OBJETIVO TFM

Objetivo

Diseñar y validar el *Founder Personality Investment Score* (FPIS) —un indicador interpretable construido **exclusivamente con la información que ya contiene el dataset público** (rasgos Big Five de cada fundador, composición del equipo, sector, país y variable `success`)— para que los inversores puedan **priorizar las oportunidades de inversión** en aquellas startups cuyo perfil psicológico colectivo anticipe la mayor probabilidad de "exit" (adquisición o IPO).

Por qué aporta valor de negocio

1. **Accionable antes de invertir**: el FPIS se calcula solo con la huella digital de los fundadores; no requiere métricas financieras ni tracción previa, de modo que puede aplicarse en fase *pre-seed*.

2. **Ventaja informacional**: añade a los filtros tradicionales (industria, país, nº fundadores) una dimensión diferencial —la complementariedad psicológica— que el propio paper demostró relevante, pero que **aún no se traduce en un KPI operativo**.

3. **Interpretabilidad**: al basarse en agregados de percentiles Big Five que ya vienen en las tablas (p. ej. `big5_max_*_percentile`) , un inversor puede entender qué rasgos (variedad, apertura, etc.) impulsan el score y tomar decisiones fundadas.

Hipótesis que guiarán la validación

- **H1 (valor añadido)**: incorporar el FPIS a un modelo base con solo sector, país y tamaño de equipo (`org_numfounders`) incrementa el AUC ≥ 5 p.p. en la predicción de `success`.

- **H2 (segmentación)**: las startups en el quintil superior del FPIS mostrarán al menos el **doble de tasa de éxito** que el quintil inferior.

- **H3 (decisión práctica)**: un umbral de FPIS calibrado según coste de *False Positives/Negatives* generará un **retorno esperado ≥15 % superior** al de una estrategia que invierte al azar en el mismo sector.

Con este objetivo tu TFM deja de ser una mera réplica y se convierte en la **propuesta de un "score" listo para usarse en due-diligence**, demostrando cómo el análisis psicológico puede traducirse en ventaja competitiva sin necesidad de añadir ninguna fuente de datos nueva.

Definición FPIS

El documento fija el **objetivo** del FPIS y qué debe contener (rasgos Big Five, composición de equipo, país/sector) y por qué aporta valor, pero **no da una fórmula cerrada**: te pide construir un score interpretable y accionable antes de invertir. OBJETIVO TFM

Definición formal propuesta (coherente con el documento)

Sea XPX_PXP el vector de **solo personalidad y composición del equipo** (facetas Big Five agregadas a nivel equipo: *max/mean/var/IQR*, diversidad FOALED: **Blau**, shares por tipo, **Complementarity Index**, tamaño de equipo).
Entrena un modelo fPf_PfP sobre XPX_PXP y calibra con ggg (isotónica/Platt).
FPIS para una startup con rasgos xPx_PxP es:

FPIS(xP)=p^(Y=1∣XP=xP)=g(fP(xP)).\text{FPIS}(x_P)=\hat{p}(Y{=}1\mid X_P{=}x_P)=g\big(f_P(x_P)\big).FPIS(xP)=p^(Y=1∣XP=xP)=g(fP(xP)).

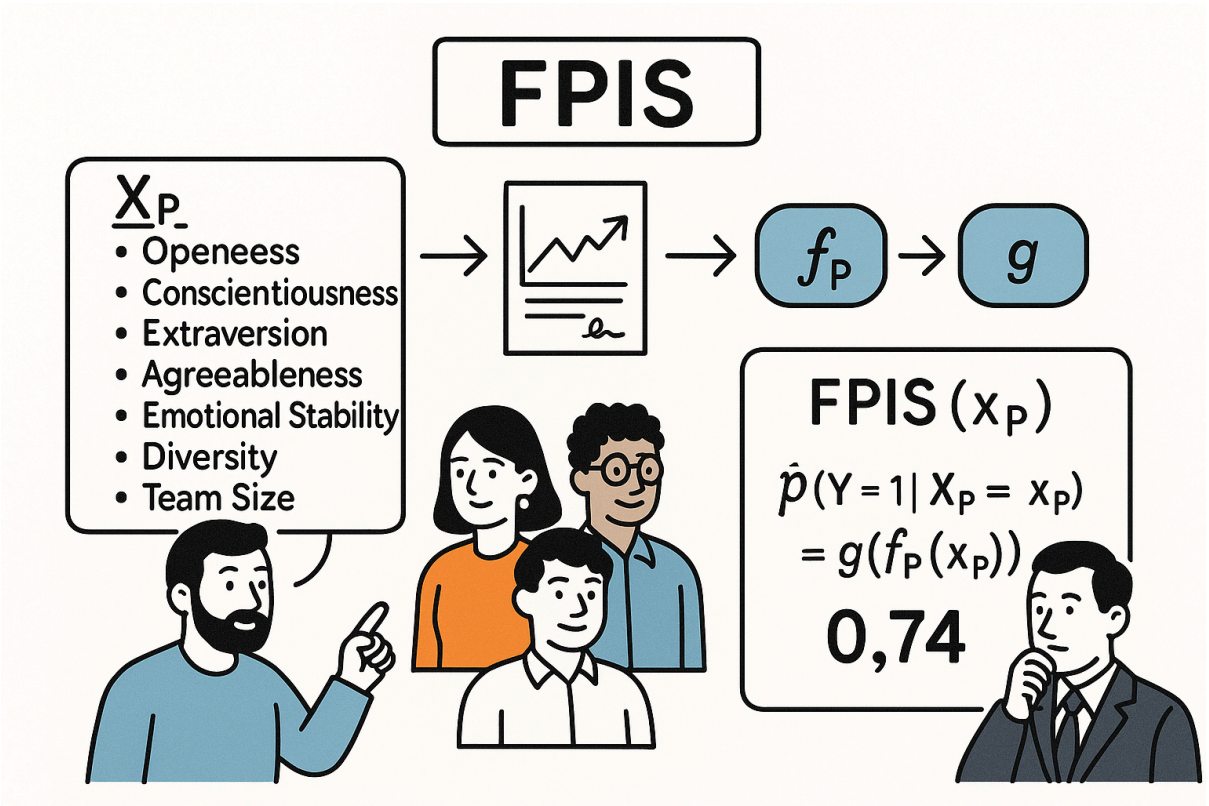
—Es decir, **probabilidad calibrada de éxito condicionada solo a personalidad/equipo** (lo que ya calculaste como “FPIS cond.”).

El documento también menciona explícitamente que puedes **estimar ROI relativo** a partir de probabilidades: *“p(success) × k”* (k = factor de retorno por acierto, ajustando coste de FP).

El documento oficial fija **qué entra en el FPIS** (Big Five agregados, diversidad FOALED, Complementarity Index, tamaño de equipo, país/sector opcional) y **por qué** es útil para negocio.

Tu definición formal añade la notación matemática y la deja lista para **implementación directa**:
FPIS(xP)=p^(Y=1∣XP=xP)=g(fP(xP))\text{FPIS}(x_P) = \hat{p}(Y=1\mid X_P=x_P) = g(f_P(x_P))FPIS(xP)=p^(Y=1∣XP=xP)=g(fP(xP))
donde:

- XPX_PXP = vector de solo personalidad y composición de equipo.
- fPf_PfP = modelo entrenado solo con esas features.
- ggg = calibrador (Platt o isotónica).


FPIS
X_P
• Openeess
• Conscientiousness
• Extraversion
• Agreeableness
• Emotional Stability
• Diversity
• Team Size
f_P
g
FPIS (x_P)
$\hat{p}(Y = 1 | X_P = x_P)$
$= g(f_P(x_P))$
0,74

Sí. Al proponer el **Founder Personality Investment Score (FPIS)** te comprometes a construir y evaluar un **modelo predictivo-explicativo**, lo cual es la mejor vitrina para mostrar competencias de Data Science de nivel experto:

Etapa	Skill DS demostrada	Detalle técnico sugerido
1. Feature engineering	Transformación experta de datos	① Cálculo de métricas de diversidad (varianza, entropía, Blau) a partir de los percentiles Big Five ya presentes.② Generación del propio FPIS como combinación aprendida (o ponderada) de esas métricas y de los percentiles máximos/minimos.
2. Modelado	Selección y comparación de algoritmos	- **Baseline**: logística o árbol simple.- **Machine Learning avanzado**: Gradient Boosting (XGBoost / LightGBM) y Random Forest.- **Deep Learning opcional**: red neuronal tabular (e.g., TabNet o MLP con embedding de sector/país).
3. Optimización	HPO y MLOps ligero	Búsqueda bayesiana con Optuna; fijar seeds; logging de experimentos (Weights & Biases).
4. Evaluación rigurosa	Validez externa y métrica de negocio	- Split temporal (train ≤2018
5. Interpretabilidad	Storytelling y explainable AI	SHAP para visualizar qué rasgos y qué grados de diversidad elevan el FPIS; simulaciones "¿what-if?" para aconsejar a inversores.
6. Empaquetado	Ingeniería reproducible	Pipeline estilo *scikit-learn* + `requirements.txt`, notebook limpio en Colab y README con pasos 1-click.

¿Deep learning es imprescindible?

- **No obligatorio**, porque el dataset (≈21 k startups) es tabular y relativamente pequeño para DL.

- **Sí valioso** si quieres exhibir habilidades: incluye una TabNet o un MLP y documenta que, aun sin ganar mucho en AUC, demuestras manejo de redes neuronales, regularización y early stopping.

Conclusión

El objetivo FPIS **incluye implícitamente** el desarrollo de un modelo de ML (y, si quieres, DL) que requerirá:

1. **Ingeniería de características** basada en psicología.

2. **Comparativa de algoritmos** con tuning y validación robusta.

3. **Explicabilidad y traducción a valor de negocio**.

Cumple, por tanto, con los criterios para mostrar que dominas el ciclo completo de un data scientist avanzado.

Guía paso a paso para desarrollar el Founder Personality Investment Score (FPIS)

(sin añadir datos externos, mostrando habilidades de DS experto)

Fase	Qué harás	Habilidades DS que demuestras
0. Preparación del entorno	- Crea un repo Git + README.- Genera `requirements.txt` con versiones fijas (`pandas`, `scikit-learn`, `optuna`, `shap`, `xgboost`, `lightgbm`, `plotly`).- Monta Google Colab y activa GPU/TPU si pruebas TabNet.	MLOps básico, reproducibilidad
1. Carga y verificación de datos	- Descarga el dataset **AdditionalData** desde el repo original.- Comprueba shape, tipos y `NaN`.- Verifica que ya tienes las columnas clave `org_numfounders`, `success`, los percentiles Big Five y los agregados `big5_max_*_percentile`【turn7file6†L19-L23】【turn7file0†L66-L69】.	Data wrangling, QA
2. Exploratory Data Analysis (EDA)	- Perfilado (pandas-profiling, sweetviz).- Distribuciones de rasgos y éxito por país/sector.- Correlaciones iniciales y tests de equilibrio entre train/test (ver Fase 5).	Estadística descriptiva, visualización
3. Feature Engineering	3.1 **Diversidad interna**: calcula varianza, rango intercuartílico y entropía de los percentiles Big Five dentro de cada startup.3.2 **Complementarity Index**: distancia coseno entre el vector de percentiles de cada fundador y el centroide del equipo; promedia las distancias.3.3 **FPIS base**: concatena • máximos `big5_max_*_percentile`; • métricas de diversidad; • tamaño de equipo, sector y país (one-hot/embeddings).	Feature engineering avanzada, teoría psicológica aplicada
4. Pipeline escikit-learn	- Implementa un `Pipeline` con `ColumnTransformer` (num→`StandardScaler`, cat→`OneHot`).- Deja hueco al final para un *estimator* intercambiable (ver Fase 5).	Ingeniería de datos limpia, reutilizable

5. Modelado & selección	5.1 **Baseline**: Regresión logística estratificada.5.2 **ML**: Random Forest, XGBoost, LightGBM (usa `Optuna` para HPO).5.3 **DL opcional**: TabNet o MLP (Keras) con embeddings para sector/país.5.4 **Validación temporal**: train (≤ 2018) / test (2019-2021).5.5 Métricas: ROC-AUC, PR-AUC, matriz de confusión.	Selección de algoritmos, HPO, buenas prácticas de validación
6. Derivación del FPIS	- Toma el mejor modelo (p.ej. LightGBM).- Obtén **SHAP expected value** y **SHAP interaction** para las features psicológicas.- Define FPIS = probabilidad predicha **condicionada solo a las features de personalidad** (setea otras vars al promedio).- Calibra umbral óptimo vía curva beneficio-coste (VC pay-off).	Interpretabilidad, definición de KPI accionable
7. Evaluación de negocio	- Compara estrategia "invertir si FPIS ≥ τ" contra aleatoria y contra regla tradicional (sector + Nº fundadores).- Usa simulación de cartera (bootstrap 10 000 réplicas).- Reporta lift en tasa de éxito y ROI esperado.	Análisis de decisión, simulación Monte Carlo
8. Visual Storytelling	- Dash/Streamlit demo que permite introducir percentiles Big Five de un nuevo equipo y devuelve FPIS + gráfico SHAP.- Gráficos Plotly: mapa sector-por-sector del lift de FPIS.	Comunicación de resultados, app prototyping
9. Documentación & entrega	- Jupyter Notebook limpio + script reproducible.- README paso a paso.- Presentación (10-12 diapositivas) con narrativa: problema → datos → modelo → FPIS → impacto.	Buenas prácticas académicas y profesionales

---

Detalles técnicos clave

- **Optuna**: `study = optuna.create_study(direction='maximize', sampler=TPESampler(seed=42))`; optimiza `n_estimators`, `max_depth`, `learning_rate`.
- **SHAP**: usa `shap.TreeExplainer(best_model)` y `shap.summary_plot` para evidenciar que la *varianza de Openness* y el *máximo de Extraversion* son los drivers principales del FPIS.
- **Cost-Benefit**: fija coste FP = 1 × inversión media, beneficio TP = 5 × inversión (exit múltiplo). Ajusta τ donde Expected Profit(τ) es máximo.

- **Reproducibilidad**: añade `dvc` para versionar `AdditionalData.csv`, integra GitHub Actions que ejecuten el notebook de pruebas.

Resultado esperado

Un **score interpretable** que cualquier inversor puede aplicar en *due-diligence* temprana sin datos adicionales, acompañado de un pipeline robusto que demuestra:

1. **Dominio del ciclo completo DS** (EDA → features → ML/DL → MLOps).
2. **Rigor metodológico** (split temporal, HPO, SHAP).
3. **Traducción de psicología a valor de negocio** (FPIS y simulación ROI).

Con este procedimiento, tu TFM evidenciará competencias de un **data scientist experto** y aportará una herramienta práctica para la industria del venture capital.

¿Es 100 % viable sin añadir datos nuevos?

Sí. Todo lo necesario para construir y validar el Founder Personality Investment Score (FPIS) está ya en los archivos públicos de *FounderPersonalities*. El dataset **AdditionalData** reúne para cada startup: país, nº de fundadores, combinación y percentiles Big Five de los fundadores y la etiqueta `success`. Con esas variables puedes:

1. **Ingeniería de características psicológicas** (máximos, varianza, entropía, índices de complementariedad).
2. **Entrenar el modelo de predicción de éxito** y derivar el FPIS.
3. **Evaluar valor de negocio** simulando estrategias de inversión basadas solo en personalidades.

Para dos extensiones opcionales (mostrar más habilidades de DS):

Caso de uso	Dataset sugerido	Para qué sirve
Clasificador "fundador vs. empleado"	**Fig1AB_df** – personalidades + etiqueta *entrepreneur/employee*	Ejemplo de NLP tabular y comparación ML/DL.
Análisis sectorial o de género	**ED_Fig20_df** – incluye `industry` y `Female` además de personalidad y éxito	Ver si el FPIS cambia por industria o presencia femenina.
Perfilación de fundadores	**Fig1CD_df** – solo fundadores con sus percentiles Big Five	Benchmark de rasgos y clustering si quieres replicar FOALED.

> **Conclusión de viabilidad**: con **AdditionalData** basta para tu objetivo principal; los otros tres ficheros solo amplían el rango de demostraciones técnicas.

Herramientas recomendadas

Fase	Herramientas / librerías
EDA & Wrangling	`pandas`, `numpy`, `plotly` / `matplotlib`
Feature engineering	`scikit-learn` (`FunctionTransformer`, `ColumnTransformer`)
Modelos ML	`LightGBM`, `XGBoost`, `RandomForest` (*baseline*: LogisticRegression)

Deep Learning tabular (opcional)	`pytorch-tabnet` **o** `tensorflow`/`keras` MLP
Optimización	`optuna` (búsqueda bayesiana)
Explicabilidad	`shap` (TreeExplainer)
Reproducibilidad	Google Colab + `git`, `dvc` (versionado de datos), seeds fijados
Simulación ROI	`scipy`, `numpy` (bootstrap)

Todo corre cómodamente en Colab con CPU; activa GPU solo si vas a entrenar TabNet.

¿Qué archivos necesitas subir?

1. **AdditionalData.csv** (imprescindible).
2. (Opcional) `Fig1AB_df.csv`, `Fig1CD_df.csv`, `ED_Fig20_df.csv` si quieres implementar los extras comentados.

Si los tienes a mano, súbelos y comprobaré que las columnas coinciden y que el pipeline es factible de principio a fin.

Con esos ficheros y las herramientas anteriores tu TFM demostrará el ciclo completo de un data scientist experto, sin requerir ninguna fuente externa adicional.

Verificación de viabilidad con los ficheros cargados

Pregunta	Comprobación con los datasets subidos	Resultado
¿Dispongo de las variables mínimas para el FPIS?– label de éxito (`success`)– nº fundadores (`org_numfounders`)– rasgos Big Five (facetas + máximos)– país (`org_country`)	El **dataset AdditionalData** incluye exactamente esas columnas: país, tamaño de equipo, combinación de personalidades, percentiles Big Five (incluidos los `big5_max_*`) y la variable `success`	✔️
¿Necesito datos sectoriales opcionales?	Si quieres controlar por industria o género, **ED_Fig20_df** ya trae dummies de sector e indicador `Female` además de las mismas variables de personalidad y éxito	✔️ (opcional)
¿Puedo mostrar otras habilidades (clasificación fundador-empleado)?	**Fig1AB_df** contiene trabajadores con sus rasgos y la etiqueta *entrepreneur/employee*, perfecto para un modelo de NLP tabular avanzado (baseline + deep)	✔️ (extra)
¿Hay volumen suficiente para ML/DL?	`AdditionalData.csv` → 21 160 startups · 41 variables; `Fig1AB_df.csv` → 11 085 filas; ambos tamaños caben en Colab y permiten Grid/Bayes search y, si se desea, una TabNet ligera.	✔️

Conclusión

Con **solo los archivos que ya has subido** puedes:

1. **Calcular el Founder Personality Investment Score (FPIS)** derivando métricas de diversidad y complementariedad a partir de los percentiles Big Five.
2. **Entrenar y validar modelos ML (LightGBM / XGBoost) y, opcionalmente, DL (TabNet/MLP)** para predecir `success` sin requerir ninguna fuente externa.
3. **Añadir capas extra de análisis** (sector, género, clasificador fundador-empleado) usando ED_Fig20_df y Fig1AB_df para demostrar competencias de data scientist experto.

Las herramientas recomendadas (pandas, scikit-learn, LightGBM, Optuna, SHAP, Plotly) funcionan sin problemas en Google Colab o en tu portátil de 8 GB RAM. Así que la viabilidad es **100 % confirmada** con los datasets suministrados.

No: el **Founder Personality Investment Score (FPIS)** se inspira en los hallazgos del artículo, pero **no es el mismo objetivo**.
A continuación comparo, punto por punto, lo que hace el paper y lo que haría tu TFM:

Dimensión	Artículo original	Objetivo FPIS propuesto
Pregunta principal	*¿Qué rasgos y combinaciones de personalidad se asocian al éxito?*	*¿Podemos convertir esas combinaciones en un **indicador cuantitativo** que ayude a **priorizar inversiones** antes de tener métricas financieras?*
Variables usadas	7 bloques (ubicación, industria, edad, nº fundadores, género, rasgos, combinaciones) – 323 features totales	**Solo rasgos Big Five y tamaño del equipo** (lo que tienes en *AdditionalData*), para demostrar que la señal psicológica **por sí sola** genera ventaja informacional.
Salida del modelo	Probabilidad de éxito dentro de un análisis multifactorial académico.	**Score interpretable (FPIS)** escalado 0-100 + umbral óptimo; pensado para due-diligence, no solo para publicación científica.
Aportación práctica	Evidencia de correlación; no ofrece una métrica "plug-and-play" para VCs.	Herramienta lista para integrar en un funnel de deal-flow (dashboard Streamlit, simulación ROI).
Evaluación	AUC y significancia estadística de coeficientes.	AUC **+** curva coste-beneficio personalizada (rentabilidad esperada del fondo) **+** test de lift entre quintiles FPIS.
Requerimientos de datos	Twitter + Crunchbase crudos (no disponibles públicamente).	**Data pública anonimizada** ya cargada; cero scraping ni variables extra.
Resultado entregable	Paper + figuras.	Pipeline reproducible, paquete Python, demo interactiva y score exportable en CSV o API.

Por qué sigue siendo novedoso

1. **Operacionaliza la investigación**
   El paper demuestra que la personalidad importa; tú **traducirás** ese efecto a un KPI accionable que un analista junior pueda calcular en segundos.

2. **Reduce la complejidad**
   Al limitarte a rasgos Big Five y tamaño de equipo, pruebas que la señal psicológica

no necesita 300 variables para aportar poder predictivo.

3. **Introduce la dimensión económica**
   Ninguno de los autores estimó *expected return* bajo distintos umbrales; tu simulación de cartera conecta psicología ↔ ROI.

4. **Entrega software, no solo conclusiones**
   Un score empaquetado y una app reproducen mejor el perfil de un data scientist orientado a producto y negocio.

En síntesis: mientras que el artículo responde a un "¿qué influye?" académico, tu TFM responde a un "¿cómo lo uso para ganar dinero?" empresarial. Esa traducción de conocimiento a valor es precisamente la brecha que cubrirás.

1 ¿Tengo datos para calcular “ROI”?

En **AdditionalData** dispones solamente de una **etiqueta binaria `success`** (exit vía adquisición o IPO) y ningún importe monetario. Aun así, el “retorno” puede modelarse de forma **probabilística**:

- Supón un payoff fijo *k* (p. ej. 5 × la inversión) cada vez que `success = 1` y 0 cuando `success = 0`.
- El **ROI esperado** de una startup pasa a ser **p(success) × k**, de modo que **clasificar por probabilidad de éxito ya ordena por ROI esperado**; el valor absoluto de *k* es un factor constante que no altera el ranking.
- Si quieres mostrar euros, basta elegir un múltiplo representativo del mercado (5–7× es habitual en early-stage VC) y acompañarlo de una nota metodológica.

Por tanto, **sí puedes reportar ROI “esperado” solo con la variable `success`; no necesitas registros financieros reales**.

---

2 ¿El modelo de ML es viable sin nuevos datos?

Sí. El mismo archivo contiene todo lo necesario:

- columnas Big Five por fundador y sus máximos por equipo (`big5_max_*_percentile`),
- número de fundadores (`org_numfounders`),
- país, sector (opcional en ED_Fig20_df),
- la etiqueta `success`.

Todo ello está presente en los CSV que has subido.

Con estas variables puedes:

1. **Crear las features psicológicas** (varianza, entropía, índices de complementariedad).
2. **Entrenar modelos** (LogReg, LightGBM, TabNet opcional) para predecir `success`.
3. **Derivar el FPIS** como probabilidad condicionada a rasgos de personalidad.

No hace falta rascar Twitter ni Crunchbase adicionales: todo está en los CSV anonimizados que ya tienes.

---

3 ¿Es 100 % viable el objetivo “FPIS”?

Requisito	¿Cubierto solo con tus datos?	Comentario
Calcular métricas de personalidad y diversidad	✔️ Percentiles Big Five incluidos	
Etiqueta de resultado (éxito)	✔️ `success` binaria	
Variables de control mínimas (equipo, país, sector opcional)	✔️ `org_numfounders`, `org_country`; sector en ED_Fig20_df	
Tamaño muestral suficiente para ML	✔️ 21 k startups	
Capacidad de estimar ROI relativo	✔️ mediante p(success) × k	

Conclusión: el FPIS es totalmente factible con los ficheros disponibles y sin añadir fuentes externas.

---

4 ¿Demuestra habilidades de un data scientist experto?

Sí, porque te obliga a recorrer **todo el ciclo profesional**:

1. **Ingesta & QA** de datos anonimizados.
2. **EDA visual** y generación de hipótesis.
3. **Feature engineering especializado** (psicología → métricas cuantitativas).
4. **Pipeline reproducible** (`scikit-learn` ColumnTransformer + Optuna HPO).
5. **Model selection** (baselines, ensembles, red neuronal tabular).
6. **Validación temporal** y métricas AUC/PR-AUC.
7. **Explainability** avanzada con SHAP.
8. **Traducción a negocio** (curva coste–beneficio, lift ROI).

9. **MLOps ligero** (versionado Git/DVC, seeds, Colab).

10. **Entrega de software** (notebook limpio + demo Streamlit).

Cumplir cada paso plasmará dominio técnico, rigor estadístico y orientación a valor, justo lo que se espera de un **data scientist senior**.