Insights

Interpretación del arbol como ejemplo de la matriz de confusión:

- TN (Verdaderos Negativos = 8442): Personas que realmente ganan ≤50K y fueron predichas correctamente como ≤50K.
- FP (Falsos Positivos = 186): Personas que realmente ganan ≤50K, pero el modelo predijo erróneamente que ganan >50K.
- FN (Falsos Negativos = 2817): Personas que realmente ganan >50K, pero el modelo predijo erróneamente que ganan ≤50K.
- TP (Verdaderos Positivos = 63): Personas que realmente ganan >50K y el modelo lo predijo correctamente.

En conclusión el modelo predijo en muchos casos que había 186 personas que ganaban más de 50k cuando realmente eran solo 63. Esto indica que el modelo predice una proporción de los positivos como positivos pero muchos negativos también lo predice como positivos y hace que no sea muy bueno para usarlo. Por otro lado los otros dos modelos también fueron bastante similares en sus resultados y con lo cual ninguno es un modelo muy preciso.

Pre procesamiento elegido

Para el pre procesamiento elegimos:

- **Input missing values:** Eliminar filas con datos faltantes, para poder ver únicamente aquellas entradas de la base de datos que estén completas
- **Select relevant features:** 78% y usando el índice de gini ya que es una medida estadística que cuantifica la desigualdad en la distribución de una variable.
- **Randomize:** Clases, lo que evita que el modelo aprenda patrones artificiales derivados del orden original de los datos y garantiza una división más justa entre entrenamiento y prueba
- **Remove sparse features:** Ceros y 4%, se eliminaron columnas que contenían muchos ceros o tenían muy poca variabilidad de menos del 4% de valores no nulos o distintos ya que este tipo de columnas pueden causar sobreajuste si el modelo trata de encontrar patrones donde no los hay.

Comparación entre modelo e interpretación de métricas

Estas son las siguientes métricas que obtuvimos:

- **Árbol:** Obtuvimos un AUC de 0,5 con lo cual el modelo no es muy bueno distinguiendo entre clases, un Accuracy de 0,74 lo que quiere decir que el modelo es medianamente acertado frente a sus predicciones, F1 del 0,65 con lo cual hay un cierto balance de clases, Precisión del 0,63 y recall del 0,74. En general es un modelo que puede llegar a ser útil si es que no hay nada mejor

- **Logistic Regresión:** Obtuvimos un AUC de 0,5 con lo cual el modelo no es muy bueno distinguiendo entre clases, un Accuracy de 0,75 lo que quiere decir que el modelo es medianamente bueno, F1 del 0,65 un valor medianamente ok, Precisión del 0,56 la cual es la más baja de los 3 modelos y por ende el modelo predijo varios positivos que en realidad eran negativos y recall del 0,75. Este modelo es el peor de los 3 por su precisión, lo que indica que predijo varios positivos (gente que gana más de 50K) cuando en realidad eran negativos(gente que ganaba menos de 50k) esto puede llevar a un error a la hora de tomar decisiones en base al resultado del censo pensando que hay menos personas que ganan menos de 50k que las que realmente hay

- **Neural Network:** Obtuvimos un AUC de 0,5 con lo cual el modelo no es muy bueno distinguiendo entre clases, un Accuracy de 0,75 lo que quiere decir que el modelo es medianamente bueno haciendo predicciones, F1 del 0,64 al igual que los otros dos Precisión del 0,64 y recall del 0,75 lo que significa que se identifican los casos positivos como positivos la mayor parte del tiempo pero dado a el valor del recall esto indica que en la mezcla de positivos hay muchos que si son positivos pero otros tantos que no, y puede ser malinterpretado por los analistas. Generalmente es muy similar al modelo 1, nada muy bueno pero tampoco muy malo.

En general todos los modelos terminaron con valores muy similares con lo cual salvo en algunas métricas particulares usar cualquiera de los 3 sería casi indiferente de cara al resultado, pero generalmente no son modelos buenos para hacer análisis dado al gran porcentaje de error que manejan.