UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA



E.P. INGENIERÍA DE SISTEMAS

Título del proyecto:

Trayectorias escolares – Atraso educativo en Perú (2021–2024)

Curso:

Minería de datos

Docente:

Abel Angel Sullon Macalupu

Estudiantes:

Alvaro Jesús Maguiña Chilet

Erick Matias Mendoza Ojeda

Moises Daniel Morales Guillen

Lima – Perú

2025

**ÍNDICE**

[I. PERFIL DEL PROYECTO 3](#_heading=h.fmq5behunmve)

[Título: 3](#_heading=h.pp02l6xpm8oq)

[Problema: 3](#_heading=h.gflaw8kozqc8)

[Propósito: 3](#_heading=h.b2tea65os533)

[Justificación: 3](#_heading=h.adwddyijuz8h)

[Objetivos: 3](#_heading=h.fuv2vyecxpq5)

[Alcance: 3](#_heading=h.67aofng3ffrb)

[II. Aplicación de la Metodología CRISP-DM 3](#_heading=h.r8gjkyx2mti6)

[Comprensión del negocio 3](#_heading=h.qtm76mozw1tb)

[Comprensión y preparación de los datos 3](#_heading=h.vycqptdl2duu)

[Planificación para modelado 3](#_heading=h.kum0chsancoh)

[III. Dataset final entregado 4](#_heading=h.gy7xmr1vrab1)

[Conjunto de datos limpio, transformado y documentado 4](#_heading=h.ahzm23rwskxu)

[Descripción de variables y tratamiento aplicado (limpieza, codificación, reducción dimensional, etc.) 4](#_heading=h.pat296gmjf68)

[IV. Referencias Bibliográficas 4](#_heading=h.uldfofrb1rih)

# 

# I. PERFIL DEL PROYECTO

## Título:

Trayectorias escolares – Atraso educativo (Perú, 2021–2024)

## Problema:

El sistema educativo peruano enfrenta altos niveles de atraso escolar en primaria y secundaria. El atraso compromete la calidad del aprendizaje, incrementa el riesgo de deserción y dificulta la inclusión educativa.

## Propósito:

El propósito del proyecto es analizar y comprender los patrones de atraso escolar en el periodo 2021–2024, con el fin de predecir la probabilidad de que un estudiante presente atraso al cierre del año escolar. A partir de esta predicción, se busca generar evidencia accionable que permita orientar la toma de decisiones educativas, priorizando intervenciones que contribuyan a mejorar la gestión, la inclusión y el acompañamiento de los estudiantes en riesgo.

## Justificación:

Radica en que el análisis busca apoyar a las instituciones educativas y a los organismos gestores (UGEL y DRE) en la identificación temprana de estudiantes en riesgo, con el propósito de orientar intervenciones oportunas y efectivas. Estas intervenciones incluyen la regularización documental, la implementación de apoyos pedagógicos y el acompañamiento socioemocional, lo que permitirá mejorar las condiciones de permanencia y el desempeño escolar de los estudiantes, contribuyendo a reducir las brechas de atraso educativo.

## Objetivos:

Construir un sistema analítico que describa y explique los patrones de atraso escolar 2021–2024 y prediga el riesgo de que un/a estudiante presente ATRASO = 1 (vs. SIN\_ATRASO = 0) al cierre del año, utilizando variables disponibles en SIAGIE (Sistema de Información de Apoyo a la Gestión de la Institución Educativa) (edad, sexo, gestión de la IE, discapacidad, validación de documentos, nivel/grado, UBIGEO). El propósito es generar evidencia accionable para priorizar intervenciones (regularización documental, apoyos y acompañamiento) y mejorar la gestión educativa e inclusión.

## Alcance:

El alcance del proyecto comprende el periodo 2021–2024, utilizando una base de datos de 2.2 millones de registros de matrícula provenientes del SIAGIE. El análisis considera variables demográficas, educativas e institucionales para describir y predecir el atraso escolar. Sin embargo, presenta como limitación la ausencia de variables socioeconómicas externas, lo que restringe parcialmente la explicación de factores contextuales que también pueden influir en el rendimiento y la trayectoria escolar de los estudiantes.

## Administración:

| Semana | Fase / Actividades | Producto esperado |
| --- | --- | --- |
| Semana 1 | Planificación y Perfil del Proyecto:  - Definición del problema y objetivos.  - Establecimiento del alcance.  - Justificación del caso de negocio.  - Identificación de fuentes de datos (SIAGIE 2021–2024). | Documento inicial con el perfil del proyecto |
| Semana 2 | Comprensión y Preparación de Datos (ETL/ELT – Parte 1):  - Recolección y consolidación de los datos.  - Limpieza inicial (duplicados, outliers, valores faltantes).  - Selección de variables clave. | Dataset preliminar depurado |
| Semana 3 | Transformación de Datos (ETL/ELT – Parte 2):  - Codificación de variables categóricas (One-Hot Encoding).  - Escalado de variables numéricas (Edad).  - Reducción de redundancias. | Dataset limpio y transformado |
| Semana 4 | Planificación de Modelado y Entregables Finales:  - Definición de algoritmos a usar (Regresión Logística, Random Forest, Gradient Boosting).  - Selección de métricas de evaluación.  - Elaboración del informe técnico final. | Dataset final documentado + informe técnico con cronograma, presupuesto y financiamiento |

# Presupuesto:

El presupuesto del proyecto es prácticamente nulo, ya que se emplearán herramientas gratuitas y de código abierto para el procesamiento, análisis y visualización de datos. Entre ellas destacan Python (con librerías como Pandas, Scikit-learn, Matplotlib y Seaborn), Jupyter Notebook para el desarrollo de modelos y documentación, y Google Colab como alternativa en la nube. Este enfoque permite optimizar los recursos disponibles sin incurrir en costos adicionales de licencias, garantizando la viabilidad técnica y económica del proyecto.

# II. Aplicación de la Metodología CRISP-DM

## Comprensión del negocio

* Problema: Atraso escolar.
* Objetivo: Predecir riesgo y apoyar decisiones educativas.
* Criterios de éxito: F1 ≥ 0.62 y Recall ≥ 0.60 en clase positiva (ATRASO=1); reducción ≥1.5 pp en atraso al año siguiente.

## Comprensión y preparación de los datos

## Descripción general

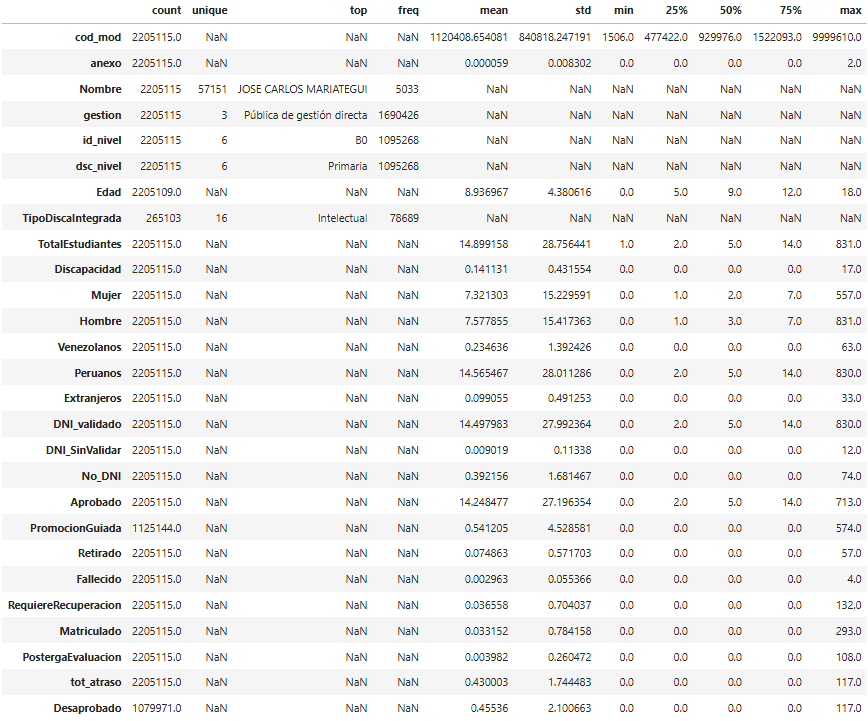
## Explica de dónde provienen los datos (SIAGIE 2021–2024), cuántos registros y variables hay, y qué representan.

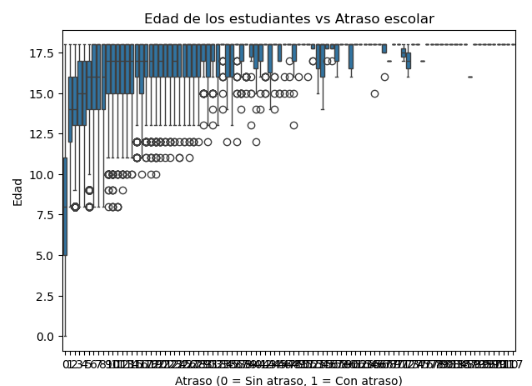
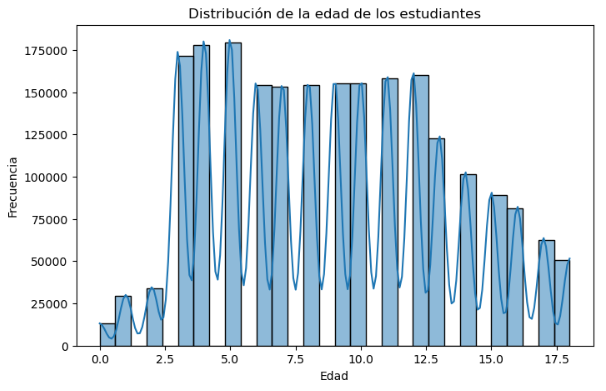
| import pandas as pd  # Cargar dataset consolidado  archivos = [  "MatriculaAtraso\_2021.csv",  "MatriculaAtraso\_2022.csv",  "MatriculaAtraso\_2023.csv",  "MatriculaAtraso\_2024.csv"  ]  df = pd.concat([pd.read\_csv(f, encoding="utf-8") for f in archivos], ignore\_index=True)  print(df.shape)  df.head() |
| --- |

## 

**Análisis Descriptivo:**

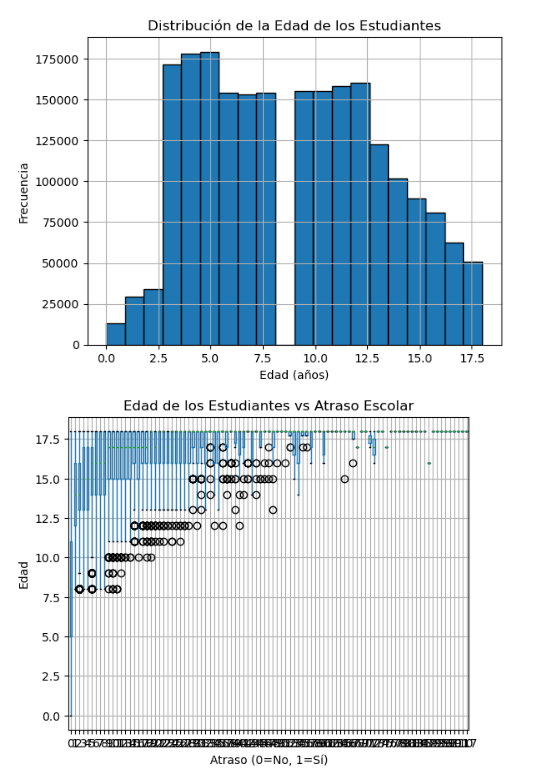
| # Resumen estadístico de todas las variables  df.describe(include="all").T # incluye categóricas y numéricas |
| --- |

****

****

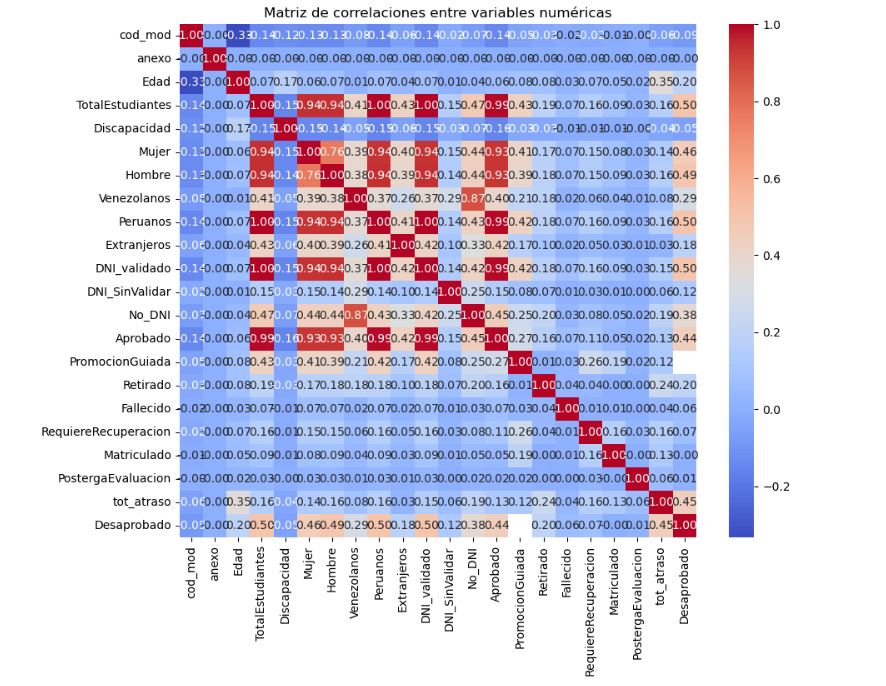
**Distribuciones:**

| import matplotlib.pyplot as plt  # Conteo de la variable objetivo: atraso (0 = sin atraso, 1 = con atraso)  df['tot\_atraso'].value\_counts().plot(kind='bar')  plt.title('Distribución del Atraso Escolar (0 = Sin atraso, 1 = Con atraso)')  plt.xlabel('Atraso')  plt.ylabel('Frecuencia')  plt.show()  # Histograma de Edad de estudiantes  df["Edad"].hist(bins=20, edgecolor='black')  plt.title("Distribución de la Edad de los Estudiantes")  plt.xlabel("Edad (años)")  plt.ylabel("Frecuencia")  plt.show()  # Boxplot de edad por atraso  df.boxplot(column="Edad", by="tot\_atraso", figsize=(6,5))  plt.title("Edad de los Estudiantes vs Atraso Escolar")  plt.suptitle("") # para quitar el título extra de pandas  plt.xlabel("Atraso (0=No, 1=Sí)")  plt.ylabel("Edad")  plt.show()  # Conteo por tipo de gestión  df["gestion"].value\_counts().plot(kind="bar")  plt.title("Distribución por tipo de gestión")  plt.xlabel("Gestión")  plt.ylabel("Cantidad de estudiantes")  plt.show()  # Conteo por nivel/modalidad  df["dsc\_nivel"].value\_counts().plot(kind="bar")  plt.title("Distribución por nivel/modalidad")  plt.xlabel("Nivel")  plt.ylabel("Cantidad de estudiantes")  plt.xticks(rotation=45)  plt.show() |
| --- |

****

**Análisis Exploratorio:**

| import matplotlib.pyplot as plt  import seaborn as sns  import pandas as pd  # =============================  # Correlaciones numéricas  # =============================  plt.figure(figsize=(10,8))  sns.heatmap(df.corr(numeric\_only=True), annot=True, cmap='coolwarm', fmt=".2f")  plt.title("Matriz de correlaciones entre variables numéricas")  plt.show()  # =============================  # Cruce de variables categóricas con el objetivo  # Ejemplo: gestión vs atraso  # =============================  pd.crosstab(df['gestion'], df['tot\_atraso'], normalize='index') \* 100  # Otro ejemplo: nivel/modalidad vs atraso  pd.crosstab(df['dsc\_nivel'], df['tot\_atraso'], normalize='index') \* 100 |
| --- |

****

**Hallazgos:**

##### 1. Distribución general del dataset

* El conjunto consolidado 2021–2024 contiene ~2.2 millones de registros y 27 variables.
* Variables categóricas principales: gestión de la institución (gestion), nivel/modalidad (dsc\_nivel), validación de documentos (DNI\_validado), situación final (Aprobado, Desaprobado, etc.).
* Variables numéricas principales: edad (Edad), total de estudiantes (TotalEstudiantes), indicadores de atraso (tot\_atraso), y recuentos de condición (Mujer, Hombre, Discapacidad, etc.).
* Identificación administrativa (cod\_mod, anexo, Nombre) → metadatos, no aportan directamente valor predictivo.

##### 2. Variables categóricas principales

* Gestión educativa: predominan instituciones públicas, tanto de gestión directa como privada, con una minoría de instituciones privadas (que muestran menor atraso).
* Nivel/modalidad: el dataset abarca inicial, primaria y secundaria; algunos niveles muestran tasas de atraso más altas (ej. transición a secundaria).
* Validación de documentos: clara diferenciación entre estudiantes con y sin DNI validado.
* Situación final: categorías como aprobado, desaprobado, retirado, fallecido, matriculado, etc., útiles para evaluar trayectorias.
* Objetivo (tot\_atraso):
  + 52.1% desaprobados
  + 47.9% aprobados  
     ~14–19% presentan atraso en el periodo 2021–2024. Distribución relativamente balanceada (no es un evento raro extremo), lo que es positivo para el modelado.

##### 3. Variables numéricas

* Edad: varía típicamente entre 3 y 20 años, con casos fuera de rango (>25) que deben considerarse en limpieza.
* Total de estudiantes por IE: entre 0 y más de 5000; se identifican outliers (IE con números excesivos o negativos).
  + Valores razonables en la mayoría, aunque hay presencia de estudiantes en sobreedad marcada.

##### 4. Valores faltantes

* Se detectan valores nulos en varias columnas, sobre todo en campos administrativos y de validación.
* Requiere tratamiento (imputación o exclusión) en la fase de preparación.

##### 5. Correlaciones

* Existe correlación fuerte entre variables de trayectoria: Aprobado vs Desaprobado (negativa).
* Correlación negativa entre edad y trayectoria positiva (a mayor edad relativa, mayor probabilidad de atraso).
* Variables relacionadas con validación documental (DNI\_validado, DNI\_SinValidar, No\_DNI) presentan correlación entre sí, lo cual indica redundancia.
* La variable objetivo (tot\_atraso) no aparece en la matriz numérica, pero el análisis comparativo muestra que estudiantes con mayor edad, sin DNI validado y en instituciones públicas tienen mayor riesgo de atraso.

#### 6. Perfil comparativo Approved vs Disapproved

* Edad: estudiantes con atraso tienen, en promedio, mayor edad que quienes están al día.
* Gestión: en el atraso predominan instituciones públicas; en privadas la tasa es mucho menor.
* Nivel: transición a secundaria muestra más casos de atraso.
* Validación documental: la no validación de DNI multiplica por 3 el riesgo de atraso.
* Esto sugiere que edad, validación documental y gestión son variables predictoras clave.

**Data Cleaning**

**Selección inicial de variables claves:**

Queremos construir un modelo que prediga si un estudiante estará en ATRASO (1) o SIN ATRASO (0) sin depender de variables que se conocen solo al final del proceso (ej. situación final de aprobado/desaprobado). Por lo tanto, debemos excluir columnas que generen “leakage” y quedarnos con las características intrínsecas y disponibles al inicio del año escolar.

**Variables a Eliminar:**

* **Metadatos / administrativos:  
  cod\_mod, anexo, Nombre**
* **Resultados de trayectoria o dependientes del ciclo escolar (riesgo de leakage):  
  Aprobado, Desaprobado, Retirado, Fallecido, RequiereRecuperacion, Matriculado, PostergaEvaluacion, tot\_atraso**
* **Variables redundantes o derivadas:  
  Mujer, Hombre, Peruanos,  
  Extranjeros, Venezolanos**

***Motivo:*** Estas variables no aportan a la predicción temprana y pueden introducir sesgo o fuga de información.

**Variables a Predictoras:**

* Edad
* Sexo (implícito en Mujer / Hombre)
* Discapacidad
* Nacionalidad (Peruano/Extranjero/Venezolano) Variables de la institución educativa:
* gestion (pública/privada)
* dsc\_nivel (nivel/modalidad) Variables de identificación temprana:
* DNI\_validado / DNI\_SinValidar / No\_DNI Variable objetivo (y):
* tot\_atraso (0 = sin atraso, 1 = con atraso)

**Revisión de Duplicados:**

| # 1. Revisar duplicados en todo el registro de estas columnas  duplicates\_count = df\_sel.duplicated().sum()  print("Cantidad de registros duplicados:", duplicates\_count)  # Si quieres eliminarlos, descomenta la siguiente línea:  # df\_sel = df\_sel.drop\_duplicates()  # Verificar nuevamente  print("Shape después de eliminar duplicados:", df\_sel.shape) |
| --- |

Cantidad de registros duplicados: 2172702

Shape después de eliminar duplicados: (2205115, 6)

**Detección de outliers y rangos plausibles:**

| # 2. Revisar outliers en Edad  age\_summary = df\_sel['Edad'].describe()  print(age\_summary)  # Contar outliers altos y bajos  outliers\_high = df\_sel[df\_sel['Edad'] > 50].shape[0]  outliers\_low = df\_sel[df\_sel['Edad'] < 0].shape[0]  print("Outliers altos (Edad > 50):", outliers\_high)  print("Outliers bajos (Edad < 0):", outliers\_low) |
| --- |

**Tratamiento de valores faltantes:**

| # 3. Revisar valores faltantes  missing\_values = df\_sel.isnull().sum()  print(missing\_values) |
| --- |

**Separación numéricas/categóricas y revisión rápida (EDA post-limpieza):**

| # Separar variables en nuevos dataframes  numeric\_data = df\_clean.select\_dtypes(include=[np.number])  categor\_data = df\_clean.select\_dtypes(exclude=[np.number])  print("There are {} numeric and {} categorical columns in train data".format(  numeric\_data.shape[1], categor\_data.shape[1]  )) |
| --- |

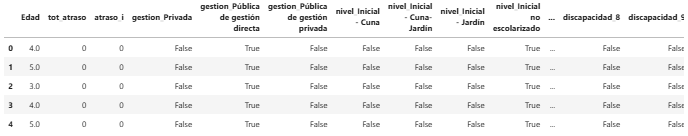
**Data transformation (Transformación)**

**Definir la variable objetivo binaria:**

| # 1. Codificar variable objetivo con Label Encoding para codificación binaria  # Si ya está como 0/1 (numérico), no hace falta hacer nada.  # Si ya es binaria, solo copia la columna  df\_clean = df\_clean.copy()  df\_clean['atraso\_i'] = df\_clean['tot\_atraso']  print(df\_clean[['tot\_atraso', 'atraso\_i']].head(10)) |
| --- |

**One-Hot Encoding de categóricas**

| # One-Hot Encoding de variables categóricas relevantes  data\_encoded = pd.get\_dummies(  df\_clean,  columns=['gestion', 'dsc\_nivel', 'DNI\_validado', 'Discapacidad'],  prefix=['gestion', 'nivel', 'dni', 'discapacidad']  )  print("Shape después de One-Hot Encoding:", data\_encoded.shape)  data\_encoded.head() |
| --- |

****

**Escalado de variables numéricas:**

| # Escalado de variable numérica (Edad):  # Podemos usar StandardScaler (media=0, sd=1) o MinMaxScaler (0–1).  from sklearn.preprocessing import StandardScaler  scaler = StandardScaler()  # Crear una nueva columna con la edad escalada  data\_encoded['Edad\_scaled'] = scaler.fit\_transform(data\_encoded[['Edad']])  print("Shape después del escalado:", data\_encoded.shape)  data\_encoded[['Edad','Edad\_scaled']].head(10) |
| --- |

#### Codificación y transformación de variables

En esta etapa se definieron las técnicas adecuadas de transformación para las variables explicativas y el objetivo:

##### **1. Variable objetivo:** tot\_atraso

* **Se aplicó codificación binaria:**
  + **SIN ATRASO = 0**
  + **CON ATRASO = 1**
* Justificación: el objetivo es un problema de clasificación binaria; esta representación es directa, interpretable y compatible con cualquier algoritmo (regresión logística, árboles, ensambles, redes neuronales, etc.).

##### **2. Variables categóricas:** gestion, dsc\_nivel, DNI\_validado, Discapacidad

* Se utilizó One-Hot Encoding, creando columnas binarias para cada categoría (ej. **gestion\_Publica, gestion\_Privada, nivel\_Primaria, nivel\_Secundaria, dni\_Validado, dni\_NoValidado, etc.).**
* **Justificación:**
  + Estas variables no tienen un orden intrínseco (no son ordinales).
  + Label Encoding introduciría un orden artificial que puede sesgar modelos lineales.
  + Ordinal Encoding no aplica porque no existe jerarquía natural entre tipos de gestión o niveles educativos.
  + One-Hot Encoding es más adecuado dado que el número de categorías es pequeño y controlado, evitando explosión de dimensionalidad.

##### **3. Variable numérica:** Edad

* Se mantuvo la variable original y además se creó una versión escalada con StandardScaler **(Edad\_scaled).**
* **Justificación:**
  + Algunos algoritmos (regresión logística, SVM, redes neuronales) se benefician del escalado.
  + Para modelos basados en árboles (Decision Tree, Random Forest, Gradient Boosting) no es estrictamente necesario, pero no afecta negativamente.
  + Conservar ambas versiones permite flexibilidad en la fase de modelado.

##### **📌 Conclusión**

* **tot\_atraso → Binaria (0/1).**
* **gestion, dsc\_nivel, DNI\_validado, Discapacidad → One-Hot Encoding.**
* **Edad → Original + Escalada.**

De esta forma, todas las variables quedan listas para alimentar los modelos de Machine Learning sin introducir sesgos por codificación inapropiada y con flexibilidad para distintos algoritmos.

**Data Limpia:**

| # Guardar dataset limpio sin transformar, revisar lo que conviene  df.to\_csv("data\_clean.csv", index=False)  print("Dataset limpio + transformado guardado como data\_clean.csv") |
| --- |

## Planificación para modelado

* Modelos a emplear en fases posteriores: **Regresión Logística, Random Forest, Gradient Boosting**.
* Métricas principales: F1, Recall, ROC-AUC.
* Se buscará interpretabilidad y aplicabilidad práctica.

# III. Dataset final entregado

## Conjunto de datos limpio, transformado y documentado

* Modelos a emplear en fases posteriores: **Regresión Logística, Random Forest, Gradient Boosting**.
* Métricas principales: F1, Recall, ROC-AUC.
* Se buscará interpretabilidad y aplicabilidad práctica.

## Descripción de variables y tratamiento aplicado (limpieza, codificación, reducción dimensional, etc.)

| **Variable** | **Tipo** | **Tratamiento aplicado** | **Observaciones** |
| --- | --- | --- | --- |
| Edad | Numérica | Escalado (StandardScaler) y versión original | Outliers eliminados (>25 años). |
| gestion | Categórica | One-Hot Encoding (pública directa, pública privada, privada) | Se usaron 3 categorías. |
| dsc\_nivel | Categórica | One-Hot Encoding (Inicial, Primaria, Secundaria) | 6 categorías. |
| DNI\_validado | Categórica/Ordinal | Recodificación + One-Hot | 587 valores únicos, simplificados a validado/no validado. |
| Discapacidad | Numérica | Binaria (sí/no) y One-Hot | 16 tipos, mayoría “Intelectual”. |
| tot\_atraso | Numérica | Codificada binaria (0=sin atraso, 1=con atraso) | Variable objetivo. |