Uso de las técnicas *DownSampling* y *UpSampling* para abordar el desequilibrio de datos en la predicción de personas propensas a sufrir accidentes cerebrovasculares

Use of *DownSampling* and *UpSampling* techniques to address data imbalance in predicting stroke-prone individuals

Nelson del Castillo Collazo
IIMAS, UNAM, Ciudad de México, México
nelson.delcastillo@iimas.unam.mx
ORCID: 0000-0002-4187-5511

Juan Antonio Contreras Arvizu
IIMAS, UNAM, Ciudad de México, México
antonio.contreras@iimas.unam.mx
ORCID: 0000-0003-1375-5499

Adalberto Joel Durán Ortega
IIMAS, UNAM, Ciudad de México, México
joel.duran@iimas.unam.mx
ORCID: 0000-0003-4272-7294

doi: https://doi.org/10.36825/RITI.12.25.007

Recibido: Abril 11, 2024
Aceptado: Junio 11, 2024

Resumen: Se emplean las técnicas de balanceo de datos *DownSampling* y *UpSampling* aplicadas a un conjunto relacionados con individuos propensos a tener un accidente cerebrovascular. El propósito de este trabajo es demostrar la importancia que tiene la aplicación de las técnicas de *DownSamplig* y *UpSampling* cuando nos encontramos con datos que presentan desbalance; haciendo una comparación entre las dos técnicas mencionadas y analizando el comportamiento de las medidas que se calculan en la matriz de confusión cuando se crea el modelo de predicción. El conjunto de datos está compuesto por 4981 registros, de ellos 4773 pertenecen a la clase de los que no han sufrido un accidente cerebrovascular y 248 a la clase que sí lo han tenido. Se encontró que para este conjunto de datos la mejor técnica para tratar el desbalance es la de *UpSampling* con la mayor de sus réplicas y en el momento en que se va a evaluar el modelo es importante, no solo basarse en su *Exactitud,* sino también en otras medidas que resultan de la matriz de confusión, esto para lograr un mejor análisis de los resultados que se obtienen.

Palabras clave: *Desbalance de Datos, DownSampling, UpSampling, Bosques Aleatorios, Aprendizaje de Máquinas.*

Abstract: The *DownSampling* and *UpSampling* data balancing techniques are used applied to a set related to individuals prone to having a stroke. The purpose of this work is to demonstrate the importance of the application of *DownSampling* and *UpSampling* techniques when we find data that present imbalance; making a comparison between the two mentioned techniques and analyzing the behavior of the measures that are calculated in the confusion matrix when the prediction model is created. The data set is composed of 4981 records, of which 4773 belong to the class of those who have not suffered a stroke and 248 to the class that has had one. It was found that for this data set the best technique to treat the imbalance is *UpSampling* with the largest of its replicas and when the model is going to be evaluated it is important not only to base it on its Accuracy, but also on other measures that result from the confusion matrix, this to achieve a better analysis of the results obtained.

Keywords: *Data Imbalance, DownSampling, UpSampling, Random Forest, Machine Learning.*

1. Introducción

Es conocida la importancia que tiene el manejo de los datos para cualquier investigación que esté relacionada con el aprendizaje de máquinas. Crear un modelo de pronóstico empleando datos que no están balanceados puede provocar que el modelo produzca resultados sesgados cuando se utilice, incluso con el empleo de otros datos que tengan iguales variables; por eso la importancia de abordar el desbalance de datos empleando algunas de las técnicas que se describen en diferentes textos [1]. El propósito de este trabajo es demostrar la importancia que tiene la aplicación de las técnicas de *DownSamplig* y *UpSampling* cuando nos encontramos con datos que presentan un desbalance muy marcado; haciendo una comparación entre las dos técnicas mencionadas y analizar el comportamiento de las medidas que se calculan en la matriz de confusión cuando se crea el modelo de predicción. Los datos están divididos en dos *clases*, la *clase No* y la *clase Si*, esta última es la que tiene menor cantidad de elementos. En el uso de la técnica de *DownSampling* se tomaron dos muestras aleatorias distintas para la generación del modelo y para el caso de *UpSampling* se replicaron los datos en diferentes cantidades; también en este caso se tomaron dos muestras aleatorias diferentes. Se obtuvieron resultados donde se puede apreciar la necesidad de no solo apoyarnos en la *Exactitud* (*Accuracy*) del modelo para saber si con él se obtuvo un buen pronóstico sino también en otras medidas que resultan de mucha importancia.

2. Materiales y métodos

Los datos fueron tomados de la página de Kaggle [2] los cuales están relacionados con accidentes cerebrovasculares, estos corresponden a dos categorías de *clases* para predecir si un individuo es propenso a sufrir un ataque de este tipo. La cantidad de datos con las que se trabajó fueron 4981 registros, de ellos 4773 pertenecen a la *clase No (individuos con valores de cero)* y 248 a la *clase Si (individuos con valores de uno)* como se puede ver en la Figura 1, en este caso se puede apreciar a primera vista que existe un marcado desbalance de datos.

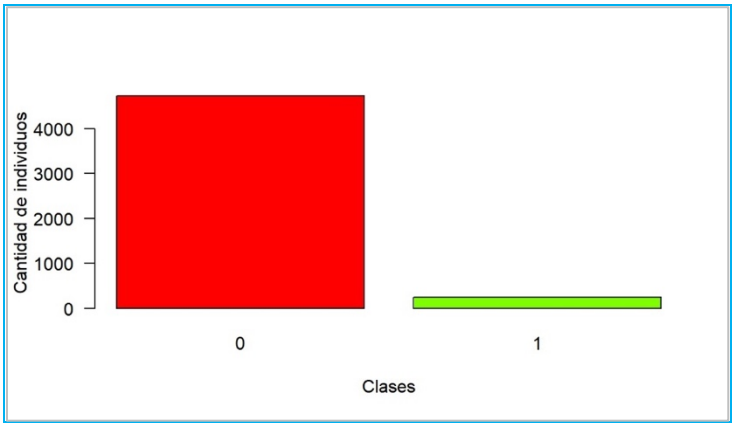

Figura 1. Cantidad de individuos por *clase*. Fuente: Elaborada por el autor.

El conjunto de datos está formado por 11 variables, 10 de ellas independientes y la oncena es la que se busca pronosticar, es decir, la variable dependiente donde se identifica si un individuo puede sufrir o no un ataque cerebrovascular. En la Tabla 1 se puede apreciar las variables que integran este conjunto de datos, incluida la variable *stroke* que es la que se intenta predecir.

Tabla 1. Nombres de las variables que intervienen en el estudio.

No.	Nombre de la variable	Descripción
1	*gender:*	Género
2	*age:*	Edad
3	*hypertension:*	Si sufre de hipertensión
4	*heart_disease:*	Si tiene algún tipo de cardiopatía
5	*ever_married:*	Si ha estado casado alguna vez
6	*work_type:*	Tipo de trabajo
7	*Residence_type:*	Tipo de residencia
8	*avg_glucose_level:*	Nivel promedio de glucosa en sangre
9	*bmi:*	Índice de masa corporal
10	*smoking_status:*	Estado del fumador
11	*stroke:*	Accidente cerebrovascular

Fuente: Elaborada por el autor.

En la Tabla 2 se puede ver como se distribuyen los datos según los tipos de *clases*. Es importante mencionar que la *clase No* se identifica con aquellos que no han sufrido un accidente mientras que en la *clase Si* lo han sufrido. Es importante mencionar que se hicieron para cada conjunto de datos diferentes corridas del método Bosques Aleatorios modificando en cada caso el porciento de datos que se emplearon para el entrenamiento en la creación del modelo, los cuales fueron: 70%, 75% y 80%, el resto de los datos, en cada caso, se empleó para validar el modelo, siguiendo el mismo orden anterior fue del 30%, 25% y 20% respectivamente.

Tabla 2. Distribución de los datos según las *clases Si* y *No* y el porciento de datos empleados en el entrenamiento.

	Training 70%			**Training 75%**			**Training 80%**	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	1419	0	**No**	1183	0	**No**	496	0
Si	0	74	**Si**	0	62	**Si**	0	49

Fuente: Elaborada por el autor.

Para el análisis de los datos y la aplicación del método de clasificación se empleó el lenguaje R en su versión 4.3.0, utilizando la función *randomforest()* que se basa en el algoritmo de bosque aleatorio desarrollado por Breiman y Cutler [3], tanto para clasificación como para regresión, pasándole parámetros fundamentales como: el conjunto de datos de entrenamiento, la creación de quinientos árboles aleatorios e indicando tener en cuenta la importancia de los predictores. La interfaz de desarrollo fue *RStudio* en su versión 2023.12.1.

2.1. Método de clasificación y desbalance de datos

El método de Bosques Aleatorios (*Random Forest*) [4] es una técnica muy empleada en la actualidad cuando se aplican métodos de aprendizaje de máquinas, tanto en clasificación como regresión, pues es un algoritmo bastante sencillo que ocupa poco tiempo de cálculo, generando muy buenos resultados en los pronósticos.

Este método se basa en la generación de bosques de árboles aleatorios, es decir, se generan un grupo de árboles aleatorios, en este caso fue de 500, los cuales se combinan permitiendo obtener una precisión mucho más exacta, por eso el nombre de bosque. Para profundizar en este tema se puede consultar [5], [6] y [7].

En cuanto al desbalance de datos se emplearon dos técnicas: *DonwSampling* y *UpSampling*. La primera de ellas consiste en reducir el tamaño de la *clase* mayor a la misma cantidad de datos que la *clase* de menor, de esta manera se igualan las cantidades y se aplica el método de clasificación deseado. En el segundo caso se procede de manera contraria, se replica la *clase* de menor cantidad de datos hasta alcanzar el tamaño de la *clase* de mayor cantidad y luego se aplica el método de pronóstico [8], [9] y [10].

En nuestro caso para la técnica de *DownSampling* se empleó la *clase Si* que es la de menor cantidad de datos y se tomaron dos muestras aleatorias del mismo tamaño de la *clase No*, cada una de ellas se combinaron con los datos de la *clase Si* y se procesaron empleando diferentes porcientos de datos para el entrenamiento del modelo, los cuales se mencionaron anteriormente.

En el caso de la técnica de *UpSamplin* se fueron aumentando las réplicas de la *clase Si* con el objetivo de ver como varían los resultados con respecto a la cantidad de repeticiones de los datos de esa *clase*. Las réplicas fueron: doble, triple, cuádruple y finalmente una repetición de 18 veces; para cada caso se tomaron dos muestras aleatorias de la *clase No* y se procesaron empleando también los porcientos de datos propuestos para el entrenamiento del modelo.

3. Resultados y discusión

En este caso se hizo el procesamiento de todos los datos sin aplicar ninguna técnica de desbalance de datos con el objetivo de ver el comportamiento del pronóstico. A continuación, la Tabla 3 muestra los resultados obtenidos de la matriz de confusión para la primera muestra, estos resultados se pueden comparar con los obtenidos en la Tabla 2. Asimismo, se presenta la Tabla 4 donde están las diferentes medidas que se obtienen de la matriz de confusión.

Tabla 3. Resultados obtenidos en el pronóstico.

	Training 70%			**Training 75%**			**Training 80%**	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	1417	74	**No**	1182	62	**No**	945	49
Si	2	0	**Si**	1	0	**Si**	1	0

Fuente: Elaborada por el autor.

Tabla 4. Medidas que se obtienen de la Matriz de Confusión.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	94.91	94.94	94.97
Sensibilidad (%):	99.86	99.92	99.89
Especificidad (%):	0.00	0.00	0.00
VP+ (%):	95.04	95.02	95.07
VP- (%):	0.00	0.00	0.00

Fuente: Elaborada por el autor.

Se puede apreciar en la Tabla 3 que en el caso de la *clase No* es muy bien pronosticada en los tres casos, pero la *clase Si* que es la que nos indica la posibilidad de que individuo pueda sufrir un accidente cerebrovascular en todos los casos son falsos negativos, es decir, el método es incapaz de predecir ningún individuo con la posibilidad de tener un accidente cerebrovascular aunque los resultados de la Tabla 4 indican un porciento alto en la Exactitud del modelo (94.91%, 94.94% y 94.97% respectivamente), esta medida indica que si se toma un nuevo conjunto de datos (con las mismas variables) que se desee predecir esas serían las probabilidades de que la predicción fuera correcta.

Se toma la *clase No* como la *clase positiva* y se tiene en cuenta que la sensibilidad es el porciento de valores positivos que son clasificados como positivos, la especificidad es el porciento de negativos que son clasificados como

negativos, los valores de predicción positivos (VP+) indican la probabilidad de que un valor sea positivo si resultó positivo en la predicción y los valores de predicción negativos (VP-) indican la probabilidad de que un valor sea negativo si resultó negativo en la predicción [11].

A partir de lo anterior se puede ver que la especificidad para los tres casos es cero, lo que nos indica que la probabilidad de predecir un valor de la *clase Si* (*clase negativa*) es prácticamente cero. Evidentemente esto está dado por el desbalance de datos, pues la *clase No* supera en cantidad de datos cerca de 19 veces a la *clase Si* trayendo como consecuencia que el modelo sea incapaz de hacer un buen pronóstico para conocer si un individuo puede tener un accidente cerebrovascular.

3.1. DownSampling

Una vez seleccionado de forma aleatoria el conjunto de datos de la *clase No* para las dos muestras, cada una de ellas se combinó con los datos de la *clase Si* de menara independiente. Para la muestra uno se obtienen los resultados que se muestran en las Tablas 5, 6 y 7.

Tabla 5. Distribución de los datos según las clases *Si* y *No*.

	Training 70%			**Training 75%**			**Training 80%**	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	74	0	**No**	62	0	**No**	49	0
Si	0	74	**Si**	0	62	**Si**	0	49

Fuente: Elaborada por el autor.

Tabla 6. Resultados obtenidos en el pronóstico en la primera muestra.

	Training 70%			**Training 75%**			**Training 80%**	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	58	17	**No**	47	12	**No**	37	9
Si	16	57	**Si**	15	50	**Si**	12	40

Fuente: Elaborada por el autor.

Tabla 7. Medidas que se obtienen de la Matriz de Confusión en la primera muestra.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	77.70	78.23	78.57
Sensibilidad (%):	78.38	75.81	75.51
Especificidad (%):	77.03	80.65	81.63
VP+ (%):	77.33	79.66	80.43
VP- (%):	78.08	76.92	76.92

Fuente: Elaborada por el autor.

En este caso baja considerablemente el valor de la Exactitud del modelo (77.70%, 78.23% y 78.57%) pero se puede apreciar que el valor de la especificidad aumenta con respecto al que se obtuvo en la Tabla 4, esto nos indica que a pesar que la Exactitud del modelo es baja comparada con la que se obtuvo en el pronóstico empleando todos los datos el porciento de elementos de la *clase* negativa (*clase Si*) aumenta de cero a 77.03%, 80.65% y 81.63% respectivamente; en este caso los valores de VP- (valores de predicción negativos) los cuales indican la probabilidad de que un valor sea negativo si resultó negativo en la predicción también aumentan entre el 76% y el 78%.

Si consideramos que el objetivo del modelo es predecir los datos iniciales que se presentan en la Tabla 5, los resultados anteriores, aunque no reproducen la tabla con precisión, si se puede observar que ya la *clase Si* tiene menos falsos negativos lo que nos indica hasta este momento que la técnica de *downSampling* en este caso funciona mejor que trabajar con el conjunto de datos total, el cual muestra un gran desbalance de datos. En las Tablas 8 y 9 se presentan

los resultados que se obtienen de la segunda muestra. Se observa que los resultados son muy parecidos a los que se obtuvieron en la muestra uno.

Tabla 8. Resultados obtenidos en el pronóstico en la muestra dos.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	55	16	**No**	48	10	**No**	38	12
Si	19	58	**Si**	14	52	**Si**	11	37

Fuente: Elaborada por el autor.

Tabla 9. Medidas que se obtienen de la Matriz de Confusión en la muestra dos.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	76.35	80.65	76.53
Sensibilidad (%):	74.32	77.42	77.55
Especificidad (%):	78.38	83.87	75.51
VP+ (%):	77.46	82.76	76.00
VP- (%):	75.32	78.79	77.08

Fuente: Elaborada por el autor.

3.2. UpSampling

Se seleccionaron los datos de la *clase No* que se corresponden con el doble, el triple y el cuádruple de la *clase Si*, para estos casos se tomaron dos muestras aleatorias; finalmente se replicó 18 veces la *clase Si* y se tomó una muestra del mismo tamaño de la *clase No*, obteniendo los resultados siguientes:

3.2.1. *UpSampling* - doble

Se reprodujo dos veces la *clase Si*, se tomaron dos muestras del mismo tamaño de la *clase No* y se combinaron de manera independiente con los elementos de la *clase Si*.

En la Tabla 10 se muestran los datos iniciales del comportamiento de las dos clases para los tres conjuntos de datos y en las Tablas 11 y 12 los resultados obtenidos al correr el modelo.

Tabla 10. Distribución de los datos según las clases *Si* y *No* para el *UnSampling* doble.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	148	0	**No**	124	0	**No**	99	0
Si	0	148	**Si**	0	124	**Si**	0	99

Fuente: Elaborada por el autor.

Tabla 11. Resultados obtenidos en el pronóstico en la muestra uno.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	113	8	**No**	94	2	**No**	74	2
Si	35	140	**Si**	30	122	**Si**	25	97

Fuente: Elaborada por el autor.

Tabla 12. Medidas que se obtienen de la Matriz de Confusión en la muestra uno.

Training:	70%	75%	80%
Exactitud (%):	85.47	87.10	86.36
Sensibilidad (%):	76.35	75.81	74.75
Especificidad (%):	94.59	98.39	97.98
VP+ (%):	93.39	97.92	97.37
VP- (%):	80.00	80.26	79.51

Fuente: Elaborada por el autor.

Aquí se aprecia como la Exactitud del modelo aumenta con respecto a las dos muestras de la técnica de *DownSampling*, están en un rango entre el 85% y 87%, la Especificidad aumenta también y se comporta mejor que las obtenidas en las dos muestras del *DownSampling,* estando en un rango que oscila entre el 94% y 98%, el VP- también mejora, aunque modestamente, encontrándose en un rango entre el 79% y 80%.

Para esta muestra el mejor conjunto de datos es donde se emplea el 75% de ellos para entrenar el modelo, siendo la Exactitud del 87.1% y la Especificidad del 98.39%. En el pronóstico solo dos individuos son clasificados como falsos negativos y acertando en el pronóstico de 122 de ellos. Se debe mencionar que en el pronóstico de la *clase No,* en este caso los falsos positivos, representan el 24.19%, mientras que para el conjunto de datos de 70% y 80% fueron del 23.65% y 25.25% respectivamente. En las Tablas 13 y 14 se muestran los resultados obtenidos en la segunda muestra.

Tabla 13. Resultados obtenidos en el pronóstico en la muestra dos.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	108	18	**No**	89	10	**No**	71	2
Si	40	130	**Si**	35	114	**Si**	28	97

Fuente: Elaborada por el autor.

Tabla 14. Medidas que se obtienen de la Matriz de Confusión en la muestra dos.

Training:	70%	75%	80%
Exactitud (%):	80.41	81.85	84.85
Sensibilidad (%):	72.97	71.77	71.72
Especificidad (%):	87.84	91.94	97.98
VP+ (%):	85.71	89.90	97.26
VP- (%):	76.47	76.51	77.60

Fuente: Elaborada por el autor.

En esta muestra, el mejor resultado estuvo donde se emplea el 80% de los datos para entrenar el modelo. La Exactitud y la Especificidad estuvieron por debajo de los resultados obtenidos en la muestra uno, pero aun así, se consideran mejores a los obtenidos en las muestras empleadas en el *DownSampling*. Los falsos positivos representaron el 27.03%, 28.23% y 28.28% para el conjunto de datos que se emplearon en el entrenamiento del modelo (70%, 75% y 80% respectivamente).

3.2.2. *UpSampling* - triple

Se reprodujo tres veces la *clase Si*, se tomaron dos muestras del mismo tamaño de la *clase No* y se combinaron de manera independiente con los elementos de la *clase Si*. En la Tabla 15 se muestran los datos iniciales del comportamiento de las dos clases para los tres conjuntos de datos de entrenamiento y en las Tablas 16 y 17 los resultados obtenidos al correr el modelo.

Tabla 15. Distribución de los datos según las clases *Si* y *No* para el *UpSampling* triple.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	223	0	**No**	186	0	**No**	148	0
Si	0	223	**Si**	0	186	**Si**	0	148

Fuente: Elaborada por el autor.

Tabla 16. Resultados obtenidos en el pronóstico en la muestra uno.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	186	3	**No**	154	0	**No**	124	3
Si	37	220	**Si**	32	186	**Si**	24	145

Fuente: Elaborada por el autor.

Tabla 17. Medidas que se obtienen de la Matriz de Confusión en la muestra uno.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	91.03	91.40	90.88
Sensibilidad (%):	83.41	82.80	83.78
Especificidad (%):	98.65	100	97.97
VP+ (%):	98.41	100	97.64
VP- (%):	85.60	85.32	85.80

Fuente: Elaborada por el autor.

Aquí se aprecia como la Exactitud del modelo aumenta con respecto a la muestra primera de *UpSampling* doble, están en un rango entre 90% y 91%, la Especificidad aumenta también y se comporta mejor que las obtenidas anteriormente, estando en un rango entre el 98% y 100%, el VP- también mejora, siendo el mejor que se ha obtenido hasta el momento.

Para esta muestra el mejor conjunto de datos es donde se emplea también el 75% de ellos para entrenar el modelo, siendo la Exactitud del 91.4% y la Especificidad del 100%. En el pronóstico todos los individuos fueron clasificados correctamente, 186 de 186 individuos que se debían pronosticar. Se debe mencionar que en el pronóstico de la *clase No,* en este caso los falsos positivos, representan el 17.20%, mientras que para el conjunto de datos de 70% y 80% fueron del 16.59% y 16.22% respectivamente. En las Tablas 18 y 19 se muestran los resultados obtenidos en la segunda muestra.

Tabla 18. Resultados obtenidos en el pronóstico en la muestra dos.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	188	6	**No**	156	0	**No**	156	0
Si	35	217	**Si**	30	186	**Si**	30	186

Fuente: Elaborada por el autor.

Tabla 19. Medidas que se obtienen de la Matriz de Confusión en la muestra dos.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	90.81	91.94	92.91
Sensibilidad (%):	84.30	83.87	85.81
Especificidad (%):	97.31	100	100
VP+ (%):	96.91	100	100

VP- (%):	86.11	86.11	87.57

Fuente: Elaborada por el autor.

En esta segunda muestra el mejor resultado estuvo donde se emplea el 80% de los datos para entrenar el modelo. La Exactitud y la Especificidad estuvieron muy parecidos a los resultados obtenidos en la muestra uno. Para la *clase No,* los falsos positivos estuvieron muy parecidos también a los encontrados en la muestra uno a excepción del grupo de entrenamiento del 80% que subió a 20.27%.

3.2.3. *UpSampling* – cuádruple

Se reprodujo cuatro veces la *clase Si*, se tomaron dos muestras del mismo tamaño de la *clase No* y se combinaron de manera independiente con los elementos de la *clase Si*. En la Tabla 20 se muestran los datos iniciales del comportamiento de las dos clases para los tres conjuntos de datos y en las Tablas 21 y 22 los resultados obtenidos al correr el modelo.

Tabla 20. Distribución de los datos según las clases *Si* y *No* para el *UpSampling* cuádruple.

	Training 70%		**Training 75%**		**Training 80%**	
	No	**Si**	**No**	**Si**	**No**	**Si**
No	297	0	248	0	198	0
Si	0	297	0	248	0	198

Fuente: Elaborada por el autor.

Tabla 21. Resultados obtenidos en el pronóstico en la muestra uno.

	Training 70%		**Training 75%**		**Training 80%**	
	No	**Si**	**No**	**Si**	**No**	**Si**
No	247	4	208	0	166	0
Si	50	293	40	248	32	198

Fuente: Elaborada por el autor.

Tabla 22. Medidas que se obtienen de la Matriz de Confusión en la muestra uno.

Training:	**70%**	**75%**	**80%**
Exactitud(%):	90.91	91.94	91.92
Sensibilidad(%):	83.16	83.87	83.84
Especificidad(%):	98.65	100	100
VP+(%):	98.41	100	100
VP-(%):	85.42	86.11	86.09

Fuente: Elaborada por el autor.

Aquí se aprecia como la Exactitud y la Especificidad del modelo se mantiene muy parecida a los resultados obtenidos con la técnica de *UpSampling* triple en ambas muestras, la Exactitud está en un rango entre 90% y casi el 92% y la Especificidad entre el 98% y 100%, el VP- apenas disminuye alrededor de 1%.

Para esta muestra el mejor conjunto de datos es donde se emplea el 75% de ellos para entrenar el modelo, siendo la Exactitud del 91.94% y la Especificidad del 100%. En el pronóstico todos los individuos fueron clasificados correctamente, 248 de 248 individuos que se debían pronosticar. Se debe mencionar que en el pronóstico de la *clase No* representan el 16.13%, mientras que para el conjunto de datos de 70% y 80% fueron del 16.84% y 16.16% respectivamente.

Los resultados obtenidos al cuadruplicar los datos de la *Clase Si* son muy parecidos en general a los obtenido en el caso triplicar esa *clase*. Esto podría significar que seguir replicando la *clase Si* de esta manera los resultados obtenidos irían variando muy poco. En las Tablas 23 y 24 se muestran los resultados obtenidos en la segunda muestra.

Tabla 23. Resultados obtenidos en el pronóstico en la muestra dos.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	252	8	**No**	211	4	**No**	170	0
Si	45	289	**Si**	37	244	**Si**	28	198

Fuente: Elaborada por el autor.

Tabla 24. Medidas que se obtienen de la Matriz de Confusión en la muestra dos.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	91.08	91.73	92.93
Sensibilidad (%):	84.85	85.08	85.86
Especificidad (%):	97.31	98.39	100
VP+ (%):	96.92	98.14	100
VP- (%):	86.53	86.83	87.61

Fuente: Elaborada por el autor.

En esta segunda muestra, el mejor resultado estuvo donde se emplea el 80% de los datos para entrenar el modelo. La Exactitud estuvo algo mayor pero no es representativa la diferencia al igual que la Especificidad que se comportó algo menor para los tres conjuntos de datos. La *clase No* se comportó muy parecida también a los encontrados en la primera muestra a excepción del grupo de entrenamiento del 80%, que bajó al 14.14%, mientras que los valores para el grupo de 70% y 75% fueron del 15.15% y 14.92% respectivamente.

3.2.4. *UpSampling* – replica de 18 veces la clase Si

Viendo los resultados que se obtuvieron en las réplicas anteriores de la *clase Si,* donde los resultados comienzan a mejorar desde la opción doble hasta la cuádruple, se decidió replicar 18 veces esta *clase* con vistas a conocer las diferencias que se podrían encontrar con respecto a las anteriores, viendo el comportamiento de la predicción del modelo y las medidas de la matriz de confusión, conociendo que es posible que se produzca un sobreajuste del modelo [12] y [13]. En este caso solo se seleccionó una sola muestra.

En la Tabla 25 se muestran los datos iniciales del comportamiento de las dos clases para los tres conjuntos de datos y en las Tablas 26 y 27 los resultados obtenidos al correr el modelo.

Tabla 25. Distribución de los datos según las clases *Si* y *No* para el *UpSampling* de 18 réplicas.

	Training 70%			Training 75%			Training 80%	
	No	**Si**		**No**	**Si**		**No**	**Si**
No	1339	0	**No**	1116	0	**No**	892	0
Si	0	1339	**Si**	0	1116	**Si**	0	892

Fuente: Elaborada por el autor.

Tabla 26. Resultados obtenidos en el pronóstico.

Training 70%	Training 75%	Training 80%

	No	Si
No	1289	0
Si	50	1339

	No	Si
No	1078	0
Si	38	1116

	No	Si
No	865	0
Si	27	892

Fuente: Elaborada por el autor.

Tabla 27. Medidas que se obtienen de la Matriz de Confusión.

Training:	**70%**	**75%**	**80%**
Exactitud (%):	98.13	98.3	98.49
Sensibilidad (%):	96.27	96.59	96.97
Especificidad (%):	100	100	100
VP+ (%):	100	100	100
VP- (%):	96.40	96.71	97.06

Fuente: Elaborada por el autor.

Aquí se aprecia como la Exactitud y la Especificidad del modelo aumentan de manera considerable con respecto a las réplicas anteriores, estando en el rango del 98% en los tres conjuntos de datos. La especificidad está al máximo y el VP- está en el orden del 96%, el cual es mayor también a las réplicas anteriores. La Sensibilidad aumenta mucho estando en el rango también del 96%, recordemos que la Sensibilidad es el porciento de valores positivos que son clasificados como positivos y VP+ indican la probabilidad de que un valor sea positivo si resultó positivo en la predicción, en este caso el VP+ es del 100% en todos los conjuntos de datos.

Para esta muestra el mejor conjunto de datos es donde se emplea el 80% de ellos para entrenar el modelo, siendo la Exactitud del 98.49% y la Especificidad del 100% y la Sensibilidad es del 96.47%. En el pronóstico todos los individuos fueron clasificados correctamente, 892 de 892 individuos que se debían pronosticar. Se debe mencionar que en el pronóstico de la *clase No,* en este caso los falsos positivos, representan solo el 3.03%, mientras que para el conjunto de datos de 70% y 75% fueron del 3.73% y 3.41% respectivamente, los más bajos hasta el momento exceptuando los resultados del modelo sin emplear ninguna técnica para mitigar el desbalance de datos.

Los resultados obtenidos son los mejores en cualquiera de las técnicas empleados y de las réplicas realizadas. A continuación, se muestra la Tabla 28 donde se puede apreciar cómo variaron los resultados en cada caso, seleccionando solo el mejor resultado por prueba realizada.

Tabla 28. Variación de los mejores resultados en las pruebas realizadas.

Medidas	***A***	***B***	***C***	***D***	***E***	***F***
Training:	80%	80%	75%	75%	80%	80%
Exactitud (%):	94.97	80.65	87.10	91.40	92.93	98.49
Sensibilidad (%):	99.89	77.42	75.81	82.80	85.86	96.97
Especificidad (%):	0.00	83.87	98.39	100	100	100
VP+ (%):	95.07	82.76	97.92	100	100	100
VP- (%):	0.00	78.79	80.26	85.32	87.61	97.06
Falsos + (%):	0.20	22.45	25.25	16.22	14.14	3.03
Falsos - (%):	100	24.49	2.02	2.03	0	0

Fuente: Elaborada por el autor.

Donde:

A: Todos los datos sin aplicar ninguna técnica de desbalance de datos

B: Mejor resultado de las dos muestras aplicando la técnica de *DownSampling*

C: Mejor resultado de las dos muestras aplicando la técnica de *UpSampling* para la réplica doble

D: Mejor resultado de las dos muestras aplicando la técnica de *UpSampling* para la réplica triple

E: Mejor resultado de las dos muestras aplicando la técnica de *UpSampling* para la réplica cuádruple
F: Mejor resultado de la muestra aplicando la técnica de *UpSampling* para la réplica de 18 veces

Se presenta el gráfico con los resultados de la Tabla 28 en la Figura 2.

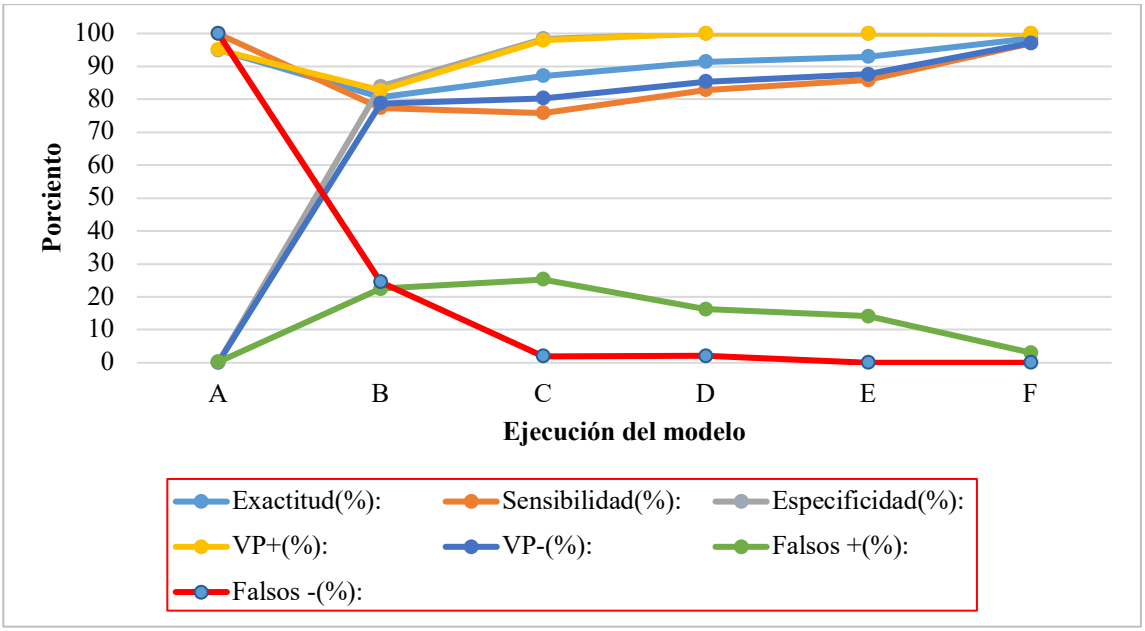

Figura 2. Variación de los mejores resultados en las mejores realizadas. Fuente: Elaborada por el autor.

4. Conclusiones

Los mejores resultados, tanto en el pronóstico de las diferentes *clases* como en las medidas de la matriz de confusión, se obtuvieron cuando se emplea la técnica de *UpSampling* con 18 réplicas. La repetición de los datos puede ser contraproducente porque podría generar un modelo sobreajustado, lo que traería como consecuencia errores en los pronósticos. Se recomienda probar el modelo con otros conjuntos de datos que estén compuestos por las mismas variables para ver si el este es capaz de hacer un buen pronóstico y las medidas de la matriz de confusión también sean correctas.

Para este conjunto de datos la mejor técnica para tratar el desbalance fue la de *UpSampling* en todas sus réplicas. Si la *clase* de menor cantidad de datos, en este caso la *clase Si,* tuviera un tamaño mayor es posible que también con la técnica de *DownSampling* se obtuvieran buenos resultados.

Los mejores resultados se obtuvieron en los grupos de entrenamiento del 80% y del 75%, con una cantidad de casos de 4 y 2 respectivamente. Esto no quita que para un conjunto de datos donde la *clase* que influya en el desbalance tenga mayor cantidad de elementos también se pueda emplear el conjunto de datos de entrenamiento del 70%.

Cuando se va a evaluar un modelo no solo se debe basar en la Exactitud del este (*Accuracy*), sino que se recomienda, que se analicen en su conjunto el resto de las medidas que resultan de la matriz de confusión para realizar un análisis más completo de los resultados que se obtengan.

Se deben emplear estas técnicas de desbalance de datos en donde la *clase* minoritaria tenga, al menos, entre veinte y treinta registros, multiplicado por la cantidad de variables que interviene en el pronóstico. En la medida que tenga más registros se repetirán los datos con menos frecuencia, logrando así una mayor confianza en los resultados del pronóstico. Un conjunto de datos pequeño puede producir resultados sesgados.

5. Referencias

[1] Martinelli, J. E. (2022). *Clasificación de datos desbalanceados. Su aplicación en la predicción de bajas de beneficiarios de un servicio de salud privado*. Facultad de Informática, Universidad Nacional de La Plata, Argentina. https://sedici.unlp.edu.ar/bitstream/handle/10915/147410/Documento_completo.pdf?sequence=1&isAllowed=

[2] Kaggle (2022). *Brain Stroke Dataset Classification Prediction*. https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset

[3] Breiman, L. (2001). Random Forest. *Machine Learning*, *45* (1), 5-32. http://dx.doi.org/10.1023/A:1010933404324

[4] Del Castillo Collazo, N. (2020). Predicción en el diagnóstico de tumores de cáncer de mama empleando métodos de clasificación. *Revista de Investigación en Tecnología de la Información (RITI), 8* (15), 96-104. https://doi.org/10.36825/RITI.08.15.009

[5] Cirillo, A. (2017). *R Data Mining. Implement data mining techniques through practical use cases and real-world datasets.* Packt Publishing Ltd.

[6] Villalba, F. (2018). *Aprendizaje supervisado en R*. https://fervilber.github.io/Aprendizaje-supervisado-en-R/bosques.html

[7] Sotaquirá, M. (2021). *Los Bosques Aleatorios: Clasificación y Regresión*. https://www.codificandobits.com/blog/bosques-aleatorios/

[8] Carrasco Calle, R. A. (2021). *¿Cómo manejar el desbalance de datos?* https://datasciencepe.substack.com/p/como-manejar-el-desbalance-de-datos

[9] Cruz-Reyes H., Reyes-Nava A., Rendón-Lara E., Alejo R. (2018). Estudio del desbalance de clases en bases de datos de microarrays de expresión genética mediante técnicas de Deep Learning. *Research in Computing Science*, *147* (5), 197–207. http://dx.doi.org/10.13053/rcs-147-5-15

[10] Landa Cosio, N. A. (2021). *Cómo actuar ante el desbalance de datos*. https://medium.com/@nicolasarrioja/c%C3%B3mo-actuar-ante-el-desbalance-de-datos-a0d64f2b9619#:~:text=Downsampling%20consiste%20en%20quitar%20puntos,que%20la%20clase%20menos%20

[11] Aldás, J., Uriel, E. (2017). *Análisis Multivariante aplicado con R* (2da Ed.). Ediciones Paraninfo.

[12] Amazon Web Services (AWS). (2024). *¿Qué es el sobreajuste?* https://aws.amazon.com/es/what-is/overfitting/

[13] IBM. (2024). *¿Qué es el sobreajuste?* https://www.ibm.com/mx-es/topics/overfitting