**Introducción al lenguaje de programación R**

**Proyecto Final**

La carpeta “Proyecto Final” contiene los archivos a utilizarse en el proyecto final, mismo que debe presentarse para culminar satisfactoriamente el curso. La carpeta contiene los archivos “student\_mat.csv” y “Students dictionary.txt” que hacen referencia a la base de datos y al diccionario de variables, respectivamente.

**ACTIVIDADES**

Se ha planteado un problema de regresión lineal con una base de datos que debe ser pre procesada previo al proceso de modelización. El objetivo es predecir la calificación final de un grupo de estudiantes en su clase de matemática. La variable a predecir es “**G3**”. Contamos con un sinnúmero de atributos económicos y sociales que podrían servir como predictores. El archivo “Students dictionary.txt” presenta el detalle completo de las variables con las que se va a trabajar.

El proyecto seguirá las siguientes actividades:

* 1. Importar la base de datos
  2. Utilizando las funciones para datos perdidos, compruebe si existen datos perdidos en la base de datos.
  3. Analizar la estructura de la base de datos y verificar que cada variable corresponda a la especificación deseada: numeric, factor, character, etc. Para ello es necesario trabajar con la función str() sobre la base de datos importada y el diccionario de variables. La idea es que el tipo de columna corresponda al tipo de variable indicada por el diccionario.
  4. Utilizando la variable “sex”, cree 2 bases de datos: una base de hombres y otra de mujeres. Utilice la indexación o la función subset().
  5. Utilizando las dos bases de datos del punto anterior, analice algunas estadísticas descriptivas para las **variable G3**. Compare brevemente el comportamiento de hombres y mujeres. ¿Observa algo interesante?
  6. Utilizando la base de datos original (sin segmentar por sexo), realice 5 gráficos estadísticos cruzando la variable a predecir **G3** con las variables más relevantes de la base de datos, de acuerdo a su criterio.
  7. Utilizando la base de datos original, realice el gráfico estadístico apropiado para cruzar las variables **G3** y ausencias (**absences**). ¿Qué puede observar en esté gráfico?
  8. Utilizando la base de datos original, realice el gráfico estadístico apropiado para cruzar las variables **G3** y número de matrículas previas (**failures**). ¿Qué puede observar en esté gráfico?
  9. Utilizando la base de datos original, estime tres modelos de regresión lineal simple utilizando variables que a su criterio podrían explicar el rendimiento escolar **G3**.
  10. Interpretar los resultados de cada unos de los modelos en términos de: significancia estadística de los coeficientes, impacto de la variable **X** sobre **G3**, coeficiente de determinación (**Rˆ2**) y Error Estándar de los Residuos (**RSE**).
  11. Emplee la función sample() para segmentar la base de datos original en dos partes. La primera parte deberá contener una muestra aleatoria con el 75% de los datos originales y se guardará en un objeto llamado **train**. La parte restante (25%) deberá guardarse en un objeto llamado **test**.
  12. Utilizando la base de datos **train**, estime un modelo de regresión múltiple para explicar **G3**. Verifique que el modelo propuesto tenga coeficientes estadísticamente significativos de forma individual (prueba t) y global (prueba F).
  13. Interpretar los resultados del modelo obtenido en el punto anterior en términos de: significancia estadística de los coeficientes, impacto de las variables **X** sobre **G3**, coeficiente de determinación (**Rˆ2**) y Error Estándar de los Residuos (**RSE**).
  14. Utilizando el modelo de regresión lineal múltiple estimado en el punto l, realice una predicción utilizando la base de datos **test.**
  15. Con las predicciones del punto anterior y los valores observados de **G3** en la base de datos **test**, calcule el Error Cuadrático Medio (MSE).

**INDICACIONES GENERALES**

Como resultado de las actividades anteriormente mencionadas, se elaborará un informe que detalle y explique cada uno de los pasos seguidos en las actividades del proyecto. A manera de anexo, se incluirá el código de R empleado para el análisis.

El proyecto puede realizarse de forma individual o en parejas y debe entregarse hasta el día **miércoles 21 de marzo de 2018** via correo electrónico: [eduardo.marin@taryanalytics.com](mailto:eduardo.marin@taryanalytics.com)