UNIVERSIDAD
EL BOSQUE

Ingeniería de Sistemas
Computación Científica y Analítica
Taller 3

Julian Andrés León Fonseca
Julian Eduardo Santos Niño
David Mauricio Valoyes Porras

1. **Del script de forest fires:**

    a. ¿Se desea resolver el problema utilizando aprendizaje supervisado o no supervisado? ¿Es un problema de clasificación o de regresión?

RTA: El script está resolviendo el problema mediante regresión (específicamente regresión lineal), lo que a su vez implica que está utilizando aprendizaje supervisado continuo.

    b. ¿Qué interpretación le puede dar a los resultados obtenidos?

RTA: Considerando que el error absoluto del modelo es del 12%, se puede decir que el modelo en general tendría una asertividad promedio del 88%. Pero al analizar los boxplots, e identificar que hay varios valores atípicos, e incluso que algunos de estos se alejan mucho de los bigotes, eso implica que hay datos que se alejan también mucho de la media, lo cual a su vez puede conllevar a que el error al predecir dichos casos sea mayor, lo que aumentaría el valor del error promedio. Lo anterior supone que, al eliminar los valores atípicos, el error promedio podría ser menor.

2. **Del script de recursos humanos (rrhh):**

    a. ¿Cuál es la clase donde el modelo más se equivoca? ¿Por qué?

RTA: Considerando los resultados presentados en la matriz de confusión, la clase frente a la cual el módelo más se equivoca es al momento de predecir los salarios altos, en contraparte la clase para la cual el modelo es más asertivo (preciso) es en la predicción de los salarios bajos.

    b. ¿Cuál cree que es el propósito del parámetro *max_depth* usado al momento de instanciar el modelo de árbol de decisión?

RTA: El parámetro max_depth se utiliza para definir (o limitar) la profundidad de crecimiento del árbol de decisión, es decir, que tantos niveles se crean luego del nodo principal, estando los nodos hoja en el último nivel (es decir el valor que se le dé al parámetro + 1).

c. Para este caso particular, ¿por qué cree que es difícil obtener un buen clasificador?

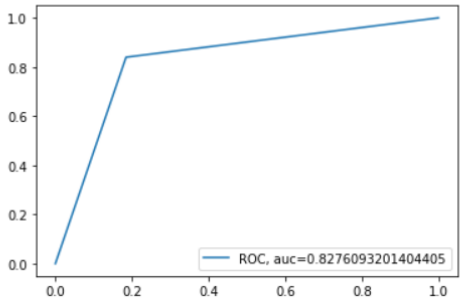
RTA: Hay mucha variabilidad en los valores de las variables, es decir que cada variable puede tomar muchos valores distintos, por lo cual un árbol de decisión se puede "quedar corto" para considerar todas las posibilidades, o por el otro lado se tendría un árbol de decisión con muchas ramificaciones (árboles no binarios), lo que llevaría a un modelo no eficiente. Incluso, el buscar que el árbol pueda manejar tal conjunto de posibilidades podría conllevar un caso de overfitting.

3. **Identificación de géneros musicales:** Tenga en cuenta que hay dos scripts: *music.ipynb* y *music-multiclass.ipynb*. En el primero se intenta crear un modelo clasificador solo para dos clases (caso binario) y en el segundo se entrena uno para todas las clases (géneros musicales) del dataset.

   a. Para el caso binario (*jazz and blues y soul and reggae*) ¿Es posible obtener mejores métricas entrenando un modelo basado en Random Forest?

RTA: En primer lugar, se desarrolló (construyó) un modelo basado en Random Forest, obteniendo los siguientes resultados:

```
=====Confusion Matrix (Random Forrest):=====

[[1060  240]
 [ 193 1012]]

=====Classification Report (Random Forrest):=====

              precision    recall  f1-score   support

           0       0.85      0.82      0.83      1300
           1       0.81      0.84      0.82      1205

    accuracy                           0.83      2505
   macro avg       0.83      0.83      0.83      2505
weighted avg       0.83      0.83      0.83      2505


=====ROC Curve:=====
```



Considerando los resultados obtenidos, y comparándolos con los resultados del modelo original (presicion, recall y F1), se puede evidenciar que no necesariamente el modelo Random Forest tiene una mejoría considerable (apenas presenta una mejoría de 0.01 - 0.03 en diferentes aspectos). Lo anterior puede deberse a que el modelo es binario, por lo cual reduce el conjunto de posibilidades con las cuales trabajar, haciendo que sea lo suficientemente eficiente.

b. Escoja otro par de géneros, entrene un conjunto de modelos y documente los resultados del mejor que se haya obtenido.

RTA: Se cambió “jazz and blues” por “classic pop and rock”, y se entrenaron modelos de: regresión logística, un clasificador MLP, un modelo neuronal, un modelo SVM y un modelo de Random Forest, donde el modelo con los mejores resultados fué Random Forest, siendo estos resultados los siguientes:

```
[40]  =====Confusion Matrix (Random Forrest):=====

[[7161    8]
 [1150   55]]

=====Classification Report (Random Forrest):=====

              precision    recall  f1-score   support

           0       0.86      1.00      0.93      7169
           1       0.87      0.05      0.09      1205

    accuracy                           0.86      8374
   macro avg       0.87      0.52      0.51      8374
weighted avg       0.86      0.86      0.80      8374
```

c. Para el caso multi-clase, ¿cuál es la clase para la que el modelo más se equivoca? ¿Por qué?

RTA: En el caso multi clase el modelo se equivoca principalmente para la clase “hip hop”, e inferencialmente la razón de la equivocación radica en que este género, en cuanto a sus atributos, y especialmente a los valores de estos, tiene una concordancia alta con los demás géneros, lo cual implica que el modelo se “confunde”.

d. Para el caso multi-clase, el modelo basado en red neuronal parece estar mayoritariamente sesgado hacia un género particular ¿Cuál género cree que es?

RTA: El modelo parece estar principalmente sesgado hacia el “classic por and rock”, puesto que es la clase para la cual menos se equivoca. Lo anterior puede implicar que el conjunto de datos relaciona principalmente registros de ese género, y que por ende el modelo está especializado en la predicción de esa clase (overfitting).

4. **Segmentación de cajas de compensación familiar (subsidio):**

a. ¿Qué cajas de compensación parecen ser mayoritariamente diferentes a las demás?

RTA: Las cajas de compensación que parecen ser mayoritariamente diferentes a las demás, son aquellas que pertenecen al cluster número 3 (nombre del cluster en esta ejecución), es decir las cajas de compensación: CAFAM y Familiar del Valle del Cauca. Las anteriores conclusiones se obtienen luego de realizadas las siguientes consultas:

```
[14] subsidio_df['cluster'].value_counts()

1    24
0    14
2     3
3     2
Name: cluster, dtype: int64
```

```
subsidio_df.loc[subsidio_df['cluster'] == 3]
```

	Código	CCF	Empresas Afiliadas	Total Afiliados Cajas Compensacion Familiar	Trabajadores Afiliados Dependientes	Afiliados Facultativos Independientes	Afiliados Pensionados	Afiliados Fidelidad	No Afiliados con Derecho a Subsidio	Personas Cargo	Total Población Cubierta	cluster
14	21	Caja de Compensacion Familiar CAFAM	45216	837046	662904	151251	8958	6788	7145	808245	1645291	3
36	57	Caja de Compensacion Familiar del Valle del Ca...	46221	603885	590416	7130	6339	0	0	723626	1327511	3

b. ¿A partir de qué características utilizadas para el entrenamiento del modelo se podría explicar la razón por la que las cajas anteriores fueron agrupadas en clusters tan pequeños?

RTA: Principalmente el hecho de que sus atributos son atípicos en comparación con los atributos de los otros cluster (es decir que sus valores se salen de los rangos normales). Lo anterior implica que al no poderse agrupar en uno de los cluster grandes existentes, requieren de un nuevo cluster para su asociación.

c. ¿Se pueden obtener resultados más homogéneos utilizando cantidades diferentes de clusters para el entrenamiento? Entienda homogeneidad como clusters con cantidades similares de instancias de datos.

RTA: Sí, mientras más clusters haya, es posible que las cajas de compensación se puedan repartir de forma más homogénea (donde cada cluster agrupa a menos cajas de compensación). El problema radica en que cada cluster tendría una cantidad cada vez menor de cajas de compensación, por lo cual a medida que se aumenta la

homogeneidad, igualmente se tiende a que la cantidad de clusters sea mayor. Lo último implica que para tener una homogeneidad total básicamente se tendría la misma cantidad de clusters que de cajas de compensación.