MAESTRÍA EN CIENCIA DE DATOS
Análisis Cuantitativo
Examen 3

Nombre: _____ Código: _____ Nota: _____

Profesor: **Santiago Ortiz - Henry Velasco** Grupo: **01** Fecha: _____ de 20__

Notas:

- Todas las respuestas, gráficas, tablas y operaciones deben ser debidamente justificadas.
- La información que sea obtenida de alguna fuente debe ser citada y referenciada en el documento a entregar.

1) Considere el conjunto de datos **"Boston Housing Data"** presentados en Harrison and Rubinfeld (1978). Defina como variable respuesta a la columna *MEDV*. Realice una partición 80-20, donde el primer 80 % de los datos son datos de entrenamiento y el restante 20 % son datos para prueba.

   Genere los modelos de regresión por regularización **Ridge**, **LASSO** y **Elastic-Net** para los datos de entrenamiento. Encuentre los valores óptimos de α^* y λ^* junto a su respectiva gráfica de evolución de los coeficientes de regresión. Compare los modelos en términos de la selección de variables, interprete los coeficientes y escriba la ecuación ajustada de regresión para cada caso. Finalmente, realice una predicción con las observaciones de prueba y determine cual de los tres modelos es el mejor en capacidad predictiva **(RMSE)**.

2) El conjunto de datos **"YearPredictionMSD"** contiene información sobre canciones de música popular y el año en que se grabaron. Incluye 515345 observaciones y 90 características, como la intensidad media del sonido, la varianza del espectro de frecuencia y la correlación entre las características espectrales. El objetivo es predecir el año en que se grabó la canción.

   - Carque el conjunto de datos usando la función `read_csv` del paquete **pandas** y el como primer argumento el Link, use como segundo argumendo *header = None*.
   - Divida el conjunto de datos en características o variables explicativas X y variable objetivo Y, tenga en cuenta que se quiere modelar el año en que se grabó la canción.
   - Reduzca la dimensión de las variables. Para ello, use un modelo de regresión **LASSO** con un coeficiente de penalización de 10, para extraer características importantes del conjunto de variables explicativas.
   - Con el conjunto de variables reducido, ajuste un modelo de regresión OLS e interprete su significancia y su R^2_{adj}.
   - Revise los supuestos de los errores, y con los hallazgos del ítem anterior, concluya sobre la conveniencia de usar este modelo para predecir el año de grabación de la canción.

3) El conjunto de datos conocido como **"California Housing Dataset"** puede ser cargado del paquete **sklearn**. La variable objetivo es el valor medio de la vivienda para los distritos de California, expresado en cientos de miles de dólares ($100000). Este conjunto de datos se derivó del censo de EE.UU. de 1990, usando como unidad de censo el grupo de bloques. Un grupo de bloques es la unidad geográfica más pequeña para la que La Oficina del Censo de EE.UU. publica datos de muestra (un grupo de bloque generalmente tiene una población de 600 a 3000 personas).

   Un hogar es un grupo de personas que residen dentro de una casa. Dado que el promedio. El número de habitaciones y dormitorios en este conjunto de datos se proporciona por hogar, estas

columnas pueden tomar valores sorprendentemente grandes para grupos de bloques con pocos hogares y muchas casas vacías, como centros vacacionales.

- Lea el conjunto de datos usando la función **fetch_california_housing** del paquete **sklearn.datasets**, guardelos en una variable llamada **california_housing** y con el comando *print(california_housing.DESCR)* observe la descripción general del dataset y en especial qué es cada una de las variables de entrada.
- Separe las variables explicativas X de la variable respuesta Y, para acceder a ellas use los comandos *california_housing.data* y *california_housing.target*. Considere la conveniencia de incluir las variables Longitud y Latitud al modelo. Haga un análisis exploratorio de las correlaciones entre las variables y comente al respecto.
- Ajuste un modelo de regresión **Elastic-Net** con un coeficiente de penalización pequeño, iterativamente ajuste este valor para eliminar variables explicativas y corregir el problema de multicolinealidad, en cada iteración calcule las correlaciones de las variables explicativas y pare cuando no se encuentren correlaciones altas.

4) El fichero de datos `'Dengue_Data.xlsx'` contiene información epidemiológica de los casos de Dengue en el Departamento de Antioquia. Estos datos contienen tanto información socio-económica como clínica de las personas que resultaron infectadas y desarrollaron Dengue o Dengue Hemorrágico. Para una completa descripción de los datos y/o fenómeno estudiado, remítase al siguiente artículo *Identification of Hazard and Socio-Demographic Patterns of Dengue Infections in a Colombian Subtropical Region from 2015 to 2020: Cox Regression Models and Statistical Analysis*. Realizar.

   - Utilizando solo las variables socio-demográficas, ajuste un modelo Logit y los modelos Logit-Ridge, Logit-LASSO y Logit-Enet (con sus parámetros óptimos, por supuesto) para predecir si una persona va a desarrollar "DENGUE" o "DENGUE GRAVE". Interprete los resultados de cada modelo y compárelos; defina que variables son las más importantes par predecir el estado categórico modelado, muestre los gráficos de penalidad y de evolución de coeficientes. Concluya sobre el fenómeno estudiado y a información del artículo.
   - Realice el mismo ejercicio anterior, solo que ahora considere como variables explicativas las variables de tipo clínico/médico para modelar si una persona requiere o no ser hospitalizada. Realice los mismos análisis y procedimientos. Concluya en función de la información presentada en el artículo.

Pautas

- Entregar un documento de `RMarkdown/Jupyter` (en PDF) con la solución y rutinas de código empleadas (fecha máxima de entrega: Noviembre 13 hasta las 23:30). Enviar por el buzón Intu asignado.
- El documento debe contener todos los procedimientos, códigos y gráficos necesarios que den debida justificación a lo realizado. Sin embargo, consolide el documento única y exclusivamente con información relevante, evite mostrar salidas de códigos innecesarias, warnings, errores, etc.
- Realizar en equipos conformados por 3-4 participantes (mandatorio).