# Proyecto de Predicción de Churn para Waze - Informe Final

## Overview del Proyecto

El objetivo principal fue desarrollar un modelo de Machine Learning capaz de predecir la probabilidad de que un usuario deje de usar la app de Waze (churn). Para esto, se trabajó con un dataset real que contiene información sobre la actividad de los usuarios, su interacción con zonas favoritas, dispositivos y patrones de navegación.

## Metodología

### 1. Preprocesamiento

* Se eliminaron registros con valores nulos en la variable objetivo label.
* Se codificó la variable objetivo: retained como 0 y churned como 1.
* Se convirtieron las variables categóricas a variables dummies (iPhone vs Android).
* Se dividieron los datos en entrenamiento (80%) y test (20%).

### 2. Modelado y Evaluación

Se probaron tres enfoques principales:

#### a) **Random Forest (básico y balanceado)**

* Precisión: hasta 83%
* Recall para churn: muy bajo inicialmente (7% mejorado con SMOTE y grid search)

#### b) **XGBoost**

* Mejor ROC-AUC inicial sin balanceo: 0.684
* Mejor recall para clase minoritaria: 23% (en usuarios recientes)

#### c) **Random Forest + SMOTE + GridSearchCV**

* ROC-AUC: 0.742
* F1 para clase churn: 0.44 (mejor resultado logrado)
* Mejores parámetros: n\_estimators=200, max\_depth=5, min\_samples\_split=5, class\_weight=balanced

### 3. Segmentación de usuarios recientes (<500 días desde onboarding)

* Se identificó que los usuarios recientes que interactúan más con su zona favorita tienen mayor probabilidad de ser correctamente identificados.
* Las variables como total\_navigations\_fav1, sessions, drives y n\_days\_after\_onboarding fueron clave.

## Key Insights

* **Interacción con “zonas favoritas” es altamente predictiva del churn.**
* **Usuarios recientes (menos de 500 días) son más susceptibles al abandono, pero también más fácilmente detectables.**
* **SMOTE mejoró la detección de la clase minoritaria (churn), elevando recall de 0.12 a 0.49.**
* **Ajustar el umbral a 0.3 también mejora la detección de churn, aunque sacrifica algo de precisión.**
* **El modelo optimizado mediante GridSearch alcanzó el mejor equilibrio entre precisión y recall para decisiones estratégicas.**

## Recomendaciones

* **Implementar modelo Random Forest optimizado en entornos de producción.**
* **Monitorear métricas en usuarios nuevos y recientes para acción temprana.**
* **Lanzar estrategias de retención focalizadas en usuarios con baja interacción en zonas favoritas.**
* **Ajustar el umbral de alerta según el contexto del negocio (mayor recall o mayor precisión).**
* **Iterar y reevaluar trimestralmente el modelo con datos nuevos.**

## Next Steps

* Documentar el pipeline de producción.
* Exportar el modelo entrenado.
* Crear dashboard de monitoreo.
* Establecer umbrales de alerta operativa.
* Validar resultados en un A/B test antes del despliegue global.

Este informe está listo para ser compartido con stakeholders técnicos y de negocio.