# Proyecto Final Data Analitycs

# Antonio Martínez Urbán

Resumen Ejecutivo

El presente proyecto aborda la problemática de la **gestión y análisis ineficiente de los datos de leads y cursos** en una empresa de formación. Anteriormente, la información crítica sobre el comportamiento de los clientes potenciales y el rendimiento de los programas de formación se encontraba dispersa, inconsistente y no era fácilmente accesible para la toma de decisiones estratégicas.

La solución implementada consiste en un **pipeline de Extracción, Transformación y Carga (ETL)** , desarrollado en Python utilizando la librería Pandas, para limpiar, normalizar y estructurar los datos. Posteriormente, se realizó un **Análisis Exploratorio de Datos (EDA)** detallado para extraer “insights” clave. El resultado final es un **Dashboard interactivo en Tableau**, que centraliza los Indicadores Clave de Rendimiento (KPIs) más significativos, ofreciendo una visión integral del embudo de ventas, la rentabilidad de los cursos y la efectividad de las fuentes de leads.

**Resultados clave:** Se ha logrado unificar los datos de leads y cursos, identificando las **fuentes de lead más rentables y con mayor tasa de conversión**, optimizando la visibilidad del **ciclo de vida del lead** para detectar **puntos de abandono**, revelando los **cursos más demandados y con mayor impacto económico**. Esta solución permite a la dirección de marketing y ventas tomar decisiones basadas en datos para maximizar la conversión y los ingresos, así como tomar las decisiones correctas para mejorar la calidad del servicio.

Descripción del Caso de Negocio

La empresa de formación genera una considerable cantidad de datos relacionados con sus interacciones con clientes potenciales (leads) y la oferta de sus programas educativos (cursos). Sin embargo, la gestión de esta información presentaba desafíos significativos:

**Fragmentación de Datos:** La información de leads, contactos, seguimientos y cursos residía en diferentes hojas de cálculo impidiendo una visión holística.

**Calidad de Datos Variable:** Existían inconsistencias, duplicados, valores nulos o formatos incorrectos que comprometían la fiabilidad de cualquier análisis.

**Falta de Insights Accionables:** Sin un proceso estructurado, era difícil responder a preguntas críticas como: ¿Cuál es la fuente de leads más efectiva? ¿Dónde se pierden los leads en el embudo? ¿Qué cursos generan más ingresos reales? ¿Cuánto tiempo toma convertir un lead?

**Toma de Decisiones Intuitiva:** Las estrategias de marketing y ventas se basaban, en parte, en la experiencia y la intuición, lo que limitaba la capacidad de optimización y escalabilidad.

Este proyecto se justifica por la necesidad de pasar de una gestión de datos reactiva a una **analítica proactiva y predictiva**. El objetivo era transformar el dato en un activo estratégico que soporte la optimización de procesos, la asignación eficiente de recursos y el crecimiento de los ingresos.

**Pipeline ETL (Extracción, Transformación, Carga)**

El pipeline ETL fue diseñado para procesar los datos desde su origen hasta un formato limpio y estructurado, listo para el análisis y la visualización.

**Extracción**

En un principio los datos iban a ser extraídos con la API de Google Cloud para poder manejar la fuente de origen de los datos de Google Sheets pero tras varias extracciones inconsistentes por perdida de valores, se decide extraer de las fuentes raw, representadas por diferentes archivos CSV (simulando extracciones de bases de datos o sistemas operativos). Estas tablas contenían información sobre:

Leads: Detalles demográficos, de origen y de interés inicial de los leads.

Cursos: Atributos de cada curso ofrecido (nombre, precio, duración, etc.).

Contactos: Registro del primer contacto de un lead (fecha, medio, etc.).

Seguimientos: Historial de interacciones y el estado final del lead, incluyendo la conversión.

**Transformación (Implementado en Transformacion.ipynb con Python/Pandas)**

Esta fue la fase más intensiva, enfocada en la limpieza, normalización y enriquecimiento de los datos:

**Normalización de Nombres de Columnas:** Todos los nombres de columnas fueron convertidos a minúsculas y se reemplazaron espacios por guiones bajos para facilitar su manejo.

**Tratamiento de Datos Nulos:**

**Numéricos:** Se implementaron estrategias de imputación (ej., media o mediana) o eliminación según la columna y el impacto en la integridad del dato.

**Categóricos:** Se rellenaron con "Desconocido" o se manejaron según el contexto.

**Conversión de Tipos de Datos:**

**Fechas:** Columnas como fecha\_registro, fecha\_contacto, fecha\_matricula fueron convertidas al tipo de dato datetime.

**Numéricos:** edad, presupuesto\_estimado, precio\_base, precio\_descuento, precio\_final, lead\_score, rating, empleabilidad fueron convertidos a tipos numéricos apropiados (Int64, Decimal).

**Booleanos:** La columna conversión (en seguimientos) se aseguró como tipo booleano (True/False). **Esta columna en concreto ha sido tremendamente relevante para poder analizar el resto de los datos**.

**Limpieza de Cadenas de Texto:**

Se aplicó strip() para eliminar espacios en blanco al inicio/final de todas las cadenas de texto.

Se normalizó el formato de campos como nombre, apellidos, ciudad, provincia a formato título (title()).

Email y fuente fueron convertidos a minúsculas para consistencia.

**Validaciones Específicas de Negocio:**

**Teléfonos:** Se eliminó el sufijo ".0" y se validó la longitud máxima de 9 dígitos.

**Edad:** Se establecieron rangos válidos (ej., 18-99 años) para eliminar entradas erróneas.

**Presupuesto:** Se adaptó el formato de coma decimal (1.234,56 a 1234.56) para asegurar la correcta interpretación numérica.

**Urgencia Formación y Lead Score:** Se limitaron a sus escalas esperadas (1-5 y 0-100, respectivamente).

Se eliminaron registros duplicados en la tabla leads basándose en el email para asegurar la unicidad de cada cliente potencial.

**Enriquecimiento de Datos:**

**conversion:** Columna booleana creada en seguimientos indicando True si el lead se matriculó, False en caso contrario.

**precio\_final:** Columna calculada en seguimientos representando el coste real del curso para el lead, considerando posibles descuentos.

**Carga**

Los datos transformados fueron exportados a un directorio data\_clean en formato CSV, con un separador de ;, para facilitar su importación en Tableau. Los archivos generados fueron: leads\_final.csv, cursos\_final.csv, seguimiento\_final.csv, primer\_contacto\_final.csv, y tecnologias\_final.csv.

**Análisis de los Datos (EDA y Hallazgos Clave)**

El Análisis Exploratorio de Datos (EDA), realizado con Python y las librerías Pandas, Matplotlib, Seaborn y Plotly, permitió una comprensión profunda de los patrones y relaciones dentro de los datos limpios.

**Análisis Descriptivo y de Distribuciones**

Se observó la distribución de leads por fuente, con "Redes Sociales" y "Referidos" como los principales canales de adquisición.

La mayoría de los leads se concentran en rangos de edad de 25-44 años, indicando un público objetivo joven profesional.

La distribución de lead\_score mostró una concentración en valores medios, con pocos leads de muy alta o muy baja puntuación.

**Hallazgos Clave y Relaciones**

**Rendimiento de Fuentes de Leads:**

Se identificó que **"Webinars" y "Eventos"** presentan las tasas de conversión más altas, aunque con un menor volumen de leads. Esto sugiere una mayor calidad de los leads provenientes de estos canales.

"Redes Sociales" y "Publicidad Online", a pesar de generar un gran volumen, tienen tasas de conversión moderadas, lo que indica la necesidad de optimizar la cualificación de leads en estas fuentes.

**Eficiencia del Ciclo de Vida del Lead:**

El embudo de conversión reveló un **"cuello de botella" significativo entre la fase de "Primer Contacto" y "Seguimiento"**, con una notable caída de leads en esta transición. Esto sugiere que las estrategias o la calidad de las interacciones iniciales podrían ser un área clave de mejora.

El **tiempo promedio hasta la matrícula para leads convertidos** fue de 18 días, mientras que la **duración promedio del proceso general de seguimiento** fue de 11 días. Esto destaca la diferencia entre la duración del ciclo de ventas exitoso y el promedio global del proceso.

**Rentabilidad y Popularidad de Cursos:**

El curso de **Data Science** fue el que generó los mayores ingresos totales, siendo de los que tuvo más matrículas, gracias a su precio final más bajo.

**Demografía y Comportamiento:**

Las ciudades con mayor número de leads y conversiones fueron Barcelona, Alicante y Málaga lo que podría indicar mercados prioritarios para futuras campañas.

Se observó una correlación positiva entre un lead score alto, una urgencia formacion elevada y una mayor probabilidad de conversión.

**Impacto de Negocio y Recomendaciones**

El dashboard implementado en Tableau proporciona visibilidad sobre el rendimiento del proceso de adquisición y conversión de leads. El valor de negocio generado incluye:

**Toma de Decisiones Basada en Datos:** Permite a la gerencia de marketing y ventas fundamentar sus decisiones con métricas precisas, en lugar de la intuición.

**Optimización de la Inversión en Marketing:** Al conocer la tasa de conversión y el ingreso por lead por fuente, se pueden reasignar presupuestos a los canales más rentables.

**Mejora de la Eficiencia del Proceso de Ventas:** La identificación de cuellos de botella en el embudo permite acciones correctivas específicas en las etapas críticas del ciclo de vida del lead.

**Estrategia de Producto y Precios:** El análisis de la rentabilidad y popularidad de los cursos ofrece información valiosa para la mejora continua de la oferta de formación y las estrategias de precios.

**Cualificación de Leads Mejorada:** La correlación entre lead score, urgencia formación y conversión puede utilizarse para priorizar leads de mayor potencial.

**Siguientes pasos para el negocio (Next steps):**

**Investigar el Punto de Abandono "Primer Contacto a Seguimiento":** Analizar las grabaciones de llamadas, scripts de contacto o feedback de comerciales para entender por qué se pierden tantos leads en esta transición.

**Duplicar Estrategias de Adquisición de "Webinars" y "Eventos":** Dado su alta tasa de conversión, se recomienda aumentar la inversión y la frecuencia de estos canales.

**Refinar la Cualificación de Leads en "Redes Sociales" y "Publicidad Online":** Implementar formularios más detallados o llamadas de cualificación iniciales para reducir el volumen de leads de baja calidad que llegan a estas fuentes.

**Desarrollar Programas de Éxito para Cursos Rentables:** Asegurar que los cursos con buena demanda como Data Analytics mantengan su atractivo y empleabilidad.

**Utilizar el Lead Score en la Asignación de Comerciales:** Priorizar los leads con mayor lead score a los comerciales con mejor rendimiento en conversión aunque esto puede dar lugar a falta de motivación por parte del resto de comerciales que menos convierten.

**Conclusiones y próximos pasos para el Proyecto.**

En este proyecto se ha establecido un marco robusto para el análisis del rendimiento de leads y cursos, que proporciona valor a la empresa.

Como **próximos pasos y extensiones del proyecto**, se plantean las siguientes iniciativas:

**Automatización Completa del Pipeline ETL:** Utilización de sistema para la extracción de los datos vía API para Google Sheets correctamente y su procesamiento con funciones complejas automatizando el proceso.

**Integración de Datos de Costos:** Incorporar datos de inversión en marketing y publicidad por fuente y campaña para calcular el Costo de Adquisición de Cliente (CAC) y el Retorno de la Inversión (ROI).

**Análisis Predictivo de Conversión:** Desarrollar modelos de Machine Learning para predecir la probabilidad de conversión de un lead, permitiendo una priorización más inteligente de los esfuerzos de venta.

**Análisis de Atribución de Marketing:** Investigar cómo los diferentes puntos de contacto (fuentes, campañas) contribuyen a la conversión a lo largo del recorrido del cliente.

**Dashboard de retención y satisfacción:** Si se dispone de datos post-matrícula, extender el análisis para incluir la satisfacción del estudiante (valoración experiencia) y la retención.

**Refinamiento del Lead Score:** Evaluar si el lead score actual es óptimo o si se pueden incorporar más variables para una puntuación más precisa.

**Mejora del dashboard interactivo:** Para un mejor análisis de los resultados y una mayor claridad en los datos.