**PREDICCIÓN DEL ÉXITO ACADÉMICO, EN EDUCACIÓN SUPERIOR, UTILIZANDO ÁRBOLES DE DECISIÓN**

|  |  |  |  |
| --- | --- | --- | --- |
| Juan Pablo Giraldo Ramirez  Universidad Eafit  Colombia  jpgiraldo1@eafit.edu.co | Luis Fernando Posada Cano  Universidad Eafit  Colombia  lfposadac@eafit.edu.co | Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co | Mauricio Toro  Universidad Eafit  Colombia  mtorobe@eafit.edu.co |

# **RESUMEN**

Se debe diseñar un algoritmo 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. Por la cuál con este proyecto queremos definir el éxito académico como la probabilidad que tiene un estudiante de obtener este puntaje, este éxito lo podemos medir, por ejemplo, la empleabilidad del egresado, el salario, la felicidad en su trabajo, etc. Hay problemas relacionados que también se basan en arboles de decisión los cuales abordan temas de éxito académico tanto en universidades como en colegios. El algoritmo a utilizar es el CART los cuales son arboles de clasificación y de regresión; Los resultados obtenidos referente al consumo de memoria, tiempo y exactitud del algoritmo verdaderamente son muy prometedores, aunque no son los mejores con varios ajustes al algoritmo y con más entrenamientos se podría ir corrigiendo. En conclusión, pensamos que este tipo de algoritmos son muy útiles para estos tipos de trabajos, aunque hay unos que podrían tener mayor exactitud y mejor rendimiento referente a tiempo y memoria.

## **Palabras clave**

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

# **1. INTRODUCCIÓN**

Debido a la expansión tecnológica y gracias a la aplicación de diferentes métodos predictivos se ha comprobado como en otros campos los árboles de decisión han tomado gran importancia la hora de enfrentarse a nuevos desafíos. En América Latina desde mediados del siglo XXI se hizo constante la preocupación sobre la calidad de la educación, la deserción, y el éxito académico, por lo cual se han tomado varias medidas en otras áreas del conocimiento para así darle una solución a esta problemática, medidas que no han sido lo suficiente. Por tanto, se pretende con este proyecto contribuir con nuevas pruebas basadas en Arboles de decisión y el resultado de las pruebas saber 11 dando un enfoque en un país subdesarrollado como Colombia.

# **1.1. Problema**

Se debe diseñar un algoritmo, basado en árboles de decisión, y los datos del saber 11; Teniendo en cuenta variables sociodemográficas y académicas que pueden ser de gran impacto para el rendimiento académico, y su éxito posteriormente, esto para predecir si un estudiante tendrá un puntaje total, en las pruebas saber pro, por encima del promedio o no.

**1.2 Solución**

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran aplicabilidad, son muy eficientes en su aplicación, se codifican realmente de una manera muy sencilla, hemos utilizado como tal los métodos conocidos como métodos de caja blanca ya que evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad y aplicabilidad. El algoritmo a utilizar es el CART, es muy bueno, es muy sencillo, y nos da facilidad a la hora de implementarlo con cualquier tipo de datos.

**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 del rendimiento académico aplicando técnicas de minería de datos**

Uno de los retos que tienen que enfrentar las instituciones de educación superior para ofrecer una mayor calidad educativa, es mejorar el rendimiento académico de los estudiantes; El bajo rendimiento académico de los estudiantes que lleva a la desaprobación de los cursos, es uno de los problemas que debe enfrentar las instituciones educativas superiores. La presente investigación tiene como objetivo aplicar las TMD de regresión logística, árboles de decisión, redes bayesianas y redes neuronales usando los datos académicos de los estudiantes matriculados en el curso de Estadística General de la UNALM de los semestres 2013 II y 2014 I, con la finalidad de predecir la clasificación final (Desaprobado o Aprobado) de los estudiantes matriculados en el curso. Se aplicarán métricas a partir de la matriz de confusión para validar e identificar el mejor modelo que permita predecir la clasificación de nuevos estudiantes matriculados en el curso; y apoyar a los profesores a identificar a los estudiantes con problemas académicos para brindarles un asesoramiento oportuno y efectivo.

Uno de los AD de clasificación de mayor aplicación, es el algoritmo C4.5 que fue propuesto por (Quinlan, 1993). El C4.5, va generando el árbol con la estrategia “Divide y Vencerás”, que consiste en un proceso recursivo y descendentemente para ir asignado los atributos a los nodos, dividiendo en tantas ramas como valores tenga la variable.

El proceso de inferencia con el C4.5,se hce con la finalidad de predecir la clasificación del resultado final del rendimiento académico para un estudiante como Desaprobado o Aprobado; se realiza a través de evaluar el conjunto de reglas que genera el C4.5, haciendo los recorridos desde el nodo raíz hasta cada uno de los nodos terminales.

Se aplica el método C4.5 que se implementa en WEKA con el algoritmo J48. Se usa un nivel de confianza de 0,15 y con poda. Se presenta el conjunto de reglas generadas con el árbol J48, con un tamaño de 17 nodos y 9 hojas. Se observa que las variables más importantes para predecir el resultado de un estudiante en el curso de EG son en orden de jerárquico el promedio acumulado, situación del curso, número de veces que lleva el curso de diferencial, nota en el curso diferencial y sexo. Las variables que no influyen son número de veces que llevo matemáticas y situación en el curso de integral.[1]

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

En este artículo 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 de educación media, que presentaron las pruebas Saber 11° en los años 2015 y 2016.

Para la construcción del modelo de clasificación con árboles de decisión se utilizó la herramienta WEKA (Hall et al, 2011) y su algoritmo J48, el cual implementa al algoritmo C.45. El algoritmo J48 se basa en la utilización del criterio de ganancia de información (information gain).

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T1061680A17, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 1.061.680 estudiantes, quienes presentaron las prueba Saber 11° en los periodos 2015 y 2016, donde se escogió el atributo puntaje global (puntaje\_global\_cuali) como clase, se puede observar que este clasifica correctamente a 711.116 instancias, que corresponde a un porcentaje de precisión del 67%; mientras que 350.564 instancias fueron incorrectamente clasificadas, correspondiendo a un porcentaje del 33%.[2]

## **2.3** **Estudio del poder predictivo del puntaje de admisión sobre el desempeño académico universitario.**

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. Método: la población objetivo fueron los estudiantes matriculados a la Universidad de Caldas en el quinquenio 2008-2012. Se aplica el análisis de regresión múltiple. Resultados: en términos globales, el puntaje de admisión evidencia una asociación estadísticamente significativa con el desempeño académico. Conclusión: consistente con otras investigaciones, 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.

Se recurrirá a un análisis de regresión múltiple, mediante el procedimiento de pasos sucesivos, lo cual acude a una construcción iterativa mediante la eliminación secuencial de variables que no muestran un poder predictivo significativo sobre la variable respuesta, que en nuestro caso es el promedio de la carrera. Para ello se usará el software estadístico SPSS versión 19.0. Puntualmente, el análisis de regresión tiene que ver con encontrar la mejor relación entre la variable respuesta y la o las variables predictoras. [3]

**2.4 Modelo predictivo de deserción estudiantil basado en arboles de decisión**

Este artículo muestra la construcción de un modelo predictivo de deserción estudiantil, para pronosticar la probabilidad, que un estudiante abandone su programa académico, mediante técnicas de clasificación, basadas en árboles de decisión. La metodología utilizada, se basa en Knowledge Discovery in Database (KDD), con cinco etapas: selección, procesamiento, transformación, minería de datos y evaluación. Aplicando el algoritmo, Classification and Regression Tree (CART) de la herramienta R, se construyó un árbol con cuatro niveles de profundidad y mismo número reglas, que evalúan a los posibles desertores. Llevando a concluir que las variables nivel y notas tienen mayor influencia en la deserción.[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 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 |

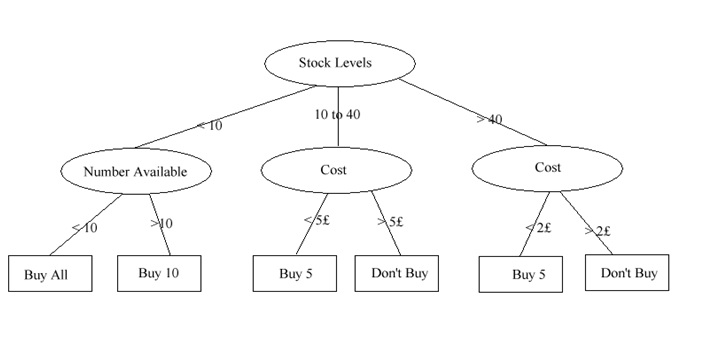
## **Tabla 1.** Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

## **3.2 Alternativas de algoritmos de árbol de decisión**

**3.2.1 C4.5**

C4.5 construye árboles de decisión desde un grupo de datos de entrenamiento de la misma forma en que lo hace ID3, usando el concepto de entropía de información. Los datos de entrenamiento son un de ejemplos ya es un vector donde representan los atributos o características del ejemplo. Los datos de entrenamiento son aumentados con un vector donde representan la clase a la que pertenece cada muestra.

En cada nodo del árbol, C4.5 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. El algoritmo C4.5 divide recursivamente en sublistas más pequeñas.

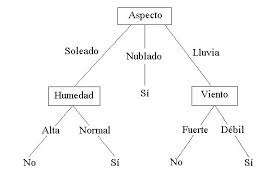


**3.2.2 ID3**

El algoritmo ID3 es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, ( el atributo a clasificar ) es el objetivo, el cual es de tipo binario ( positivo o negativo, sí o no, válido o inválido, etc. ).

De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.ID3 realiza esta labor mediante la construcción de un árbol de decisión.



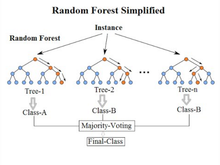
**3.2.3 Random Forest**

Random forest (o random forests) también conocidos en castellano como '"Bosques Aleatorios"' es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.

El algoritmo para inducir un random forest fue desarrollado por Leo Breiman1​ y Adele Cutler y Random forests es su marca de fábrica. El término aparece de la primera propuesta de Random decision forests, hecha por Tin Kam Ho de Bell Labs en 1995. El método combina la idea de bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho,2​3​ Amit y Geman,4​ para construir una colección de árboles de decisión con variación controlada.

La selección de un subconjunto aleatorio de atributos es un ejemplo del método random subspace, el que, según la formulación de Ho, es una manera de llevar a cabo la discriminación estocástica5​ propuesta por Eugenio Kleinberg.

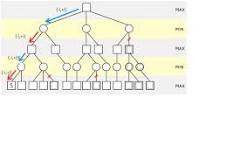
En muchos problemas el rendimiento del algoritmo random forest es muy similar a la del boosting, y es más simple de entrenar y ajustar. Como consecuencia, el Random forest es popular y ampliamente utilizado.



**3.2.4 CART**

El algoritmo CART es el acrónimo de Classification And Regression Trees (Árboles de Clasificación y de Regresión) fue diseñado por Breiman et al. (1984). Con este algoritmo, se generan árboles de decisión binarios, lo que quiere decir que cada nodo se divide en exactamente dos ramas.

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



## **4. DISEÑO DE LOS ALGORITMOS**

## La estructura de datos utilizada son los arboles de decisión y regresión, estos árboles nos dan la posibilidad de atreves de una serie de datos, el programa lo lea, los datos pasen por ciertos filtros y el programa va dividiendo los datos almacenándolos en listas, para al final del programar determinar una situación específica en nuestro caso es la predicción del éxito en las pruebas saber pro teniendo en cuenta varias condiciones sociales; El algoritmo utilizado es el CART donde este según la impureza de Gini y los datos de la tabla 1 va calcular la probabilidad de excelencia académica.

## (https://github.com/lfposadac/ST0245-001/tree/master/proyecto)

## **4.1 Estructura de los datos**

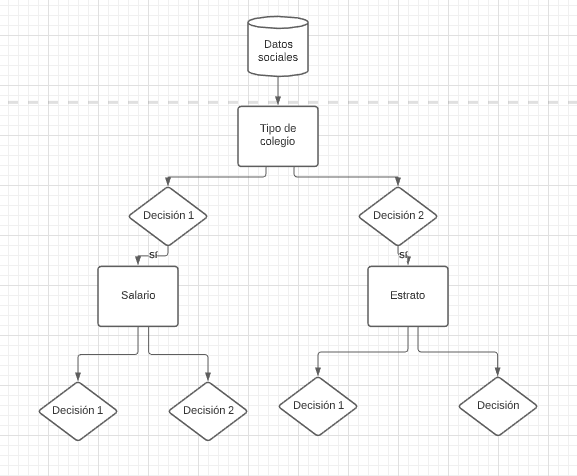
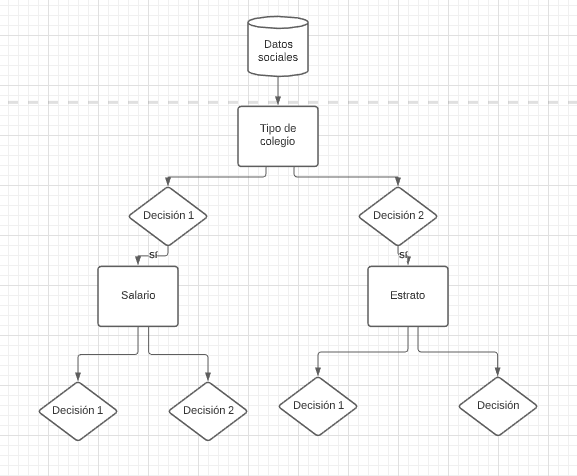
La estructura es un árbol de predicción binaria en la cual se tiene varios criterios que según la respuesta o el resultado a ese criterio, van pasando por los demás datos socilaes de las personas, luego de hacer esto las veces necesarias y pasar por cada uno de los criterios se podría predecir y tomar una decisión luego de esa predicción. 

Figura 1: Un árbol de decisión binario para predecir el éxito en la educación superior según sus condiciones sociales.

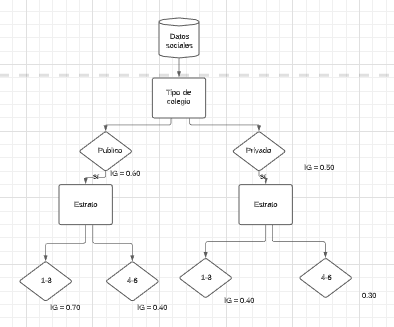
**4.2 Algoritmos**

Como tal el diseño del algoritmo es para leer una gran cantidad de datos, estos datos organizarlos, luego pasar por ciertos criterios donde en cada uno él va calcular la impureza de Gini, que según los datos pasados a este van a predecir que entre menos pureza más éxito va tener la persona en las pruebas de educación superior.



**4.2.1 Entrenamiento del modelo**

Se ingresaban una serie de datos donde a partir de cada suceso nos podría llevar a una decisión y así sucesivamente.



**Figura 2:** Entrenamiento de un árbol de decisión binario usando CART, donde mostramos el éxito de una persona según la impureza de Gini.

**4.2.2 Algoritmo de prueba**

Al validar los datos, pasarlos por cada una de los nodos del árbol, se podría tomar una decisión o predecir el éxito de las personas que es por lo cual se realiza el proyecto, estos e hace con gran exactitud ya que es por esto que el árbol crea estos nuevos nodos filtrando los más posible para la predicción final.

**4.3 Análisis de la complejidad de los algoritmos**

A partir de lo visto en complejidad en cada una de las clases pudimos analizar referente a tiempo y memoria la complejidad de cada uno de los algoritmos en notación O.

En cada uno de los algoritmos mediante las diferentes funciones dadas en clase por el profesor se sacó la complejidad, Los algoritmos a partir de cada uno de los entrenamientos se han tratado de optimizar y que mejore su rendimiento.

|  |  |
| --- | --- |
| **Algoritmo** | **La complejidad del tiempo** |
| Entrenar el árbol de decisión | O(N^2\*M\*2^M) |
| Validar el árbol de decisión | O(N\*M) |

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

N= Número de filas de los datos ingresados

M= Número de columnas de los datos ingresados

|  |  |
| --- | --- |
| **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.

N= Número de filas de los datos ingresados

M= Número de columnas de los datos ingresados

**4.4 Criterios de diseño del algoritmo**

Este tipo de algoritmo se escoge a partir de que para nosotros el algoritmo CART es el más viable para este tipo de problemas, ya que con este podemos optimizar de manera muy adecuado la memoria y el tiempo, Para la estructura de datos que nos tocó este semestre que es la de árbol binarios este tipo de algoritmos son los más sencillos de implementar, además de esto nos sentimos más cómodos a la hora de manipularlos. Este algoritmo también nos permite de manera más sencilla realizar la complejidad en cada uno de los casos.

**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 4*** |
| *Exactitud* | 0.75 | 0.98 | 0.94 |
| *Precisión* | 0.79 | 0.84 | 0.93 |
| *Sensibilidad* | 0.80 | 0.82 | 0.91 |

## **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 4*** |
| *Exactitud* | 0.77 | 0.78 | 0.79 |
| *Precisión* | 0.76 | 0.75 | 0.81 |
| *Sensibilidad* | 0.70 | 0.75 | 0.93 |

## **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 4*** |
| ***Tiempo de entrenamiento*** | **65.5 s** | **75.4 s** | **89.1 s** |
| ***Tiempo de validación*** | **60.1 s** | **70.8 s** | **87.6 s** |

## **Tabla 5:** Tiempo de ejecución del algoritmo

## **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 | 30 MB | 71 MB | 149 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.](http://docs.oracle.com/javase/7/docs/technotes/guides/visualvm/profiler.html) Para Python, use C-profiler.

## **6. DISCUSIÓN DE LOS RESULTADOS**

Los resultados obtenidos referente a la exactitud del algoritmo nos deja muy a gusto, no son los deseados por que nos gustaría a ver diseñado un algoritmo que nos brindara una mayor exactitud, la sensibilidad del algoritmo nos preocupaba un poco, pero también es una sensibilidad que no es mala por esto en trabajos futuros deseamos poder optimizar este algoritmo.

El modelo del algoritmo no está sobre sobre ajustado, se hicieron pocos entrenamientos con los diferentes datasets brindados, cabe destacar que mediante más entrenamientos mejor optimizado va estar el algoritmo. Los datos utilizados para los entrenamientos son parecidos a los datos utilizados para la verificación.

Referente al consumo de tiempo y memoria podemos señalar que los resultados obtenido son muy prometedores.

Sobre la utilización del algoritmo se puede decir que es mejor para brindar becas a las personas con mejor probabilidad de éxito.

**6.1 Trabajos futuros**

Primero que todo en nuestro algoritmo nos gustaría mejorar la optimización y rendimiento a partir del código, hacerlo de forma más recursiva para tener la menos complejidad posible.

Como se dijo anteriormente sería la optimización del código ya que nos gustaría que su implementación no tuviera gran complejidad en términos de memoria y tiempo, pero deseamos obtener una mayor exactitud.

Sinceramente pienso que el uso de un bosque aleatorio si no es implementado de una forma correcta se puede agregar una mayor complejidad al algoritmo deseado.

# **AGRADECIMIENTOS**

Agradecemos la asistencia brindada y la solución de cada una de las dudas a Simón Marín monitor de la materia, a Santiago Espinoza persona que nos apoyó explicándonos algoritmo y la utilización de este. Hacemos una mención muy especial al guía de la materia, el gran Mauricio Toro Bermudez por sus enseñanzas, la pasión y entrega para cada una de sus clases.

# **REFERENCIAS**

1 Menacho, C. *Predicción del rendimiento académico aplicando técnicas de minería de datos,* Recuperado el 16 de agosto del 2020: Dialnet-PrediccionDelRendimientoAcademicoAplicandoTecnicas-6171237.pdf

2. Timarán, R. Caicedo, J y Hidalgo, A. *Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°*, Recuperado el 16 de agosto del 2020: Dialnet-ArbolesDeDecisionParaPredecirFactoresAsociadosAlDe-6976577.pdf .

3. Tapasco, O. Ruiz, F y Osorio D. *Estudio del poder predictivo del puntaje de admisión sobre el desempeño académico universitario,* Recuperado el 16 de agosto del 2020: <https://www.redalyc.org/jatsRepo/1341/134149931008/html/index.html>

4. Cuji, B. Gavilanes, W y Sanchez, R. *Modelo predictivo de deserción estudiantil basado en arboles de decisión,* Recuperado el 20 de agosto del 2020: <https://www.revistaespacios.com/a17v38n55/a17v38n55p17.pdf>