ALGORITMOS BASADOS EN ÁRBOLES DE DECISIÓN Y EN DATOS PRUEBA SABER 11

Miguel Ángel Martínez Flórez Universidad Eafit Colombia mamartinef@gmail.com	Pablo Maya Villegas Universidad Eafit Colombia pmayav@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

El objetivo de este informe es analizar y solucionar el problema de diseño de un algoritmo cuya principal función es gestionar un manejo basado en árboles de decisión y en los datos del Saber 11, para predecir si un estudiante tendrá un puntaje total, en las pruebas Saber Pro, por encima del promedio o no.

En un futuro cercano, el papel de la tecnología será un factor clave en el proceso de transformación digital de la educación en Colombia. Esta transformación se conoce como Educación 4.0. En el pasado, se han estudiado qué factores influyen en la deserción académica, cuáles son sus causas y motivaciones, y se han utilizado algoritmos para predecir la deserción

No obstante, es poco lo que se ha logrado para predecir el éxito académico en educación superior. El éxito puede medirse de muchas formas; por ejemplo, la empleabilidad del egresado, el salario de los egresados, la felicidad del trabajo de los egresados, entre otros.

Palabras clave

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

1. INTRODUCCIÓN

Gracias a que nos encontramos actualmente en una era donde la informática es más que indispensable, se hace imprescindible el factor de desarrollar herramientas informáticas nuevas y revolucionarias que permitan de manera eficiente, tener un mejor manejo en gestión de volúmenes con alta información.

Este es el caso de un sistema capaz de obtener cierto resultado a base de los datos del Saber 11, para predecir si un estudiante tendrá un puntaje total, en las pruebas Saber Pro, por encima del promedio o no.

1.1. Problema

A través del análisis de algoritmos y árboles de decisión , llegar a implementar este sistema resaltará las predicciones de los exámenes de la prueba saber 11, incluyendo temas como el éxito académico en los pregrados de América latina. .

1.2 Solución

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

Para solucionar el problema utilizaremos un árbol CART, este va a separar los datos según sus características más importantes y ver cómo estas afectan el promedio final del Saber 11 para así identificar cuáles son las que verdaderamente influyen.

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

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

En este sistema se presentan los resultados obtenidos al aplicar el modelo de clasificación basado en árboles de decisión, con el fin de detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo, los resultados son obtenidos de las pruebas saber 11 en los años 2015 y 2016.

la investigación es de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. con ayuda de la metodología CRISP-DM se seleccionó, de las bases de datos del ICFES.Concluyeron con patrones descubiertos que ayudaran en los procesos de toma de decisiones del ministerio de educación nacional, junto con las instituciones que velen por la calidad de la educación en colombia .

3.2 SABER 11 y rendimiento universitario: un análisis del progreso en el plan de estudios

Este artículo emplea como elemento de reflexión las pruebas icfes de egreso de la educación secundaria y su posible influencia en los posibles desempeños universitarios de los estudiantes. Este sistema se realizó inicialmente con un barrido sobre algunas pruebas existentes a nivel mundial y algunos estudios que han intentado atribuir cierta importancia en sus resultados.

Anteriormente, se presentó una propuesta metodológica que intenta describir las condiciones de entrada de los estudiantes que ingresaron a una universidad colombiana en un periodo de tiempo, y a través de una análisis se proyecta poder observar su progreso académico. Se concluye que a partir de asociaciones entre los desempeños en pruebas de estado y los posteriores rendimientos académicos, siendo este último un análisis a través del tiempo.

3.3 ¿Cuál es la brecha de la calidad educativa en Colombia en la educación media y en la superior?

inicialmente el objetivo de encontrar cual es la brecha de la calidad de la educación colombiana en los niveles medio y superior, se examinaron los resultados de las pruebas saber 11 saber pro a través de modelos jerárquicos en los que se contrastaron factores individuales, familiares,etc. con ayuda del análisis de brecha basado en la desigualdad entre planteles y otro basado en la eficacia escolar.

No obstante , se compararon los hallazgos entre áreas del conocimiento, carreras profesionales y departamentos. teniendo resultados como que el 11% de las variaciones del puntaje en la educación media y el 27,8% en la educación superior se explican por las diferencias entre planteles, y que la mayor parte de estas se debe a los factores individuales.

3.4 Factores socioeconómicos y educativos asociados con el desempeño académico, según nivel de formación y género de los estudiantes que presentaron la prueba SABER PRO 2009

La presente investigación estudia el poder predictivo del puntaje de admisión a la universidad y de otras variables relacionadas con el proceso de admisión, sobre el desempeño académico del estudiante, representado en el promedio de la carrera.

El puntaje de admisión muestra una baja correlación, aunque significativa, con el desempeño en la universidad. Además emergen otras variables que influyen favorablemente en el buen desempeño, tales como el hecho de ser mujer, el ingresar joven a la universidad, el provenir de estratos más altos y aplicar a políticas de admisión en donde se reconoce el historial académico del aspirante.

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 graduados 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 cómo 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 usados para construir automáticamente un árbol de decisión binario. *(En este semestre, ejemplos de tales algoritmos son ID3, C4.5 y CART).*

3.2.1 CART

Los árboles de clasificación y regresión Cart son una alternativa al análisis tradicional de clasificación y/o discriminacion a la predicción tradicional (regresión). Entre las ventajas de estos árboles Cart podemos destacar su robustez a outliers, la invarianza en la estructura de sus árboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad.

3.2.2 ID3

El modelo de clasificación, también conocido como ID3, significa (inducción del árbol de decisiones), Un sistema de aprendizaje supervisado que utiliza una estrategia de clasificación "divide y vencerás", implementa métodos y técnicas para realizar procesos inteligentes que representan conocimiento y aprendizaje, automatizando así las tareas. Hoy en día es común escuchar que la palabra "inteligencia" o algunos de sus sinónimos se aplican a máquinas, sistemas, procesos e incluso productos del hogar que son ampliamente conocidos en todo el mundo. Su origen se debe a la aparición de la inteligencia artificial y los sistemas expertos.

3.2.3 C4.5

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 (depth-first). El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos.

3.2.4 CLS

Un árbol de decisiones es un diagrama de los posibles resultados de una serie de decisiones relacionadas. Permite a las personas u

organizaciones comparar posibles acciones en función de sus costos, probabilidades y beneficios. Se pueden utilizar para llevar a cabo una lluvia de ideas informal o para diseñar un algoritmo para predecir matemáticamente la mejor opción. Los árboles de decisión generalmente comienzan desde un solo nodo y luego se ramifican hacia posibles resultados. Cada uno de estos resultados crea otros nodos, que se pueden dividir en otras posibilidades. Esto le da una forma de árbol.

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 que vamos a usar para este proyecto, es un árbol de decisión ID3, porque es capaz de tomar decisiones con gran precisión. . por ejemplo, nuestro tema principal en la predicción es diseñar un algoritmo basado en árboles de decisión, para predecir si un lote de café caturra está infectado con la roya o no; con ayuda de las principales variables que vamos a usar como lo son : iluminación, temperatura ambiental, humedad del suelo, humedad ambiental, temperatura del suelo, ph del suelo; tenemos no solo la información necesaria para tener una predicción completamente eficiente si no que tenemos muchos otros factores los cuales podemos utilizar en la predicción.

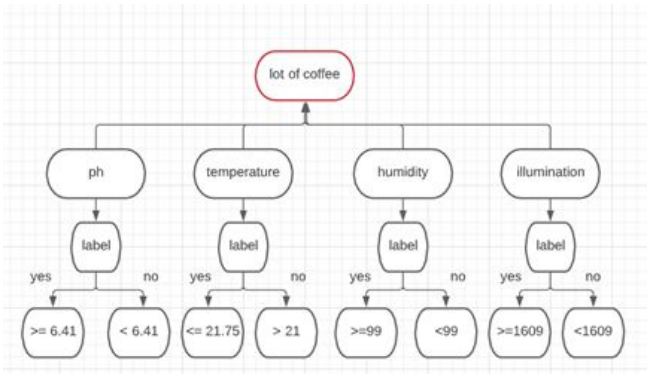

Figura 1: Ejemplo de el uso de los valores para la creación del árbol.

4.2 Algoritmos

Se testean varios intervalos en los valores del modelo y se selecciona el que tenga el valor gini ponderado más bajo, usando este método llegamos a que cuando la humedad del suelo es mayor a 67 entonces significa que esas plantas están aseguradas de no sufrir la enfermedad de roya (68), se divide el árbol y quedamos con un nuevo nodo de 304 valores para buscar el siguiente valor que pueda disminuir su impureza.

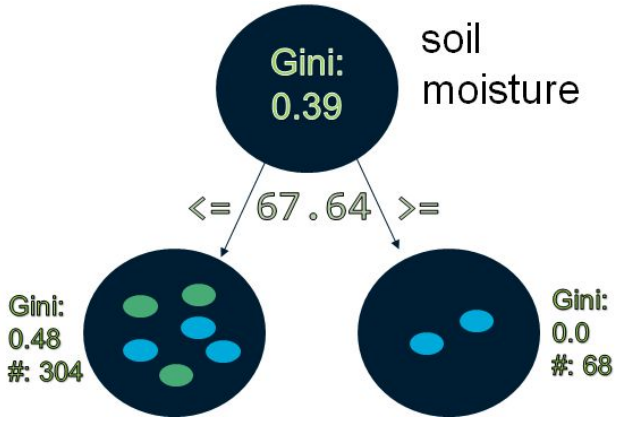

Figura 2: Una separación de nodo mostrando que valores se tomaron en cuenta.

4.2.1 Entrenamiento del modelo

El algoritmo va a probar cuál valor es el más efectivo para separar el árbol, y cuando lo separe va a tomar los hijos impuros y a repetir el proceso, va a guardar el orden de las instrucciones y así cuando reciba otro archivo de datos podrá recrearlas para predecir con precisión.

4.2.2 Algoritmo de prueba

Para probar que el prototipo del algoritmo está funcionando comparamos los valores del gini ponderado, el gini de los nodos hijos, el punto de inflexión de los hijos y la cantidad de valores con que cada hijo quedó. Primero probamos de entre los 6 valores el que tenia el gini más bajo era la humedad y que un nodo hijo tenía 0 impureza, esto puede suceder si el algoritmo le da 0 valores, pero fue diseñado para que no pasará, entonces este nodo hijo con 0 impureza es por que tiene 68 valores los cuales son iguales, esta decisión fue correcta por que en la primera inflexión se filtra 18% de los valores.

4.3 Análisis de la complejidad de los algoritmos

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

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

El algoritmo fue elegido por su simpleza y flexibilidad, para poder tomar en cuenta tantos valores distintos y con tantas implicaciones la mejor opción es tomar un acercamiento general y eficiente. Al plazo de desarrollar el proyecto se va a refinar y mejorar el algoritmo pero se buscará la eficiencia como principal objetivo.

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 preajustado? ¿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.*Adobe Acrobat Reader 7, Asegúrate de que el texto de las secciones de referencia es está alíneado a la derecha y no justificado. http://www.adobe.com/products/acrobat/.*

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.

3.1 Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. REVISTA DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN, 9(2), 363-378.

3.2 Gabalán-Coello, Jesús; Vásquez-Rizo, Fredy Eduardo SABER 11 y rendimiento universitario: un análisis del progreso en el plan de estudios Ciencia, Docencia y Tecnología, vol. 27, núm. 53, noviembre, 2016, pp. 135-161 Universidad Nacional de Entre Ríos Concepción del Uruguay, Argentina

3.3 Marly Tatiana Celis, Óscar Andrés Jiménez y Juan Felipe Jaramillo Maestría en Economía, Universidad de Manizales