1. **¿Qué es la validación cruzada repetida?**

Es una técnica que consiste en dividir el conjunto de datos varias veces en subconjuntos de entrenamiento y prueba, aplicando la validación cruzada k-fold en múltiples repeticiones para obtener una estimación más robusta y confiable del rendimiento del modelo, reduciendo la varianza del resultado.

1. **¿Explica para qué sirve la partición del conjunto de datos en entrenamiento (80%) y prueba (20%)?**

Sirve para construir el modelo con el conjunto de entrenamiento (80%) y luego evaluar su capacidad predictiva con datos nuevos e independientes (20%) que no fueron utilizados en el ajuste. Esto permite medir la generalización del modelo.

1. **¿A qué se deben las diferencias observadas entre los modelos generados con AIC y los generados con BIC?**

Las diferencias se deben a los criterios de penalización: el AIC penaliza con menor intensidad la complejidad del modelo, por lo que tiende a seleccionar modelos más complejos; mientras que el BIC penaliza con mayor severidad, favoreciendo modelos más simples y conservadores.

1. **¿Qué son los modelos penalizados?**

Son modelos de regresión que incorporan un término de penalización a la función de pérdida para evitar el sobreajuste y controlar la complejidad del modelo.

1. **¿En qué consisten los modelos penalizados?**

Consisten en ajustar un modelo de regresión minimizando una función que combina el error de predicción con una penalización basada en los coeficientes del modelo. Esta penalización puede ser de tipo L1 (LASSO) o L2 (Ridge), entre otras.

1. **¿Cuándo es necesario usar los modelos penalizados?**

Cuando se tienen muchas variables explicativas, posibles problemas de multicolinealidad o se desea mejorar la capacidad de generalización del modelo evitando el sobreajuste.

1. **¿Por qué en ocasiones se hace necesario estandarizar los parámetros de un modelo?**

Porque muchas técnicas penalizadas y de regularización (como LASSO o Ridge) son sensibles a la escala de las variables. Estandarizar permite que todas las variables contribuyan de forma equitativa al modelo.

1. **¿Explica en qué consiste el análisis tipo II y de qué sirve?**

Por otro lado, a la hora de determinar la utilidad de las variables input, puede resultar de utilidad obtener lo que se conoce como suma de cuadrados de tipo III.

Esta suma de cuadrados permite cuantificar cuánto aumenta la suma de cuadrados de los errores (en el conjunto de datos de entrenamiento) debido a la eliminación de cada variable input. La suma de cuadrados de tipo III se obtiene para cada variable input y, por tanto, nos permite hacernos una idea sobre la utilidad de las mismas.

En particular, cuando esta cantidad es pequeña, la variable correspondiente carece de poder predictivo pues el modelo no empeora significativamente al eliminarla del mismo.

1. **¿A partir del R2 en los datos de entrenamiento y prueba cómo se puede concluir sobre la calidad del modelo, así como su estabilidad?**

Si el R² es alto en entrenamiento pero bajo en prueba, el modelo está sobreajustado. Si el R² es consistente y razonablemente alto en ambos conjuntos, el modelo es estable y generaliza bien. La diferencia entre ambos ayuda a evaluar la estabilidad.

1. **¿Cuáles son los fundamentos del modelo LASSO?**

LASSO (Least Absolute Shrinkage and Selection Operator) penaliza la suma de los valores absolutos de los coeficientes. Tiende a hacer que algunos coeficientes sean exactamente cero, lo que equivale a una selección automática de variables.

1. **¿Cuáles son los fundamentos del modelo Ridge?**

Ridge penaliza la suma de los cuadrados de los coeficientes. Reduce la magnitud de todos los coeficientes pero no los elimina. Es útil para manejar la multicolinealidad y estabilizar los estimadores.

1. **¿Qué significa la línea horizontal que se añade a la gráfica en la validación cruzada para determinar el modelo penalizado óptimo?**

Dado que la validación cruzada es un proceso aleatorio, para reducir aún más las posibilidades de sobreajuste, se puede recurrir a la **regla 1se** (*one standard error rule*), que consiste en seleccionar el modelo más sencillo que proporcione una métrica de calidad que se encuentre a menos de una desviación típica del óptimo.

1. **¿Qué indican los índices accurate, Kappa, Sensitivity y Specificity?**

Accuracy: proporción de predicciones correctas.

Kappa: elimina el efecto de los aciertos que se pueden producir al azar.

Sensitivity: proporción de verdaderos positivos detectados.

Specificity: proporción de verdaderos negativos correctamente clasificados.

1. **¿Qué representa la curva rho y el AUC de un modelo?**

La curva rho muestra el comportamiento del error (o de la métrica de validación) en función del parámetro de penalización. El AUC (Área Bajo la Curva ROC) mide la capacidad del modelo para discriminar entre clases: mientras más cercano a 1, mejor desempeño.

1. **¿Qué es la curva ROC one vs all?**

Es una versión de la curva ROC usada en clasificación multiclase, donde cada clase se evalúa frente al resto de clases (uno contra todos). Permite analizar el rendimiento discriminante de cada clase individualmente.

1. **¿Cuáles son las diferencias teóricas entre el modelo de Poisson y la binomial negativa?**

Poisson: asume que la media y la varianza son iguales (equidispersión).

Binomial negativa: permite que la varianza sea mayor que la media (sobredispersión). Es más flexible para datos de conteo con alta variabilidad.

1. **¿Por qué es importante verificar la existencia de sobredispersión?**

Porque si existe y se utiliza un modelo de Poisson (que asume equidispersión), se subestima la varianza, lo que puede llevar a errores en las inferencias (valores p erróneos y confianza excesiva en los resultados). En esos casos, se recomienda usar modelos que permiten sobredispersión, como la binomial negativa.