Profesor: Dr. Oldemar Rodríguez Rojas
CA-0404 Modelos Lineales
Regresión mediante el método KNN
Fecha de Entrega: Jueves 16 de septiembre a las 1pm

Tarea Número 3

- Las tareas serán revisadas en clase, no pueden ser enviadas por correo.
- Quienes no se presenten a la revisión de la tarea tendrán un cero de nota.
- Las tareas son estrictamente de carácter individual, tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso, es decir, el promedio de las notas obtenidas en la tareas será la nota final del curso.
- Todos los ejercicios tienen el mismo valor.

- **Pregunta 1:** [10 puntos] Explique detalladamente la diferencia entre un problema de regresión y uno de clasificación. Basado en su experiencia laboral o académica comente 2 ejemplos de problemas de clasificación y 2 de regresión que conozca y la oportunidad que ve en resolverlos.

- **Pregunta 2:** [15 puntos]

  1. Programe en lenguaje `R` una función que reciba como entrada la matriz de confusión (para el caso 2×2) que calcule y retorne en una lista: la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), los Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (NP).
  2. Supongamos que tenemos un modelo predictivo para detectar Fraude en Tarjetas de Crédito, la variable a predecir es `Fraude` con dos posibles valores `Sí` (para el caso en que sí fue fraude) y `No` (para el caso en que no fue fraude). Supongamos la matriz de confusión es:

	`No`	`Sí`
`No`	83254	15
`Sí`	879	4

     - Calcule la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), los Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (NP).
     - ¿Es bueno o malo el modelo predictivo? Justifique su respuesta.

- **Pregunta 3:** [25 puntos] En este ejercicio vamos a usar la tabla de datos `wine.csv`, que contiene variantes del vino "Vinho Verde". Los datos incluyen variables de pruebas fisicoquímicas y sensoriales realizadas a dicho vino.

  La tabla contiene 1599 filas y 12 columnas, las cuales se explican a continuación.

  - `fija.acidez`: Acidez fija, ácidos presentes después de la destilación.
  - `volatil.acidez`: Acidez volátil, esta determina si el vino tendrá un sabor avinagrado.

- `citrica.acidez`: Acidez cítrica, esta es propiamente de las uvas y no por fermentación.
- `residual.azucar`: Azúcar residual del vino que no fue fermentada.
- `cloruros`: Cloruros, uno de los principales componentes salinos del vino.
- `libre.sulfuro.dioxido`: Dióxido de azufre libre, antimicrobiano natural.
- `total.sulfuro.dioxido`: Dióxido de azufre total, suma del dióxido libre y combinado.
- `densidad`: Densidad.
- `pH`: Potencial de hidrógeno, este reduce la sensación de acidez en el vino.
- `sulfitos`: Sulfitos, conservan los aromas y actúa como desinfectante.
- `alcohol`: Alcohol presente.
- `calidad`: Calidad del vino.
- `tipo`: Tinto o Blanco (variable a predecir).

Para esto realice lo siguiente:

1. Cargue la tabla de datos `wine.csv` en `R`.
2. ¿Es este problema equilibrado o desequilibrado? Justifique su respuesta.
3. Use el método de K vecinos más cercanos en `traineR` (con los parámetros por defecto) para generar un modelo predictivo para la tabla `wine.csv` usando el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las categorías. ¿Son buenos los resultados? Explique.
4. Repita el item 3, pero esta vez, usando el menú de `Poder Predictivo` seleccione las 6 variables que, según su criterio, tienen mejor poder predictivo. ¿Mejoran los resultados?
5. Genere un Modelo Predictivo usando K vecinos más cercanos para cada uno de los siguientes núcleos: `rectangular`, `triangular`, `epanechnikov`, `biweight`, `triweight`, `cos`, `inv`, `gaussian` y `optimal` ¿Cuál produce los mejores resultados?

- **Pregunta 4:** [25 puntos] Un cliente nos contrata esta vez para aplicar un modelo no paramétrico con el fin de estudiar una posible oportunidad de negocio, y para ver si le es rentable quiere una predicción de las ventas potenciales de asientos de niños para autos en su tienda. Para ello nos proporciona la tabla de datos `AsientosNinno.csv` que contiene detalles de ventas de asientos de niños para auto en una serie de tiendas similares a las del cliente, y además los datos incluyen variables que definen características de la tienda y su localidad. La tabla de datos está formada por 400 filas y 13 columnas. Seguidamente se explican las variables que conforman la tabla.

  - `Ventas`: Ventas de asientos de niños para autos en cada localidad, en miles de unidades.
  - `PrecioCompt`: Precio promedio por asiento de niño cobrado por la competencia en cada localidad.
  - `Ingreso`: Nivel de ingreso promedio de los habitantes de la región, en miles de dólares.
  - `CercaniaEsc`: Índice que indica que tan cercana está la tienda a zonas escolares.
  - `Publicidad`: Presupuesto que asigna cada tienda a publicidad, en miles de dólares.

- `Poblacion`: Tamaño de la población en cada región, en miles.
- `Precio`: Precio cobrado por la tienda por los asientos de niño para auto.
- `CalidadEstant`: Indica la calidad de ubicación de los asientos de niño en los estantes de la tienda.
- `Edad`: Edad promedio de los habitantes de la localidad.
- `Educacion`: Años de aducación promedio de los habitantes de cada región.
- `Urbano`: Indica si la tienda está localizada en una zona urbana o no (1 = Sí, 0 = No).
- `USA`: Indica si la tienda está ubicada en Estados Unidos o no (1 = Sí, 0 = No).
- `Desarrollo`: Índice de desarrollo de cada localidad.

Cargue la tabla de datos en `R` y no elimine los NA. En caso de ser necesario, recodificar las variables de forma adecuada. Para medir el error tome un 20 % de la tabla de datos. Realice lo siguiente:

1. Genere un modelo de regresión con KNN usando cada unos de los kernels disponibles. Identifique el kernel que da un mejor resultado en la tabla de testing e interprete las medidas de error.
2. Esta tabla contiene algunas variables muy correlacionadas, descarte al menos 2 variables predictoras e indique la razón. Corra nuevamente una regresión con KNN e identifique el kernel que da mejores resultados. ¿Mejora el resultado respecto al modelo seleccionado en el inciso anterior?
3. Para los modelos de KNN, Regresión Lineal Múltiple, Lasso yRidge póngalos a competir y obtenga un modelo ganador ¿Cuál de estos modelos prefiere usar para interpretar los resultados obtenidos?

- **Pregunta 5:** [25 puntos] En este ejercicio usaremos la tabla de datos que viene en el archivo `Uso_Bicicletas.csv`. Este es un conjunto de datos de usuarios de la empresa de alquiler de bicicletas por horas *Capital Bike* en Washington D.C. Las columnas de la tabla de datos son:

  - `Fecha`: No es una variable,.
  - `Estacion`: Estación del año, es una variable categórica ordinal por lo que se puede dejar como numérica.
  - `Hora`: Hora del día, es una variable categórica ordinal por lo que se puede dejar como numérica.
  - `Feriado`: Si el día es feriado o no, es una variable categórica.
  - `DiaSemana`: Día de la semana, es una variable categórica ordinal por lo que se puede dejar como numérica.
  - `DiaTrabajo`: Si el día es de trabajo o no, es una variable categórica.
  - `TipoClima`: Tipo de clima, es una variable categórica.
  - `SensacionTermica`: Sensación térmica, es una variable numérica.
  - `TemperaturaReal`: Temperatura real, es una variable numérica.
  - `Humedad`; Humedad relativa, , es una variable numérica.

- `VelocidadViento`: Velocidad de viento, es una variable numérica.
- `UsuariosCasuales`: Cantidad de usuarios en día pero que no están registrados como clientes. Es una variable numérica.
- `UsuariosRegistrados`: Cantidad de usuarios en día que sí están registrados como clientes. Es una variable numérica.
- `TotalUsuarios`: Cantidad total de usuarios en día.

La variable a predecir es `TotalUsuarios`.

1. Cargue la tabla de datos en `R`. Asegúrese de codificar adecuadamente las variables y de ignorar las columnas `Fecha`, `UsuariosCasuales` y `UsuariosRegistrados`. Además asegúrese de seleccionar la variable `TotalUsuarios` como la variable a predecir. Use para entrenar el modelo el 80 % de los datos.
2. Corra un modelo de KNN en `traineR` con los parámetros por defecto e incluyendo todas las variables predictoras. Interprete las medidas de error.
3. Repita el item 2, pero esta vez seleccione las 4 variables que, según su criterio, tienen mejor poder predictivo. ¿Mejoran los resultados?
4. Genere un Modelo de Regresión usando K vecinos más cercanos para cada uno de los siguientes núcleos: `rectangular`, `triangular`, `epanechnikov`, `biweight`, `triweight`, `cos`, `inv`, `gaussian` y `optimal` ¿Cuál produce los mejores resultados?

Entregables: Debe entregar un documento autreproducible HTML con todos los códigos y salidas, incluya pruebas de ejecución de las funciones programadas. No olvide poner un título para cada pregunta. Las demostraciones las puede entregar en papel a mano.