Evaluación de modelo de regresión logística para la predicción de bancarrota en bancos norte americanos

UNIVERSIDAD EAFIT

Presentación del equipo

Stefanny Escobar
Preparación de los datos e investigación

Sara Gallego
Revisión de datos e investigación

Isabella Navarro
Correcciones y conclusiones

https://github.com/StefannyEscobar/Bankruptcy_Prediction

UNIVERSIDAD EAFIT

Problemática

La crisis del 2008 en Estados Unidos causó problemas económicos mundiales, uno de estos siendo que empresas quebraran y muchos inversionistas perdieran su dinero.

Figura 1 periódico crisis [8]

Objetivo general

Evaluar la efectividad del modelo de regresión logística en el fenómeno de la bancarrota en los bancos estadounidenses.

Figura 2 banco[8]

Justificación

- Cerca del 25% de las importaciones colombianas provienen de Estados Unidos.
- El motivo de esta investigación es ayudar a los inversionistas Colombianos al invertir su dinero en bancos estadounidenses.

Figura 3 Bancos americanos [9]

Conceptos básicos

Liquidez:

Capacidad para cumplir con los compromisos a corto y mediano plazo. Primer análisis para conocer si la empresa cuenta con dinero.

Rentabilidad:

Capacidad de una inversión de generar utilidad.

Calidad de credito:

Es la capacidad de poder cumplir con los pagos al momento de adquirirlos.

Eficiencia:

Sacar el máximo provecho de los recursos.

Conceptos básicos

Solvencia:

Capacidad de cumplir con los pagos independiente si es en forma inmediata o posterior.

Productividad:

Calcula cuántos bienes y servicios se han producido por cada factor utilizado.

Razón de aplacamiento:

Es la razón entre el capital de la empresa y las deudas.

Rotación de activos:

Mide el nivel de eficiencia con la que una empresa utiliza sus activos para generar ingresos.

Fig 1 matriz de confusión

Margen operativo:

Cuantifica el porcentaje de ingresos por ventas que la empresa convierte en beneficios, antes de descontar impuestos e intereses.

Matriz de confusión:

Permite visualizar el desempeño de un algoritmo de aprendizaje supervisado.

Figura 4 matriz de confusión [10]

Regresión logística

- Método estadístico
- Permite establecer una relación y obtener una estimación estadística ajustada de probabilidad de la ocurrencia de un evento(Binario)

Matemáticamente

- Tenemos la variable dependiente ***y***, las variables independientes X_i ,los parámetros del modelo W_i,

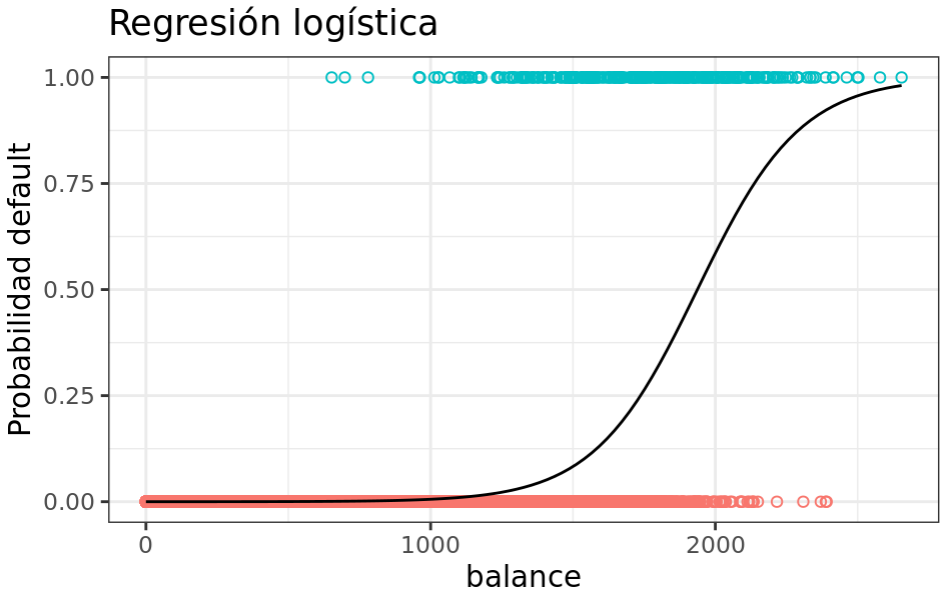


Figura 5 Interpretación modelo [11]

$$S(w^t x) = \frac{1}{1 + e^{-w^t x}}$$

¿Cómo funciona internamente?

- Tenemos la variable dependiente ***y***, con 0 y 1 (dicotómico)
- El vector de variables independientes X_i
- El vector transpuesto de parámetros W_i
- Combinación lineal entre los parámetros y variables características

$$F(w^t x) = 1 - \frac{1}{1 + e^{w^t x}}$$

- Luego las probabilidades del modelo estan dadas por:

$$p = \frac{e^{w^t x}}{1 + e^{w^t x}}$$

X: v.a quebrado o no quebrado

Función Sigmoide

- La predicciones de este modelo vienen calibradas entre 0 y 1, gracias a la función sigmoide.

$$S(t) = \frac{1}{1 + e^{-t}}$$

- La combinación lineal en la función sigmoide resultante es:

$$S(w^t x) = \frac{1}{1 + e^{-w^t x}}$$

- Esta permite que la probabilidad este correctamente calibrada.

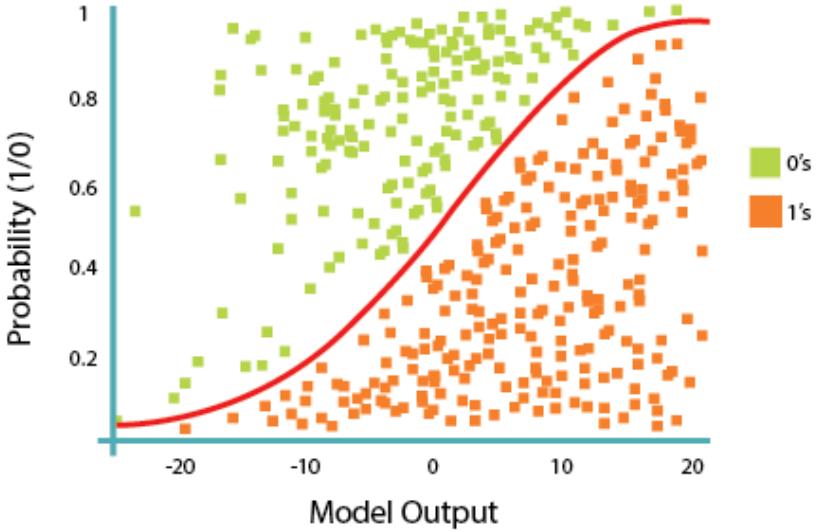


Figura 6 Modelo Calibrado [12]

$$1 \text{ sis } (w^t x) \geq 0,5$$
$$0 \text{ sis } (w^t x) < 0,5$$

Variables e hipotésis

Neurona artificial:

W se calcula atraves del entrenamiento

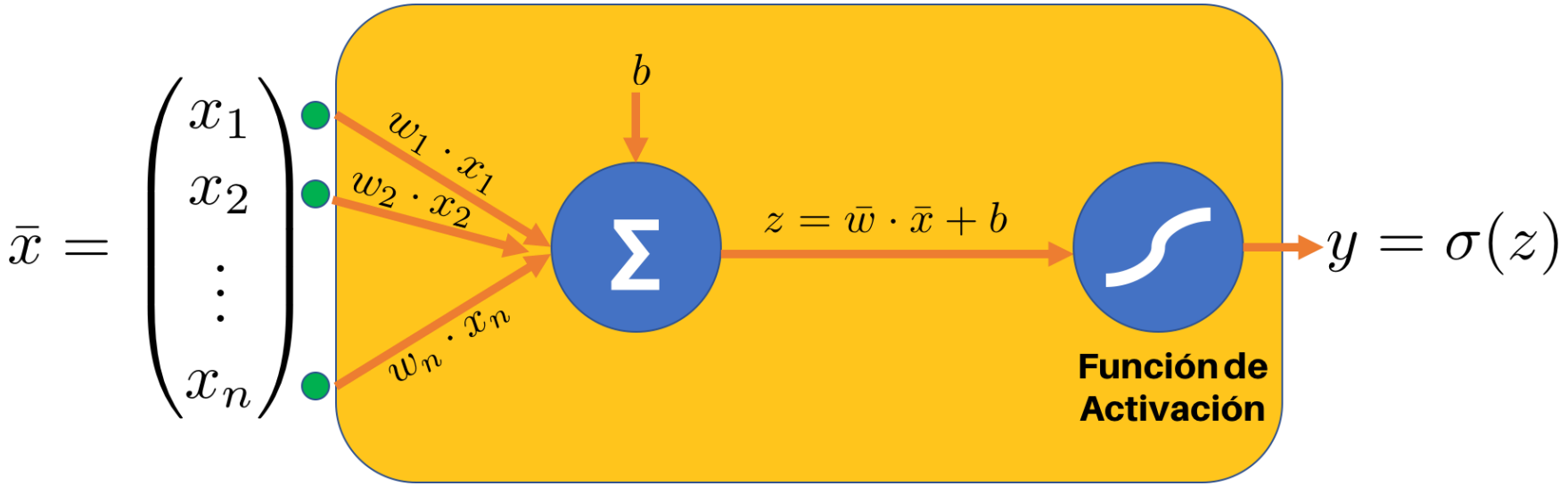


Figura 7 Regresión lógistica para la clasificación [13]

Metodología

- Los datos fueron tomados del laboratorio financiero y Kaggle
- 120 bancos no quebrados y 71 quebrados
- Muestra inicial 200 bancos de 2007 a 2017
- Modelo de regresion logística binaria, para analizar los datos se uso matrices de confusión, precisión, recall, F1 score y support

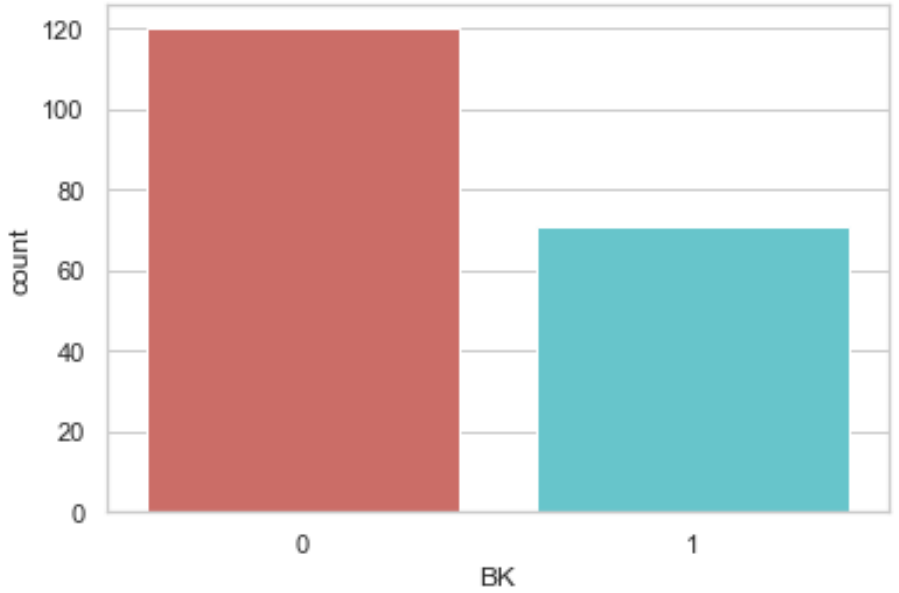


Figura 8 Plot de datos

Indicadores financieros

EPS (Earnings per share)

- Ganancia generada por cada acción de la compañía.

$$= \frac{Profit - Dividends}{\#\ of\ shares\ outstandimg}$$

Market Book Ratio

- Evaluar si la compañía está infravalorada o sobrevalorada.

$$= \frac{Book\ Value}{Market\ Value}$$

- Ratio > 1 undervalued
- Ratio < 1 overvalued

Tobin´s Q

- Indicador básico de rentabilidad y de beneficios a largo plazo.

$$= \frac{Equitity\ Market\ Value}{Equitity\ Book\ Value}$$

- Ratio > 1 overvalued
- Ratio < 1 undervalued

Return on Equity

- La eficiencia que tiene una compañía de generar ganancias.

$$= \frac{Profit}{Equity}$$

Variables e hipotésis

- Para las variables X_i Tomamos indicadores financieros como: EPS, Liquidity, profitability…

- Se implementó el modelo con herramientas de machine learning, librerias como imblearn, sklearn, pandas… de esta manera el modelo internamente identifica pesos e importancia de las categorías.

BK	Data Year - Fiscal	Tobin's Q	EPS	Liquidity	Profitability	Productivity	Leverage Ratio	Asset Turnover	Operational Margin	Return on Equity	Market Book Ratio	Assets Growth	Sales Growth	Employee Growth
0	2007.0	2.545333	1.884333	0.145250	-0.836250	0.069583	0.647833	1.103917	-0.863417	0.012333	1201.467333	1.571833	0.131683	0.059908
1	2010.0	1.772423	-2.652958	-0.257085	-2.940634	-0.254437	-3.371113	1.250690	-2.685085	-44.089324	52.931451	-0.142563	0.445592	0.004056

Fig 7 Promedio de los indicadores de bancarrota y no bancarrota

UNIVERSIDAD EAFIT

Implementación del modelo

Librerias usadas:

Pandas, Numpy, sklearn, matplotlib, seaborn y imblearn

```
import pandas as pd
import numpy as np
from sklearn import preprocessing
import matplotlib.pyplot as plt
plt.rc("font", size=14)
from sklearn.metrics import classification_report, confusion_matrix
import seaborn as sns
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
```

Lectura de datos y análisis general de datos:

```
pgf = pd.read_csv("Bancarrota.finalll.csv")
data= pgf.dropna()
print(data)
print(list(data.columns))
```

```
data['BK'].value_counts()
```

```
0    120
1     71
```

```
count_no_sub = len(data[data['BK']==0])
count_sub = len(data[data['BK']==1])
pct_of_no_sub = count_no_sub/(count_no_sub+count_sub)
print("percentage of healthy bank", pct_of_no_sub*100)
pct_of_sub = count_sub/(count_no_sub+count_sub)
print("percentage of bankruptcy", pct_of_sub*100)
```

```
percentage of healthy bank 62.82722513089005
percentage of bankruptcy 37.17277486910995
```

UNIVERSIDAD EAFIT

Implementación del modelo

Implementación de regresión logística con sklearn:

```
X_train, X_test, y_train, y_test = train_test_split(train.iloc[:,:14],
                                                    train.iloc[:,14] ,test_size=0.20,
                                                    shuffle=True)
```

```
from sklearn.linear_model import LogisticRegression
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
predictions = logmodel.predict(X_test)
```

Evaluación de modelo:

```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.85	0.92	0.88	24
1	0.85	0.73	0.79	15
accuracy			0.85	39
macro avg	0.85	0.82	0.83	39
weighted avg	0.85	0.85	0.84	39

Matriz de confusión:

```
import itertools
cm = confusion_matrix(y_test, predictions)
plt.imshow(cm, cmap=plt.cm.Reds, interpolation='nearest')
plt.colorbar()
plt.title('Matriz de Confusión')
plt.xlabel('Predicción')
plt.ylabel('Real')
thresh = cm.max() / 2.
for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
    plt.text(j, i, format(cm[i, j], 'd'),
             horizontalalignment='center')
plt.show()
```

UNIVERSIDAD EAFIT

Direcciones de trabajo futuras

Empresas

Instituciones

Personas

Figura 9 Bankruptcy [14]

Matriz de Confusión y sus métricas

	precision	recall	f1-score	support
0	0.92	0.96	0.94	25
1	0.92	0.86	0.89	14
accuracy			0.92	39
macro avg	0.92	0.91	0.92	39
weighted avg	0.92	0.92	0.92	39

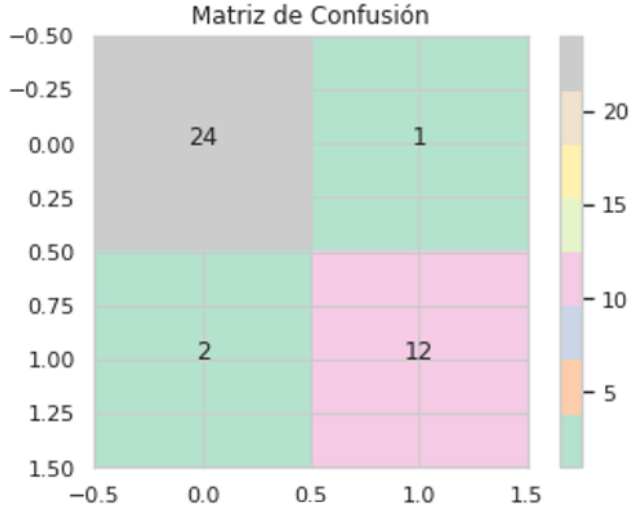


Figura 10 resultados obtenidos

Precisión:

$$= \frac{VP}{VP + FP}$$

Recall:

$$= \frac{VP}{VP + FN}$$

macro avg:

$$= (0.5 * score\ class\ 0) + (0.5 * score\ class\ 1)$$

f1-score:

$$= \frac{2*(precisión*recall)}{precisión+recall}$$

Weighted avg:

$$= (\frac{support}{total} * score\ class\ 0) + (0.5 * score\ class\ 1)$$

Accuracy:

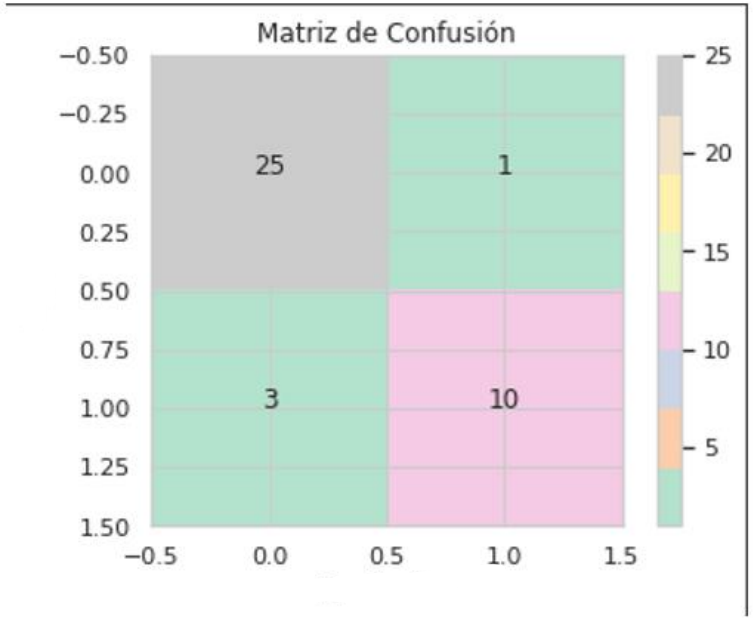
$$= \frac{VP + VN}{VP + FP + FN + VN}$$

Clase 0: bancos sin bancarrota

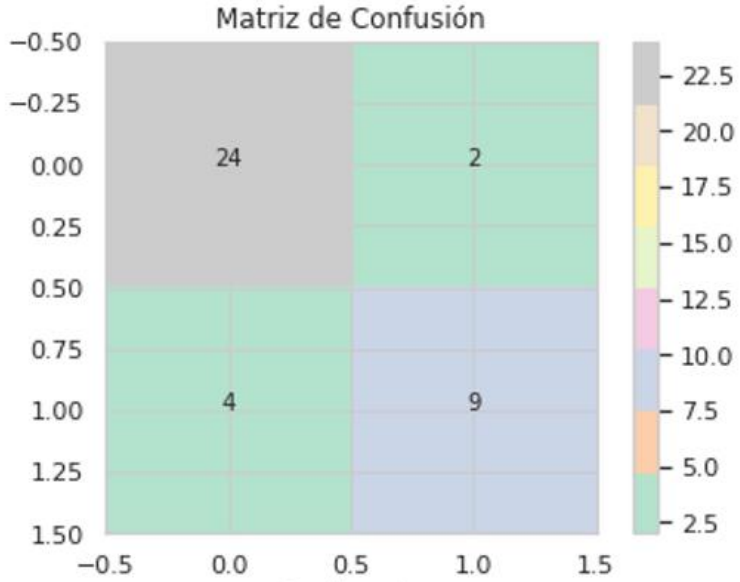
Clase 1: bancos en bancarrota

Conclusiones

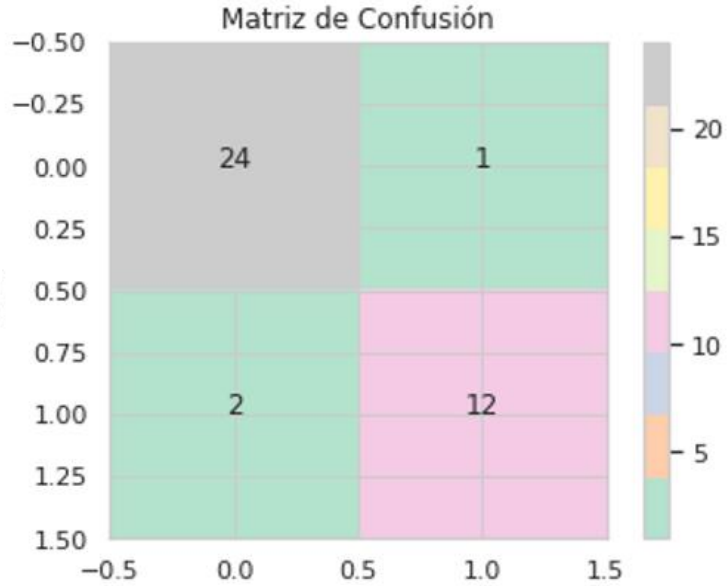
	precision	recall	f1-score	support
0	0.89	0.96	0.93	26
1	0.91	0.77	0.83	13
accuracy			0.90	39
macro avg	0.90	0.87	0.88	39
weighted avg	0.90	0.90	0.90	39



	precision	recall	f1-score	support
0	0.86	0.92	0.89	26
1	0.82	0.69	0.75	13
accuracy			0.85	39
macro avg	0.84	0.81	0.82	39
weighted avg	0.84	0.85	0.84	39



	precision	recall	f1-score	support
0	0.92	0.96	0.94	25
1	0.92	0.86	0.89	14
accuracy			0.92	39
macro avg	0.92	0.91	0.92	39
weighted avg	0.92	0.92	0.92	39



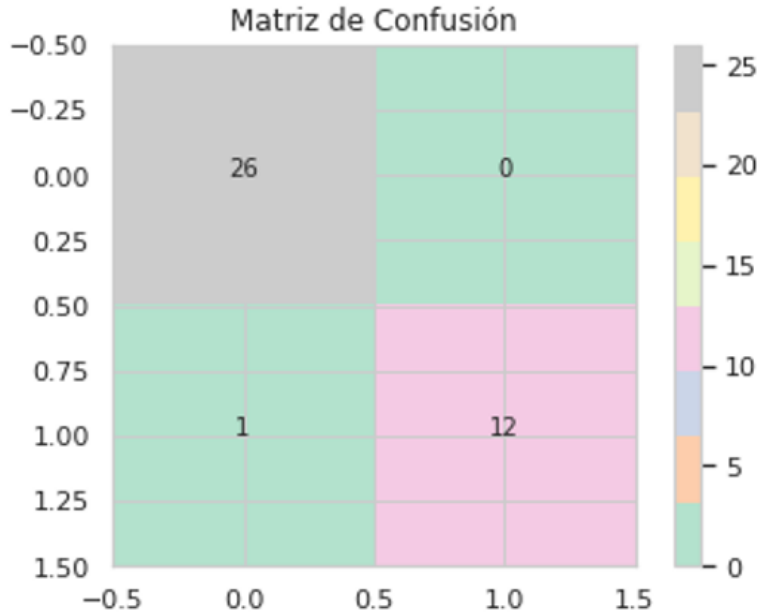
UNIVERSIDAD EAFIT

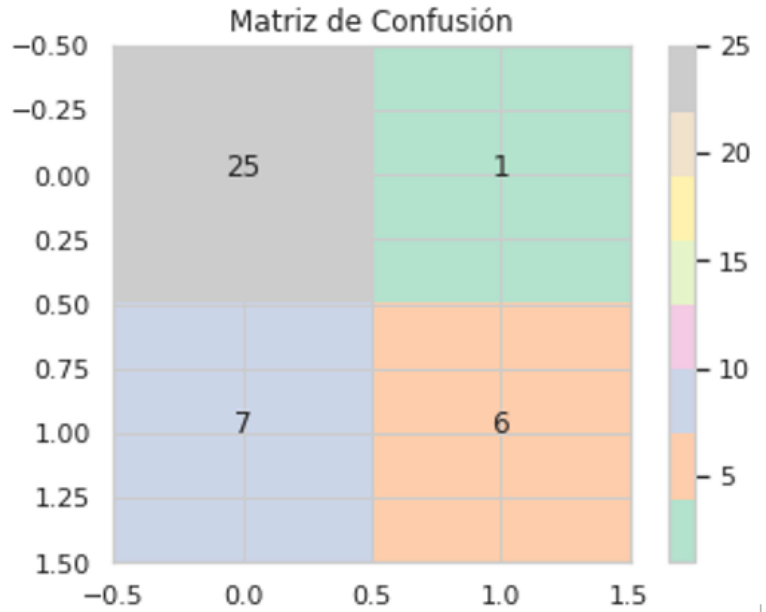
Conclusiones

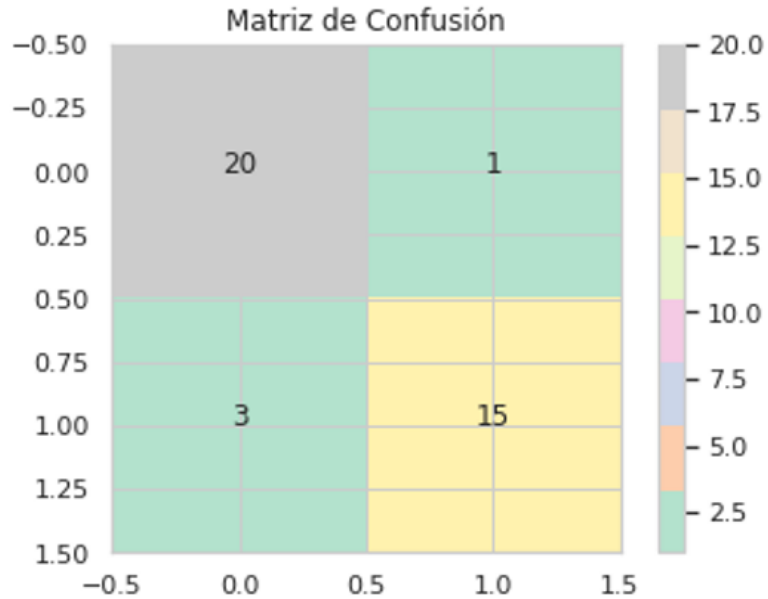
	precision	recall	f1-score	support
0	0.96	1.00	0.98	26
1	1.00	0.92	0.96	13
accuracy			0.97	39
macro avg	0.98	0.96	0.97	39
weighted avg	0.98	0.97	0.97	39

	precision	recall	f1-score	support
0	0.78	0.96	0.86	26
1	0.86	0.46	0.60	13
accuracy			0.79	39
macro avg	0.82	0.71	0.73	39
weighted avg	0.81	0.79	0.77	39

	precision	recall	f1-score	support
0	0.87	0.95	0.91	21
1	0.94	0.83	0.88	18
accuracy			0.90	39
macro avg	0.90	0.89	0.90	39
weighted avg	0.90	0.90	0.90	39







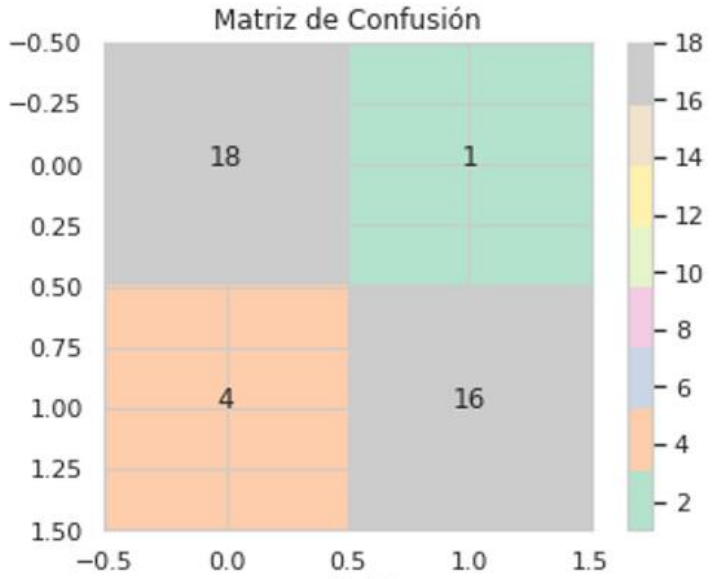
UNIVERSIDAD EAFIT

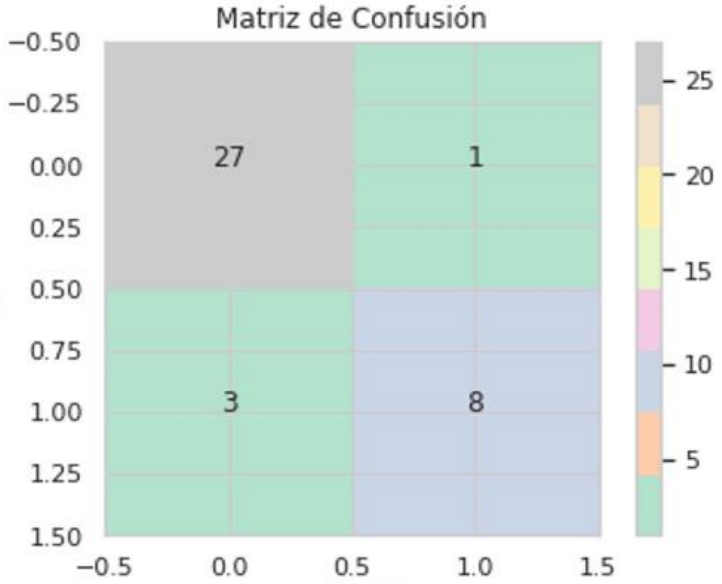
Conclusiones

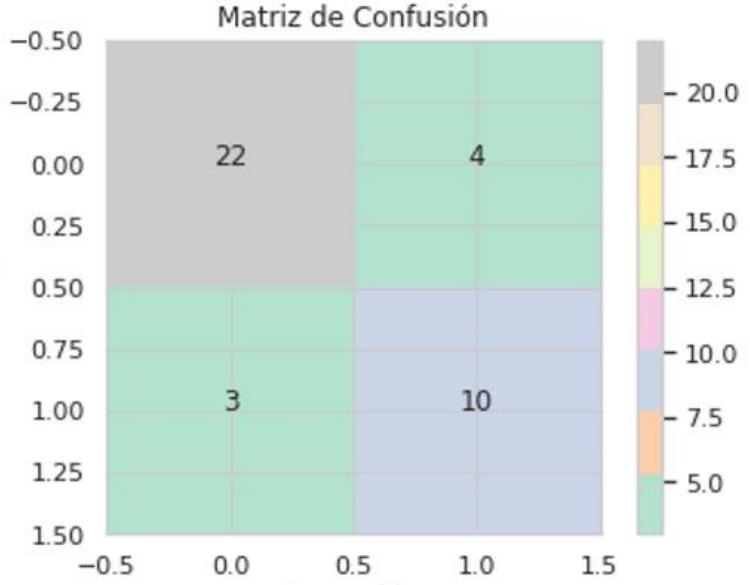
	precision	recall	f1-score	support
0	0.82	0.95	0.88	19
1	0.94	0.80	0.86	20
accuracy			0.87	39
macro avg	0.88	0.87	0.87	39
weighted avg	0.88	0.87	0.87	39

	precision	recall	f1-score	support
0	0.90	0.96	0.93	28
1	0.89	0.73	0.80	11
accuracy			0.90	39
macro avg	0.89	0.85	0.87	39
weighted avg	0.90	0.90	0.89	39

	precision	recall	f1-score	support
0	0.88	0.85	0.86	26
1	0.71	0.77	0.74	13
accuracy			0.82	39
macro avg	0.80	0.81	0.80	39
weighted avg	0.82	0.82	0.82	39







UNIVERSIDAD EAFIT

Resultados

Después de evaluar el modelo 9 veces se obtuvo en promedio (clase 0):

- Una **precisión** de **0.87** indicando que hubo pocos falsos positivos.
- Un **recall** de **0.94** indicando que hubo muy pocos falsos negativos.
- Un **f1-score** de **0.9** indicando que la precisión del modelo es muy buena.

Después de evaluar el modelo 9 veces se obtuvo en promedio (clase 1):

- Una **precisión** de **0.85** indicando que hubo pocos falsos positivos.
- Un **recall** de **0.75** indicando que hubo pocos falsos negativos.
- Un **f1-score** de **0.81** indicando que la precisión del modelo es buena.

Conclusiones

- Como se puede visualizar, los resultados de la clase 0 son mejores a los de la clase 1. Esto se debe a que la clase 0 (bancos sin bancarrota) tiene más datos que la clase 1 (bancos en bancarrota).

- **Nuestro aporte:**

Los datos utilizados en este modelo fueron del 2007-2017 logrando que el modelo sea más eficaz.

- El modelo es muy eficaz prediciendo cuales bancos se van a quebrar y cuales no.
- Se desea que el modelo sea utilizado para predicción de bancarrota en los bancos estadounidenses por los inversionistas Colombianos.

Referencias

- 1 HealthBigData. (2019).La matriz de confusión y sus metricas [fig 3]. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/
- FINANZAS “¿Qué es y cómo se logra la eficiencia financiera?” .13 de marzo del 2018. URL: https://rpp.pe/campanas/contenido-patrocinado/que-es-y-como-se-logra-la-eficiencia-financiera-noticia-1110108
- 2 Mª Visitación García Jiménez, Jesús Mª Alvarado Izquierdo y Amelia Jiménez Blanco “La predicción del rendimiento académico: regresión lineal versus regresión logística” 2000. Vol. 12, Supl. nº 2, pp. 248-252.
- 3 Juan Camilo Vega, Edgar Guillermo Rodríguez Díaz, Alexandra Montoya R. “Metodología de evaluación del clima organizacional a través de un modelo de regresión logística para una universidad en Bogotá, Colombia” Revista CIFE: Lecturas de Economía Social, ISSN-e 2248-4914, Vol. 14, Nº. 21, 2012, págs. 63-88.
- 4 Clavijo, Maria A V. 2022. “Regresión Logística Robusta Para La Clasificación de Residuos Sólidos.” OSF. April 1. doi:10.17605/OSF.IO/CW6UP.
- “Modelo de regresión logística para estimar la dependencia según la escala de Lawton y Brody.”.Septiembre (2010). Vol. 36. Núm. 7
- 5 Barrios. I."La matriz de confusión y sus métricas"en 2019 URL: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas
- 6Rodrigo 2016."Regresión logística simple y múltiple" el URL : https://www.cienciadedatos.net/\\documentos/27_regresion_logistica_simple_y_multiple.
- 7Programming foundation. I."Module 4 \\Logistic regression". URL:https://learn.theprogrammingfoundation.org/gett\\ing_started/intro_data_science/module4
- 8Bankruptcy [Fig 9]. https://www.thebluediamondgallery.com/legal/images/bankruptcy.jpg
- 9Miguel Sotaquirá. Codificandobits. (2020).La Neurona Artificial y la Regresión Logística
- Investopedia URL:https://www.investopedia.com/1
- [10]https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/
- [11]https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple
- [12] https://www.statdeveloper.com/regresion-logistica-en-python/
- [13][fig 8].https://www.codificandobits.com/blog/regresion-logistica-y-neurona-artificial/
- [14[Bancos americanosFig 3 https://www.eleconomista.es/mercados-cotizaciones/noticias/11299016/06/21/Los-bancos-estadounidenses-pagaran-mas-de-2000-millones-de-dolares-mas-en-dividendos-trimestrales.html

UNIVERSIDAD EAFIT

¡GRACIAS!

- El profesor Henry Laniado por su apoyo en el Proyecto.
- Monitores del laboratorio financiero que ayudaron a obtener los datos.
- La directora de Simat por su ayuda.

Todos los autores agradecen a la Vicerrectoría de Descubrimiento y Creación de la Universidad EAFIT por su apoyo en esta investigación.