UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS, SEDE BOGOTÁ
DEPARTAMENTO DE ESTADÍSTICA
Instrucciones segunda parte parcial 3 de análisis de regresión

OJO: ¡LEER ANTES DE ENVIAR!

- **Fecha límite de entrega:** domingo 4 de diciembre, 11:59pm (No se calificarán entregas después de la fecha y hora, y su calificación será 0.0).
- **Formato:** Se debe enviar un archivo en formato PDF con las respuestas escritas en un editor de texto y ecuaciones, y un archivo EXCEL solo con las predicciones solicitadas. (Penalidad por incumplimiento: -2.0 sobre la nota final)
- **Nombre del archivo:** -Nombrar el archivo PDF como "Parcial 3 Reg Grupo" *[Inserte el número de su grupo de trabajo tal y como aparece en la hoja de cálculo en el Drive].* (Penalidad por incumplimiento: -0.2 sobre la nota final).
  -Nombrar el archivo EXCEL como "pred_Reg_Grupo" *[Inserte el número de su grupo de trabajo tal y como aparece en la hoja de cálculo en el Drive].*
  (Penalidad por incumplimiento: -0.2 sobre la nota final).
  -Nombrar el archivo de código igual que el PDF. (Penalidad por incumplimiento: 0.0 en la nota final).
- **Mecanismo de entrega:** El informe y las predicciones deben ser subido al classroom UNA SOLA VEZ por uno de los estudiantes del grupo. Absténganse de enviar los trabajos a través del correo electrónico. (Penalidad por incumplimiento: -0.2 sobre la nota final).

Observaciones generales:

- Esta parte la podrán desarrollar en los mismos grupos del proyecto.
- La entrega consiste en hacer llegar las predicciones del mejor modelo que ustedes obtuvieron en el archivo **pred.xlsx**, nombrarlo adecuadamente y realizar un informe que describa brevemente el procedimiento que siguieron para llegar a ese modelo.
- Es **obligatorio** escribir el informe con un editor de texto y de ecuaciones.
- También es **obligatorio** enviar el archivo de código con la construcción del modelo que generó las predicciones entregadas.

Reglas generales:

1. Archivos. Ustedes recibieron tres archivos:
   a. **data.xlsx** contiene 4665 observaciones de las variables explicativas y de la variable respuesta. Los datos corresponden a las mediciones de energía eólica producida por un parque en Texas, Estados Unidos. La variable respuesta (**pow**) es la cantidad de energía entregada en cada hora (en MWh) y las variables explicativas corresponden a condiciones metereológicas/climáticas medidas en el mismo instante. La Tabla 1 (al final) muestra una breve descripción de las variables y sus convenciones.

b. **data2.xlsx** contiene solamente las mediciones de las variables explicativas para 1972 instantes de medición (la variable **pow** no está disponible). Esa información la deberán usar para predecir la cantidad de energía que se generó en cada instante de los allí mostrados.
c. **pred.xlsx** contiene únicamente un identificador de las mediciones presentes en **data2.xlsx** y una columna denominada “predicción” para que ustedes coloquen la predicción que obtuvieron con su mejor modelo. Solo se puede reportar una predicción para cada fila. Una vez ingresen sus predicciones, deberán modificar el nombre de este archivo de acuerdo con las indicaciones dadas arriba.

2. Evaluación. Para todos los grupos, la calificación de la parte B del parcial 3 corresponderá a la evaluación del informe 3 y haber entregado sus mejores predicciones. No será un ítem de evaluación la calidad de las predicciones que entreguen. Sin embargo, los grupos que obtengan la mejor habilidad predictiva promedio (medida con el error cuadrático medio entre las predicciones en **pred.xlsx** y los valores observados), tendrán bonificaciones adicionales.
    a. El equipo de la categoría 1 (ver más abajo) con el menor error cuadrático medio obtendrá una calificación de 5.0 en todo el corte, siempre y cuando haya entregado el informe.
    b. El segundo equipo de la categoría 1 (ver más abajo) con el menor error cuadrático medio obtendrá una calificación de 5.0 en la parte B del tercer corte y un bono de una unidad (1.0) en la parte A, siempre y cuando haya entregado el informe. Si este equipo le gana en desempeño al mejor grupo de la categoría 2 o si no hay candidatos en la categoría 2, el segundo equipo de la categoría 1 recibirá los mismos beneficios que el primer equipo de la categoría 1.
    c. El equipo de la categoría 2 (ver más abajo) con el menor error cuadrático medio obtendrá una calificación de 5.0 en la parte B del tercer corte y un bono de una unidad (1.0) en la parte A, siempre y cuando haya entregado el informe.

3. Categorías de participación. En esta competencia, se puede participar en dos categorías diferentes. Sólo se puede escoger una de las dos por grupo.

    a. Categoría 1. En esta categoría, solo podrán usar modelos vistos en la clase. Es decir,
        i. Modelos de regresión lineales en los parámetros.
        ii. Modelos de regularización (ridge, LASSO)*
        iii. El modelo aditivo general (GAM) con una o varias relaciones marginales modeladas de manera no paramétrica*
        iv. Modelos lineales generalizados.

    Nota 1: (*) Recuerden que, en los modelos con asterisco, hay hiperparámetros involucrados que deben ser calibrados haciendo nuevamente una partición del entrenamiento, de modo que se calibren varios modelos del mismo tipo con diferentes valores del hiperparámetro

con la primera parte; y se use la segunda parte para seleccionar un valor del hiperparámetro para cada tipo de modelo.

Nota 2: Es permitido usar transformaciones funcionales de las variables explicativas.

Nota 3: También se pueden combinar estas técnicas como el grupo considere pertinente y usar los mecanismos de selección automática.

Nota 4: Recuerden que, una vez que seleccionan el mejor modelo, deben reestimar sus parámetros usando todos los datos disponibles y luego, predecir los valores para los instantes que hacen parte de la competencia.

b. Categoría 2. En esta categoría podrán usar otros modelos no vistos en clase (redes neuronales, bosques aleatorios, árboles de decisión, SVR, SVM, etcétera) además de los ya mencionados.

4. <u>Detalles del informe.</u>

a. Integrantes (nombre, programa, correo)

b. Categoría de participación.

c. Criterio de habilidad predictiva. Justifiquen su elección.

d. Mecanismo de validación o de validación cruzada. Justifiquen su elección.

e. Construcción de modelos. Cuenten, de manera breve, cómo construyeron los modelos a considerar. ¿Hicieron transformaciones de variables? De las técnicas de la categoría 1, ¿cuáles usaron? ¿Combinaron las técnicas entre ellas? Si pertenecen a la categoría 2, ¿qué otras técnicas consideraron?

f. Completen la siguiente tabla con tres de los modelos que estimaron (los de mejor desempeño):

Modelo	**Descripción** *¿Cómo fue construido ese modelo? ¿Con qué técnicas?*	**Criterio predicción sobre testeo**
1		
2		
3		

g. ¿Qué variables aparecen en el modelo de mejor habilidad predictiva?

Tabla 1. Descripción de las variables disponibles.

avg_ws80m1	Velocidad promedio del viento extrapolada a 80m (m/s) por el método 1
sd_ws80m1	Desv. Estándar de la velocidad del viento extrapolada a 80m (m/s) por el método 1
med_ws80m1	Velocidad mediana del viento extrapolada a 80m (m/s) por el método 1
riq_sp80m1	Rango intercuartílico de la velocidad del viento extrapolada a 80m (m/s) por el método 1
avg_ws80m2	Velocidad promedio del viento extrapolada a 80m (m/s) por el método 2
sd_ws80m2	Desv. Estándar de la velocidad del viento extrapolada a 80m (m/s) por el método 2
med_ws80m2	Velocidad mediana del viento extrapolada a 80m (m/s) por el método 2
riq_sp80m2	Rango intercuartílico de la velocidad del viento extrapolada a 80m (m/s) por el método 2
avg_ws30ft	Velocidad promedio del viento extrapolada a 10m (m/s)
sd_ws30ft	Desv. Estándar de la velocidad del viento extrapolada a 10m (m/s)
med_ws30ft	Velocidad mediana del viento extrapolada a 10m (m/s)
riq_sp30ft	Rango intercuartílico de la velocidad del viento extrapolada a 10m (m/s)
avg_wd	Dirección promedio del viento (rad)
med_wd	Dirección mediana del viento (rad)
avg_wg	Velocidad promedio de ráfaga (m/s)
med_wg	Velocidad mediana de ráfaga (m/s)
avg_30ft	Temperatura promedio a 10m (C)
avg_hum	Porcentaje de humedad relativa (%)
wind_speed	Velocidad del viento a la altura de la turbina (m/s)
pow	Energía eólica (MWh)