Nom: Prénom:

L3 informatique. Bases de données avancées.
Seconde chance. Partie J.-F. COUCHOT.

Seule une feuille A4 de notes manuscrites personnelle est autorisée. Tous les moyens de communication sont interdits. Toutes les réponses doivent être justifiées. Sans justification, une réponse est considérée comme fausse.

1 k-anonymat et métriques

On considère les données de la figure 3 qui sont un très petit extrait (issues des recherches [1]) de celles recueillies dans différents hôpitaux, cliniques communautaires et maternités au Bangladesh et donnant un niveau de risque sur les maternités de patientes. On a les attributs suivants :

— âge : âge en années de la patiente enceinte ;
— pression systolique : valeur supérieure de la pression artérielle en mmHg ;
— pression diastolique : valeur inférieure de la pression artérielle en mmHg ;
— glycémie : exprimée en termes de concentration molaire, en mmol/L ;
— fréquence cardiaque : fréquence cardiaque normale au repos, en battements par minute ;
— niveau de risque : niveau d'intensité du risque de la grossesse.

L'objectif est de mettre en place de l'apprentissage supervisé sur ces données. Mais comme vous devez vous conformer au RGPD, les données doivent être nettoyées. On considère comme admis que l'âge est un quasi-identifiant.

1. La figures 1 propose une version 3-anonyme de cet extrait de données. Quelle généralisation a été mise en place ? Quel est le pourcentage des données supprimées ?

1. Ahmed M., Kashem M.A., Rahman M., Khatun S. (2020) Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT). In : Kasruddin Nasir A. et al. (eds) InECCE2019. Lecture Notes in Electrical Engineering, vol 632. Springer, Singapore.

Age	Pr. systolique	Pr. Diastolique	Glycémie	Fréq. Card.	Risque
{15,19}	76	49	6.4	77	faible
	120	80	6.8	70	faible
	90	70	7.8	80	faible
	120	80	7.0	70	moyen
	120	80	7.0	70	moyen
{20,22, 23,25}	100	90	7.5	88	faible
	120	85	7.5	88	faible
	140	90	6.8	70	élevé
	130	70	7.01	78	moyen
	120	90	7.8	60	moyen
	140	100	7.2	80	élevé
{32,35,38}	140	90	18.0	88	élevé
	100	70	7.5	66	faible
	110	70	7.9	80	moyen
*	100	65	7.5	66	faible
*	120	80	13.0	70	élevé

FIGURE 1 – Un exemple de 3-anonymat

Nom: Prénom:

2. Donner ci-dessous une autre version 3-anonyme de cet extrait de données. On généralisera l'âge à l'aide d'un intervalle d'amplitude 8 à partir de la valeur 15. Vous avez droit à 15 % de données supprimées.

Age	Pr. systolique	Pr. Diastolique	Glycémie	Fréq. Card.	Risque

3. Pour ces deux versions 3-anonyme, calculer la valeur de Loss.

4. Pour ces deux versions 3-anonyme, calculer le nombre moyen normalisé d'éléments par classe.

5. Selon, vous quelle est la généralisation qui préserve le mieux l'utilité des données ? Pourquoi ?

6. Comparer les deux valeurs de Loss. Est-ce cohérent avec votre intuition ? Que faut-il remettre en cause ?

Nom: Prénom:

2 Apprentissages machine et vie privée

Un apprentissage machine va être mis en place. Ici c'est de l'inférence Bayésienne naïve qui est choisie.

1. Les métriques utilisées dans la classification sont souvent basées sur une matrice de confusion. Expliquer une des deux matrices de confusion donnée à la figure 2.

2. Expliquer ce qu'est la F-mesure dans le cadre d'une classification. Pourquoi cette métrique est-elle privilégiée par rapport à la précision ou la sensibilité ?

3. La figure 2 donne des métriques relatives aux classifications réalisées sur des données issues du fichier complet des risques sur les maternités de patientes et sur des données 10-anonymes. Comparer ces résultats.

4. Pourquoi la classification semble-t-elle plus "exacte" sur les données 10-anonymes que sur les données réelles ?

```
% classification bayesienne naive sur données orginales
=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,654    0,055    0,813      0,654    0,725      0,645    0,862     0,773     high
                 0,933    0,525    0,543      0,933    0,687      0,432    0,758     0,637     low
                 0,137    0,075    0,474      0,137    0,212      0,099    0,633     0,423     mid
Weighted Avg.    0,595    0,250    0,593      0,595    0,540      0,379    0,745     0,603
=== Confusion Matrix ===
   a   b   c   <-- classified as
 178  63  31 |   a = high
   7 379  20 |   b = low
  34 256  46 |   c = mid

% classification bayesienne naive sur données 10 anonymes avec niveau de généralisation égal à 2
=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,909    0,398    0,604      0,909    0,726      0,511    0,812     0,713     low
                 0,301    0,093    0,616      0,301    0,404      0,265    0,707     0,507     mid
                 0,724    0,057    0,824      0,724    0,771      0,697    0,881     0,788     high
Weighted Avg.    0,658    0,205    0,667      0,658    0,631      0,480    0,796     0,665
=== Confusion Matrix ===
   a   b   c   <-- classified as
 369  29   8 |   a = low
 201 101  34 |   b = mid
  41  34 197 |   c = high
```

FIGURE 2 – Sorties des classifications bayésiennes naïves sur les données originales, ou celles 10-anonymes

Age	Pr. systolique	Pr. Diastolique	Glycémie	Fréq. Card.	Risque
15	76	49	6.4	77	faible
15	120	80	6.8	70	faible
19	90	70	7.8	80	faible
19	120	80	7.0	70	moyen
19	120	80	7.0	70	moyen
20	100	90	7.5	88	faible
22	120	85	7.5	88	faible
23	140	90	6.8	70	élevé
23	130	70	7.01	78	moyen
23	120	90	7.8	60	moyen
25	140	100	7.2	80	élevé
32	140	90	18.0	88	élevé
35	100	70	7.5	66	faible
38	110	70	7.9	80	moyen
55	100	65	7.5	66	faible
56	120	80	13.0	70	élevé

FIGURE 3 – Extrait de données médicales sur les risques liées aux maternités