# Pontificia Universidad Javeriana - Wikipedia, la enciclopedia libre

**APRENDIZAJE DE MÁQUINA** - **PARCIAL 2**

**Fecha: abril 25 de 2025. Duración: 3 horas**

# REGLAMENTO DE ESTUDIANTES de la PONTIFICIA UNIVERSIDAD JAVERIANA

*Faltas disciplinarias graves y gravísimas:*

Artículo 123 inciso (d): El fraude en actividades, trabajos y evaluaciones académicos y la posesión o utilización de material no autorizado en los mismos. (Falta disciplinaria grave)

Artículo 124 inciso (b): Todas las modalidades de plagio. (Falta disciplinaria gravísima)

Artículo 124 inciso (e): La suplantación en una evaluación académica, en exámenes preparatorios, en trabajos de grado y tesis. (Falta disciplinaria gravísima)

*Sanciones Disciplinarias:*

Artículo 127: Las faltas graves serán sancionadas con amonestación escrita con cargo a la hoja de vida del estudiante y la imposición de matrícula condicional durante el tiempo necesario para cumplir la condición.

Artículo 128: Adicional a la sanción disciplinaria, el fraude en actividades, trabajos y evaluaciones académicos se sancionará académicamente con la pérdida de la asignatura, la cual será calificada con nota definitiva de cero punto cero (0.0).

NO SE PERMITE EL USO DE HERRAMIENTAS DE CORREO, CHAT, DRIVES, o HERRAMIENTAS DE IA COMO CHATGPT, DEEPSEEK, ETC. SOLO PUEDE USAR LA PLATAFORMA DE BS ASÍ COMO GOOGLE PARA CONSULTA DE INFORMACIÓN GENERAL.

NO SE PERMITE USO DE CELULARES O TABLETS. PUEDE USER APUNTES DE CLASE.

IMPORTANTE

* Debe generar un notebook para cada punto 1 y 2, responda las preguntas planteadas dentro de una celda de texto en el notebook. Nombre cada notebook así: “**Punto X – nombre del estudiante**”. **Plazo para cargar ambos notebooks, así como la hoja Excel del punto 2: 10 am.**
* Para el entrenamiento y prueba de los modelos, **solo puede utilizar los comandos vistos en clase en R**.
* Para todos los comandos que involucren una selección aleatoria, por ejemplo, dividir el dataset entre entrenamiento y prueba, o seleccionar registros de un dataset, fije siempre la semilla con los **últimos 3 números de su cédula** de ciudadanía.
* Debe preparar los datos adecuadamente para ser usados con el modelo solicitado.

**PUNTO 1 (25). REGRESIÓN LINEAL**

Un sistema de bicicletas compartidas es un servicio en el que las bicicletas se ponen a disposición de los particulares para su uso compartido a corto plazo a cambio de un precio. Los sistemas de bicicletas compartidas permiten tomar prestada una bicicleta de una estación que suele estar controlada por computador, en el que el usuario introduce la información del pago y el sistema la desbloquea. Esta bicicleta puede devolverse en otra estación perteneciente al mismo sistema.

Se pide modelar la demanda de bicicletas compartidas con las variables independientes disponibles. La Administración usará el modelo para comprender cómo varía exactamente la demanda en función de las distintas características. En consecuencia, podrá manipular la estrategia empresarial para satisfacer los niveles de demanda y cumplir las expectativas de los clientes. Además, el modelo será un buen medio para que la Administración comprenda la dinámica de la demanda de un nuevo mercado.

Descripción del dataset (archivo: **demanda\_diaria.csv**):

1. indice: consecutivo
2. fecha: fecha del alquiler
3. temporada:
   * 1: temporada seca (diciembre a febrero)
   * 2: primera temporada de lluvia (marzo, abril)
   * 3: temporada normal (mayo a agosto)
   * 4: segunda temporada de lluvia (septiembre a noviembre)
4. año: año del alquiler (0: 2018, 1:2019)
5. mes: mes del alquiler (1 a 12)
6. festivo: si el día del alquiler es festivo o no
7. dia: día de la semana en que se hizo el alquiler (0: domingo, … , 6: sábado)
8. laboral: si el día del alquiler no es ni fin de semana ni festivo es 1, en caso contrario es 0
9. clima:
   * 1: mayormente despejado, pocas nubes, parcialmente nublado
   * 2: mayormente nublado con lluvia ligera o ninguna
   * 3: mayormente nublado con lluvia fuerte
10. temp: Temperatura en grados Celsius.
11. atemp: Temperatura de sensación, en grados Celsius.
12. humedad: nivel de humedad.
13. viento: Velocidad del viento.
14. casual: número de usuarios ocasionales
15. registrado: número de usuarios registrados
16. conteo: número total de bicicletas de alquiler, incluidos los usuarios ocasionales y los registrados

Usted deberá revisar el dataset, eliminar aquellas variables innecesarias, y formular el modelo que permita predecir la demanda (representada por la variable conteo). Las variables finales deben tener los formatos correctos.

1. (4) Plantee un modelo preliminar para predecir la demanda (llámelo **modelo0**).
2. (5) Seleccione las variables que sean significativas a un nivel de significancia = 0.01, y plantee un nuevo modelo (llámelo **modelo1**). Calcule las métricas de desempeño (RMSE, R2) tanto en entrenamiento como en prueba.
3. (5) Plantee un modelo de regularización RIDGE, con las mismas variables del modelo1, usando validación cruzada. Pruebe con el lambda mínimo y el lambda a una desviación estándar. Compare el desempeño con ambos lambdas y escoja el lambda que conlleve a un menor overfitting, justifique su respuesta (llame el modelo final **modelo2**).
4. (5) Plantee un modelo de regularización LASSO , con las mismas variables del modelo1, usando validación cruzada. Compare el desempeño con ambos lambdas y escoja el que conlleve a un menor overfitting, justifique su respuesta (llame el modelo final **modelo3**).
5. (3) Plantee un modelo de regularización ELASTIC NET usando las mismas variables del modelo1, optimizando el parámetro alfa (llámelo **modelo4**).
6. (3) Entre los 5 modelos desarrollados, seleccione el mejor modelo teniendo en cuenta las métricas generadas en entrenamiento y prueba.

**PUNTO 2 (25). REGRESIÓN LOGÍSTICA**

Bogotá se encuentra sumida en una severa sequía, y las autoridades gubernamentales han tomado la decisión drástica de cortar completamente el suministro de agua durante un mes si no se presenta un plan sólido de contingencia para esta crisis ambiental. La Alcaldía decide que lo mejor es dar una bonificación en el precio de las facturas para aquellos hogares que muestren un comportamiento ejemplar en el uso racional del agua, (más que simplemente tomar el número de m3 consumidos).

Usted ordena la recopilación de información importante en una muestra aleatoria de 838 hogares en la ciudad. El dataset resultante, luego de una selección adecuada de variables, cuenta con las siguientes variables (archivo: **bonificación.csv**):

* Personas\_Hogar: Personas que viven en el hogar
* Subsidio\_agua : Monto en $ que recibe de subsidio /mes
* Tipo\_jardin : 0 no tiene , 1 pequeño , 2 grande
* Tiempo en ducha : minutos promedio en ducha por persona /mes
* Cantidad de veces que se lava loza: promedio por día / mes
* Cantidad de litros tomados/persona: promedio de litros por persona / mes
* Lavadora: cantidad de usos en lavadora / mes
* Drenaje\_diario : promedio veces que se descarga cada sanitario al día /mes
* Metros\_inmueble: metros del inmueble
* Bono: 1 se da la bonificación, 0 no se da la bonificación
* Mascotas: número de mascotas
* Fiestas\_mes: cantidad de fiestas/mes

1. (5) Plantee un modelo para predecir si un hogar debe recibir bonificación. Utilice un cutoff de 0.5. Calcule las métricas de desempeño (accuracy, F1-score para la clase positiva) tanto en entrenamiento como en prueba.
2. (3) De las variables presentes en el modelo final, si tuviera que eliminar una variable de las que quedaron en el modelo final, ¿cuál eliminaría y porqué?
3. (8) Tomando como probabilidad base la probabilidad de obtener una bonificación, determine la probabilidad final que se obtiene al aumentar en 3 el número promedio de veces que se lava la loza diariamente, suponiendo que las demás variables no tienen ningún cambio. Realice este cálculo en una hoja de Excel, luego en el notebook valide el número obtenido (muestre que en los cálculos de R aparece dicho valor).
4. (5) Se selecciona un grupo de 2000 hogares para evaluar la implementación de la estrategia. Se ha estimado que por cada hogar categorizado correctamente se estima un ahorro promedio de $800, mientras que por cada hogar mal categorizado se estima un costo de $350. Basado en el cutoff de 0.5, ¿cuál sería la ganancia esperada en términos monetarios para este modelo?
5. (4) Si el cutoff se disminuye a 0.3, ¿cómo afectaría el precision y el recall? ¿Cómo afectaría la ganancia obtenida en el ítem anterior? Justifique sus respuestas. Valídelas con los resultados obtenidos en el notebook.