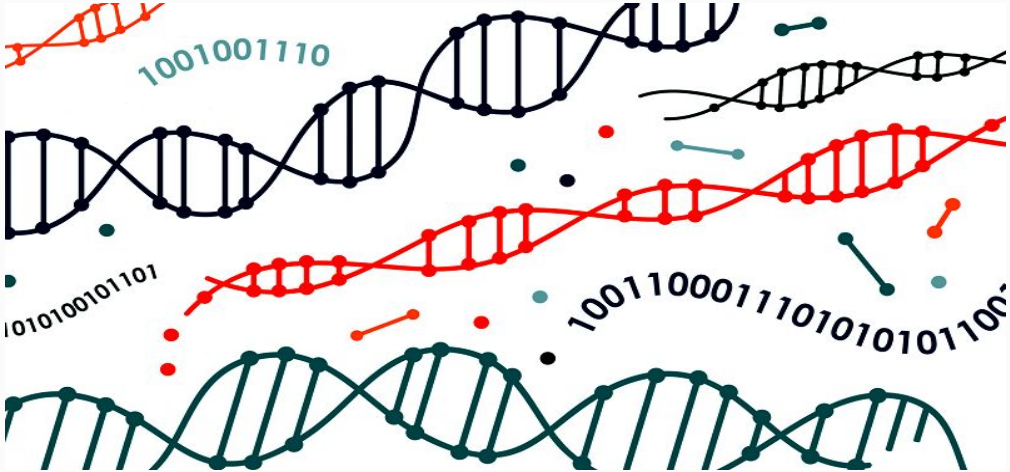


Selección de atributos en Aprendizaje de Máquina con Algoritmos Genéticos

Diplomado de Desarrollo de Aplicaciones con inteligencia Artificial

Agosto 2020

Autor: Jorge Rodríguez Castillo

CONTEXTO

La alta dimensionalidad en un conjunto de datos puede afectar el desempeño en algoritmos de aprendizaje de máquina. Por lo cual algunas de las razones para reducir la dimensionalidad de datos son:

- Nos interesa identificar y eliminar atributos irrelevantes.
- No necesariamente usar la mayor cantidad de variables nos brinda un mejor modelo.
- Se mejora el rendimiento computacional.
- Se reduce la complejidad, facilita la comprensión del modelo y sus resultados.

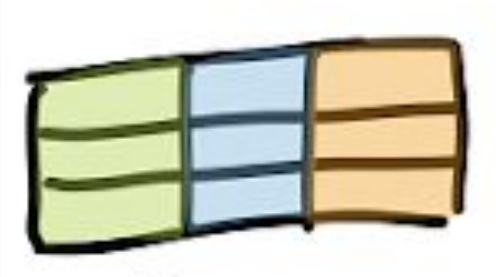
PROBLEMA

¿De qué manera con el uso de algoritmos genéticos se pueden seleccionar atributos para un modelo de aprendizaje de máquina?

HIPÓTESIS

Mediante el uso de algoritmos genéticos con representación binaria, se encuentra un subconjunto de atributos relevantes para un modelo de aprendizaje de máquina.

CONJUNTO DE DATOS



CONJUNTO DE DATOS

El conjunto de datos usado es Bank Marketing Data Set, está relacionado con campañas directas de marketing de un banco Portugués.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

16 Atributos
(6 numéricos, 10 categóricos)

45,210 registros

*Se realizó un pre procesamiento de los datos

TÉCNICA BASE

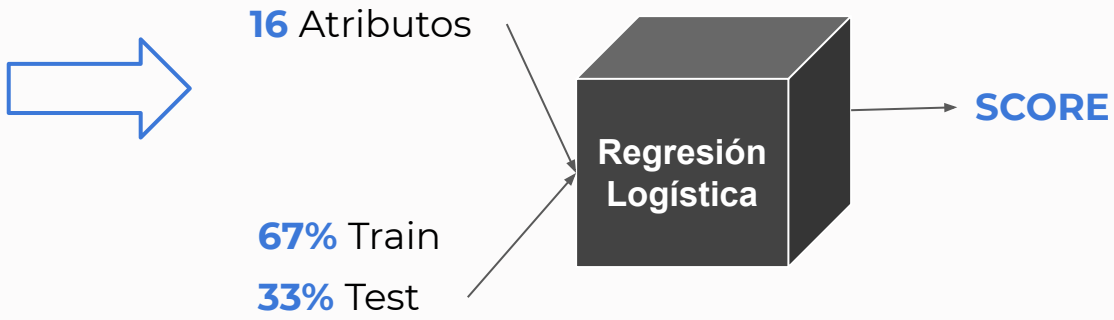
TÉCNICA BASE

Una técnica sencilla para realizar tareas de clasificación binaria es el uso de Regresión Logística:

Probabilidad de cada instancia:

$$\hat{p} = h_\theta(X) = \sigma(X^T\theta)$$

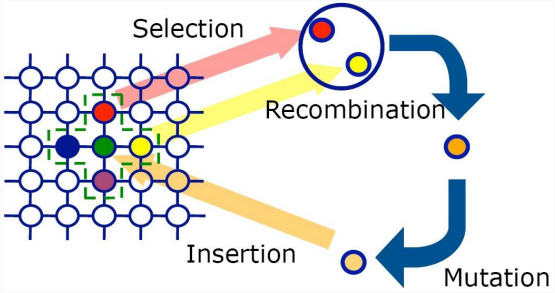
$$\sigma(t) = \frac{1}{1+e^{-t}}$$



Clasificación de cada instancia:

$$\hat{y} = \begin{cases} 0 & \text{si } \hat{p} < 0.5 \\ 1 & \text{si } \hat{p} \geq 0.5 \end{cases}$$

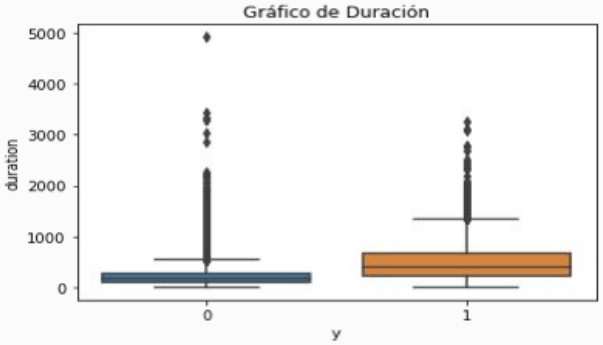
ENFOQUE PROPUESTO

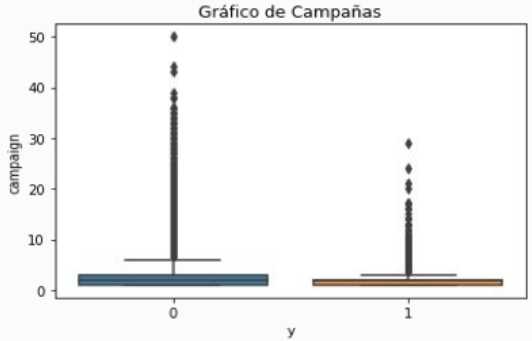


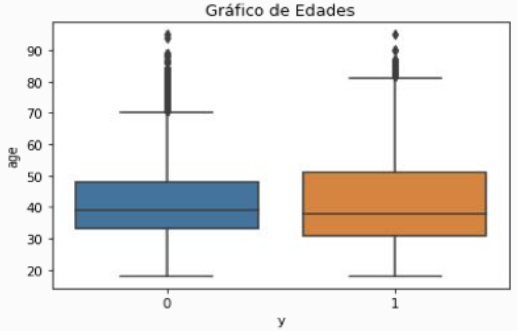
ENFOQUE PROPUESTO

Previo a usar el algoritmo genético, se realiza un pequeño pre procesamiento

- Eliminamos datos con valores ‘unknown’, luego de esto el conjunto de datos pasa de tener **45,210** a **30,907**
- A los outliers le cambiamos actualizamos sus valores con la media de los datos, lo aplicamos sobre las columnas **‘duration’, ‘campaign’, ‘age’**



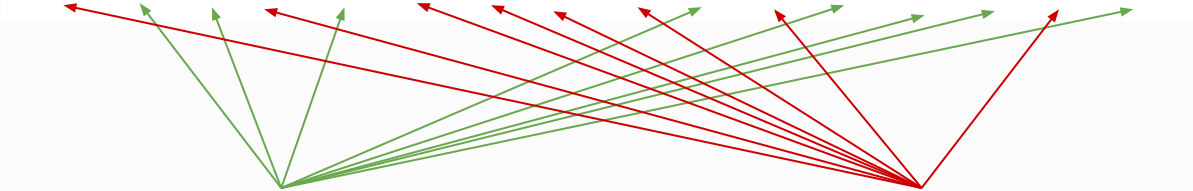




ENFOQUE PROPUESTO

El enfoque propuesto es el uso de algoritmos genéticos mediantes representación binaria.

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
0	1	1	0	1	0	0	0	0	1	0	1	1	1	0	1



Atributos seleccionados

$\{x_1, x_2, x_4, x_9, x_{11}, x_{12}, x_{13}, x_{15}\}$

Atributos no seleccionados

$\{x_0, x_3, x_5, x_6, x_7, x_8, x_{10}, x_{14}\}$

ENFOQUE PROPUESTO

La representación de la población inicial son 50 individuos **(cromosomas)**.
Inicialmente los fitness no son evaluados **(-1)**

Individuos	Cromosoma
Individuo 1	[0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0]
Individuo 2	[1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0]
Individuo 3	[1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1]
…	…
…	…
Individuo n	[1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1]

ENFOQUE PROPUESTO

Para la selección de individuos se realiza por medio de algoritmo de ***Tournament y Roulette***.

- En el caso de Tournament se realizan ***NT = 3*** y generamos 2 individuos.
- En el caso Roulette se realizan 2 veces y se obtienen los individuos en base a la sumatoria de fitness de cada individuo.

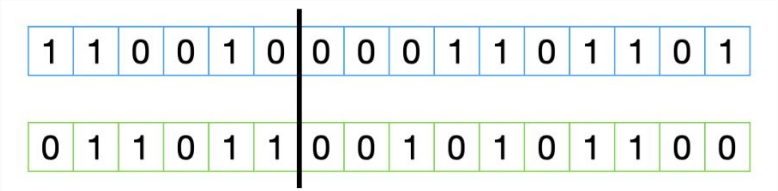
Para la selección de nueva generación se realiza por ***Elitismo***

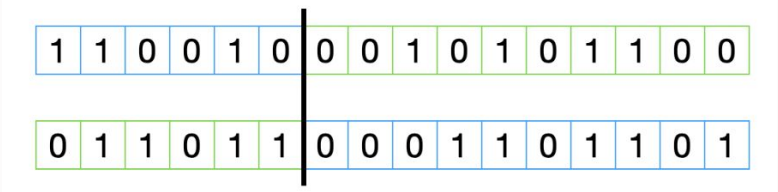
- Seleccionamos ***“n”*** individuos que tengan un mejor valor de ***fitness***

ENFOQUE PROPUESTO

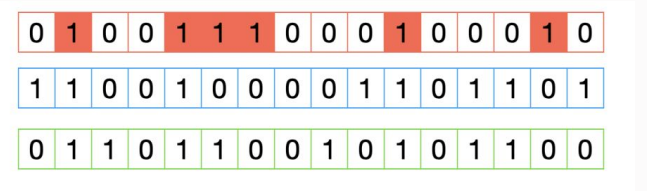
Los algoritmos de cruzamiento que se usaron fueron: OnePoint y Uniform.
Ambos con ***PC = 0.5***

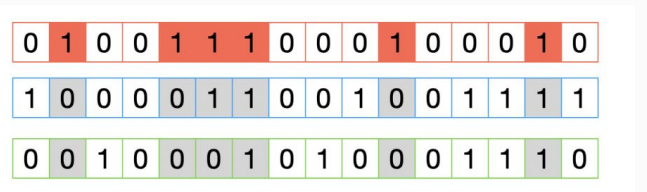
OnePoint





Uniform

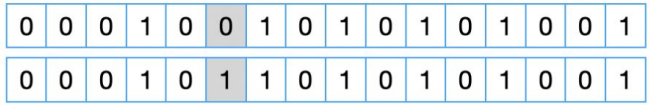




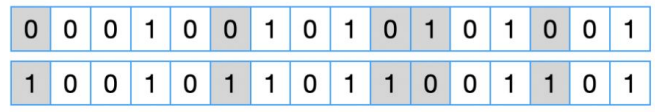
ENFOQUE PROPUESTO

Los algoritmos de mutación que se usaron fueron: Bitflip, Bitwise & Inversion.
Por defecto ***PM = 0.5***

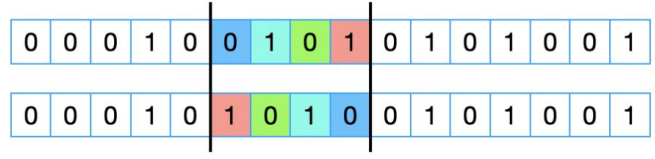
Bitflip



Bitwise



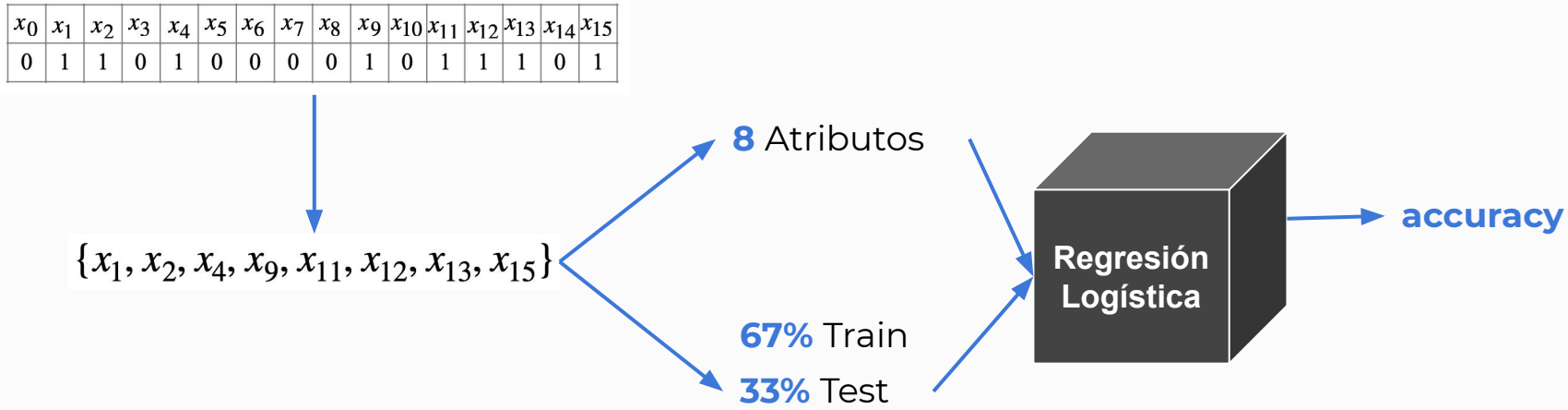
Inversion



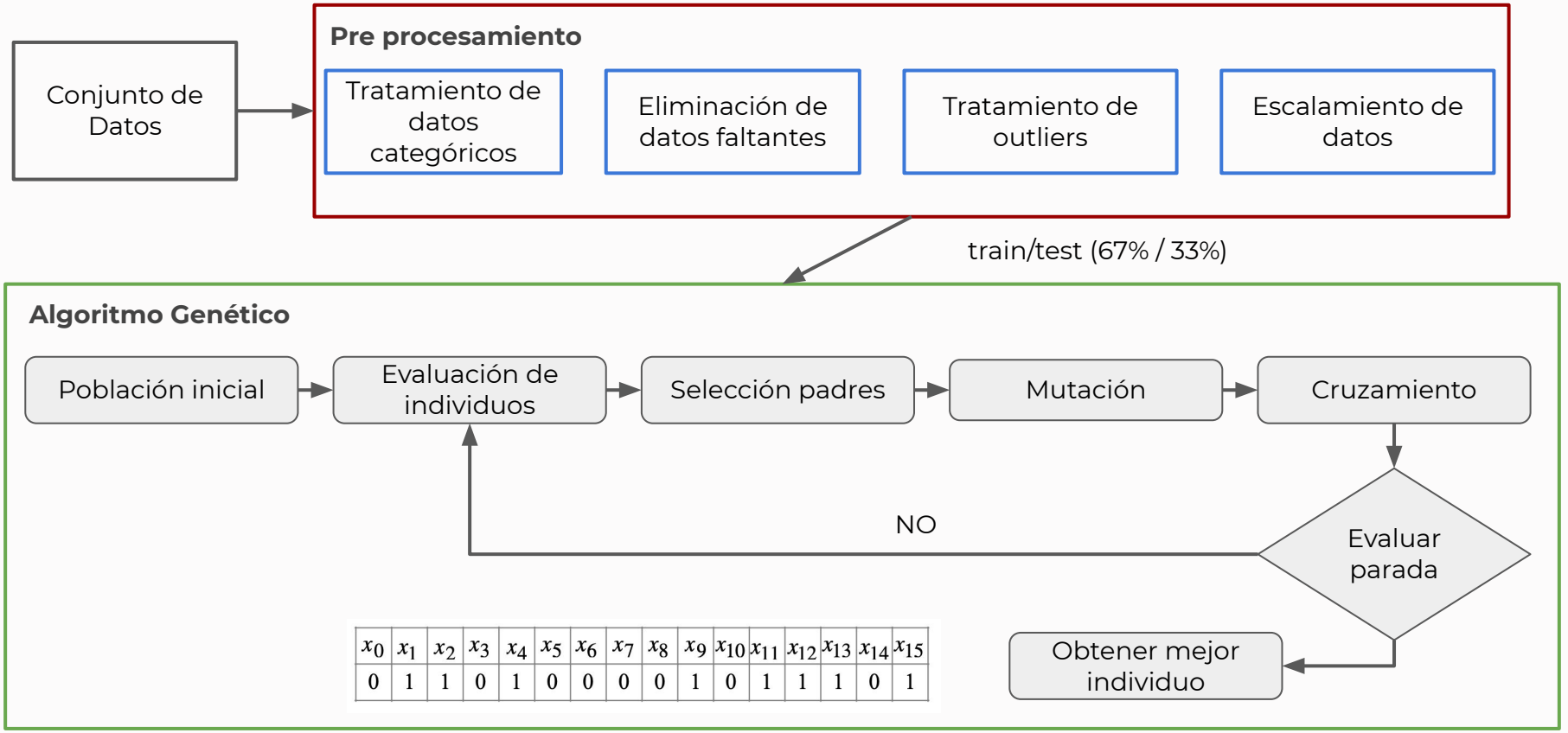
ENFOQUE PROPUESTO

La función de fitness que se usa es la métrica de exactitud: ***"accuracy_score"***

$$Acurracy = \frac{TP + TN}{TP + FP + FN + TN}$$

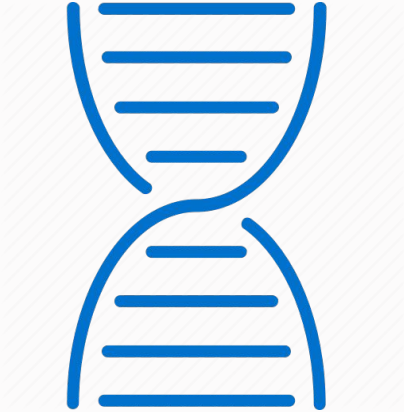


ENFOQUE PROPUESTO



x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
0	1	1	0	1	0	0	0	0	1	0	1	1	1	0	1

EXPERIMENTACIÓN Y RESULTADOS



EXPERIMENTACIÓN

Nro	Pop	Tipo de cruzamiento	Tasa de Mutación	Mutación	Padres
1	50	uniform	0.0	bitflip	roulette
2	50	uniform	0.5	bitflip	roulette
3	50	uniform	0.75	bitflip	roulette
4	50	uniform	0.0	bitwise	tournament
5	50	uniform	0.5	bitwise	tournament
6	50	uniform	0.75	bitwise	tournament
7	50	onepoint	0.0	inversion	roulette
8	50	onepoint	0.5	inversion	roulette
9	50	onepoint	0.75	inversion	roulette
10	50	onepoint	0.0	bitflip	tournament
11	50	onepoint	0.5	bitflip	tournament
12	50	onepoint	0.75	bitflip	tournament

100 Generaciones

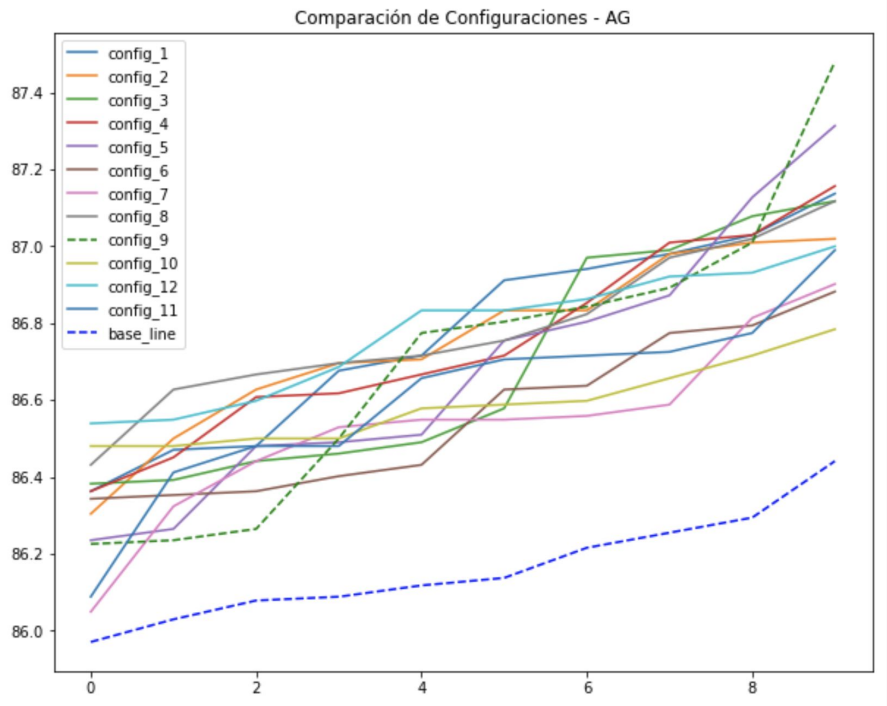
10 Ejecuciones por generación

RESULTADOS

Nro	Atributos seleccionados	Total	Fitness
0	**['age', 'job', 'marital', 'education', 'default', 'balance', 'housing', 'loan', 'contact', 'day', 'month', 'duration', 'campaign', 'pdays', 'previous', 'poutcome']**	**16**	**0.8644**
1	['marital', 'education', 'balance', 'housing', 'month', 'duration', 'pdays', 'poutcome']	8	0.8713
2	['age', 'job', 'marital', 'education', 'default', 'loan', 'contact', 'duration', 'campaign', 'pdays', 'poutcome']	11	0.8701
3	['age', 'job', 'marital', 'education', 'default', 'day', 'month', 'duration', 'campaign', 'pdays', 'previous', 'poutcome']	12	0.8711
4	['age', 'education', 'housing', 'duration', 'campaign', 'pdays', 'poutcome']	7	0.8715
5	['default', 'housing', 'loan', 'day', 'month', 'duration', 'campaign', 'pdays', 'previous', 'poutcome']	10	0.8731
6	['job', 'education', 'default', 'housing', 'loan', 'contact', 'duration', 'campaign', 'pdays', 'previous']	10	0.8679
7	['age', 'marital', 'default', 'balance', 'housing', 'loan', 'contact', 'duration', 'campaign', 'pdays', 'previous', 'poutcome']	12	0.8690
8	['education', 'housing', 'loan', 'contact', 'day', 'month', 'duration', 'pdays', 'previous', 'poutcome']	10	0.8711
9	**['age', 'marital', 'default', 'contact', 'day', 'duration', 'pdays', 'poutcome']**	**8**	**0.8748**
10	['age', 'job', 'education', 'default', 'balance', 'housing', 'loan', 'contact', 'day', 'duration', 'pdays', 'poutcome']	12	0.8678
11	['age', 'job', 'education', 'default', 'housing', 'day', 'month', 'duration', 'pdays', 'poutcome']	10	0.8699
12	['age', 'education', 'default', 'housing', 'loan', 'contact', 'day', 'duration', 'pdays', 'poutcome']	10	0.8700

Técnica base:
Regresión Logística

Técnica propuesta:
Algoritmo Genético



CONCLUSIONES

- Usando algoritmos genéticos se encontró un subconjunto de atributos que igual o mejora la métrica de exactitud de un 86.4% a 87.4%
- Reducimos la dimensionalidad de 16 a 8 atributos sin perder exactitud e interpretabilidad.
- Los algoritmos genéticos se pueden usar tanto en problemas de clasificación como de regresión
- El uso de algoritmos genéticos nos permiten adaptar la configuración de distintos parámetros para nuestro problema en específico

SUGERENCIAS

- Probar distinto modelos de aprendizaje de máquina como Árboles de Decisión, SVM, etc.
- Usar otra métrica para obtener el fitness de los individuos, por ejemplo usar log_loss en el caso de Regresión Logística.
- Usar técnicas como SAGA-II para enfocar el problema como multi-objetivo.

GRACIAS