Modelos y simulación de sistemas II

Departamento de Ingeniería de Sistemas Facultad de Ingeniería

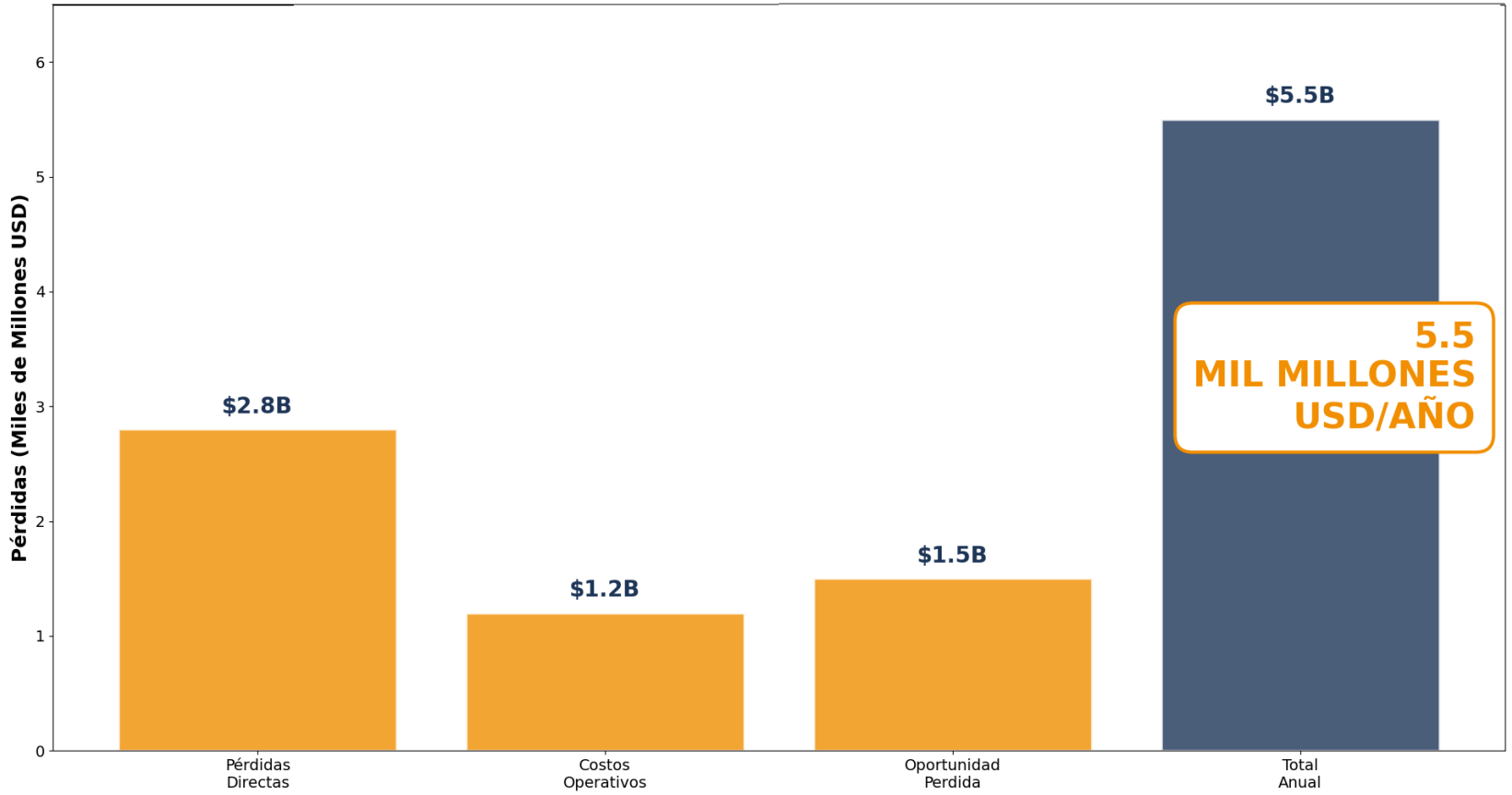
UNIVERSIDAD DE ANTIOQUIA

1803

INTEGRANTES:
Juan David Arismendy
Laura Tobón

PREDICCIÓN DE CANCELACIONES EN RESERVACIONES DE HOTELES

UNIVERSIDAD DE ANTIOQUIA

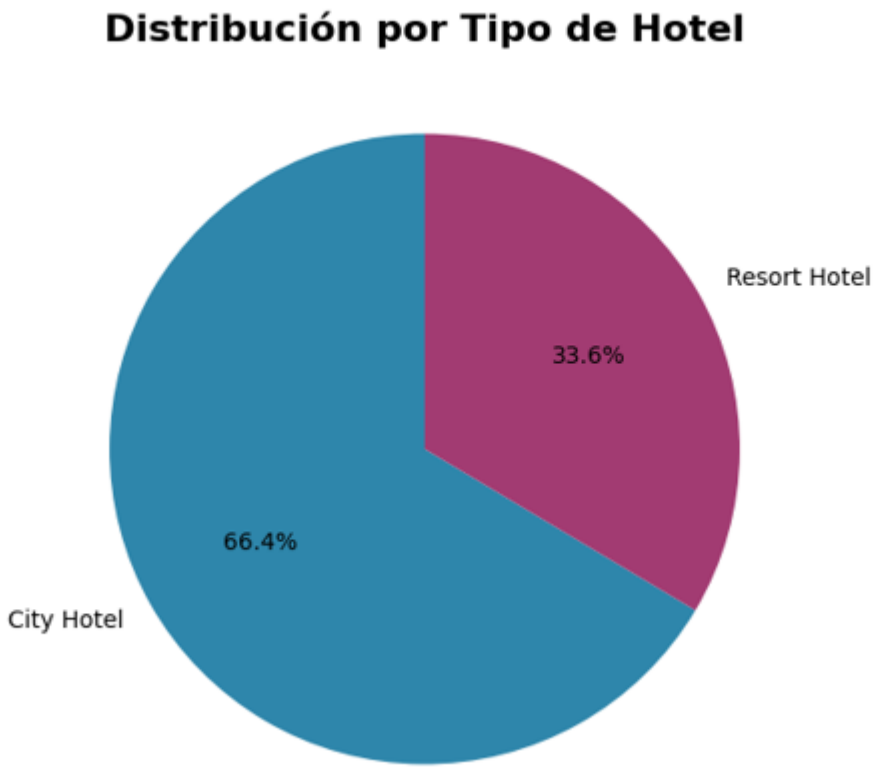


Las cancelaciones representan un problema de **millones de dólares** para la industria hotelera.

Nuestro objetivo era claro: crear un sistema que prediga cancelaciones, usando técnicas de machine learning y optimización de características.

PREDICCIÓN DE CANCELACIONES EN RESERVACIONES DE HOTELES

UNIVERSIDAD DE ANTIOQUIA

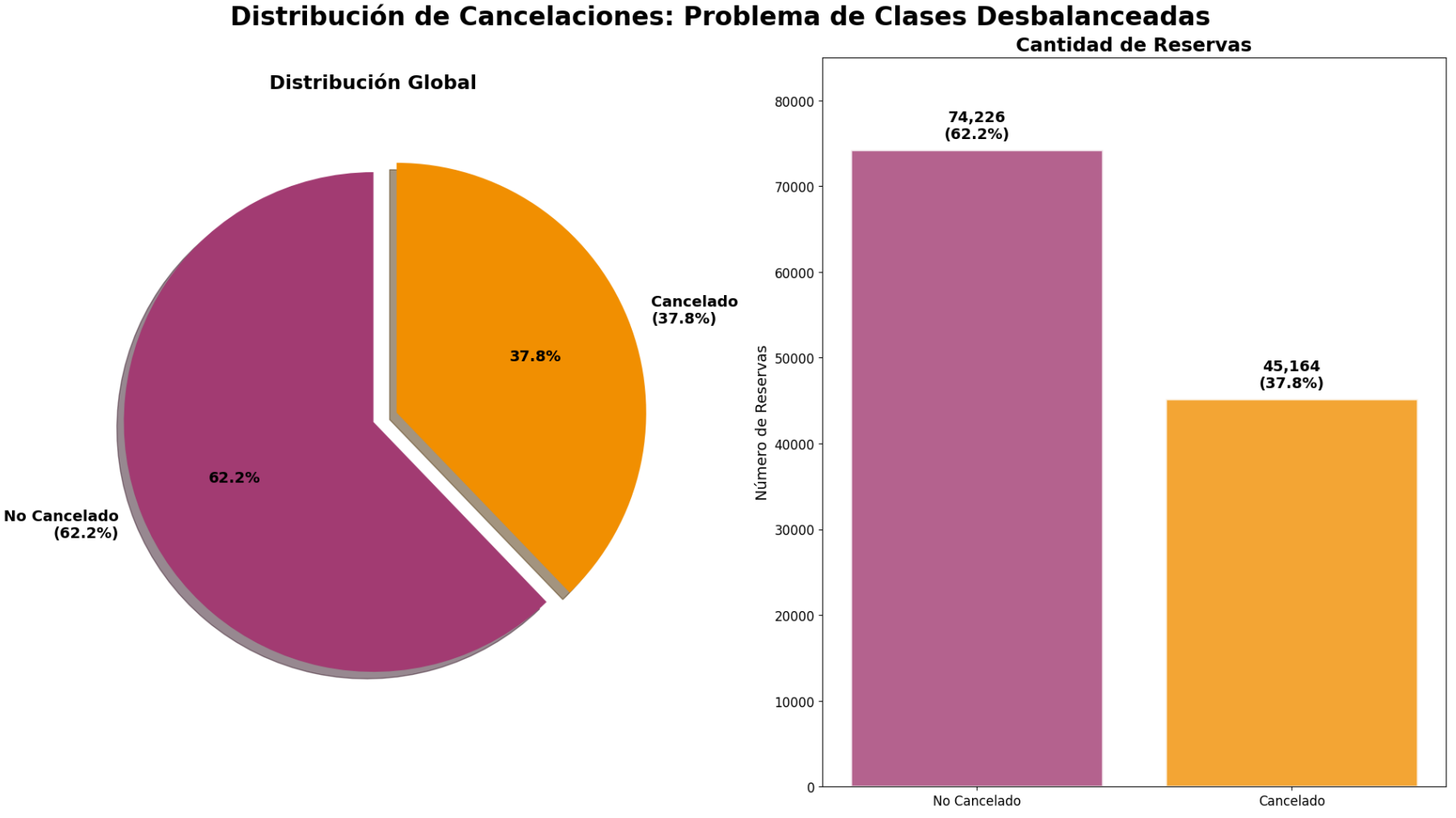


Usamos un dataset real de **119,390 reservas hoteleras** de Kaggle, con información detallada de hoteles urbanos y resort entre 2015-2017.

Este incluye **36 características** como tiempo de anticipación, tipo de cliente, tarifa diaria, y por supuesto, si la reserva fue cancelada o no.

METODOLOGÍA

UNIVERSIDAD DE ANTIOQUIA



- Presencia de clases ligeramente desbalanceadas
- Requiere técnicas como SMOTE para balanceo sintético
- Modelos a probar: Regresión Logística, K-Nearest Neighbors, Random Forest, Redes Neuronales, y Support Vector Machines.

METODOLOGÍA

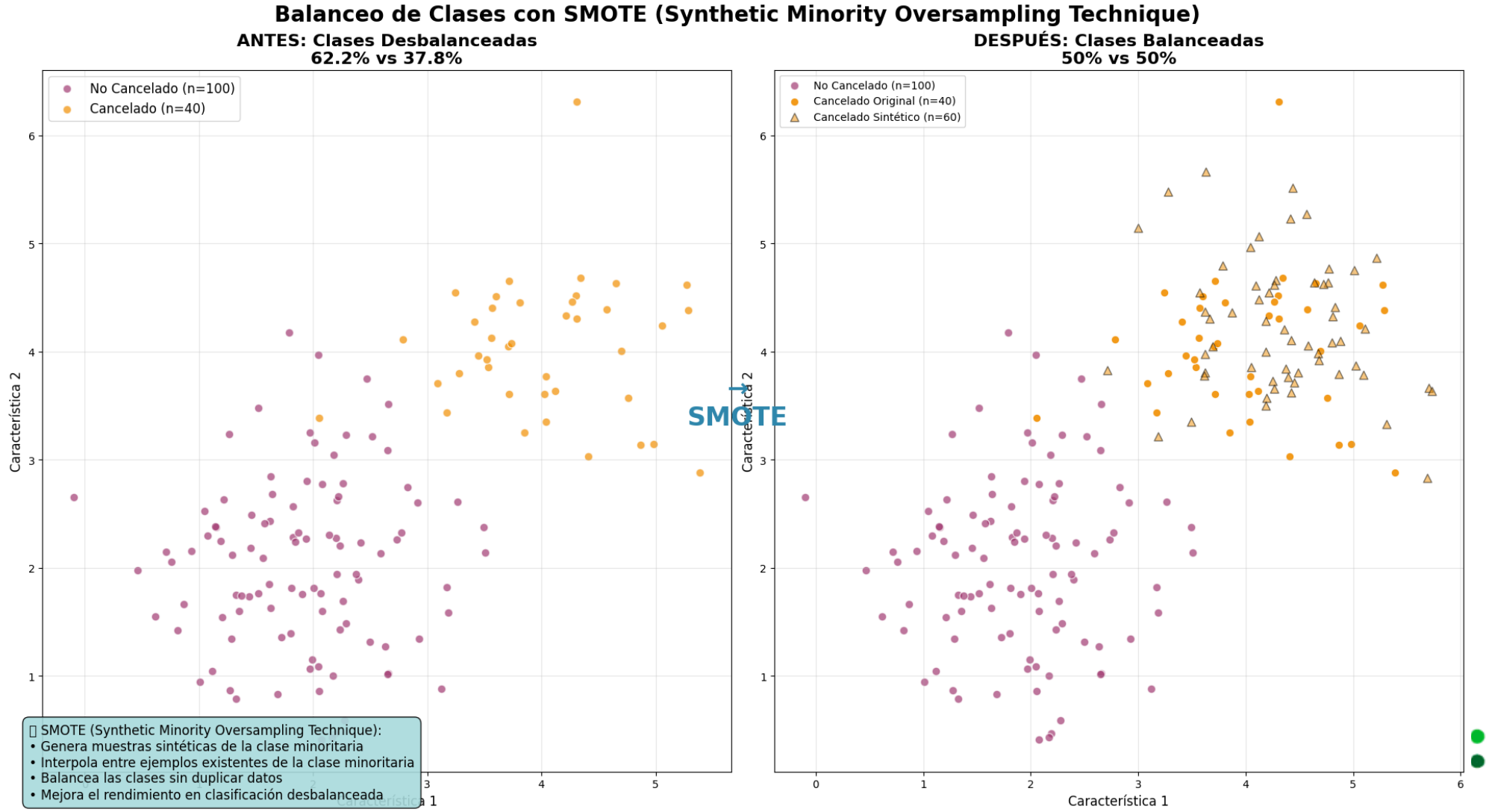
UNIVERSIDAD DE ANTIOQUIA

Modelo	Accuracy Promedio ± Intervalo de Confianza	Precision Promedio ± Intervalo de Confianza	Recall Promedio ± Intervalo de Confianza	F1-Score Promedio ± Intervalo de Confianza	AUC-ROC Promedio ± Intervalo de Confianza	Hiperparámetros Óptimos
Regresión Logística	0.825 ± 0.018	0.769 ± 0.026	0.768 ± 0.024	0.768 ± 0.023	0.900 ± 0.017	C=1
K-Vecinos más Cercanos (KNN)	0.814 ± 0.025	0.779 ± 0.038	0.712 ± 0.033	0.744 ± 0.033	0.880 ± 0.015	n_neighbors=11, metric= manhattan, weights='distance'
Random Forest	0.863 ± 0.010	0.838 ± 0.019	0.791 ± 0.016	0.814 ± 0.014	0.933 ± 0.014	n_estimators=100, max_features= 0.5, max_depth= 30, min_samples_leaf = 2
Red Neuronal Artificial	0.813 ± 0.014	0.800 ± 0.040	0.677 ± 0.027	0.733 ± 0.023	0.892 ± 0.021	hidden_layer_sizes = 50, activation = relu, solver = adam, alpha = 0.01, learning_rate = constant
Máquina de Vectores de Soporte (SVM)	0.846 ± 0.013	0.797 ± 0.022	0.797 ± 0.016	0.797 ± 0.016	0.923 ± 0.013	C = 10, kernel =rbf, gamma = scale

- Cada modelo pasó por un proceso de **Grid Search** con validación cruzada estratificada de 5 folds para encontrar los mejores hiperparámetros.

METODOLOGÍA

UNIVERSIDAD DE ANTIOQUIA
1803



Para el desbalance de clases: **SMOTE** (Synthetic Minority Oversampling Technique)

Se aplicaron dos pasos de **preprocesamiento automático**: **StandardScaler** y **OneHotEncoder**.

F1-Score como métrica principal: balancea precisión y recall → problemas con clases desbalanceadas.

MÉTRICAS DE DESEMPEÑO

UNIVERSIDAD DE ANTIOQUIA

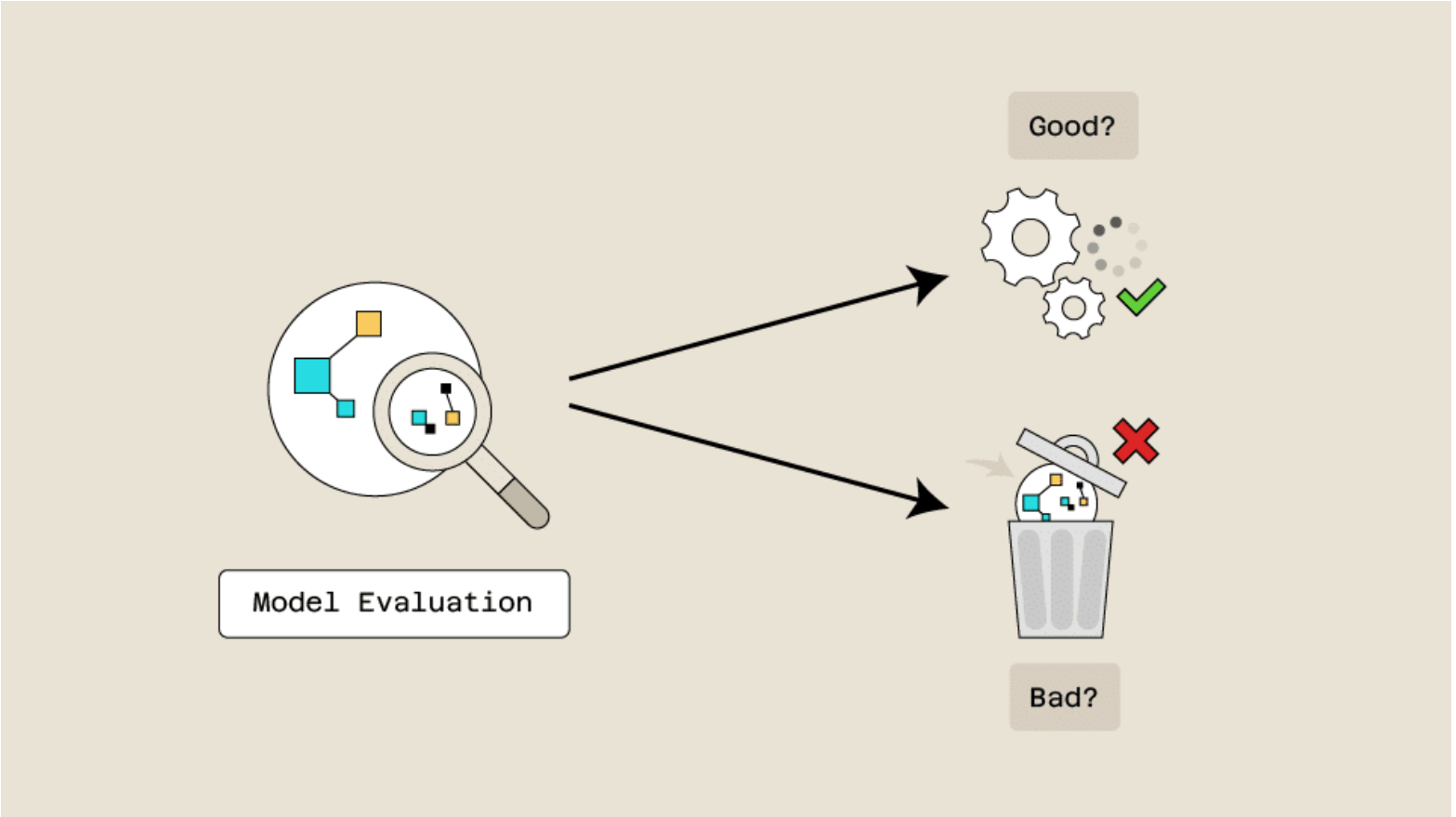
Exactitud (Accuracy)

Presición (Precision)

F1-score

Sensibilidad (Recall)

ROC-AUC



RESULTADOS

UNIVERSIDAD DE ANTIOQUIA

Tabla de Rendimiento de Modelos

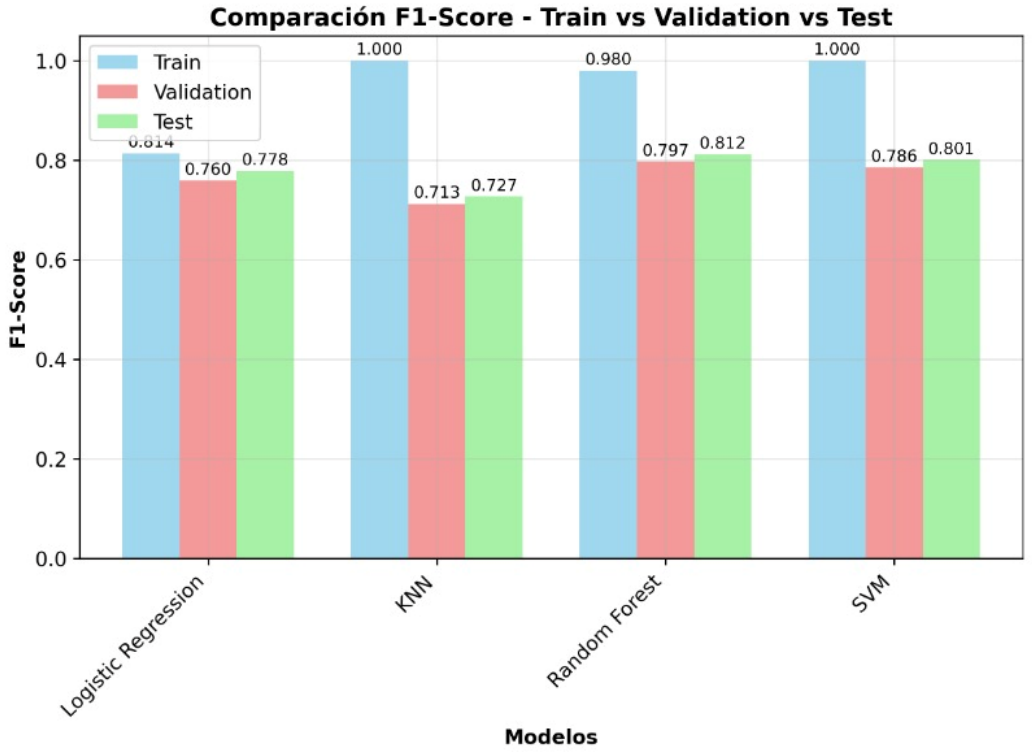
Modelo	F1-Score	AUC-ROC	Accuracy	Precision	Recall
Logistic Regression	0.768 ± 0.023	0.900 ± 0.017	0.825 ± 0.018	0.769 ± 0.026	0.768 ± 0.024
KNN	0.744 ± 0.033	0.880 ± 0.015	0.814 ± 0.025	0.779 ± 0.038	0.712 ± 0.033
Random Forest	0.814 ± 0.014	0.933 ± 0.014	0.863 ± 0.010	0.838 ± 0.019	0.791 ± 0.016
MLP	0.733 ± 0.023	0.892 ± 0.021	0.813 ± 0.014	0.800 ± 0.040	0.677 ± 0.044
SVM	0.797 ± 0.016	0.923 ± 0.013	0.846 ± 0.013	0.797 ± 0.022	0.797 ± 0.016

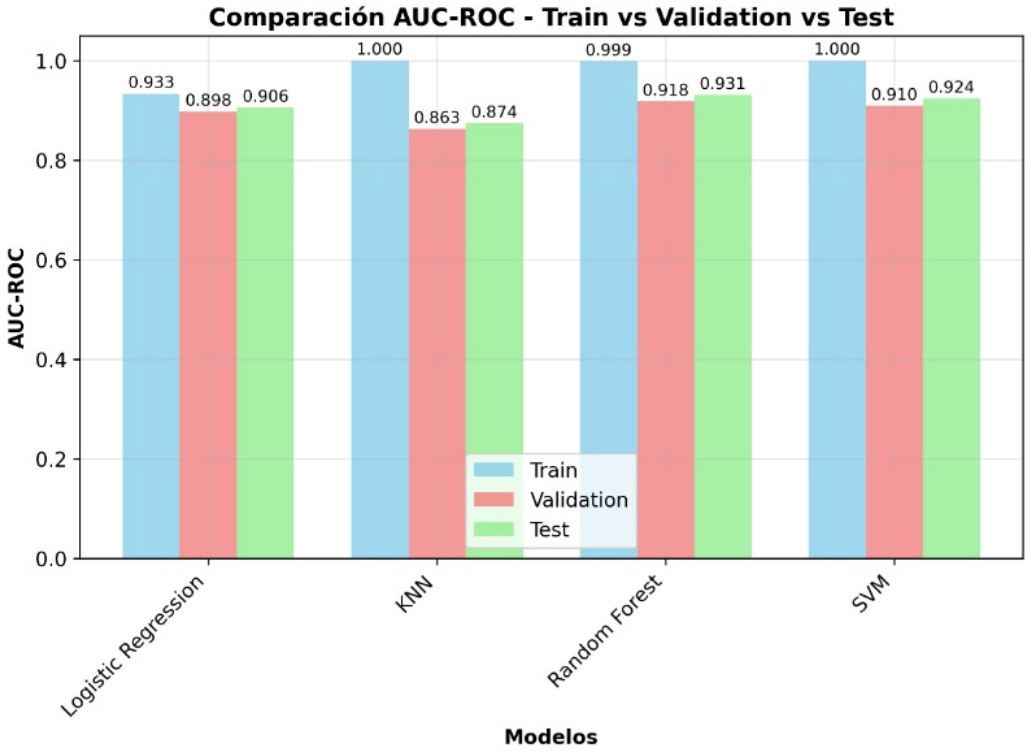
Random Forest lidera con un F1-Score de **0.814** y AUC-ROC de **0.933**.

SVM queda en segundo lugar con **0.797** de F1-Score, seguido por Regresión Logística con **0.768**. KNN y MLP tuvieron rendimientos menores.

RESULTADOS

UNIVERSIDAD DE ANTIOQUIA

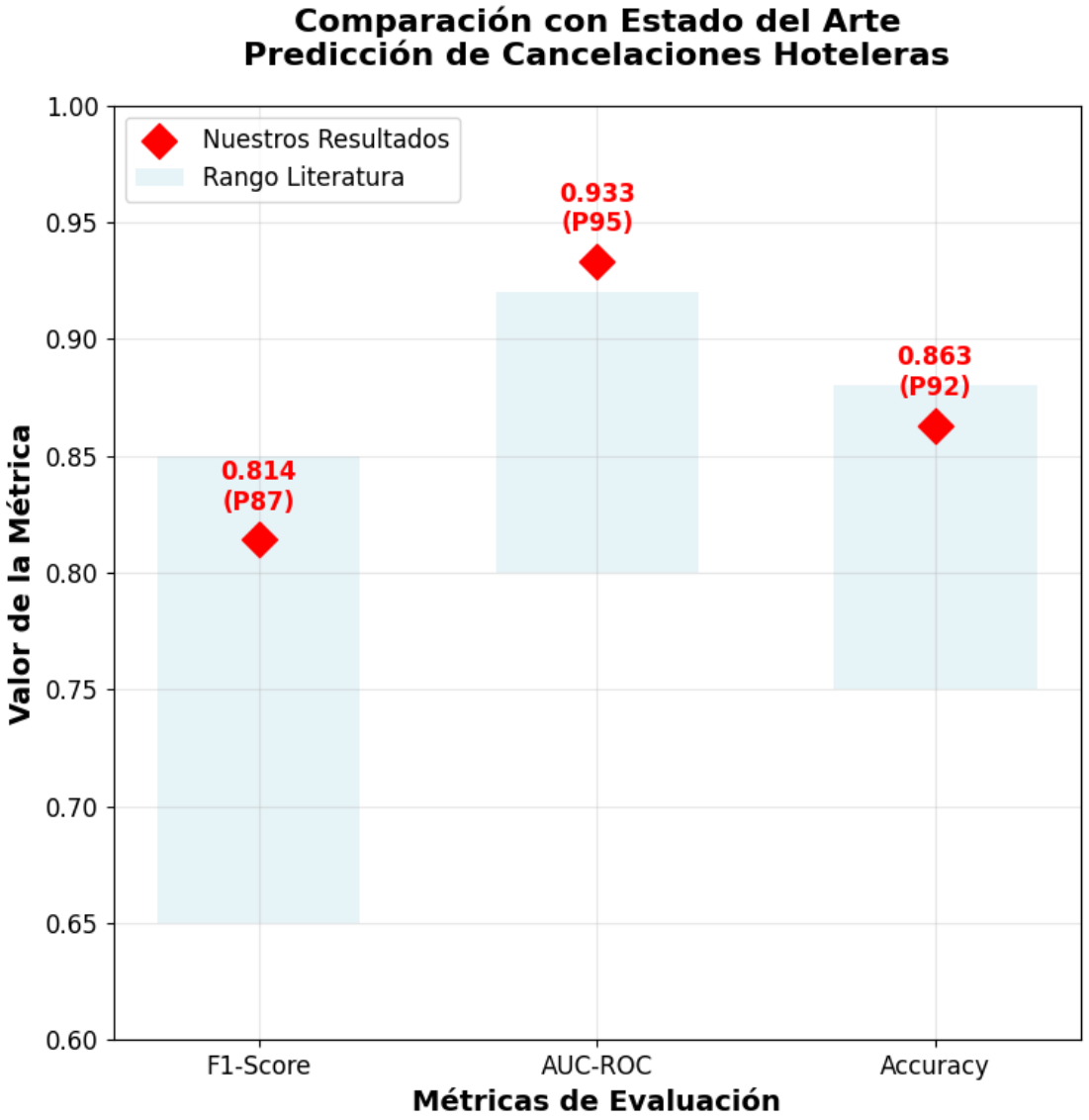




Modelos óptimos
Random Forest
F1-Score: 0.812 AUC-ROC: 0.931
SMV
F1-Score: 0.801 AUC-ROC: 0.924

COMPARACIÓN CON EL ESTADO DEL ARTE

UNIVERSIDAD DE ANTIOQUIA



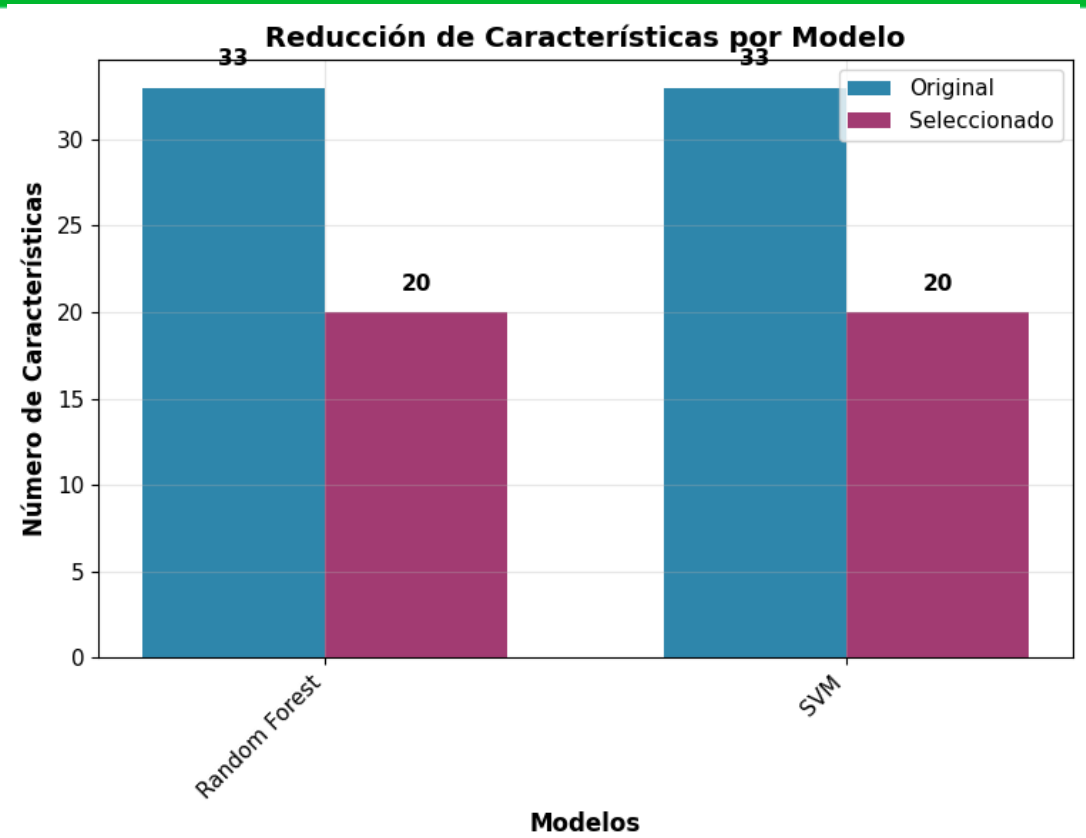
Mientras la literatura reporta F1-Scores entre 0.65-0.85, nosotros logramos **0.814**.

En AUC-ROC, la literatura va de 0.80-0.92, y nosotros alcanzamos **0.933**, ubicándonos en el **percentil 95.**

Esto significa que nuestro modelo es competitivo

TÉCNICAS AVANZADAS

UNIVERSIDAD DE ANTIOQUIA



Variables con mayor correlación con la variable objetivo (is_canceled):
Más correlacionada: lead_time (+0.29) → a mayor anticipación, más probable la cancelación.
Menos correlacionada: total_of_special_requests (–0.23) → si el cliente pide cosas, es menos probable que cancele

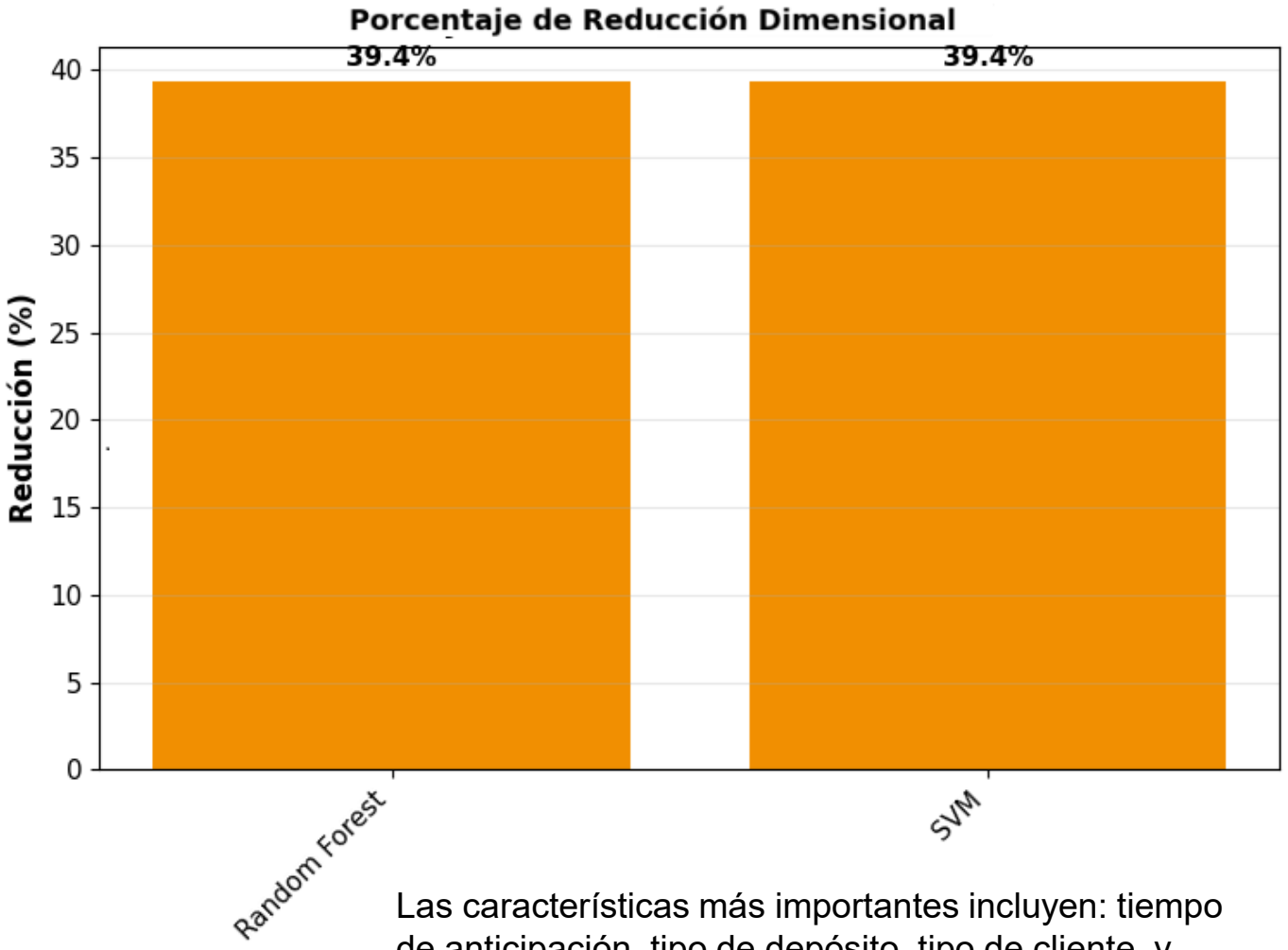
La **selección secuencial forward** redujo las características de 33 a 20 - una **reducción del 39.4%** con lo que eliminamos ruido y nos enfocamos en las características más predictivas.

ANÁLISIS DE DEGRADACIÓN DE RENDIMIENTO:
RandomForest
F1-Score original: 0.814
Selección Secuencial: 0.743 (degradación: 8.6%)

SVM
F1-Score original: 0.797
Selección Secuencial: 0.772 (degradación: 3.1%)

TÉCNICAS AVANZADAS

UNIVERSIDAD DE ANTIOQUIA



Las características más importantes incluyen: tiempo de anticipación, tipo de depósito, tipo de cliente, y tarifa diaria promedio.

Por otro lado, **PCA** con 95% de varianza explicada logró una **reducción del 39,4%** con **4.3% de pérdida** en Random Forest.

Esto es crucial para implementaciones con recursos limitados - podemos mantener casi el mismo rendimiento con menos características.

ANÁLISIS DE DEGRADACIÓN DE RENDIMIENTO:
RandomForest
F1-Score original: 0.814
PCA: 0.779 (degradación: 4.3%)

SVM
F1-Score original: 0.797
PCA: 0.781 (degradación: 2.0%)

Conclusiones

UNIVERSIDAD DE ANTIOQUIA

Modelos óptimos

Random Forest
F1-Score: 0.812 AUC-ROC: 0.931

SMV
F1-Score: 0.801 AUC-ROC: 0.924

=== JUSTIFICACIÓN DEL CRITERIO DE SELECCIÓN ===
F1-Score (media armónica de precisión y recall)
Justificación:
- El dataset es desbalanceado, por lo que accuracy puede ser engañoso.
- F1-Score balancea precisión y recall, penalizando tanto falsos positivos como falsos negativos.
- Es el criterio más robusto para problemas de clasificación desbalanceada.

🔍 SELECCIÓN SECUENCIAL:
✅ Mejor modelo: SVM (F1: 0.772)
✅ Ventajas: Mantiene características originales, más interpretable

⚠ Desventajas: Mayor degradación de rendimiento

📉 PCA:
✅ Mejor modelo: SVM (F1: 0.781)
✅ Ventajas: Menor degradación, componentes ortogonales

⚠ Desventajas: Pérdida de interpretabilidad

UNIVERSIDAD DE ANTIOQUIA

@UdeA @universidaddeantioquia @UdeA

@universidaddeantioquia @UdeA