Université Echahid Hama Lakhdar, El-oued
Institut des sciences exactes
Departement d'informatique
1^{er} *Master* : Intelligence Artificielle et Systémes distribués
Semestre : 1.2021

— TP1 : Analyse de données —

---

1 Exercice1 : 4pts

Soit, la série statistique S={5,7,10,15,19,21,21,22,22,23,23,23,23,23,24,24,24,24,25}

- Calculer la médiane.
- Quelles sont les quartiles Q_0, Q_1, Q_2, Q_3, Q_4.
- Calculer le rang interquartile IQR, $IRQ = Q_3 - Q_1$.
- Calculer : $Q_3 + 1.5 \times IRQ$ et $Q_1 - 1.5 \times IRQ$

Les données qui se trouvent hors l'intervalle $]Q_1 - 1.5 \times IRQ..Q_3 + 1.5 \times IRQ[$ sont les données aberrantes (Outlier), les données de faible aberrantes $<Q_1 - 1.5 \times IRQ$ et les données d'aberrantes élevées $> Q_3 + 1.5 \times IRQ$.

- Trouver les données d'aberrants faibles et élevé.
- Dessiner Boxplot en R de la série S et expliquer votre résultat.

2 Exercice2 : 12pts

On considère les données, data.csv, train.csv et test.csv tell que :

- data.csv : Les données totales (Figure 1).
- train.csv : Les données utilisées pour la création du modèle.
- test.csv : les données utilisées pour tester le modèle.

Les attribues des données $X0$ et $X1$ numérique et $y \in \{0, 1\}$

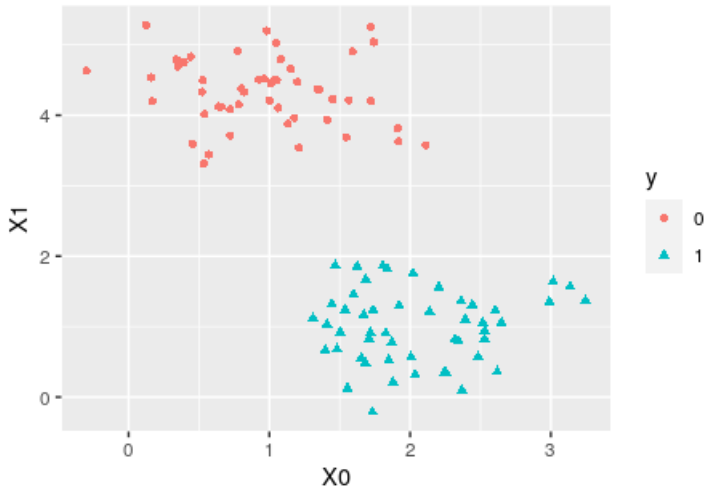

Figure 1 –

1. Construire le modèle (**linear_model**) de régression linéaire multiple de la forme $y = f(X0, X1)$.
2. La forme de régression logistique de la forme :

$$\mathbb{P}(y = 1|X = (X0, X1)) = \frac{\exp(\beta +_0 X_0 + \beta_1 X_1)}{1 + \exp(\beta +_0 X_0 + \beta_1 X_1)} \quad (1)$$

   Construire le modèle (**logistic_model**) de la régression logistique.
3. Prédire la classe de X=(2.75,1.5)
   - Par le modèle **linear_model**
   - Par le modèle **logistic_model**

4. Prédire les classes de données test.csv
   - Par le modèle **linear_model**
   - Par le modèle **logistic_model**
5. On veut évaluer les résultats des modèles **linear_model** et **logistic_model** par deux méthodes Confusion Matrix (Confusion Matrix) et la courbe CRO (ROC curve).
   - Évaluer la qualité des modèles **linear_model** et **logistic_model** par Confusion Matrix et ROC curve.
6. Quel est le modèle le plus judicieux {argumenter votre réponse}
7. Après la fin de ce Tp comment voyer vous le modèle linéaire et le modèle logistique.

3 La qualité de présentation des réponses 04 pts

- Pour assurer la meilleure présentation, je vous conseille d'utiliser : **R notebook** et LaTeX.