Práctica 1

Jorge Humberto Sierra Florido
2123065656@cua.uam.mx
UAM Cuajimalpa
Ingeniería en Computación
Ciudad de México, México

María de Jesús Sánchez Zepeda
2153068423@cua.uam.mx
UAM Cuajimalpa
Ingeniería en Computación
Ciudad de México, México

RESUMEN

Aquí va el abstract de la práctica...

ACM Reference Format:
Jorge Humberto Sierra Florido and María de Jesús Sánchez Zepeda. 2021. Práctica 1. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1. INTRODUCCIÓN

En el análisis de datos es importante la limpieza de los datos previos a realizar modelos predictivos con ellos. Con limpieza nos referimos al tratamiento de los datos faltantes con el fin de no inducir un sesgo en el modelo.

Para encontrar relaciones entre dos variables continuas definidas en la misma población se puede utilizar el coeficiente de correlación y la regresión líneal. En está práctica se tienen dos conjuntos de datos de diferentes tamaños a los cuáles se les desea encontrar un modelo una ecuación que modele la relación entre variables determinadas.Para ello se limpiarán los datos de cada conjunto, posteriormente se generaran los diagramas de dispersión que nos ayudan a identificar los *outloaders*. Después se realizará una regresión lineal para posteriormente evaluar cada modelo, para ello se cálculará la eficiencia del modelo generado y el grado de error que contiene.

Texto introductorio al tema en que se enfoca la práctica y lo que se desarrollará en ella. Se debe escribir un texto que introduzca el tema de la práctica, definición del problema, los objetivos, motivación, y resultados esperados.

2. CONCEPTOS PREVIOS

Escribir conceptos teóricos empleados en el desarrollo de la práctica (fórmulas matemáticas, por ejemplo). Es un tipo de sección con todos los conceptos teóricos empleados.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

TdegC		Salnty	
Min.	1.44	Min.	28.43
1st Qu.	7.68	1st Qu.	33.49
Median	10.06	Median	33.86
Mean	10.80	Mean	33.84
3rd Qu.	13.88	3rd Qu.	34.20
Max.	31.14	Max.	37.03
NA's	10963	NA's	47354

Cuadro 1: Resumen de el dataset water.csv

3. METODOLOGÍA

Se analizaron los datasets usando el lenguaje de programación R[1] y el IDE R studio. a fin de generar los resúmenes estadísticos y gráficas requeridas, esto tras limpiarlos de información faltante.

Posteriormente se procedió al análisis de la información usando Python, ayudándonos de Scikit-Learn , Pandas y Numpy. Se requirió investigar la función LinearRegression, para aplicarla a los datasets y generar los modelos

Finalmente se evaluaran los modelos y se compararan la regresion lineal normal y la ponderada.

4. RESULTADOS

Se inicio la practica revisando los datasets proporcionados, estos corresponden a los datos de temperatura y salinidad medidas, así como las características de algunos vehículos.

El primer paso fue cargar los datos en R a fin de poder analizar los datasets. para esta carga se usaron las instrucciones del Código 1

Listing 1: lectura de los datos

```
data = read.csv("water.csv", header=TRUE)
data2 = read.table("mtcars.txt", header=
    TRUE , sep = " ")
```

Una vez cargada la información se procedió a obtener el resumen de la misma con la siguiente instrucción del Código 2, generando la salida mostrada en la Tabla 1

Listing 2: Resumen de los datos

```
summary(data)
```

En la Tabla 1 se observaron 10963 registros vacíos en el campo de temperatura y 47354 en la salinidad. Se eligió como

[1]R es un lenguaje orientado a computo estadístico y de gráficas, https://www.r-project.org/

TdegC		Salnty	
Min.	1.44	Min.	28.43
1st Qu.	7.72	1st Qu.	33.50
Median	10.06	Median	33.86
Mean	10.79	Mean	33.84
3rd Qu.	13.83	3rd Qu.	34.18
Max.	31.14	Max.	37.03

Cuadro 2: Resumen limpio de el dataset water.csv

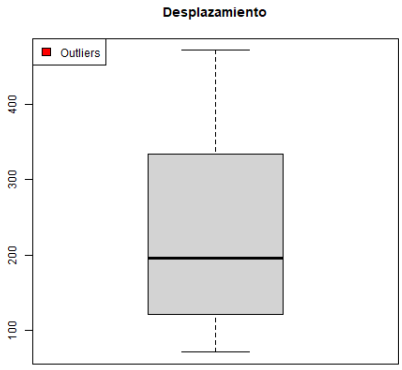

Figura 1: Esto es una imagen de ejemplo

alternativa dejar los elementos con campos NA, remplazándolos por su Mediana, esto para no afectar el dataset. Este remplazo se realizo usando el Código 3

Listing 3: Código de emplazo de los NA

```
data$Salnty[is.na(data$Salnty)] =
    33.86
data$TdegC[is.na(data$TdegC)] =
    10.06
```

El nuevo resumen genero la Tabla **??**. Donde se observan cambios mínimos.

Respecto a los datos

Imagen de prueba

5. CONCLUSIONES Y REFLEXIONES

Conclusiones generales de la práctica. Añadir una reflexión analítica por cada miembro del equipo.

REFERENCIAS