# **TÍTULO (UNA BREVE DESCRIPCIÓN DEL PROYECTO, ENTRE 8 Y 12 PALABRAS)**

|  |  |  |  |
| --- | --- | --- | --- |
| David Madrid Restrepo  Universidad Eafit  Colombia  dmadridr@eafit.edu.co | Simón Álvarez Ospina  Universidad Eafit  Colombia  salvarezo1@eafit.edu.co | Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co | Mauricio Toro  Universidad Eafit  Colombia  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 = C**ontribución 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

# **ABSTRACT**

This research aims to predict students’ performance in statal test 'Saber PRO' through their past results and the different variables which could affect their academic performance. This problem opens up opportunities to develop in different areas, like education, tools that can contribute to improve the results of students in their college career.

Our project is related with some questions such as finding the most influential variables in students results and the methods that can be applied to improve their global score in statal tests. ¿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**

# Due to technology expansion and the expected results that some algorithms produced for the resolution of nonlinear problems [1], the application of different methods of prediction is sought by other knowledge areas where systems have increased their importance. One of the essential challenges that Latin America has gone through is the quality of the education and, despite the different efforts in others knowledge areas, sufficiently noticeable measures have not been taken to improve it. Thus, this project intends to contribute with tests and new interpretations of decision trees’ results for the development of a quality higher education and the training of competent professionals.

# **1.1. Problem**

Predicting student's academic success is a matter where researchers have been developing different solutions, which deliver varied but no determinant results. Finding an algorithm that yields an appropriate solution can help to understand which variables have the greatest impact in student’s academic performance.

**1.2 Solución**

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad (¡falta *una* cita *para este argumento!*). Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad (¡Falta *una* cita para *este argumento!*).

Explique, brevemente, su solución al problema *(En este semestre, la solución es una implementación de un algoritmo de árbol de decisión para predecir el éxito académico. ¿Qué algoritmo elegiste? ¿Por qué?)*

**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. RELATED WORKS**

**2.1 Student's academic success in their first year through decision trees.**

## Josip Mesarić and Dario Šebalj in their project were looking for solutions to predict student’s success rate in their first year through the study of variables that could affect their performance, dividing their results in two groups. For this, they used different decision trees algorithms like ID3 and J4.8, besides random forests and REPTree. In the end, they achieved a 79.35% precision with the REPTree algorithm. [1]

**2.2 Data mining and decision trees to find the most influential variables in higher education.**

## Qasem et. al. made use of different data mining techniques to find the most influential variables in the performance of students on higher education. Essentially, they used a classification method in algorithms like ID3, C4.5 and Naïve Bayes to achieve their goal. The precision they got with the 3 methods did not reach the 40% threshold. [2]

**2.3 Prediction of students’ academic performance through classification algorithms ID3 and C4.5.**

## Kalpesh Adhatrao et. al. studied the problem of how to manage the performance of the students with low efficiency through the prediction of their future results with data mining techniques. They used the algorithms ID3 and C4.5, the second one used with the purpose of complement ID3 deficiencies. They obtained successful results, achieving a 72.275% of accuracy. [3]

**2.4 Prediction of students’ final results through the algorithm of Classification And Regression Trees (CART).**

## Julianti Kasih, Mewati Ayub and Sani Susanto in their research were looking for a more practical way to predict the results of Indonesian students, which they classified in three different levels. Previously, they used discriminant analysis, where they concluded that was unpractical for the kind of problem they were studying. In this new paper, they were based in the CART method to explain its efficiency in prediction. Their research did not work with known results, since its purpose was to explain the differences between the methods that they used to achieve a simpler functional algorithm. [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 |

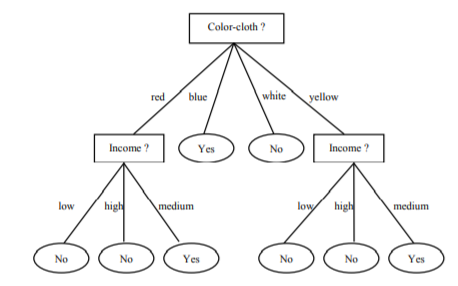
## **Tabla 1.** Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

## **3.2 Alternative ways of algorithms decision trees**

## Now, we are going to show different algorithms used to automatically build a binary decision tree.

**3.2.1 Iterative Dichotomiser 3 (ID3)**

Iterative Dochotomiser 3 (ID3), created by Ross Quinlan, is the most used algorithm to generate decision trees. It consists of a measure called entropy, which is the responsible of finding the amount of uncertain information. This makes of ID3 a greedy heuristic. This method has complications when try to work with information of continuous intervals and, by its composition, it doesn’t guarantee an optimal solution. Its complexity, in the worst scenario, is from O(h), where h is the height of the tree. [5]

The following picture represents how the algorithm works

**Figure 1:** ID3’s performance. [5]

**3.2.2 C4.5**

C4.5 is an algorithm also created by Quinlan, that searches to cover the ID3 problems. C4.5 avoid some steps that ID3 uses and focus in other options to get a better base condition. Another of C4.5 advantages is that it allows to work with continuous variables, which means that it extends its area of application. Because it uses only one recursively-call, its complexity must be O(h).

**3.2.3 Classification and regression trees (CART)**

It is a term to study both analysis of classification and regression trees. The classification trees are in charge of deliver a result of the same type of variable that is in the sample, while the regression trees are in charge of work with variables that have real values, like the students’ grades or the work time spent in an office.

**3.2.4 Random Forests**

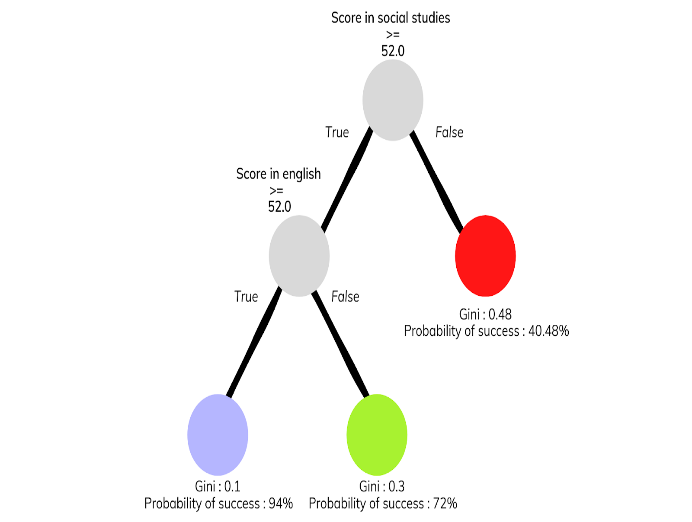
## Random Forests is an algorithm used to create multiple decision trees and get the one with the best results through a selection of a random subset of variables in the sample. Although it seems counterintuitive think that this method is efficient, in practice random forests helps to find and rank the most important variables in a regressive or a classification problem, which we will focus in this project. The complexity of the random forests can become to be O(k\*nlog(n)), where n is the number of registers and k is the number of variables.

## **4. DISEÑO DE LOS ALGORITMOS**

## En lo que sigue, explicamos la estructura de los datos y los algoritmos utilizados en este trabajo.

## **4.1 Estructura de los datos**

## Explique la estructura de datos utilizada para hacer la predicción y haga una figura que la explique. No utilice imágenes de Internet. *(En este semestre, la estructura de datos es un árbol de decisión binario)*



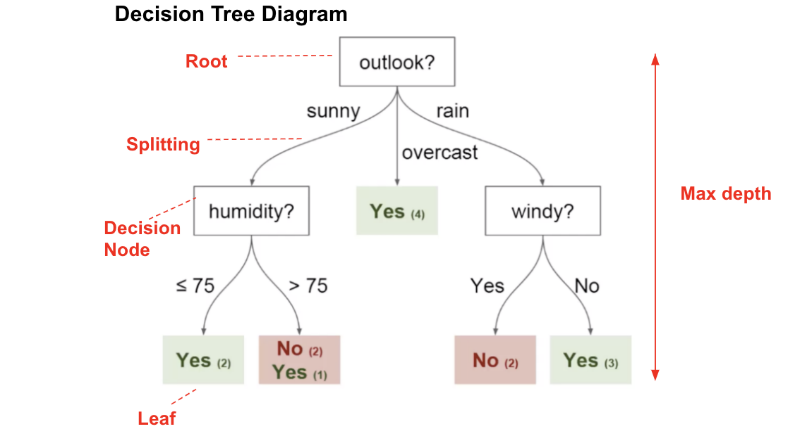
**Figura 1:** Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos violetas representan a aquellos con una alta probabilidad de éxito, los verdes con una probabilidad media y los rojos con una baja probabilidad de éxito.

**4.2 Algoritmos**

Explica el diseño del algoritmo para resolver el problema y haz una figura. No uses figuras de Internet, haz las tuyas propias. *(En este semestre, un algoritmo debe ser un algoritmo para entrenar un algoritmo de árbol de decisión como ID3, C4.5, CART y el segundo algoritmo debe ser un algoritmo para clasificar los nuevos datos utilizando dicho árbol).*

**4.2.1 Entrenamiento del modelo**

Explique, brevemente, cómo entrenó a la modelo: Esto equivale a explicar cómo su algoritmo construye automáticamente un árbol de decisión binario.



**Figura 2:** Entrenamiento de un árbol de decisión binario usando *(En este semestre, uno podría ser CART, ID3, C4.5... por favor, elija)*. En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

**4.2.2 Algoritmo de prueba**

Explique, brevemente, cómo probó el modelo: Esto equivale a explicar cómo su algoritmo clasifica los nuevos datos después de que se construya el árbol.

**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.](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.

# **REFERENCES**

1. Mesarić, J., Šebalj, D. Decision trees for predicting the academic success of studens, *Croatian Operational Research Review, (7).* Retrieved 2016, from University of Josip Juraj Strossmayer in Osijek.

2. Al-Radaideh, Q., Al-Shawakfa, E., and Al-Najjar, M. Mining Student Data Using Decision Trees. In *The 2006 International Arab Conference on Information Technology, (Jordan, 2006)*.

3. Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R. and Honrao, V. Predicting Stundets’ Performance Using ID3 and C4.5 Classification Algorithms. *International Journal of Data Mining & Knowledge Management Process 3 ().* Retrieved September, 2013, from Fr. C.R.I.T, Navi Mumbai, Maharashtra, India.

4. Kasih, J., Ayub, M. and Susanto, S. Predicting student’s final passing results using the Classification and Regression Trees (CART) algorithm. *World Transactions on Engineering and Technology Education, 11 (1).* Retrieved on 2013, from Maranatha Christian University and Parahyangan Catholic University.

5. De-lin, L., Jin, C. and Fen-xiang, M. An Improved ID3 Decision Tree Algorithm. *Proceedings of 2009 4th International Conference on Computer Science & Education.*