Aprendizaje Automático:
Tarea programada 01

María Auxiliadora Mora
Instituto Tecnológico de Costa Rica,
Programa de Ciencias de Datos,

June 21, 2023

Entrega: Subir al TEC-Digital un archivo comprimido que contenga: Cuaderno de Jupyter y los archivos de datos, el original y el resultante.
Modo de trabajo: Grupos de 2 personas.
Estudiante 1: _____
Estudiante 2: _____

Abstract

En el presente trabajo se repasarán aspectos básicos del preprocesamiento de datos.

1 Preprocesamiento de datos

Para realizar el trabajo se utilizará un conjunto de datos generado por la Facultad de Ciencias de la Información y la Computación Donald Bren de la Universidad de California en Irvine (datos adjuntos). El conjunto de datos puede ser utilizado para entrenar modelos para predecir la edad de los abulones (moluscos también conocidos como orejas de mar) a partir de mediciones físicas. Comúnmente, la edad de un abulón se determina cortando la concha a través del cono, tiñéndola y contando el número de anillos a través de un microscópio, una tarea que requiere mucho tiempo. Sin embargo, es posible utilizar datos morfológicos del individuo, que son más fáciles de obtener y permiten predecir la edad este. Una descripción detallada de los datos está disponible en:
https://archive.ics.uci.edu/ml/datasets/abalone.

Utilice el conjunto de datos y prepare un cuaderno Jupyter para realizar actividades de preprocesamiento de los datos. Utilizando celdas de texto y de código, realice las siguientes acciones. Para cada requerimiento debe insertarse primero una celda de texto en la que se explica qué se va a realizar y por qué elige ese método y una celda de código en Python en la cual realiza el procesamiento.

Recuerde que puede usar la biblioteca scikit-learn.
Los requerimientos son:

1. Cargar los datos y visualizar su contenido, utilice los gráficos y funciones que considere aportan para entender cómo se comportan las variables. (5 puntos)
2. El atributo Shucked es ordinal, preprocéselo y justifique su selección. (5 puntos)
3. El atributo Length tiene valores faltantes, realice imputación de datos y justifique su selección. (5 puntos)
4. El atributo Rings tienen valores atípicos (outliers) proponga cómo corregirlo (investigue) y hágalo, justifique su selección. (5 puntos)
5. El atributo Sex es nominal, preprocéselo y justifique su elección. (5 puntos)
6. Muestre un histograma de cada variable, comente lo que le parece relevante de la graficación. (5 puntos)
7. Explore el siguiente material asociado a sesgo en NLP: Chang, K. W., Prabhakaran, V., & Ordonez, V. (2019). Bias and Fairness in Natural Language Processing [tutorial]. Disponible en http://web.cs.ucla.edu/~kwchang/talks/emnlp19-fairnlp/
   (a) ¿Cuál es su opinión al respecto del tema? Respalde su respuesta con 3 referencias (3 puntos)
   (b) ¿Qué estrategia se debería definir para evitar los problemas de sesgo y equidad en proyectos de Ciencias de datos? Respalde su respuesta con 3 referencias (5 puntos)
   (c) Incluyan las referencias bibliográficas en formato APA (1 punto).