Bases de Données II, Charleroi, 11 janvier 2024

NOM + PRÉNOM :

Orientation + Année :

Cet examen contient 7 questions. Durée : exactement 2 heures et 50 minutes. Les questions sont censées être claires. Aucune clarification supplémentaire ne sera fournie pendant l'examen. Si une question vous semble ambiguë ou incomplète, veuillez formuler vos hypothèses et répondez en fonction de celles-ci. **Il est permis de détacher la dernière page.**

Un fleuriste livre des bouquets à la maison. Chaque espèce de fleur (tulipe, rose...) a un prix exprimé en centimes d'euro. Par exemple, le prix d'une rose est de 150 centimes d'euro, indépendamment de sa couleur ; voir la ligne

```
<fleur fnom="rose" prix="150"/>
```

Un bouquet rassemble des fleurs de différentes espèces et couleurs. Par exemple, le bouquet *Exotique* est composé d'un seul tournesol et trois iris bleus. Voir les lignes :

```
<bouquet bnom="Exotique">
   <fleur fnom="tournesol" couleur="jaune" nombre="1"/>
   <fleur fnom="iris" couleur="bleu" nombre="3"/>
</bouquet>
```

La figure 1 montre le document XML et la figure 2 montre le DTD.

La balise `ventes-par-jour` est utilisée pour encoder les ventes journalières. Par exemple, à la date du 18 septembre 2022, le fleuriste a vendu trois bouquets *Valentin* et six bouquets *Printemps*. Voir les lignes :

```
<date mois="sep 2022" jour="18">
  <vente bnom="Valentin">3</vente>
  <vente bnom="Printemps">6</vente>
</date>
```

Question 1 Écrivez une requête en **XPath**, aussi simple que possible, qui renvoie le nom de chaque bouquet contenant au moins deux espèces de fleur distinctes. **Pour cette question, il n'est pas permis d'utiliser des fonctions d'agrégation telles que count, min et max.** Évitez autant que possible d'utiliser les axes *parent* et *ancestor*.

Pour le document XML de la figure 1, la réponse est comme suit :

```
bnom="Exotique"
bnom="Printemps"
```

Noter que le bouquet *Belge* ne contient que des tulipes et n'est donc pas dans la réponse.

.../5

```
//bouquets/bouquet[fleur/@fnom!=fleur/@fnom]/@bnom
```

ou

```
//bouquets/bouquet[fleur[@fnom!=following-sibling::fleur/@fnom]]/@bnom
```

Notez la double imbrication : l'évaluation de `following-sibling` doit se faire par rapport à une `fleur`, et non par rapport à un `bouquet`. En français, cette requête renvoie les bouquets contenant une fleur dont l'espèce diffère de celle d'une autre fleur située ultérieurement **dans le même bouquet**.

Question 2 Écrivez une requête en **XPath** qui renvoie, de préférence sans doublons, chaque espèce de fleur présente dans au moins deux bouquets distincts. **Il n'est pas permis d'utiliser des fonctions d'agrégation telles que count, min et max. Il n'est pas non plus permis d'utiliser la fonction `distinct-values`.** Évitez autant que possible d'utiliser les axes *parent* et *ancestor*.

Pour le document XML de la figure 1, la réponse est comme suit :

```
fnom="tulipe"
fnom="rose"
```

.../5

```
//fleurs/fleur/@fnom[.=//bouquet/fleur/@fnom[.=following::bouquet/fleur/@fnom]]
```

La sous-requête `//bouquet/fleur/@fnom[.=following::bouquet/fleur/@fnom]` (que nous noterons S ci-après) récupère chaque espèce de fleur présente dans au moins deux bouquets distincts, mais elle peut générer des doublons. La requête `//fleurs/fleur/@fnom[.=`S`]` renvoie, sans doublons, chaque espèce de fleur dans S.

Une autre solution est :

```
//bouquet/fleur/@fnom[not(.=following::fleur/@fnom)][.=preceding::bouquet/fleur/@fnom]
```

En français, cette demande vérifie si, pour chaque dernière occurrence d'une espèce de fleur dans le document, cette espèce figure également dans un autre bouquet situé antérieurement dans le document.

Notez que la formulation symétrique suivante **n'est pas correcte** :

```
//bouquet/fleur/@fnom[not(.=preceding::fleur/@fnom)][.=following::bouquet/fleur/@fnom]
```

En effet, l'expression `[not(.=preceding::fleur/@fnom)]` sera constamment fausse, car toutes les espèces de fleur sont répertoriées au début du document entre `fleurs` et `/fleurs`.

Question 3 Écrivez une requête en **XQuery** qui renvoie le bouquet le plus vendu.

Pour le document XML de la figure 1, la réponse est :

```
bnom="Valentin"
```

Effectivement, le bouquet nommé *Valentin* a été vendu au total 9 fois ; aucun autre bouquet n'a atteint un nombre de ventes plus élevé.

.../10

```
let $help := for $b in //bouquet
             return <b
                     bnom='{$b/@bnom}'
                     ventes='{sum(//vente[@bnom=$b/@bnom])}'/>
for $b in $help[@ventes=max($help/@ventes)]
return $b/@bnom
```

Question 4 Écrivez un programme en **XSLT** qui affiche, pour chaque bouquet, le nombre total d'unités vendues. **Il n'est pas permis d'utiliser `xsl:for-each` et `xsl:if`. Il n'est pas non plus permis d'utiliser la fonction `distinct-values`.** L'usage de la fonction `sum` est permis. Les résultats doivent être représentés comme suit :

```
<ventes-par-bouquet>
  <ventes bnom="Valentin">9</ventes>
  <ventes bnom="Belge">8</ventes>
  <ventes bnom="Exotique">4</ventes>
  <ventes bnom="Printemps">7</ventes>
</ventes-par-bouquet>
```

Noter que 9 est obtenu comme $4 + 3 + 2$. L'ordre des bouquets n'a pas d'importance.

. . . /10

```
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<!--Pour chaque bouquet, liste le nombre d'unités vendues.-->

<xsl:template match="/">
   <ventes-par-bouquet>
      <xsl:apply-templates select="//bouquet"/>
   </ventes-par-bouquet>
</xsl:template>

<xsl:template match="bouquet">
   <ventes bnom="{@bnom}">
      <xsl:value-of select="sum(//vente[@bnom=current()/@bnom])"/>
   </ventes>
</xsl:template>

</xsl:stylesheet>
```

Question 5 Pour un problème de classification à deux classes (disons "yes" et "no"), l'exactitude (*accuracy*) d'un classificateur est défini comme $\mathsf{ACCURACY} = \frac{\mathsf{TP}+\mathsf{TN}}{\mathsf{P}+\mathsf{N}}$. Deux autres métriques abordées en cours sont $\mathsf{TPR} = \frac{\mathsf{TP}}{\mathsf{P}}$ et $\mathsf{FPR} = \frac{\mathsf{FP}}{\mathsf{N}}$. Dans l'espace ci-dessous, complétez les cases vides de manière à obtenir une matrice de confusion où la valeur d'ACCURACY dépasse 0.9, mais où le classificateur sous-jacent ne se distingue pas d'une classification aléatoire (connue en anglais sous le terme de *random guess*) en raison des valeurs pour TPR et/ou FPR.

.../4

		Classe prédite	
		yes	no
Classe observée	yes	0	5
	no	0	95

Détaillez ci-dessous les raisons pour lesquelles votre matrice proposée satisfait bien aux exigences demandées. En particulier, clarifiez quel scénario de *random guess* est représenté par votre matrice de confusion.

.../6

On obtient :

$$\mathsf{ACCURACY} = 0.95$$
$$\mathsf{TPR} = 0$$
$$\mathsf{FPR} = 0$$

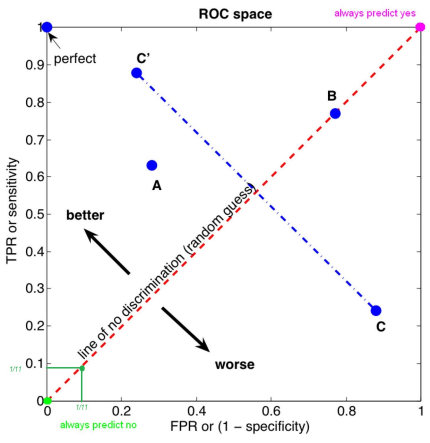
La prédiction du *random guess* est constamment "no".

Notez que ceci correspond à un scénario discuté en cours : une telle matrice de confusion est obtenue si ce *random guess* est utilisé pour prédire la fertilité d'une vache pendant 100 jours (en considérant qu'une vache est fertile un jour toutes les trois semaines).

Notez que d'autres matrices peuvent convenir, par exemple,

		Classe prédite	
		yes	no
Classe observée	yes	10	10^2
	no	10^3	10^4

Puisque $\mathsf{TPR} = \mathsf{FPR} = \frac{1}{11}$, cette matrice se trouve "sur la diagonale", et $\mathsf{ACCURACY} = \frac{10.010}{11.110} > 0.9$. Le *random guess* prédit "yes" avec une probabilité de $\frac{1}{11}$ (et par conséquent prédit "no" avec une probabilité de $\frac{10}{11}$). Si $\mathsf{P} = 110$ et $\mathsf{N} = 11.000$ (comme dans la matrice), les valeurs attendues par ce *random guess* sont bien $\mathsf{TP} = \frac{110}{11} = 10$ et $\mathsf{FP} = \frac{11.000}{11} = 1.000$.



Question 6 Voici un slide du cours à titre de rappel.

Évaluation des modèles de classification

- We observe that X instances out of n are correctly classified. (Note: there can be two or more classes.)
- $X \sim N(np, np(1-p))$, where p is the real accuracy.
- Thus, $\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$.
- Fix α.
- $Pr(-Z_{\frac{\alpha}{2}} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq +Z_{\frac{\alpha}{2}}) = 1-\alpha$
- We obtain:

$$p_{1,2} = \frac{2X + (Z_{\frac{\alpha}{2}})^2 \pm Z_{\frac{\alpha}{2}}\sqrt{(Z_{\frac{\alpha}{2}})^2 + 4X - \frac{4X^2}{n}}}{2(n + (Z_{\frac{\alpha}{2}})^2)}$$

- For example, $n = 100$, $X = 80$ (accuracy $= 0.80$)
  $\rightsquigarrow$ if $1-\alpha = 0.95$ then $p_1 = 0.711$ and $p_2 = 0.867$.

J. Wijsen | Classification | 55 / 111

Dans l'espace vide ci-dessous, expliquer la signification des symboles α et $Z_{\frac{\alpha}{2}}$.

.../2

Soit

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

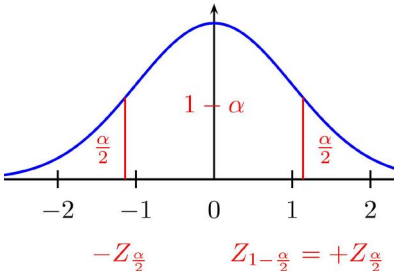
la courbe de Gauss avec $\mu = 1$ et $\sigma = 1$. [Il n'était pas nécessaire de mémoriser cette formule.] Pour un nombre réel α compris entre 0 et 1, $+Z_{\frac{\alpha}{2}}$ représente le nombre réel positif tel que

$$\int_{-Z_{\frac{\alpha}{2}}}^{+Z_{\frac{\alpha}{2}}} g(x)\, dx = 1-\alpha.$$

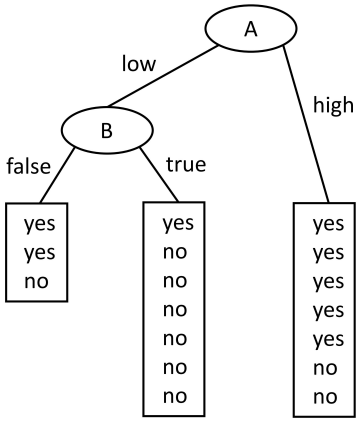
Donc,

$$\int_{-\infty}^{-Z_{\frac{\alpha}{2}}} g(x)\, dx = \frac{\alpha}{2};$$
$$\int_{-\infty}^{+Z_{\frac{\alpha}{2}}} g(x)\, dx = 1 - \frac{\alpha}{2}.$$

Graphiquement,



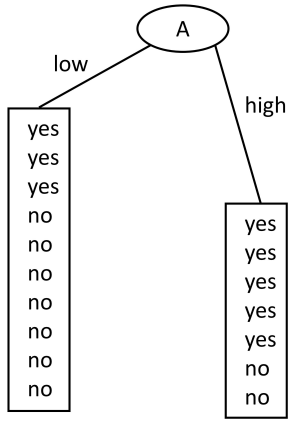
Dans l'espace vide ci-dessous, expliquez en détail comment l'algorithme C4.5 prend la décision de savoir si l'attribut B doit être élagué dans l'arbre suivant :



Inutile de mentionner que C4.5 utilise la valeur $Z_{\frac{\alpha}{2}} = 0.68$. Cependant, veuillez préciser dans votre réponse les valeurs de X et n intervenant dans les calculs de C4.5, ainsi que le(s) critère(s) utilisé(s) par l'algorithme.

/8

L'arbre sans B :



Soit

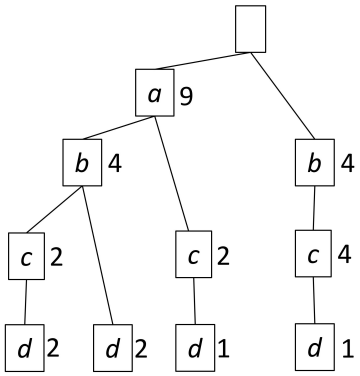
$$f(n, X) \coloneqq \frac{2X + 0.68^2 - 0.68\sqrt{0.68^2 + 4X - \frac{4X^2}{n}}}{2(n + 0.68^2)}.$$

Soient

$$\begin{aligned} \mathsf{avecB} &\coloneqq \frac{3}{10} f(3,2) + \frac{7}{10} f(7,6); \\ \mathsf{sansB} &\coloneqq f(10,7). \end{aligned}$$

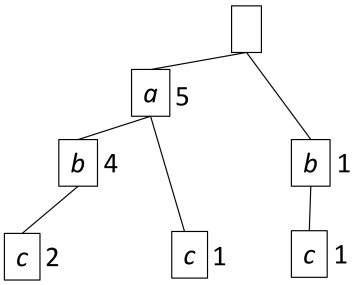
Si avecB est supérieur à sansB, l'arbre original est conservé ; sinon, l'arbre sans B est renvoyé.

Question 7 Voici l'arbre FP (connu en anglais sous le nom de *FP-tree*) associé à une base de données comprenant des transactions d'articles a, b, c, d.



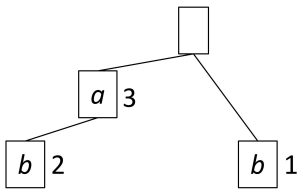
Dans l'espace vide ci-dessous, dessinez **l'arbre FP conditionnel pour** d (connu en anglais sous le nom de *conditional FP-tree for d*).

.../3



Dans l'espace vide ci-dessous, dessinez **l'arbre FP conditionnel pour** cd (connu en anglais sous le nom de *conditional FP-tree for cd).*

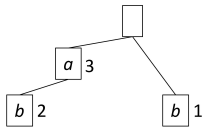
.../3



Dans l'espace vide ci-dessous, donnez le *support count* de chaque ensemble (connu en anglais sous le nom d'*itemset)* qui inclut $\{c, d\}$ en tant que sous-ensemble. Expliquez de manière concise comment ces valeurs peuvent être aisément déduites à partir des arbres précédentes.

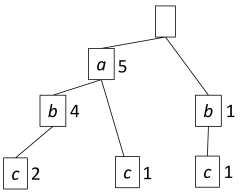
.../2

Il existe 4 sous-ensembles de $\{a, b, c, d\}$ qui incluent $\{c, d\}$, à savoir $\{c, d\}$, $\{b, c, d\}$, $\{a, b, c, d\}$, et $\{a, c, d\}$.

L'arbre 

nous indique que parmi les transactions qui contiennent c et d :

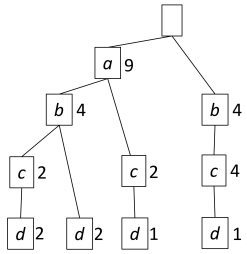
- il y en a $2 + 1 = 3$ qui contiennent aussi b, donc $\sigma(\{b, c, d\}) = 3$;
- il y en a 2 qui contiennent à la fois a et b, donc $\sigma(\{a, b, c, d\}) = 2$;
- il y en a 3 qui contiennent aussi a, donc $\sigma(\{a, c, d\}) = 3$.

L'arbre 

nous indique que parmi les transactions qui contiennent d, il y en a $2 + 1 + 1 = 4$ qui contiennent aussi c, donc $\sigma(\{c, d\}) = 4$.

Dans l'espace vide ci-dessous, donnez deux ensembles qui ne sont pas fermés (connu en anglais sous le nom de *non-closed itemsets*), ainsi que le *support count* de chacun de ces deux ensembles. Expliquez de manière concise comment ces ensembles peuvent être facilement identifiés à partir de l'un des arbres précédents.

.../2

L'arbre 

nous indique :

- parmi les 2 transactions qui incluent $\{a, b, c\}$, il y en a 2 qui contiennent aussi d. Donc, $\sigma(\{a, b, c\}) = \sigma(\{a, b, c, d\}) = 2$;
- parmi les 4 transactions qui contiennent a et b, il y en a $2 + 2 = 4$ qui contiennent aussi d. Donc, $\sigma(\{a, b\}) = \sigma(\{a, b, d\}) = 4$.

Il est correct de conclure que $\{a, b, c\}$ et $\{a, b\}$ ne sont pas fermés.

Notez que la présence de l'élément b dans une transaction ne garantit pas systématiquement la présence de c. Effectivement, parmi les 8 transactions comportant l'élément b, seules 6 incluent également c.

```xml
<?xml version="1.0" encoding="utf-8"?>
<fleuriste>
  <fleurs>
    <!-- Les prix sont en centimes d'euro -->
    <fleur fnom="tulipe" prix="100"/>
    <fleur fnom="rose" prix="150"/>
    <fleur fnom="iris" prix="250"/>
    <fleur fnom="tournesol" prix="300"/>
    <fleur fnom="dahlia" prix="120"/>
  </fleurs>
  <bouquets>
    <bouquet bnom="Valentin">
      <fleur fnom="rose" couleur="rouge" nombre="10"/>
    </bouquet>
    <bouquet bnom="Belge">
      <fleur fnom="tulipe" couleur="noir" nombre="3"/>
      <fleur fnom="tulipe" couleur="jaune" nombre="4"/>
      <fleur fnom="tulipe" couleur="rouge" nombre="6"/>
    </bouquet>
    <bouquet bnom="Exotique">
      <fleur fnom="tournesol" couleur="jaune" nombre="1"/>
      <fleur fnom="iris" couleur="bleu" nombre="3"/>
    </bouquet>
    <bouquet bnom="Printemps">
      <fleur fnom="rose" couleur="jaune" nombre="10"/>
      <fleur fnom="tulipe" couleur="jaune" nombre="4"/>
      <fleur fnom="dahlia" couleur="rose" nombre="3"/>
    </bouquet>
  </bouquets>
  <ventes-par-jour>
    <date mois="sep 2021" jour="14">
      <vente bnom="Exotique">4</vente>
    </date>
    <date mois="juin 2022" jour="30">
      <vente bnom="Valentin">4</vente>
      <vente bnom="Printemps">1</vente>
      <vente bnom="Belge">2</vente>
    </date>
    <date mois="sep 2022" jour="18">
      <vente bnom="Valentin">3</vente>
      <vente bnom="Printemps">6</vente>
    </date>
    <date mois="sep 2022" jour="19">
      <vente bnom="Valentin">2</vente>
      <vente bnom="Belge">6</vente>
    </date>
  </ventes-par-jour>
</fleuriste>
```

FIGURE 1 – Document XML

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE fleuriste [
<!ELEMENT fleuriste (fleurs, bouquets, ventes-par-jour)>
<!ELEMENT fleurs (fleur)*>
<!ELEMENT bouquets (bouquet)*>
<!ELEMENT bouquet (fleur)*>
<!ELEMENT fleur (#PCDATA)>
<!ELEMENT ventes-par-jour (date)*>
<!ELEMENT date (vente)*>
<!ELEMENT vente (#PCDATA)>
<!ATTLIST fleur fnom CDATA #REQUIRED>
<!ATTLIST fleur prix CDATA #IMPLIED>
<!ATTLIST fleur couleur CDATA #IMPLIED>
<!ATTLIST fleur nombre CDATA #IMPLIED>
<!ATTLIST bouquet bnom CDATA #REQUIRED>
<!ATTLIST date mois CDATA #REQUIRED>
<!ATTLIST date jour CDATA #REQUIRED>
<!ATTLIST vente bnom CDATA #REQUIRED>
]>
```

Figure 2 – DTD