Proyecto para predecir la excelencia de las pruebas saber pro

Juan Andres Henao Diaz	Carlos Andres Mosquera	Miguel Correa	Mauricio Toro
Universidad Eafit	Universidad Eafit	Universidad Eafit	Universidad Eafit
Colombia	Colombia	Colombia	Colombia
jahenaod@eafit.edu.co	camosquerp@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 problema que se plantea en este proyecto es observar los resultados de las pruebas saber Icfes de un número determinado de estudiantes y así por medio de ciertos algoritmos predecir su éxito académico en las futuras pruebas saber pro. La importancia de este es darse una idea gracias a los algoritmos si se darán unos resultados satisfactorios para los estudiantes que los presentaran en un futuro.

¿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

El papel de la tecnología en la educación ha sido muy importante en nuestra vida cotidiana y gracias a esta y a los programadores que se han encargado de hacer ciertos algoritmos se ha podido estudiar mas a fondo los motivos que causan la deserción académica, prediciendo resultados a futuro tomando en cuenta resultados y así identificar lo que ha ocurrido con certeza

1.1. Problema

El problema que ocurre es gracias a un algoritmo diseñado por nosotros que se encargue de predecir el rendimiento académico de los estudiantes que realizaran las pruebas saber tomando en base sus pasadas puntuaciones en las pruebas saber 11°

1.2 Solución

en este proyecto decidimos usar y enfocarnos en los arboles de decisión ya que estos son muy eficientes en su buena aplicación y a la hora de jugar con los datos y operaciones matemáticas con ellos, son bastante fáciles de explicar y desarrollar su codificación, hemos utilizado principalmente métodos denominados como caja blanca porque lo métodos apuestos, es decir, los de caja negra, carecen de explicación lo que es tedioso para sustentarse. El árbol de decisión que elegimos es el llamado CART, que es muy bueno y resulta menos tedioso de explicar que los otros existentes además de su buena maleabilidad con todos los tipos de datos primitivos

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 Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional.

El objetivo de este estudio fue descubrir patrones de desempeño académico en competencias genéricas de los estudiantes de programas profesionales en las pruebas Saber Pro-2011-2, a partir de los datos sociodemográficos, económicos, académicos e institucionales almacenados en las bases de datos del Instituto Colombiano para la Evaluación de la Educación (icfes), y utilizando técnicas de minería de datos. Los estudios realizados hasta el momento con respecto al análisis de los resultados de las pruebas Saber Pro-2011-2 se basan en información procesada mediante una investigación estadística

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

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Siguiendo la metodología CRISP-DM, se seleccionó, de las bases de datos del ICFES, la información socioeconómica, académica e institucional de estos estudiantes. Se construyó, limpió y transformó un repositorio de datos y utilizando la herramienta de minería de datos WEKA, se generaron árboles de decisión que permitieron identificar patrones asociados al buen o mal desempeño académico de los estudiantes en las pruebas Saber 11°.

3.3 Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios

Este proyecto tiene por objetivo construir modelos predictivos del rendimiento académico de los estudiantes de las diversas carreras de la FACENA de la UNNE. Las variables para incorporar en los modelos serán seleccionadas de acuerdo a los resultados obtenidos a partir de los siguientes análisis: a) Resultados del test de diagnóstico de conocimientos matemáticos previos; b) Condiciones socioeconómicas de los alumnos de las distintas carreras y datos obtenidos de encuesta directa a los alumnos de primer año.

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

Objetivo: crear un algoritmo que permita dar solución al problema de remodelación ósea de una tibia humana bajo diferentes valores de cargas mecánicas.
Métodos: se empleó 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. La red neuronal fue establecida; se aplicó el software MATLAB R2013a.

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

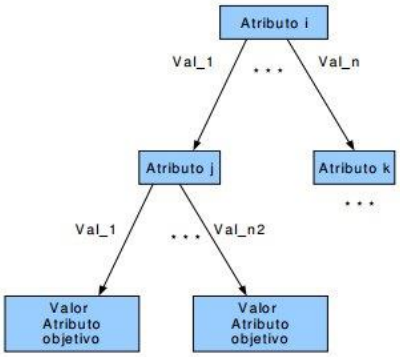
Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

3.2 Alternativas de algoritmos de árbol de decisión

En lo que sigue, presentamos diferentes algoritmos 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 Algoritmo 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 n 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.



3.2.2 Árboles B

Los árboles B son estructuras de datos de árbol que se encuentran comúnmente en las implementaciones de bases de datos y sistemas de archivos. Al igual que los árboles binarios de búsqueda, son árboles balanceados de búsqueda, pero cada nodo puede poseer más de dos hijos. Los árboles B mantienen los datos ordenados y las inserciones y eliminaciones se realizan en tiempo logarítmico amortizado.

Propiedades básicas

1. Cada nodo tiene como máximo M hijos.
2. Cada nodo (excepto raíz) tiene como mínimo (M)/2 claves.
3. La raíz tiene al menos 2 hijos si no es un nodo hoja. (según M)
4. Todos los nodos hoja aparecen al mismo nivel.
5. Un nodo no hoja con k hijos contiene k-1 elementos almacenados.

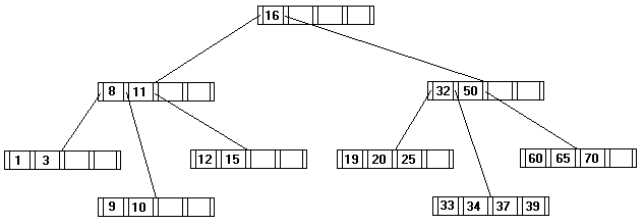
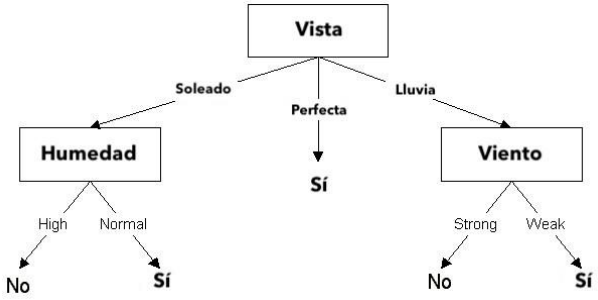


Figura 2:Inserción de un nuevo elemento en un B-árbol

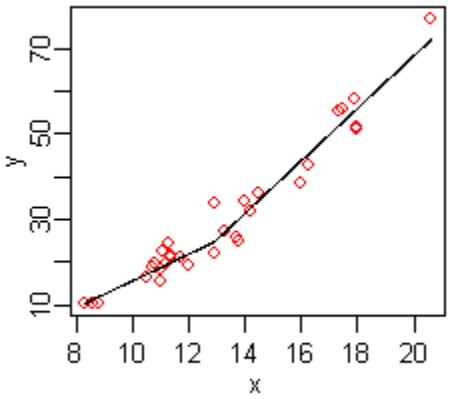
3.2.3 Arbol ID3

El ID3 permite determinar el árbol de decisión mínimo, para un conjunto de objetos. Este árbol permite que la información se mantenga en forma organizada y entendible para cualquier persona, además hace uso de una secuencia de preguntas, donde cada una de las preguntas es evaluada con el propósito de obtener la mejor respuesta posible.



3.2.4 Arboles Mars

Los arboles mars consisten en Reemplazar la división discontinua en un nodo con una transición modelada por un par de líneas directas. Al final del proceso de construcción del modelo, las líneas directas en cada nodo son reemplazadas con una función libre de obstáculos.
No requiere que nuevas divisiones dependan de divisiones antecesoras.



4. DISEÑO DE LOS ALGORITMOS

El presente diseño del algoritmo se encuentra en nuestro repositorio de GitHub donde ser encuentran todo los códigos e informes para el día de la entrega del proyecto, el proyecto fue hecho por Juan Andres Henao Diaz y Carlos Andres Mosquera

4.1 Estructura de los datos

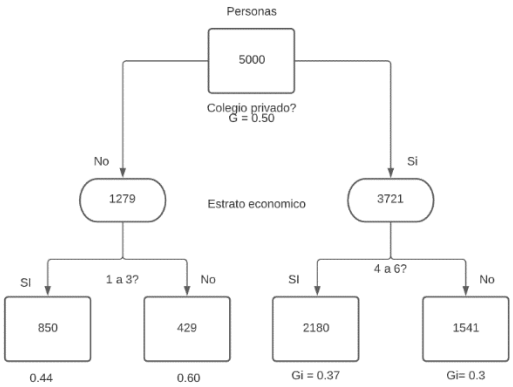


Figura 1: Este diagrama hace una pequeña noción de la estructura de datos utilizada para encontrar la impuresa de Gini y así encontrar de una manera precisa la impuresa de Gini para cada caso, lo que se convierte en saber la excelencia en las pruebas saber pro

4.2 Algoritmos

Los algoritmos diseñados por nosotros serían capaces de almacenar gran cantidad de datos (los datasets más grandes del proyecto) el algoritmo empieza organizando los datos, es decir, los estudiantes bajo ciertos criterios vistos en la figura 1 como es su colegio o estrato social lo que hace que el árbol binario calcule con certeza la impureza de Gini y entre mas baja este esta, mas probabilidad de excelencia académica en las pruebas saber pro.

4.2.1 Entrenamiento del modelo

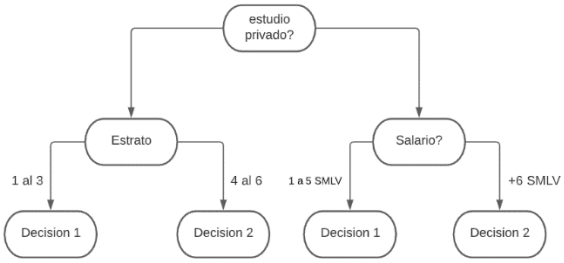


Figura 2: Nuestro árbol de decisión tipo CART será entrenado de manera que prediga los resultados acordes a lo esperado, es esta figura se ve como el árbol tomas diferentes decisiones dependiendo de los estudios, estrato social, SLMV y con estos datos y mas incluidos se puede crear una producción acertada

4.2.2 Algoritmo de prueba

Un nuevo nodo se crea y cada vez que suceda el algoritmo utiliza un nuevo criterio para crear mas de ellos y clasificar los datos de manera correcta. Esto hace el árbol mas especifico y preciso, las decisiones van basadas en la homogeneidad de los nuevos nuevos nodos creados.

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

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

Wikipedia. 2019. ID3 algorithm. (22 May 2019). Retrieved August 11, 2019 from https://en.wikipedia.org/wiki/ID3_algorithm

Juan, B.V., Árboles B*. Universidad Catolica de Oriente. (2011). https://sites.google.com/site/tutoriasarboles/arbolesb

Cisneros, G. O. (2015). Obtenido de https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=63994

negocios, I. d. (8 de 4 de 2008). *inteligencia de negocios*. Obtenido de https://inteligencianegocios.wordpress.com/tag/arboles-de-decision/

Porcel, D. l. (mayo de 2009). *SEDICI*. Obtenido de http://sedici.unlp.edu.ar/handle/10915/19846

Timaran, C. H. (1 de 2019). *Revista de investigacion desarrollo e innovacion*. Obtenido de https://revistas.uptc.edu.co/index.php/investigacion_duitama/article/view/9184

Timaran, i. c. (30 de 15 de 2015). *Universidad cooperativa de colombia*. Obtenido de https://repository.ucc.edu.co/handle/20.500.12494/1039

Wikipedia. (2020 de 2 de 8). *Wikipedia*. Obtenido de https://es.wikipedia.org/wiki/C4.5