Universidad San Francisco de Quito

Examen Final – Minería de Datos

Parte Práctica (Proyecto en Notebook)

12 de mayo de 2025

Contexto

Eres **Data Scientist** en la empresa **Interconnect**. El equipo de marketing desea anticipar la *tasa de cancelación de suscripciones* para ofrecer códigos promocionales a quienes muestren peligro de abandono. Dispones de cuatro fuentes de datos, todas ligadas por `customerID`:

- `contract.csv` — detalles de contrato (tipo, duración, método de pago, importe, etc.);
- `personal.csv` — datos demográficos y segmentación del cliente;
- `internet.csv` — servicios de Internet y características técnicas;
- `phone.csv` — servicios telefónicos y líneas asociadas.

El objetivo consiste en **predecir si un cliente abandonará la suscripción**

Entregable único

Un **único notebook** (Python) llamado `<apellido>_<nombre>_churn.ipynb` que:

1. Implementa todos los pasos de la metodología **CRISP-DM** en el mismo orden:
   a) **Business Understanding**: formulación del problema, hipótesis y KPI.
   b) **Data Understanding**: exploración inicial, data-dictionary y verificación de supuestos.
   c) **Data Preparation**: unificación de fuentes, limpieza, imputación, ingeniería de variables (incluya ejemplos de encoding, escalado, creación de features sobre servicios).
   d) **Modeling**: al menos *tres* algoritmos supervisados (uno lineal, un ensemble y uno a elección de boosting.
   e) **Evaluation**: compare modelos con **AUC-ROC** (métrica principal) y *otra métrica* elegida.
   f) **Deployment / Próximos pasos**: plan breve (máx. 5 celdas) sobre cómo integrar el modelo, monitoreo de deriva, y estrategia de descuentos.
2. Está **comentado y ordenado**: cada fase inicia con un `#` encabezado Markdown; gráficos y tablas tienen títulos y ejes claros.
3. Usa **control de versiones Git**: adjunte `link` al repositorio (público + acceso).
4. Se ejecuta de principio a fin con `kernel restart & run all` sin errores.

Requisitos técnicos mínimos

- **Validación**: utilice `train_test_split` estratificado (70-30 %) y, dentro de entrenamiento, *cross-validation* con al menos $k = 5$.
- **Manejo de desbalance**: si la clase `Churn="Yes"` es $< 25\,\%$, aplique técnica justificada.
- **Reproducibilidad**: fije semillas aleatorias (`random_state`).
- **Interpretabilidad**: incluya SHAP, Gain, *Permutation* u otra técnica para identificar *top-10 features*.

Evaluación (30 puntos)

AUC-ROC (en test)	Puntos	Comentario
< 0.75	0	Debajo de criterio mínimo
0.75 – 0.80	10	Modelo inicial aceptable
0.81 – 0.84	15	Buen desempeño
0.85 – 0.869	20	Muy buen desempeño
0.87 – 0.879	25	Sobresaliente
≥ 0.88	30	Nivel de excelencia

Ponderación global (70 ptos)

- Metodología CRISP-DM completa y bien documentada — 30 ptos
- Rendimiento según la tabla anterior — 20 ptos
- Interpretación de resultados y recomendaciones de negocio — 10 ptos
- Organización del notebook (claridad, secciones, gráficos legibles) — 5 ptos
- Control de versiones Git y reproducibilidad — 5 ptos

Entrega

Suba a la plataforma: Enlace al repositorio Git con `README.md` breve.

Sugerencia de estructura de carpetas

```
data/            # archivos CSV originales (sin subir al repo público)
notebooks/
   <apellido>_<nombre>_churn.ipynb
src/             # scripts auxiliares (opcional)
models/          # versiones serializadas (pkl, joblib)
README.md
requirements.txt
.gitignore
```

¡Éxitos!

Demuestre su dominio de Minería de Datos aplicando prácticas industriales y justifique cada decisión técnica en función del negocio de **Interconnect**.