Sécurité des Bases de Données :

Anonymisation et Analyse de Données

Le sujet utilise la base de données disponible ici :

http://archive.ics.uci.edu/ml/machine-learning-databases/00320/

Des informations sur le jeu de données sont disponible ici :

http://archive.ics.uci.edu/ml/datasets/Student+Performance

1- Téléchargez le fichier student-mat.csv On suppos que les données sensibles sont les colonnes G1, G2 et G3.

2- Déterminez combien de n-uplets sont uniques en précisant la méthode que vous avez utilisée, et listez ces n-uplets.

On souhaite effectuer l'analyse suivante : déterminer la probabilité que la moyenne d'un élève (G1+G2+G3) / 3 soit supérieure à 14, si ses deux parents Fedu et Medu ont effectué des études supérieures (valeur 4). Vous pouvez si vous le souhaitez retraiter ces données (dans un sgbd, sous excel, etc.) pour changer la valeur pour la transformer en une variable booléenne.

3- Donnez une requête SQL permettant de calculer cette probabilité sur le jeu de données. **Intégrer les données dans une base et donner le résultat exact est un bonus.**

Effectuez une k-anonymisation de ce jeu de données (k=20) en utilisant pour les données numériques des intervalles de 5, 10, 20, 40, et * et pour les données catégorielles, directement une suppression de la donnée (valeur *).

4- Combien avez-vous de classes d'équivalence ? Listez-les.

5- Pouvez-vous directement appliquer la requête de la question 1.3 ? Sinon, comment la transformer et comment l'interpréter ? **Donner le résultat de cette requête est un bonus.**

6- Exécutez une analyse de des données en clair de type **classification**, par rapport à la variable catégorielle définie en 1.2. Divisez le jeu de données en 66% de jeu d'apprentissage et 33% de jeu de test. Quelle qualité de résultats obtenez vous ? Est-ce que vos résultats sont exploitables en terme d'*explicabilité* ?

7- Exécutez la même tâche d'apprentissage sur ce même jeu de test anonymisé. Quels résultats obtenez vous ? Exécutez un apprentissage sur un jeu de données anonymisé, puis sur un jeu de test anonymisé. Quels résultats obtenez vous ? Vous pouvez comparer plusieurs algorithmes d'apprentissage.

8- Même question avec un algorithme de clustering.