Memoria del Proyecto — Predicción de Rotación de Empleados

Autora: Beatriz Velayos · **Bootcamp Big Data · AI · ML** · **Tag:** v1.0-entrega

1. Resumen

Pipeline reproducible (Docker) para predecir rotación (attrition) y explicar *drivers*. Incluye ingesta/ETL con PySpark, modelado (scikit-learn), persistencia de métricas y dashboard en Power BI. Notebooks exportados a HTML y esta memoria en PDF.

2. Objetivo

- Estimar riesgo de baja por empleado.
- Explicar factores clave (drivers) y KPIs para RR. HH.

3. Datos

- WA HR Attrition (1.470 filas), objetivo: `Attrition` (Yes/No).
- Encuesta de clima (Engagement, Satisfaction, WorkLifeBalanceSurvey, ManagerRelationship, RemoteWorkSatisfaction) unida por `EmployeeNumber`.

Filas finales: 1.470 | **Attrition positivo:** 237 | **Tasa:** 0,1612

4. Metodología

EDA: nulos, duplicados, dominios; nuevas features (`overtime_flag`, `income_yearly`, `tenure_ratio`).
ETL (PySpark): limpieza + join → Parquet.
Modelos: Regresión Logística (pipeline con escalado + OneHot) y Random Forest. Métricas: ROC-AUC, PR-AUC y F1 con umbral óptimo.

5. Resultados

- **LogReg** — AUC 0.8173 · PR-AUC 0.5627 · F1* 0.5286 · thr* 0.7324
- **RF** — AUC 0.8051 · PR-AUC 0.5364 · F1* 0.5348 · thr* 0.22

La LogReg ofrece mejor ROC-AUC y explicabilidad directa (coeficientes). Drivers y KPIs se muestran en el dashboard.

6. Dashboard (Power BI)

- **Resumen & KPIs**: métricas principales y guía rápida.
- **Segmentos & Drivers**: Top-N coeficientes (LogReg) con signo.
- **Departamentos / Job Roles**: comparativas.

7. Reproducibilidad

```
docker compose up -d
docker compose run --rm jupyter python /scripts/preprocess.py \
  --input /data/processed/employee_attrition.parquet \
  --out /output/models/transformer.joblib
docker compose run --rm jupyter python /scripts/train_ml.py \
  --input /data/processed/employee_attrition.parquet --model logreg
docker compose run --rm jupyter python /scripts/train_ml.py \
  --input /data/processed/employee_attrition.parquet --model rf
```

8. Limitaciones y futuro

- Datos de un único corte; faltan variables temporales/comportamiento.
- Ajuste de umbral por coste de error y monitorización en producción.

9. Conclusión

Se cumple el objetivo de negocio con un sistema reproducible y un dashboard interpretable (AUC≈0,82) para priorizar acciones de retención.