Curso de posgrado: “Aprendizaje automático: Fundamentos, Herramientas y Aplicaciones”
Diplomado de Telecomunicaciones y Sensado Remoto por Ondas de Radio
Trabajo práctico Nº 1
TEMA: Árboles de decisión

[PROBLEMA 1: prediciendo si una persona tiene o no tiene diabetes]

El archivo “diabetes.csv” posee datos obtenidos del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales (https://www.niddk.nih.gov/). El objetivo es predecir si un paciente tiene diabetes o no, basándose en determinadas medidas de diagnóstico incluidas en el conjunto de datos. Se impusieron varias restricciones a la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años.

Para el desarrollo del ejercicio, resliazar los siguientes apartados:

- ¿De qué tipo de dato y rango son las columnas del conjunto de datos?
- Mencionar cuáles son las columnas características y cuál es la clase que se predice.
- Categorizar las columnas que considere necesario, por ejemplo: si se posee un campo “edad” con valores entre [0-70], es posible agrupar (binning) el campo de la siguiente forma:
  - rango [0-11] → niño ; rango [12-18] → adolescente ; rango [19-25] → joven ; rango [25- 59]→adulto ; rango [60-70] → mayor
- Utilizar un porcentaje de datos adecuado para el entrenamiento y para el testeo.
- Medir la performance del modelo obtenido sobre el conjunto de datos de testeo mediante las métricas que arroja la matriz de confusión.
- Analizar diferentes modelos mediante el uso de distintos criterios de selección de nodos.

[PROBLEMA 2: Ta-Te-Ti]

El siguiente problema se trata de una clasificación binaria. Se realizó una clasificación binaria al combinar el conjunto de datos que contiene los movimientos específicos del juego Ta-Te-Ti y su resultado. Se clasifica el resultado del juego como positivo si “X” gana de acuerdo con los movimientos especificados o ingresados, de lo contrario negativo. De esta manera, se aprendió cómo debían ser los movimientos en este juego o cómo sería el resultado del juego como resultado de estos movimientos.
El archivo “tic-tac-toe.csv” contiene los datos del análisis mencionado anteriormente.

- ¿De qué tipo de dato y rango son las columnas del conjunto de datos?
- Mencionar cuáles son las columnas características y cuál es la clase que se predice.
- Utilizar un porcentaje de datos adecuado para el entrenamiento y para el testeo.
- Medir la performance del modelo obtenido sobre el conjunto de datos de testeo mediante las métricas que arroja la matriz de confusión.
- Analizar diferentes modelos mediante el uso de distintos criterios de selección de nodos.

Ayuda:
Para la transformación de los carácteres ‘x’, ‘o’ y ‘b’ de un archivo leído utilizando la librería “Pandas”, es posible la conversión de estos a números usando el método “factorize”. Este método ayuda a obtener la representación numérica de una matriz mediante la identificación de valores distintos.
Ejemplos: https://www.geeksforgeeks.org/python-pandas-factorize/

Curso de posgrado: “Aprendizaje automático: Fundamentos, Herramientas y Aplicaciones”
Diplomado de Telecomunicaciones y Sensado Remoto por Ondas de Radio
Trabajo práctico Nº 1
TEMA: Árboles de decisión

+Info (algunas referencias al tema):

Matheus, C.J., & Rendell, L.A. (1989). Constructive induction on decision trees. In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence. (pp. 645--650). Detroit, MI: Morgan Kaufmann.

Matheus, C.J. (1990). Adding domain knowledge to SBL through feature construction. In Proceedings of the Eighth National Conference on Artificial Intelligence (pp. 803--808). Boston, MA: AAAI Press.

Aha, D. W. (1991). Incremental constructive induction: An instance-based approach. In Proceedings of the Eighth International Workshop on Machine Learning (pp. 117--121). Evanston, ILL: Morgan Kaufmann.