***Trabajo Práctico N°1***

Computación Científica Actuarial

Alumnos:

Ezequiel Paladini (nro. de registro:)

Facundo López (nro. de registro: 883195)

Agustina Pacini (nro. de registro:889656)

Vassili Jigalov (nro. de registro:890016)

Docente: Del Rosso Rodrigo

Materia: Computación científica actuarial

Año/Cuatrimestre: 2020/2

Cod. Mat./Curso: 746/98

**Introducción**

El objetivo de este trabajo práctico es analizar un conjunto de datos suministrados, utilizando el programa Rstudio con las herramientas correspondientes. El dataset que se utilizara en este estudio, es sobre los distintos sabores de chocolate, que cuenta con la información relacionada con el origen y la producción de estos.

**Referencias del script:**

<https://drive.google.com/drive/folders/1TxZj-eK9bAqiOCQTJ17h9-6TuCdJ-K0M?usp=sharing>

**Ejercicio N°1:**

Una vez importado el dataset, se realizó un cambio en las variables, donde se modificaron las ciudades por sus respectivos países y se corrigieron aquellos que están mal escritos, para que se pueda clasificar según su país de origen. Asimismo, los datos que encontraban con más de un país o están mal escritos se les asignará un grupo llamado ¨Desconocidos¨.

Dado el paso anterior, se puede determinar que Estados Unidos es el país con mayor fabricación dado la muestra. Por otro lado, los países que producen más granos son República Dominicana, Ecuador, Madagascar, Perú y Venezuela.

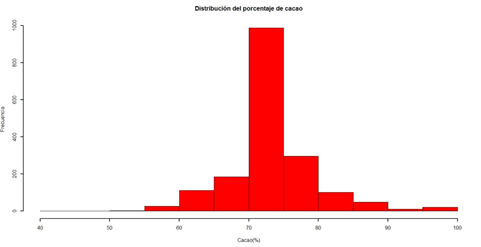
Posteriormente, se decidió armar un nuevo data frame con los datos que solamente utilizamos, eliminando aquellos que los llamamos anteriormente “Desconocidos”.

Con esta nueva estructura de datos, se logró determinar que las siguientes clasificaciones:

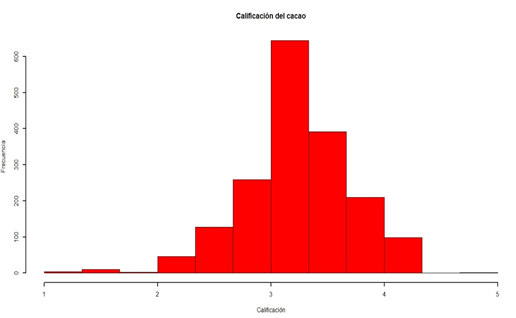
1. Chile tiene el mayor promedio de calificación.
2. Los granos cosechados en las Islas Salomón tienen el mayor promedio de calificación.

Siguiendo con el estudio, se realizó un análisis que arrojó que el coeficiente de correlación (R2) es 0.02717 con lo que se determinó que hay una correlación baja entre las variables, además presentan una relación en sentido inverso, es decir, con pendiente negativa.

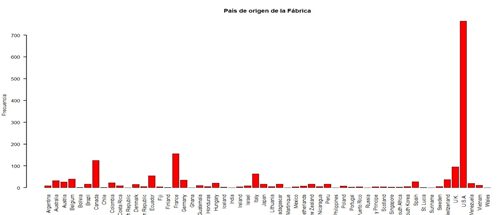
***A continuación los siguientes gráficos:***



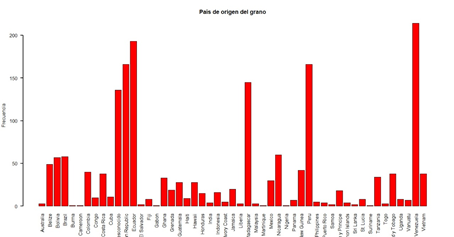
*Gráfico N°1: distribución de porcentaje de cacao*



*Gráfico N°2 : Calificación de cacao*

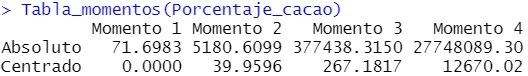


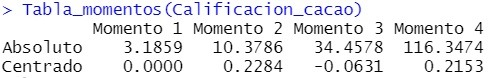
*Gráfico N°3: País de origen de la fábrica*



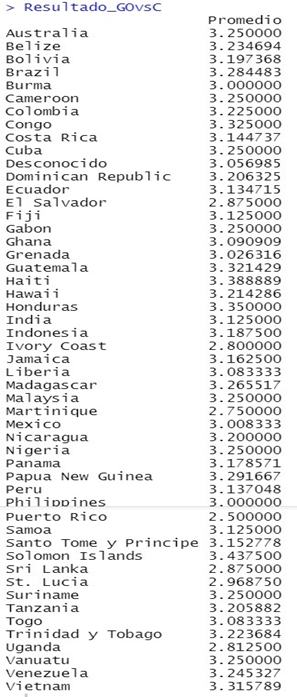
*Gráfico N° 4: País de origen del grano*

*A continuación se realizó la tabla de momentos para el porcentaje de cacao y su calificación:*

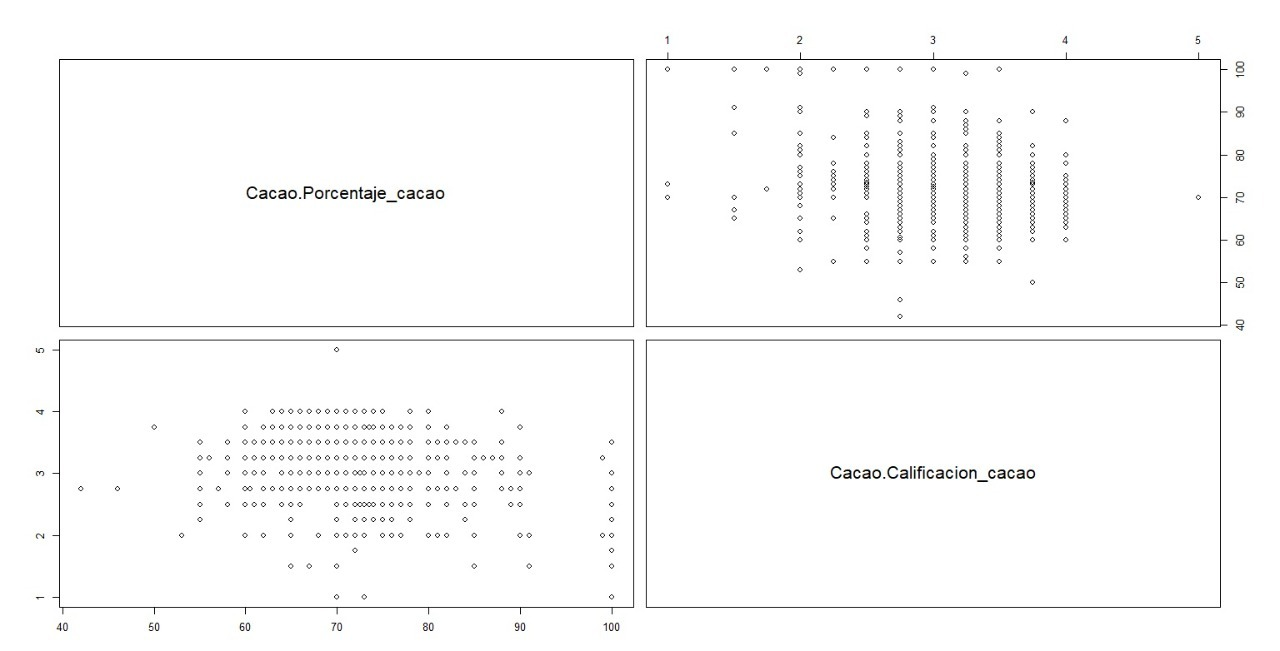


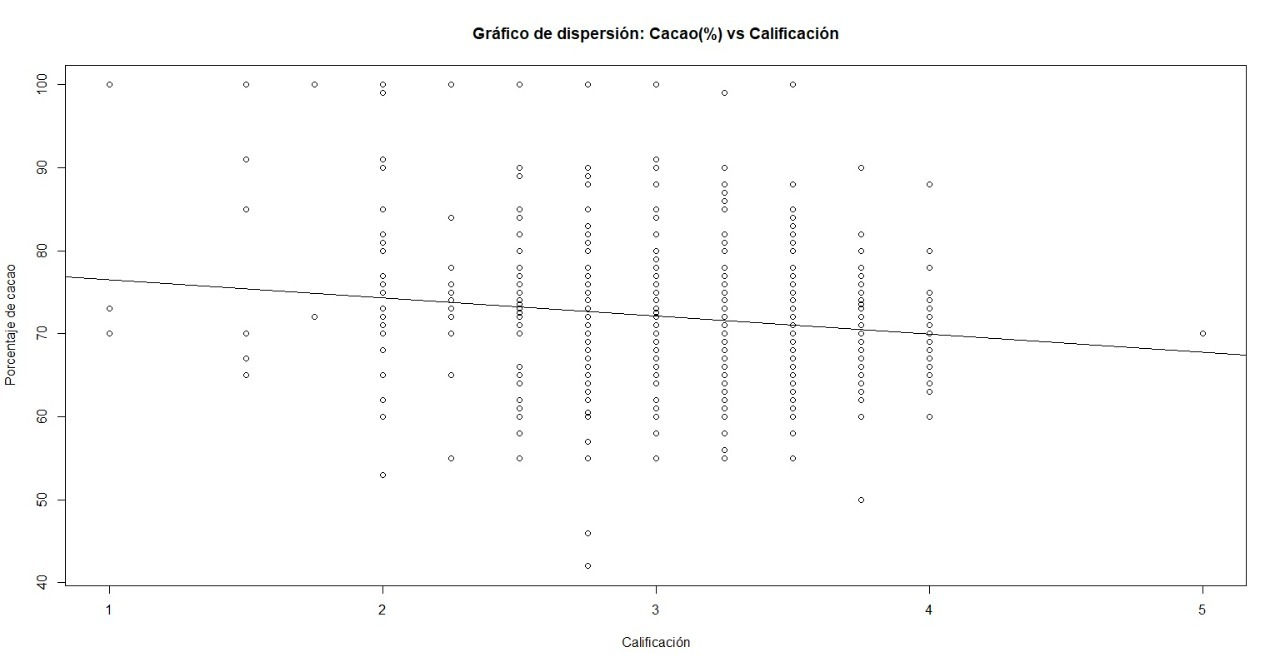


*A continuación presentamos las tablas de promedios de calificación (a la izquierda, hace referencia a los países de origen de la fábrica y a la derecha, al país de origen del grano):*



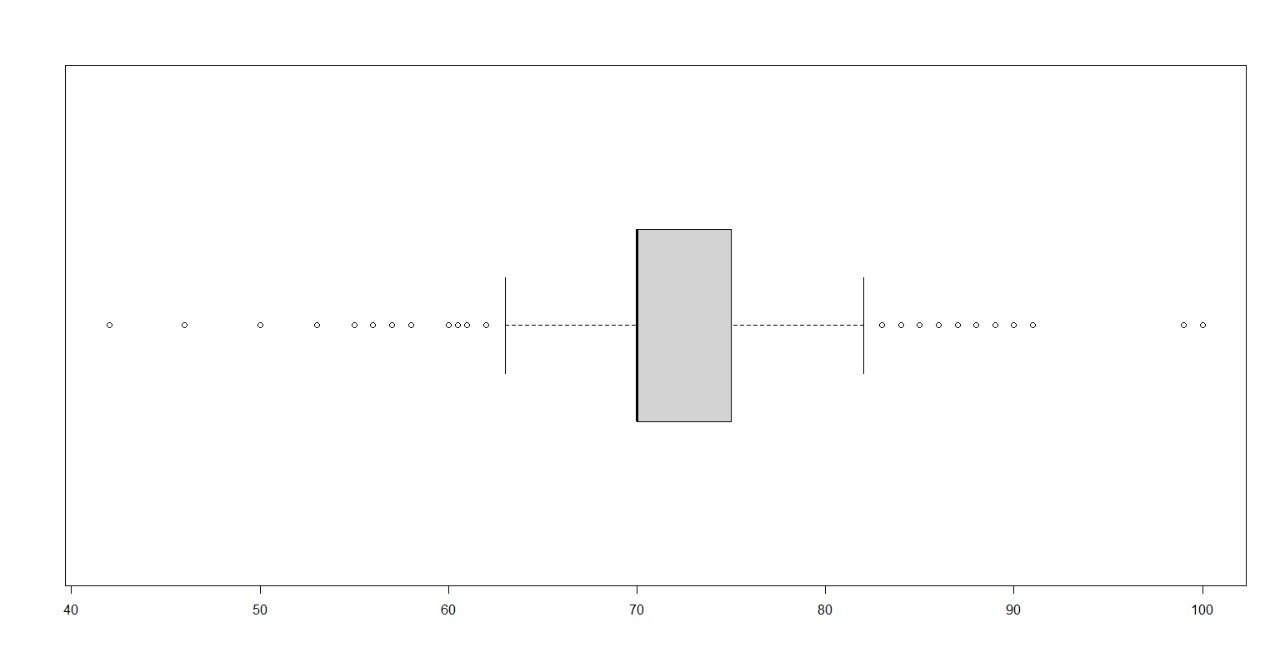
*Al comparar la relación entre porcentaje de cacao y su calificación decidimos oportuno analizar los datos y sacamos las siguientes conclusiones de los siguientes gráficos:*

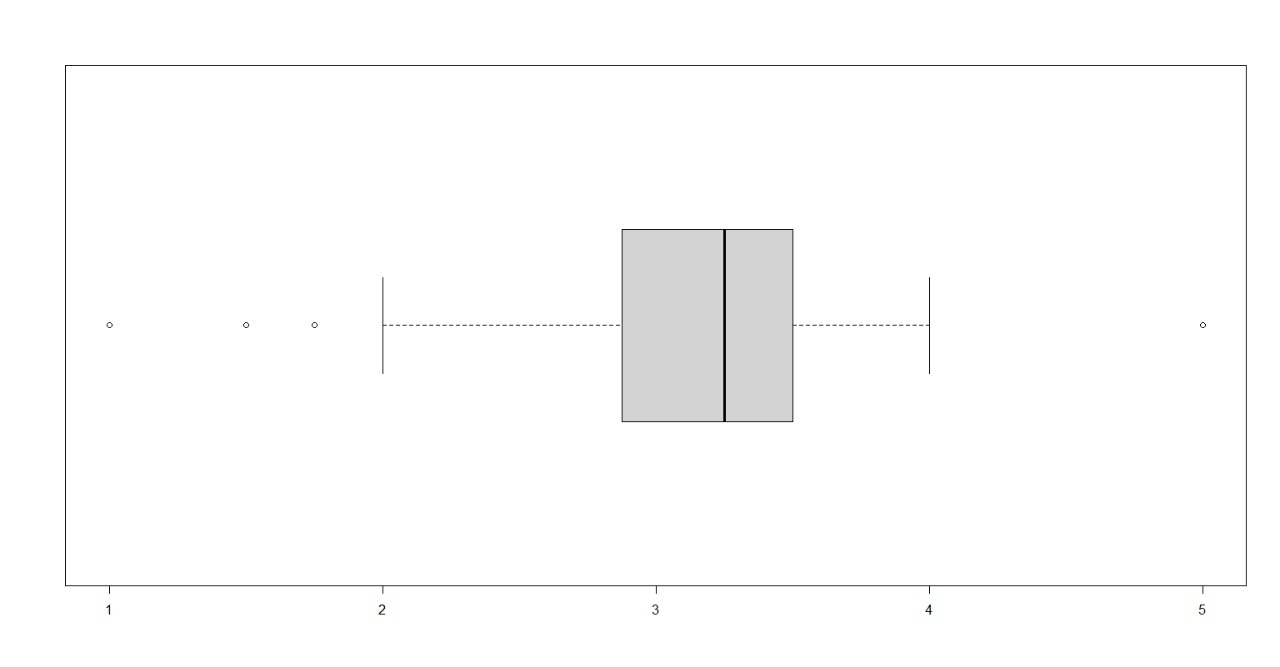




*En este último gráfico podemos ver que hay una relación indirecta entre el porcentaje de cacao y la calificación. Se podría interpretar que cuanto más amargo es el chocolate, el mismo tendrá una menor calificación. De todas maneras, se trata de una relación que no necesariamente se debe cumplir en todos los casos, ya que podemos ver que los distintos valores de la nube de puntos no se aproximan a la línea estimada.*

*Outliers de las calificaciones (Aquellos valores que están por afuera de las líneas se tratan de outliers):*





*El primer gráfico hace referencia a los outlier de la variable “porcentaje de cacao”, mientras que el segundo trata de los outliers de las calificaciones.*

*Con respecto a este último, de 1795 observaciones, se encontró la siguiente lista de outliners:*

5.00, 5.00, 1.75, 1.75, 1.50, 1.50, 1.50, 1.00, 1.00, 1.50, 1.00, 1.50, 1.75, 1.50, 1.50, 1.00, 1.50, 1.50, 1.50

**Ejercicio N°2: Desarrollo de Funciones y Regresión**

El dataset que se desarrolla en este ejercicio sobre la barra de chocolate está establecido a criterio de expertos del sabor de chocolate, por eso, es difícil estimar una predicción significativa con las variables calificativas.

Iniciamos el trabajo, utilizando la variable objetivo (V. Obj) como variable continua, y se agrupara las demás variables en intervalos cuyas calificaciones sean similares. Es por eso por lo que primero analizamos el promedio de calificación para poder determinar cada intervalo.

Una vez definidos los intervalos, reemplazamos los valores de cada variable por el número de su respectivo intervalo en el data frame de la variable continua.

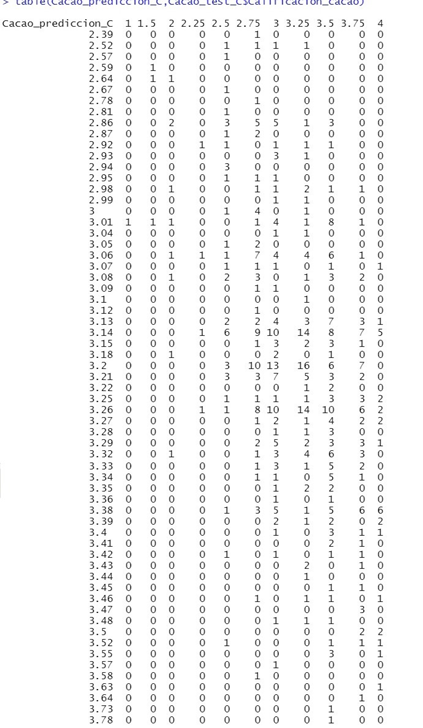
Diferente es el caso donde trabajamos la variable objetivo (V. Obj) como una variable discreta, en este modelo, las variables categóricas se deben desagregar en variables “Dummy”, que van a tomar valores entre 0 y 1. Se generan *m*-1 columnas para los distintos *m* valores que puede tomar una variable.

Posteriormente se dividió los datos en un 70% para el entrenamiento y un 30% para validación, lo cual se obtuvo de las 1795 observaciones:

1. 1258 observaciones (70%) van a ser usados para prueba
2. 537 observaciones (30%) van a ser usados para prueba

Luego se realiza un modelo para cada uno de los dataset de entrenamiento y proceder a validar el mismo con su respectiva población de entrenamiento. Este modelo consta en dos partes, por un lado, el modelo de variable continua que nos muestra un modelo válido, ya que rechaza la hipótesis nula. Podemos determinar que el 14,65% de la variable de la calificación es predicha por el modelo teniendo en cuenta las variables independientes.

A continuación, se mostrará una tabla donde veremos en la primer columna las predicciones que realizó nuestro modelo:



Cómo se mencionó, la primera columna trata de las predicciones, mientras que las demás son los posibles valores de las calificaciones que nos da la muestra. Por lo tanto, podemos ver que para la única observación de nuestro data frame que usamos para la validación y que poseía la calificación de 1, nuestro modelo predijo que su valor sería de 3,01.

Esto se debe a que nuestro modelo sólo es efectivo en un 14,65%, por lo que no siempre acertará en la predicción. De todas maneras, hubo observaciones donde nuestro modelo los predijo de forma correcta, por ejemplo, una de las observaciones de la muestra para validar tenía una calificación de 3.25 y el modelo realizó una correcta predicción.

Por otro lado, el modelo de variable objetivo-discreta se puede examinar que, al haber muy pocas calificaciones de 5 en todo el dataset, esto hace que la probabilidad de encontrar una observación cuya calificación sea exactamente de 5 sea realmente muy baja. Esto ayuda al modelo que predijo que todas las observaciones tenían una calificación de 0 y, si bien en el grupo de validación todas las observaciones también eran 0, su buena predicción se debió a la escasez de las calificaciones 5 más que una eficiencia del modelo.

Partiendo del dataset original, podemos distinguir sólo dos observaciones en el dataset original que llega a la máxima clasificación, ambas fueron creadas por la empresa Amedei en Italia con la fórmula de 70% de cacao. Pero se diferencian en el país de origen y el blend utilizado, uno de los granos proviene de Venezuela con el blend trinitario, y el otro, es desconocida ambas informaciones.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Company (Maker-if known) | Cocoa Percent | Company Location | Rating | Bean Type | Broad Bean Origin |
| Amedei | 70% | Italy | 5 | Trinitario | Venezuela |
| Amedei | 70% | Italy | 5 | Blend |  |

En cuanto a las observaciones que obtienen la mínima clasificación, no se encuentra una similitud entre ellas que sea relevante. Por un lado, dos de ellas cuenta con un chocolate de 70% de cacao, pero también las mejores clasificadas así que no creemos que sea suficiente esa relación, pero sí podemos notar que tres de ellas la locación de la compañía es en Bélgica

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Company (Maker-if known) | Cocoa Percent | Company Location | Rating | Bean Type | Broad Bean Origin |
| Callebaut | 70% | Belgium | 1 |  | Ecuador |
| Claudio Corallo | 100% | Sao Tome | 1 | Forastero | Sao Tome & Principe |
| Cote d' Or (Kraft) | 70% | Belgium | 1 |  |  |
| Neuhaus (Callebaut) | 73% | Belgium | 1 |  |  |

Una vez terminado los diferentes modelos de análisis, podemos llegar a la conclusión de que la mejor opción es el modelo de regresión lineal con variable continua porque se trata de un modelo más abarcativo, que contempla más variables a la hora de hacer una predicción.

Sí sólo nos basamos en la opción de tomar las mejores observaciones cuyas calificaciones fueron de 5, se podría presumir que los chocolates que poseen un 70% de cacao serían los mejor calificados, dados que justo sus dos observaciones poseen esta característica. Sin embargo, dentro de las cuatro observaciones que poseían una calificación de 1, encontramos que dos de ellas también poseían dicho porcentaje. Por lo tanto, no podemos sacar conjeturas que sean efectivas de esta manera, dado que es muy aleatorio.

Por otro lado, en relación a la opción de la variable objetivo discreta binaria, no es muy útil si tenemos en cuenta que el dataset original poseía sólo dos observaciones cuya calificación era de 5. Recordemos que planteamos este modelo realizando una transformación en la variable de la calificación, otorgando el valor de 1 a aquellas cuya calificación era de 5 y 0 para todas las demás. Entonces, si bien nuestro modelo predijo correctamente que todas las observaciones que guardamos para la validación eran de 0, como se mencionó anteriormente, esto se debió a la baja probabilidad de que se presente un valor que no sea 0, más a que a una buena tarea realizada por el modelo.

Atento a todo esto, consideramos que la mejor opción es utilizar la variable objetivo continua.