**Dataset Utilizados.**

1. Dataset categórico: <https://archive.ics.uci.edu/dataset/936/national+poll+on+healthy+aging+(npha)>
2. Dataset con Missing Values: <https://archive.ics.uci.edu/dataset/105/congressional+voting+records>
3. Dataset numérico (punto 6): <https://catalog.data.gov/dataset/water-quality-data-41c5e>

(Notebook limpieza de estos datos <https://deepnote.com/workspace/Dani-e9d57001-4e9b-435d-9a2a-2cd1539c64d5/project/Limpieza-Datos-KD-IAII-405e11bf-4ccd-4c99-bb69-c9e285b9c874/notebook/Notebook-1-90c7b507f7374b89a690aa6d69c4fccd?utm_source=share-modal&utm_medium=product-shared-content&utm_campaign=notebook&utm_content=405e11bf-4ccd-4c99-bb69-c9e285b9c874> )

**Análisis arboles obtenidos.**

A continuación, se muestran el numero de nodos y hojas que se tienen en los arboles resultantes de cada data set.

1. A close up of a text

   AI-generated content may be incorrect.
2. A white box with black text

   AI-generated content may be incorrect.
3. A close up of a text

   AI-generated content may be incorrect.

Vemos que el árbol mas grande fue el del dataset a). Y el mas simple o pequeño, fue el del dataset b). Pues aunque ambos tienen todos sus datos categóricos, las categorías o columnas de a) pueden tomar varios valores, por ejemplo, Physical\_Health, puede tomar valores del 1 al 5. Mientras que las categorías de b) son binarias (y, n).

**Proporciones de acierto.**

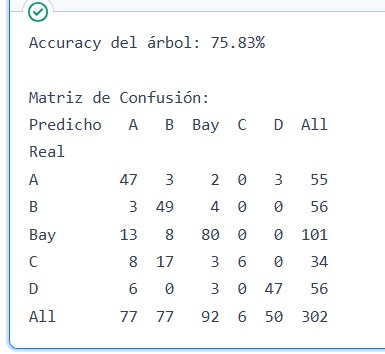
|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Reescritura** | **Cross validation** | **Leave one Out** |
| a) | **0.908** | 0.509 | 0.499 |
| b) | **0.988** | 0.960 | 0.960 |
| c) | **0.995** | 0.799 | 0.827 |

**Comparación Árbol KD.**

A continuación, tenemos el número de nodos y hojas del árbol KD. Así como su proporción de acierto en reescritura.

A screenshot of a computer code

AI-generated content may be incorrect.



|  |  |  |
| --- | --- | --- |
| **Métrica** | **Árbol KD** | **Árbol J48** |
| Proporción de acierto reescritura | 0.7583 | **0.995** |
| Nodos totales | **151** | 121 |

**Conclusiones.**

* Para los 3 árboles la proporción de acierto usando reescritura fue la mas alta, de las 3 opciones.
* Tenemos que entre mas nodos y hojas tiene el árbol resultante, tiene en general proporciones de acierto más bajas
* Entre menos valores pueden tomar las columnas categóricas, mas pequeño o simple será el árbol resultante de aquel dataset.
* En los datasets a) y c), tuvimos una caída de proporción, al cambiar de reescritura a Cross validation o Leave one Out, lo que puede indicar que en general al modelo no le va tan bien, como la reescritura nos hace creer. Así que estas dos medidas pueden ser una mejor opción a la hora de evaluar un modelo.
* Con KD, se obtuvo un árbol mas grande, que con el J48. Y a su vez obtuvo peor porcentaje de aciertos, siguiendo así el patrón de que entre mas grande y complejo el árbol peores proporciones de acierto se obtienen.