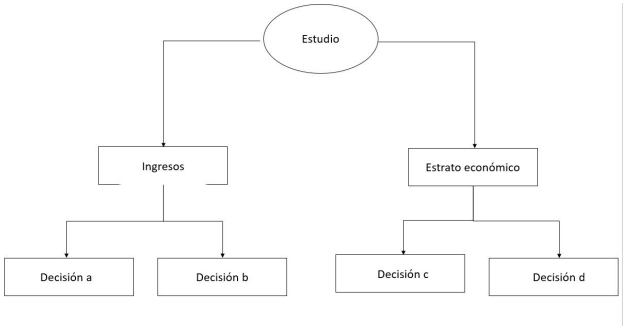


Figura 1: Representación de el arbol de decisión binario dando una idea de la estructura de datos utilizada para poder encontrar la impureza de Gini lo que nos permite saber la excelencia en las pruebas saber pro

4.2 Algoritmos

Diseñamos un algoritmo que será capaz de almacenar gran cantidad de datos, el algoritmo en un principio organizará a los estudiantes, y realizando una serie de calculos el algoritmo será capaz de calcular con precisión la impureza de y si está impureza está muy baja las probabilidades de excelencia en las pruebas saber será mucho mayor

4.2.1 Entrenamiento del modelo



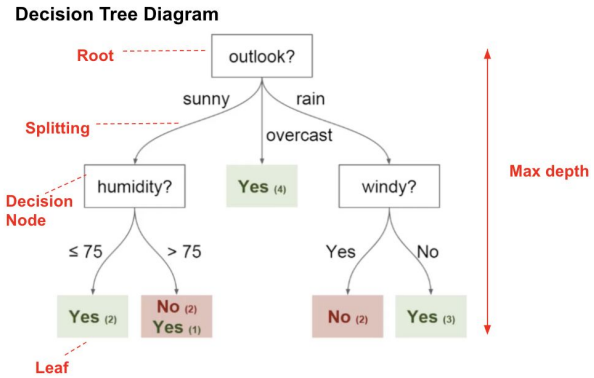


Figura 2: El arbol de decisión que decidimos utilizar es el arbol tipo CART, arbol que entrenaremos de forma que tenga la capacidad de predecir los resultados y que concuerden con lo que se espera, en la figura 2 podemos apreciar como el arbol tomará diferentes decisiones dependiendo de diversos factores como lo son el estrato social, el estudio entre otros y ya con estos datos el arbol será capaz de crear una produccion acertada

4.2.2 Algoritmo de prueba

Cada vez que se crea un nodo el algoritmo realiza un criterio nuevo que permite la creación de más de ellos y una clasificación más acertada de los datos. Gracias a esto se puede crear un arbol con mayor precisión y las decisiones se basan netamente en la homogeneidad de los nodos que se van creando.

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(N2*M2)
Validar el árbol de decisión	O(N3*M*2N)

Tabla 2: Complejidad temporal de los algoritmos de entrenamiento y prueba. *(Por favor, explique qué significan N y M en este problema.)*

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

Tabla 3: Complejidad de memoria de los algoritmos de entrenamiento y prueba. *(Por favor, explique qué significan N y M en este problema.)*

4.4 Criterios de diseño del algoritmo

Explica por qué el algoritmo fue diseñado de esa manera. Use un criterio objetivo. Los criterios objetivos se basan en la eficiencia, que se mide en términos de tiempo y consumo de memoria. Ejemplos de criterios no objetivos son: "Estaba enfermo", "fue la primera estructura de datos que encontré en Internet", "lo hice el último día antes del plazo", etc. Recuerde: Este es el 40% de la calificación del proyecto.

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	0.9
Precisión	0.7	0.75	0.9
Sensibilidad	0.7	0.75	0.9

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	0.5	0.55	0.7
Precisión	0.5	0.55	0.7
Sensibilidad	0.5	0.55	0.8

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	10.2 s	20.4 s	5.1 s
Tiempo de validación	1.1 s	1.3 s	3.3 s

Tabla 5: Tiempo de ejecución del algoritmo *(Por favor, escriba el nombre del algoritmo, C4.5, ID3)* 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	10 MB	20 MB	5 MB

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

Para medir el consumo de memoria, debería usar un generador de perfiles (*profiler*). Uno muy bueno para Java es VisualVM, desarrollado por Oracle, http://docs.oracle.com/javase/7/docs/technotes/guides/visualvm/profiler.html. Para Python, use C-profiler.

6. DISCUSIÓN DE LOS RESULTADOS

Explique los resultados obtenidos. ¿Son la precisión, exactitud y sensibilidad apropiadas para este problema? ¿El modelo está sobreajustado? ¿Es el consumo de memoria y el consumo de tiempo sib apropiados? *(En este semestre, de acuerdo con los resultados, ¿se puede aplicar esto para dar becas o para ayudar a los estudiantes con baja probabilidad de éxito? ¿Para qué es mejor?)*

6.1 Trabajos futuros

Respuesta, ¿qué le gustaría mejorar en el futuro? ¿Cómo le gustaría mejorar su algoritmo y su implementación? ¿Qué hay de usar un bosque aleatorio?

AGRADECIMIENTOS

Identifique el tipo de agradecimiento que quiere escribir: Para una persona o para una institución. Considere las siguientes pautas: 1. El nombre del profesor no se menciona porque es un autor. 2. No debe mencionar sitios web de autores de artículos que no haya contactado. 3. Debe mencionar estudiantes y profesores de otros cursos que le hayan ayudado.

Como ejemplo: Esta investigación fue apoyada parcialmente por [Nombre de la Fundación, Donante].

Agradecemos la asistencia con [técnica particular, metodología] a [nombre apellido, cargo, nombre de la institución] por los comentarios que mejoraron enormemente el manuscrito.

REFERENCIAS

La referencias se hacen con el formato de referencias de la ACM. Lea las directrices de ACM en http://bit.ly/2pZnE5g

A modo de ejemplo, consideremos estas dos referencias:

1.Pandey, M. y Sharma, VK (2013). Un algoritmo de árbol de decisiones relacionado con el análisis y la predicción del desempeño de los estudiantes. Revista internacional de aplicaciones informáticas , 61 (13).

2. Fischer, G. y Nakakoji, K. Amplificando la creatividad de los diseñadores con entornos de diseño orientados al dominio. en Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.

3. Quinlan, JR 1986. Inducción de árboles de decisión. Mach. Aprender. 1, 1 (marzo de 1986), 81–106

https://bookdown.org/content/2274/metodos-de-clasificacion.html

https://repository.ucc.edu.co/handle/20.500.12494/1039

https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=63994

http://sedici.unlp.edu.ar/handle/10915/19846

https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting