Estructuras de predicción de éxito académico en prueba saber pro.

Tomas Marin A Universidad Eafit Colombia tmarina@eafit.edu.co	Juan Andrés Vera Universidad Eafit Colombia javeraa@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN:

Con este trabajo se pretende analizar y dar predicciones sobre cómo sería la educación superior de personas en Colombia, analizando varias variables y con los resultados de los icfes estimar el rendimiento de estudiantes en su futuro próximo principalmente en las pruebas saber pro determinando así si el estudiante tendrá un óptimo rendimiento en esta prueba o no.

Es importante dar solución a este problema ya que ayudará a gran número de estudiantes a poder de cierta manera tener una idea de cómo puede ser su desempeño en este examen y así poder según sus resultados determinar que tanto debe de mejorar sus conocimiento y habilidades.

Pudimos observar que al realizar este tipo de proyectos se puede llegar a predecir los resultados de las personas de una manera muy poco ética porque se toman encuentra variables que no deberían de ser tomadas ya que no limitan la capacidad cognitiva de las personas, además se quitan o no se tienen en cuenta variables más personales y que no pueden ser tomada ya que por ejemplo alguna persona puede morir y por ende hacer que una persona con probabilidad de ganar pierda así mismo con otras variables como conseguir novia o novio antes del examen entre otras, como también nos damos cuenta de que la predicción que realiza no es 100% efectiva.

1. **INTRODUCCIÓN**:

con el pasar del tiempo nos damos cuenta de la importancia que tiene presentar pruebas que miden nuestro conocimiento principalmente nos ayuda en el ámbito laboral y académico y por esto los estudiantes siempre tienen como objetivo sacar resultados de acuerdo con sus expectativas.

un ejemplo de lo anterior son las pruebas saber pro que de una manera u otra miden que tanto hemos aprendido durante nuestra educación superior y en nuestras respectivas carreras por esto la solución a este problema trae consigo una gran ayuda para los estudiantes ya que mediante los resultados que tengan en los icfes y otras variables podrán determinar su desempeño en esta prueba y así poder saber de qué manera seria la óptima para prepararse para presentar dicha prueba.

1.1 **PROBLEMA**:

La problemática planteada es crear un árbol de decisión que ayude a los estudiantes a conocer de una manera aproximada cómo serán sus resultados en las pruebas saber pro que tienen como objetivo determinar qué tanto conocimiento fue adquirido durante la carrera, buscamos determinar si un estudiante dependiendo de variables sociodemográficas y académicas entre otras obtiene un desempeño en dicha prueba por encima del promedio o no.

1.2 SOLUCION:

La solución que se presenta para este proyecto es el uso de un árbol de decisión CART, lo hemos escogido porque es un árbol más familiar al lenguaje o al entendimiento de las personas del común ya que presenta una menor complejidad matemática y operativa.

2. TRABAJOS RELACIONADOS

2.1. Estudio sobre la prueba saber 11.

Este estudio fue realizado con el fin de detectar factores asociados a el desempeño académico de los estudiantes que presentaron la prueba saber 11.

Tubo como fin mostrarle datos concretos sobre el desempeño de los estudiantes al ministerio de educación como también a las instituciones que tienen relación con la educación en Colombia. [1]

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

Consiste en hacer una minería de datos, en este caso con árboles de decisión, reglas, redes neuronales, entre otras. Para luego hacer una predicción de datos de las Pruebas Saber Pro, que se hacen en Colombia, "En dichas pruebas, se evalúan competencias genéricas, las cuales aplican para estudiantes de todos los programas de formación, estas incluyen las áreas de lectura crítica, razonamiento cuantitativo, composición escrita, inglés y competencias ciudadanas, además de la medición de las competencias específicas" (scielo). Entonces decidieron implementar esta estrategia para predecir a un valor aproximado de los análisis de resultados del examen Saber-pro.

Este proceso lo implementaron tomando una serie de datos apropiados de entrada, para facilitar los datos que se quieren obtener, luego hacen una depuración de los datos en caso de que se encontrara alguna incoherencia.[2]

2.3 Predicción del desempeño académico usando técnicas de aprendizaje de máquinas.

En muchas instituciones de educación superior, la decisión de admisión está basada en el resultado de las pruebas ICFES Saber 11, entonces, se descubrió que existen factores que facilitan la predicción de los posibles resultados, y con esto, se pretende ayudar a mejorar cada vez la calidad de las calificaciones, al descubrir que factores son los que más afectan para que este mejore o vaya en decadencia, entonces se pretende implementar un algoritmo para la minería de datos, En este caso con árboles de decisión, tomando la información proporcionada por la misma entidad ICFES correspondiente a características sociodemográficas de los estudiantes y variables sobre los establecimientos educativos.[3]

2.4 Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees.

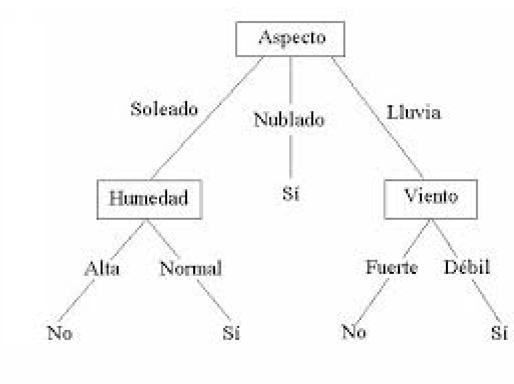
Tomando los datos del entorno familiar, las condiciones socioeconómicas, notaron que influye en el desempeño académico de los estudiantes, y con más razón afecta los resultados de las pruebas de estado. Aun así, en Colombia es limitada la investigación y los métodos que se han usado para analizar estas variables. Esta investigación está enfocada en predecir los resultados del examen de estado ICFES saber 11 del año 2016, a partir de ciertas observaciones y características de los estudiantes. Los datos proporcionados provienen de la base de datos del instituto ICFES.[4]

Alternativas de algoritmos de árbol de decisión.

3.1 Árbol ID3.

Árbol ID3 que es “inducción mediante arboles de decisión” fue creado por j Ross Quinlan.

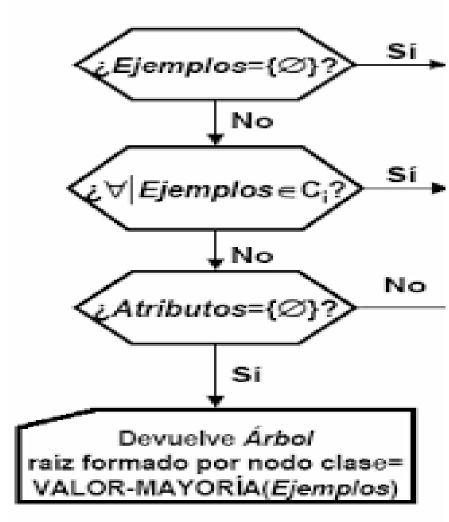
Este árbol tiene como principal función la búsqueda de hipótesis o reglas este está conformado por nodos de decisión y nodos-hojas con estas se construye un árbol no tan grande y que tiene un menor número de posibilidades ya que cuando se presentan nodos-hojas se para.[5]



3.2 Árbol C4.5.

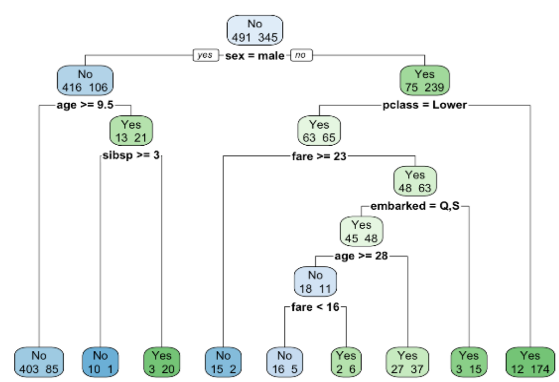
Es un árbol desarrollado por Ross Quinlan, este es una extensión o complemento del ID3 que también fue desarrollado por él.

Usa el concepto de entropía de información para la creación de árboles, este es un árbol que en si se define como un árbol clasificador [6]



3.3 Árbol C5.0

Este árbol pretende dividir los datos para así llegar a una ganancia máxima de información. Este va dividiendo la información en subinformacion y así repetidamente hasta que ya no se pueda hacer más divisiones y luego elimina las que no tiene relevancia para así optimizar la información que se usara.[7]



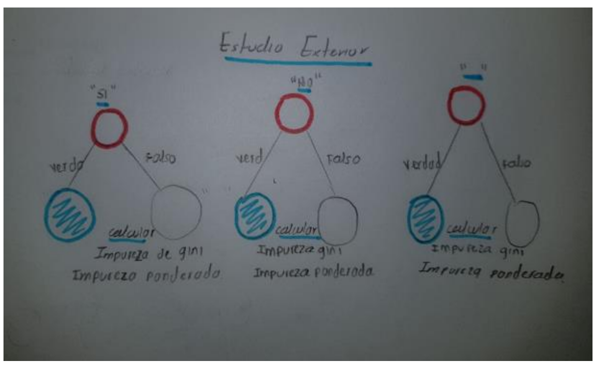
3.4 CART

Este es un árbol de clasificación y regresión puede utilizar variables numéricas fácilmente como también
este es un árbol binario y tiene bases recursivas además podemos destacar su interpretabilidad.[8]

4. DISEÑOS DE LOS ALGORITMOS.

4.1Estructura de los datos.

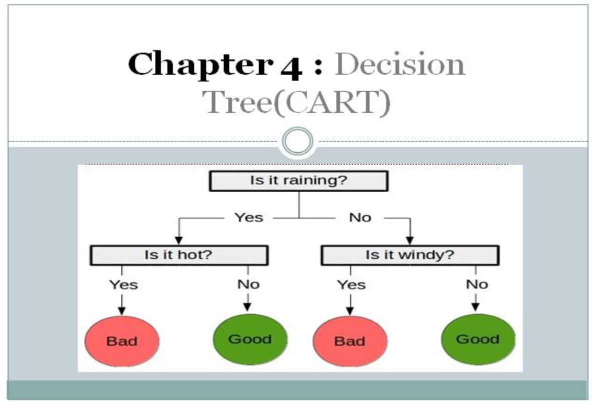
La estructura de datos que vamos a usar para este proyecto es un árbol de decisión CART el cual nos ayudara a predecir los resultados de los exámenes tomando varias variables que afectan a los estudiantes a la hora de presentar pruebas. Este tipo de árbol va generando divisiones según los mejores criterios que se tengan del conjunto de datos (según las respuestas que se tengan y su grado de impureza), determinando la impureza de Gini.



4.2 Algoritmos.

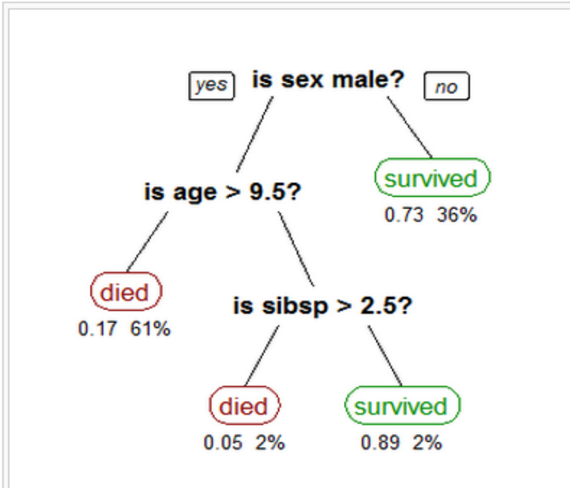
El algoritmo CART utiliza una clasificación de variables según los resultados que se puedan tener de una pregunta o dato, separa los resultados posibles que se pueden generar de esta pregunta en buenos o malo de esta manera puede generar varios caminos según las respuestas que se presenten para casos específicos.

La siguiente imagen muestra cómo funciona un árbol CART y muestra como separa las preguntas en buenas o malas según la mejor opción.



4.2.1 Entrenamiento del modelo.

Vamos a usar el árbol de decisión CART, podemos ver un ejemplo de uso de este árbol:



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

Este árbol muestra varias posibles respuestas de una pregunta dependiendo de las posibles respuestas que se puedan dar para esta pregunta ya que si por ejemplo la respuesta o los datos son de 1 a 5 se tomaría hacia la derecha los números de la respuesta es de 3 a 5 y de 1 a 2 se va hacia la izquierda y de esta manera va creando el árbol según las respuestas obtenidas dividiendo cada vez más el árbol.

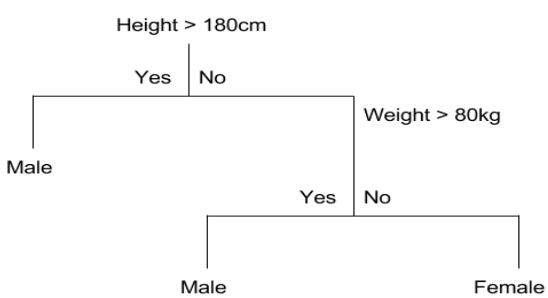
4.2.2 Algoritmo de prueba.

El seudocódigo para la creación de un árbol CART es el siguiente

Classification and Regression Tree

1. Start at the root node.
2. For each ordered variable X,
   convert it to an unordered variable X' by grouping its values in the node into a small number of intervals
   if X is unordered, then set X' = X.
3. Perform a chi-squared test of independence of each X' variable versus Y on the data in the node and compute its significance probability.
4. Choose the variable X∗ associated with the X' that has the smallest significance probability.
5. Find the split set {X∗ ∈ S∗} that minimizes the sum of Gini indexes and use it to split the node into two child nodes.
6. If a stopping criterion is reached, exit.
   Otherwise, apply steps 2–5 to each child node.
7. Prune the tree with the CART method.

```
d=0, endtree=0
Note(0)=1, Node(1)=0, Node(2)=0
while endtree<1
        if
Node(2^d – 1) + Node(2^d) + … + Node(2^(d+1) – 2) = 2 – 2^(d+1)
                endtree = 1
        else
                do i = 2^d – 1, 2^d, …, 2^(d+1) – 2
                        if Node(i) > -1
                                Split tree
                        else
                                Node(2i + 1) = -1
                                Node(2i + 2) = -1
                        end if
                end do
        end if
d = d + 1
end while
```



Este árbol va haciendo cada vez menor la impureza calculándola mediante la impureza de Gini que debe de ir disminuyendo a medida que se van creando nuevos nodos llamado nodo hijo ya que su fin es conseguir la mínima impureza de un conjunto de datos.

4.3 Análisis de la complejidad de los algoritmos

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

Tabla 2: Complejidad temporal de los algoritmos de entrenamiento y prueba.

La complejidad es esta debido a que sería el número de filas por columnas y como cada piso cuenta con 2 nodos mínimo entonces por eso sería 2 a la m.

Algoritmo	**Complejidad de memoria**
Entrenar el árbol de decisión	O(m*n* 2^m)
Validar el árbol de decisión	O(n)

Tabla 3: Complejidad de memoria de los algoritmos de entrenamiento y prueba.

La m y la n significan el número de columnas y filas respectivamente.

4.4 Criterios de diseño del algoritmo.

Decidimos hacerlo de esta manera ya que vimos que era la forma a nuestro parecer y entender mas optima y con una menor complejidad para el entendimiento de personas ajenas a el proyecto como también lo determinamos de esta manera por consejos del profesor que nos dieron una mayor claridad sobre cual algoritmo escoger.

5. RESULTADOS.

5.1 Evaluación del modelo.

5.1.1 Evaluación de modelo de entrenamiento.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Exactitud	0.5	0.5	0.6

Precisión	55.5%	55.5%	56%
Sensibilidad	0.3	0.4	0.4

Tabla 3. Evaluación del modelo con los conjuntos de datos de entrenamiento.

5.1.2 Evaluación de los conjuntos de datos de validación.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Exactitud	0.6	0.6	0.7
Precisión	56%	55%	54%
Sensibilidad	0.4	0.4	0.4

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	1569.61s	20.4 s	27.6743 s
Tiempo de validación	16.536 s	22.456 s	29.6743s

Tabla 5: Tiempo de ejecución del algoritmo *(CART)* para diferentes conjuntos de datos.

5.3 Consumo de memoria.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Consumo de memoria	190M	560 MB	1,300 MB

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

6. DISCUSIÓN DE LOS RESULTADOS.

Tenemos que el consumo de memoria y tiempo no es el mas adecuado debido a que toman un tiempo un poco prolongado, pero consideramos que por la cantidad de datos presentados no son tan desacertados.

Consideramos que este tipo de prueba o proyecto no puede ser usado para determinar becas o cosas por el estilo ya que no reflejan la realidad de la persona al final del día siguen siendo probabilidades que varían de acuerdo con muchos más factores y no algo exacto.

Consideramos que puede ser viable hacer una campaña con los estudiantes en general para que les vaya mejor a todos, pero no clasificándolos, dependiendo de esta prueba ya que no da una mirada clara a lo que puede ocurrir.

6.1 Trabajos futuros.

Nos gustaría mejorarlo con la implementación de una interfaz grafica para que sea mas visual y más cómodo ver todos los resultados, gráficos y grafos. Como también un tipo de árbol que pueda dar una mejor predicción. Además de implementar un Random Forest ya que es una mejor practica en lo referente a árboles.

AGRADECIMIENTOS

Damos agradecimiento tanto a el profesor Mauricio Toro como también a sus monitores Simón Marin y Isabel Piedrahita que siempre estuvieron dispuestos a resolver nuestras dudas y a guiarnos en la realización de este proyecto.

REFERENCIAS.

1.Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. Rev.investig. desarro. innov., 9 (2), 363-378. https://doi.org/10.19053/20278306.v9.n2.2019.9184

2. García J.R.G. – Sánchez P.A.S. – Orozco M. – Obredor S. Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia.

https://scielo.conicyt.cl/scielo.php?pid=S0718-50062019000400055&script=sci_arttext

3. Rodríguez F.J.D. – Benavides H.L.G. – Riascos A.J.V. Predicción del desempeño académico usando técnicas de aprendizaje de máquinas.

https://www.icfes.gov.co/documents/20143/234129/Prediccion+desempeno+academico+usando+un+enfoque+de+mineria+de+datos.pdf/0e5d0f1d-20ac-dffc-f3f1-88ccfde6b0bc

4. García J.D.G. – Skrita A. Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees.

http://ojs.ual.es/ojs/index.php/psye/article/view/2056/3119

5. colaboradores de Wikipedia. (2020b, abril 25). Algoritmo ID3. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Algoritmo_ID3#:%7E:text=El%20algoritmo%20ID3%20es%20utilizado,dado%20un%20conjunto%20de%20ejemplos.

Caparrini, F. S. (2013, 5 enero). Arboles decision id3. SlideShare. https://es.slideshare.net/FernandoCaparrini/arboles-decision-id3

6.colaboradores de Wikipedia. (2020, 8 febrero). C4.5. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/C4.5

7.

IBM Knowledge Center. (s. f.). IBM. https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/c50node_general.html

8. Días, J.f (2012). Comparación entre Arboles de Regresión CART y regresión Lineal. http://www.bdigital.unal.edu.co/9474/1/71269839.2013.pdf

mlejarza. (s. f.). Arboles de clasificación y regresión.https://www.uv.es/mlejarza/actuariales/tam/arbolesdecision.pdf