PREDICCIÓN DE RESULTADOS EN PRUEBA SABER PRO USANDO DATOS ACADEMICOS Y SOCIALES

Jorge Juan Araujo Universidad Eafit Colombia jjaraujoa@eafit.edu.co	José Aníbal Pinto Universidad Eafit Colombia japintof@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co	Miguel Manzur Universidad Eafit Colombia mmanzurg@eafit.edu.co

RESUMEN

Este proyecto pretende predecir los resultados que podría obtener un estudiante en las pruebas Saber PRO. Lo interesante de esto es ver qué factores influyen, afectan o ayudan a obtener un "buen puntaje". Se obtuvo este resultado organizándolos, usando PANDAS, luego se procesaron con un código inspirado en el algoritmo CART. Posteriormente la impureza de Gini nos ayudó a conocer qué preguntas eran las más influyentes. Finalmente, el código da como resultado un árbol de decisión binario donde los nodos son las preguntas, y cada pregunta divide el marco de datos en dos marcos de datos más pequeños que separan los objetos que cumplen la condición y los que no. Concluimos que el numero de libros en casa y los puntajes de ciencias sociales, química y por encima de todo el de inglés. Y el algoritmo dio como resultado en los valores de sensibilidad, exactitud y precisión, en todos, un valor medio del 80% que consideramos muy positivo. Pero en los tiempos de espera, a pesar de utilizar el mejor o uno de los mejores algoritmos como lo es el CART, no lo recomendamos para cantidades de datos masivas.

Palabras clave

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

1. INTRODUCCIÓN

Uno de los objetivos primordiales que tiene un estudiante al momento de culminar sus estudios, son las pruebas estandarizadas del estado; ya que, con estas, se mide el nivel académico del pregrado en comparación al resto del país. Esto es muy importante, ya que nos permite saber que tan bien preparados estarán, si desertarán en sus estudios, o si tendrán un salario mejor; entre muchas otras posibilidades.

1.1. Problema

Este estudio se hará tomando en cuenta factores extraacadémicos, por lo que también evaluaremos si existe o no una relación en estos al momento de obtener resultados

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad [1]. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad [2]

Elegimos el algoritmo CART ya que es un algoritmo optimo en la explicación de este proyecto, dada su simplicidad a la hora de que las personas puedan entender el por qué se toma esa decisión.

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 Predicción de accidentes viales en Cartagena, Colombia, con árboles de decisión y reglas de asociación

En esta investigación se buscó predecir la severidad de los accidentes viales en la ciudad de Cartagena por medio de los factores asociado (Individuales, climatológicos, actores viales) usando técnicas como la minería de datos y arboles de decisión J48. Luego de 10.053 registros de accidentes de tráfico entre 2016 y 2017 han obtenido unos amplios datos que han servido para obtener conclusiones. Estas han sido de utilidad al momento de tomar decisiones en materia de seguridad vial. [3]

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

Esta es una investigación con un problema de estudio muy parecida a la propuesta en este proyecto, pero con la diferencia de que, en esta, su objetivo final luego de obtener los resultados es mejorar los posibles puntajes a obtener en la prueba Saber PRO. Para ello usaron redes neuronales como técnica de minería de datos. Concluyeron que sus resultados, aunque no fueron exactos, fueron acorde a lo esperado, y recalcaron todos los espacios inexplorados que aún faltan por explotar. [1]

2.3 Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia

Detectar patrones delictivos en la ciudad de Pasto fue el objetivo de esta investigación. Ellos usaron todos los datos registrados en el municipio de víctimas violentas (fatales y no fatales), por esto, construyeron un árbol de decisión que permitió descubrir patrones, los cuales han sido utilizados por los organismos gubernamentales y de seguridad para tomar decisiones eficaces. Se aplicó Cross Industry Standard Process for Data Mining (CRISP-DM). [2]

2.4 Aplicación de árboles de decisión para la predicción del rendimiento académicos de los estudiantes de los primeros ciclos de la carrera de ingeniería civil de la Universidad Continental (U.C)

El objetivo de esta investigación fue predecir el rendimiento académico identificando las variables de los factores que más influyen en el estudiante en su rendimiento académico. La aplicación partículas que se hace en este estudio es lo interesante, ya que toma factores específicos para su selecta población (estudiantes de ingeniería civil de la U.C). Utilizaron el algoritmo J48. Tuvieron resultados positivos, lo que sirvió como base no solo para predecir el rendimiento académico, sino también otros indicadores, lo que me parece un gran ejemplo para tomar en cuenta con el futuro de este proyecto. [4]

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 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

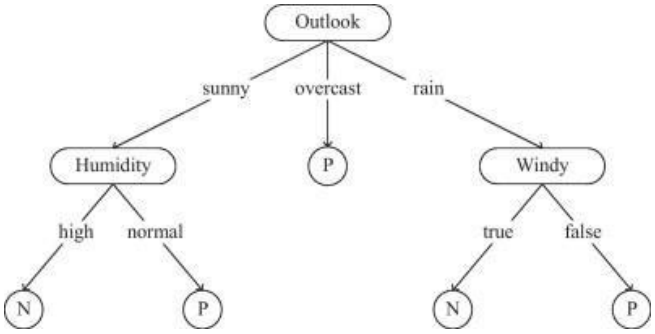
Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

3.2 Alternativas de algoritmos de árbol de decisión

3.2.1 ID3 (Induction Decision Trees)

Este algoritmo tiene como objetivo construir un árbol de decisión el cual explica instancias de una secuencia de entrada, eligiendo el mejor atributo dependiendo de una determinada heurística para luego determinar variables importantes para solucionar el problema y establecer una secuencia dentro del árbol de decisión.

Su resultado puede ser expresado como un conjunto de reglas Si-entonces. Además, es recursivo, no se realiza "backtracking"y utiliza la entropía(medida de incertidumbre del sistema). Los árboles de decisión están formados por nodos, ramas y hojas.



3.2.2 CART (Classification and Regression Trees)

Son una alternativa a la predicción tradicional. Entre sus ventajas está su interpretabilidad y la invarianza den la estructura de sus árboles de clasificación o de regresión a transformaciones de las variables independientes.

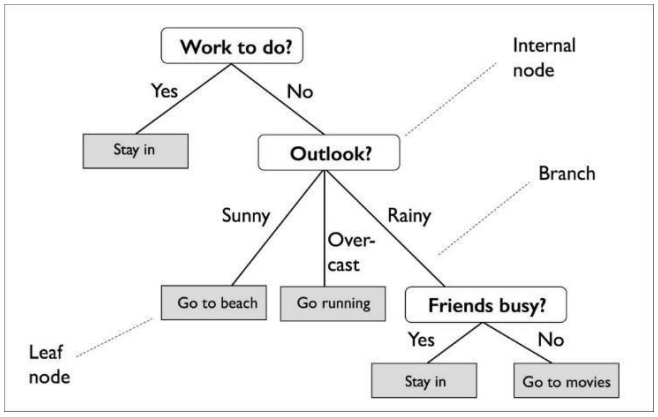
Consiste en tres pasos:

- Construir un árbol saturado

- Elegir el tamaño correcto

- Clasificar nuevos datos a partir del árbol construido

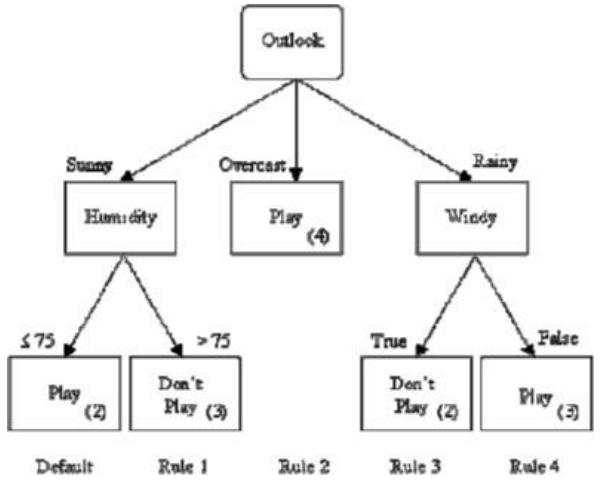
Otra de las ventajas es su más eficiente comportamiento antes situaciones de estructura alejadas de la linealidad.



3.2.3 C4.5

Es un algoritmo creado por Ross Quinlan para generar árboles de decisión como una extensión del algoritmo ID3. Estos árboles generados por C4.5 pueden ser usados para clasificación.

Este algoritmo maneja atributos continuos y discretos, maneja puntos de datos incompletos y usa el concepto de entropía de información.



3.2.4 C5.0

Este algoritmo es una mejora del C4.5 y algunos de los cambios son un aumento en la velocidad, uso mas eficiente de memoria, árboles de decisión mas pequeños, soporte para boosting, ponderación y eliminación de atributos de poca ayuda con el algoritmo Winnow.

. 4. DISEÑO DE LOS ALGORITMOS

Dado que la información de entrada con la que se entrenará y evaluará el modelo se encuentra en forma de “tabla”, es decir, hay variables correspondientes a cada columna con observaciones correspondientes a cada fila, y con tipos de datos diversos para cada una de estas variables se decide tratar con un “data frame”.

4.1 Estructura de los datos

Un árbol de decisión es un diagrama de predicciones con secuencias lógicas. Entre estos, existen los binarios, que se caracterizan por encasillar las respuestas en dos (bi) resultados diferentes, como puede ser un sí o un no, un verdadero o falso, etc.

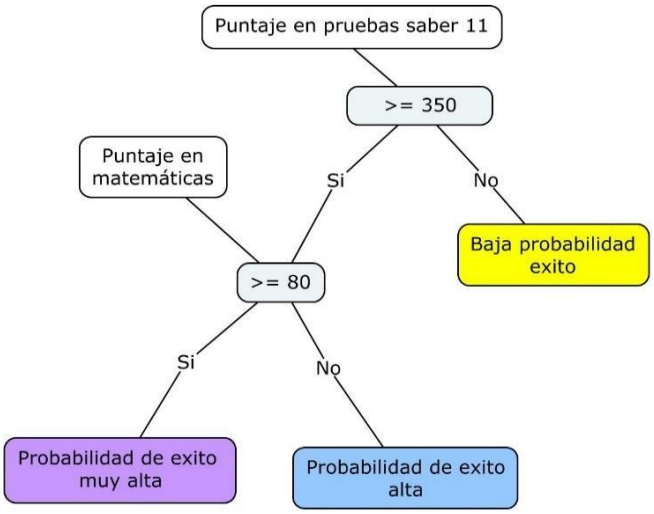


Figura 1: Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una probabilidad de éxito muy alta, los azules con una probabilidad alta y los amarillos con una baja probabilidad de éxito.

4.2 Algoritmos

El algoritmo funciona construyendo un árbol de decisión en el cual se explican instancias de una secuencia de entrada, elige el mejor atributo dependiendo de una determinada heurística y luego determina las variables importantes para solucionar el problema y establece después una secuencia dentro del árbol de decisión.

4.2.1 Entrenamiento del modelo

Utilizando un dataset que brinda información previa sobre los estudiantes, como estrato, lugar de nacimiento, tipo de colegio, etc; y la información posterior considerando especialmente el éxito o fracaso de este estudiante en la prueba. Con estos datos el algoritmo toma estas variables (estrato, lugar de nacimiento, etc.) y las evalúa según los resultados a posteriori, categorizándolo e intentando hallar una similitud entre las variables dadas a priori, y los resultados a posteriori.

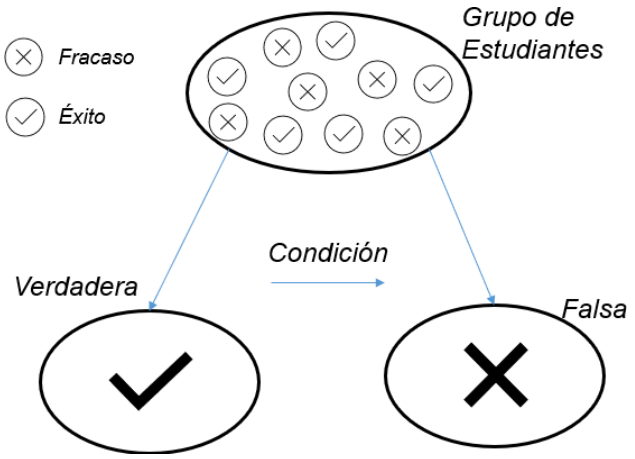


Figura 2: Entrenamiento de un árbol de decisión binario usando *CART*. En este ejemplo, mostramos un modelo para predecir si los estudiantes "Exitosos" cumplen con la condición dada.

4.2.2 Algoritmo de prueba

El árbol prueba cada una de las condiciones y guarda las condiciones que menor índice de impureza de Gini tuvieron, para luego ser tenidos en cuenta dentro de los factores a destacar, por su posible influencia directa hacia los resultados obtenidos de éxito o fracaso.

4.3 Análisis de la complejidad de los algoritmos

Para sacar las complejidades tuvimos en cuenta los ciclos que obtuvimos la seleccionar la mejor opción, crear el árbol, clasificar, ordenarlo y luego la forma en la que brindamos los datos.

Algoritmo	La complejidad del tiempo
Entrenar el árbol de decisión	O(2^x *n*m)
Validar el árbol de decisión	O(Log2 n)

Tabla 2: Complejidad temporal de los algoritmos de entrenamiento y prueba. Donde x es el numero de nodos en el árbol, n el numero de columnas y m el número de valores únicos en cada columna.

Algoritmo	Complejidad de memoria
Entrenar el árbol de decisión	O(n*m +x)
Validar el árbol de decisión	O(n)

Tabla 3: Complejidad de memoria de los algoritmos de entrenamiento y prueba. Donde x es el número de nodos en el árbol, n el número de columnas y m el número de valores únicos en cada columna.

4.4 Criterios de diseño del algoritmo

El Data Frame es una de las estructuras de datos más utilizadas en todo el mundo cuando grandes cantidades de datos se trata, ya que la eficiencia y fácil organización y entendimiento de estas, ayuda a la mejor toma de decisión posible en las empresas, ya que sus respuestas se basarán en datos, y no en especulaciones. Esta efectiva toma de decisiones está siendo vital en todas las industrias.

A su vez, la forma en la que fueron entregados los datos (matriz), es más fácil de manipularlos usando una estructura similar. Aparte, los diferentes tipos de datos (enteros, string, etc.) entregados en el archivo CSV, son fácilmente manejables y permisibles en el lenguaje de programación Python y en nuestra estructura de datos. No solamente las facilidades son a la hora de manipular la información, sino que también, el CART en comparación al C5, tiene menor consumo de memoria y menor probabilidad de clasificación errónea. [7]

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 0	*Conjunto de datos 1*	*...Conjunto de datos 2*
Exactitud	0.833	0.812	0.805
Precisión	0.836	0.813	0.803
Sensibilidad	0.830	0.812	0.809

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 0	***Conjunto de datos 1***	***...Conjunto de datos 2***
Exactitud	0.8712	0.8294	0.82
Precisión	0.8756	0.8159	0.80
Sensibilidad	0.8636	0.85	0.8461

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	225 s	439 s	1003 s
Tiempo de validación	0.706 s	0.967 s	2,148 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	171 MB	216 MB	514 MB

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

6. DISCUSIÓN DE LOS RESULTADOS

Las medidas de exactitud, precisión y sensibilidad son apropiadas, ya que nos permiten saber con más detalle que tan cercanos son los resultados con lo que podría pasar en la realidad. Valores altos en estos tres aspectos reflejan la correcta selección de las variables elegidas y su confiabilidad. Además, gracias a esto podemos ver que el modelo tiene un buen ajuste, ya que no tiene en cuenta variables que no afectan los resultados, lo contrario pasa con las que son consideradas importantes.

El algoritmo posee una buena efectividad en cuanto a espacio y memoria, debido a que los tiempos son cortos y la cantidad de memoria consumida no es muy grande, en proporción a la cantidad de datos que se abarcan.

Estos resultados pueden utilizarse para situaciones en las que el puntaje obtenido por los estudiantes tenga algún tipo de influencia, como lo son, programas de becas gubernamentales y admisiones en institutos o universidades, medición de efectividad de empresas prestadoras de servicios en cuanto a preparación pre-icfes, selectividad de trabajadores de una empresa, etc.

6.1 Trabajos futuros

En este código existe la variable prof (profundidad) que es la que determina el número de preguntas que el programa va a tener en cuenta, pero esta profundidad está dada por el usuario, por lo que el número de preguntas que toma el programa será la que uno crea que es mejor, por lo tanto, se propone, para una próxima versión, que el programa calculo la profundidad optima por sí solo.

AGRADECIMIENTOS

Agradecemos la colaboración en la de definición de la metodología que utilizamos para la realización del código a [Valentina Mendoza, Matemática egresada de la Universidad del Bosque].

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.

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.

REFERENCIAS

[1] Williams, Laurie. *"White-Box Testing" (PDF)*: 60–61, 69. *Recuperado el 11 de octubre de 2020*.

[2]. Gao, J., Tsao, H.-S. J., & Wu, Y. (2003). *Testing and quality assurance for component-based software*. Boston, MA: Artech House.

[3]. Garcia, J. y Sanchez, P. Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia. *SciELO Analytics.* 2019.

[4]. Ospina, H y Quintana, L. Predicción de accidentes viales en Cartagena, Colombia, con árboles de decisión y reglas de asociación. Universidad Javeriana. 2019.

[5]. Timaran, R. Calderon, A. Hidalgo, A. Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia. *ResearchGate.* 2017.

[6]. Camborda, M. Aplicación de árboles de decisión para la predicción del rendimiento académicos de los estudiantes de los primeros ciclos de la carrera de ingeniería civil de la universidad Continental. Universidad Nacional del Centro de Perú. 2014.

[7] Nguyen. Comparative Study of C5.0 and CART algorithms. Retrieved from http://mercury.webster.edu/aleshunas/Support%20Materials/C4.5/Nguyen-Presentation%20Data%20mining.pdf