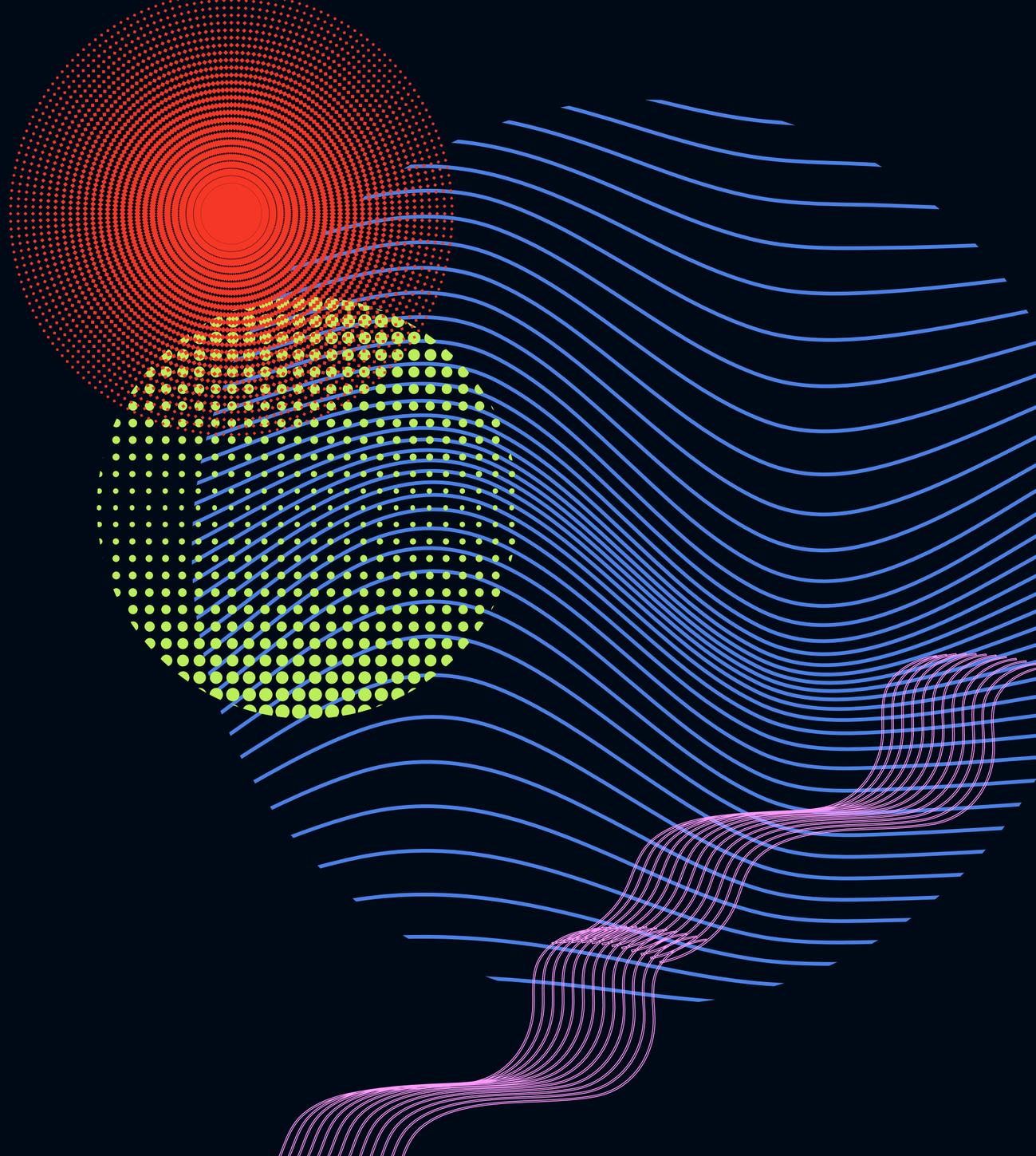
F I S C O G U A R D

HERRAMIENTA DE PREDICCIÓN DE INGRESO Y ANÁLISIS DE RIESGO FISCAL



CONTENIDOS

NUESTROS DATOS

DataSet de UCI Machine Learning Repository

- 14 variables atributo más una variable objetivo.
- Datos demográficos y socioeconómicos recogidos en el censo de EEUU

Column	Non-Null Count	Dtype
age	32561	int64
workclass	32561	object
fnlwgt	32561	int64
education	32561	object
education-num	32561	int64
marital-status	32561	object
occupation	32561	object
relationship	32561	object
race	32561	object
sex	32561	object
capital-gain	32561	int64
capital-loss	32561	int64
hours-per-week	32561	int64
native-country	32561	object
income	32561	object

LIMPPIEZA Y ANÁLISIS

EXTRACCIÓN DE VARIABLES

Variables sintéticas, Ordinales y Dummies

↓

Educacion_superior
Edad_ajustada
Extranjero
horas_puesto

ESCALADO Y TRANSFORMACIÓN

Favoreciendo igualdad de magnitudes y redistribución de desplazamientos

GENERACIÓN DE CLUSTERS

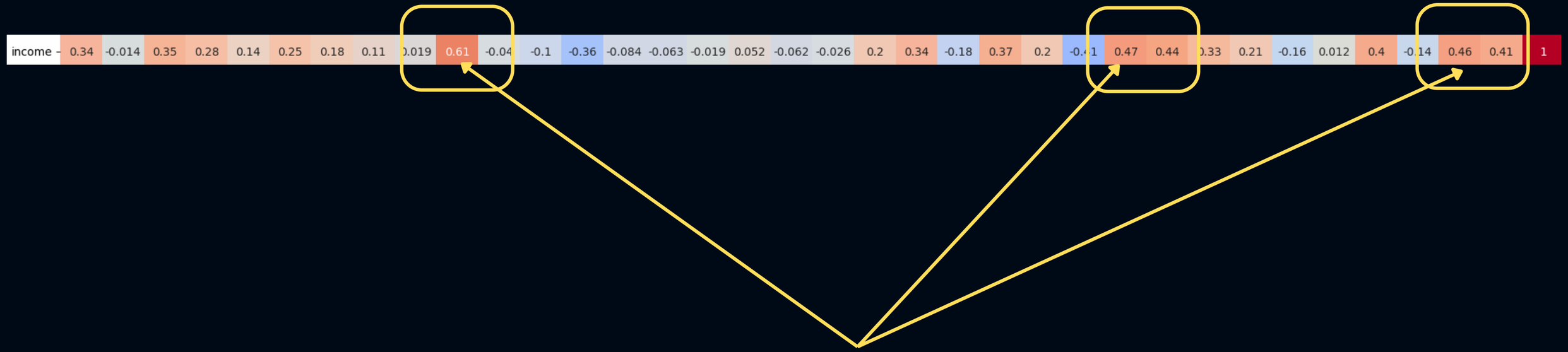
Nueva variable predictora en función de las agrupaciones de un modelo no supervisado

BALANCEO DEL TARGET

Método SMOTE

	income
0	0.75919
1	0.24081

CORRELACIONES



Despues de generar las variables necesarias y procesar bien nuestra muestra de datos, obtuvimos correlaciones interesantes

MODELADO Y EVALUACIÓN

	accuracy	recall	precision
SVC	0.843697	0.668364	0.678733
CAT	0.871027	0.700191	0.748809
LGBM	0.823123	0.735201	0.610788
DT	0.761093	0.854870	0.502808
RF	0.800246	0.792489	0.560811
XG	0.823584	0.737747	0.611287

Seleción de variables

Para decidir que variables eran relevantes y se debían incluir en según qué modelo se utilizó SelectKBest y Feature Importances (Modelos de árboles)

Modelos de clasificación y RN

Se pusieron a prueba todos los grandes conocidos en el mundo de algoritmos de clasificación.

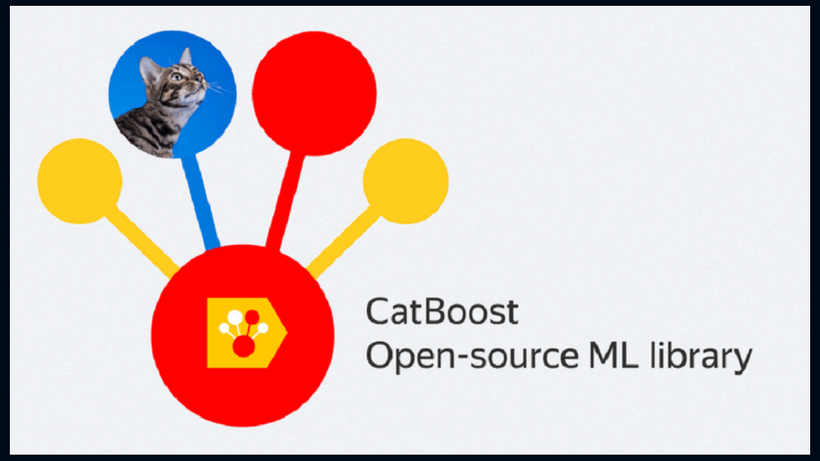
Evaluación

Resumen de lo encontrado con los diferentes modelos

*El modelo de RN no superó la eficiencia de modelos más sencillos

DECISIÓN

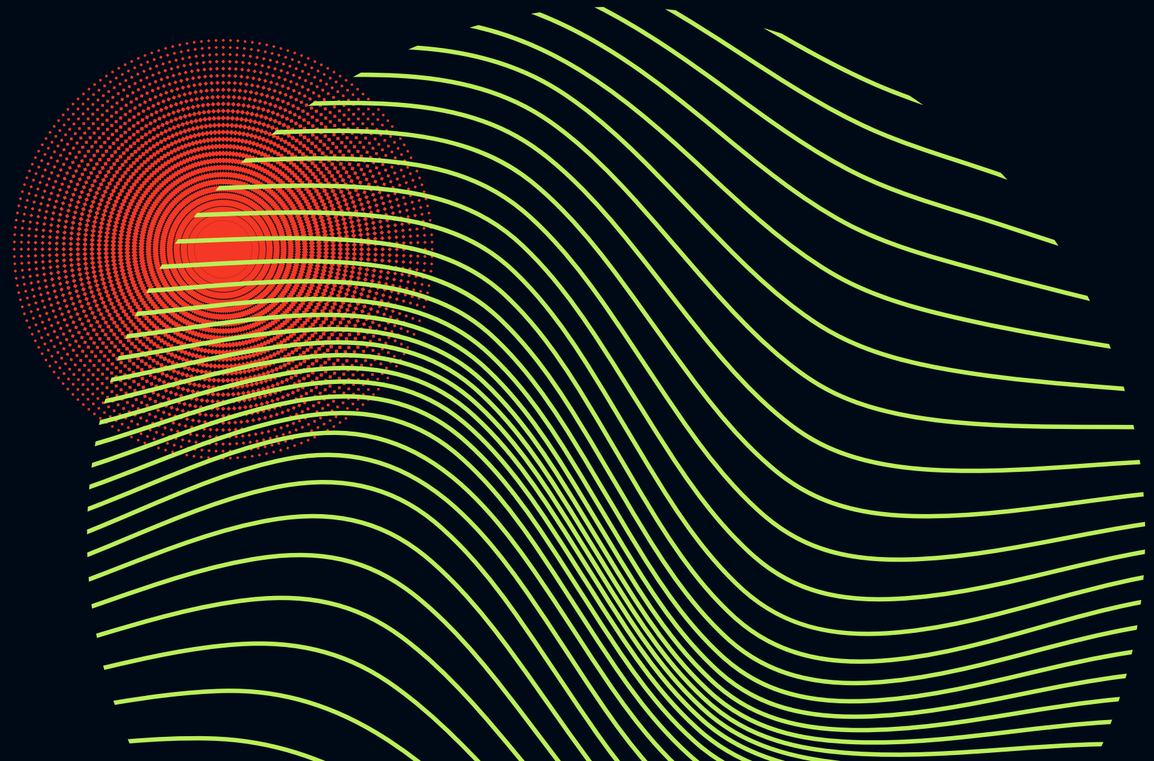
SELECCIÓN DEL MODELO



Catboostclassifier fue el modelo que mejores resultados aportó en general y por ello el seleccionado. Foco en el Recall

NUESTRA VISIÓN DE FUTURO

- Datos robustos que nos dan una base sólida sobre la que trabajar
- Clasificador “ligero” que nos asegurará identificar en un 70 % a los individuos que cobren por encima de un rango específico
- Extracción de nuevos datos de relevancia para el organismo contratador
- Desarrollo e implantación de una nueva arquitectura más elaborada para la red neuronal



GRACIAS

Todo el código y los detalles en el repositorio de GitHub