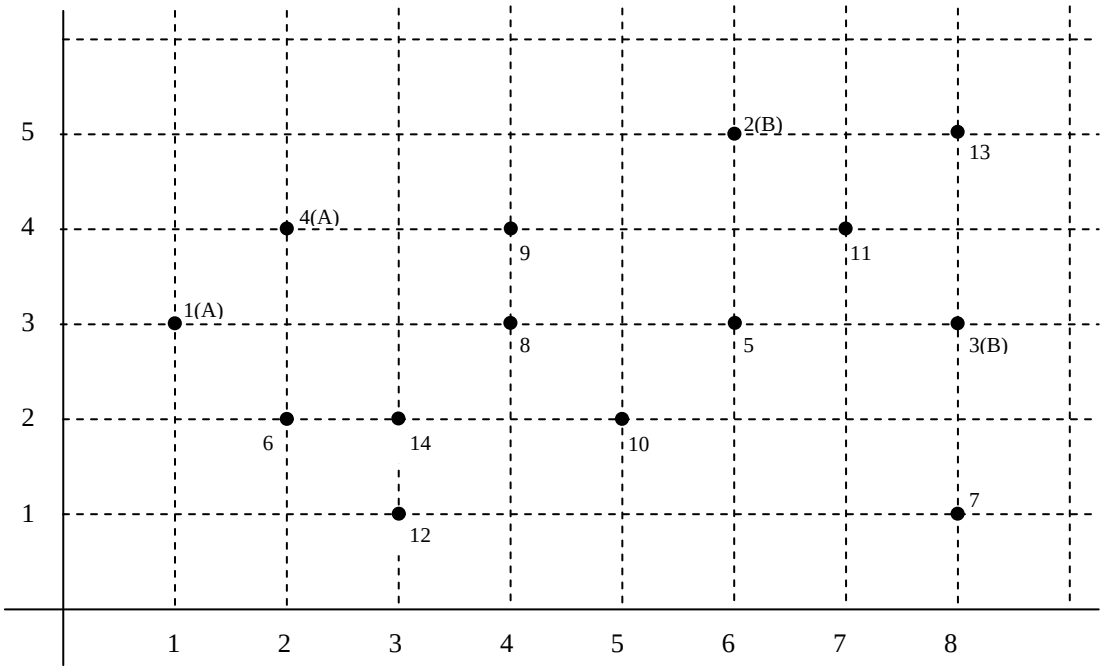
**Ministère de l'Enseignement supérieur et de la recherche scientifique**
**Université Abderrahmane Mira de Bejaïa**
**Faculté des Sciences Exactes**
**Département de Mathématiques**
**Année universitaire : 2022-2023**

| | | |
|---|---|---|
| **Enseignante : Dr. BERMAD**<br>**Durée: 1h30** | **TD 1 Data Mining :**<br>**Techniques de Classification** | **A l'intention de : L3STID** |

## Exercice 1

Dans la figure 1, les points représentent un ensemble de vecteurs de dimension 2, appartenant à 2 classes appelées A et B. L'ordre de sélection des vecteurs est indiqué par les indices situés à côté de chacun. Les points 1 à 4 sont déjà classés; on applique donc l'algorithme en commençant avec le point 5.




1. Appliquer la méthode des KPPV avec k=3. Ecrire la classe résultante à côté de chaque point. Préciser la démarche en prenant quelques points comme exemples.
2. Montrer par un exemple (tiré de la figure 1 ou proposé) que le résultat de la classification dépend de l'ordre de présentation des exemples.

**Exercice2:**

Notre département mathématiques désire utiliser l'apprentissage automatique afin d'améliorer le processus de sélection des étudiants pour l'accès à la première année master PSA pour la prochaine année 2023/2024. Pour atteindre cet objectif, le responsable de la spécialité PSA a décidé d'utiliser les résultats des étudiants inscrits en M1 PSA pour l'année en cours en se basant sur leurs résultats dans les matières: Analyse Numérique et Algèbre obtenus en Licence. La table suivante résume les données rassemblées pour l'entrainement:

| **N°** | **Analyse Numérique** | **Algèbre** | **Admis en M1 (2022/2023)** |
|---|---|---|---|
| 1 | Excellent | Moyen | Oui |
| 2 | Faible | Excellent | Non |
| 3 | Moyen | Moyen | Non |
| 4 | Moyen | Excellent | Oui |
| 5 | Faible | Faible | Non |
| 6 | Excellent | Faible | Non |

1. Calculer la décision estimée pour l'étudiant ayant les mentions (Analyse Numérique: Moyen, Algèbre: Faible) par la méthode KPPV avec K=5 en utilisant la distance de Hamming généralisée:

$$\begin{cases} D(x_i, x_j) = 1 - \frac{1}{Nb_{Att}} \sum_{k=1}^{Nb_{Att}} \frac{f(x_{ik}, x_{jk})}{Nb\ modalités} \\ \textbf{avec} \quad f(x_1, x_2) = 1 \quad \textbf{si} \quad x_{ik} = x_{jk} \quad , \quad 0 \ \textbf{sinon} \end{cases}$$

2. Construire un modèle de décision bayésien en utilisant l'estimateur de Laplace.
3. Calculer la décision obtenue par ce modèle pour l'étudiant de la question précédente.

**Exercice3:**

Etant donné l'ensemble d'apprentissage " jouer au tennis?"

| Jour | Ciel | Température | Humidité | Vent | Jouer au tennis? |
|---|---|---|---|---|---|
| 1 | Ensoleillé | Chaude | Élevée | Faible | Non |
| 2 | Ensoleillé | Chaude | Élevée | Fort | Non |
| 3 | Couvert | Chaude | Élevée | Faible | Oui |
| 4 | Pluie | Tiède | Élevée | Faible | Oui |
| 5 | Pluie | Fraîche | Normale | Faible | Oui |
| 6 | Pluie | Fraîche | Normale | Fort | Non |
| 7 | Couvert | Fraîche | Normale | Fort | Oui |
| 8 | Ensoleillé | Tiède | Élevée | Faible | Non |
| 9 | Ensoleillé | Fraîche | Normale | Faible | Oui |
| 10 | Pluie | Tiède | Normale | Faible | Oui |
| 11 | Ensoleillé | Tiède | Normale | Fort | Oui |
| 12 | Couvert | Tiède | Élevée | Fort | Oui |
| 13 | Couvert | Chaude | Normale | Faible | Oui |
| 14 | Pluie | Tiède | Élevée | Fort | Non |

Jeu de données "Jouer au tennis?"

1. Construire l'arbre de décision en utilisant l'algorithme ID3?
2. En appliquant l'algorithme des k plus proches voisins (KNN), allez-votre jouer s'il y a du soleil, beaucoup d'humidité, température moyenne et pas de vent ?

**Exercice 4 :**

1. Sur le jeu de données « jouer au tennis ? »:
2. Quelle est la classe prédite en utilisant la règle de Bayes d'une journée ensoleillé avec vent faible ?
3. Quelle est la classe d'une journée ensoleille, température de 23°C, humidité de 70 % et vent faible (jeu d'apprentissage2)?
4. Quelle est la probabilité de jouer un jour où la température est de 23°C?
5. Quelle est la probabilité de jouer un jour où l'humidité est comprise entre 60 et 75 %?

**Exercice5 :**

Etant donné l'ensemble d'apprentissage suivant:

| Numéro | Ensoleillement | Température (°F) | Humidité(%) | Vent | Jouer |
|---|---|---|---|---|---|
| 1 | Soleil | 75 | 70 | Oui | Oui |
| 2 | Soleil | 80 | 90 | Oui | Non |
| 3 | Soleil | 85 | 85 | Non | Non |
| 4 | Soleil | 72 | 95 | Non | Non |
| 5 | Soleil | 69 | 70 | Non | Oui |
| 6 | Couvert | 72 | 90 | Oui | Oui |
| 7 | Couvert | 83 | 78 | Non | Oui |
| 8 | Couvert | 64 | 65 | Oui | Oui |
| 9 | Couvert | 81 | 75 | Non | Oui |
| 10 | Pluie | 71 | 80 | Oui | Non |
| 11 | Pluie | 65 | 70 | oui | Non |
| 12 | Pluie | 75 | 80 | Non | Oui |
| 13 | Pluie | 68 | 80 | Non | Oui |
| 14 | Pluie | 70 | 96 | Non | Oui |

1. Construire l'arbre en utilisant l'algorithme C4.5?
2. Estimer la précision du modèle obtenu.