Estimación del Éxito de Videos de YouTube Mediante Enfoques de Machine Learning: Caso DelcaVideography

Mariana Agudelo Zuluaga
Andrés Mauricio Cano Campiño
Esteban Castro Castaño
Juan David Gallego Montoya
Vanessa Osorio Urrea

DelcaVideography

Agenda

01. Entendimiento del Problema – Pregunta de Negocio
02. Marco Conceptual
03. Entendimiento y Preparación de datos
04. Modelado de Variables
05. Selección de Variables
06. Regresión Logística
07. Conclusiones

¿Qué es DelcaVideography?

2014

Nace el canal **DelcaVideography**

YouTube

Medellín, Colombia

2015

2016

Se convierte en un **canal educativo** de diseño gráfico, audiovisual y fotográfico

Enfocado en CorelDraw

2017

2018

Se **reduce la longitud** de los videos **a 15-20 minutos**

2019

Antes se publicaban videos de 1 hora

2020

Se publican **2 videos semanales**

2021

2022

Logra **112 mil suscriptores** con más de **540 videos publicados** y más de **9 M de visualizaciones** desde el 2014.
Ranking en Colombia: 723

Contenido **visto por el 0.1% - 0.3% de sus suscriptores**

Entendimiento del Problema

Pregunta de negocio

¿Cómo saber si un video que se publicará será exitoso o no?

En promedio, un video publicado tiene **8 horas** de trabajo **en actividades de producción**

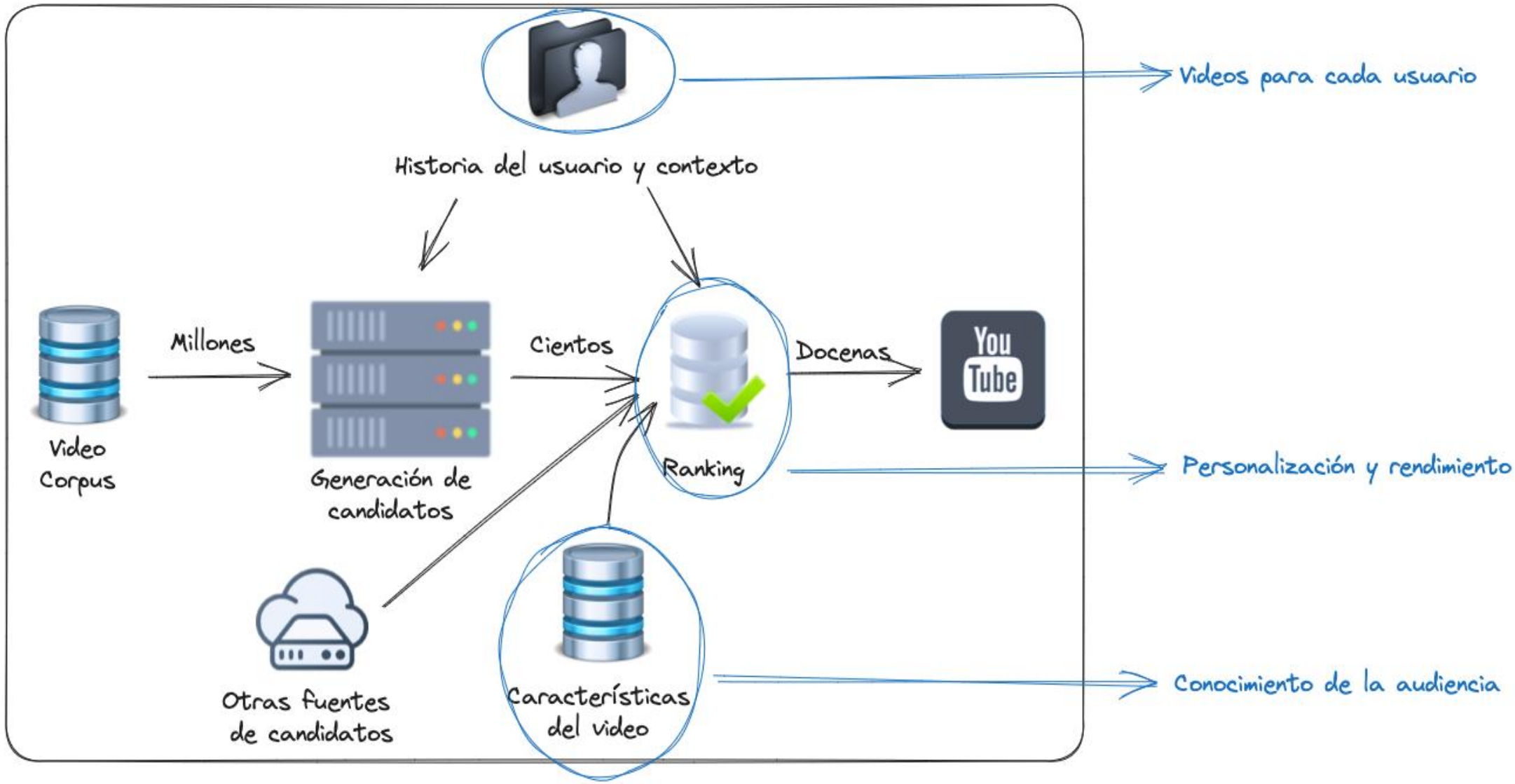
Cada **contenido** publicado es **visto por el 0.1% - 0.3% de sus suscriptores**

Ingresos del canal **dependen** en gran medida **del éxito de los videos**

Los expertos consideran que un **porcentaje de clics en las impresiones mayor al 3% hacen exitoso a un video**

No se conocen con exactitud **las variables** que afectan esta marca de éxito

Marco Conceptual



¿Qué variables son relevantes?

Literatura Científica

- Retención promedio
- Porcentaje promedio visto (APV)
- Duración promedio de visualización (AVD)
- Duración video
- Retención de audiencia relativa (RAR)
- Porcentaje Leal

CRECETUBE (SEO)

- CTR
- Retención de Audiencia
- Posicionamiento Orgánico
- Tráfico Recomendado
- Duración Media de Visualización
- Número de Suscriptores

Entendimiento y Preparación de Datos

Tu canal

DelcaVideography

Web Scraping + Studio

61 Variables (7 días)
108 Videos

Estadísticas

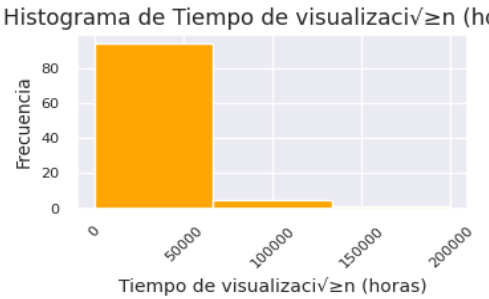
Información Pública y Privada
(16Feb21 a 30Ago22)

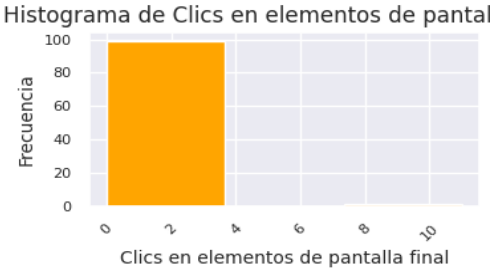
15 Variables
100 Videos

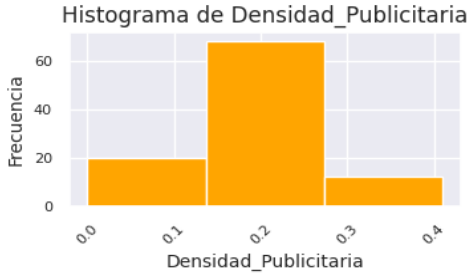
- Eliminación Variables Duplicadas
- Eliminación Variables tipo objeto
- Manejo Nulos: Variables y videos

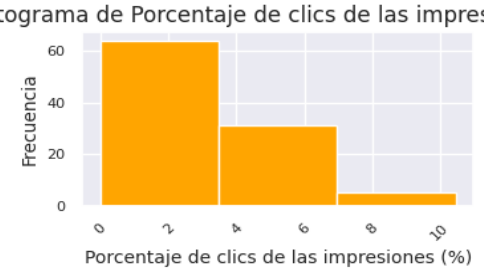
Análisis Exploratorio - Histogramas

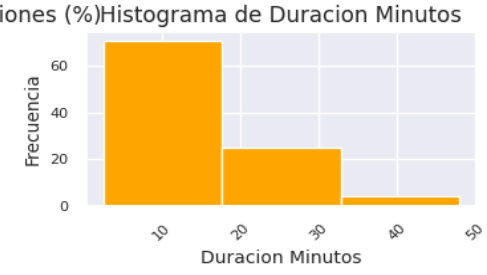
DelcaVideography

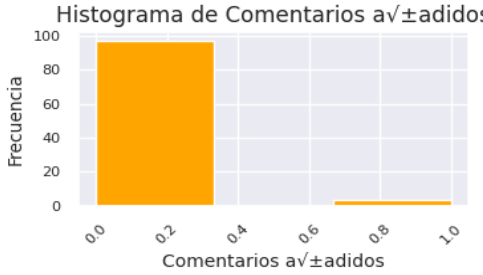


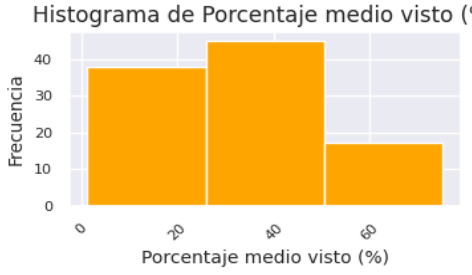


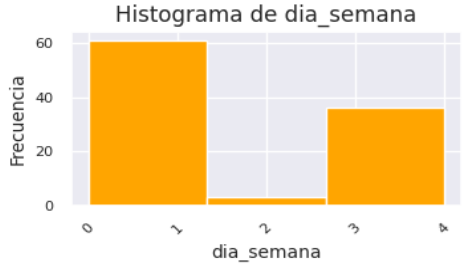


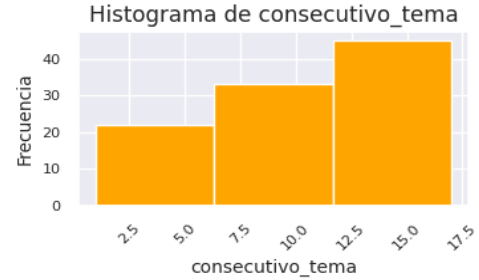




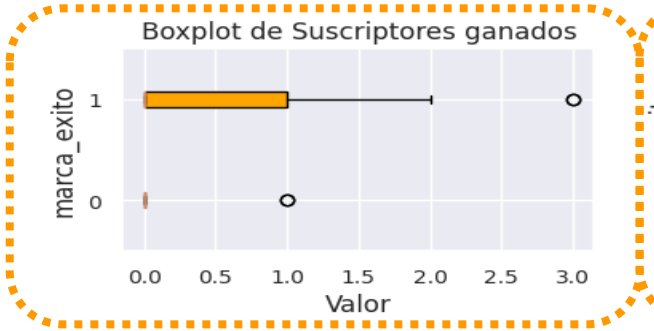


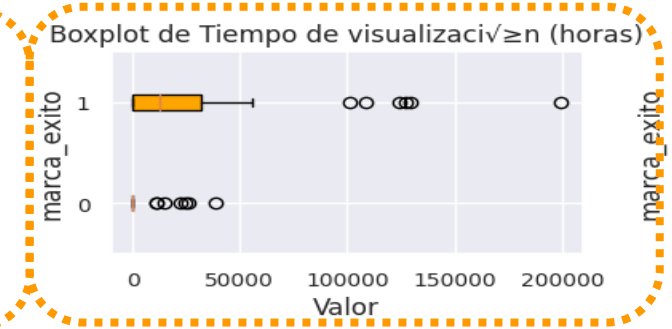


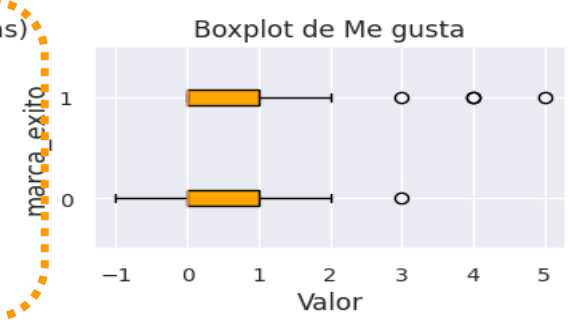


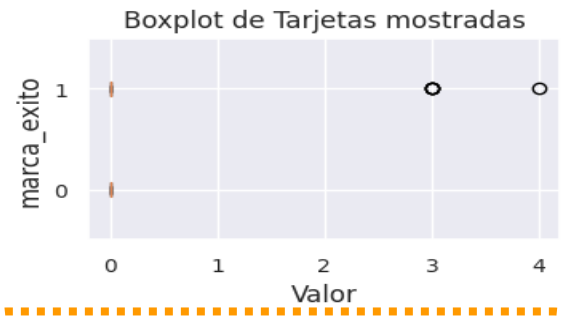


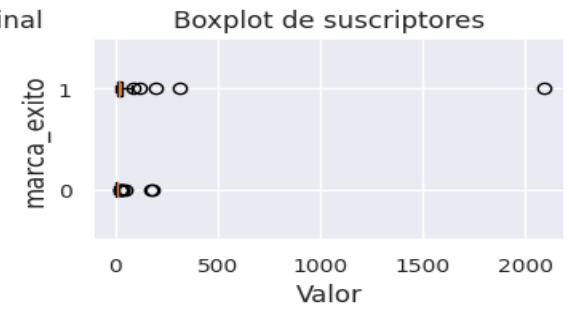
Análisis Exploratorio - Boxplot

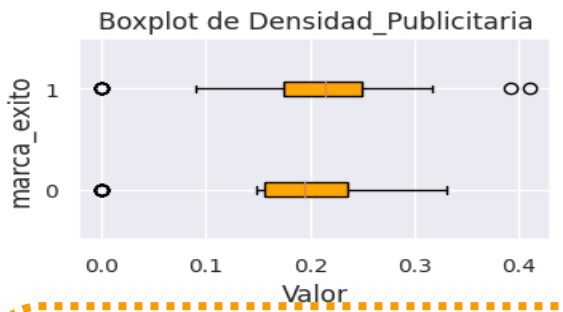


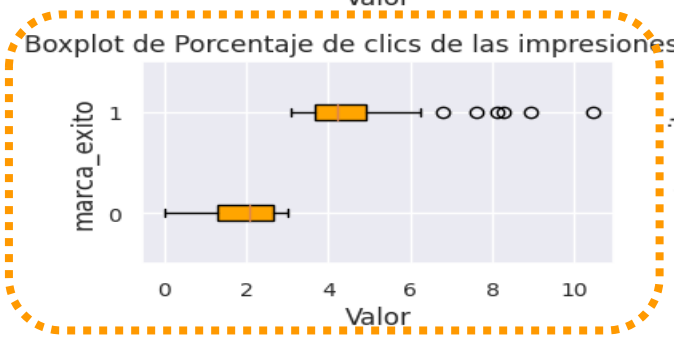


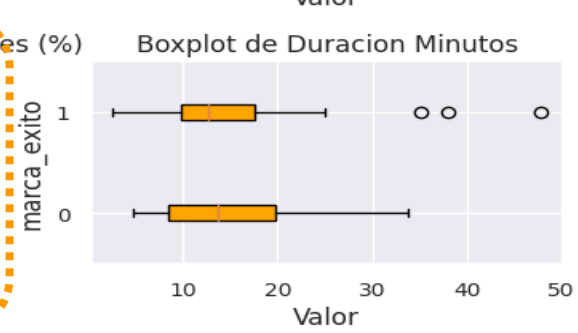


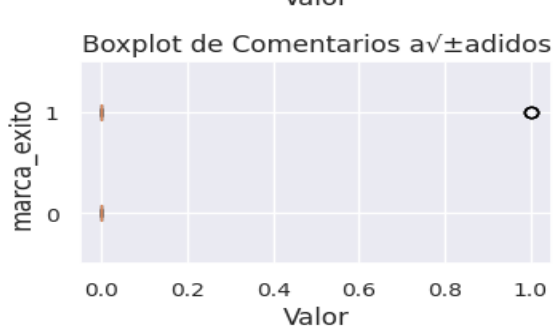


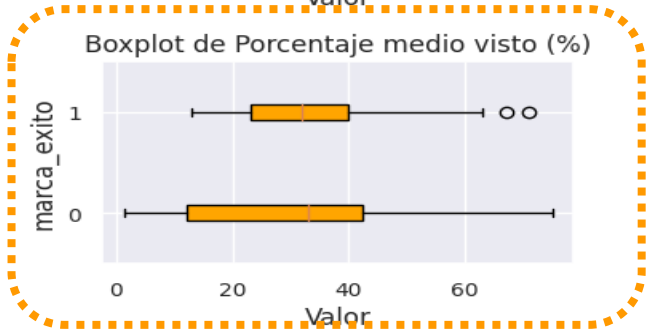


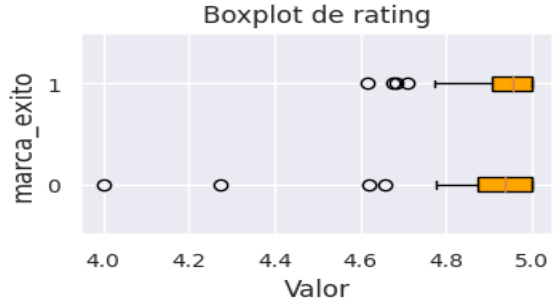


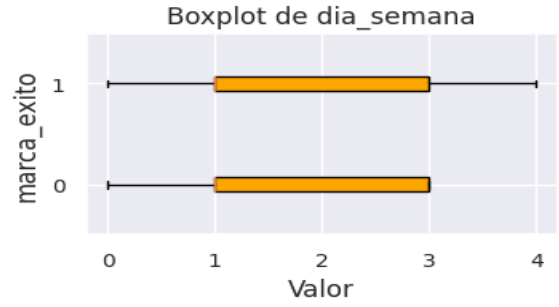


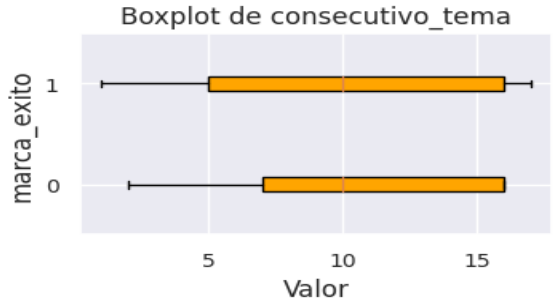


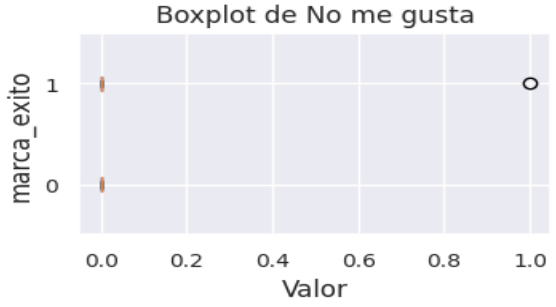




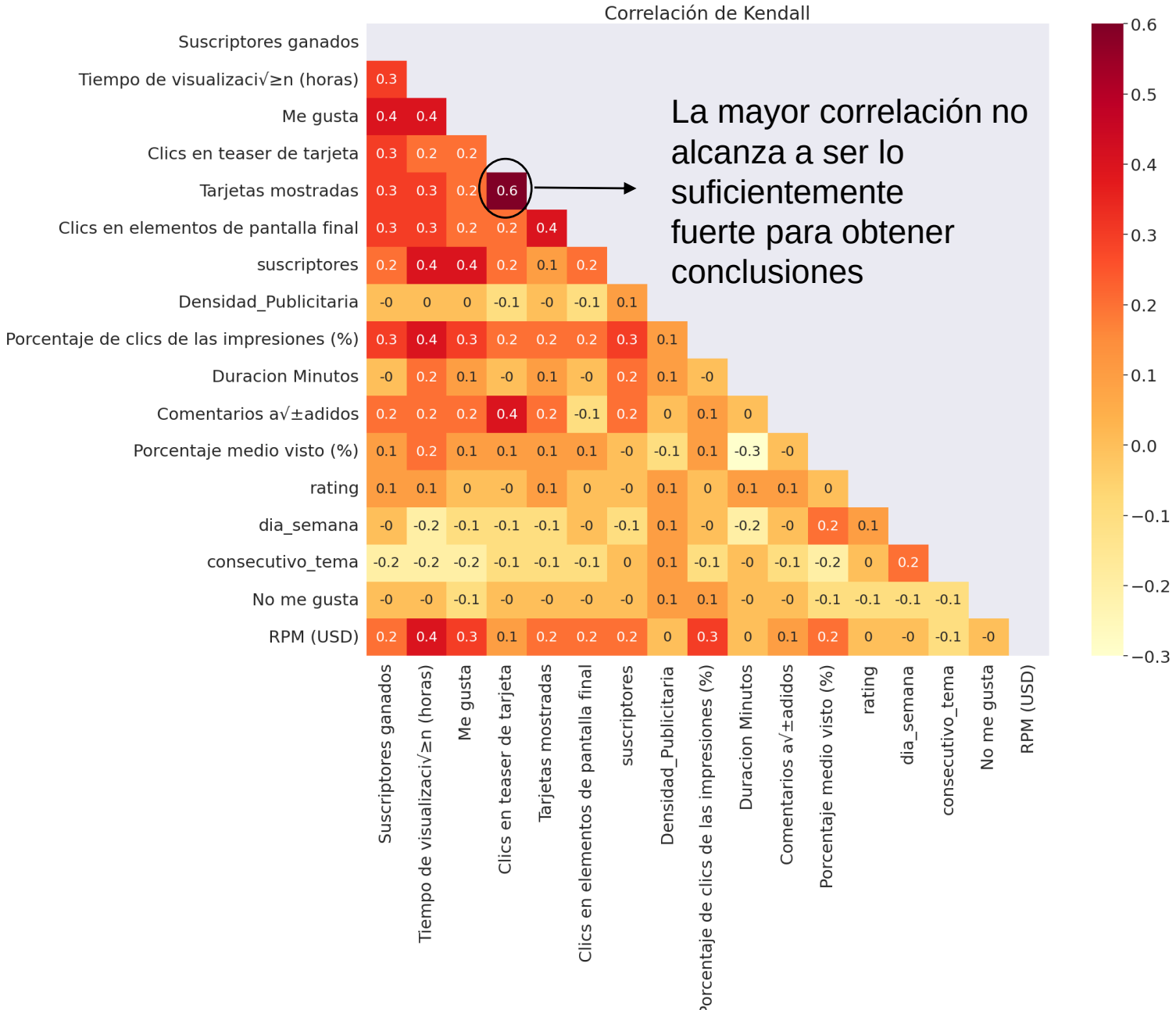








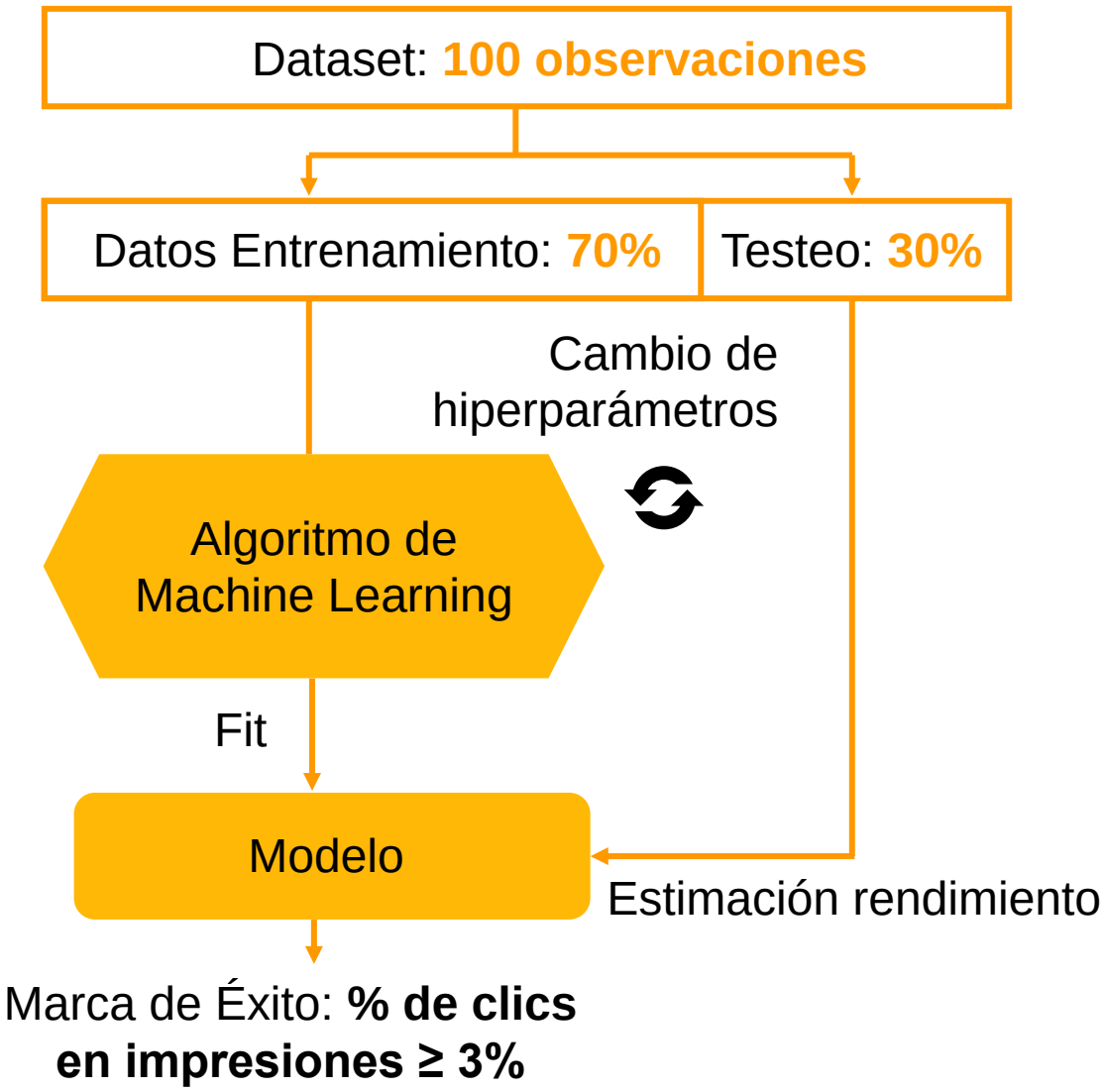
Análisis Exploratorio



Correlaciones:

1. Kendall y Spearman
2. Son moderadas
3. No son suficientes para realizar predicciones precisas.
4. No permiten explicar la variabilidad de los datos
5. Se requieren análisis adicionales

Modelado de Variables



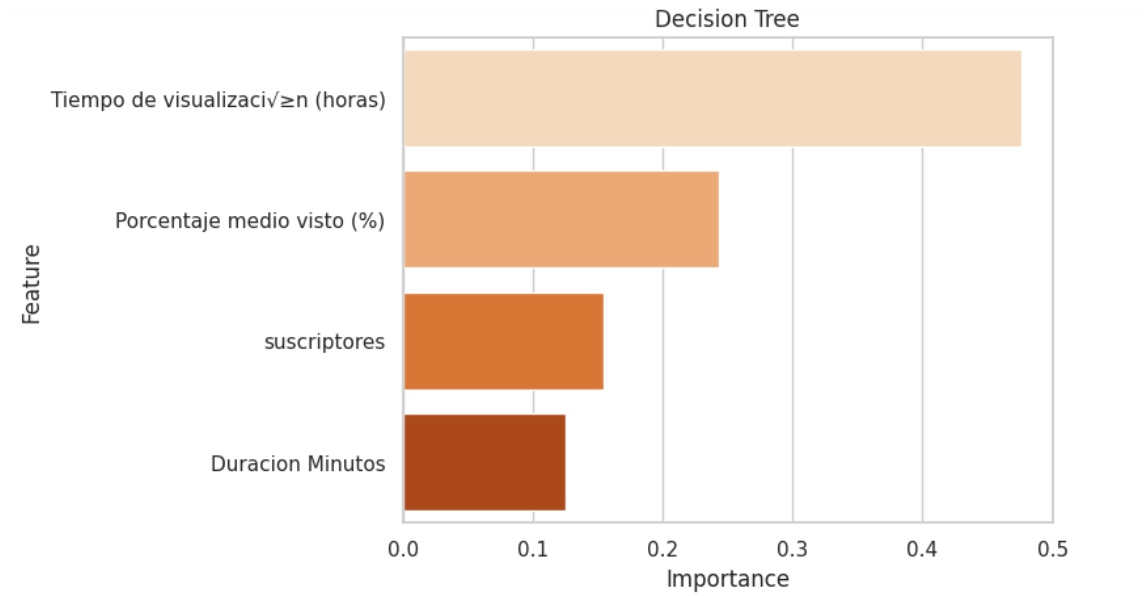
Modelado de Variables – Métricas de Evaluación

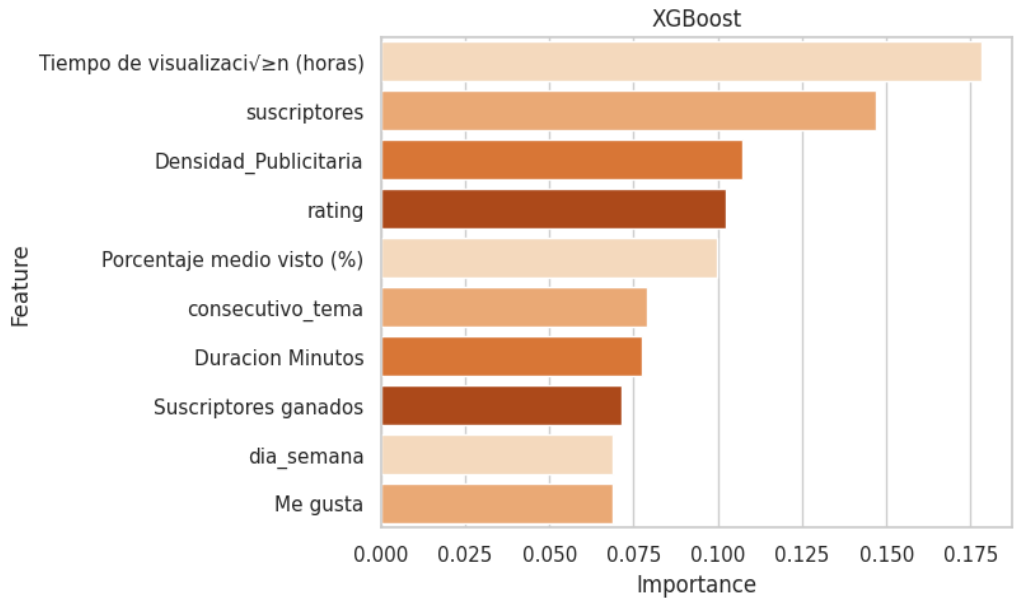
DelcaVideography

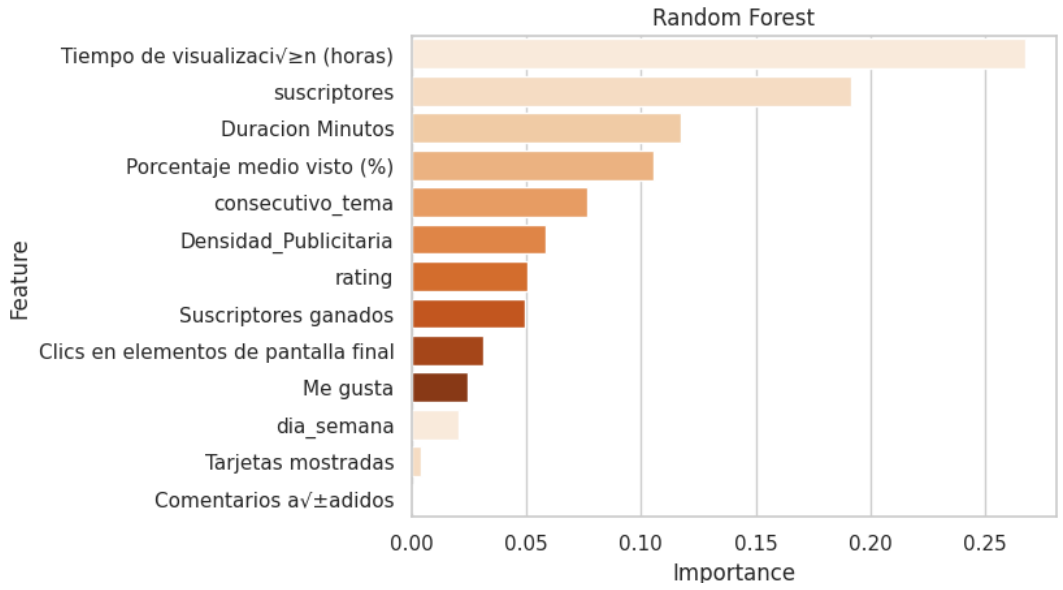
		Árbol de Decisión	XGBoost	Random Forest	Reg. Logística
Precision	No Éxito	0.85	0.79	0.80	0.83
	Éxito	0.60	0.67	0.80	0.83
Recall	No Éxito	0.81	0.90	0.95	0.95
	Éxito	0.67	0.44	0.44	0.56
F1-Score	No Éxito	0.83	0.84	0.87	0.89
	Éxito	0.63	0.53	0.57	0.67
Accuracy		0.77	0.77	0.80	0.83

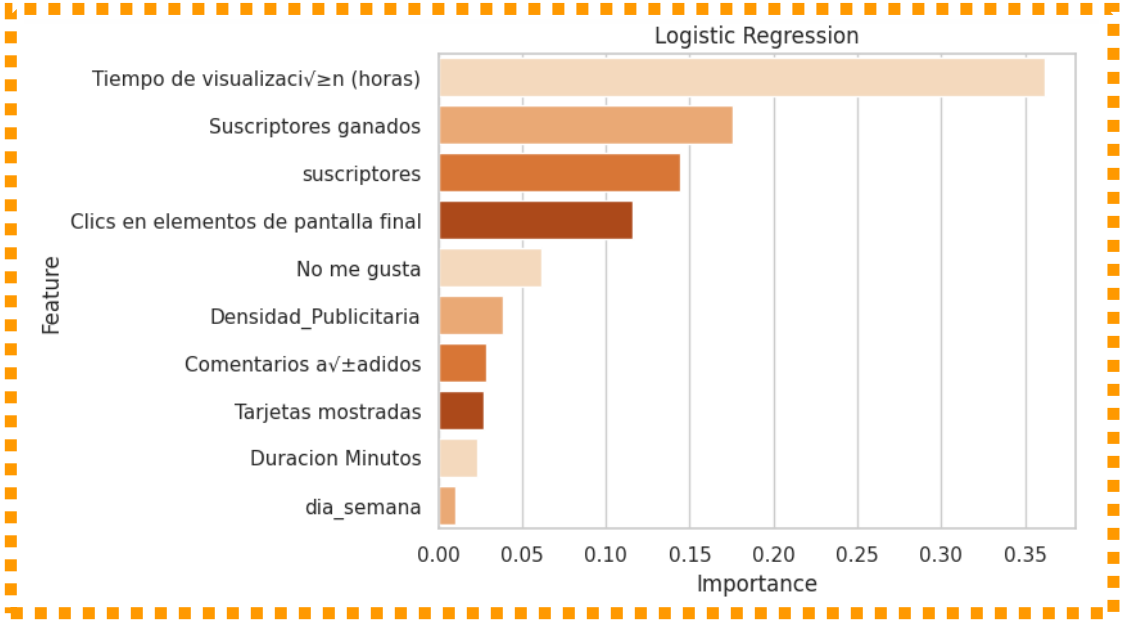
Nota: Las etiquetas de los datos se encuentran balanceadas, el 45% de los videos son exitosos (55% no exitosos).

Modelado de Variables – Importancia de Variables

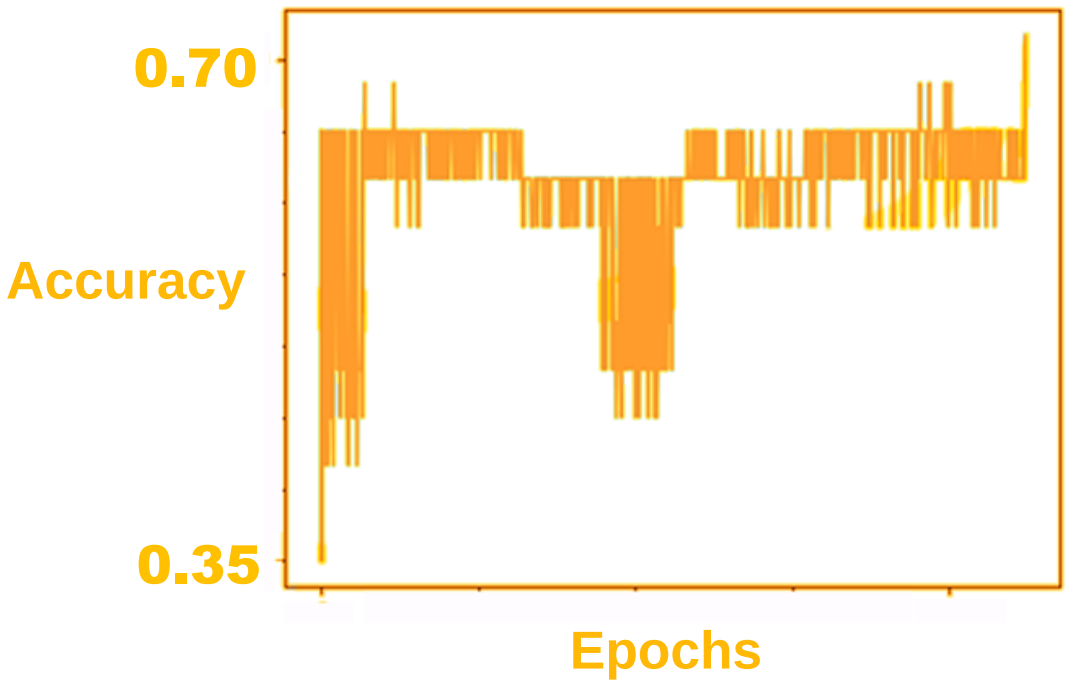








Modelado de Variables – Evolutivo



- 1000 Épocas de entrenamiento
- Genoma = [1 0 1 . . . 0 0 1] o [True False True . . . False False True]
- 32 individuos con 20 iteraciones por epoch

	Evolutivo	Tradicional
Accuracy	0.70	0.83

Regresión Logística – 3 Parámetros

Estimador

Saga

Utiliza el método de descenso de gradiente estocástico y agrega la regularización L1 (LASSO) a la función de costo

Penalidad

L1 (Lasso)

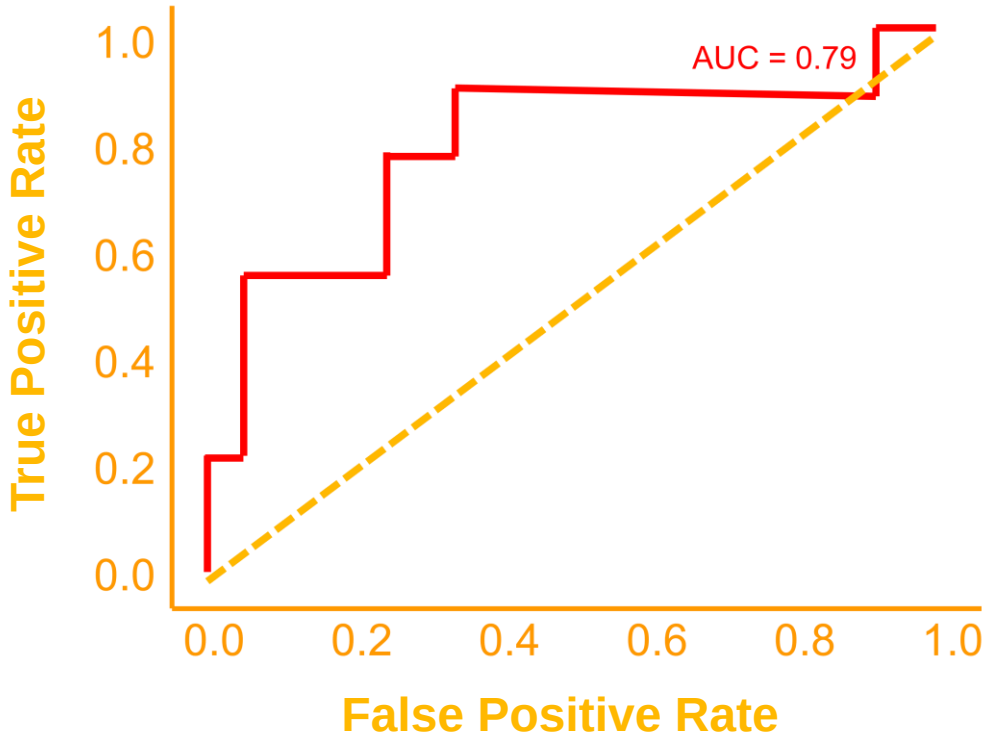
λ

Constante

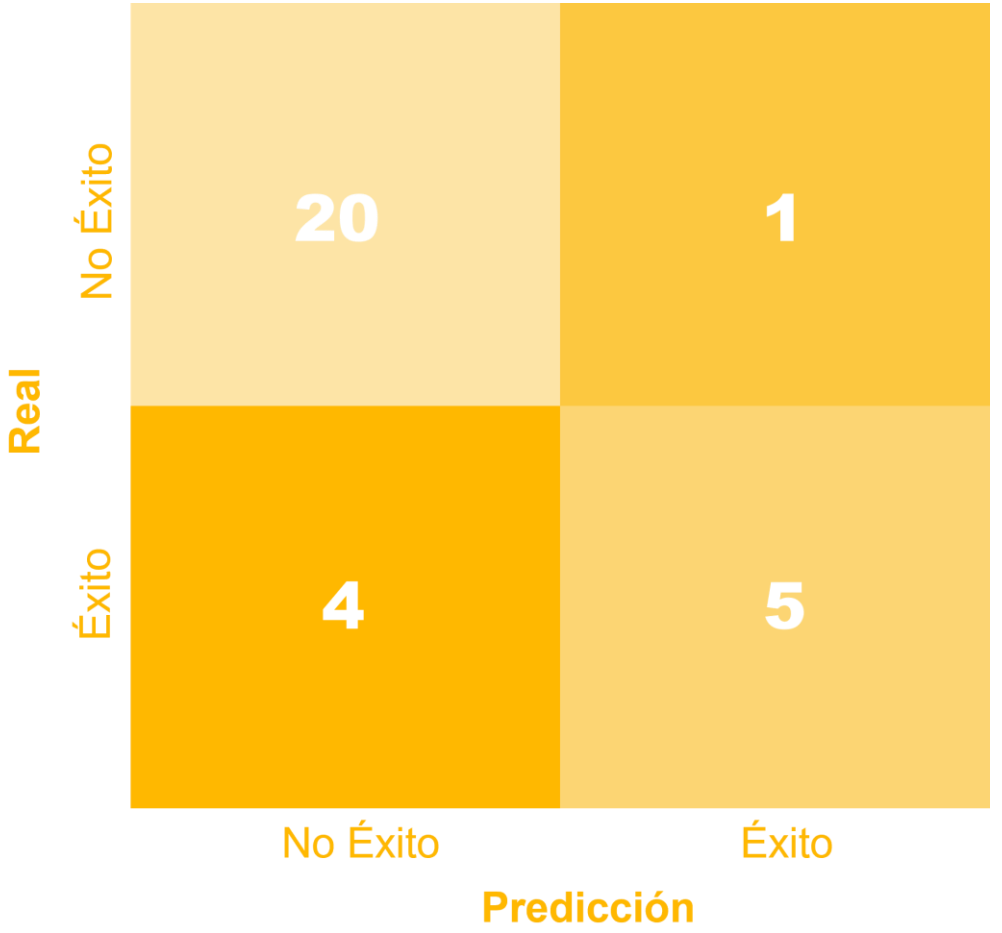
1.34

Regresión Logística

ROC



Matriz de Confusión



Conclusiones

Idoneidad de Variables

48 Variables → **15 Variables**

Gracias al filtrado de variables, se mejora la interpretabilidad del modelo e idoneidad

Aportes

Intervenciones NP

Selección de características

Se hicieron intervenciones no paramétricas para evitar las suposiciones sobre los datos y reducción del sesgo en la selección de características

Aprendizajes

- Regularización
- Importancia de Características
- Flexibilidad