ANÁLISIS DE COMPONENTES PRINCIPALES

CLASIFICACIÓN DE EMPRESAS FRAUDULENTAS: UN ESTUDIO DE CASO DE UN AUDITORÍA

YULEIDIS MESA

Clasificación de empresas fraudulentas: un estudio de caso de un Auditoría

Este documento es un estudio de caso de una visita a una empresa de auditoría externa. Se recogieron 777 datos anuales de 46 ciudades y en empresas de 14 sectores diferentes de la economía y su objetivo consiste en construir un modelo de clasificación que pueda predecir si una empresa es fraudulenta sobre la base del riesgo actual e histórico de factores (26 atributos).

Se encontraron dos registros de NaN en la base de datos, por lo que se trabajó finalmente con 775 observaciones.

Componentes principales

PC1

```
variable
TOTAL            0.254699
Score            0.290865
Inherent_Risk    0.267209
Name: PC1, dtype: float64
```

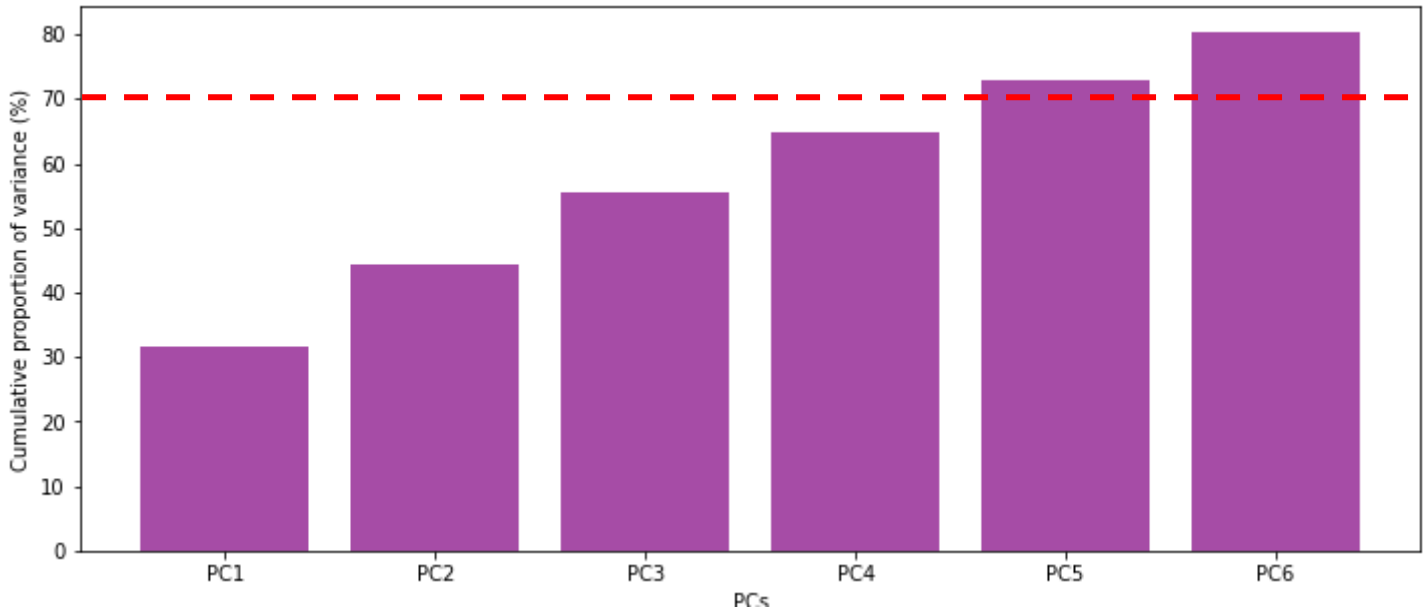
PC2

```
variable
History          0.355749
Risk_F           0.360233
CONTROL_RISK     0.425589
Name: PC2, dtype: float64
```

PC3

```
variable
PARA_B           0.392435
Risk_B           0.392700
TOTAL            0.367227
Audit_Risk       0.366455
Name: PC3, dtype: float64
```

Gráfico Scree (prueba del codo)



Al observar los cambios de forma acumulada vemos que a partir de la quinta componente se explica un poco mas del 70% de la variabilidad de datos.

Regresión

OLS Regression Results

Dep. Variable:	Risk0	**R-squared:**	0.716
Model:	OLS	**Adj. R-squared:**	0.714
Method:	Least Squares	**F-statistic:**	322.3
Date:	Sat, 04 Jun 2022	**Prob (F-statistic):**	6.44e-206
Time:	20:14:44	**Log-Likelihood:**	-57.091
No. Observations:	775	**AIC:**	128.2
Df Residuals:	768	**BIC:**	160.8
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9308	0.038	-24.498	0.000	-1.005	-0.856
Risk_B	-0.0116	0.003	-3.800	0.000	-0.018	-0.006
TOTAL	0.0063	0.002	3.451	0.001	0.003	0.010
Score_MV	0.8401	0.091	9.223	0.000	0.661	1.019
Score	0.2968	0.019	15.591	0.000	0.259	0.334
District_Loss	0.1073	0.008	13.602	0.000	0.092	0.123
History	-0.0106	0.019	-0.562	0.575	-0.047	0.026

Omnibus:	81.497	**Durbin-Watson:**	1.831
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	117.169
Skew:	0.766	**Prob(JB):**	3.61e-26
Kurtosis:	4.131	**Cond. No.**	602.

Conforme a los resultados obtenidos del análisis de componentes principales, de 26 variables, nos quedamos con 6 variables.
Se encontró un R ajustado de 0.716, lo cual indica una buena medida de ajuste, la variable Risk_B tiene una relación inversa con el riesgo de fraude en empresas y las variables TOTAL, Score_MV, Score, District_Loss e History tienen una relación directa con el riesgo.

```
[224] confusion_matrix(Y,pred)/775*100
```

```
array([[60.51612903,  0.12903226,  0.        ],
       [ 5.16129032, 32.51612903,  1.67741935],
       [ 0.        ,  0.        ,  0.        ]])
```

Finalmente, la matriz de confusión muestra que la regresión explica el 93% de las veces el comportamiento de las datos originales