

Predicción del éxito estudiantil con árboles de decisión



Presentación del Equipo



Julian Andres
Ramirez
Jimenez



Samuel Villegas
Bedoya



Miguel
Correa



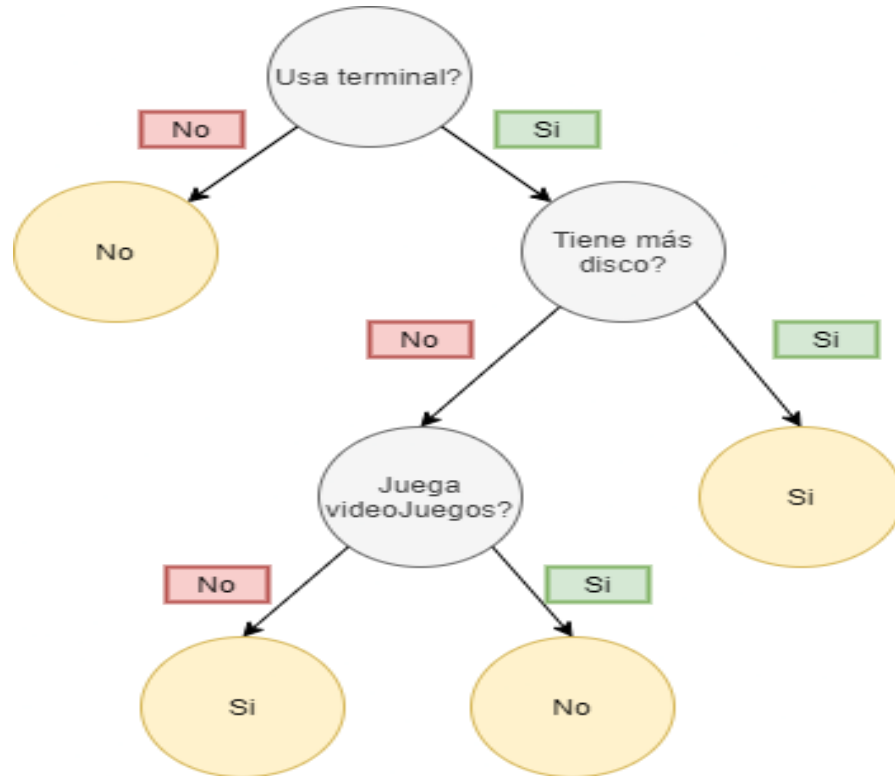
Mauricio
Toro



<http://github.com/sdvillegab/proyecto/>



Diseño del Algoritmo



Árbol resultante del estudio de la tabla con la utilización del algoritmo CART.

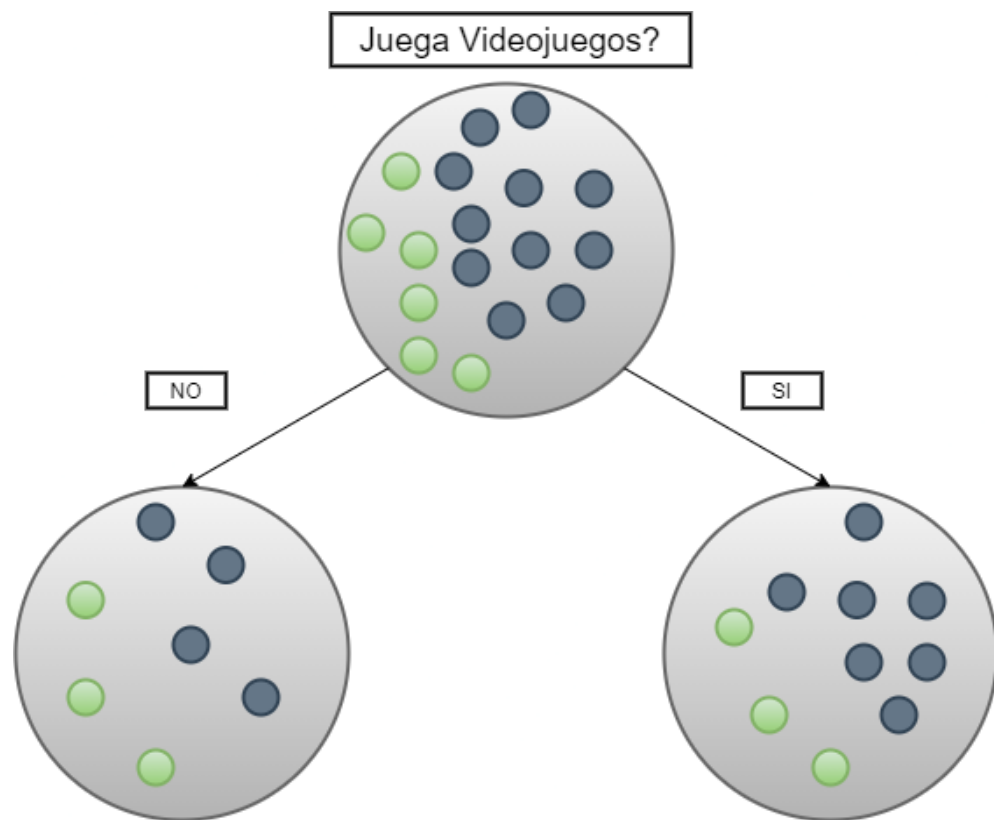
Usa terminal	Juega videojuegos	Tiene mas disco	Gusto
Si	Si	Si	Si
Si	Si	No	No
Si	No	No	Si
No	No	No	No
Si	No	Si	Si
No	No	Si	No
No	Si	Si	No
Si	Si	No	No
No	No	Si	No
No	No	No	No
Si	Si	No	No
Si	Si	Si	Si
No	Si	Si	No
Si	No	Si	Si
Si	Si	No	No
Si	Si	Si	Si
No	Si	No	No

Data para la realización de ejemplo sobre si una persona usara el SO Linux o no, por medio del algoritmo CART.

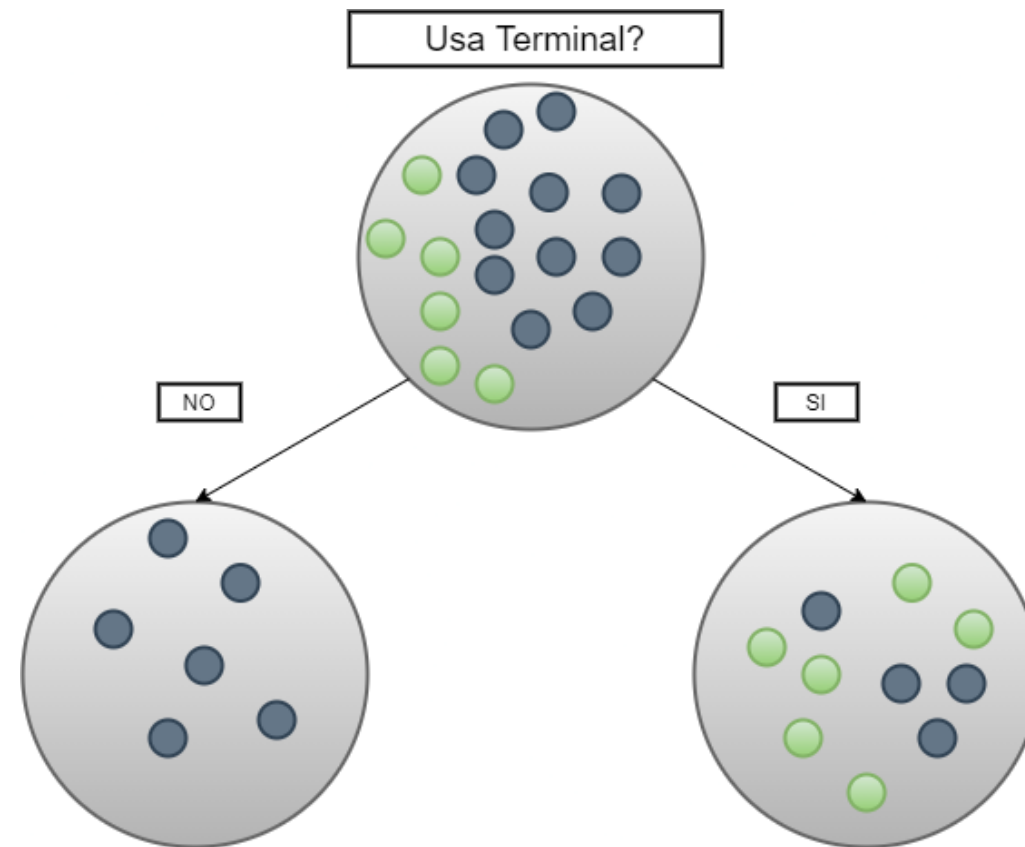
En esta parte tenemos un ejemplo básico de una implementación del algoritmo CART, este consiste en realizar un árbol de decisión con los siguientes pasos.

- Separar las reglas o condiciones que se utilizan para la realización del árbol.
- Hallar el índice de Gini ponderado de las condiciones existentes.

División de un nodo



Esta división está basada en la condición "Juega videojuegos" Para este caso, la impureza Gini de la izquierda es 0.49, la impureza Gini de la derecha es 0.42 y la impureza ponderada es de 0.44.



Esta división está basada en la condición "Usa Terminal" Para este caso, la impureza Gini de la izquierda es 0, la impureza Gini de la derecha es 0.48 y la impureza ponderada es 0.28.

Hablemos ahora del algoritmo CART, para explicarlo recordemos una famosa frase "divide y vencerás" que quizás la hemos escuchado múltiples veces y que tiene tanta validez en el contexto de la ingeniería de sistemas. En este caso buscamos dividir, pero también darle un valor agregado a esta división, el cual es buscar la condición de entre una serie de variables que mejor las separe a la hora de compararlas con la influencia que tienen sobre un resultado final, en otras palabras es buscar la variable-condición que se podría decir más importancia tiene en el resultado final, para esto se calcula el índice de Gini. Al hacer el cálculo de este, se escoge la variable cuyo índice de Gini sea menor a las

Complejidad del Algoritmo



Conserven ese título

Completen esta lámina
en la tercera entrega



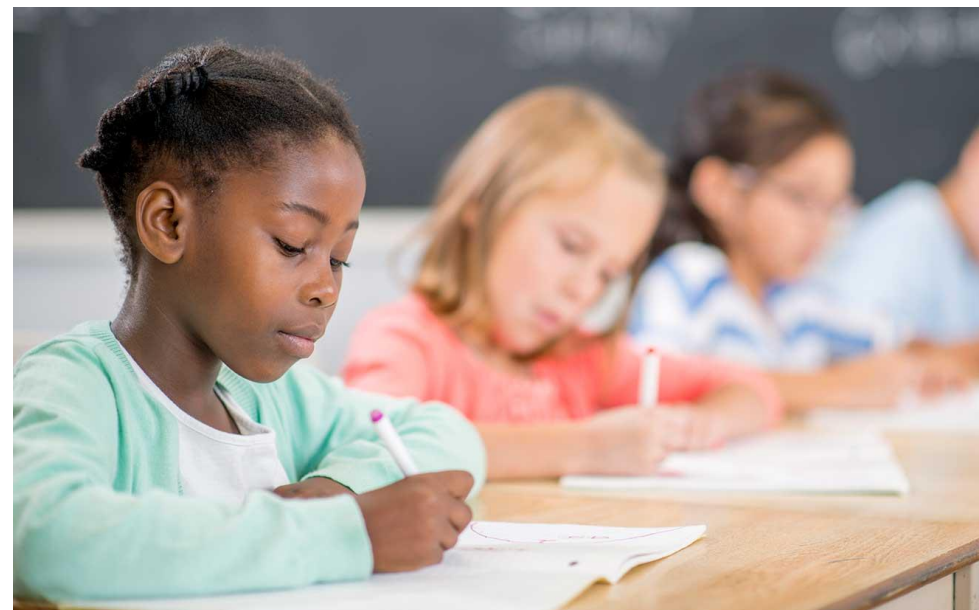
Creen esta tabla en Powerpoint. ¡No
copien pantallazos pixelados del porte
aquí!

	Complejidad en tiempo	Complejidad en memoria
Entrenamiento del modelo	$O(N^2 * M * 2^M)$	$O(N * M * 2^M)$
Validación del modelo	$O(N * M)$	$O(1)$

Complejidad en tiempo y memoria del algoritmo (En este semestre, una opción puede ser CART, ID3, C4.5, elijan uno). (Por favor, expliquen qué es N y qué es M en este problem. ¡POR FAVOR, HÁGANLO!)



Expliquen las tablas con
sus propias palabras



Incluyan una foto de alta definición
relacionada con el problema que
están modelando

Modelo de Árbol de Decisión

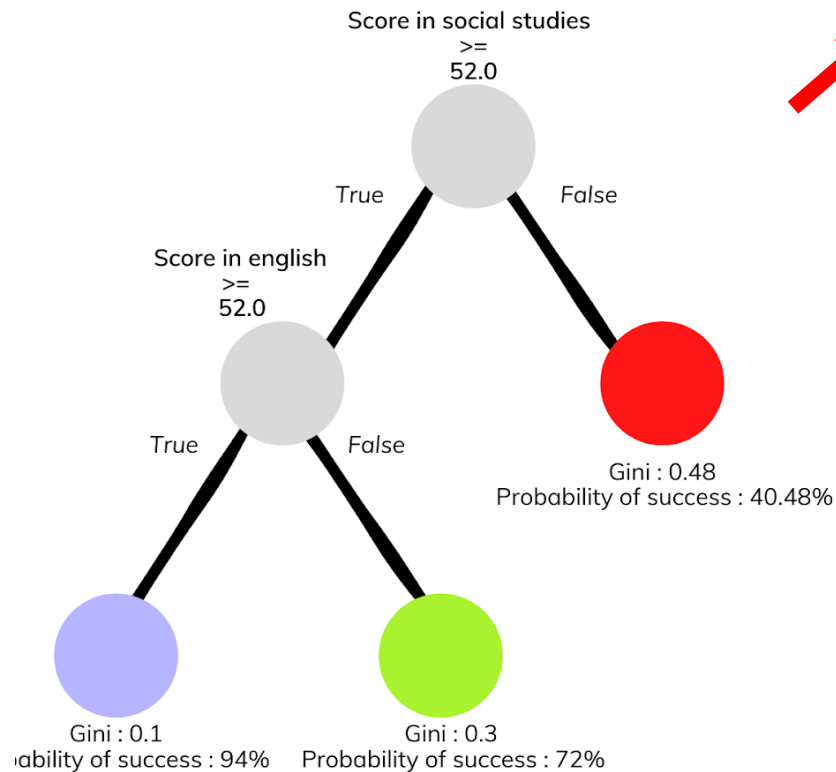
Conservan ese título

Completen esta lámina
en la tercera entrega



Usen estos colores
en sus gráficas

Crean una gráfica, en español, en
Powerpoint. ¡No copien pantallazos
pixelados del reporte técnico, por favor!



Un árbol de decisión para predecir el resultado del Saber Pro usando los resultados del Saber 11. Violeta representa nodos con alta probabilidad de éxito; verde media probabilidad; y rojo baja probabilidad.

Expliquen sus gráficos con
sus propias palabras

Características Más Relevantes



Ciencias Sociales



Inglés



Género

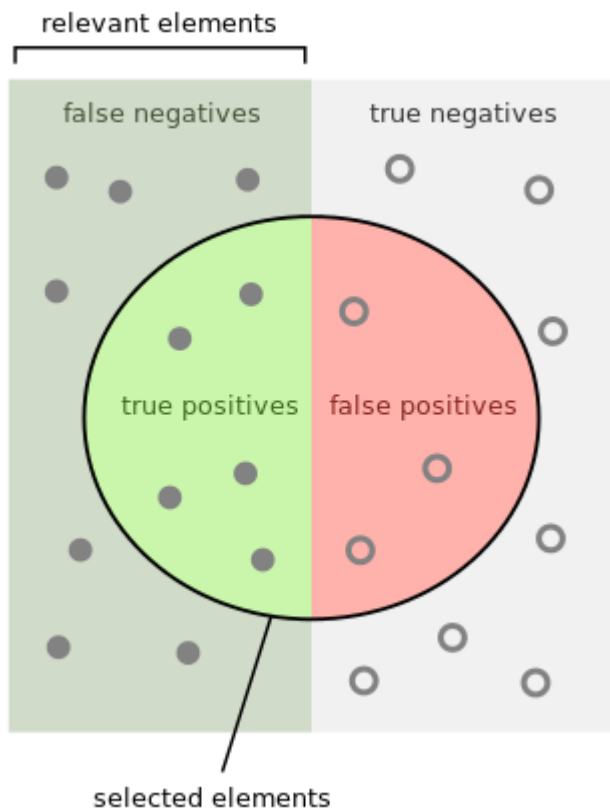
¡Usen un ícono para
representar cada
característica!

¿Es ético usar el género en
un modelo que sirve para
predecir el éxito académico?

Métricas de Evaluación

Conserven ese título

Completen esta lámina
en la tercera entrega



Usen gráficas vectorizadas, en español,
para explicar las métricas de evaluación,
de esa forma no les quedará pixelado
como las mías

How many selected
items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant
items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Expliquen la exactitud
también....
De la misma manera

Usen estos colores
para sus gráficas

Si es posible, eviten usar
ecuaciones para explicar simples
conceptos que se pueden explicar
con diagramas coloridos

Métricas de Evaluación

Conserven ese título

Completen esta lámina
en la tercera entrega



Crean la tabla en Powerpoint. ¡No copien
pantallazos pixelados del reporte, por
favor!

	Conjunto de entrenamiento	Conjunto de validación
Exactitud	0.8	0.62
Precisión	0.6	0.55
Sensibilidad	0.76	0.61

Métricas de evaluación obtenidas con el conjunto de datos de
entrenamiento de 135,000 estudiantes y el conjunto de datos
de validación de 45,000 estudiantes.



Incluyan otra gráfica en alta
definición relacionada con el
problema que están resolviendo.

Expliquen las tablas con sus
propias palabras

Consumo de tiempo y memoria

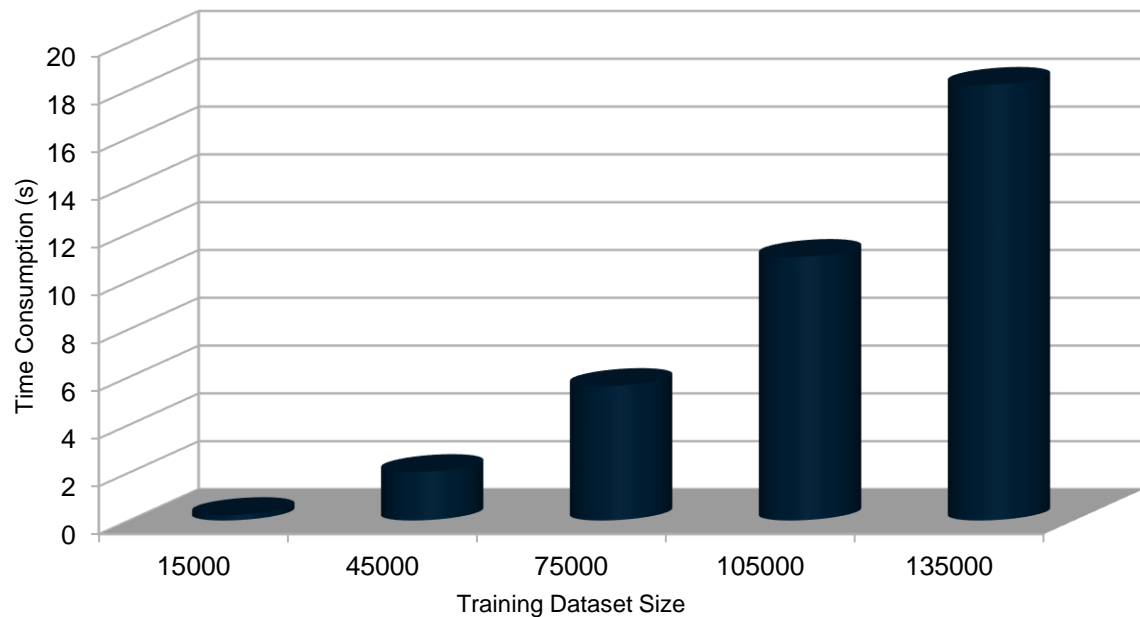
Conserven ese título

Completen esta lámina
en la tercera entrega

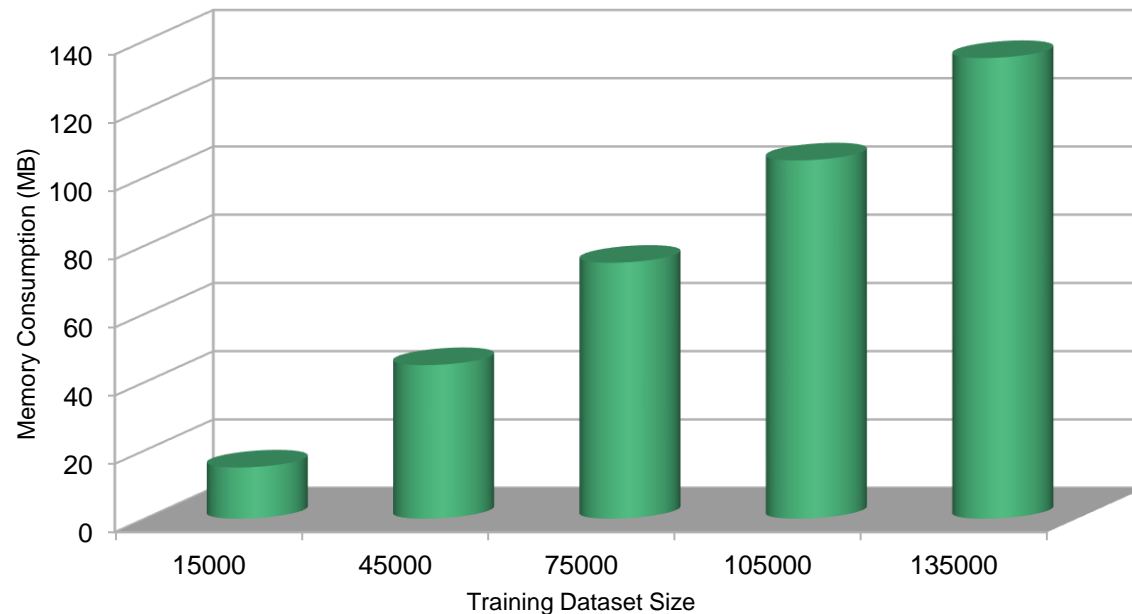


Creen las gráficas en Excel en español. ¡No
tomen pantallazos pixelados del reporte!

Usen estos colores
para sus gráficas



Consumo de tiempo



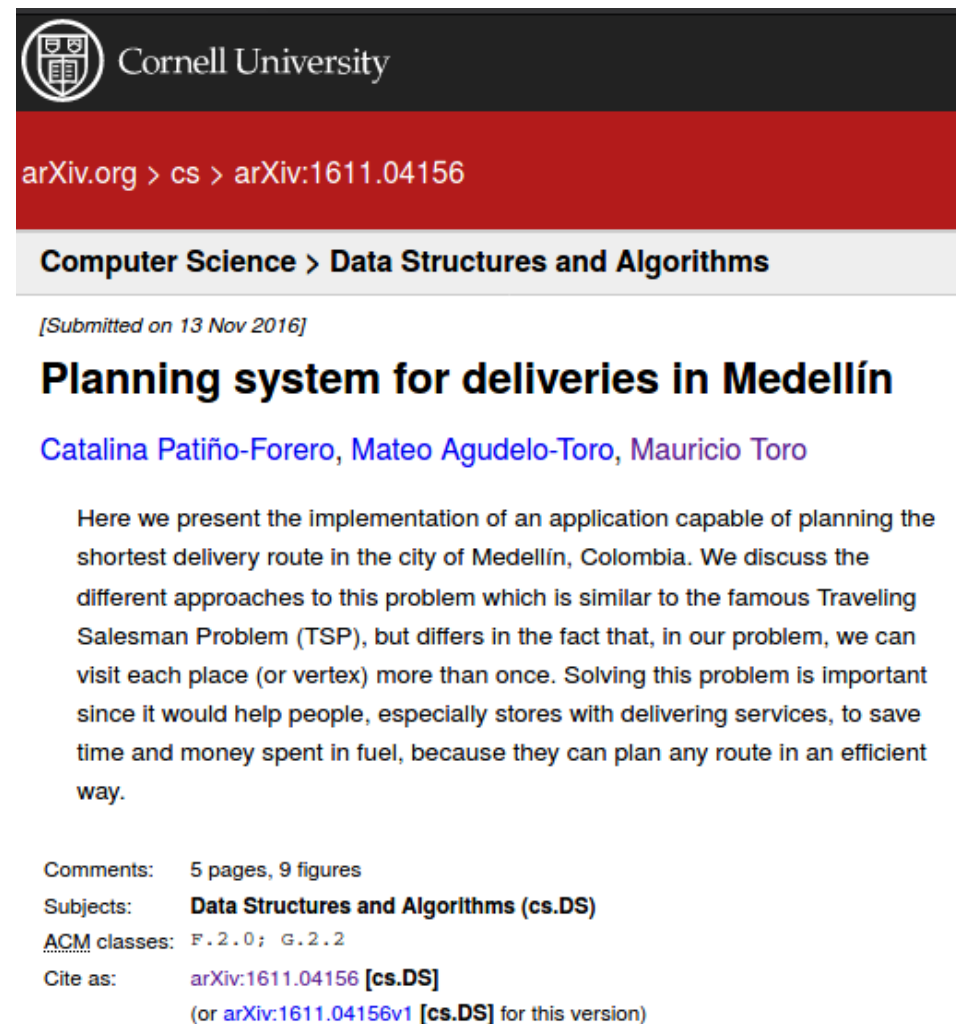
Consumo de memoria



Incluyan la citación del reporte en arXiv y
su vínculo comose muestra abajo

C. Patiño-Forero, M. Agudelo-Toro, and M. Toro. Planning
system for deliveries in Medellín. ArXiv e-prints, Nov. 2016.
Available at: <https://arxiv.org/abs/1611.04156>

Incluyan un
pantallazo



Cornell University

arXiv.org > cs > arXiv:1611.04156

Computer Science > Data Structures and Algorithms

[Submitted on 13 Nov 2016]

Planning system for deliveries in Medellín

Catalina Patiño-Forero, Mateo Agudelo-Toro, Mauricio Toro

Here we present the implementation of an application capable of planning the shortest delivery route in the city of Medellín, Colombia. We discuss the different approaches to this problem which is similar to the famous Traveling Salesman Problem (TSP), but differs in the fact that, in our problem, we can visit each place (or vertex) more than once. Solving this problem is important since it would help people, especially stores with delivering services, to save time and money spent in fuel, because they can plan any route in an efficient way.

Comments: 5 pages, 9 figures

Subjects: **Data Structures and Algorithms (cs.DS)**

ACM classes: F.2.0; G.2.2

Cite as: [arXiv:1611.04156](https://arxiv.org/abs/1611.04156) [cs.DS]
(or [arXiv:1611.04156v1](https://arxiv.org/abs/1611.04156v1) [cs.DS] for this version)

Completen esta lámina
en la tercera entrega

Digan gracias por
escucharnos

¡GRACIAS!

