Unidad 4: Problema de clasificación biológico usando el algoritmo de Naive Bayes

Data set:

Se conoce el genotipo de 697 plantas y su momento de floración en días. El fichero floweringTime.zip contiene:

- genotype.csv: Es el genotipo para cada planta (697 x 149). Hay tres posibles estados como máximo: 0, 1 y 2 que corresponden a Homozigoto dominante, Heterozigoto y Homozigoto recesivo, respectivamente.
- flowering_time.csv: Es el tiempo transcurrido hasta la floración en días para cada planta (697 x 1).

Objetivo:

Se quiere **predecir** el **tipo floración** (rápida o lenta) en función del **genotipo** de la planta.

Procedimiento:

Para resolver este problema se realizará un informe dinámico donde se aplicaran los mismos pasos que se hizo en el problema de la Unidad 2.

En este informe tiene que aparecer al principio un índice y una sección que incluya la tabla de fortalezas y debilidades del algoritmo Naive Bayes y una explicación del algoritmo. A continuación, ya se plantea la resolución del problema con el algoritmo de Naive Bayes aplicando los 5 pasos que plantea siempre el libro, como se puede ver en el ejemplo de spam data.

Puntos importantes:

- Floración *rápida* es cuando el número de días trascurrido hasta la floración es menor o igual a 40 días, en caso contrario es floración *lenta.* Se codifica la floración rápida como 0 y la floración lenta como 1.
- El dataset se dividirá en 2/3 training y 1/3 test. Para utilizar la misma serie de registros de training y de test usar como semilla inicial el valor de set.seed(12345).
- Para evaluar el rendimiento del modelo se usará la función confusionMatrix() del package caret que da más información. La categoría positiva es floración lenta.
- Se debe probar el modelo de Naive Bayes con laplace=0 y laplace=1.
- (OPCIONAL) Para aquellos de vosotros con más tiempo disponible:
    - Crear un nuevo apartado titulado “Curvas ROC” donde se obtenga las curvas ROC para el modelo de Naive Bayes con laplace=0 y laplace= 1 y el área bajo la curva. Recordar que el package ROCR sirve para hacer las curvas ROC (ver Unidad 3). Importante: Usar el argumento type =”raw” de la función predict() para obtener las probabilidades.

Es fundamental verificar que el informe es "dinámico", es decir, que se adapte a unos nuevos datos. Por ejemplo, si cuando se describe el fichero original de datos se escribe: "Nuestro fichero tiene 300 registros y 30 variables" pero después cambiamos el fichero por otro de 302 registros y 28 variables el informe debería aparecer como: "Nuestro fichero tiene 302 registros y 28 variables" automáticamente. Por tanto, el valor 300 y 30 debe ser el resultado del número de filas y de columnas del fichero, respectivamente. Este principio se debe tener en cuenta en la redacción del informe para poder hacer el informe lo más general/dinámico posible.

Para tener constancia de vuestro trabajo, cada estudiante debe empaquetar el fichero "Unidad4.Rmd" y los dos ficheros de salida "Unidad4.html" y "Unidad4.pdf" en el fichero "Unidad4.zip". Este fichero se debe de añadir a la actividad "Entrega actividad no evaluable Unidad 4" que encontraras en la sección de Contenidos.