3. Regresión con Métodos de Selección y Regularización. Para este ejercicio se utilizará la base de datos *cancer_mama* perteneciente a la Universidad de Wisconsin. Los datos contienen el diagnóstico de cáncer de mama en 569 mujeres. Las características del diagnóstico corresponden a características obtenidas mediante un análisis de imágenes digitalizadas de masas en los senos. Para mayor información pueden consultar: UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. El objetivo de este ejercicio es predecir el diagnóstico de cáncer de mamá (*Diagnosis*) con base en las características de la masa observada. Este ejercicio debe ser contestado en un cuaderno de Jupyter siendo muy claro en donde está respondiendo cada uno de los siguientes ítems y con su código asociado. Recuerde que puede utilizar todos los paquetes de Python vistos en clase.

a) Si es necesario, convierta los datos a los formatos correctos. Obtenga una descripción de la base de datos. Describa los resultados.

Solo se convirtió a formato numérico la variable Diagnosis. Esta variable no presenta un desbalance muy fuerte y por tanto no sesgaría los resultados. De 568 pacientes, el 62.7% no fueron clasificados con cancer maligno, el restante 37.2% sí.

La base de datos no presenta NA y las estadísticas descriptivas no presentan ninguna anomalía. Así, mismo con los histogramas (sin contar las binarias): fractal dimension, simetry, compactness, smoothness, área, perimeter, textura y radius parecen tener una distribución normal, el restante parece tener una χ^2.

Si se analizan las correlaciones y scatterplots, se evidencia que el diagnostico tiene una relación positiva y alta con radius, perimeter, área, concavity y concave points, por lo que es esperable que estas sean relevantes en los modelos posteriores.

b) Implemente correctamente un modelo de regresión logística como un problema de predicción, usando como variable a predecir *Diagnosis*. ¿Qué piensa de los resultados?

Al evaluarlo con la muestra de testeo el modelo tiene unos resultados positivos. Si tenemos en cuenta que el accuraccy, precisión y recall fueron de 90%, 85.24% y 86.6%, el modelo es confiable para predecir el diagnóstico: (benigno o maligno) del cáncer y específicamente los malignos, debido al alto recall y precisión, podemos decir que el modelo maneja bien esa clase.

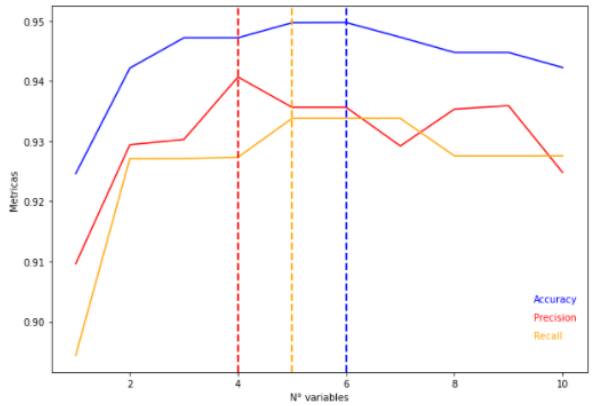
Al implementar validación cruzada (dentro de la muestra de entrenamiento), los resultados son mucho mejores (93.4%, 91% y 91.4%) y consecuentes con lo anterior. De la matriz de confusión vale recalcar que el porcentaje de Falsos Negativos de 8 y de Falsos Positivos de 9. Esto toma relevancia en un contexto donde se quiere predecir cáncer.

Por los coeficientes, se deduce que el radius, el perimeter y compactness afectan negativamente el diagnóstico y lo contrario ocurre con textura, área, concavity y concavity points. Las variables con coeficientes bajos se esperan que sea suprimidas con el método Lasso.

c) Ajuste una regresión logística con selección de variables usando métodos de selección de variables *Forward, Backward y Stepwise* para encontrar el mejor modelo. Para cada uno de los métodos calcule la matriz de confusión, *accuracy*, *precision*, *recall*. ¿Qué diferencia observa entre las métricas de desempeño de los diferentes modelos? Explique.

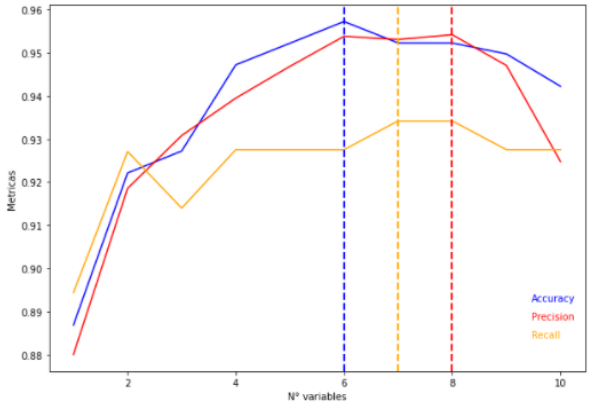
Para cada método se selecciono un conjunto de variables en función de las métricas, es decir, en total se seleccionaron 9 modelos (ej: Foward con accuracy, con precisión y recall). En las siguientes gráficas se resumen los resultados:

G1: METRICAS DE MODELOS CON VARIABLES SELECCIONADAS CON FOWARD



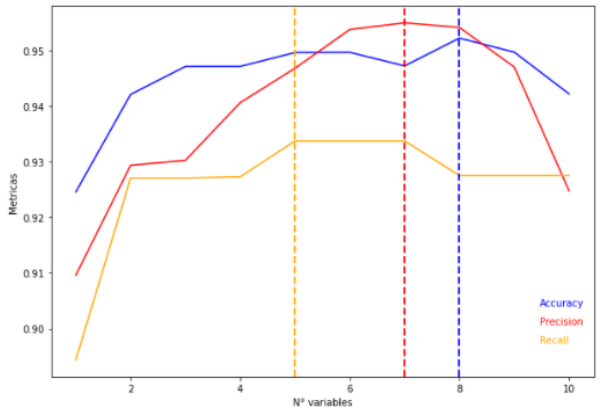
Fuente: UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. Elaboración propia.
Las líneas punteadas muestran el número de variables óptimas según la métrica.

G2: METRICAS DE MODELOS CON VARIABLES SELECCIONADAS CON BACKWARD



Fuente: UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. Elaboración propia.
Las líneas punteadas muestran el número de variables óptimas según la métrica.

G3: METRICAS DE MODELOS CON VARIABLES SELECCIONADAS CON STEPWISE



Fuente: UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. Elaboración propia.
El número de variables óptimos de Recall y Accuracy coinciden

En la siguiente tabla se resumen las métricas de los mejores modelos:

T1: métricas de los modelos óptimos según método selección – métrica

	VRFA	VRFP	VRFR	VRBA	VRBP	VRBR	VRSA	VRSP	VRSR
VN	103	99	103	104	103	104	103	103	103
FP	8	12	8	8	8	8	8	8	8
FN	7	8	7	7	6	7	6	7	7
VP	53	52	53	53	54	53	54	53	53
AS	0,912281	0,883041	0,912281	0,918129	0,918129	0,918129	0,918129	0,912281	0,912281
PS	0,868852	0,8125	0,868852	0,883333	0,870968	0,883333	0,870968	0,868852	0,868852
RS	0,883333	0,866667	0,883333	0,883333	0,9	0,883333	0,9	0,883333	0,883333

Fuente: UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. Elaboración propia
Las filas son las métricas y las columnas los modelos. En amarillo están los modelos con mejor métrica y en gris aquellos escogidos con la misma métrica que se evalúa.

En esta tabla la casilla (1,1) se lee como los verdaderos positivos del mejor modelo Foward utilizando la métrica de Accuracy o la (7,9) como el Recall del mejor modelo Stepwise seleccionado con Recall.

Si tenemos en cuenta la métrica Accuracy, los mejores modelos son aquellos que utilizaron Backward y Stepwise-Accuracy ; lo mismo para Precision con Backward-Precision y Backward-Recall; y Recall presenta un empate con Backward y Stepwise. Evidentemente, aquellos modelos que fueron seleccionados con una métrica determinada tienen mejor desempeño en la misma dentro del método de selección.

El modelo de Stepwise-Accuracy parece tener el mejor desempeño ya que tiene el porcentaje más bajo de Falsos Negativos (importante en un contexto de cancer), el Recall y Accuracy más alto. Además de que Precision es cercano al mejor.

d) Presente en una tabla las variables de la regresión del punto (b) y las variables finales de los tres métodos de selección.

	Logit	VRFA	VRFP	VRFR	VRBA	VRBP	VRBR	VRSA	VRSP	VRSR	TOTAL
radius	1	1	1	1	1	1	1	1	1	1	10
texture	1	1	0	1	1	1	1	1	1	1	9
perimeter	1	0	0	0	0	0	0	0	0	0	1
area	1	0	0	0	0	1	1	1	0	0	4
smoothness	1	1	0	1	1	1	1	1	1	1	9
compactness	1	1	0	1	0	1	0	1	0	1	6
concavity	1	0	1	0	1	1	1	1	1	0	7
concave points	1	1	1	1	0	0	1	0	1	1	7
symmetry	1	0	0	0	1	1	1	1	1	0	6
fractal dimension	1	1	1	0	1	1	0	1	1	0	7

La cantidad de modelos que seleccionaron las variables de textura, radius, concave points y concavity es consecuente con la magnitud de los coeficientes de la regresión Logit inicial. Es interesante ver como el modelo VRSA (el mejor seleccionado anteriormente) selecciona 8 de las 10 de variables.

e) Implemente las técnicas de regularización Ridge y Lasso bajo un problema de clasificación y determine cuál tiene mejor desempeño según las métricas de desempeño descritas en el inciso (c). ¿el modelo elegido contiene todas las variables de la base de datos? ¿Por qué?

Accuracy: Ridge (92.9%) Lasso (91.2%)

Precision: Ridge (91.3%); Lasso (88.3%)

Recall: Ridge (88.3%) Lasso (86.6%)

Matriz Lasso
Verdaderos negativos: 104
Falsos Positivos: 7
Falsos Negativos: 8
Verdaderos Positivos: 52

Matriz Ridge
Verdaderos negativos: 106
Falsos Positivos: 5
Falsos Negativos: 7
Verdaderos Positivos: 53

Modelo escogido: Ridge. Este se comporta mejor en todas las métricas. Este tiene todas las variables, ya que, a diferencia de Lasso, Ridge utiliza un método de suavización en el parámetro lambda (lo que impide una solución es esquina) y todas las variables tengan un peso distinto en orden de importancia.