Trabajo Práctico Nro. 4

Sistemas de Soporte para la Toma de Decisiones

La entrega de los ejercicios deberá realizarse en formato. ipynb. Se recomienda incorporar comentarios explicativos que fundamenten las decisiones adoptadas, las suposiciones consideradas y cualquier otro aspecto relevante para la comprensión del trabajo realizado.

Calidad de los Datos

1. Limpieza de los datos del dataset Fifa21 en Python con Pandas.
    a. Eliminar las columnas que no aportan datos útiles como la URL de la foto y la URL de la fuente.
    b. Dar formato a los datos de las columnas `Club`, `Hits`, `Weight`, `Height`, `Joined`, `Value`, `Release Clause` y `Wage`. Analice cuál es el mejor formato de datos para cada columna y aplique las transformaciones necesarias. Considere la posibilidad de modificar el nombre de la columna.
2. Identificar si el dataset tiene valores duplicados. ¿Es necesario eliminarlos?
3. Investigar las siguientes técnicas para mejorar un dataset con valores faltantes (missing values):
    a. Descarte o *listwise deletion.*
    b. Imputación simple.
    c. Imputación múltiple.
    d. Interpolación.
    e. Modelado predictivo.
4. Identificar los valores faltantes de la columna `Hits` y completar la columna utilizando algunas de las técnicas mencionadas.
5. Basado en la columna `Value`, extraer en un nuevo archivo .csv una muestra con el percentil 25% de los jugadores más valiosos de la FIFA en 2021.

Aprendizaje Automatizado

Preprocesamiento de datos

El ejercicio número 2 de este apartado tiene carácter entregable y debe ser realizado de forma individual. La fecha límite para la entrega es el 08/09/2025 a las 23:59 hs. Cabe señalar que dicha fecha podrá ser modificada por la cátedra si así se considerara necesario.

1. A la hora de preparar los datos antes de ser entrenados existen tres acciones básicas a realizar. Explicar en qué consiste cada una de las siguientes y cómo se implementarían en Python:
    a. *Mean subtraction.*
    b. *Normalisation*.
    c. *Standardization.*

2. En el análisis de datos multidimensionales, resulta fundamental aplicar métodos de reducción de dimensionalidad que permitan preservar la mayor cantidad de información significativa posible.
    a. Lleve a cabo una investigación detallada sobre las técnicas PCA y t-SNE, analizando sus fundamentos teóricos, las principales diferencias entre ambas y las aplicaciones específicas de cada una.
    b. Aplicar estas técnicas sobre un dataset a elección.
    c. ¿Cuántos componentes se deben usar para explicar la variabilidad del 70% y 80% de los datos en PCA?
    d. Graficar los resultados obtenidos.
    e. ¿Qué ventajas tiene aplicar la reducción de dimensionalidad sobre nuestro dataset?