	Arbres de decisió	UAB

Observacions:

En aquesta pràctica es presenta el problema dels arbres de decisió. El problema s'organitza en tres nivells de dificultat: A (sobre 10) i és la dificultat màxima, B (sobre 8) que és la dificultat mitja i C (sobre 6) que és la dificultat més baixa. En aquest projecte hi ha un apartat més (A+), de manera que aquesta pràctica està valorada sobre 12.

Per aprovar la pràctica és requisit necessari completar satisfactòriament la part C del problema, demostrant així una comprensió fonamental de la matèria. La superació de la mateixa estarà condicionada a la presentació d'una documentació i una defensa adequades. La màxima puntuació s'aconsegueix resolent els problemes de tots els nivells (A, B, C). S'ha de tenir en compte que no es pot fer la part A sense haver fet abans la part B.

Objectius de la pràctica:

Els objectius d'aquesta pràctica són:

- Assimilar conceptes de guany d'informació, entropia i discriminació de grans conjunts de dades
- Millorar la comprensió dels algoritmes d'inferència d'arbres ID3, C4.5 i la seva posterior avaluació
- Utilització de bases de dades reals
- Utilització de les tècniques de validació de resultats en dades reals.

Materials per a la sessió:

1. Cada grup de pràctiques ha d'utilitzar la base de dades que figura en la següent taula:

Grups 1	Grups 2	Grups 3	Grups 4
Link	Link	Link	Link

Avaluació sessió de seguiment

En la sessió de seguiment caldrà tenir:

- Analitzat i codificat l'estructura de dades i la visualització d'aquestes.
- El criteri de divisió pot ser qualsevol, p.ex: "error de classificació".
- Dissenyat el conjunt de proves a realitzar i les mesures que s'usaran per avaluar els algorismes
- Com a mínim ser capaços de llegir la base de dades

Exercici 1: **(C)**

En aquest exercici es demana implementar l'algorisme ID3 i el criteri de separació de C4.5 (Guany d'informació i Gini) per a atributs categòrics que ens permeti determinar el tipus d'entitat d'una nova mostra. També es demana la visualització de l'arbre, és a dir, la representació de l'arbre que es crea per poder determinar la classificació. Aquesta visualització no cal que sigui molt complexa.

Si la base de dades té 'missing values', és a dir, hi ha valors buits, cal que es faci un tractament simple d'aquests, com podria ser l'eliminació de les files que els continguin.
Si la base de dades té variables contínues, per aquest apartat, es podrà discretitzar els seus valors utilitzant la funció qcut(...) de la llibreria pandas de python. Aquesta funció assigna a la instància un valor en funció del quartil al qual pertany el valor.

Finalment cal validar la base de dades amb els mètodes apresos a classe:

- Cross-validation
- Leave one out
- ...

La nota tindrà en compte que s'hagi fet una avaluació acurada, així com la justificació de la mètrica utilitzada.

Exercici 2: **(B)**

En aquest exercici es demana la implementació de 'missing values' avançat. Tant en la fase d'aprenentatge com de predicció.

Exercici 3: **(A)**

En aquest cas, cal aplicar el tractament d'arguments continus explicat a la classe de teoria.

Recordeu que també cal validar aquesta base de dades.

Exercici 4: **(A+)**

En aquest exercici es demana la implementació de tècniques de reducció de l'overfitting aplicant la tècnica del pruning o la implementació d'un Random Forest.

Caldrà fer una comparativa entre el sistema original i aplicant la tècnica triada per a evitar l'overfitting.