UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MAESTRÍA EN CIENCAS DE DATOS

APRENDIZAJE AUTOMÁTICO

MAESTRO: JOSÉ ALBERTO BENAVIDES VÁZQUEZ

TAREA #2

ALUMNO: EDWIN MARTÍN ROMERO SILVA

MATRÍCULA: 1731276

Tarea 2

Para esta tarea utilicé la función read_csv para mi base de datos, la cual está relacionada con crédito, y le di algo de limpieza previo a los análisis que realizaremos en esta materia.

Esto es necesario para facilitar el análisis y además me ayuda a conocer más los datos con los que trabajaré durante el tetra.

Paso 1) Leer base de datos en formato csv, especificando el tipo de dato.

Obtuve esta base de datos de Kaggle y estaba dividida en 2 partes: TRAIN y TEST. Intenté leer ambas y unirlas, pero noté que la variable Target de la base TEST se encontraba vacía. Por lo que tuve que desecharla y usaré únicamente la base TRAIN (Aunque probablemente esta la tenga que dividir en 2 nuevamente).

```
df = pd.read_csv(dir_input + 'train.csv', usecols = columnas, dtype = tipos_de_datos)
```

Paso 2) Simplificar variables categóricas.

El objetivo de esta sección es agrupar los posibles valores de las variables tipo string, en un número reducido de categorías (siempre y cuando sean similares). Utilizaré GROUP BY y la función de agregación COUNT para este paso.

```
categoricas.groupby('Application Type')[['Application Type']].count()
categoricas.groupby('Verification Status')[['Verification Status']].count()
categoricas.groupby('Employment Duration')[['Employment Duration']].count()
categoricas.groupby('RELACION_VIVIENDA')[['RELACION_VIVIENDA']].count()
categoricas.groupby('Sub Grade')[['Sub Grade']].count()
categoricas.groupby('Sub Grade2')[['Sub Grade2']].count()
categoricas.groupby('Term')[['Term']].count()
categoricas.groupby('Payment Plan')[['Payment Plan']].count()
categoricas.groupby('Loan Title')[['Employment Duration']].count().sort_values('Employment Duration',
categoricas.groupby('Loan Title2')[['Loan Title2']].count()
```

La variable 'Loan Title' tiene más de 100 categorías, son demasiadas, por lo que cree la variable 'Loan Title 2', la cual esta simplificada en 3 categorías:

- Credit Card Refinancing
- Debt Consolidation
- Otros

	Loan Title2
Loan Title2	
CREDIT CARD	31173
DEBT CONSOLIDATION	28638
OTHER	7652

En mi experiencia en Crédito, los comportamientos de impago de los clientes difieren mucho del tipo de crédito, por lo que probablemente tendré que dividir mi análisis en 2: Credit Card y Debt Consolidation.

La variable Sub Grade tiene 35 categorías, por lo que cree la variable Sub Grade 2, la cual esta simplificada en solo 7:

Sub Grade2	Sub Grade2
SUB_GRADE_A	10690
SUB_GRADE_B	18313
SUB_GRADE_C	16250
SUB_GRADE_D	11093
SUB_GRADE_E	6251
SUB_GRADE_F	3372
SUB_GRADE_G	1494

Las otras variables categóricas tienen una cantidad pequeña de categorías, por lo que no hice modificaciones.

Durante esta sección también noté que la variable 'Employment Duration' realmente indicaba si la casa del solicitante de crédito es rentada o propia, por lo que le cambié el nombre a 'RELACION_VIVIENDA'.

Paso 3) Crear variables nuevas y eliminar en caso de ser necesario

Con las variables que ofrece la base, podemos crear unas nuevas. Cree 3 relacionadas con los montos del préstamo, con la hipótesis de que cuando un cliente pide un préstamo muy alto en comparación con el que realmente puede pagar, es un mal cliente.

Loan Amount: Monto solicitado.
Funded Amount: Monto financiado.
Funded Amount Investor: Monto autorizado.

```
df['Dif_monto_sol_monto_fin'] = df['Loan Amount']-df['Funded Amount']
df['Dif_monto_sol_monto_aut'] = df['Loan Amount']-df['Funded Amount Investor']
df['Dif_monto_fin_monto_aut'] = df['Funded Amount']-df['Funded Amount Investor']
```

También eliminé 3 variables, 2 de ellas porque son IDs, y otra porque es una constante para todos los registros, por lo que no tiene poder predictivo.

```
df = df.drop(columns = ['Payment Plan', 'ID', 'Batch Enrolled'])
```

Paso 4) Filtrar, hacer subconjuntos de datos y agregar índices.

Con base en lo comentado en el Paso 2, segmenté mi base en 2: DEBT CONSOLIDATION y CREDIT CARD. También reinicié el índice, ya que al usar los filtros los índices dejan de ser consecutivos.

```
df_cc = df[df['Loan Title2'] == 'CREDIT CARD'].reset_index(drop = True)
df_dc = df[df['Loan Title2'] == 'DEBT CONSOLIDATION'].reset_index(drop = True)
```

Para finalizar, exporté ambas bases en formato csv.

```
df_cc.to_csv('df_cc.csv')
df_dc.to_csv('df_dc.csv')
```