Laboratorio # 4: Regression

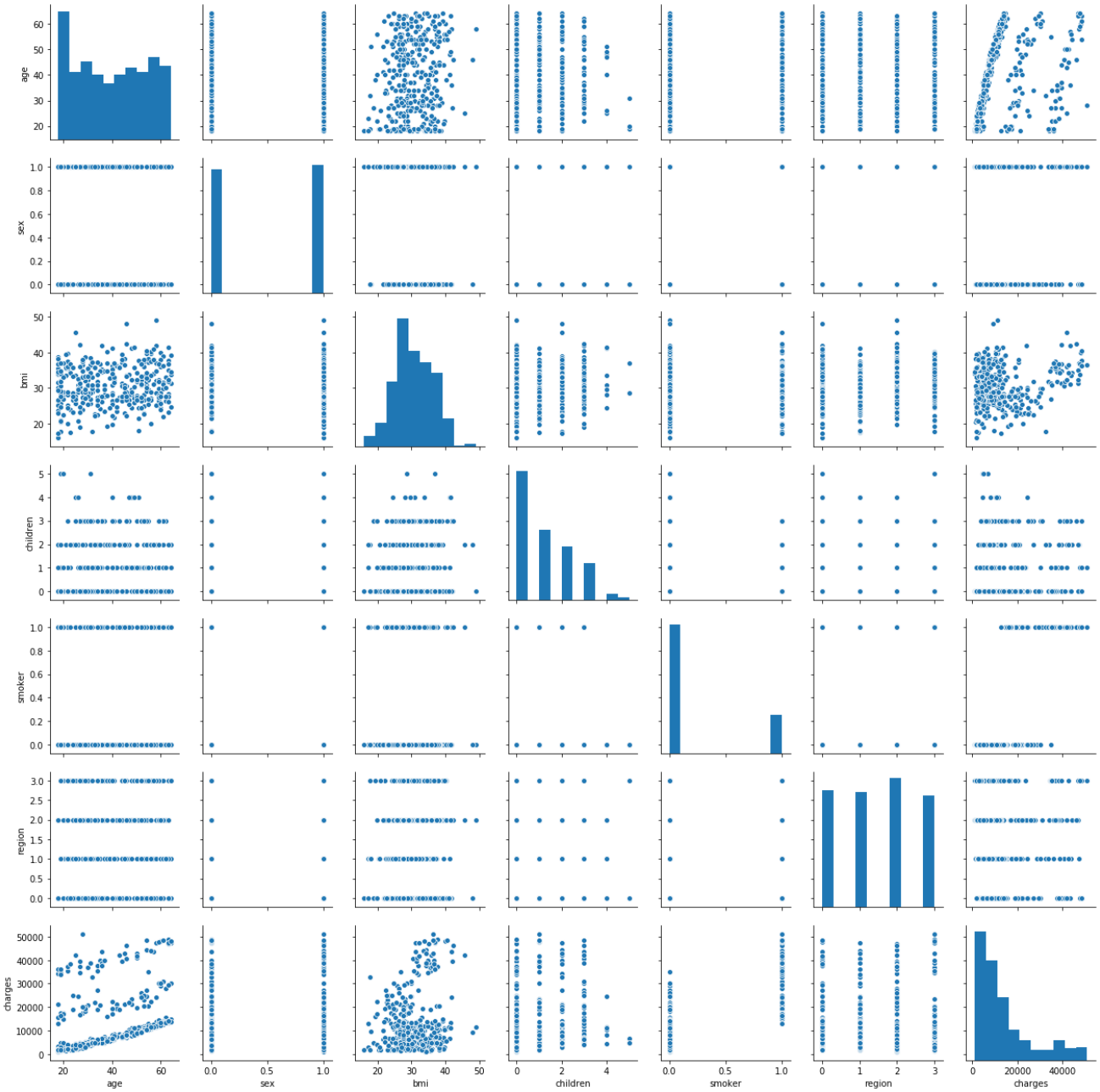
Presentado por:
Cristina María Bautista Silva (carné No. 161260)

Minería de datos

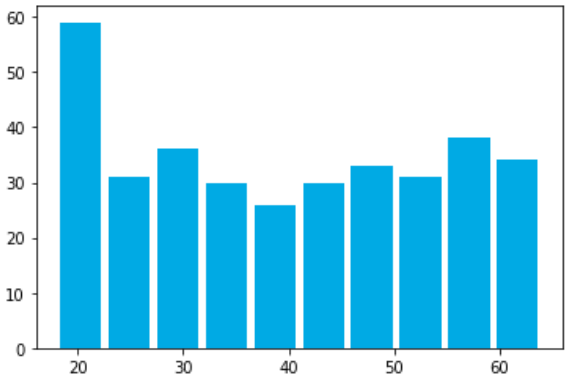
Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
Guatemala, abril de 2021

1. Información del dataset
    a. Tiene 7 columnas
    b. Todas las columnas son numéricas
    c. Tienen 348 filas
    d. 5 de las columnas son int64 y 2 de las columnas son float64
    e. No tenían valores null las columnas
    f. Datos interesantes:
        i. En la edad la persona más joven tiene 18, mientras que la más grande tiene 64, y la media se encuentra en 39.59
        ii. La muestra tienen casi un 50% de ambos sexos
        iii. Hay personas de 4 distintas regiones, realmente el dataset no dice cuales son las regiones, pero si puede ver claramente que la muestra divide a sus integrantes entre las 4 regiones
        iv. Hay muy pocas personas que tienen charges de más de 50,000. La mayoría se en el intervalo entre 0 y 10,000
2. Hipótesis u objetivo
   El objetivo es el usar el dataset para entrenar un modelo, para luego evaluarlo, aplicarle tanto regresión lineal como regresión polinómica. Para hacer lo anterior primero hay que explorarlo, para conocer más de el, además que hay que verificar que no haya valores del tipo null que no permitan hacer un buen trabajo.
3. Solución y Exploración: incluye la lógica detrás de la manipulación que realiza a los datos, retos encontrados y cualquier información relevante del proceso de exploración y delimitación del problema. Incluya gráficos y visualizaciones generadas.
    a. Lo primero fue ver como era el dataset, que tan largo era, y que tipo de datos tenía, con las funciones de head, info y describe, obtuve información importante. A lo que se puede ver que ninguna de las columnas tiene datos de tipo null, aunque para corroborar se llamo a la función de isnull() la cual dio falso, por lo que no había datos null.
    b. Lo segundo fue hacer un pairplot, para visualizar y ver la distribución de los campos, se puede apreciar en el apartado 4 imagen a. Un dato curioso es que cuando las tablas se muestras como muestras de si mismas tanto en el eje x como el eje y son histogramas, mientras que el resto son gráficos de dispersión.
    c. Los tercero fue conocer las distribuciones de las columnas como edad, sexo, región y charges, que se puede ver en el apartado 4 imágenes b, c y d. Con lo que se recopilaron varios datos como los presentados en el apartado 1 inciso f.
    d. Se realizo tanto el label Encoder como el One Hot Encoder. Lo siguiente fue separar nuestro training del test. Se uso el StandardScaler.
    e. Lo más desafiante fue hacer el Modelo de Least Square, sin la librería, me guíe con un video de Youtube para entender el paso a paso de la formula que había que programar para el ejercicio. Luego use la librería de sklearn para aplicar LinearRegression. Para el tercer modelo volví a obtener un x/y train además del x/y test para tomar todos los campos a diferencia del primero que solo tiene a bmi.
    f. Por último, realizar las gráfica de puntos con línea de dispersión con Regresión Lineal y luego con Regresión Polinómica, la segunda realizo el mismo procedimiento que en el de Regresió Lineal, se agregaron MAE, MSE y RMSE para completar el ejercicio.

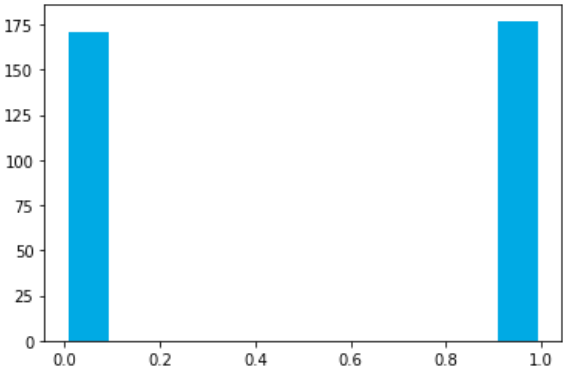
4. Resultados
    a. Pairplot de todas las columnas



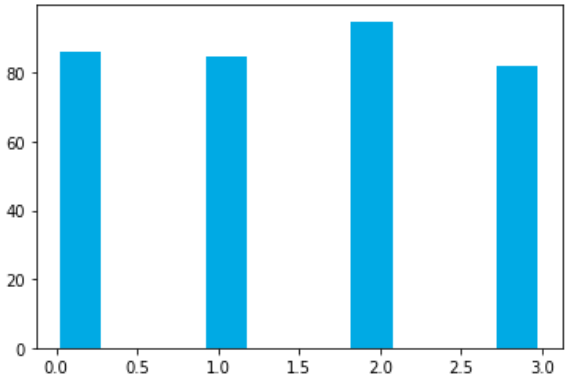
    b. Histograma de edades



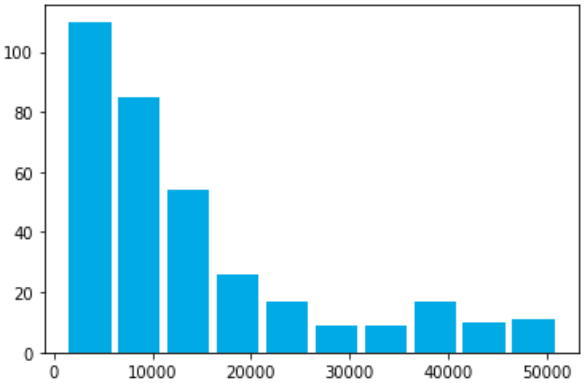
c. Histograma de sexo



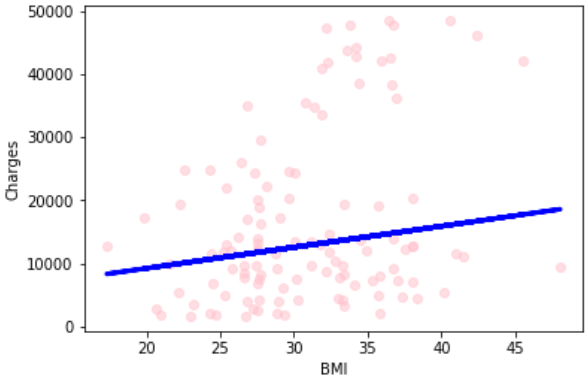
d. Histograma de la región



e. Histograma de charges



f. Gráfica puntos de test con línea de regresión



g. Gráfica puntos de test con regresión polinomial