Análisis Predictivo de Riesgo de Ictus mediante Modelos Lineales y Redes Neuronales Artificiales

Grupo GI_02
Gonzalo Cruz Gómez

23 de noviembre de 2025

Resumen

Este informe presenta el estudio y comparación de modelos de machine learning para la predicción de ictus (*stroke*). Utilizando el Healthcare Stroke Dataset, se ha diseñado un una comparativa para contrastar la eficacia de modelos lineales con diferentes características frente a modelos basados en redes neuronales (*MLP*). La metodología consiste en un proceso previo exhaustivo de análisis exploratorio de datos y una fase de preprocesamiento para después elaborar modelos lineales aplicando regularización y selección de características con el objetivo de compararlos con un MLP. Por último se ha elegido un modelo final y se han extraido una serie de conclusiones, las cuales evidencian la dificultad de elección de modelos en el ámbito médico.

1. Introducción y definición del problema

El **ictus** es una de las principales causas de mortalidad a nivel global. La identificación de pacientes de alto riesgo, basada en diferentes biomarcadores, así como en datos demográficos, es un aspecto clave para una intervención médica temprana.

El objetivo de este estudio es entrenar, optimizar y validar modelos de clasificación binaria, capaces de diferenciar entre aquellos pacientes propensos a sufrir un ictus y aquellos que no lo son, utilizando variables como la edad, niveles de glucosa, tipo de trabajo, si dicho paciente fuma, índice de masa corporal (BMI) y antecedentes de hipertensión o cardiopatías.

El problema técnico principal está en la naturaleza del dataset, el cual presenta un gran desbalance de clases (existen muy pocos casos de ictus en el dataset) y la existencia de datos faltantes. Estas características implican una fase de preprocesamiento exhaustiva para lidiar con ellas.

2. Metodología

La arquitectura del experimento se ha diseñado para abordar el compromiso entre sesgo y varianza, buscando modelos que generalicen correctamente ante datos no vistos. Esta tarea ha sido pensada para buscar modelos que generalicen correctamente, primando la predicción de verdaderos positivos (Recall alta), pero tratando de mantener el número de falsos positivos lo más bajo posible.

2.1. Análisis exploratorio

En el análisis exploratorio encontramos tres principales problemas: que la variable BMI tenía valores faltantes, que teniamos un gran desbalance de clase y que gender tenía tres categorías. Estas tres cosas debían ser solucionadas en la fase de preprocesamiento. También se estudiaron las relaciones entre variables del dataset y de las mismas con el target, encontrando relaciones significativas como que aquellos con hipertensión y con enfermedades cardíacas eran más propensos a sufrir ictus, o que aquellos que fuman o han fumado también tienen más riesgo. Todas estas relaciones tienen sentido con lo que hay escrito sobre strokes en la literatura.

2.2. Preprocesamiento

La calidad de nuestro modelo depende principalmente de la calidad de nuestros datos, por ello esta parte tiene mucha importancia. Primeramente se trató la presencia de valores faltantes en BMI, utilizamos una imputación mediante un algoritmo de K-nn para el cual se buscó la mejor k mediante cross validation. Una vez obtenida la k, se imputaron todos los valores faltantes en base a los k vecinos mas cercanos. Después, como teníamos dos variables cuya distribución estaba muy sesgada (bmi y glucose level) y tenían colas muy largas, se aplicó una transformación logarítmica para mantener la distribución original acortando dichas colas. También se hizo una codificación de las variables categóricas mediante One Hot Encoding, permitiéndonos así utilizarlas en nuestros modelos. Asimismo se trató el desbalanceo de clase utilizando SMOTE para generar nuevas muestras de la clase minoritaria en nuestro conjunto de entrenamiento. Por último, se hizo un escalado de las variables mediante una estandarización Z-score. Este último paso es clave para poder utilizar nuestros datos en los modelos sin introducir sesgos.

2.3. Modelos lineales

Como nuestra tarea es de clasificación binaria, se eligió un modelo de regresión logística, que además es un modelo fácilmente interpretable que nos permite ver el riesgo asociado a cada variable. Después obtamos por realizar una **selección de características con métodos wrapper**. Para evitar la "maldición de la dimensionalidadz eliminar ruido, se emplearon métodos secuenciales tanto forwards como backwards. Estos métodos iterativos añaden o elimina variables basándose en la mejora del rendimiento del modelo, garantizando así el principio de parsimonia: buscar el modelo más simple posible que explique suficientemente nuestros datos. Por último se utilizó **regularización** buscando reducir la importancia de las variables menos relevantes en el modelo, previniendo sobreajuste y multicolinearidad. Se compararon modelos utilizando L1, L2 y Elastic net, eligiendo el que mejores predicciones consiguió.

2.4. Modelos no lineales: Perceptrón Multicapa (MLP)

Con el objetivo de encontrar relaciones complejas no lineales entre las características de los datos, se implementó una red neuronal de tipo MLP. Su uso se fundamenta en el Teorema de Aproximación Universal, que nos indica que un MLP con al menos una capa oculta y suficientes neuronas puede aproximar cualquier función continua compleja. El MLP deforma el espacio de características mediante las funciones de activación no lineales, permitiendo trazar fronteras más complejas que un modelo lineal.

Para su implementación, primeramente se entrenó un modelo con una sola capa oculta eligiendo entre modelos con diferente numero de neuronas y eligendo aquel que tuviera mejor F1-score como criterio de parada (fig. 8). Después se hizo un MLP profundo con varias capas ocultas iterando sobre distintas configuraciones de capas y neuronas hasta encontrar la mejor (fig. 9). Por último se realizó una selección de características usando métodos wrapper mediante la implementación de MLP. Una vez se seleccionaron dichas características (fig. 11), se procedió a la implementación de un esquema lineal de regresión logística para hacer predicciones.

3. Análisis de Resultados

A continuación, se presentan los resultados obtenidos para el conjunto de datos de validación.

Tabla 1: Figuras de mérito por modelo aplicado

Modelo	**Recall**	**Precision**	**F1-Score**	**ROC-AUC**	**Specificity**
Regresión Logística	0.78	0.13	0.23	0.76	0.74
Wrapper Lineal	0.88	0.14	0.24	0.84	0.73
Regularización (Elastic Net)	0.78	0.13	0.22	0.83	0.73
MLP 1 Capa	0.90	0.13	0.22	0.83	0.69
MLP Profundo	0.84	0.12	0.21	0.83	0.68
Selección MLP + Reg. Logística	0.78	0.12	0.21	0.83	0.72

Regresión Logística: Este es el modelo base, que nos sirve como referencia. Como podemos ver en la la Tabla 1, presenta una recall relativamente alta(0.78), igual que la AUC (0.76) y la especificidad (0.74), pero presenta una precisión baja, es decir, tiene muchos falsos positivos. Es un modelo que prioriza detectar correctamente los verdaderos positivos sobre los falsos positivos como se puede ver en la Fig. 1 del anexo . Algo que no es ideal, pero que si es preferible en el contexto de la medicina.

Wrapper Lineal: La selección de características consiguió eliminar ruido, incrementando todas las figuras de mérito excepto la especificidad. Tenemos un modelo con una recall muy alta (0.88), una AUC también muy alta (0.84) y la precision ha aumentado ligeramente. Además, siguiendo el principio de parsimonia, este modelo no solo es mejor prediciendo, sino que lo hace con menos variables, por lo que es un modelo mejor en todos los aspectos. A pesar de ello sigue teniendo muchos falsos positivos como podemos ver en la Fig. 2.

Regularización (Elastic Net): De los tres modelos que probamos (L1, L2 y Elastic net), el modelo Elastic net fue el mejor en términos del F1-score. A pesar de ser un modelo sólido, es básicamente igual que el modelo base de regresión logística. Es un modelo con peor recall y precisión que el modelo de selección de características, prediciendo peor los casos de stroke como se puede ver en la Fig. 3

MLP 1 Capa: La introducción de un modelo no lineal (con 10 neuronas y función de activación tanh) mejora sustancialmente la capacidad del modelo de predecir los casos de pacientes con stroke (recall de 0.9, la más alta) pero a cambio el modelo tendrá una especificidad más baja debido a que para predecir mejor los verdaderos positivos, genera muchos falsos negativos como se puede ver en la Fig. 4. En el caso de que la capacidad de predecir la clase 1 fuera la única cosa en la que nos debiéramos fijar, este modelo sería el más adecuado.

MLP Profundo: Aumentar la profundidad de la red neuronal no nos aportó ningún beneficio como se puede ver en la Tabla 1. El modelo elegido fue un modelo con función de activación ReLu y 3 capas de 50, 25 y 10 neuronas respectivamente. Este modelo es peor en todos los aspectos. Esto se puede deber a que la introducción de más capas en el MLP haya hecho que el modelo sobreajuste, empeorando su capacidad de generalización. Predice peor los casos positivos y tambien genera muchos más falsos positivos (Fig. 5)

Selección MLP + Regresión Logística: Este enfoque híbrido de selección de variables mediante un MLP y su aplicación mediante una regresióon logística no mejora en absoluto la predicción hecha por los modelos anteriores, tan solo mejora a la regresión logística base. Es peor que la selección de características del modelo Wrapper lineal no solo en las figuras de mérito, sino que también utiliza más variables que este para hacerlo.(fig. 6)

4. Elección final del modelo, conclusiones y trabajo a futuro

La elección del modelo final es el **modelo de regresión logística con selección de variables mediante métodos Wrapper**. Es el modelo que mejores fiiguras de mérito tiene salvo la recall que es solo 0.02 menor que el MLP de una capa. En todos los modelos hemos priorizado siempre la figura de mérito del **Recall**, ya que el cosete de tener falsos negativos (no detectar un ictus) es mucho mayor que el de tener un falso positivo y enviar a un paciente sano a realizarse más pruebas. Es por ello que priorizamos el Recall sobre la Accuracy o la Precision.

A pesar de esto, si la diferencia en el recall no es demasiado grande, podemos elegir un modelo que tenga recall muy similar pero una precision un poco superior. Esto significa que tendríamos menos falsos positivos, evitando así una posible saturación del sistema sanitario debido al exceso de pacientes teniendo que realizarse pruebas.

Además, la elección del modelo logístico frente al MLP también viene dada por una de las ventajas principales del modelo logístico: su fácil interpretabilidad y su transparencia. Es sencillo ver cuales son los factores que determinan que la predicción sea buena frente a la dificultad que tiene en este sentido el MLP el cual es una caja negra. También, al utilizar nuestro modelo logístico menos variables cumplimos con el principio de parsimonia. Por todo ello es por lo que hemos elegido como modelo final el de regresión logística con selección de variables. Finalmente hemos utilizado este modelo en test con resultados no tan óptimos como en validación, pero aun así siguen siendo buenos resultados para nuestro problema como podemos ver en la Fig. 7

Por último, para evolucionar y mejorar este estudio podemos tomar una serie de acciones ded cara a futuro. Primeramente, poder computar la incertidumbre de las predicciones sería interesante, ya que podríamos establecer una serie de niveles de riesgo para los pacientes. Esto podría hacerse con modelos gausianos para clasificación o con una regresión isiotónica. También se podría estudiar la inclusión de variables socioeconómicas como estrés o condiciones laborales, así como la inclusión de otro tipo de patologías que pudieran estar relacionadas con los ictus. Por último, podría estudiarse la aplicación de otros modelos de predicción como SVM o modelos ensamblados como XGBoost que podrían ayudarnos con la baja precision.

Referencias Bibliográficas

[1] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1-5. Recuperado de: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html

[3] GeeksforGeeks. (s.f.). Regularization in Machine Learning. Recuperado de: https://www.geeksforgeeks.org/machine-learning/regularization-in-machine-learning/

[4] Hannerz, H., Albertsen, K., Burr, H., et al. (2018). Long working hours and stroke among employees in the general workforce of Denmark. *Scandinavian Journal of Public Health*, 46(3), 368–374. doi: 10.1177/1403494817748264

[5] Nakayama, H., Jørgensen, H. S., Raaschou, H. O., & Olsen, T. S. (1994). The influence of age on stroke outcome: The Copenhagen Stroke Study. *Stroke*, 25(4), 808–813. doi: 10.1161/01.STR.25.4.808

[6] Shah, R. S., & Cole, J. W. (2010). Smoking and stroke: the more you smoke the more you stroke. *Expert Review of Cardiovascular Therapy*, 8(7), 917–932. doi: 10.1586/erc.10.56

[7] World Health Organization. (2020). *The top 10 causes of death*. Recuperado de: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

A. Anexos

A.1. Figuras

A continuación se presentan las figuras referenciadas en el informe

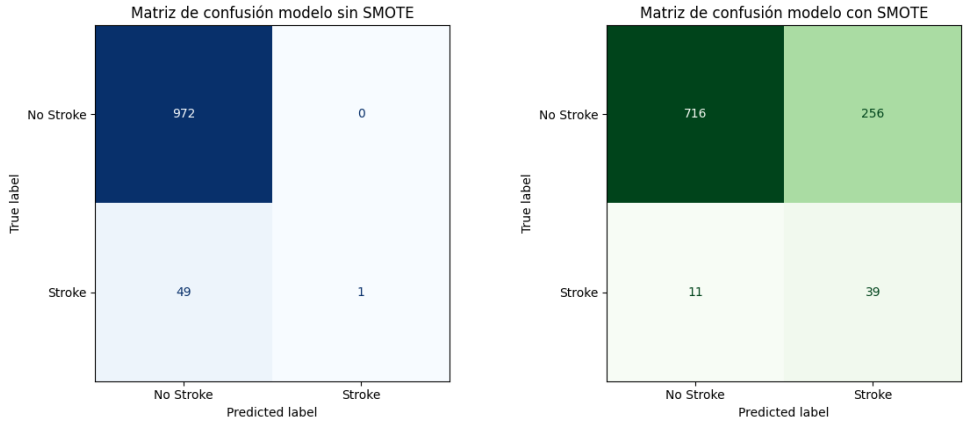

Figura 1: Matriz de confusión para la regresión logística base.

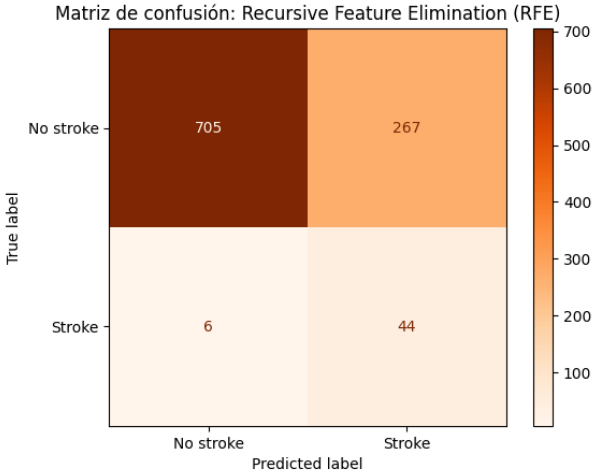

Figura 2: Matriz de confusión para el modelo wrapper Lineal.

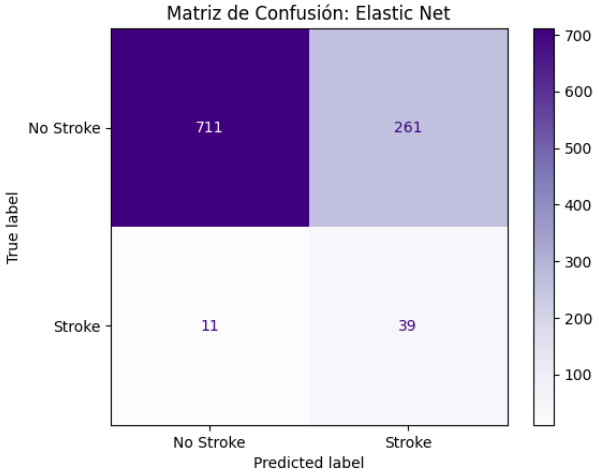

Figura 3: Matriz de confusión para la regularización elastic net.

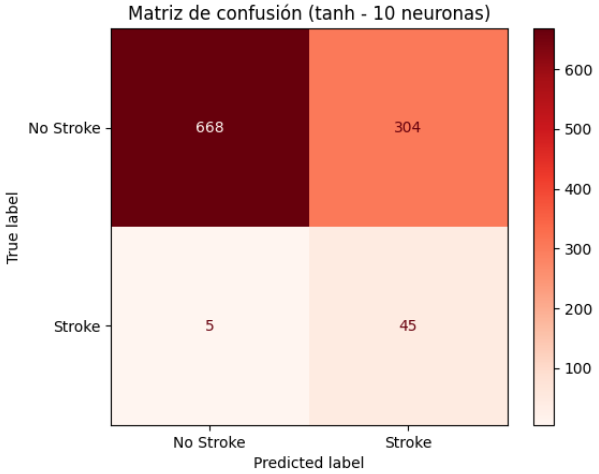

Figura 4: Matriz de confusión del MLP de 1 capa.

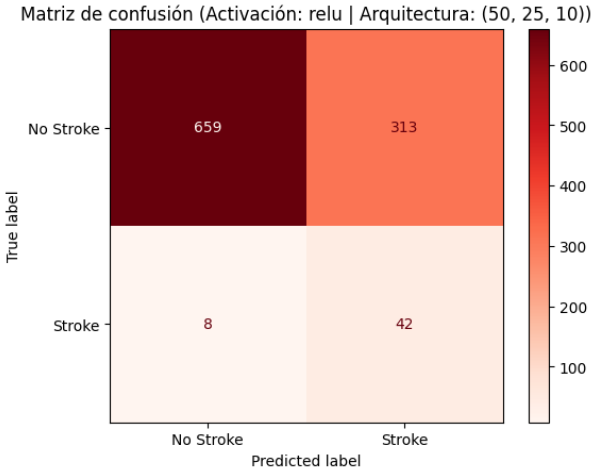

Figura 5: Matriz de confusión del MLP multicapa.

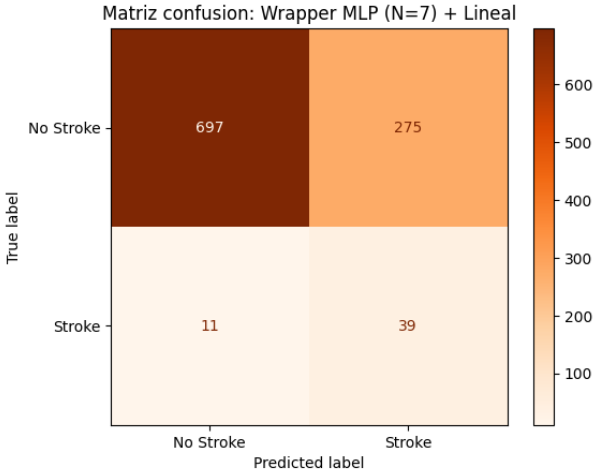

Figura 6: Matriz de confusion de MLP + regresión logística).

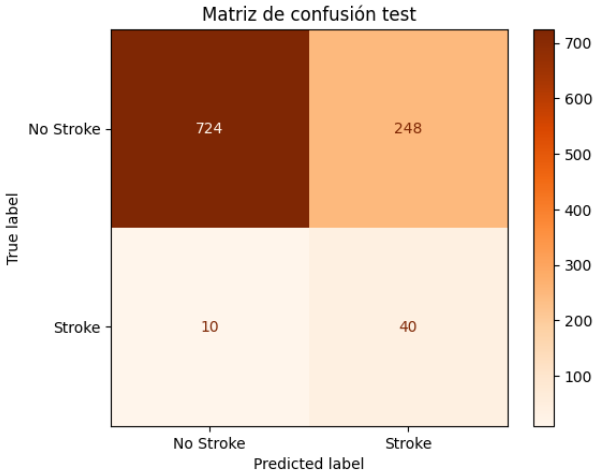

Figura 7: Matriz de confusión del modelo final en test.

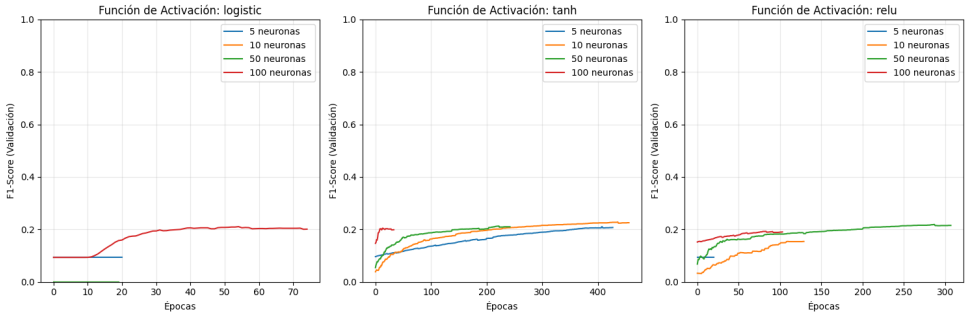

Figura 8: Curva de aprendizaje MLP de una capa.

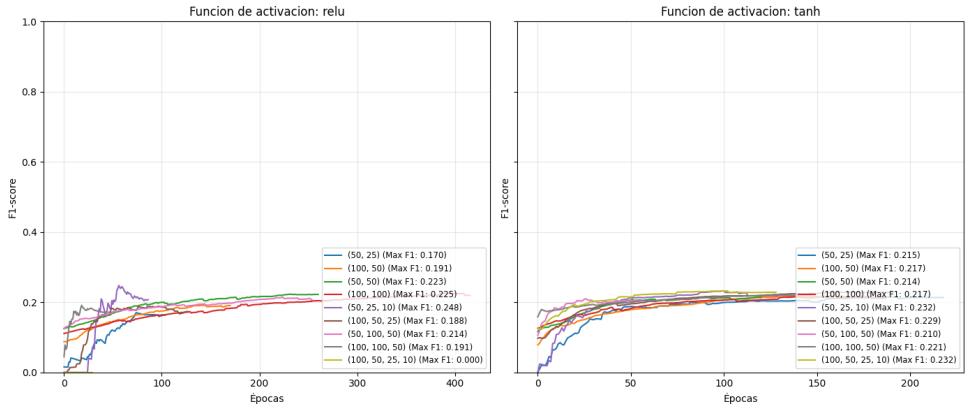

Figura 9: Curva de aprendizaje MLP profundo.

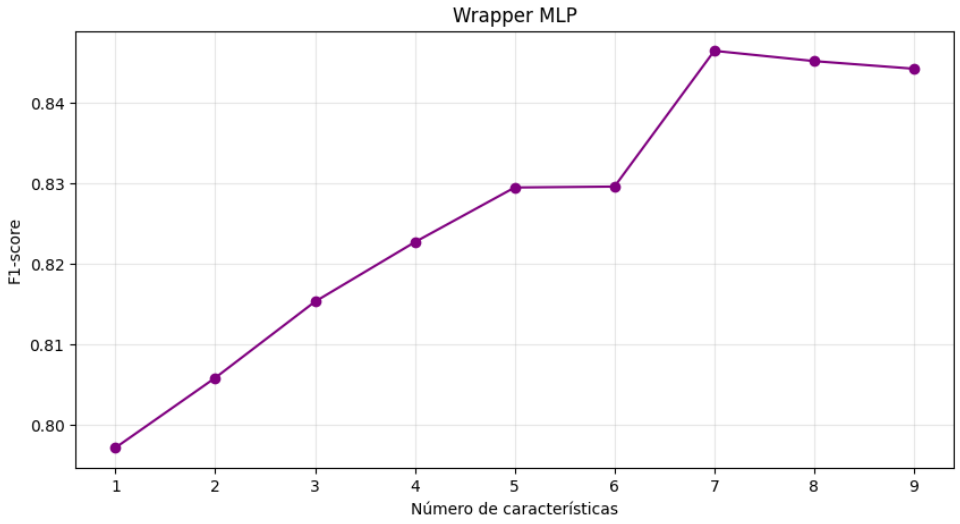

Figura 10: Selección de variables MLP + logistica.

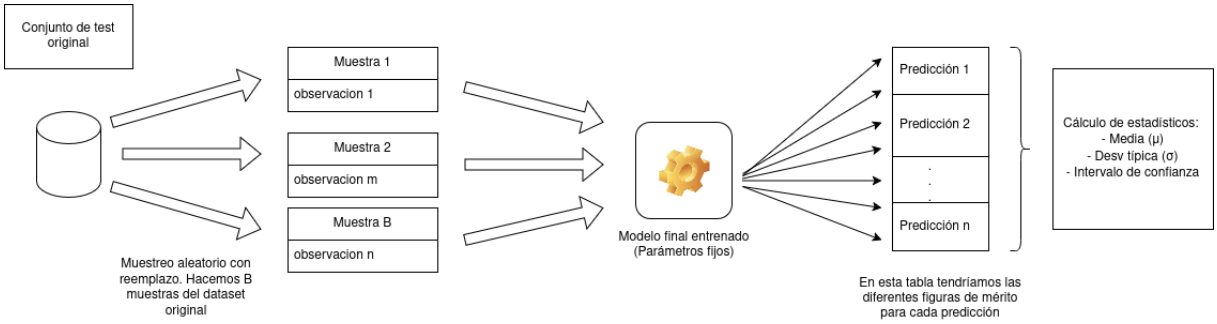

Figura 11: Diagrama obtención de estadísticos de las figuras de mérito en test.