Test 1 seminar ML – 1 noiembrie 2021, ora 16:00 (100 minute)

1. (1.8pt) Fie variabile aleatoare discrete *X, Y*, având distribuția comună de probabilitate conform cu tabelul de mai jos:

	X=0	*X=1*
Y=0	*3/16*	*5/16*
Y=1	***p***	*5/32*
Y=2	*3/32*	***q***

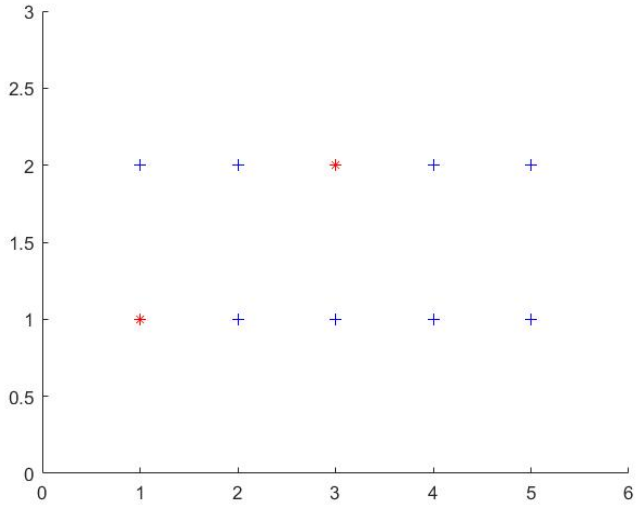
   a) Calculați *p* și *q* astfel încât cele două variabile să fie independente. Folosind valorile găsite pentru *p* și *q*, verificați condiția de independență pentru toate valorile variabilelor *X* și *Y*.
   b) Calculați *Var[X]* și $E[Y^2]$.

2. (2.2pt) Tabelul de mai jos descrie șapte înregistrări pozitive și negative pentru persoane cărora banca le-a acordat (sau nu le-a acordat) un card de credit în funcție de Gen, Venit și Vârstă. Atributul Vârstă este atribut continuu.

Gen	Venit	Vârstă	Aprobat
M	Mic	22	Da
M	Mare	32	Da
M	Mare	38	Da
M	Mare	32	Nu
M	Mare	30	Nu
F	Mic	24	Nu
F	Mare	32	Da

   a) Elaborați și desenați arborele de decizie învățat de algoritmul ID3 pentru datele din acest tabel. Pentru fiecare nod al arborelui specificați câștigul de informație sau entropia condițională medie.
   b) Schimbați valoarea unuia din atributele de intrarea de mai sus, pentru una din instanțe, astfel încât arborele învățat are cel puțin un nod suplimentar.

3. (1.2pt) Figura de mai jos prezintă un set de date cu două intrări X_1 şi X_2, variabile cu valori reale şi o variabilă de ieșire Y care poate lua două valori (marcate + şi * pe grafic).



Care va fi eroarea de clasificare la cross-validare cu metoda "Leave-One-Out" pe acest set de date atunci când folosim algoritmul ID3? (Indicaţi datele incorect clasificate.) În fiecare situație, doar desenați granițele de decizie.

4. (0.8pt) Marcați cu *adevărat* sau *fals* fiecare din afirmațiile de mai jos:
   a) Pentru orice $A, B \subseteq \Omega$ independente, ***P(AUB)=P(A)+P(B)-P(A)P(B)***
   b) Fie X o variabilă aleatoare pentru care media este $E[X]=m$ şi varianţa este $Var[X]=s$. Fie $c \in \mathbb{R}$ atunci $E[(X-c)^2] = (m-c)^2 - s$.

   Pentru fiecare afirmație adevărată, faceţi demonstraţia proprietăţii respective. Pentru fiecare afirmaţie falsă, daţi fie un contraexemplu, fie o justificare riguroasă.
   c) Învățăm un arbore de decizie folosind algoritmul ID3 standard, fără pruning. Atributele de intrare (X_1,X_2, ...,X_m) sunt categoriale, iar atributul de ieșire (Y) este de asemenea categorial. Marcați cu A (adevărat) sau F (fals) următoarele afirmații și dați în fiecare caz o explicație succintă:
      - ✓ Adâncimea maximă a arborelui de decizie este de cel mult m.
        Notă: Dacă arborele este format doar din nodul rădăcină, ceea ce corespunde cazului în care toate exemplele de antrenament sunt identic clasificate, atunci se consideră că adâncimea arborelui este 0.
      - ✓ Dacă sunt R exemple de antrenament, atunci adâncimea maximă a arborelui de decizie este de cel mult $1 + log_2 R$.