Actividad: Resolver un problema de clasificación

**Objetivos de la actividad**

Con esta actividad vas a conseguir resolver un problema de clasificación aplicando los modelos basados en árboles de decisión y mediante técnicas de ensamble. Además, conseguirás evaluar la importancia de las variables para distintos modelos y realizar comparaciones entre distintas técnicas. Asimismo, aumentarás tu práctica sobre las librerías *statsmodels* y *scikit-learn* de Python al crear un conjunto de datos mediante simulación.

**Descripción de la actividad y pautas de elaboración**

El primer paso consiste en crear un conjunto de datos ficticio. Generaremos nuestro conjunto a través de un número semilla compuesto por ocho cifras a través de nuestro número de identificación. Si el número de identidad tiene más de esas cifras, nos quedaremos con las primeras ocho empezando desde la derecha. Si tiene menos de ocho cifras replicaremos las primeras hasta obtener exactamente ocho. Además, para evitar los dígitos cero y uno, si alguna de las cifras es menor que 2 la sustituiremos por ese número. Aplicando estos cambios tendremos el número del documento de identidad preparado para la resolución de la actividad. Veamos algunos ejemplos:

* Ejemplo 1:

1234567890 34567890 34567892

* Ejemplo 2:

12345678 22345678

* Ejemplo 3:

304156 30415630 32425632

Para lo anterior se puede crear una función que lo realice automáticamente o hacerlo manualmente.

Una vez que tengamos el número de la semilla vamos a crear el conjunto de datos para el problema de clasificación con la función *sklearn.datasets.make\_classification* de la librería *scikit-learn*.

Accede a la librería de *scikit-learn* en el siguiente enlace: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html>

* Crea el conjunto de datos empleando los siguientes argumentos:
  + n\_samples = 200 + 10 · primer dígito semilla
  + n\_features = 10 + segundo dígito semilla + tercer dígito semilla
  + n\_informative = 10 + segundo dígito semilla
  + n\_redundant = 2
  + shift = 10 · cuarto dígito semilla
  + random\_state = semilla
  + shuffle = False

Entonces, se pide, de forma justificada:

* ¿Qué significa cada uno de los parámetros de entrada en la función *make\_classification* usados para crear nuestro *dataset*?
* Describe tu conjunto de datos, aplica los métodos *info, describe, head, tail* y los que creas adecuados y obtén un análisis gráfico de todas las variables (predictoras y respuesta).
* Divide el conjunto de datos en 200 observaciones para el entrenamiento y el resto para realizar la validación de los distintos métodos de clasificación aplicados.
* Obtén un modelo de clasificación mediante un árbol de decisión utilizando los argumentos por defecto.
* Ahora indica que el nivel de profundidad máximo permitido es de 3. ¿Qué diferencias observas con el árbol anterior?
* Ahora indica que el mínimo número de observaciones que debe contener cada nodo hoja es de 5. ¿Qué diferencias observas con respecto a los árboles anteriores?
* Realiza los siguientes modelos de ensamble:
  + Con reemplazamiento (*bagging*).
  + Sin reemplazamiento (*pasting*).
  + Realizando *Random Forest* (fijando el número de nodos hoja máximo a 4).
  + *Gradient Boosting*.
* Analiza la importancia de las variables de cada uno de estos métodos.
* No olvides añadir las referencias utilizadas para la elaboración del trabajo al final de la memoria.

**Extensión**: se deben entregar dos archivos:

* Un archivo ipynb con las operaciones realizadas. Formato en el aula.

apellido1\_apellido2\_actividad.ipynb

Ejemplo: gonzalez\_sanchez\_actividad.ipynb

* Un archivo pdf con el análisis de los resultados obtenidos. Máximo 10 páginas.

apellido1\_apellido2\_actividad.pdf

Ejemplo: gonzalez\_sanchez\_actividad.pdf

**Rúbrica:**

|  |  |  |  |
| --- | --- | --- | --- |
| Resolver un problema de clasificación | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | El conjunto de datos y su descriptivo son correctos. | 2 | 20% |
| Criterio 2 | Los modelos de clasificación mediante árboles de decisión son correctos. | 2 | 20% |
| Criterio 3 | Se aplican correctamente las técnicas de ensamble. | 3 | 30% |
| Criterio 4 | La importancia de las variables se muestra correctamente. | 2 | 20% |
| Criterio 5 | La entrega, la presentación, el formato y bibliografía son correctos. | 1 | 10% |
|  |  | **10** | **100 %** |