DATA SCIENTIST

EDA

- Drop en el Target (Graduated / Dropout)

- Data Frame Limpio (Sesgo de conocimiento de negocio)

Train Test Split

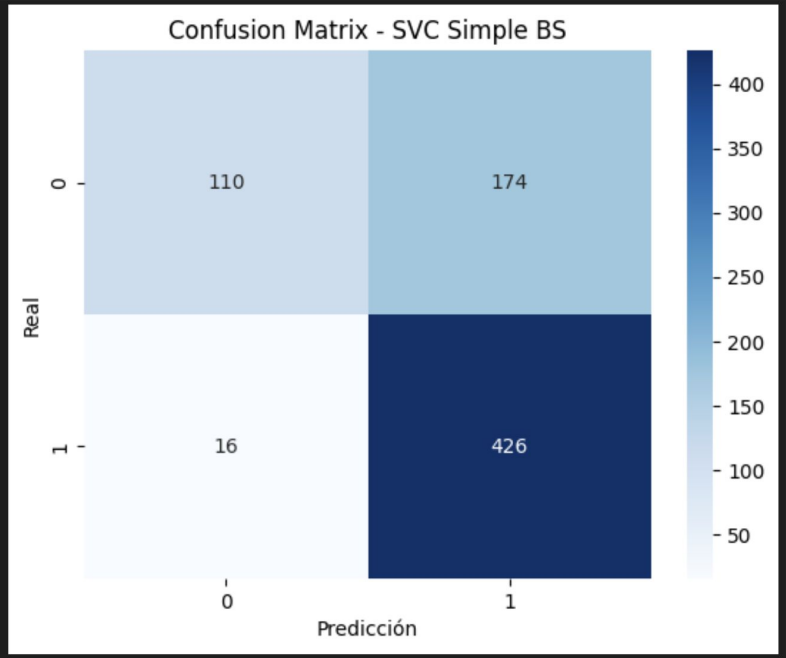
- 80/20

- seed = 42

- Stratify para el target (desbalanceado 60/40)

BASELINE

SVC → F1-score: 0.7077



1 → Graduated

0 → Dropout

1ra Vuelta (Inicialización):

- Inicio los otros 4 modelos supervisado

- Optimizo baseline (SVC)

- Inicio modelo No Supervisado (K-Means)

1ra Vuelta:

	Modelo	Ranking (F1 / silhoutte)
0	XGBoost	0.881937
1	Random Forest	0.877523
2	CatBoost	0.874975
3	Logistic Regression	0.830030
4	SVC (Optimizado)	0.826787
5	Baseline SVC	0.707707
6	KMeans (k=5)	0.215067

- Hiperparametros simples

- K-Means (k = 5)

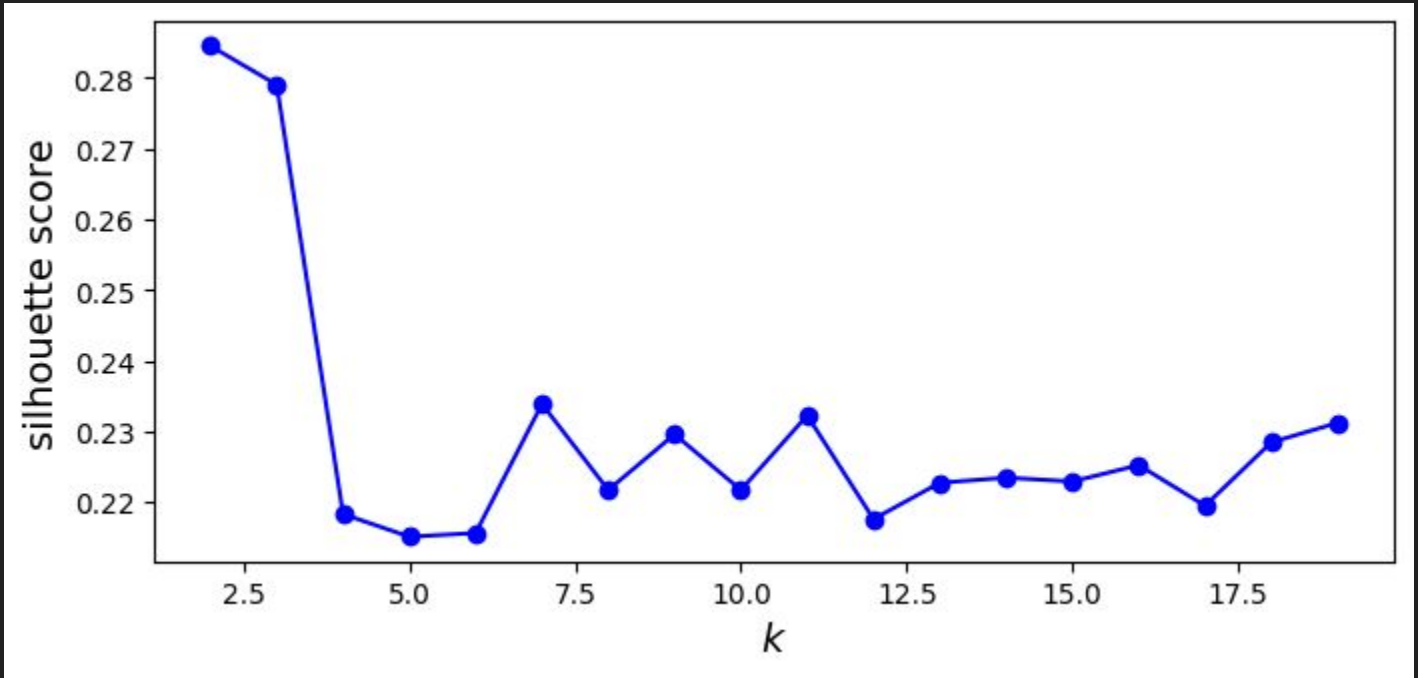
2da Vuelta (Optimización)

- Score → F1_weighted

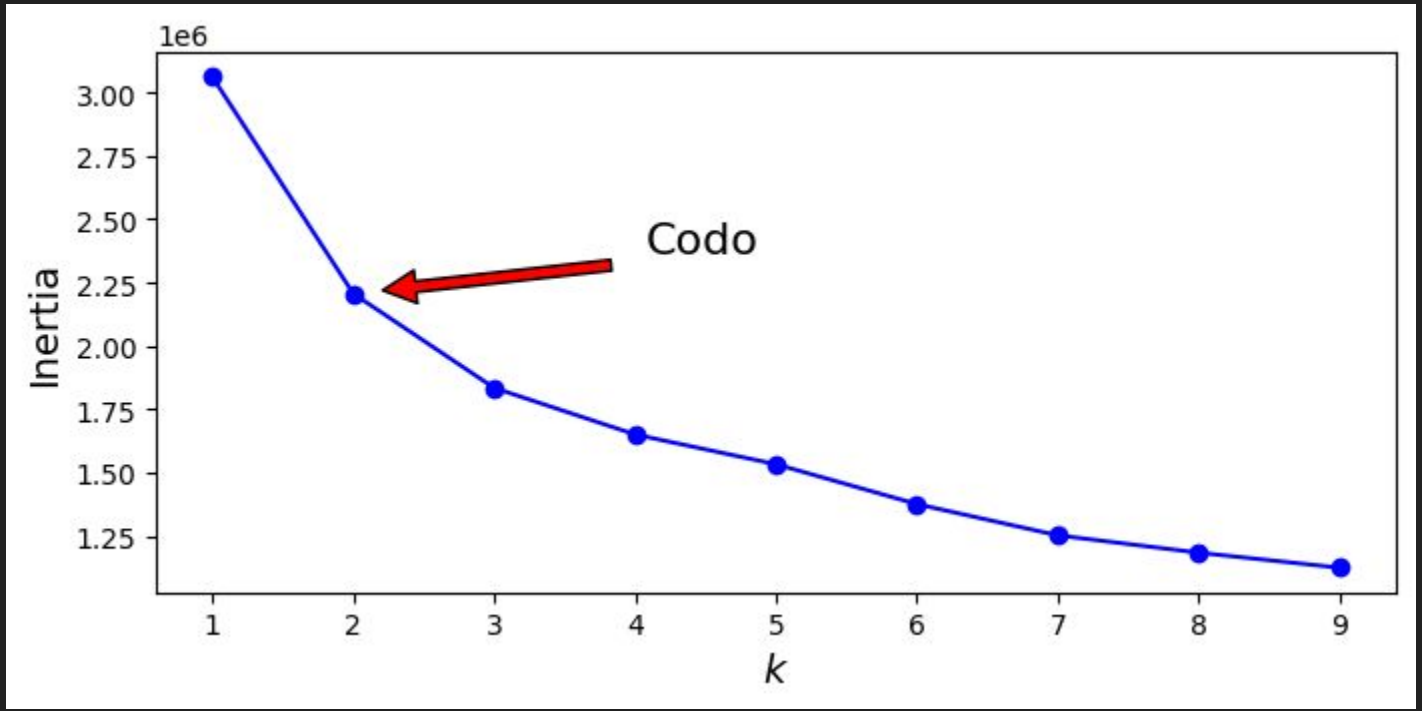
- Hiperparametros parcial

- Busque el mejor k para K-Means

2da Vuelta (K-Means)



2da Vuelta (K-Means)



2da Vuelta:

	Modelo	Ranking (F1 Weighted / silhoutte)
0	XGBoost	0.883404
1	CatBoost	0.883280
2	Logistic Regression	0.875233
3	Random Forest	0.872166
4	SVC (Optimizado)	0.850012
5	Baseline SVC	0.707707
6	KMeans(k=2)	0.284515

- F1_weighted

- Hiperparametros optimizados parcialmente

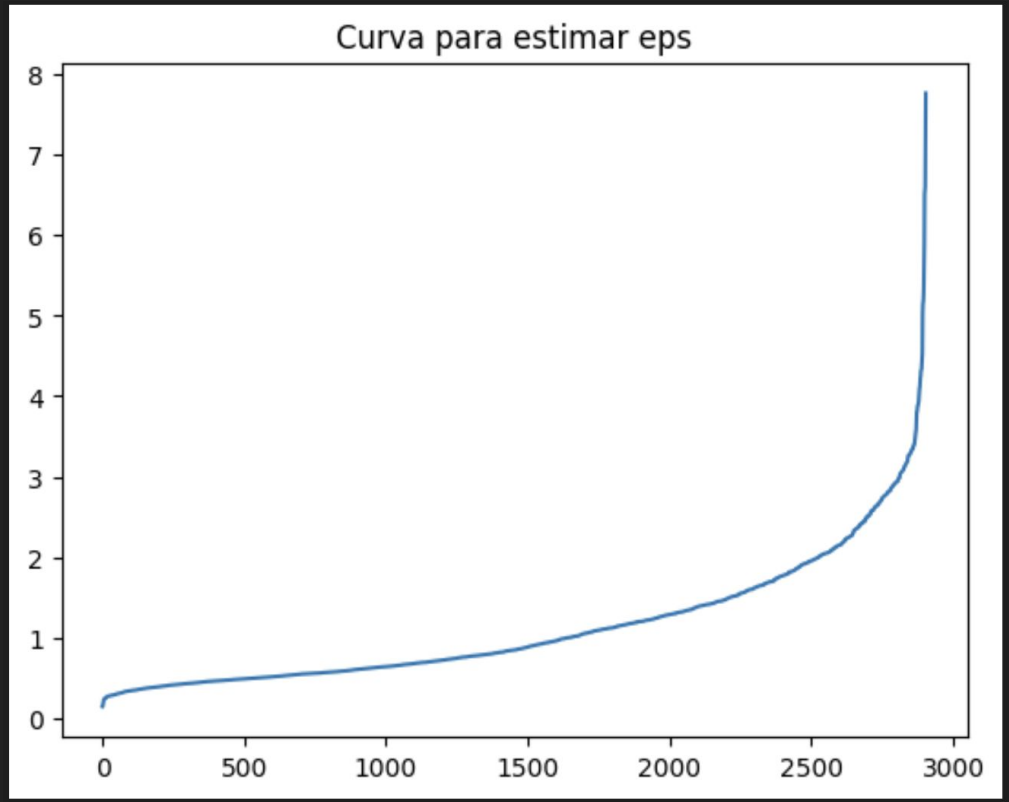
- K-Means (k = 2) 0.21→0.28

3ra Vuelta (Optimización)

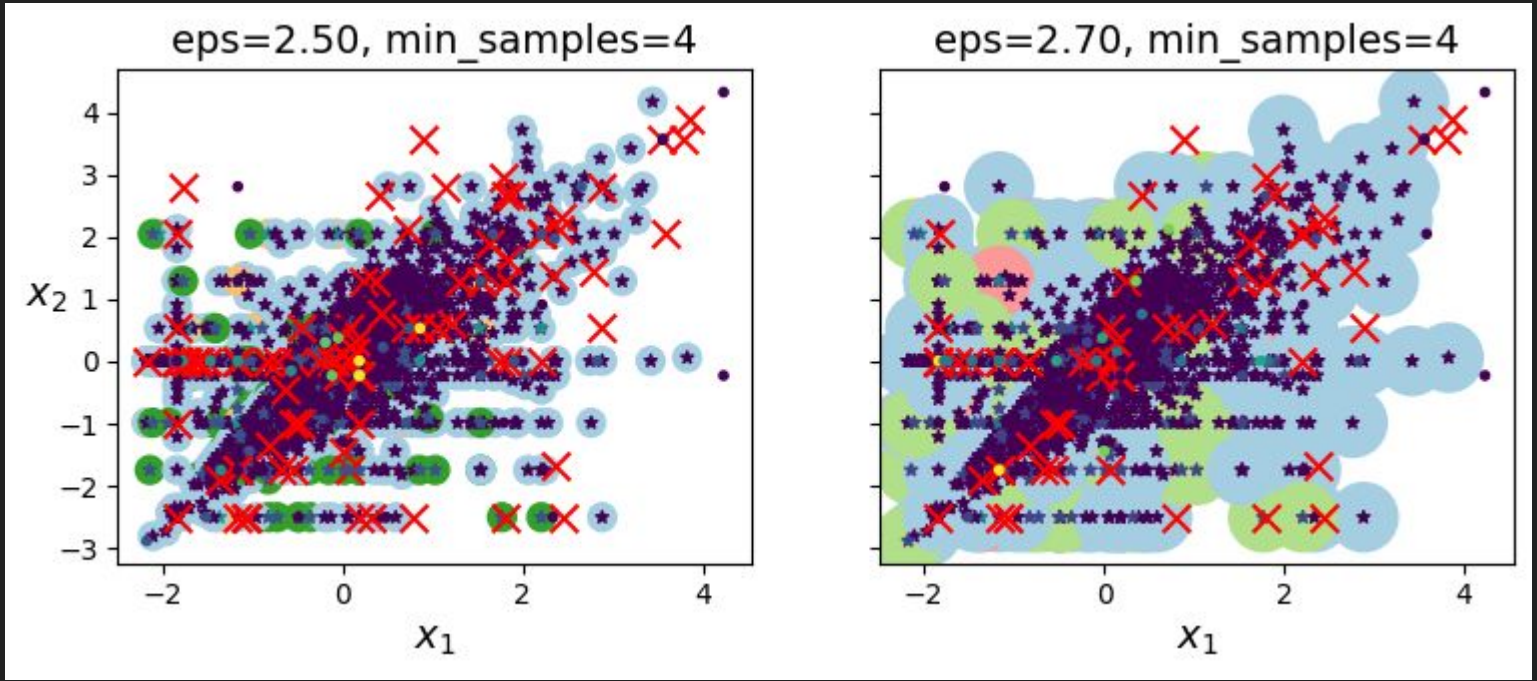
- Optimización Final de Hiperparametros

- ¿DBSCAN > K-Means?

3ra Vuelta (DBSCAN):



3ra Vuelta (DBSCAN):



3ra Vuelta:

	Modelo	Ranking (F1 Weighted / silhoutte)
0	Logistic Regression	0.883645
1	XGBoost	0.883525
2	CatBoost	0.871892
3	Random Forest	0.870962
4	SVC (Optimizado)	0.853182
5	Baseline SVC	0.707707
6	DBSCAN (eps = 2.7)	0.217711

- F1_weighted

- Optimización Final

- DBSCAN

(eps = 2.7, min_samples = 4)

K-Means > DBSCAN

0.28 > 0.21

4ta Vuelta (Optimización)

- TOP 3 a lo largo de las vueltas

- Random Forest, Logistic Regresion, CatBoost, XGBoost.

- No pude optimizar más el RF

4ta Vuelta:

	Modelo	Ranking (F1 Weighted)
0	XGBoost	0.884988
1	CatBoost	0.878996
2	Logistic Regression	0.876953
3	Baseline SVC	0.707707

	Modelo	Ranking (AUC)
0	Logistic Regression	0.925562
1	CatBoost	0.918775
2	XGBoost	0.918504

¿Qué features prioriza RF, LR, XGBoost y CatBoost?

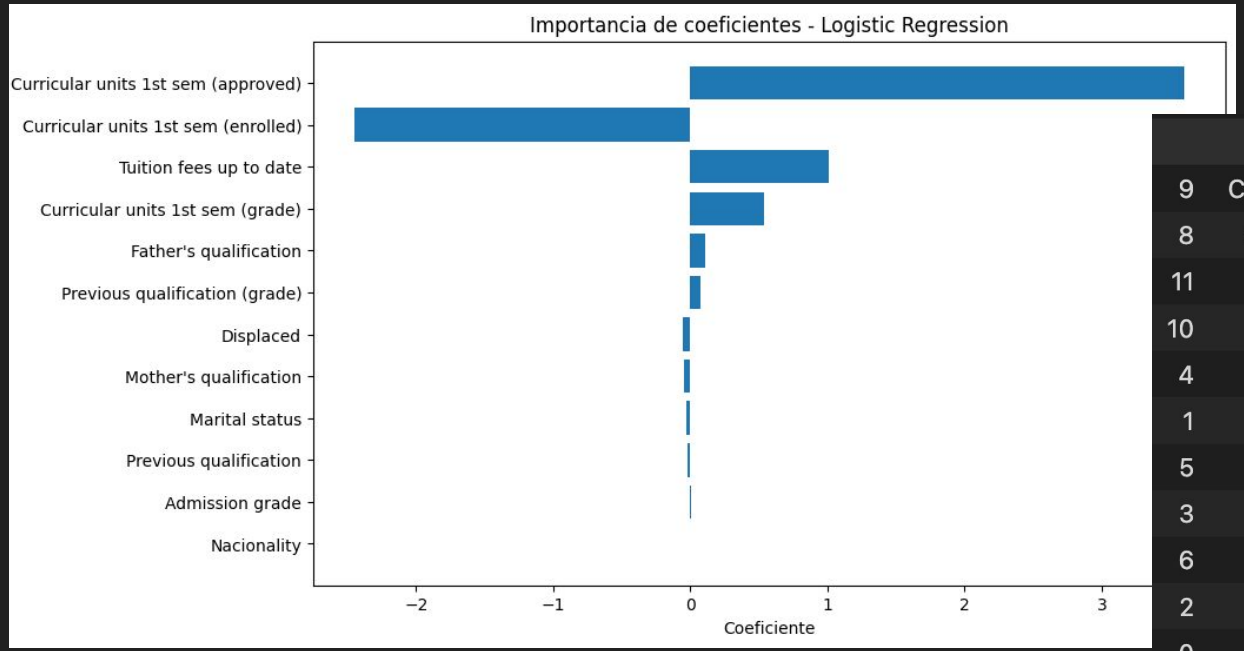
FEATURE IMPORTANCES

Random Forest:

	feature	ranking
0	Admission grade	1
1	Previous qualification (grade)	1
8	Curricular units 1st sem (enrolled)	1
9	Curricular units 1st sem (approved)	1
10	Curricular units 1st sem (grade)	1
11	Tuition fees up to date	1
4	Father's qualification	2
3	Mother's qualification	3
2	Previous qualification	4
5	Displaced	5
6	Marital status	6
7	Nacionality	7

FEATURE IMPORTANCES

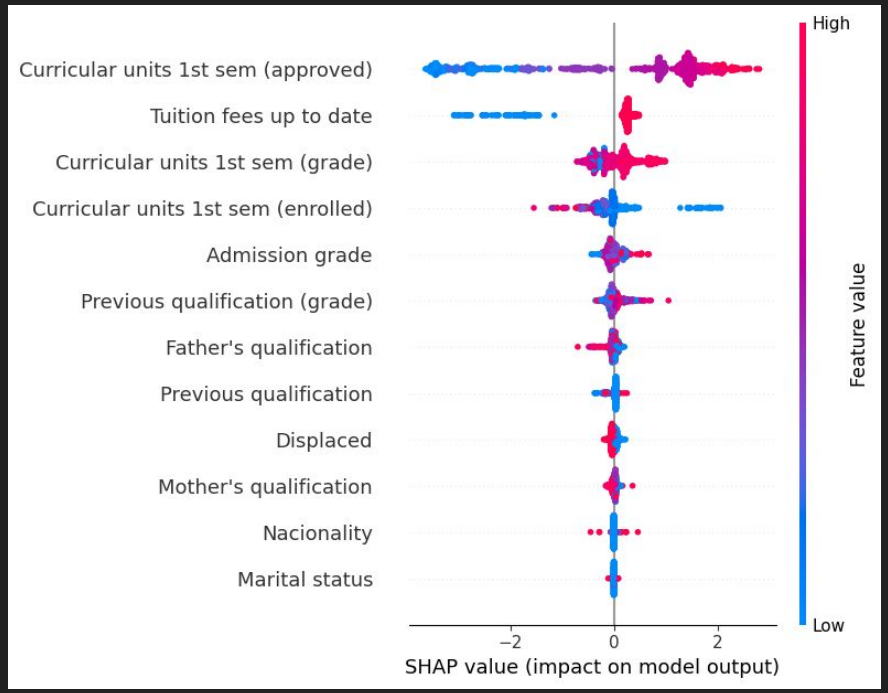
Regresión Logística:



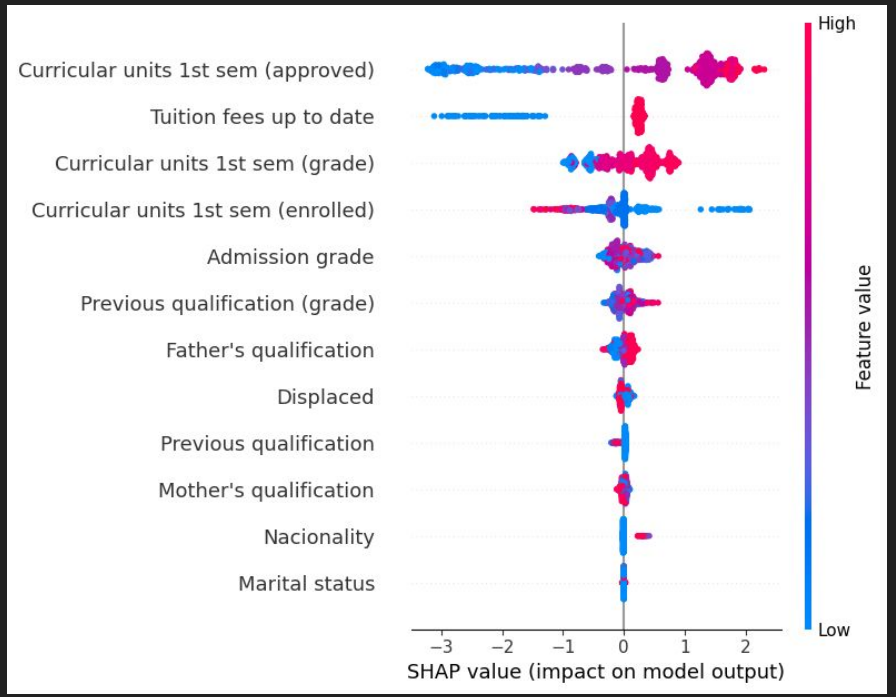
	Feature	Coefficient
9	Curricular units 1st sem (approved)	3.599241
8	Curricular units 1st sem (enrolled)	-2.446455
11	Tuition fees up to date	1.012954
10	Curricular units 1st sem (grade)	0.542185
4	Father's qualification	0.107911
1	Previous qualification (grade)	0.078671
5	Displaced	-0.053173
3	Mother's qualification	-0.046338
6	Marital status	-0.029302
2	Previous qualification	-0.022134
0	Admission grade	0.009668
7	Nacionality	0.001043

FEATURE IMPORTANCES

CatBoost (shap):

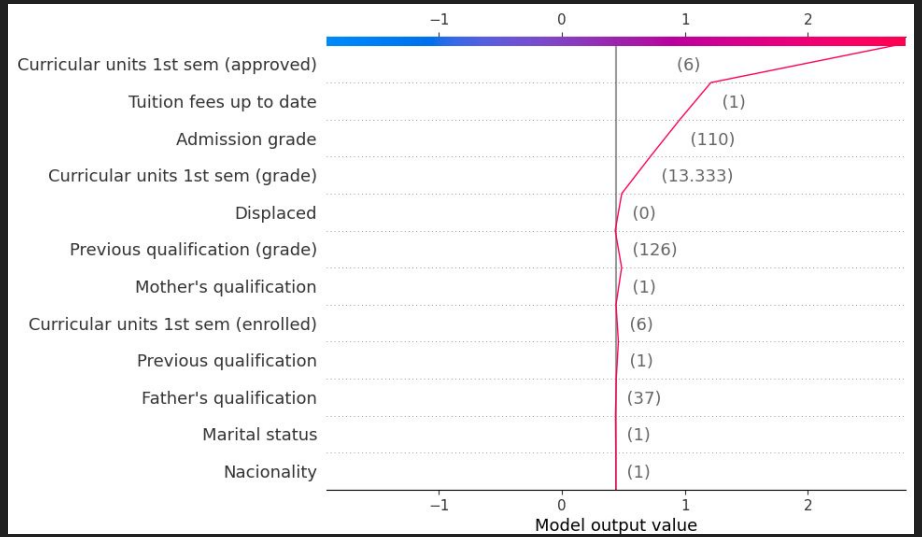


XGBoost (shap):

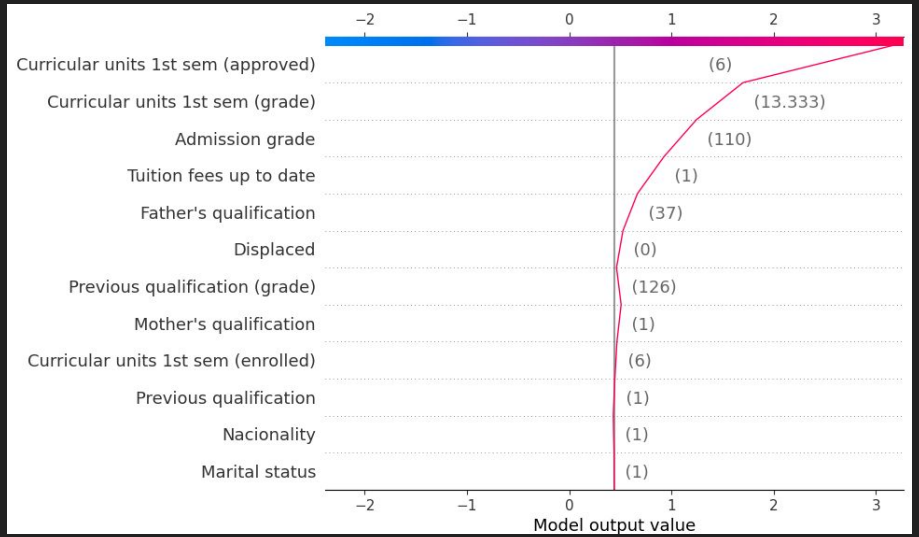


FEATURE IMPORTANCES

CatBoost (shap):

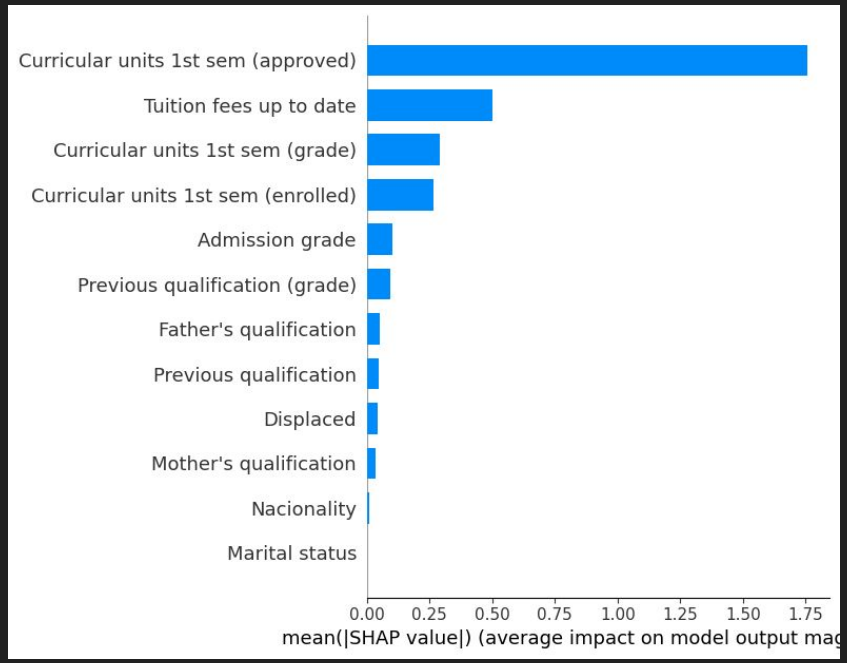


XGBoost (shap):

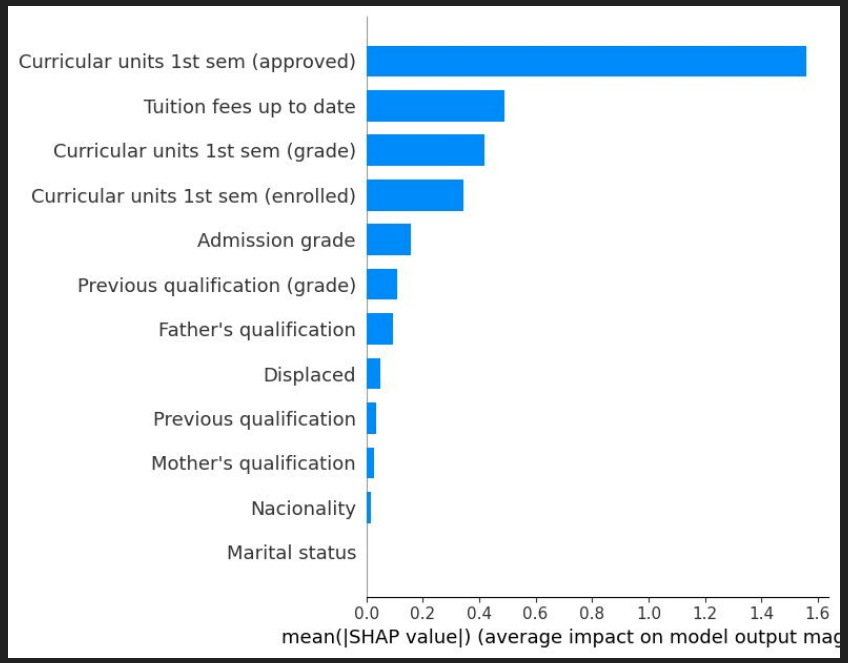


FEATURE IMPORTANCES

CatBoost (shap):

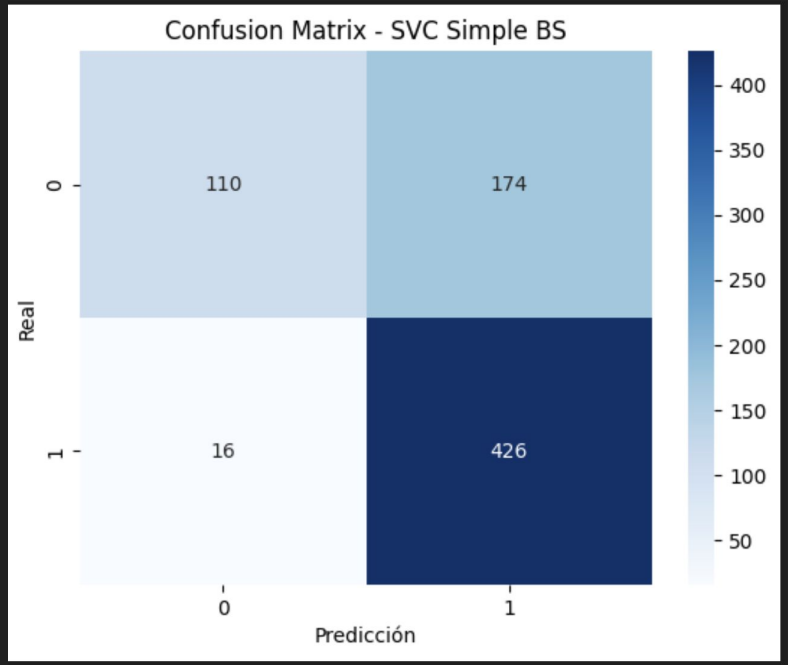


XGBoost (shap):



Baseline

SVC → F1-score: 0.7077

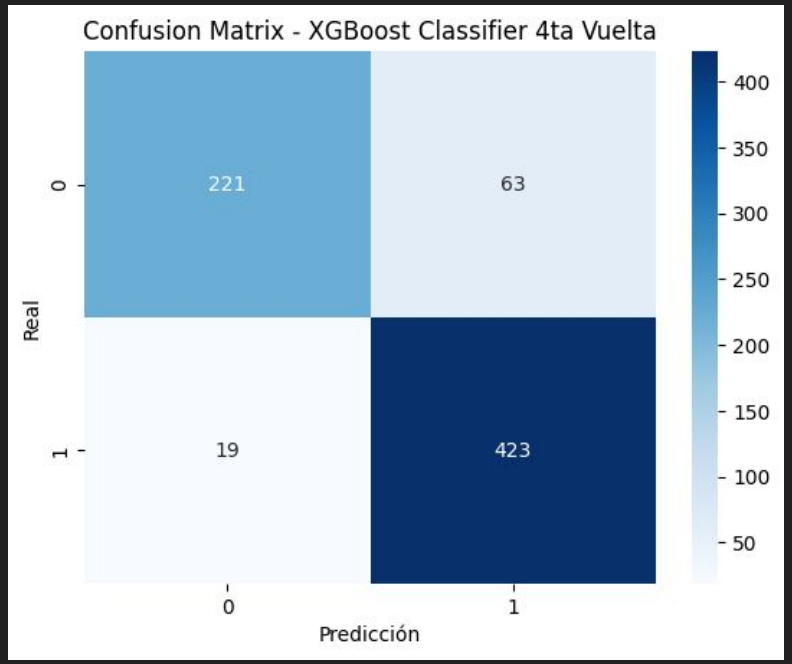


1 → Graduated

0 → Dropout

Modelo Final

XGBoost → F1_weighted: 0.8849



SVC → F1-score: 0.7077

XGBoost → F1_weighted: 0.8849

0.707 → 0.884

Mejora ≈ 0.2

Mejoras a futuro

- No Supervisado → Identificar diferentes perfiles a mitad de curso

-No Supervisado → Identificar diferentes perfiles en la matriculación

(Riesgo ético por sesgos)

- Probar con un 70/30 split
- Implementar mejor los feature importance
- Otra metodología, para mejorar con respecto a necesidades de negocio
- Otra métrica como curva de precision-recall de cada clase