Maestría en Inteligencia Artificial Aplicada

Materia: Proyecto Integrador

Profesor Titular: Dra. Grettel Barceló Alonso / Dr. Luis Eduardo Falcón Morales

Asesor de Proyecto: Dr. Carlos Alberto Villaseñor Padilla

Avance 1. Análisis exploratorio de datos

Equipo 10

David García Robles A01152606

David Nava Jiménez A01168501

José Antonio Hernández Hernández A01381334

Fecha: 11 de Mayo de 2025

Optimización de ventas en Nacional Monte de Piedad

Avance 1. Análisis exploratorio de datos

Contenido

Introducción

El presente documento tiene la finalidad de exponer el análisis de datos exploratorio (EDA) realizado sobre el dataset compartido por Nacional Monte de Piedad. Esta institución cuenta con una infraestructura robusta de almacenamiento de datos basada en plataformas como Oracle, Databricks y un data lake que concentran información estructurada y no estructurada proveniente de múltiples fuentes internas.

El objetivo del análisis es comprender la estructura general del dataset, evaluar la calidad y robustez de los datos disponibles, e identificar aquellas variables clave que podrían aportar valor en la construcción de un modelo de machine learning. Este proceso incluye la detección de valores nulos, la distribución de las variables numéricas, la presencia de valores atípicos, así como la evaluación de correlaciones que puedan dar lugar a ingeniería de características más eficiente.

Además, se busca validar que el volumen y representatividad de los datos sea adecuado para dividirse en subconjuntos de entrenamiento, validación y prueba, asegurando así condiciones optimas para el desarrollo y la generalización de modelos predictivos en etapas posteriores.

1.1 Importación de librerías

```
from google.colab import drive
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('ventas.csv', encoding='iso-8859-1', low_memory=False)


df.head()
```

	SUCURSAL	ESTADO_SUCURSAL	CLAVE_OPERACION	OPERACION	PARTIDA	ORIGEN	DESCRIPCION_PARTIDA	GRAMAJE	KILATAJE	AVALUO_COMPLEMENTARIO	...	IVACOM_PASECOM	COM_EXHIBICION	IVACOM_EXHIBICION	INTERES_DEPRECUP	IVAINT_DEPRECUP	FECHA_MAX_DEP_RECUP	FECHA_CARGA	num_particion	imp_minusvalia	imp_cancelacion_in
0	1005	CIUDAD DE MÃ□XICO	VP	Venta al Publico	181615421	SIVA	176231504-1 ANILLO TIPO:DAMA DISEÃ□O:CABUJON D...	4	14	0	...	NaN	NaN	NaN	NaN	NaN	NaN	2024-05-03T05:01:25.637Z	202405.0	NaN	Na
1	1005	CIUDAD DE MÃ□XICO	VP	Venta al Publico	181616741	MIDAS	174275433-1 ANILLO ORO AMARILLO 14K PESO 1.60 ...	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	2024-05-03T05:01:25.637Z	202405.0	NaN	Na
2	1005	CIUDAD DE MÃ□XICO	VP	Venta al Publico	181616743	MIDAS	173896016-1 MEDIA CHURUMBELA ORO AMARILLO 14K...	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	2024-05-03T05:01:25.637Z	202405.0	NaN	Na
3	16	AGUASCALIENTES	VP	Venta al Publico	181662772	SIVA	1 COLLAR TIPO:ROSARIO DISEÃ□O:ESFERAS LISAS D...	23.9	10	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	2024-05-03T05:01:25.637Z	202405.0	NaN	Na
4	278	CIUDAD DE MÃ□XICO	VP	Venta al Publico	181809632	SIVA	1 ACCESORIOS TIPO:LLAVERO DISEÃ□O:GRABADO DE O...	12.4	8	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	2024-05-03T05:01:25.637Z	202405.0	NaN	Na

5 rows × 64 columns

Interpretación del DataFrame

Durante la carga del archivo CSV, se detectó que algunos caracteres especiales como acentos y símbolos fueron mal interpretados, generando valores con caracteres especiales. Aunque el archivo original estaba en UTF-8 al utilizar pandas.read se forzó a la codificación alternativa que fue Latin-1.

```
print("Número de filas:", df.shape[0])                                    # Imprimir numero de
filas
print("Número de columnas:", df.shape[1])                                 # Imprimir numero de
columnas
```

Número de filas: 1048575

Número de columnas: 64

1.2 Valores Faltantes

```
df.info()                                                    # Imprimir información del conjunto de datos
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 64 columns):
 #   Column                     Non-Null Count    Dtype
---  ------                     --------------    -----
 0   SUCURSAL                   1048533 non-null  object
 1   ESTADO_SUCURSAL            1048538 non-null  object
 2   CLAVE_OPERACION            1048516 non-null  object
 3   OPERACION                  1048511 non-null  object
 4   PARTIDA                    1048516 non-null  object
 5   ORIGEN                     1048516 non-null  object
 6   DESCRIPCION_PARTIDA        1048436 non-null  object
 7   GRAMAJE                    941340 non-null   object
 8   KILATAJE                   866459 non-null   object
 9   AVALUO_COMPLEMENTARIO      915717 non-null   object
 10  FACTOR_HECHURA             941352 non-null   object
 11  FACTOR                     941343 non-null   object
 12  VALOR_MONTE                1048243 non-null  object
 13  VALOR_MONTE_ACTUALIZADO    846157 non-null   object
 14  AVALUO_COMERCIAL           933994 non-null   object
 15  PRESTAMO                   1048248 non-null  object
 16  PRECIO_VENTA_INICIAL       1048245 non-null  object
 17  PRECIO_VENTA_FINAL         1048240 non-null  object
 18  FECHA_EMPENO               1048225 non-null  object
 19  FECHA_COMERCIALIZACION     1048223 non-null  object
 20  VALOR_ANCLA_ORO            846135 non-null   object
 21  RAMO                       1048221 non-null  object
 22  SUBRAMO                    1048208 non-null  object
 23  REFRENDOS_REALIZADOS       1048216 non-null  object
 24  INCREMENTO                 857161 non-null   object
 25  DESPLAZAMIENTO_COMERCIAL   857112 non-null   object
 26  VALUADOR                   1020141 non-null  object
 27  FECHA_HORA_MOV             1048145 non-null  object
 28  GASTOSOPERACION            910814 non-null   object
 29  DEMASIA                    1019744 non-null  object
 30  INTERES                    1018565 non-null  object
 31  IVAINTERESDEPOSITO         116 non-null      object
 32  IVAINTERESALMONEDA         114 non-null      object
 33  IVAGASTOSOPERACION         122 non-null      object
 34  INTERESALMONEDA            900852 non-null   object
 35  DES_EXT                    1008647 non-null  object
 36  IVA_DESEXT                 100 non-null      object
 37  IMPORTE_VENTA              1048177 non-null  object
 38  PRODUCTO                   1048178 non-null  object
 39  TASA_OFERTA                1048172 non-null  object
 40  CANAL                      1048087 non-null  object
 41  DIAS_ALMONEDA              1048110 non-null  object
 42  RANGO_DIAS_ALMONEDA        1048062 non-null  object
 43  PRECIO_VENTA_FINAL_SID     1048108 non-null  object
 44  TIPO_PRENDA                1048014 non-null  object
 45  FCH_CARGA                  1048097 non-null  object
 46  FECHA_EMPENO_OK            11 non-null       object
 47  FECHA_HORA_MOV_OK          11 non-null       object
```

Interpretaciones

Podemos observar que todas las columnas tienen tipo de dato "object", por lo que en la etapa de preprocesamiento vamos a modificarlas de acuerdo con el tipo de dato que se utiliza en el negocio.

```
print("\n Valores faltantes;\n", df.isnull().sum())                                    # Imprimir valores
faltantes
missing_percentage = df.isnull().sum() / len(df) * 100                                 # Imprimir porcentaje de
valores faltantes
print("\n Porcentaje de valores faltantes;\n", missing_percentage)
```

```
 Valores faltantes;
 SUCURSAL                    42
ESTADO_SUCURSAL              37
CLAVE_OPERACION              59
OPERACION                    64
PARTIDA                      59
                            ...
FECHA_MAX_DEP_RECUP      376751
FECHA_CARGA                 487
num_particion               496
imp_minusvalia          1048575
imp_cancelacion_int     1048575
Length: 64, dtype: int64

 Porcentaje de valores faltantes;
 SUCURSAL                   0.004005
ESTADO_SUCURSAL             0.003529
CLAVE_OPERACION             0.005627
OPERACION                   0.006104
PARTIDA                     0.005627
                             ...
FECHA_MAX_DEP_RECUP        35.929810
FECHA_CARGA                 0.046444
num_particion               0.047302
imp_minusvalia            100.000000
imp_cancelacion_int       100.000000

Length: 64, dtype: float64
```

Interpretaciones

Se identificaron columnas con distintos niveles de valores faltantes. La mayoría de las variables presentan una proporción mínima de valores nulos (menor al 1%), lo cual indica una buena cardinalidad general del dataset. Sin embargo, existen algunas variables que destacan por su alta proporción de dato faltantes:

- Las columnas imp_minusvalia e imp_cancelacion_int presentan el 100% de valores nulos, por lo que es probable que no aporten información útil en su estado actual y podrían ser candidatas a ser eliminadas, salvo que se justifique su retención por motivos de negocio o enriquecimiento posterior.
- La variable FECHA_MAX_DEP_RECUP tiene un 35.95% de valores faltantes lo que indica una cantidad significativa de ausencias.
- Otras columnas como FECHA_Carga, num_particion, OPERACIÓN y CLAVE_OPERACION muestran una proporción baja de nulos <0.05% por lo que podrían mantenerse tras una imputación básica.

1.3 Análisis Descriptivo (univariante)

De acuerdo con la naturaleza del negocio, se procedió a realizar un análisis de cada variable que conforma el dataset, se identificaron el tipo real de dato que contiene cada columna

Columna	Tipo de dato
SUCURSAL	Cualitativa
ESTADO_SUCURSAL	Cualitativa
CLAVE_OPERACION	Cualitativa
OPERACION	Cualitativa
PARTIDA	Cualitativa
ORIGEN	Cualitativa
DESCRIPCION_PARTIDA	Cualitativa
GRAMAJE	numérica/continua
KILATAJE	numérica /discreta
AVALUO_COMPLEMENTARIO	numérica /continua
FACTOR_HECHURA	Cualitativa
FACTOR	Cualitativa
VALOR_MONTE	numérica/continua
VALOR_MONTE_ACTUALIZADO	numérica/continua
AVALUO_COMERCIAL	numérica/continua
PRESTAMO	numérica/continua
PRECIO_VENTA_INICIAL	numérica/continua
PRECIO_VENTA_FINAL	numérica/continua
FECHA_EMPENO	Cualitativa
FECHA_COMERCIALIZACION	Cualitativa
VALOR_ANCLA_ORO	numérica/continua
RAMO	Cualitativa
SUBRAMO	Cualitativa
REFRENDOS_REALIZADOS	numércia/continua
INCREMENTO	numérica/discreta
DESPLAZAMIENTO_COMERCIAL	numérica/discreta

VALUADOR	Cualitativa
FECHA_HORA_MOV	Cualitativa
GASTOSOPERACION	sin datos
DEMASIA	numérica/continua
INTERES	numérica/continua
IVAINTERESDEPOSITO	sin datos
IVAINTERESALMONEDA	sin datos
IVAGASTOSOPERACION	sin datos
INTERESALMONEDA	numérica/continua
DES_EXT	numérica/continua
IVA_DESEXT	sin datos
IMPORTE_VENTA	numérica/continua
PRODUCTO	Cualitativa
TASA_OFERTA	numérica/continua
CANAL	Cualitativa
DIAS_ALMONEDA	numérica/continua
RANGO_DIAS_ALMONEDA	Cualitativa
PRECIO_VENTA_FINAL_SID	numérica/continua
TIPO_PRENDA	Cualitativa
FCH_CARGA	Cualitativa
FECHA_EMPENO_OK	sin datos
FECHA_HORA_MOV_OK	sin datos
FECHA_COMERCIALIZACION_OK	sin datos
CUSTODIA	sin datos
SALDO_INSOLUTO	numérica/continua
COM_ALMACENAJE	sin datos
IVACOM_ALMACENAJE	sin datos
COMPASE_COMERCIALIZACION	sin datos

Interpretaciones

A diferencia del análisis estadístico que se obtuvo previamente, este análisis nos permitió identificar visualmente y conforme a los argumentos proporcionados por los directores de Monte de Piedad, cuales datos realmente pueden ser catalogados como cuantitativos o cualitativos. Con el fin de tener un mejor panorama de aquellos campos clave que formaran parte de las transformaciones.

1.4 Análisis Descriptivo (univariante)

```
print("\n Estadisticas numéricas del conjunto de datos;\n")
print(df.describe())                                                    # Imprimir estadisticas
numericas del conjunto de datos
```

Estadisticas numéricas del conjunto de datos;

```
       num_particion  imp_minusvalia  imp_cancelacion_int
count   1.048079e+06             0.0                  0.0
mean    2.024072e+05             NaN                  NaN
std     3.158538e+00             NaN                  NaN
min     2.024010e+05             NaN                  NaN
25%     2.024050e+05             NaN                  NaN
50%     2.024070e+05             NaN                  NaN
75%     2.024100e+05             NaN                  NaN
max     2.024120e+05             NaN                  NaN
```

```
print("\n Estadisticas categoricas del conjunto de datos;\n")
print(df.describe(include=['O']))                                    # Imprimir estadisticas
categoricas del conjunto de datos
```

Estadisticas categoricas del conjunto de datos;

```
       SUCURSAL    ESTADO SUCURSAL CLAVE OPERACION         OPERACION  \
count   1048533             1048538         1048516           1048511
unique      458                 163             113               101
top        1001  CIUDAD DE MÃ□XICO              VP  Venta al Publico
freq      107819              313839         1047780           1047780

                                       PARTIDA   ORIGEN  \
count                                  1048516  1048516
unique                                 1036647      110
top     Buen Estado Sin Personalizar / Sin Abollar     SIVA
freq                                        20   941333

          DESCRIPCION_PARTIDA GRAMAJE KILATAJE AVALUO_COMPLEMENTARIO  ...  \
count                 1048436  941340   866459                915717  ...
unique                 934957    4430      151                  3077  ...
top     BROQUELES  0.1GR  14K       2       14                     0  ...
freq                      263   28127   379165                906637  ...

       COM_ALMACENAJE         IVACOM_ALMACENAJE COMPASE_COMERCIALIZACION  \
count              69                        10                       32
unique             54                         8                       30
top            202405  2024-05-12T05:01:22.463Z                   202405
freq                6                         2                        2

                 IVACOM_PASECOM COM_EXHIBICION IVACOM_EXHIBICION  \
count                        66         301790                 9
unique                       41          56110                 9
top     2024-05-13T09:10:50.693Z           70.5            202405
freq                         17            770                 1

       INTERES_DEPRECUP IVAINT_DEPRECUP FECHA_MAX_DEP_RECUP  \
count                 7              56              671824
unique                7              56                 329
top              202405      16/04/2024          23/05/2024
freq                  1               1               22457

                    FECHA_CARGA
count                   1048088
unique                      220
top     2024-05-13T09:10:50.693Z
freq                     110787

[4 rows x 61 columns]
```

Interpretaciones

La única variable numérica con valores registrados es num_participation, que parece actuar como una especia de identificador por bloque o segmento. Presenta una distribución dispersa std=3.16 lo cual indica que los registros pertenecen a un rango estrecho de particiones.

En cuanto a las variables categóricas, tienen un numero elevado de valores únicos que analizaremos mas adelante durante la frecuencia de variables y cardinalidad.

1.5 Frecuencia de variables categoricas

```
cat_cols = df.select_dtypes(include=['object']).columns.tolist()                    # Obtener columnas categoricas


for col in cat_cols:                                                                # Imprimir frecuencia de variables categoricas
    print(f"Frecuencia de {col}:")
    print(df[col].value_counts())
    print("\n")
```

```
Frecuencia de SUCURSAL:
SUCURSAL
1001                                  107819
0                                      42823
1005                                   13683
28                                     13577
12                                      9061
                                       ...
Relojes Comerciales - Gama Media Baja      1
0.91                                       1
CO                                         1
0.3849                                     1
GRAMOS 2.2 ....\""                         1
Name: count, Length: 458, dtype: int64


Frecuencia de ESTADO_SUCURSAL:
ESTADO_SUCURSAL
CIUDAD DE MÃ©XICO                                                                313839
ESTADO DE MÃ©XICO                                                                115442
VERACRUZ                                                                          92393
JALISCO                                                                           65015
NUEVO LEÃ³N                                                                       64662
                                                                                  ...
3.3                                                                                   1
MONEDA DE 50 PS ORO NACIONAL ALO 1821 1947 GRS 41.6. EN ESTUCHE DE ACRÃlLICO GRS TT 46.8      1
2004210                                                                               1
18                                                                                    1
155                                                                                   1
Name: count, Length: 163, dtype: int64


Frecuencia de CLAVE_OPERACION:
CLAVE_OPERACION
VP            1047780
DN                547
14                 26
10                 12
0                  12
                ...
5.4                 1
41.67               1
14/08/2024          1
30/05/2024          1
25/04/2024          1
Name: count, Length: 113, dtype: int64


Frecuencia de OPERACION:
OPERACION
Venta al Publico           1047780
Devolucion Mercancia           547
0                               55
Cumplido                        10
Alhajas                          6
                              ...
```

Interpretaciones

En el análisis de frecuencias, permitió identificar la concentración, calidad y diversidad de los valores en variables categóricas clave del dataset.

- Variables como OPERACIÓN y CLAVE_OPERACION están fuertemente dominadas por valores únicos, como “Venta al Público” y “VP”, lo cual indica baja variabilidad y poca aportación informativa.
- En CANAL, mas del 95% de los registros corresponden a “Cumplido” lo que sugiere un comportamiento sin mucha dispersión.
- SUCURSAL, ESTADO_SUCURSAL, PRODUCTO o FACTORES presentan valores numéricos, fechas o textos inconsistentes dentro de campos categóricos, lo cual indica errores de captura o mezcla de variables. Esto representa una oportunidad durante el preprocesamiento de los datos.

1.6 Cardinalidad de las variables categóricas

```
print("\n Cardinalidad de las variables categoricas;\n")
for col in cat_cols:                                                    # Imprimir cardinalidad
de variables categoricas
    print(f"Cardinalidad de {col}: {df[col].nunique()}")
    print("\n")
```

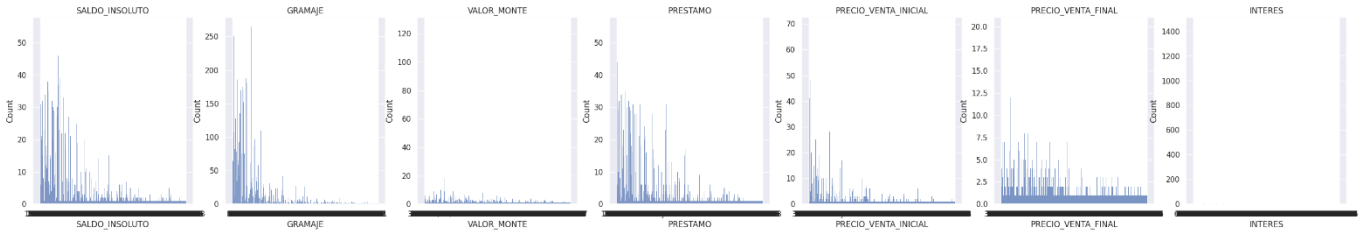
```
Cardinalidad de CLAVE_OPERACION: 113
Cardinalidad de OPERACION: 101
Cardinalidad de PARTIDA: 1036647
Cardinalidad de ORIGEN: 110
Cardinalidad de DESCRIPCION_PARTIDA: 934957
Cardinalidad de GRAMAJE: 4430
Cardinalidad de KILATAJE: 151
Cardinalidad de AVALUO_COMPLEMENTARIO: 3077
Cardinalidad de FACTOR_HECHURA: 159
Cardinalidad de FACTOR: 132
Cardinalidad de VALOR_MONTE: 38714
Cardinalidad de VALOR_MONTE_ACTUALIZADO: 37482
Cardinalidad de AVALUO_COMERCIAL: 41660
Cardinalidad de PRESTAMO: 28458
Cardinalidad de PRECIO_VENTA_INICIAL: 41464
Cardinalidad de PRECIO_VENTA_FINAL: 43238
Cardinalidad de FECHA_EMPENO: 1635
Cardinalidad de FECHA_COMERCIALIZACION: 1382
Cardinalidad de VALOR_ANCLA_ORO: 1514
Cardinalidad de SUBRAMO: 70
Cardinalidad de REFRENDOS_REALIZADOS: 79
Cardinalidad de INCREMENTO: 115
Cardinalidad de DESPLAZAMIENTO_COMERCIAL: 62
Cardinalidad de VALUADOR: 855
Cardinalidad de FECHA_HORA_MOV: 283
Cardinalidad de GASTOSOPERACION: 70
Cardinalidad de DEMASIA: 132103
Cardinalidad de INTERES: 140278
Cardinalidad de IVAINTERESDEPOSITO: 103
Cardinalidad de IVAINTERESALMONEDA: 37
Cardinalidad de IVAGASTOSOPERACION: 64
Cardinalidad de INTERESALMONEDA: 94409
Cardinalidad de DES_EXT: 27780
Cardinalidad de IVA_DESEXT: 44
Cardinalidad de IMPORTE_VENTA: 43189
Cardinalidad de PRODUCTO: 54
Cardinalidad de TASA_OFERTA: 129
Cardinalidad de CANAL: 53
Cardinalidad de DIAS_ALMONEDA: 1693
Cardinalidad de RANGO_DIAS_ALMONEDA: 33
Cardinalidad de PRECIO_VENTA_FINAL_SID: 43217
Cardinalidad de TIPO_PRENDA: 94
Cardinalidad de FCH_CARGA: 225
Cardinalidad de FECHA_EMPENO_OK: 9
Cardinalidad de FECHA_HORA_MOV_OK: 11
Cardinalidad de FECHA_COMERCIALIZACION_OK: 13
Cardinalidad de CUSTODIA: 23
Cardinalidad de SALDO_INSOLUTO: 29298
Cardinalidad de COM_ALMACENAJE: 54
Cardinalidad de IVACOM_ALMACENAJE: 8
Cardinalidad de COMPASE_COMERCIALIZACION: 30
Cardinalidad de IVACOM_PASECOM: 41
Cardinalidad de COM_EXHIBICION: 56110
Cardinalidad de IVACOM_EXHIBICION: 9
Cardinalidad de INTERES_DEPRECUP: 7
Cardinalidad de IVAINT_DEPRECUP: 56
Cardinalidad de FECHA_MAX_DEP_RECUP: 329
Cardinalidad de FECHA_CARGA: 220
```

Interpretaciones

Tras el análisis exploratorio, se identificó una amplia variedad de variables categóricas con niveles de cardinalidad alta, media y baja. Destacan variables como TIPO_PRENDA, CANAL y PRODUCTO, las cuales presentan una estructura adecuada y serán clave para los procesos de visualización, segmentación y modelado en etapas posteriores.

En contraste, variables como ESTADO_SUCURSAL presentan una cardinalidad inesperadamente alta (163 categorías). Este comportamiento sugiere posibles problemas de formato en el archivo fuente, ya que conceptualmente esta variable debería contener un máximo de 32 valores únicos (los 31 estados de México más la Ciudad de México). Durante el análisis se detectaron filas que rompen la estructura del DataFrame, desplazando los valores de columna, lo cual confirma la necesidad de un proceso de preprocesamiento y limpieza estructural más riguroso.

1.7 Gráficos (histogramas)



Interpretaciones

Las variables anteriormente analizadas, se transformarán en numéricas (int o float) dependiendo de su naturaleza en la etapa de preprocesamiento. Lo que podemos observar es que la mayoría tienen un sesgo a la derecha, por lo que se tendrían que normalizar con un escalamiento para obtener una distribución más normal que ayude al modelo de machine learning.