Avance del proyecto semana 5

Alvaro Pabón
Jorge López
Maria Ana Ortiz

1. **Estadísticas descriptivas de los datos “limpios”, acompañado de una descripción de las mismas y del proceso de limpieza de datos. Recuerden describir el tipo de variables con las que cuentan, sus estadísticos descriptivos, qué hicieron con los datos faltantes, qué hicieron con las variables categóricas, etc. Incluyan en su descripción una comparación detallada entre la base de datos original y la base de datos final.**

   Proceso de limpieza: Tenemos 8925 variables perdidas en sueldo básico, con una primera iteración eliminamos las columnas de cédula y sueldo básico, posteriormente eliminamos las 284 filas de rango_de_edad que contiene valores perdidos. Adicionalmente las variables categóricas género y rango de edad las convertimos a dummies. Y finalmente estandarizamos los datos. Como resultado obtenemos un dataframe con 30482 filas y 43 columnas. Los datos iniciales tenían 30766 filas y 37 columnas.

```
Int64Index: 30482 entries, 0 to 30765
Data columns (total 43 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   vlr_cuota                             30482 non-null  int64
 1   vlr_libranza                          30482 non-null  int64
 2   puntaje                               30482 non-null  int64
 3   numero_obligaciones_activas           30482 non-null  int64
 4   cartera_bancaria                      30482 non-null  int64
 5   cartera_bancaria_numero_creditos      30482 non-null  int64
 6   cartera_bancaria_valor_inicial        30482 non-null  int64
 7   cartera_bancaria_valor_saldo          30482 non-null  int64
 8   cartera_bancaria_valor_cuotas         30482 non-null  int64
 9   cartera_bancaria_valor_mora           30482 non-null  int64
 10  informacion_tdc_numero_tdc            30482 non-null  int64
 11  informacion_tdc_valor__cupos          30482 non-null  int64
 12  informacion_tdc_valor__utilizado      30482 non-null  int64
 13  informacion_tdc_valor__cuotas         30482 non-null  int64
 14  informacion_tdc_valor_mora            30482 non-null  int64
 15  sector__real_numero_creditos          30482 non-null  int64
 16  sector__real_valor_inicial            30482 non-null  int64
 17  sector__real_valor_saldo              30482 non-null  int64
 18  sector__real_valor_cuotas             30482 non-null  int64
 19  sector__real_valor_mora               30482 non-null  int64
 20  sector__cooperativo_numero_creditos   30482 non-null  int64
 21  sector__cooperativo_valor_inicial     30482 non-null  int64
 22  sector__cooperativo_valor_saldo       30482 non-null  int64
 23  sector__cooperativo_valor_cuotas      30482 non-null  int64
 24  sector__cooperativo_valor_mora        30482 non-null  int64
 25  obligaciones_mora_30                  30482 non-null  int64
 26  obligaciones_mora_60                  30482 non-null  int64
 27  obligaciones_mora_90                  30482 non-null  int64
 28  obligaciones_mora_120                 30482 non-null  int64
 29  ultimo_ano_moras_30                   30482 non-null  int64
 30  ultimo_ano_moras_60                   30482 non-null  int64
 31  ultimo_ano_moras_90                   30482 non-null  int64
 32  ultimo_ano_moras_120                  30482 non-null  int64
 33  genero_F                              30482 non-null  uint8
 34  genero_M                              30482 non-null  uint8
 35  genero_sin info                       30482 non-null  uint8
 36  rango_de_edad_18-21                   30482 non-null  uint8
 37  rango_de_edad_22-28                   30482 non-null  uint8
 38  rango_de_edad_29-35                   30482 non-null  uint8
 39  rango_de_edad_36-45                   30482 non-null  uint8
 40  rango_de_edad_46-55                   30482 non-null  uint8
 41  rango_de_edad_56-65                   30482 non-null  uint8
 42  rango_de_edad_66+                     30482 non-null  uint8
dtypes: int64(33), uint8(10)
```

2. **Basándose en la retroalimentación que recibieron de su propuesta, describan en detalle el algoritmo que plantearon utilizar.**

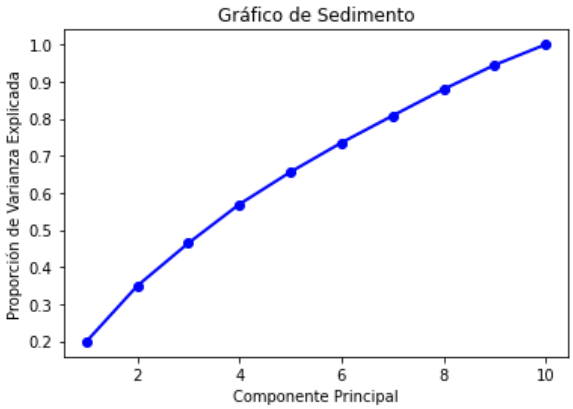
   Utilizaremos el algoritmo de PCA para reducción de dimensiones, tenemos 43 variables después del proceso de transformación.

3. **Haciendo uso de los datos limpios, implementen el algoritmo propuesto y presente los resultados preliminares utilizando tablas y/o visualizaciones. Describan los criterios de decisión que tomaron para obtener y elegir los resultados expuestos. Por ejemplo: si utilizaron PCA, deben describir qué criterio(s) utilizaron para escoger el número de componentes principales.**

```
x = StandardScaler().fit_transform(datos3)
pca = PCA(n_components=10)
principalComponents = pca.fit_transform(x)
principalDf = pd.DataFrame(data = principalComponents,
                           columns = ['PC1', 'PC2','PC3','PC4','PC5','PC6','PC7','PC8','PC9'
                           index = datos3.index)
principalDf
```

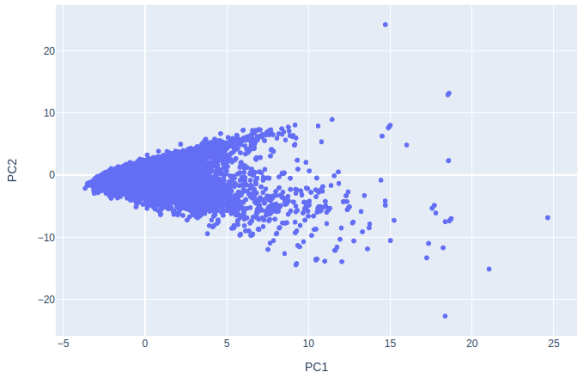
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
2	-1.390839	-0.749603	-0.329395	-0.428545	0.586823	-0.896036	-1.437463	1.614943	0.019080	2.377108
3	-1.768795	0.333966	-0.477290	0.220181	0.053834	-0.393847	-0.026658	0.309380	-1.466823	0.464486
4	-1.634599	0.205053	-0.030770	-1.106741	0.528095	0.298266	-0.639572	-1.114514	0.385124	0.289364
5	-0.092650	1.276212	-0.657859	0.543447	-0.615390	-0.800367	0.845186	0.364900	-1.540274	0.881928
6	0.561952	-2.227831	0.242249	-0.504037	-0.027209	-1.401389	-0.005204	0.853172	-0.314194	-0.565141
...	...	...	...	...	...	...	...	...	...	...
30761	-1.060371	0.769789	-0.569759	0.147931	-0.556592	-0.342695	-0.751745	0.768033	-1.263792	0.770024
30762	-1.606357	-0.863034	-0.306689	-1.294331	1.584550	0.827747	-2.532814	-0.220857	1.009210	1.922752
30763	-0.469165	0.188069	-0.370512	-0.463110	-0.089738	0.156900	-1.767821	-0.426848	2.070731	-0.583839
30764	1.628688	-1.986806	-0.195026	-6.063718	9.594451	-3.004432	-2.713031	6.432162	7.492095	15.379904
30765	1.051788	2.390759	-1.275379	0.268689	-0.327910	-0.130797	0.165048	0.461145	-1.132384	1.110244

21819 rows × 10 columns



array([0.19878882, 0.34939532, 0.46446475, 0.56929857, 0.65621168, 0.73525471, 0.80784004, 0.87925023, 0.94465004, 1.])

Inicialmente vamos a utilizar 9 componentes ya que decidimos tener en cuenta hasta que llegue a 90% aprox.



Como un primer analisis podemos ver las variables más importantes en el primer y segundo componente:

	Pesos PC1
numero_obligaciones_activas	0.338518
informacion_tdc_valor__utilizado	0.298005
informacion_tdc_numero_tdc	0.280164
informacion_tdc_valor__cupos	0.263671
informacion_tdc_valor__cuotas	0.230725
ultimo_ano_moras_30	0.219928
ultimo_ano_moras_60	0.217851
ultimo_ano_moras_90	0.211996
cartera_bancaria_numero_creditos	0.207874
sector__real_numero_creditos	0.202863

	Pesos PC2
puntaje	0.329310
informacion_tdc_valor__cupos	0.249059
informacion_tdc_numero_tdc	0.233876
sector__real_valor_saldo	0.212539
sector__real_valor_inicial	0.210015
informacion_tdc_valor__utilizado	0.197339
numero_obligaciones_activas	0.197013
cartera_bancaria_numero_creditos	0.122834
sector__cooperativo_valor_saldo	0.095542
sector__cooperativo_valor_inicial	0.094762