Nom: Prénom:

M2 ISL. Sécurité avancée
Protection de la vie privée.

Tous les documents sont autorisés. Épreuve individuelle qui ne peut être réalisée que par vous ! Toute communication est interdite. Toutes les réponses doivent être justifiées. **Sans justification, une réponse est considérée comme fausse.**

1 Etude des données

Dans toute cette partie, on considère le jeu de données de la FIGURE 1. Cette table contient 12 enregistrements précisant les attributs tels que l'âge, le niveau de revenu, le fait d'être salarié ou non et si la personne a déjà souffert d'un trouble cardiaque.

Nom	Âge	Revenu	Salarié.e	Trouble cardiaque
Adams	jeune	élevé	Non	Non
Baker	jeune	élevé	Non	Non
Cook	mûr	élevé	Non	Oui
Dodd	sénior	moyen	Non	Oui
Engel	sénior	faible	Oui	Oui
Flynn	sénior	faible	Oui	Non
Grady	mûr	faible	Oui	Oui
Hayes	jeune	moyen	Non	Non
Irons	jeune	faible	Oui	Oui
Jones	sénior	moyen	Oui	Oui
Knapp	jeune	moyen	Oui	Oui
Lord	mûr	moyen	Non	Oui

FIGURE 1 – Données concernant des patients fictifs d'une clinique étudiant les troubles cardiaques.

1. Si vous vouliez garantir du 2-anonymat, quel ensemble d'attributs considéreriez-vous comme quasi-identifiants ?
2. Cette table est-elle 2-anonyme ?
3. Construire l'histogramme de l'attribut Revenu pour les personnes souffrant de trouble cardiaque.

2 Apprentissage supervisé sans PVP

Dans cette partie, on souhaite estimer si une personne se présentant à nous risque de souffrir d'un trouble cardiaque.

4. Pourquoi cette section se nomme-t-elle "Appprentissage supervisé…" ?
5. Peut-on mettre en place une démarche de régression linéaire sur ces données pour cet apprentissage ?
6. Se présente Gladys, jeune, avec un revenu élevé et qui est salariée. Mettre en place complètement une démarche d'apprentissage bayésien naïf pour estimer si cette personne a davantage de chances de ne pas souffrir d'un trouble cardiaque que d'en souffrir.
7. Discuter de la pertinence de l'estimation précédente en regard des probabilités obtenues en dernière étape.

3 Apprentissage supervisé avec PVP

A partir d'ici, on veut protéger la confidentialité des données. Dans cet exercice, on va utiliser un algorithme de classification binaire qui vérifie la confidentialité différentielle.

8. Pourquoi la classification bayésienne naïve discrète revient-elle à construire un histogramme par attribut pour chaque valeur de l'attribut à estimer ? Montrer alors que l'on doit construire 6 histogrammes dans l'exemple du contrôle pour estimer la valeur de "Trouble cardiaque" étant donnée n'importe quelle valeur des autres attributs.
9. Pourquoi la sensibilité de la construction d'un histogramme vaut-elle 1 ?
10. On souhaite un budget total ϵ de 1 pour l'exemple précédent. Au vu du nombre d'histogrammes à construire (cf. question 8), quel serait le budget de construction pour chacun de ceux-ci ? Citez le théorème que vous utilisez.

4 Codes PVP

12. On suppose que le jeu de données est stocké dans le fichier `datacontrole.csv`. Donner le code complet qui répond à la question 6 sans confidentialité différentielle puis avec en exploitant la bibliothèque `diffprivlib`. On pourra exploiter le code suivant pour transformer les données catégorielles textuelles en des données numériques discrètes :

```
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
df['Age']= label_encoder.fit_transform(df["Age"])
```

5 Nettoyage a priori

Dans cette section, chaque personne a la possibilité de nettoyer elle-même ses données.

13. Sur quelle hypothèse de confiance sont basés les algorithmes de traitement des données vérifiant la confidentialité différentielle ?
14. Lorsqu'une personne nettoie elle-même ses données, cette hypothèse est-elle encore nécessaire ?
15. Vous allez mettre en place un algorithme de réponse randomisée (possiblement généralisé) avec un ϵ' valant 1 pour le traitement de chaque attribut. Montrer que le budget global ϵ pour le nettoyage individuel des données de chaque utilisateur vaut 4.
16. Pour les attributs "Revenu" et "Trouble Cardiaque", détaillez les paramètres du mécanisme $\mathcal{M}_{GRR}$ que vous mettriez en place.
17. On suppose la base de données largement plus remplie (elle contient 4245 données). Pour l'attribut "Revenu", on observe après nettoyage par $\mathcal{M}_{GRR}$ les effectifs suivants :
    — 1254 revenus faibles,
    — 1842 revenus moyens et
    — 1149 revenus élevés.
    Donnez une estimation des effectifs de ces valeurs avant le nettoyage par $\mathcal{M}_{GRR}$.