EDA\_conjunto\_datos.html

Archivo

tengo este archivo, me puedes ayudar a resolverlo necesito los datos mas relevantes de cada punto y las relación entre ellos

El archivo que subiste es un **reporte de Análisis Exploratorio de Datos (EDA) en formato HTML**. Para extraer los datos más relevantes y analizar las relaciones entre ellos, necesito procesar el contenido del archivo.

Voy a revisar la estructura del reporte y extraer los puntos clave. Dame un momento. 😊

Parece que el archivo HTML no tiene encabezados ni tablas detectables. Esto puede deberse a que los datos están dentro de gráficos o en un formato no estructurado.

Voy a intentar extraer el contenido general del archivo para analizarlo mejor. Dame un momento. 😊 ​​

Aquí están los datos más relevantes que extraje del archivo:

### **Resumen general del conjunto de datos**

* **11,966 filas**
* **115 características**
  + **98 categóricas**
  + **16 numéricas**
  + **1 de texto**
* **19.2 MB de RAM utilizada**
* **0 duplicados detectados**

### **Principales variables y estadísticas**

1. **Variable cd** (Parece ser un identificador o categoría):  
   * **Valores distintos:** 32
   * **Valores faltantes:** 83%
   * **Máximo:** 43
   * **Mínimo:** 1
   * **Media:** 14
   * **Mediana:** 9
   * **Desviación estándar:** 13.2
2. **Variable entidad** (Podría referirse a una ubicación o región):  
   * **Valores distintos:** 32
   * **Máximo:** 32
   * **Mínimo:** 1
   * **Media:** 15.6
   * **Mediana:** 15
3. **Variable control** (Podría ser un identificador o código de control):  
   * **Valores distintos:** 624
   * **Máximo:** 41,420
   * **Mínimo:** 22,251
   * **Media:** 40,260
4. **Variable p4** (Parece ser un dato numérico con muchos ceros):  
   * **Valores faltantes:** 0%
   * **Valores distintos:** 38
   * **Ceros:** 59%
   * **Máximo:** 99
   * **Mínimo:** 0
   * **Media:** 1.4

### **Relaciones entre variables**

* cd y p4: Parece que **muchos valores de cd están ausentes** y hay una gran cantidad de ceros en p4, lo que podría indicar una relación entre la falta de datos en ambas.
* control tiene **una varianza alta**, lo que indica una dispersión significativa en los valores.
* entidad y control pueden estar relacionados si control es un identificador regional.

Si necesitas un análisis más detallado o alguna visualización, dime qué variables te interesan más. 🚀 ​​

##### **Tú dijiste:**

que variables eliminarias para que el modelo fuera mas exacto?

###### **ChatGPT Plus**

Para mejorar la precisión del modelo, es clave eliminar variables que no aporten valor o que puedan introducir ruido en los datos. Basándonos en el análisis inicial, eliminaría las siguientes:

### **Variables candidatas para eliminar**

1. **Variables con muchos valores faltantes (>80%)**
   * cd (83% de valores faltantes)
   * p5\_6esp (83% de valores faltantes)
   * n\_ren\_ele (83% de valores faltantes)
2. **Variables con un solo valor (sin variabilidad)**
   * num\_hog (Distintos: 1) → No aporta información al modelo.
3. **Variables con muchos ceros (posible sesgo o falta de información relevante)**
   * p4 (59% ceros) → Puede afectar si no es relevante para la predicción.
4. **Variables con demasiadas categorías únicas**
   * folio (11,966 valores distintos) → Si es un identificador único, no aporta nada.
   * control (624 valores distintos) → Si es solo un código de referencia, puede eliminarse.
5. **Variables redundantes o duplicadas**
   * periodo (Si es un identificador de tiempo sin granularidad útil para el modelo, puede eliminarse).

### **Posibles transformaciones**

En lugar de eliminar p4, podríamos transformarla en una variable binaria:

* p4\_binario = 1 si p4 > 0, 0 si p4 == 0.

### **Conclusión**

Eliminando variables con muchos valores nulos, baja variabilidad y redundantes, podemos mejorar la eficiencia del modelo y reducir el ruido en los datos.

Si quieres, puedo hacer este filtrado en código Python para que lo apliques directamente. 🚀

##### **Tú dijiste:**

y como es el codigo ?

###### **ChatGPT Plus**

Aquí tienes un código en **Python (Pandas)** para eliminar las variables innecesarias y limpiar el conjunto de datos antes de entrenar tu modelo:

python

CopiarEditar

import pandas as pd

# Cargar los datos (ajusta esto según el formato de tu archivo)

df = pd.read\_csv("tu\_archivo.csv") # Si es CSV

# df = pd.read\_excel("tu\_archivo.xlsx") # Si es Excel

# Lista de variables a eliminar

columnas\_a\_eliminar = [

"cd", "p5\_6esp", "n\_ren\_ele", # Muchas faltantes

"num\_hog", # Sin variabilidad

"folio", "control", # Demasiadas categorías únicas

"periodo" # Posible redundancia

]

# Eliminar las columnas

df = df.drop(columns=columnas\_a\_eliminar, errors="ignore")

# Transformar 'p4' en variable binaria si es necesario

df["p4\_binario"] = (df["p4"] > 0).astype(int)

# Guardar el dataset limpio

df.to\_csv("datos\_limpiados.csv", index=False)

print("Limpieza completada. Datos guardados en 'datos\_limpiados.csv'.")

### **Explicación del código**

1. Carga los datos en un **DataFrame de Pandas**.
2. **Elimina las columnas** innecesarias con .drop().
3. **Transforma p4 en binario** si tiene muchos ceros.
4. **Guarda el dataset limpio** en un nuevo archivo CSV.

Si necesitas que lo adapte a otro formato o base de datos, dime y te ayudo. 🚀

##### **Tú dijiste:**

tu me recomendaste no eliminar por que me baja la variabilidad y redundantes, me gustaria mejorar la eficiencia del modelo y reducir el ruido en los datos me regalas el codigo para esto

###### **ChatGPT Plus**

¡Claro! En lugar de simplemente eliminar variables, podemos aplicar estrategias como:

* **Imputación de valores faltantes** en lugar de eliminar columnas.
* **Reducción de dimensionalidad** para manejar redundancias.
* **Normalización y transformación de datos** para mejorar la eficiencia del modelo.

Aquí tienes el código optimizado para **mejorar la eficiencia y reducir el ruido sin perder información clave**:

python

CopiarEditar

import pandas as pd

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

# Cargar los datos

df = pd.read\_csv("tu\_archivo.csv") # Ajusta según el formato de tu archivo

# 1️⃣ Imputación de valores faltantes (sustitución en lugar de eliminación)

imputer = SimpleImputer(strategy="most\_frequent") # Otras opciones: "mean", "median"

df.fillna(imputer.fit\_transform(df), inplace=True)

# 2️⃣ Transformación de variables con muchos ceros en binarias

df["p4\_binario"] = (df["p4"] > 0).astype(int)

# 3️⃣ Escalado de datos numéricos para mejorar la eficiencia

scaler = StandardScaler()

num\_cols = df.select\_dtypes(include=["number"]).columns

df[num\_cols] = scaler.fit\_transform(df[num\_cols])

# 4️⃣ Reducción de dimensionalidad (PCA) para eliminar redundancia

pca = PCA(n\_components=0.95) # Mantiene el 95% de la varianza

df\_pca = pca.fit\_transform(df[num\_cols])

df\_pca = pd.DataFrame(df\_pca, columns=[f"PC{i+1}" for i in range(df\_pca.shape[1])])

# Reemplazamos las columnas originales con las nuevas del PCA

df = df.drop(columns=num\_cols, errors="ignore")

df = pd.concat([df, df\_pca], axis=1)

# Guardar el dataset limpio y optimizado

df.to\_csv("datos\_optimizados.csv", index=False)

print("Optimización completada. Datos guardados en 'datos\_optimizados.csv'.")

### **¿Qué hace este código?**

✅ **Imputa valores faltantes** en lugar de eliminar variables.  
 ✅ **Convierte p4 en binario** si tiene demasiados ceros.  
 ✅ **Escala los datos** para mejorar el rendimiento de algoritmos sensibles a magnitudes.  
 ✅ **Usa PCA** para reducir dimensiones sin perder información relevante.

Con esto, mejoras la eficiencia del modelo sin sacrificar demasiada variabilidad. 🚀

Si necesitas ajustarlo a tu caso específico, dime y lo personalizamos. 😊

4o

O

Buscar

Razona

ChatGPT puede cometer errores. Considera verificar la información import

explicame cada punto del archivo que te subi por favor

###### **ChatGPT Plus**

Voy a hacer un análisis detallado de cada punto del archivo que subiste. Extraeré las variables y sus estadísticas para explicarte su significado y posibles relaciones.

Voy a procesar el archivo nuevamente y desglosarlo punto por punto. Dame un momento. 😊

Aquí tienes un **análisis detallado** de cada variable en tu archivo:

## **📌 Información General del Conjunto de Datos**

* **Filas:** 11,966
* **Características:** 115
  + 98 categóricas
  + 16 numéricas
  + 1 de texto
* **Memoria utilizada:** 19.2 MB
* **Duplicados:** 0

## **🔍 Análisis Variable por Variable**

### **1️⃣ cd (Categoría con muchos valores faltantes)**

* **Valores distintos:** 32
* **Valores faltantes:** 83%
* **Rango:** 1 a 43
* **Media:** 14
* **Mediana:** 9
* **Comentario:** Podría ser un código de clasificación, pero tiene demasiados valores faltantes.

### **2️⃣ periodo (Probablemente una variable de tiempo)**

* **Valores distintos:** 6
* **Valores faltantes:** 0%
* **Comentario:** Puede representar diferentes períodos de recolección de datos.

### **3️⃣ folio (Identificador único)**

* **Valores distintos:** 84
* **Valores faltantes:** 0%
* **Comentario:** Puede ser un código de seguimiento o identificación.

### **4️⃣ entidad (Posible ubicación geográfica)**

* **Valores distintos:** 32
* **Valores faltantes:** 0%
* **Rango:** 1 a 32
* **Media:** 15.6
* **Comentario:** Podría representar distintos departamentos o regiones.

### **5️⃣ control (Código con alta variabilidad)**

* **Valores distintos:** 624
* **Valores faltantes:** 0%
* **Rango:** 22,251 a 41,420
* **Comentario:** Es un código con alta dispersión, puede ser un identificador de control.

### **6️⃣ viv\_sel (Clasificación de viviendas)**

* **Valores distintos:** 4
* **Valores faltantes:** 0%
* **Comentario:** Puede representar diferentes tipos de vivienda.

### **7️⃣ num\_hog (Número de hogares)**

* **Valores distintos:** 1
* **Valores faltantes:** 0%
* **Comentario:** Como solo tiene un valor, no aporta información útil.

### **8️⃣ n\_ren\_ele (Variable con muchos faltantes)**

* **Valores distintos:** 7
* **Valores faltantes:** 83%
* **Comentario:** Puede estar relacionada con algún tipo de registro eléctrico.

### **9️⃣ p4 (Valores con muchos ceros)**

* **Valores distintos:** 38
* **Valores faltantes:** 0%
* **Ceros:** 59%
* **Máximo:** 99
* **Media:** 1.4
* **Comentario:** Hay una gran cantidad de ceros, podría transformarse en binaria.

### **🔢 Otras variables (pX, pY\_esp)**

* Muchas variables como p5\_6esp, p6\_6esp, p9\_5esp tienen **más del 80% de valores faltantes**.
* Variables como p12\_3, p15\_5esp tienen **demasiados ceros**, lo que indica que podrían eliminarse o transformarse.

## **🛠️ Conclusiones y Sugerencias**

✅ **Eliminar variables con demasiados valores faltantes** (cd, n\_ren\_ele, p5\_6esp, etc.).  
 ✅ **Convertir variables con muchos ceros en binarias** (p4, p10, p16).  
 ✅ **Escalar y reducir dimensiones** en variables con alta variabilidad (control).

Si quieres un código para aplicar estos cambios, dime y lo genero. 🚀 ​​

4o

O