Trabajo Práctico N° 1

Computación Científica Actuarial

Docente: Rodrigo Del Rosso - Colaboradores: Santiago Silva - Auza - Sanchez Gavier - Ponce

Deadline: 22/11/2020

Consideraciones generales

El objetivo de este primer trabajo práctico es simular una situación de la vida real en la que se solicita analizar un conjunto de datos y posteriormente entregar un reporte exponiendo los resultados.

Es en base a esa idea que se solicita, aparte de entregar el script con los comandos en R, un muy breve informe describiendo las variables del dataset, comentando toda información que considere relevante acerca de las mismas. Y lo mismo para el modelo de regresión que se solicita al final del segundo ejercicio.

No olviden que tienen el grupo para hacer todas las consultas que consideren necesarias tanto entre ustedes como a nosotros.

Pueden usar cualquier paquete o diseñar cualquier función adicional que consideren necesario, siempre indicando el uso de los mismos.

- Algunas cuestiones prácticas:

  1. La fecha de entrega es inclusive. Tienen hasta las 23:59 de ese día para entregar el trabajo.
  2. Cualquier entrega tardía será penalizada, descontando un 30% de la nota obtenida.
  3. El trabajo es grupal. Con un mínimo de 2 y un máximo de 4 integrantes. Por lo tanto seleccionen bien a lxs compañerxs del grupo para que no haya free riders. Aquel grupo conformado en este trabajo será el utilizado en el siguiente trabajo. Por lo tanto no se admiten modificaciones de integrantes.
  4. Deberán entregar el informe en formato pdf en el siguiente **link** con el siguiente formato: "Informe_TP1_Grupo1.pdf" y cualquier otro archivo que deseen subir lo deberán hacer en el mismo formato.
  5. El script en formato .R deberán ubicarlo como link a algún repositorio en el pdf. El campus no permite el ingreso de un archivo en dicho formato.
  6. El número de grupo será asignado cuando completen los integrantes en el siguiente **formulario**.

Descripcción del dataset:

El dataset seleccionado consiste de evaluaciones realizadas por expertos sobre distintos sabores de chocolate, así como información relacionada al origen y producción de los mismos. El mismo fue confeccionado con la intención de tratar de responder a una serie de preguntas las cuales se van a plantear más adelante en las consignas.

El criterio de evaluación fue el siguiente:

- 5 = Elite (trasciende los límites de lo ordinario)
- 4 = Premium (desarrollo superior de sabor, cáracter y estilo)
- 3 = Satisfactorio (3.0) a Remarcable (3.75) (bien hecho, con atributos especiales)
- 2 = Decepcionante (Sabor aceptable, pero presenta una o más falencias significativas)
- 1 = Implacentero (muy desagradable al gusto)

Resulta importante destacar que al ser un dataset público, entendemos que existen notebooks o trabajos que hayan utilizado el mismo. Por lo cual resulta importante aclarar que **cualquier similitud significativa con los mismos resultará en que la calificación para el trabajo sea insuficiente**.

Consignas

Ejercicio N°1: Análisis Exploratorio del Dataset y Preparación de Datos

1. Importar el **dataset**, analizar la distribución de las variables del mismo, los momentos absolutos y centrados de las mismas, indentificar si alguna presenta outliers. El dataset tiene datos faltantes (NA), indicar cómo trataría los mismos y porqué.
2. Luego analizar las relaciones entre las mismas y la calificación que se le dio a cada barra de chocolate, indicar y comentar que relaciones se pudieron observar, en el caso que se hayan podido encontrar relaciones.
3. Como último paso de este primer ejercicio la consigna es: mediante todo el entendimiento que se tiene sobre el dataset, producto de los análisis realizados en los incisos anteriores, se tratará de dar respuesta a las tres preguntas por las cuales se construyó este dataset.

a. ¿Dónde se producen los mejores granos de cacao?

b. ¿Qué paises producen las barras de cacao mejor con mejor calificación?

c. ¿Qué relación hay entre el porcentaje de cacao en una barra y su calificación?

Ejercicio N°2: Desarrollo de Funciones y Regresión

El objetivo final de este ejercicio será entrenar un modelo que permita entender que combinación de las variables sobre las que se posee información produce la barra de chocolate con la mejor calificación. Para esto se plantearon las siguientes consignas:

1. Hay varios enfoques posibles al momento de definir la variable objetivo: uno es dejarla como una variable continua y la otra es transformarla en una variable discreta, para eso se solicita en una copia del dataset original, transformar la variable 'rating' a una variable que tome el valor 1 si la calificación fue igual a 5 y 0 en caso contrario. Luego guardar tanto el dataset original como el dataset con la variable transformada en una lista, donde cada dataset es un elemento de la misma.

2. En segundo lugar, hay que analizar como tratar a las variables que se utilizarán para realizar los modelados, y las transformaciones que se tienen que realizar para utilizar las mismas, en ambos casos hay variables categóricas y hay que definir un tratamiento para las mismas. En el caso del dataset que se utilizará para modelar la variable objetivo de forma continua lo que se sugiere es identificar o tratar de ordenar los niveles dentro de las variables categóricas en grupos donde las calificaciones son similares y aplicar una transformación numérica. Es decir, si se identifica n grupos donde el grupo 1 tiene las peores calificaciones y el grupo n tiene las mejores calificaciones, entonces se tiene que crear una variable que si el elemento pertenece al grupo 1 tome el valor 1, si pertenece al grupo 2 tome el valor 2 y así sucesivamente hasta el grupo n.

Para el dataset en el que la variable objetivo es binaria, el desafío es distinto: las variables categóricas se deben desagregar en variables "dummy" donde se generan $m-1$ columnas para los m distintos valores que puede tomar una variable. Mientras que para las variables continuas se puede hacer un ejercicio de agruparlas similar al que solicitó realizar en el inciso anterior.

Es por esto que se solicita según corresponda, realizar las transformaciones comentadas anteriormente, y guardar los resultados nuevamente en una lista con dos elementos: uno el dataset con variable objetivo continua y las demás variables transformadas según corresponda y lo mismo con el datset que tiene la variable objetivo expresada de forma binaria.

3. Luego hay que dividir el dataset en una población de entrenamiento y validación. Debido a la escasa cantidad de observaciones, el ratio recomendado es 70/30, pero se puede utilizar otro, siempre y cuando se justifique el motivo. A partir de esto se pide que de la lista que contiene los dos datasets solicitados en el inciso anterior, realizar la particion de los datos y guardar los resultados en una lista compuesta a su vez por dos listas: una con las particiones del dataset original y otra con las particiones del dataset con la variable objetivo transformada.

4. Realizar un modelo para cada uno de los datasets de entrenamiento y luego proceder a validar el mismo con su respectiva población de entrenamiento. Para la variable objetivo continua se recomienda utilizar una regresión lineal, mientras que para el caso de la variable objetivo binaria se recomienda una regresión logística.

Aclaraciones extra:

I. Si se desease utilizar otro tipo de modelo, se puede siempre y cuando se aclaren las motivaciones para utilizar el mismo.

5. Como un enfoque alternativo y aprovechado el entendimiento superior que se posee sobre el dataset luego de todos los análisis que se realizaron sobre el mismo así como las distintas formas de agrupar los niveles de las variables que lo componen, lo que se solicita es, partiendo del dataset original, previo a todas las transformaciones realizadas: intentar definir las mismas combinaciones que producen una barra de chocolate con la máxima calificación posible, pero sin recurrir a modelos, utilizando

simplemente el conocimiento que ya se tiene de como se distribuyen las variables con respecto a la calificación, dentro de cada variable qué atributo/elemento (o conjunto de elementos en caso de no ser uno solo) de la misma tiene la mejor calificación y cual/es tiene/n la peor.

6. Analizar y comparar los resultados de los 3 enfoques distintos que se vieron e indicar si sabiendo lo que saben luego de todo el desarrollo anterior y tuvieran que resolver el mismo problema, por cual enfoque comenzarían?.