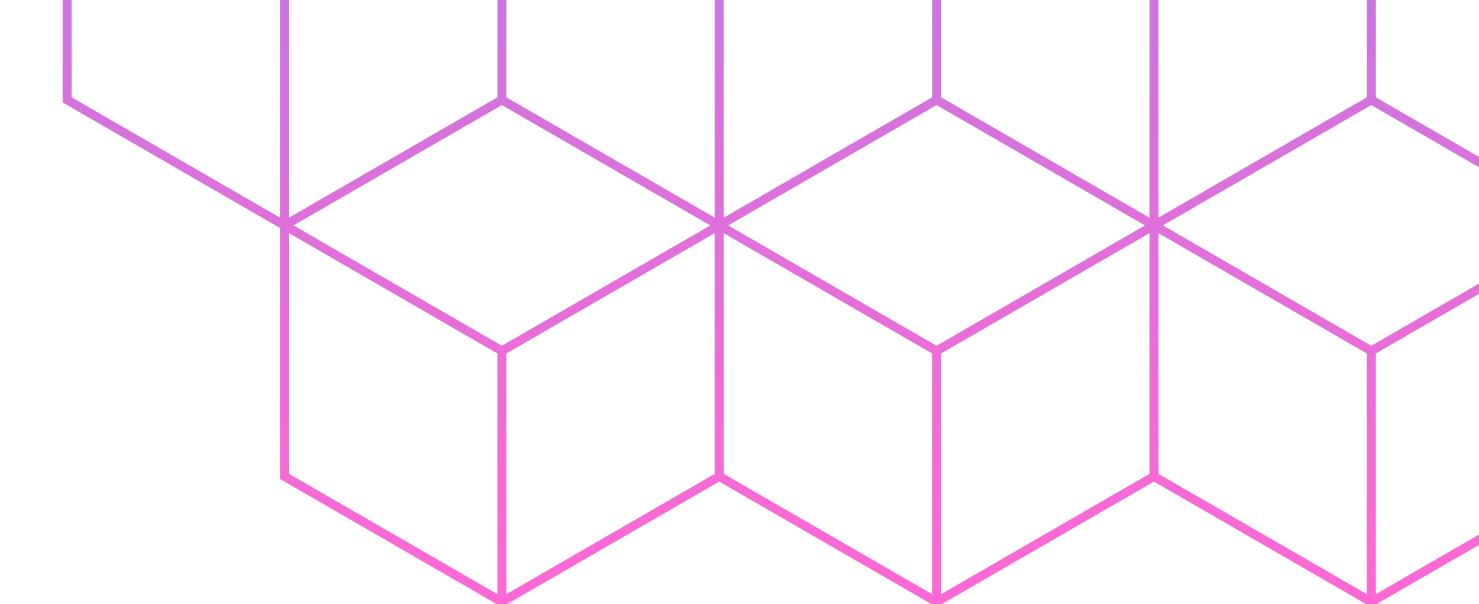


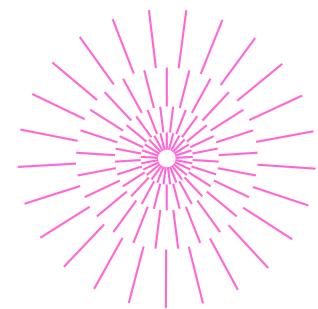
Data Science
Camada #25570
Dovale, Feü, Palacio, Parada



Modelo Machine Learning basado en Clasificación para el Otorgamiento de Crédito

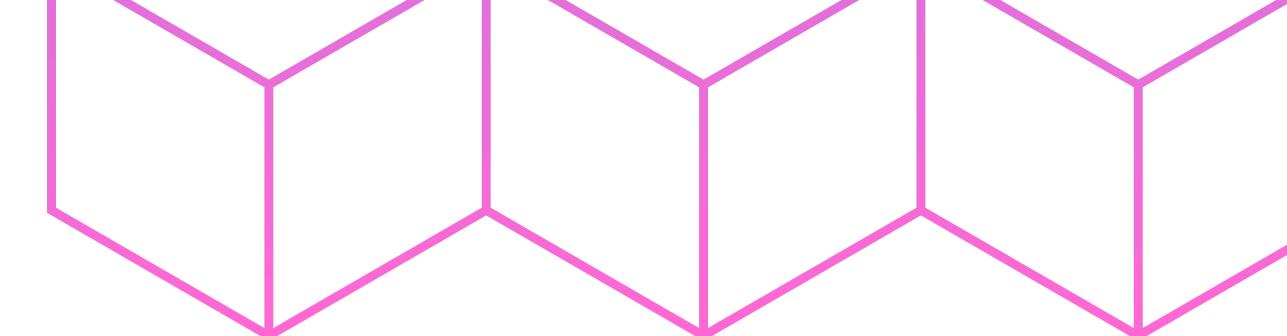
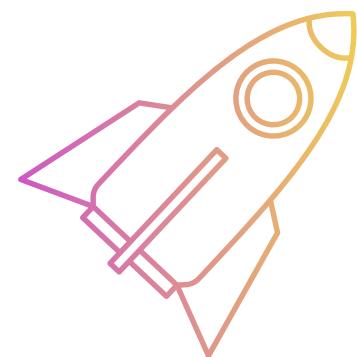
Integrantes: Dovale, Feü, Palacio, Parada

CODER HOUSE



Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

The Team



MARIA DOVALE
Innovation Lead



LAURA FEU
Data & Analytics Developer



GABRIEL PALACIO
Data Visualization



LAURA PARADA
Data Analyst

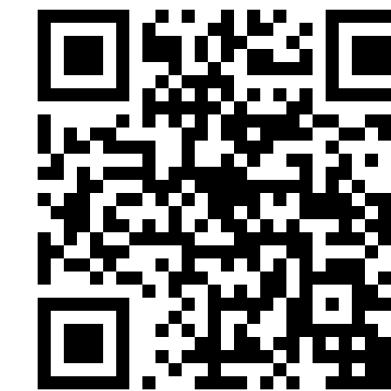


Tabla de contenidos

Descripción del caso

Tabla de versionado

Palabras clave

Objetivos del modelo

Descripción de los datos

Consideraciones

Hallazgos encontrados por el EDA

Algoritmo elegido

Métricas de desempeño

Iteraciones de Optimización

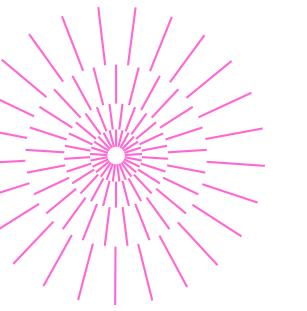
Métricas finales del Modelo Optimizado

Conclusiones

Futuras líneas

TOC

Descripción del caso



Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

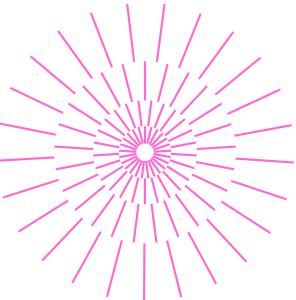
Existe una problemática común en la industria bancaria en cuanto al otorgamiento de crédito para personas naturales, estos usuarios al solicitar un préstamo muchas veces deben esperar por tiempos prolongados para poder obtener una respuesta por parte del banco y en ocasiones la necesidad de conocer esta respuesta es de carácter urgente.

Con el fin de acelerar estos tiempos y semi-automatizar estos procesos en la industria bancaria, hemos decidido crear un modelo para el otorgamiento de créditos bancarios que tenga en cuenta las características individuales de cada usuario que solicita el crédito.

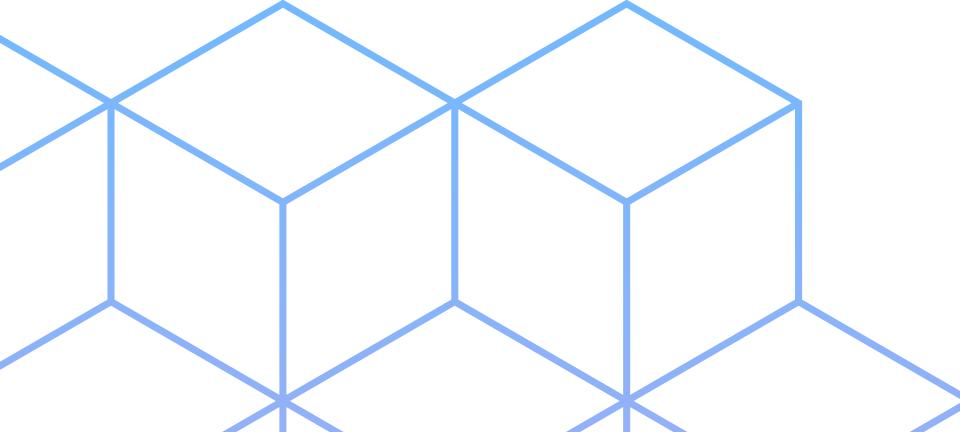
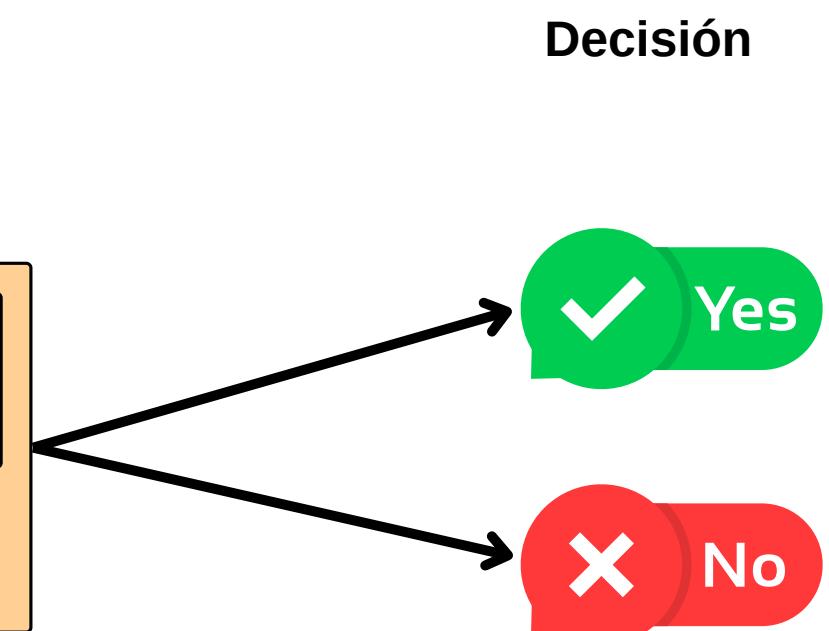
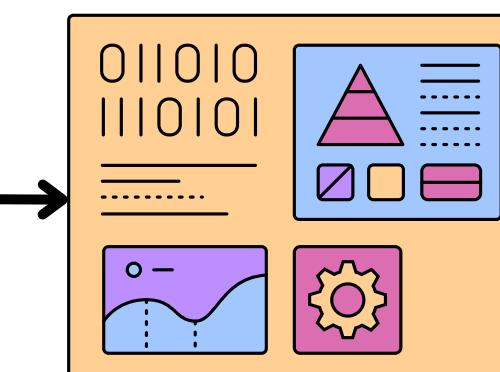
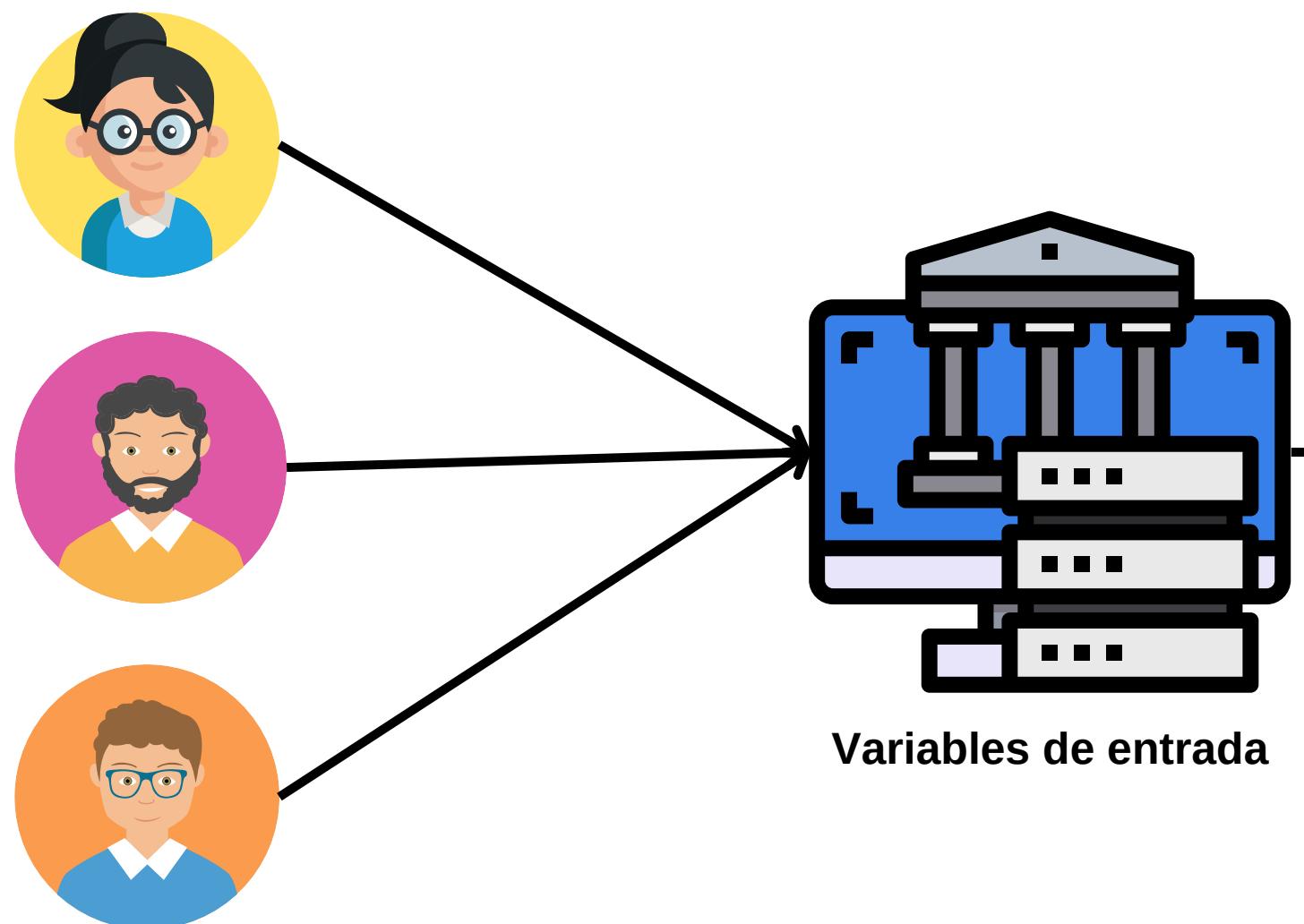
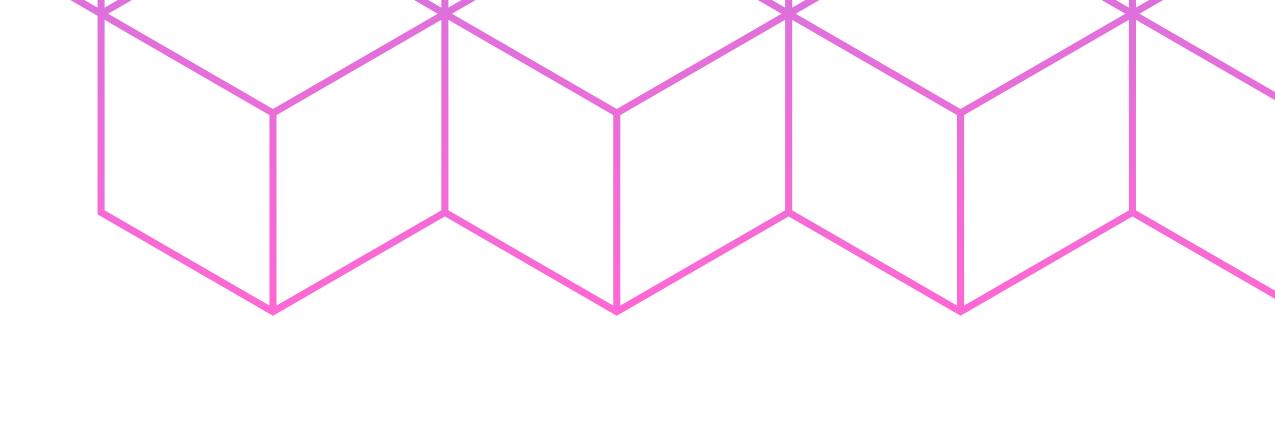


[BACK TO AGENDA](#)

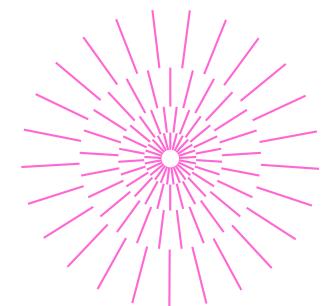
Descripción del caso



Data Science
Camada #25570
Dovale, Feü, Palacio, Parada



BACK TO AGENDA

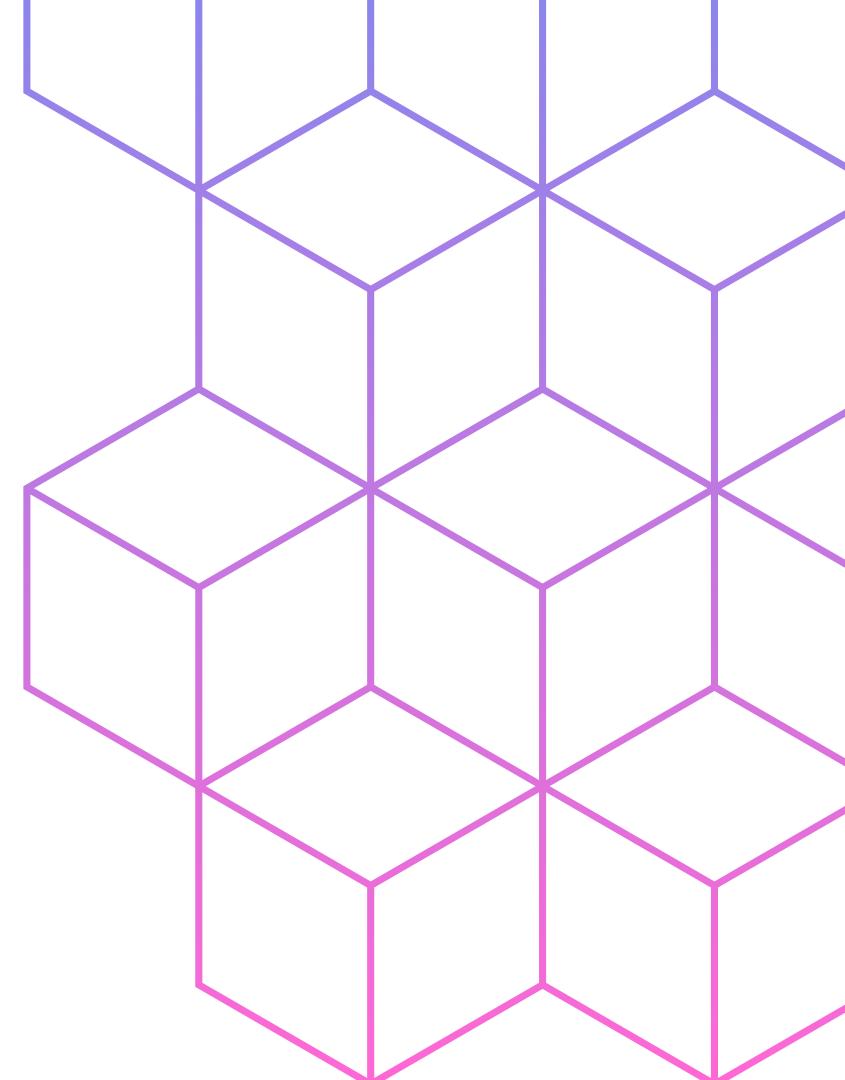


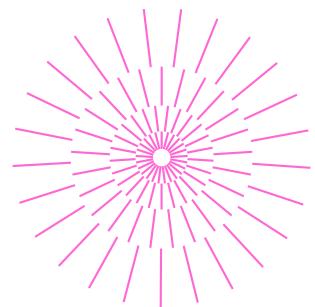
Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

Key Words

Machine Learning, Modelo Scoring, riesgo de crédito, otorgamiento de crédito, sistema financiero.

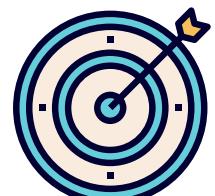
[BACK TO AGENDA](#)



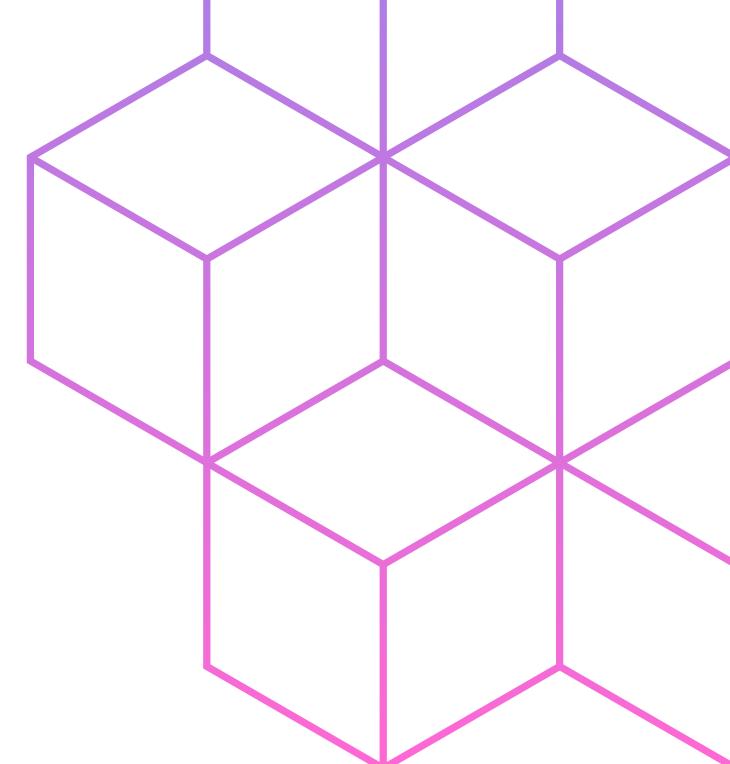


Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

Objetivo General



Crear un modelo de Machine Learning que determine si un usuario es candidato favorecido o no para recibir un préstamo bancario.

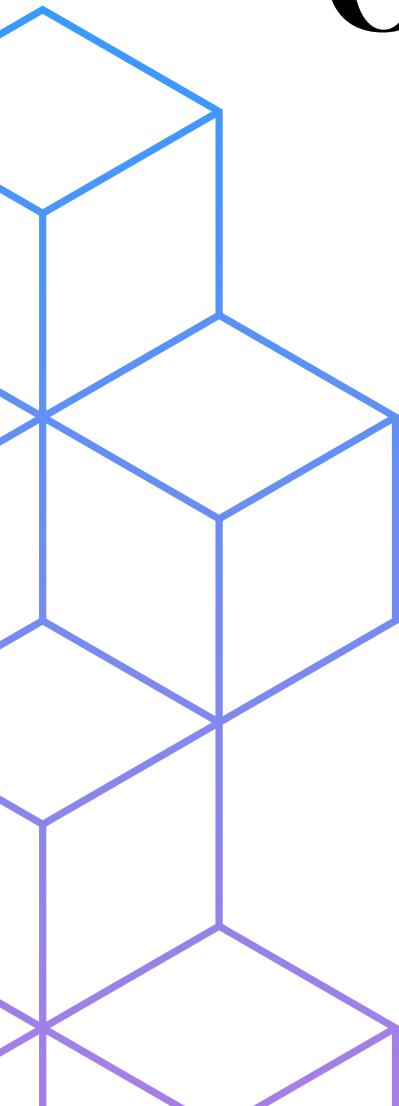


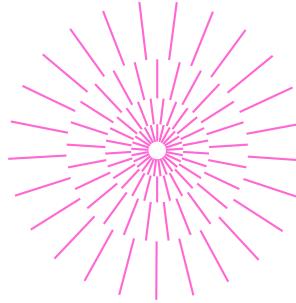
Objetivos Específicos



- Analizar varios datasets con información bancaria para elegir uno que cumpla con la mayoría de las variables de entrada más relevantes para los asesores comerciales.
- Realizar un EDA al dataset seleccionado.
- Realizar un análisis univariado, bivariado y multivariado del set de datos.
- Realizar una clasificación para el diagnóstico de otorgamiento de crédito a partir de datos de entrada
- Brindar la respuesta de otorgamiento del crédito como una variable decisora (YES/NO)

[BACK TO AGENDA](#)





Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

Descripción de los Datos

En la búsqueda de un dataset con el cuál trabajar encontramos 3 relacionados con la temática crediticia y préstamos:

01

bank_df_1.csv

02

credit_risk_dataset.csv

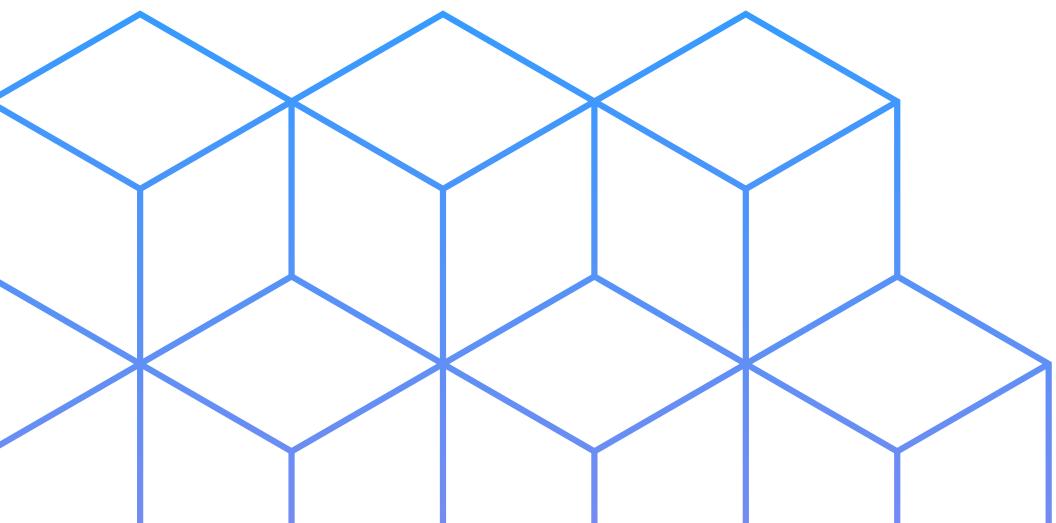
03

Loan.csv

Se procedió a realizar una exploración de cada uno de ellos para identificar cuál de los tres cumplía con la mayor cantidad de variables relevantes a la hora de hacer un análisis crediticio.

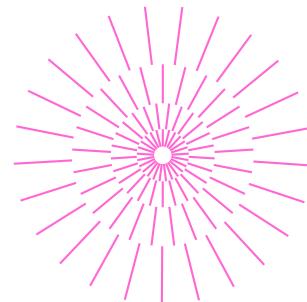
Tomamos cada dataset y vemos cómo está compuesto:

- Estructura del dataframe (rows and columns)
- Conteo del total de registros
- Nombres de las columnas
- Describe (para ver datos estadísticos)
- Tipo de dato
- Verificar si hay datos nulos
- Datos únicos



BACK TO AGENDA

Descripción de los Datos



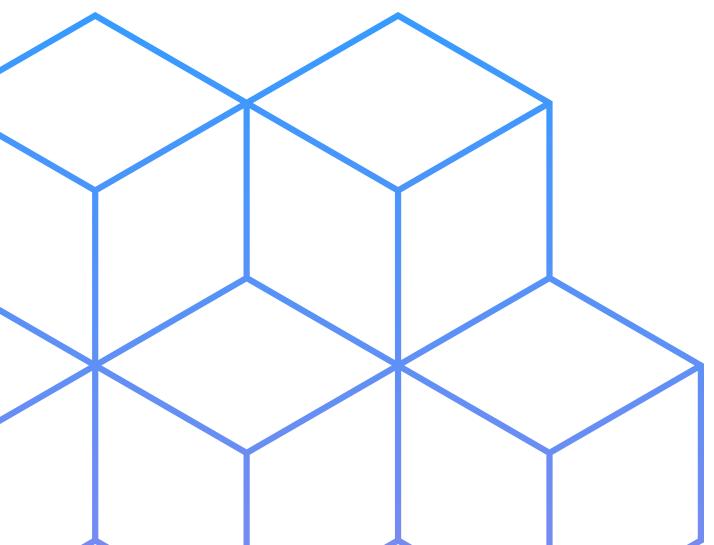
Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

	DATASET 1	DATASET 2	DATASET 3
Nombre Dataset	<u>bank_df_1.csv</u>	<u>credit_risk_dataset.csv</u>	<u>Loan.csv</u>
Cantidad de filas	438557	32581	148670
Columnas	18	12	37

Luego revisamos las variables más relevantes a la hora de solicitar un crédito y las comparamos con todas las que nos brinda cada dataset, de esta forma definimos con cuál trabajar teniendo en cuenta también la tabla anterior de cómo estaban compuestos y la cantidad de datos brindados.

Para hacer el siguiente análisis tuvimos en cuenta la siguiente bibliografía:

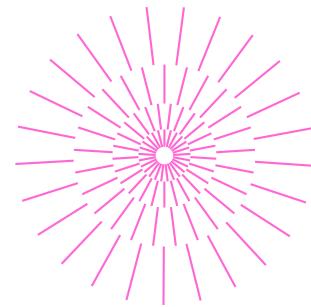
[Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera](#)
el cual sugiere las variables adecuadas a tener en cuenta al momento de tomar una decisión de otorgamiento de crédito.



[BACK TO AGENDA](#)

DATASET 1	DATASET 2	DATASET 3	VARIABLE
X	X	✓	Oficina: ubicación de solicitud de préstamos
X	✓	X	Categoría: Calificación de cada cliente por su historial de crédito.
X	✓	✓	Monto: Volumen del préstamo concedido. 5 categorías
✓	✓	✓	Garantía: Personal/real
X	✓	✓	Reestructurado: 0/1
✓	✓	✓	Edad
✓	X	X	Ocupación
✓	X	X	Nivel educativo
✓	✓	X	Ingreso total
X	X	X	Estrato social
✓	X	X	Antigüedad laboral
✓	X	X	Estado civil
✓	X	✓	Género
✓	X	X	Persona a cargo
✓	✓	X	Tipo de vivienda
✓	X	X	Tipo de contrato
✓	✓	X	Antigüedad en la institución
12	8	6	TOTAL

[BACK TO AGENDA](#)



Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

Descripción de los Datos

	DATASET 1
Nombre Dataset	<u>bank_df_1.csv</u>
Cantidad de filas	438557
Columnas	18
Bibliografia	12/18

La tabla comparativa anterior nos permitió establecer según las variables recomendadas por la bibliografía cuál de los tres, era el dataset más adecuado para realizar la clasificación y decisión de otorgamiento de crédito.

Además de la bibliografía mencionada previamente, contamos con la opinión de experto en banca que confirma que ese tipo de datos son los que se suelen obtener al momento en el cual una persona natural aplica para un crédito bancario.

[BACK TO AGENDA](#)

Descripción de las Variables

A continuación se encuentra un listado de las variables con las que cuenta nuestro DataSet:

NOMBRE DE LA VARIABLE	DESCRIPCIÓN BREVE	NOMBRE DE LA VARIABLE	DESCRIPCIÓN BREVE
ID	Número de cliente	AGE	Edad calculada en base a DAYS_BIRTH
CODE_GENDER	Género	DAYS_EMPLOYED	Cuenta regresiva desde hoy, Si es positivo, implica que la persona está desempleada
FLAG_OWN_CAR	¿Tiene auto?	YEARS_EMPLOYED	Años calculados en base a DAYS_EMPLOYED
FLAG_OWN_REALTY	¿Tiene propiedades?	FLAG_MOBIL	¿Tiene teléfono celular?
CNT_CHILDREN	Cantidad de hijos	FLAG_WORK_PHONE	¿Tiene teléfono corporativo?
AMT_INCOME_TOTAL	Ingresos anuales	FLAG_PHONE	¿Tiene teléfono?
NAME_INCOME_TYPE	Categoría de Ingresos	FLAG_EMAIL	¿Tiene e-mail?
NAME_EDUCATION_TYPE	Nivel de Educación	OCCUPATION_TYPE	Ocupación
NAME_FAMILY_STATUS	Estado civil	CNT_FAM_MEMBERS	Tamaño de la familia
NAME_HOUSING_TYPE	¿Dónde vive?	TOTAL_SCORE	Puntaje total obtenido
DAYS_BIRTH	Conteo regresivo del día actual hasta el día que nació (-1 significa ayer)	APPROVED	¿Crédito aprobado? 1:SI / 0: NO

Variables principales

Nuestro experto en banca nos sugirió tener en cuenta las siguientes variables que señalamos en naranja para realizar el **scoring** y nos brindó ejemplos de los puntajes que suelen brindarse en base a las mismas para saber si el crédito se otorga o no, lo cual observamos en la variable **APPROVED**

NOMBRE DE LA VARIABLE	DESCRIPCIÓN BREVE	NOMBRE DE LA VARIABLE	DESCRIPCIÓN BREVE
ID	Número de cliente	AGE	Edad calculada en base a DAYS_BIRTH
CODE_GENDER	Género	DAYS_EMPLOYED	Cuenta regresiva desde hoy, Si es positivo, implica que la persona está desempleada
FLAG_OWN_CAR	¿Tiene auto?	YEARS_EMPLOYED	Años calculados en base a DAYS_EMPLOYED
FLAG_OWN_REALTY	¿Tiene propiedades?	FLAG_MOBIL	¿Tiene teléfono celular?
CNT_CHILDREN	Cantidad de hijos	FLAG_WORK_PHONE	¿Tiene teléfono corporativo?
AMT_INCOME_TOTAL	Ingresos anuales	FLAG_PHONE	¿Tiene teléfono?
NAME_INCOME_TYPE	Categoría de Ingresos	FLAG_EMAIL	¿Tiene e-mail?
NAME_EDUCATION_TYPE	Nivel de Educación	OCCUPATION_TYPE	Ocupación
NAME_FAMILY_STATUS	Estado civil	CNT_FAM_MEMBERS	Tamaño de la familia
NAME_HOUSING_TYPE	¿Dónde vive?	TOTAL_SCORE	Puntaje total obtenido
DAYS_BIRTH	Conteo regresivo del día actual hasta el día que nació (-1 significa ayer)	APPROVED	¿Crédito aprobado? 1:SI / 0: NO

Scoring

NOMBRE DE LA VARIABLE	VALUE	SCORE
FLAG_OWN_CAR	Y	100
FLAG_OWN_CAR	N	0
FLAG_OWN_REALTY	Y	100
FLAG_OWN_REALTY	N	0
AMT_INCOME_TOTAL	0	0
AMT_INCOME_TOTAL	< 100.000	100
AMT_INCOME_TOTAL	100.000 & 250.000	150
AMT_INCOME_TOTAL	> 250.000	200
NAME_INCOME_TYPE	Working	150
NAME_INCOME_TYPE	Commercial associate	100
NAME_INCOME_TYPE	State servant	100
NAME_INCOME_TYPE	Pensioner	100
NAME_INCOME_TYPE	Student	0

NOMBRE DE LA VARIABLE	VALUE	SCORE
NAME_EDUCATION_TYPE	Secondary / secondary special	0
NAME_EDUCATION_TYPE	Higher education	50
NAME_EDUCATION_TYPE	Incomplete higher	0
NAME_EDUCATION_TYPE	Lower secondary	50
NAME_EDUCATION_TYPE	Academic degree	100
NAME_HOUSING_TYPE	House / apartment	100
NAME_HOUSING_TYPE	With parents	0
NAME_HOUSING_TYPE	Municipal apartment	50
NAME_HOUSING_TYPE	Rented apartment	50
NAME_HOUSING_TYPE	Office apartment	50
NAME_HOUSING_TYPE	Co-op apartment	50

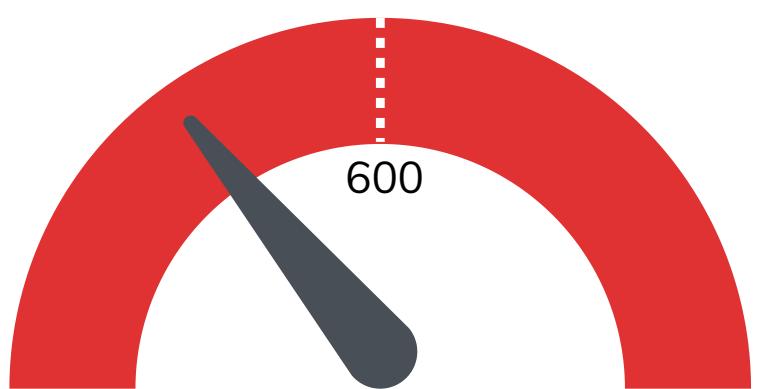
Scoring

NOMBRE DE LA VARIABLE	VALUE	SCORE
DAY_S_BIRTH	> 21 years	0
DAY_S_BIRTH	21 & 60	100
DAY_S_BIRTH	> 60	50
DAY_S_EMPLOYED	< 1 year	0
DAY_S_EMPLOYED	2 & 5 year	50
DAY_S_EMPLOYED	5 year	100
DAY_S_EMPLOYED	-1001	0

Nuestro banco tiene el criterio que si el puntaje obtenido es menor a 600 el crédito no se aprueba, en cambio si es mayor o igual a 600 si se aprueba

De esta manera cada uno de los clientes puede obtener un puntaje (score) basado en sus datos y así definir si el préstamo es aprobado o no lo cual se ve reflejado en la variable "APPROVED"

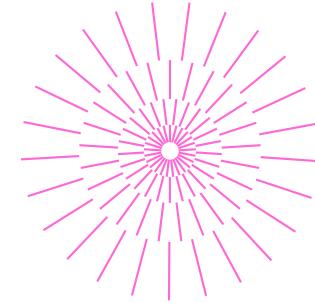
NO APROBADO



APROBADO



Nuestra VARIABLE TARGET es
"APPROVED"



Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

Comentarios

Nuestro dataset fue renombrado luego de los ajustes realizados y será este el utilizado para hacer el entrenamiento a continuación.

WE ARE
READY

	DATASET 1
Nombre Dataset	<u>bank.csv</u>
Cantidad de filas	438557
Columnas	22



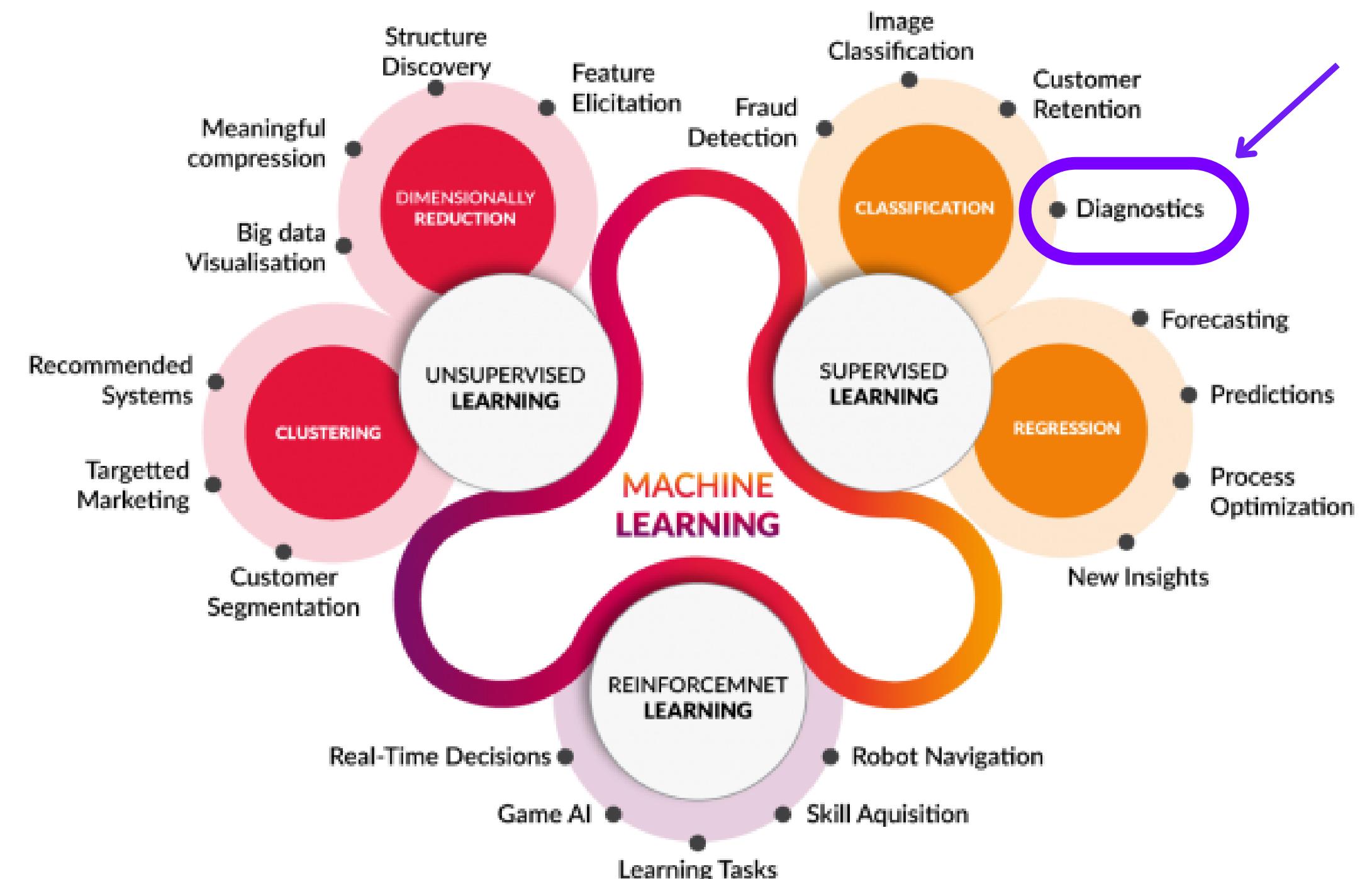
BACK TO AGENDA

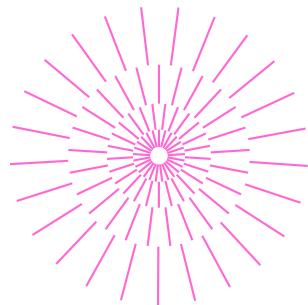
Algoritmo Elegido

Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

El algoritmo elegido es de Aprendizaje Supervisado cuyo objetivo es realizar Clasificación para el Diagnóstico para el otorgamiento de créditos bancarios.

[BACK TO AGENDA](#)

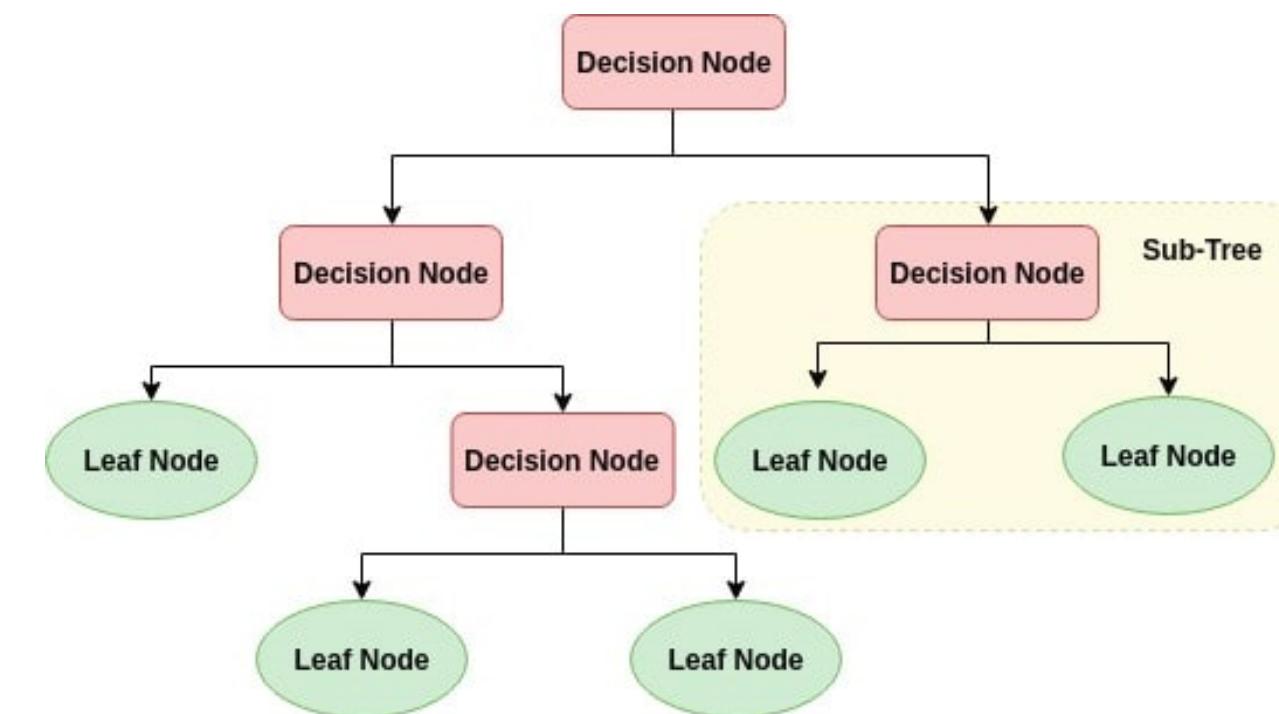




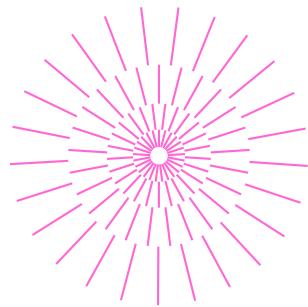
Arbol de Decisión

Siendo que nuestro objetivo es obtener una predicción de si se le otorgará o no un crédito a una persona, uno de los modelos que consideramos más convenientes por su simplicidad, es el árbol de decisión.

Los algoritmos de aprendizaje basados en árboles se consideran uno de los mejores y más utilizados métodos de aprendizaje supervisado. Los métodos basados en árboles potencian los modelos predictivos con alta precisión, estabilidad y facilidad de interpretación.

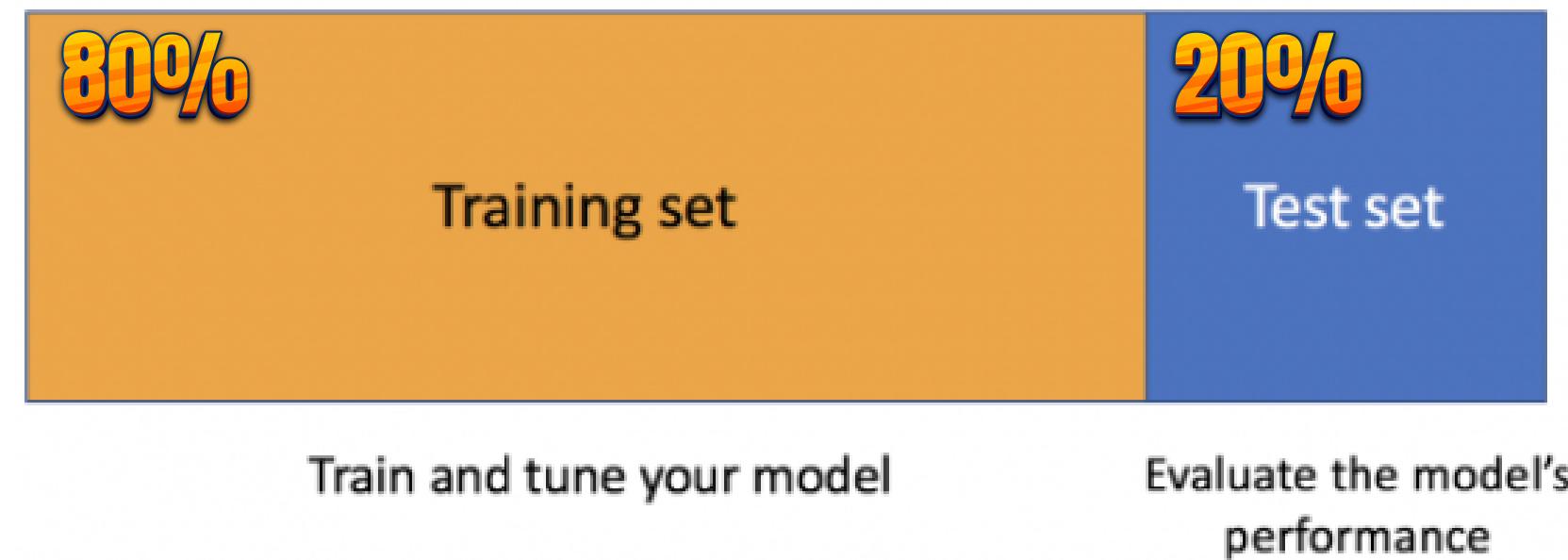


<https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/>

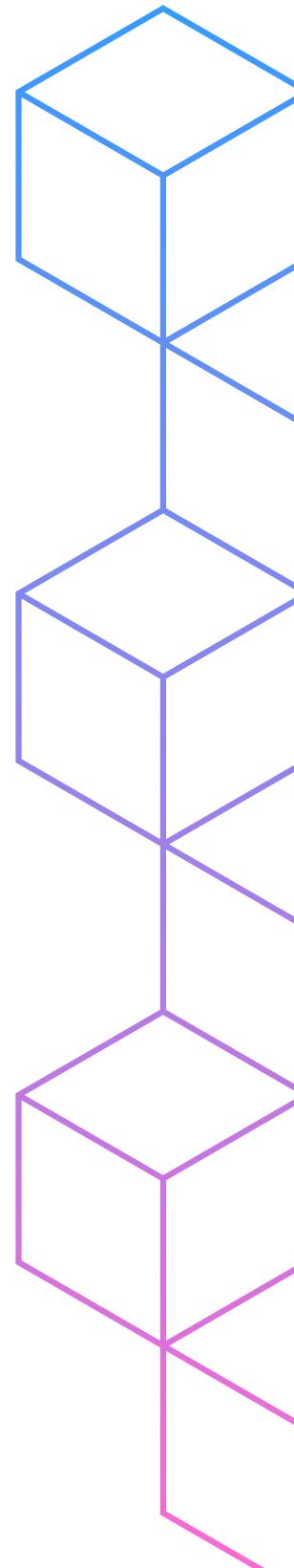


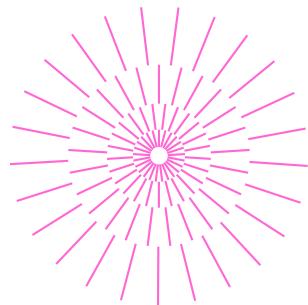
Arbol de Decisión

Validar un modelo de aprendizaje es esencial en la práctica y para poder evaluarlo correctamente, hay que realizar “split de datos”, es decir, separar nuestro dataset original en “Datos de Entrenamiento”, que serán usados justamente para entrenar a nuestro modelo y en “Datos de Test o de Testing” que serán aquellos datos que utilizaremos para evaluar la performance de nuestro modelo.



En nuestra práctica decidimos tomar el 80% de datos para el entrenamiento y el 20% de datos para el test





Data Science
Camada #25570
Dovale, Feü, Palacio, Parada

En el documento ejecutivo, se puede apreciar el desarrollo del modelo elegido, las problemáticas que surgieron y cómo se fueron resolviendo.

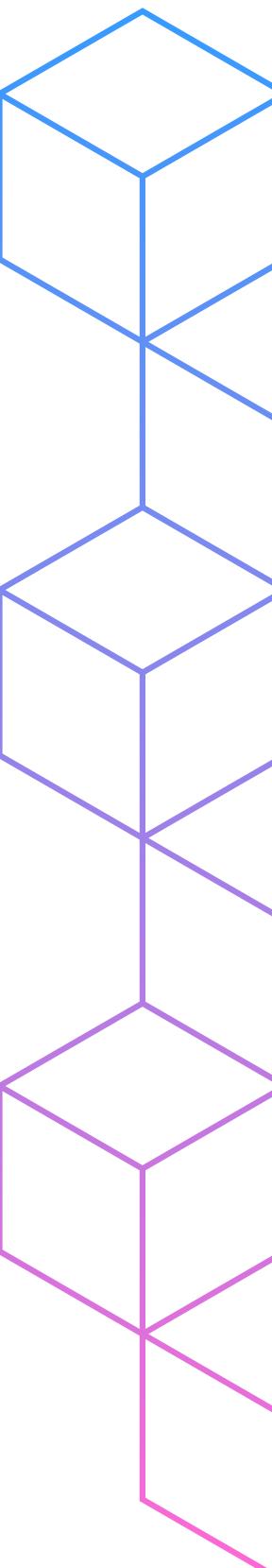
Arbol de Decisión - Desarrollo



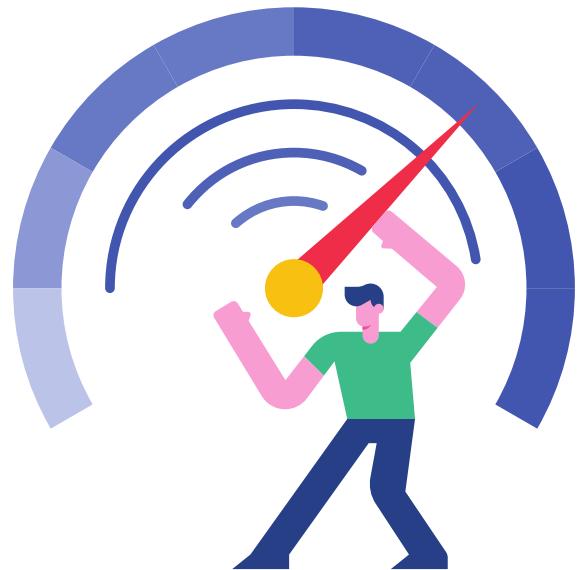
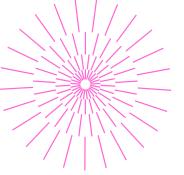
Modelo Machine Learning basado en Diagnóstico para el Otorgamiento de Crédito

Integrantes:

Dovale María
Feü Laura
Palacio Gabriel
Parada Laura

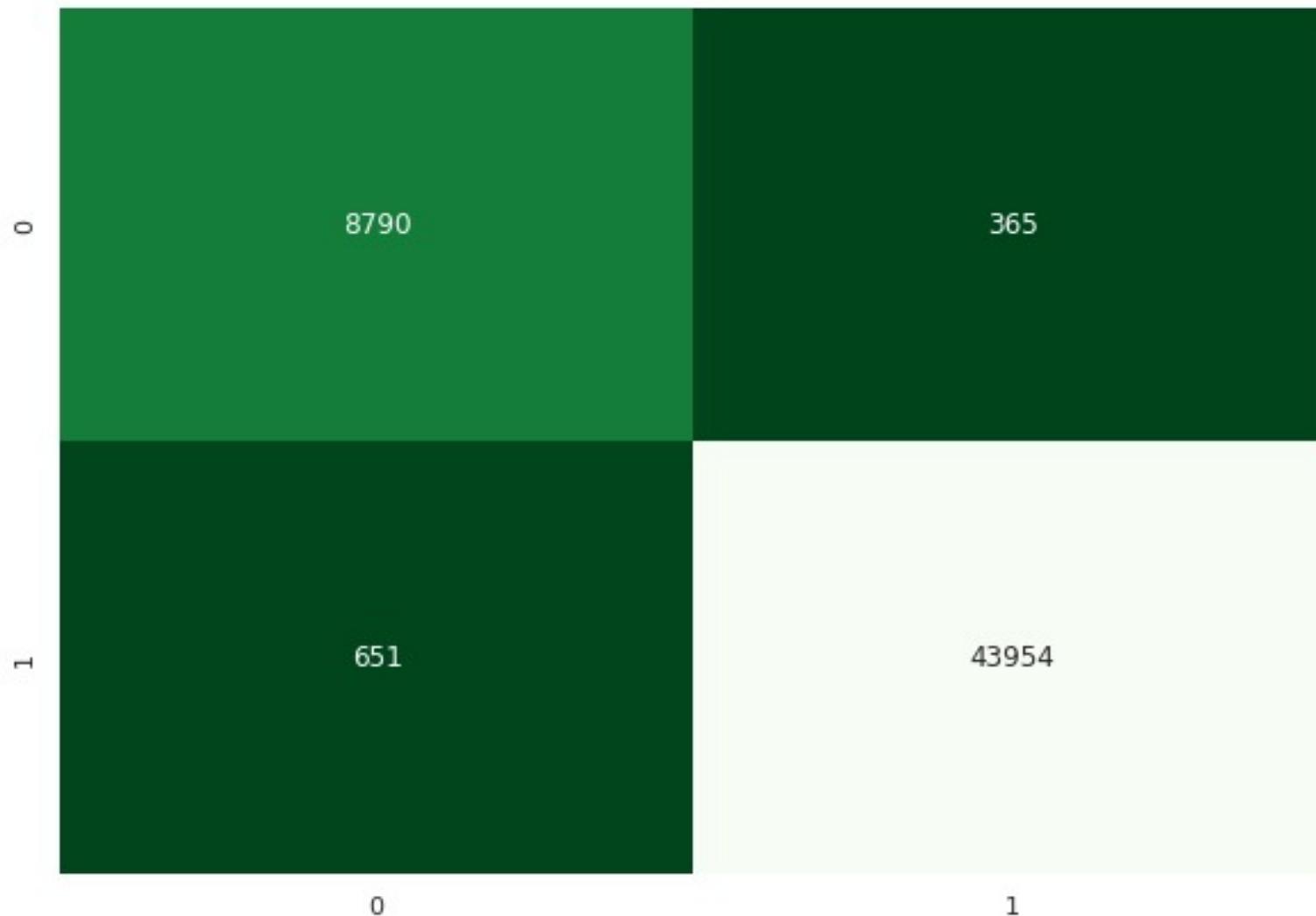


[BACK TO AGENDA](#)



Métricas de Desempeño

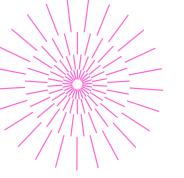
Vemos que ahora si el resultado que nos marca el accuracy de nuestro modelo es basado en el aprendizaje de todas las variables del dataset, en lugar de tomar sólo una como nos ocurrió la primera vez.



Podemos confirmar que nuestro modelo mejora su precisión al 98.11% cuando realizamos un entrenamiento de variables.

```
Accuracy score for test data is: 0.9811011904761905
          PREDICCIÓN NO APROBADO    PREDICCIÓN APROBADO
NO APROBADO                           8790                  365
APROBADO                            651                   43954
```

[BACK TO AGENDA](#)



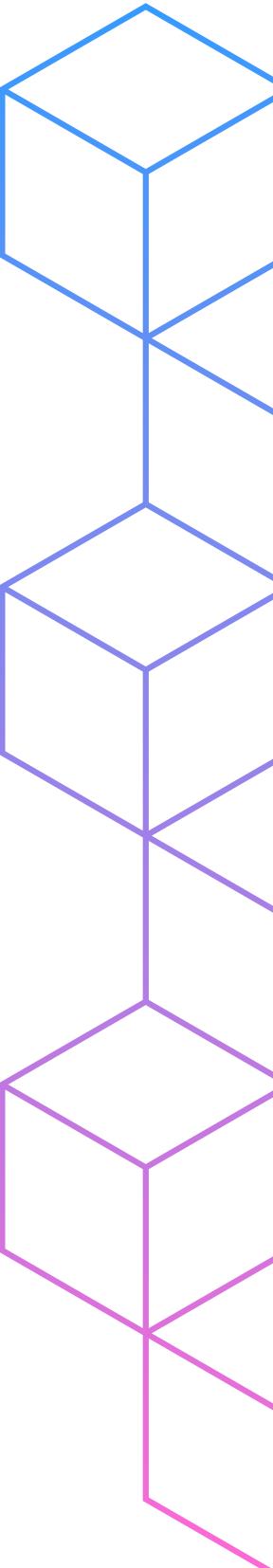
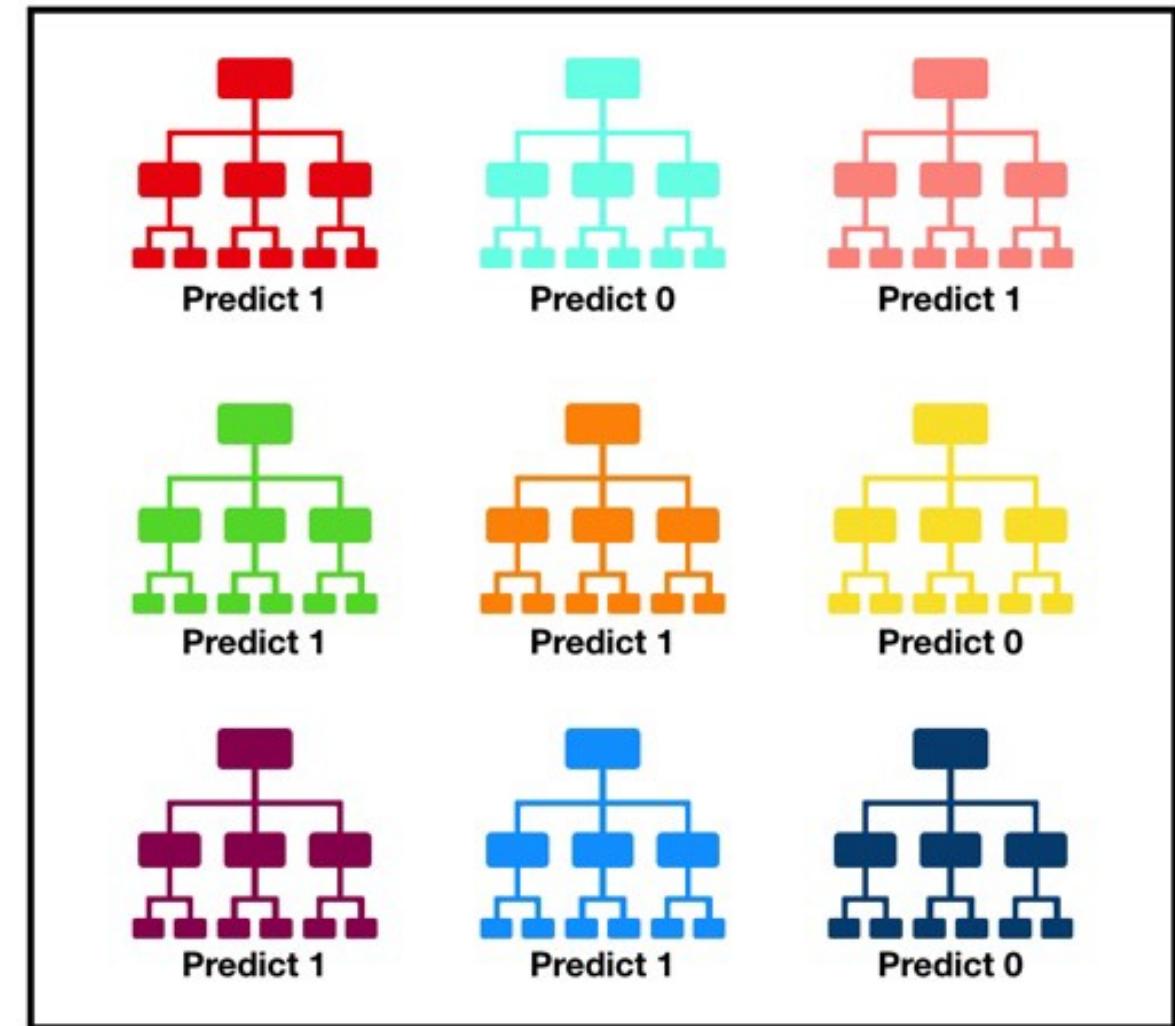
A continuación:

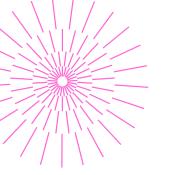
Random forest, como su nombre lo indica, consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto.

La razón de este maravilloso efecto es que los árboles se protegen entre sí de sus errores individuales.



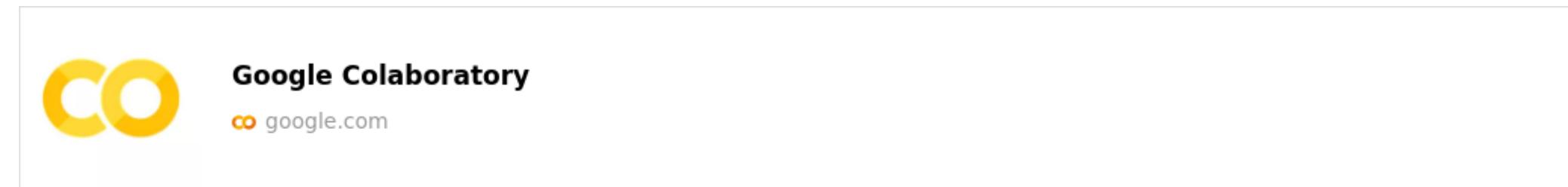
[BACK TO AGENDA](#)





Iteraciones de Optimización 3 - Random Forest

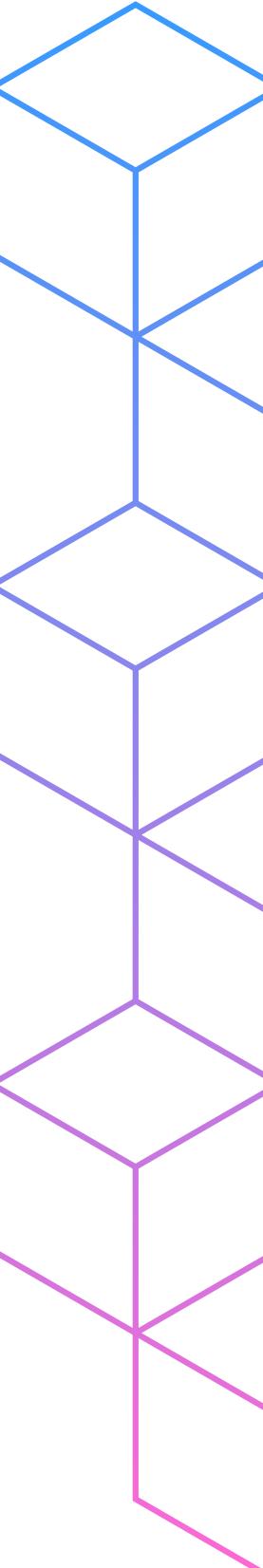
En el siguiente Google Colaboratory trabajamos el modelo de Random Forest:

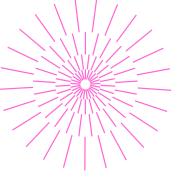


En el mismo hicimos varias pruebas con diferentes n_estimators

- 1000
- 500
- 50
- 5

[BACK TO AGENDA](#)



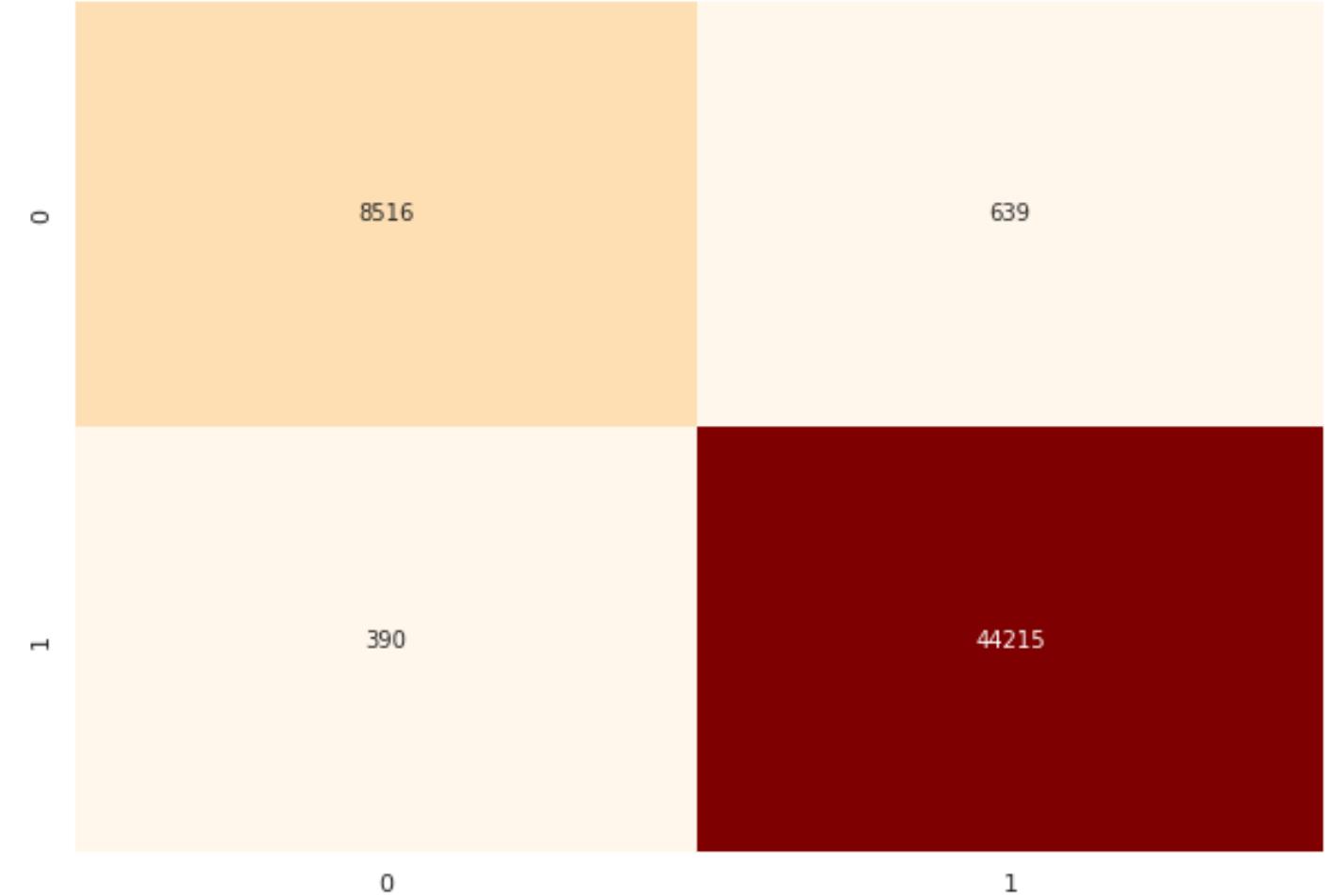


Métricas de Desempeño - Random Forest

Esta matriz de confusión nos resulta de 1 entrenamiento del modelo de Random Forest teniendo en cuenta un `n_stimator = 1000`.

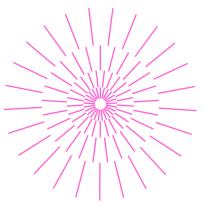
Podemos notar que tenemos un Accuracy de 98,08%.

La cantidad de falsos positivos es de 639 y los falsos negativos es de 390



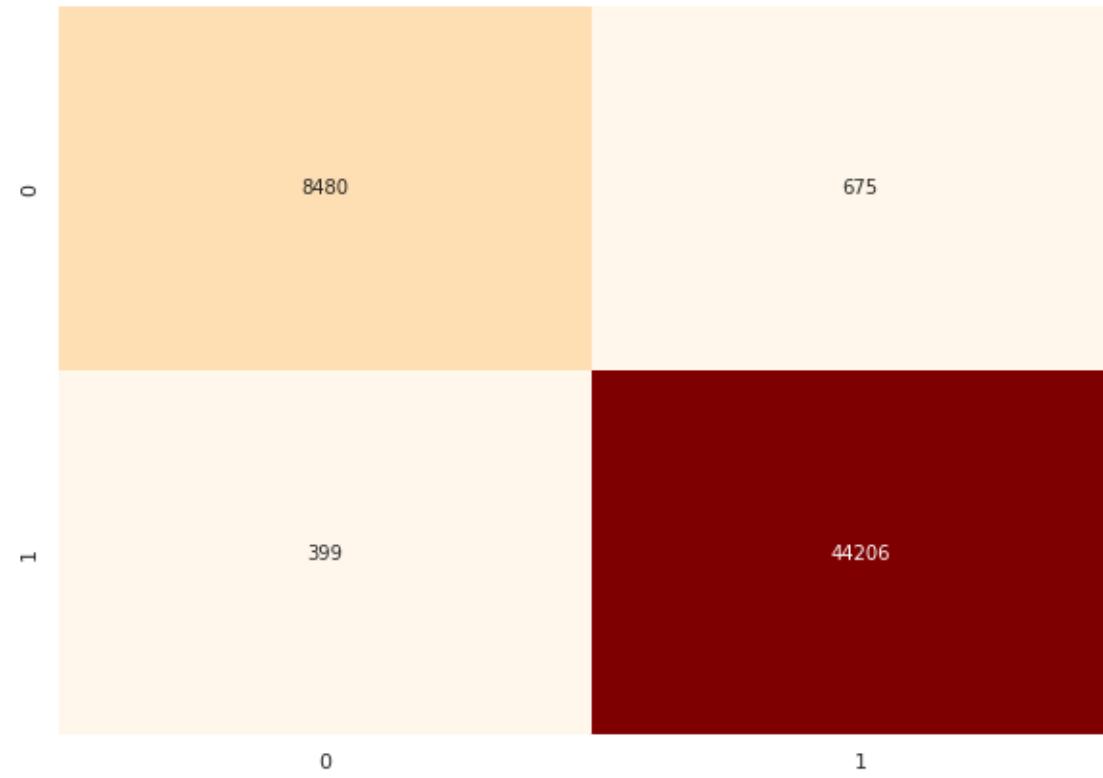
		PREDICCION NO APROBADO	PREDICCION APROBADO
ACTUAL	NO APROBADO	8516	639
	APROBADO	390	44215
Accuracy score for test data is: 0.980859375			

[BACK TO AGENDA](#)

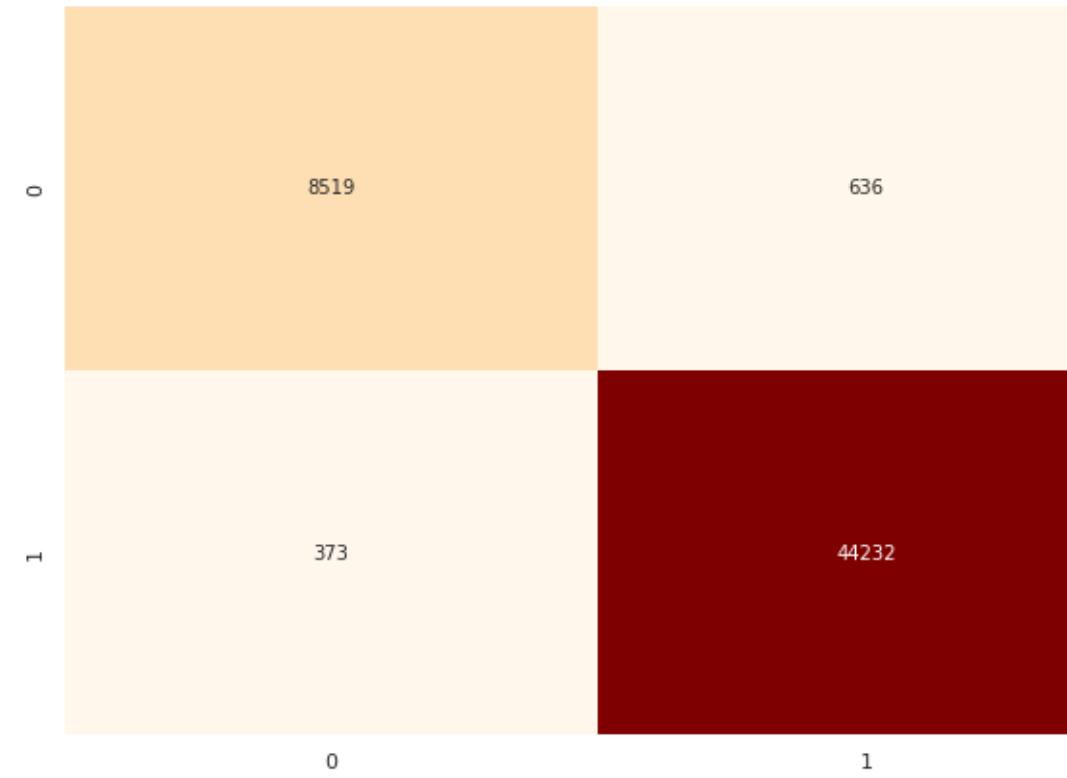


Otras Métricas de Desempeño - Random Forest

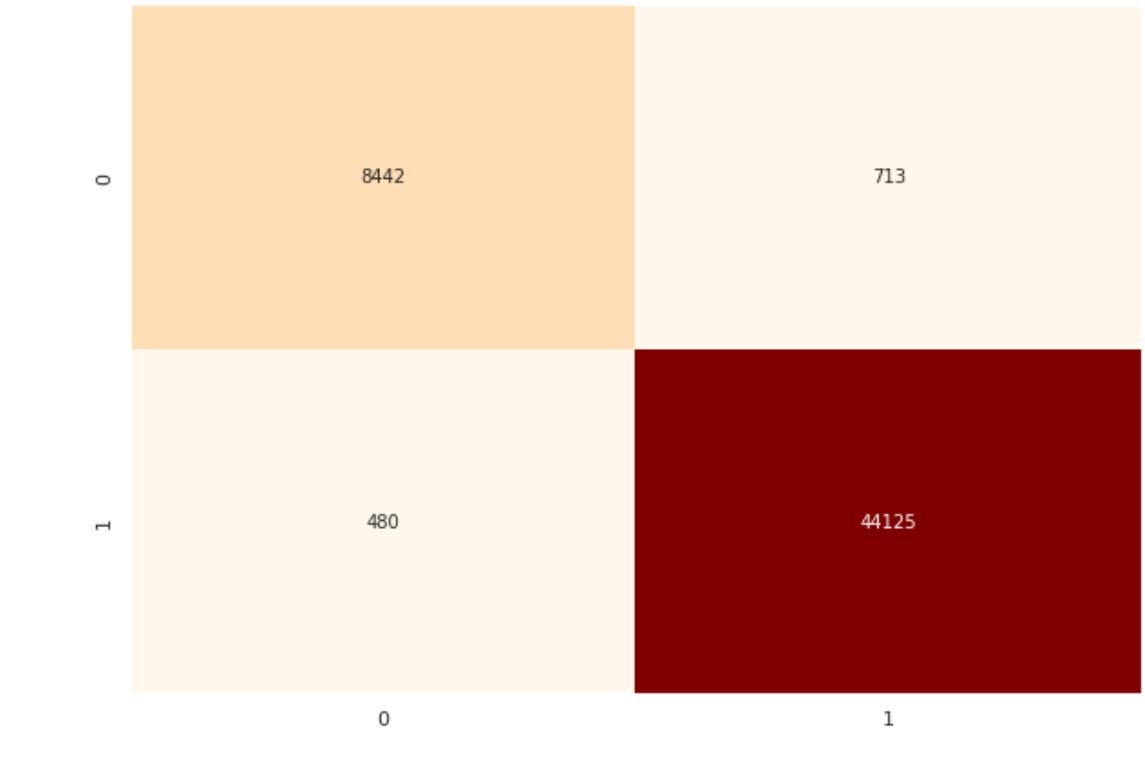
Analizamos otros resultados del Random Forest con el fin de comprobar las diferencias existentes en cada corrida y entender como afecta a las métricas. Vemos aquí matrices de confusión teniendo en cuenta:



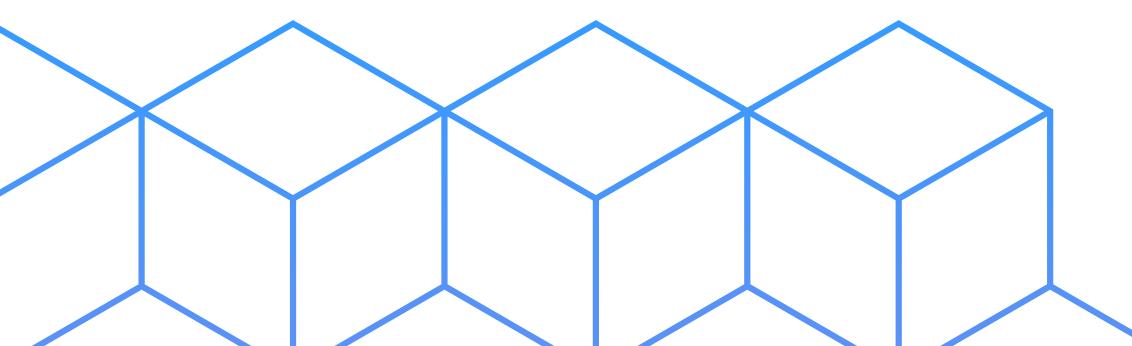
n_estimators = 500

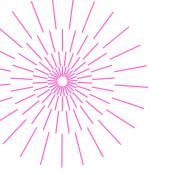


n_estimators = 50



n_estimators = 5





Conclusiones

Notamos que entre las 4 pruebas que se realizaron no obtuvimos mucha diferencia entre el margen 50 a 1000 pero notamos que el que se realizo con el número más alto dio el mejor resultado, por lo que decidimos quedarnos que el de 1000

Comparando los 2 algoritmos

Recordemos los resultados de los algoritmos que entrenamos.

En el caso del "Árbol de decisión" fueron:

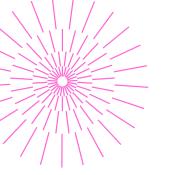
- Accuracy de 98,11%
- Falsos Positivos: 365
- Falsos Negativos: 651

Y en el caso de del "Random Forest" :

- Accuracy de 98,08%
- Falsos Positivos: 639
- Falsos Negativos: 390

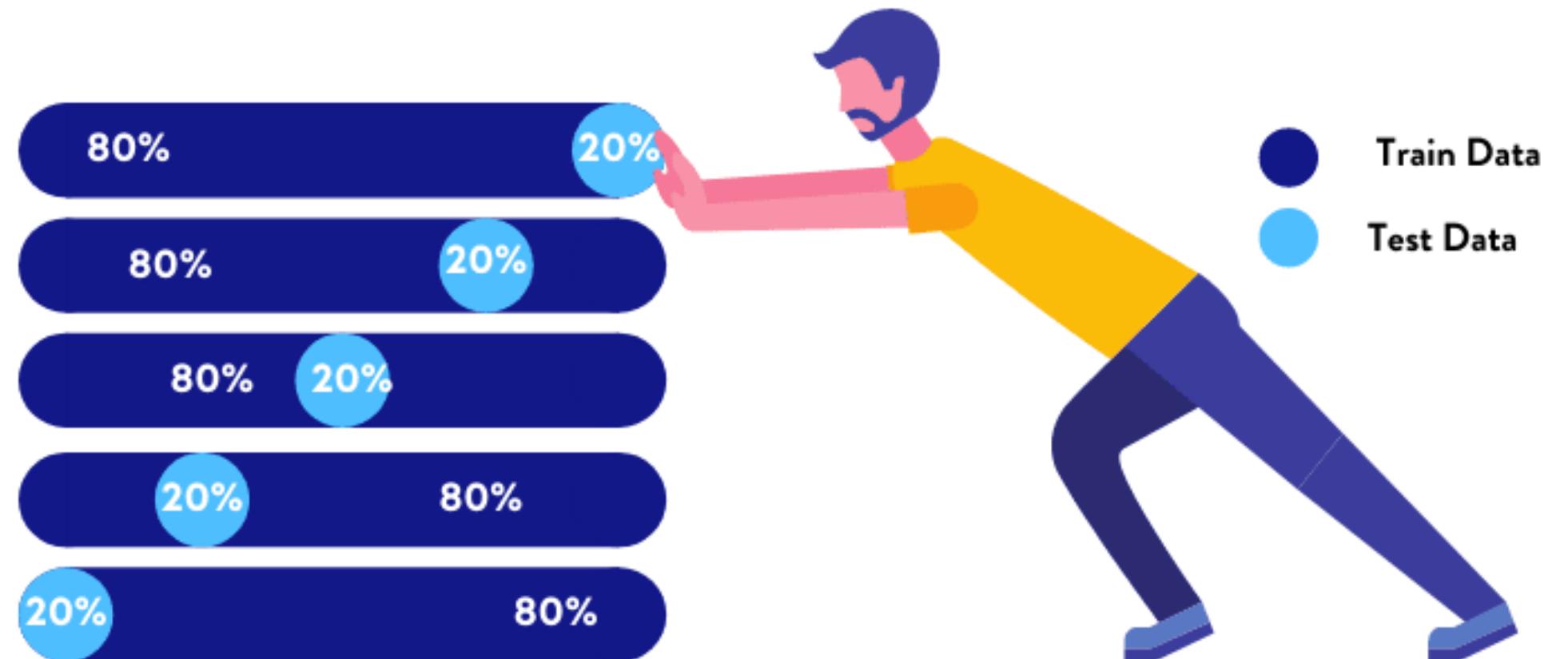
Teniendo esto en cuenta, la conclusión a la que llegamos es que, a priori el algoritmo que mejor se aadecua al plan de negocio es el Árbol de Decisión, dado que los Falsos Positivos son menos que los del Random Forest. Ya que es mejor que el banco reciba reclamos por no poder sacar un crédito, que otorgar un prestamo a alguien que no pueda pagarlos a futuro.



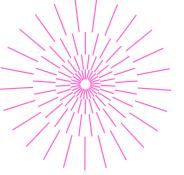


Cross Validation

Una de las técnicas más empleadas para probar la eficacia de un modelo de Machine Learning es la “cross-validation” o validación cruzada. Este método también es un procedimiento de “re-sampling” (remuestreo) que permite evaluar un modelo incluso con datos limitados.

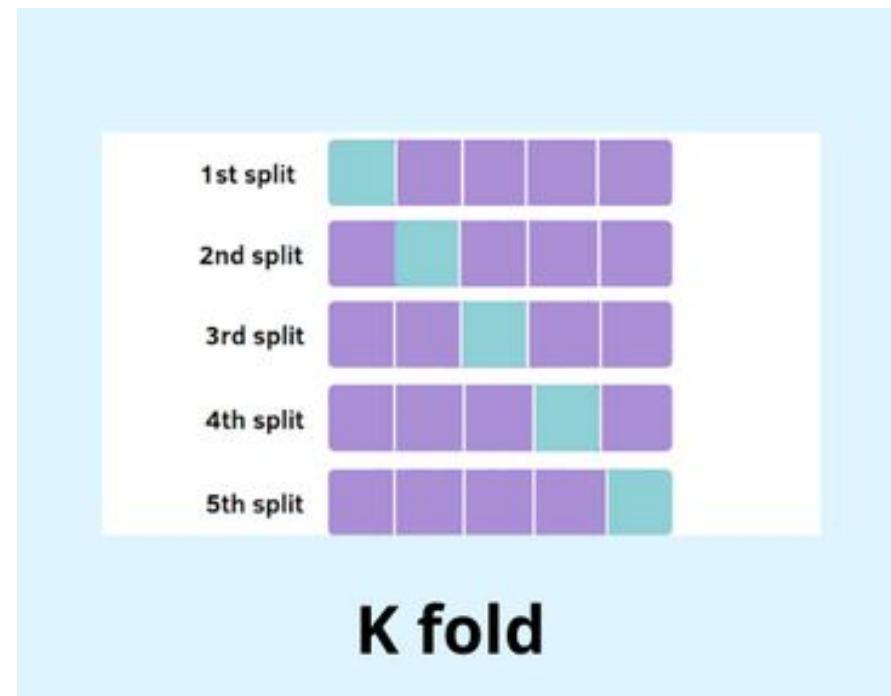


[BACK TO AGENDA](#)



K-FOLD

Permite garantizar que todas las observaciones de la serie de datos original tengan la oportunidad de aparecer en la serie de entrenamiento y en la serie de prueba. En caso de datos de entrada limitados, resulta uno de los mejores enfoques.

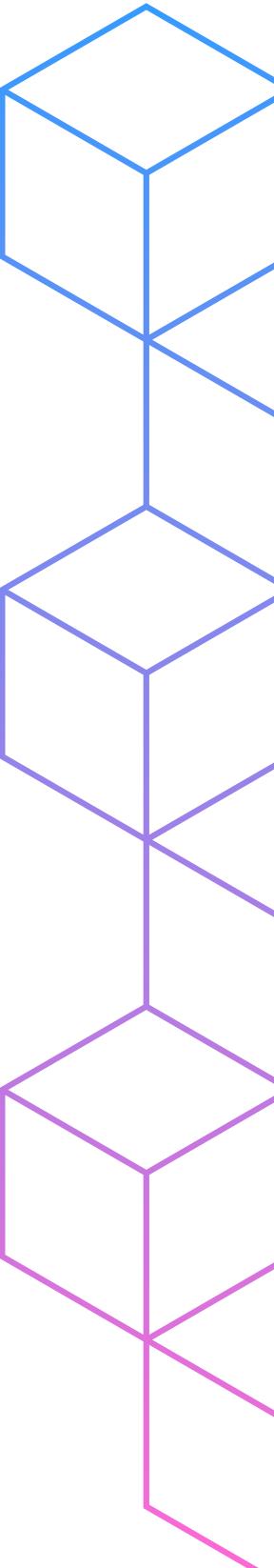


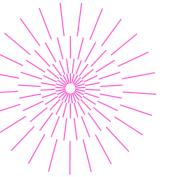
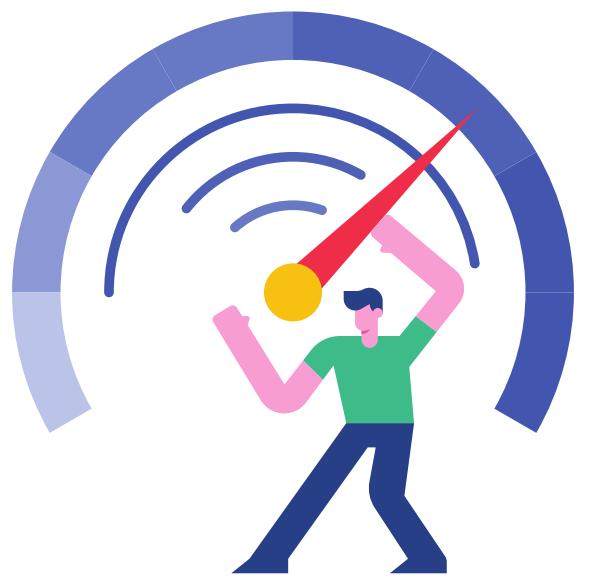
```
print("Cross validation score es %.5f ± %0.2f" %  
(CV_scores.mean(), CV_scores.std()))
```

Cross validation score es 0.98033 ± 0.00

Con este resultado confirmamos que el modelo Árbol de decisión en comparación con el modelo Random Forest tiene un mejor performance y que se ajusta mejor a los resultados esperados para este modelo de negocio, donde la agilidad y la seguridad en que un cliente cumple con los requisitos indicados para obtener un crédito..

[BACK TO AGENDA](#)

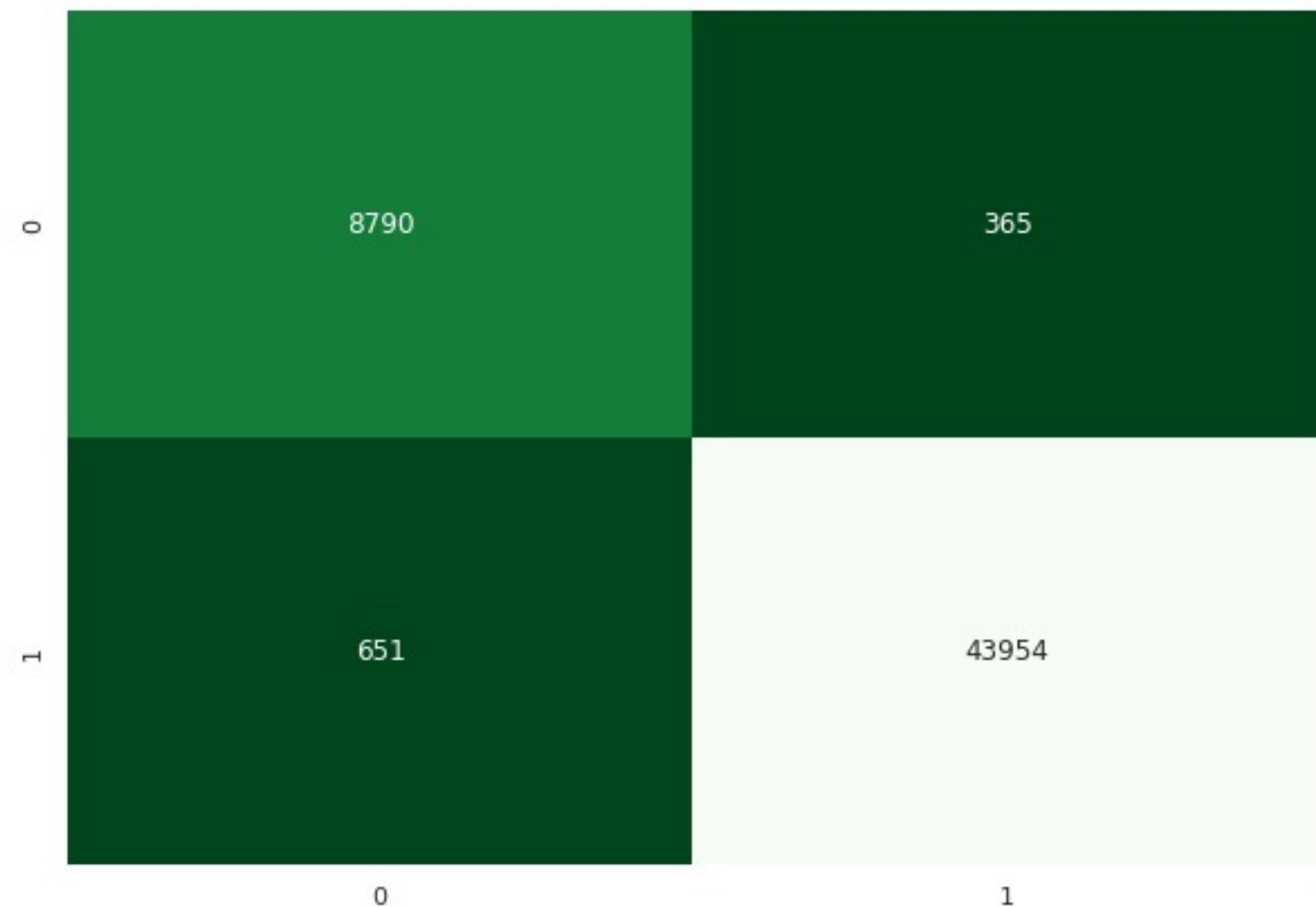




Métricas finales del Modelo Optimizado

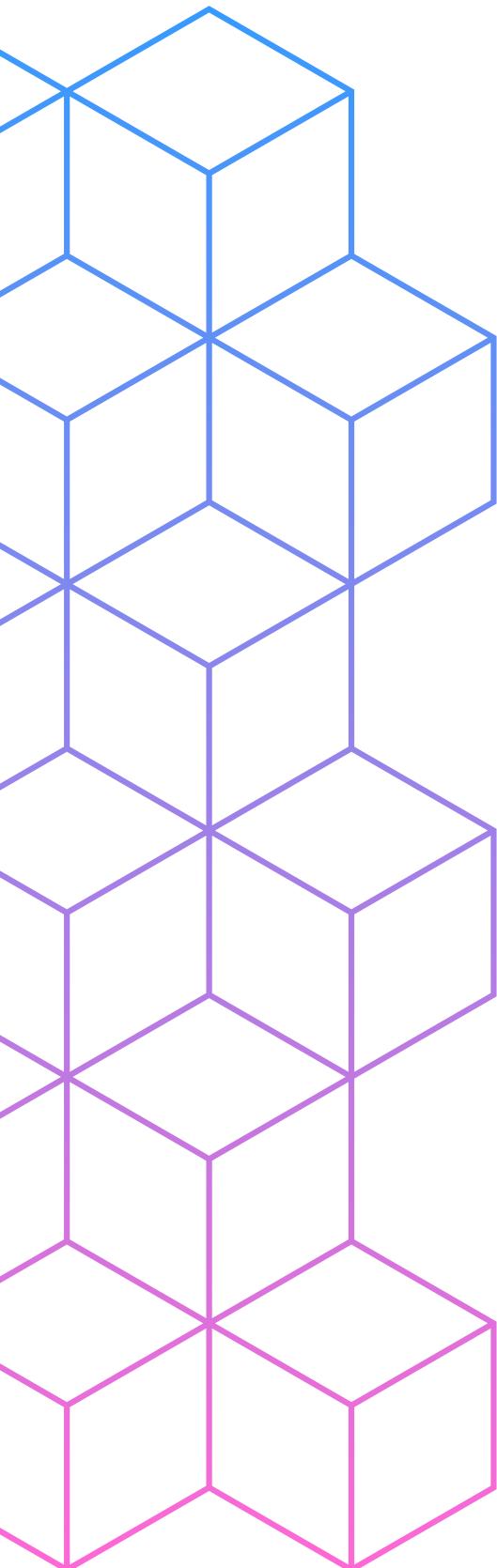
Podemos confirmar que nuestro modelo mejora su precisión al 98.11% cuando realizamos un entrenamiento de variables.

		PREDICCIÓN NO APROBADO	PREDICCIÓN APROBADO
NO APROBADO	8790	365	
	651	43954	
		0.9811011904761905	



[BACK TO AGENDA](#)

Conclusiones



Si bien obtuvimos un modelo funcional y bastante básico, las posibilidades de aplicación y mejora son muchas para este modelo dentro del mercado (banca), entre esas posibilidades de alcance está la de poder asignar también el monto de crédito a dar al usuario y no limitarnos a una respuesta binaria, otro posible alcance del modelo sería el de segmentar a los usuarios que obtuvieron crédito y de acuerdo a sus características poder aplicar modelos cross-sell/up-sell y los que no obtuvieron créditos derivarlos a otros productos que se ajustan a sus características/scoring y re-evaluarlos después de un tiempo determinado.

Como experiencia de esta cursada de Data Science entendimos la importancia de esta disciplina la cual es clave en el proceso de toma de decisiones hoy día donde los grandes volúmenes de información son materia bruta que podemos refinar y aprovechar en muchas industrias para brindar mayores y mejores soluciones en ventas, atención al cliente, optimización y personalización de resultados, etc.

