Análisis Predictivo de Riesgo de Ictus mediante Máquinas de Vectores de Soporte (SVM)

Grupo GI_02
Gonzalo Cruz Gómez

10 de enero de 2026

Resumen

En este informe se presenta la tarea 2 de la asignatura de aprendizaje automatico II, que consiste en la aplicación de modelos de machine learning para la predicción de ictus (stroke). Tras una primera parte con modelos de regresión logística y MLP, en esta segunda tarea se evalúa la eficacia de las Máquinas de Vectores de Soporte (SVM) para la predicción de ictus. Se han comparado esquemas lineales (SVM lineal) frente a esquemas no lineales (SVM con kernel RBF). La metodología incluye la optimización de hiperparámetros y la selección de características mediante métodos wrapper. Los resultados muestran una clara superioridad de los modelos lineales para este problema, logrando una sensibilidad (Recall) del 80 %, mientras que los modelos no lineales complejos sufrieron un sobreajuste a la clase mayoritaria y no lograron predecir con precisión los casos de ictus.

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 principal radica en el desbalance de clases. En este contexto, priorizamos la métrica del **Recall (Sensibilidad)**, ya que el coste de un falso negativo (no detectar un paciente enfermo) es muy alto, aunque esto implique asumir una menor precisión. Aún asi, un gran numero de falsos positivos podria saturar el sistema por lo que en esta tarea vamos a utilizar el F1-score que nos da un balance entre ambas

2. Metodología

2.1. Preprocesamiento

Se partió de los datos preprocesados en la Tarea 1 a los cuales se le aplicó: Imputación de valores faltantes (BMI) mediante KNN, una transformación logarítmica de variables sesgadas, se hizo una codificación One-Hot y por último para el desbalanceo de clases, se aplicó SMOTE en el conjunto de entrenamiento, además de una normalización de los datos.

2.2. Máquinas de Vectores de Soporte (SVM)

Se exploraron dos enfoques de SVM buscando la mejor frontera de decisión:

2.2.1. SVM Lineal

Como nuestra tarea es de clasificación aplicamos primeramente un SVM lineal. Se ajustaó el hiperparámetro C mediante cross-validation hasta encontrar el modelo con mejor F1-score. Además, con el SVM lineal, podemos ver los pesos de cada una de las variables a la hora de predecir, dándonos una gran interpretabilidad del modelo. Además, comparamos el mejor modelo con el modelo base de la tarea 1, una regresión logística con todas las variables. Para traernos el modelo, en vez de entrenarlo de nuevo, decidimos utilizar pickle para guardar los pesos de la regresión logística y ahorrar coste computacional.

2.2.2. SVM No Lineal

Después aplicamos modelos SVM de kernel no lineal, se hizo un cross validation para buscar el mejor modelo posible en función del F1-score, utilizando diferentes kernels (polinómico de diferentes grados, RBF, sigmoide) y diferentes hiperparámetros para cada kernel. Todo ello para buscar capturar relaciones no lineales complejas que el modelo lineal no pudiera encontrar.

2.3. Selección de Características (Wrapper)

Para reducir la dimensionalidad y eliminar ruido, se buscó emplear una selección de características mediante métodos wrapper con un SVM. En este caso, aplicamos una eliminación recursiva de características con validación cruzda. Para ello se utilizó LinearSVC, que es un modelo SVM muy optimizado, lo cual permite que la selección de características sea rápida a pesar de tener que entrenar un gran número de modelos. Además, el uso de un SVM lineal permite que el modelo sea más interpretable.

Después se aplicó dicha selección de características tanto al modelo no lineal anterior como al modelo lineal, para así ver qué efecto tiene dicha selección de características en el modelo final, ya que siempre buscamos que se cumpla el principio de parsimonia, priorizando el uso de modelos lo más simples e interpretables posible y más aún en un ámbito tan crítico como es la medicina. En este caso se acabaron seleccionando 12 variables, reduciendo la complejidad del modelo significativamente.

3. Análisis de Resultados

A continuación, se comparan distintos modelos aplicados en esta práctica..

Cuadro 1: Comparativa de Figuras de Mérito (Conjunto de Validación)

Modelo	**Recall**	**Especif.**	**Precisión**	**F1-Score**	**ROC-AUC**
Regresión Logística (Base)	0.7800	0.7366	0.1322	0.2261	0.8300
SVM Lineal	**0.8000**	0.7181	0.1274	0.2198	0.8332
SVM No Lineal (RBF)	0.0800	**0.9506**	0.0769	0.0784	0.6624
SVM Lineal (Wrapper)	**0.8000**	0.7181	0.1274	0.2198	**0.8313**
SVM No Lineal (Wrapper)	0.2800	0.9053	0.1321	0.1795	0.7024

3.1. Análisis del SVM Lineal

El **SVM Lineal** tuvo muy buen rendimiento, alcanzando un **Recall del 80 %**. Además, la aplicación de la selección de características fue muy buena ya que se consiguió con menos variables tener exactamente el mismo rendimiento. De esta forma se tiene un modelo más simple y eficiente sin sacrificar nada de capacidad predictiva, es decir, las variables eliminadas de este modelo eran ruido.

3.2. Análisis del SVM no lineal

El modelo no lineal, que finalmente fue un modelo con kernel RBF, C=10, gamma = 1, tuvo un rendimiento muy malo, con un recall extremadamente bajo (0.08), indicandonos un sobreajuste a la clase mayoritaria e ignorando la minoritaria, el problema es que en nuestro caso queremos justo lo contrario. Esto también nos indica que buscar fronteras no lineales en nuestro modelo termina por ser perjudicial para el mismo, además de que estos modelos no lineales son mucho menos interpretables.

La selección de variables consiguió que nuestro modelo no lineal rindiera mucho mejor pero aun así sigue sin ser un rendimiento bueno, mucho menos para uso clínico y comparado con el modelo lineal

4. Conclusiones y Elección del Modelo Final

Primero vamos a ver como rindieron los modelos de esta segunda práctica frente a los de la primera:

Cuadro 2: Comparativa de modelos tarea 1 y tarea 2

Modelo	**Recall**	**Especif.**	**F1-Score**	**AUC**
— *Tarea 1: Modelos de Referencia* —				
Regresión Logística Base	0.78	0.74	0.23	0.76
Reg. Logística + Wrapper	**0.88**	0.73	**0.24**	**0.84**
MLP (1 Capa)	0.90	0.69	0.23	0.83
— *Tarea 2: Modelos SVM* —				
SVM Lineal	**0.80**	0.72	0.22	0.83
SVM No Lineal (RBF)	0.08	**0.95**	0.08	0.66
SVM Lineal + Wrapper (12 vars)	0.80	0.72	0.22	0.83

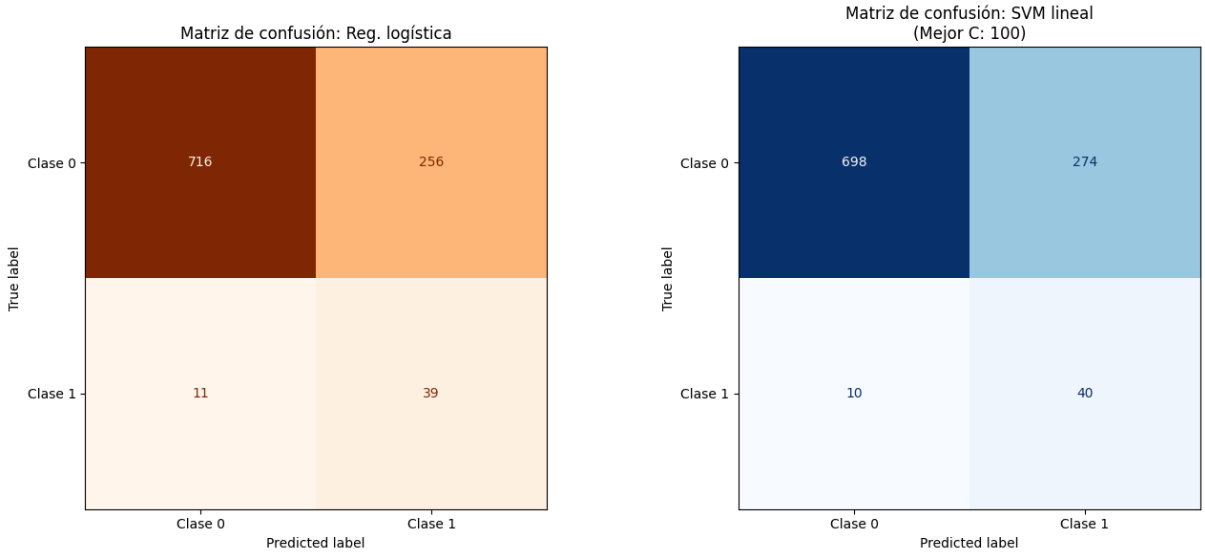
Primeramente, vemos que los modelos de la práctica anterior superan con creces a los de esta, sobre todo en recall, pero también en F1-score, que es en lo que más nos fijamos. El mejor modelo de las dos prácticas sigue siendo la Regresión logística con selección de características de la primera parte, el cual tenia muy buen recall y además era un modelo extremadamente simple e interpretable.

Después, respecto a los modelos de esta práctica únicamente, vamos que la selección de características fue muy efectiva, haciendo que los modelos sean más robustos y eliminando el ruido, además de simplificándolos. Claramente el enfoque lineal es el correcto para nuestro problema, ya que las fronteras de decisión lineales fueron mucho más robustas y efectivas para detectar los casos de ictus.

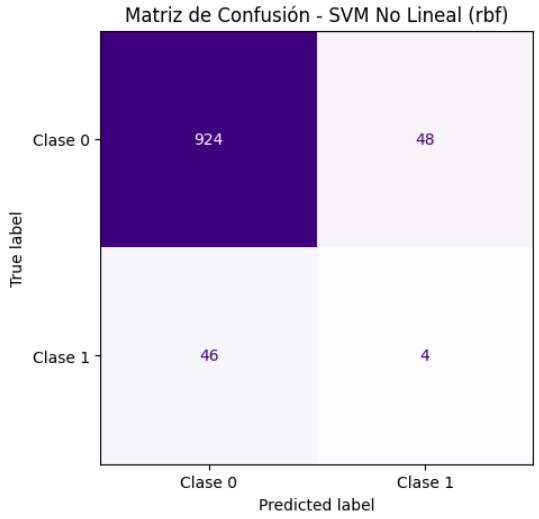
Finalmente, se eligió de esta práctica el modelo SVM lineal con selección de características, debido a su uso de un conjunto reducido de variables junto a una buena capacidad de detección de pacientes con riesgo. Este modelo se entrenó combinando train + validación y se probó en test dando resultados muy similares, por lo que no hubo sobreajuste en el modelo.

A. Anexos: Gráficas del Estudio

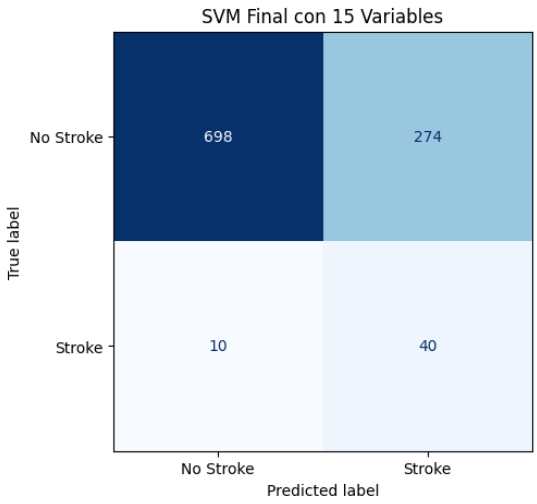
A.1. Matrices de Confusión de los Modelos



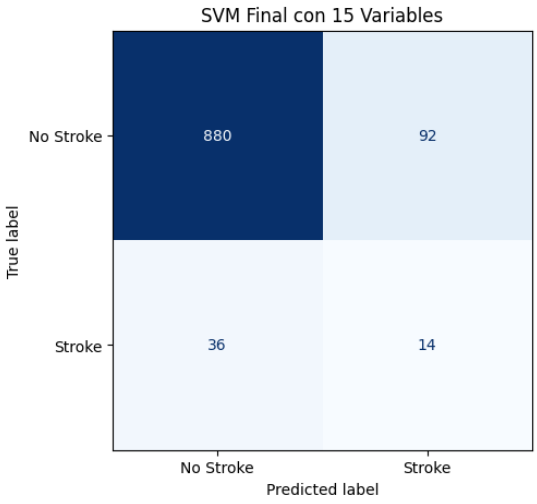
(a) Regresión Logística y SVM lineal



(b) SVM No Lineal (RBF)



(c) SVM Lineal (Wrapper)



(d) SVM No Lineal (Wrapper)

Figura 1: Matrices de confusión de los 5 modelos evaluados en el conjunto de validación.

A.2. Análisis de Variables y Coeficientes

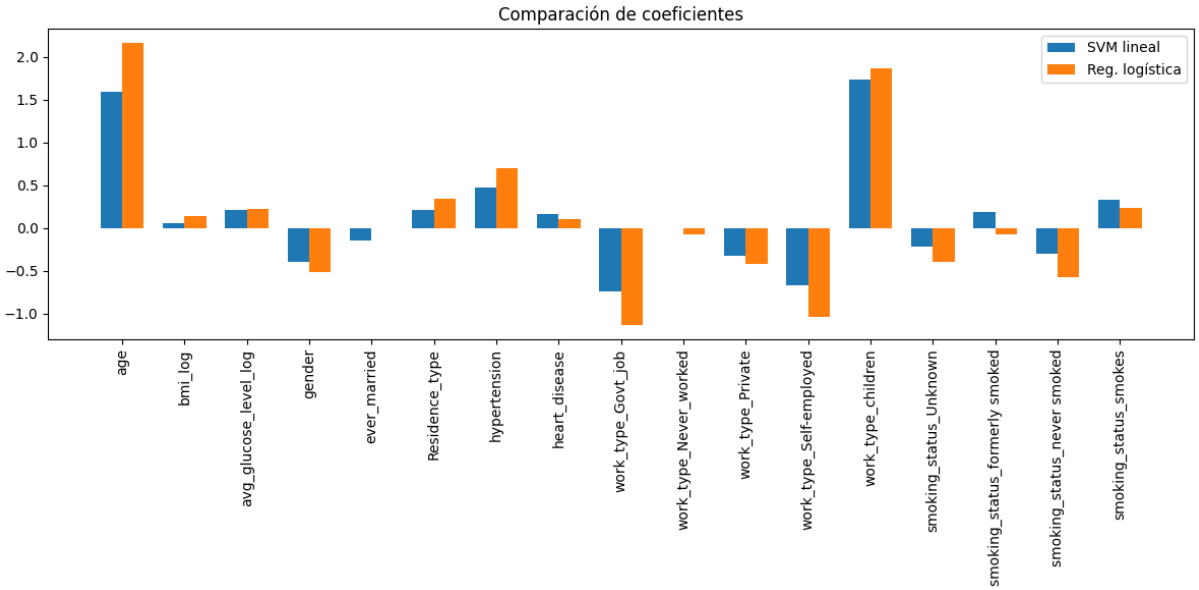


Figura 2: Comparación de coeficientes entre la legresión logística y el SVM Lineal. Se observa una asignación de importancia similar en ambos modelos lineales.

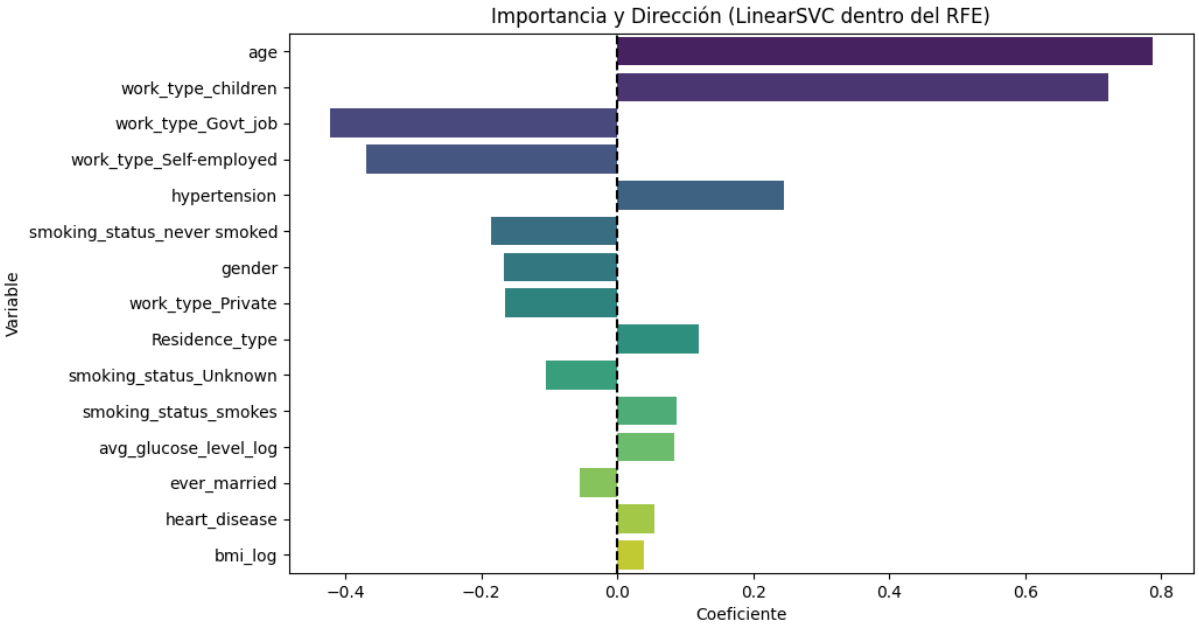


Figura 3: Importancia de las variables seleccionadas en el Modelo Final (SVM Wrapper). La edad (*age*) y el nivel de glucosa destacan como factores principales.