**Proyecto 6: Activación de conceptos y Regresión Lineal**

**Objetivos de aprendizaje y competencias**

1. Objetivos de aprendizaje:

a. Clasificar tipos de variables.

b. Realizar análisis exploratorio de datos.

c. Establecer un criterio adecuado para el cambio de nulos y outliers.

d. Interpretar información estadística con gráficos como distplot, histogramas, etc.

f. Entender e interpretar el concepto de correlación o covarianza mediante gráficos de dispersión, heatmap, matriz de covarianza, etc.

g. Conocer la regresión lineal y sus fundamentos.

h. Interpretar los parámetros estimados en la regresión.

i. Conocer y ser capaz de interpretar estadísticos de bondad de ajuste y coeﬁcientes.

j. Implementar regresión lineal con statsmodels

2. Competencias que se desarrollan:

a. Analizar un problema para formular una solución.

b. Idear un algoritmo para resolver un problema..

c. Presentar los resultados de un proyecto.

**Descripción**

Realice un análisis exploratorio sobre la BBDD de la salud financiera (Credit.csv) mediante el uso de tablas y librerías de gráficos. Si ud lo estima conveniente, genere una limpieza de los datos. En base a lo anterior, deberá generar Regresiones Lineales donde se incluya conclusiones basadas en su análisis.

Para efectos prácticos del ejercicio, nos interesa saber sobre los determinantes de la variación en el promedio de deuda en tarjeta de crédito. A lo largo del ejercicio se generarán ecuaciones econométricas para inspeccionar los efectos de distintos atributos.

**Antecedentes**

En esta actividad trabajaremos con un registro de clientes de un banco e información sobre su salud financiera. La base contenida en el csv `Credit` contiene las siguientes variables:

\* `Income`: Ingreso anual disponible en miles de dólares.

\* `Limit`: Límite del crédito.

\* `Rating`: Ranking del crédito.

\* `Cards`: Cantidad de tarjetas de crédito.

\* `Age`: Edad del usuario.

\* `Education`: Años de educación del usuario.

\* `Gender`: ¿Hombre o mujer?

\* `Student`: ¿Es estudiante?

\* `Married`: ¿Está casado?

\* `Ethnicity`: Etnicidad del usuario (Caucásico, Afroamericano, Asiático)

\* `Balance`: Promedio de deuda en tarjeta de crédito de usuario.

**Instrucciones:**

Repaso de contenidos previos

1. Realizar un análisis exploratorio de los datos identificando la presencia de datos nulos y establezca una normativa para tratar dichos datos, identificar variables categóricas y cuantitativas.

2. Identifique y establezca un protocolo para tratar outliers.

3. Generar análisis exploratorio basándose en variables de su interés.

4. Generar análisis de correlación. Generar Matriz y scatterplots.

5. Generar una función que dibuje un histograma y que además identifique de manera visual la media y mediana. Esta función debe ser aplicada para las variables ‘Balance’, ‘Income’, ‘Rating’ y ‘cards’. Según los gráficos generados, como es el comportamiento de estas variables.

Regresión Lineal

6. Generar 3 regresiones lineales simple para la variable “Balance” (variable dependiente) utilizando como variables independientes ‘Student’, ‘Income’, ‘Rating’ respectivamente (una para cada regresión). Identificar coeficientes de la recta y ver cuales de estas tres explica de mejor forma el fenómeno.

**7. Genere una nueva regresión lineal utilizando todos los regresores (todas las columnas). Indique cuales son sus coeficientes. ¿Cual es el que más repercute en el modelo?**

8. Utilizando `lmplot` de `seaborn`, inspeccione el comportamiento de los ~~tres~~ principales determinantes. ¿Qué comportamiento observa?.

Bonus:

9. Depure el modelo utilizando solo los regresores más significativos (pvalues =< 0.025) Genere un nuevo modelo con solo estos regresores. ¿Qué diferencias encuentra con respecto al modelo de la pregunta 7.?

**Formato de entrega:**

* Notebook formato informe. Esto quiere decir:
  + Ordenado por preguntas.
  + Que corra de principio a fin. Si este presenta error, se corregirá hasta antes del error.
  + Código bien documentado. Esto hace referencia a que deben contestar las preguntas en el notebook. Además, incluir sus conclusiones si la pregunta así lo requiere.
* Diapositivas de presentación (formato pdf).

**Plazos de entrega:**

* Entregar notebook y diapositivas **Jueves 21/11/2019 hasta las 23:59 hrs.**
* Presentación **Viernes 22/11/2019 desde las 09:00 hrs.**