M2 ISL Sécurité appliquée - TP1.
k-anonymat et apprentisage.

Jean-François COUCHOT
couchot[arobase]femto-st[point]fr

22 octobre 2021

1 Introduction

1.1 Objectifs et données

Ce TP s'inspire largement du TD de B. NGUYEN et P. CLEMENTE [1]. Il s'appuie sur un ensemble de données [2] de l'Institut National du Diabète et des Maladies Digestives et Rénales (USA).

L'objectif de l'ensemble de données est de prédire (algorithmiquement) si une patiente d'origine indienne Pima est ou non diabétique, sur la base de certaines mesures diagnostiques.

Dans ce TP, nous allons réaliser des prédictions sur les données dites originales, les données 2-anonymisées et celles qui seront 5-anonymisées. Nous comparerons les résultats de prédictions pour voir en quelle mesure cette "anonymisation" a perturbé l'apprentissage.

1.2 Les outils

— Weka [3] (pour Waikato Environment for Knowledge Analysis) est une suite de logiciels open source d'apprentissage automatique développée à l'université de Waikato (Nouvelle-Zélande). Elle est notamment capable de prétraiter, les données, de les regrouper (data clustering), de réaliser de la classification statistique, ...

— ARX [4] est un logiciel open source complet permettant de protéger des données personnelles sensibles dans un jeu de données. Il prend en charge une grande variété de modèles de protection de la vie privée et de risques, de méthodes de transformation des données et de méthodes d'analyse de l'utilité des données de sortie.

Exercice 1.1. *Mise en place du TP*

1. *Récupérer le jeu de données et renommer le fichier en* `diabetes.csv`.
2. *Télécharger les deux outils dans une des versions utilisables sur votre OS.*

2 Apprentissage sur des données brutes

Exercice 2.1. *Initialisation du projet*

Après avoir lancé Weka, choisir le bouton explorer pour charger le fichier `diabetes.csv`. *En cliquant sur un des attributs (comme à la figure 1), on peut accéder à son histogramme. Le jeu de données comporte 768 patientes caractérisées par 9 attributs dont* Outcome *que l'on souhaite prédire. Les 8 autres sont :*

— *le nombre de fois où la patiente a été enceinte (Pregnancy) ;*
— *son taux de glucose après ingestion au bout de 2h (Glucose) ;*
— *sa tension artérielle (BP en mm Hg) ;*
— *l'épaisseur de la peau de son triceps (TricepsThickness en mm) ;*
— *la prise d'insuline au bout de 2h (Insulin en mu U/ml) ;*
— *l'indice de masse corporelle (BMI en (kg/m)^2) ;*
— *la fonction pedigree de diabete (DiabetesPedigree) ;*

1. http://benjamin-nguyen.fr/ENS/4ASTI-EA-BIGDATA-SECU/TD_ARX_WEKA.pdf
2. https://members.femto-st.fr/jf-couchot/sites/femto-st.fr.jf-couchot/files/content/diabetes.txt
3. WEKA : https://waikato.github.io/weka-wiki/downloading_weka/
4. ARX : https://arx.deidentifier.org/downloads/

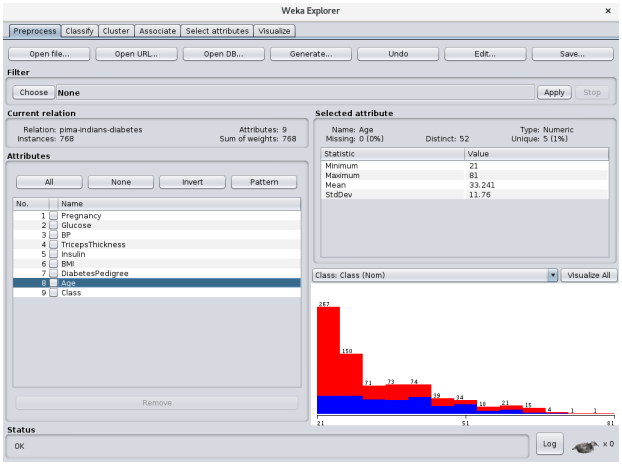

FIGURE 1 – Explorer Weka sur les données PIMA

		classe estimée	
		P	N
Classe réelle	P	TP	FN
	N	FP	TN

FIGURE 2 – Matrices de confusion

— *son âge en années (AGE).*

Selon la littérature[5], les prédicteurs les plus “efficaces” pour ce jeu de données sont les forets aléatoires et la classification baysienne naïve. Dans ce TP on se concentrera sur ce deuxième type de classificateur.

Pour mesurer l’efficacité d’un système de classification, on va évaluer les éléments de la matrice 2x2 de confusion données à la figure 2. Dans cette matrice, chaque ligne L correspond à une classe réelle, chaque colonne C correspond à une classe estimée. La cellule ligne L, colonne C contient le nombre d’éléments de la classe réelle L qui ont été estimés comme appartenant à la classe C.

— P et N signifient respectivement Positive et Négative ;
— TP et TN signifient respectivement True Positive et True Négative ; ces nombre contabilisent les prédictions correctes ;
— FP et FN signifient respectivement False Positive et False Négative ; ces nombre contabilisent les prédictions erronées.

Naturellement, l’objectif est de réduire au maximum les réponses erronées, c’est à dire rendre aussi petites que possible les valeurs de FP et FN.

Quelques indicateurs d’efficacité sont généralement utilisés, ici dans le cas d’une prédiction binaire :

— Sensibilité (Recall en anglais) : $\frac{TP}{TP+FN}$;
— Spécificité : $\frac{TN}{TN+FP}$;
— Pertinence : $\frac{TN+TP}{TN+TP+FN+FP}$.
— Précision : $\frac{TP}{TP+FP}$.
— F-Mesure : $\frac{2 \times \text{Recall} \times \text{Precision}}{\text{Precision} + \text{Recall}}$

On s’intéressera dans ce TP principalement à la F-mesure, comme une mesure agrégée.

5. Benbelkacem, S., & Atmani, B. (2019, April). Random forests for diabetes diagnosis. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE.

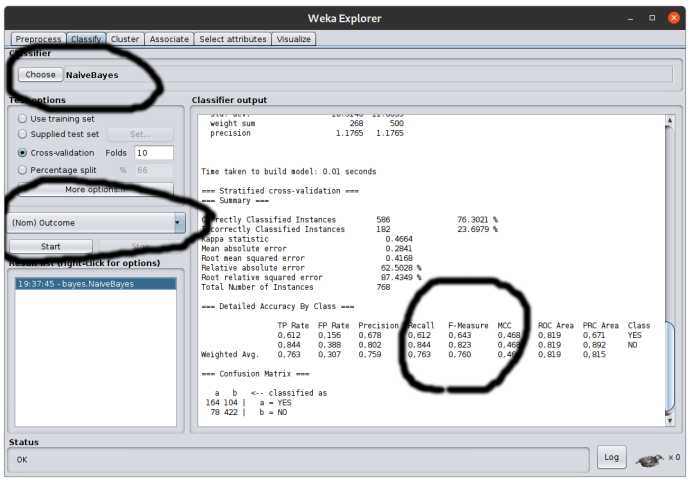

FIGURE 3 – Prédiction de Outcome par classification naïve bayésienne

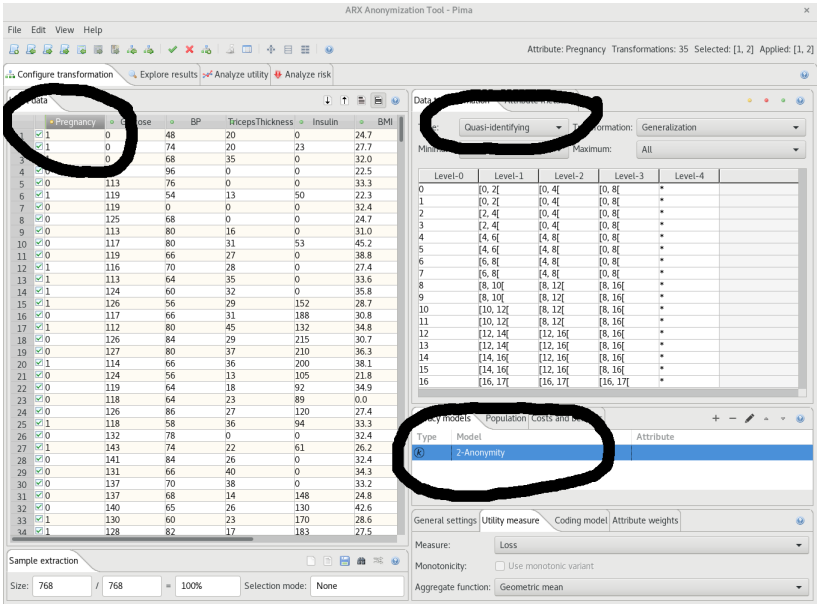

FIGURE 4 – Définir un Quasi IDentifiant, et le 2-anonymat

Exercice 2.2. *Prédiction par réseau bayésien naïf sur les données originales*

1. *On pourra essayer de comprendre ce qu'est la méthode générale de classification naïve bayésienne donnée sur Wikipedia.*
2. *Dans Weka, choisir l'onglet "classify", puis choisir (Choose) l'algorithme de classification naïve de Bayes. Vérifier que c'est bien l'attribut Outcome que l'on cherche à prédire (cf. partie centre-gauche de la figure 3).*
3. *On notera $U_{max} = 0.76$ la valeur de la F-mesure obtenue avec ce predicteur.*

3 2-anonymisation et 5-anonymisation par généralisation

3.1 Mise en place

Dans ce jeu de données, deux attributs peuvent être considérées comme des Quasi IDentifiants : Pregnancy et Age. On va exploiter ARX pour construire efficacement des données k anonymes en généralisant ces Quasi-IDentifiants.

Exercice 3.1. *Import et modification des types*

1. *Importer tout d'abord le jeu de données* `diabetes.csv` *dans l'outil ARX.*
2. *Préciser que les attributs BMI et DiabetesPedigreeFunction sont des nombres décimaux (et pas de chaînes de caractères) dont le séparateur est le point (i.e. en langue anglaise). Ceci se fait dans l'onglet « Attribute metadata ».*

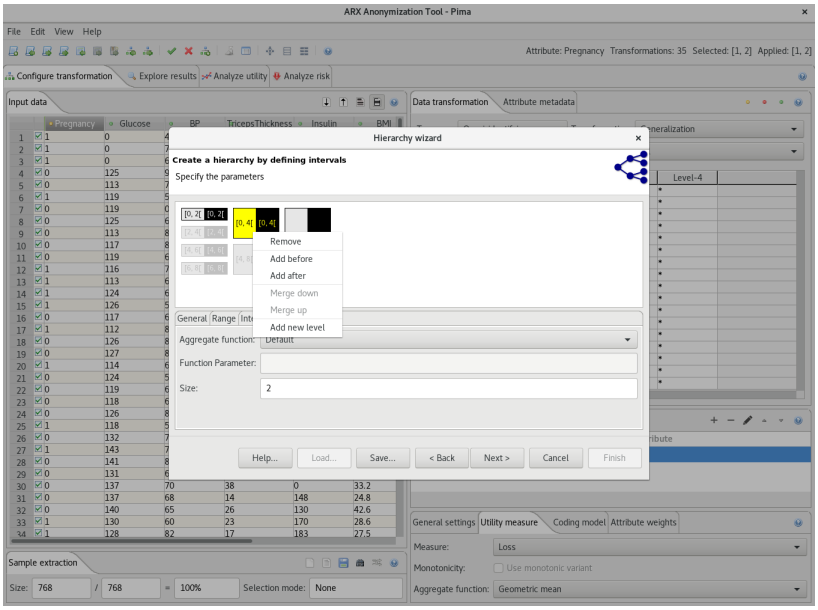

FIGURE 5 – Définir une hiérarchie de généralisation

Exercice 3.2. *Hiéarchies de généralisation*

1. *Spécifier que le type de Pregnancy est Quasi-identifying, comme cela l'est précisé à la figure 4.*
2. *Définir la hiérarchie de généralisation de cet attribut :*
   (a) *on sélectionne l'attribut Pregnancy,*
   (b) *on sélectionne le menu Edit > Create hierachy,*
   (c) *on précise que l'on va raisonner par intervalles,*
   (d) *on précise que le premier intervalle est [0,2[*
   (e) *on ajoute un nouveau niveau de taille 2, en cliquant avec le bouton droit de la souris, comme cela l'est précisé à la figure 5.*
   (f) *Vous devriez avoir 5 niveaux de généralisation : le niveau 0, où rien n'est modifié, les niveaux 1 (amplitude 2), 2 (amplitude 4) et 3 (amplitude 8) et le niveau 4 qui est la généralisation globale (*).*
3. *Définir de même dans ARX la généralisation correspondant à l'attribut Age. Vous devriez avoir 7 niveaux de généralisation.*
4. *Combien il y a-t-il de possibilités de généralisation au total ? Expliquer.*

3.2 2-anonymat et 5-anonymat

Exercice 3.3. 2-*anonymat*

On Choisit d'abord le modèle de vie-privé correspondant au k anonymat et on fixe k à 2, comme représenté en bas à la figure 4.

1. *Dans ARX, choisir le modèle de 2 anonymat et demander à l'outil de générer un modèle 2-anonyme, (Edit>Anonymize) Cette étape est réalisée instantanément.*
2. *Dans longlet « Utility mesure » sous le champ « 2-anonymity », dire quelle metrique est utilisée pour comparer les résultats.*
3. *Dans l'onglet Explore results apparaissent les deux solutions :*
   — $[4, 4]$ *: à quelle généralisation cela correspond-il ? Quel est le score ? Que siginifie celui-ci ?*
   — $[4, 6]$ *: mêmes questions.*

On va autoriser quelques suppressions de données (5%) pour aboutir au k-anonymat. Cela va permetre d'enlever les données les plus problématiques, c'est à dire celles qui sont difficilement joignables à d'autres.

Exercice 3.4. 2-*anonymat en autorisant quelques suppressions*

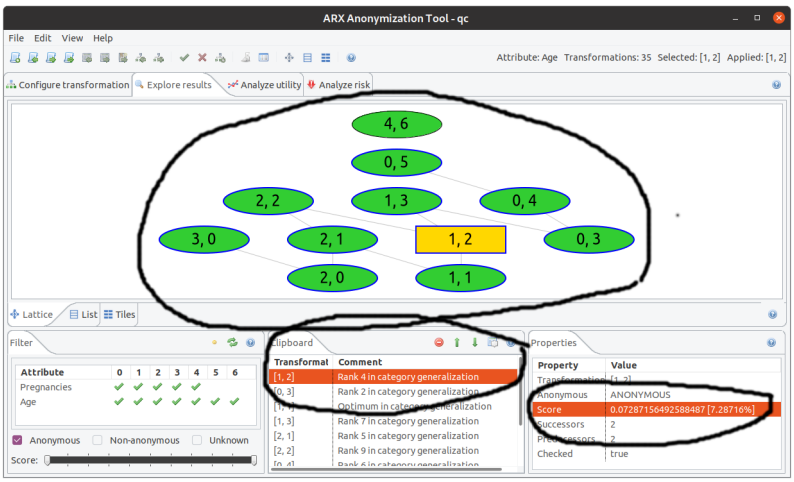

FIGURE 6 – Treillis de généralisation

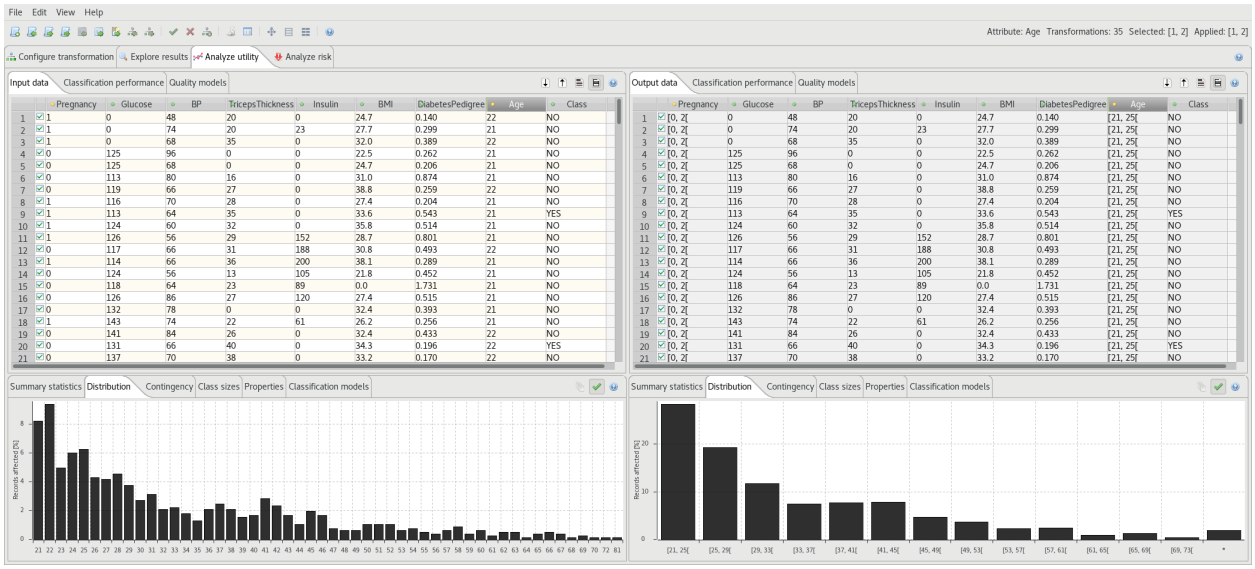

FIGURE 7 – Distributions de l'attribut Age

1. *Fixer ce paramètre de suppression autorisée dans les paramètres généraux (General settings) d'ARX (juste en dessous de k-anonymity.*
2. *Relancer alors la demande d'anonymisation.*
3. *La figure 6 (en haut) montre un extrait de ce treillis en mettant en jaune (1,2) celui qui possède le meilleur score (en bas) en terme de perte d'information. A priori, c'est cette généralisation qui est la plus intéressante.*
4. *Quelle est l'amplitude des classes du nombre d'enfants, de l'âge ? Vérifier ceci en regardant la distribution de ces deux attributs dans l'onglet Analyse utility, comme représenté à la figure 7*
5. *Pour réaliser ce 2-anonymat, certains enregistrements ont été supprimés (voir l'onglet Class sizes, à droite de Distribution). Combien ? Cela est-il significatif par rapport au jeu de données gloabl ?*

Il reste à exporter les données 2 anonymisées pour pouvoir les analyser ultérieurement.

Exercice 3.5. *Export de données 2-anonymes*

1. *Utilisez File > Export Data dans un fichier nommé* `diabete_k_2.csv` *pour réaliser cet export.*

Exercice 3.6. *5-anonymat en autorisant quelques suppressions*

1. *Changer le modèle de protection de la vie privée pour du 5-anonymat tout en autorisant quelques suppressions.*
2. *Combien de données on été supprimées ?*
3. *Exporter les données dans un fichier nommé* `diabete_k_5.csv`.

4. *La stratégie de généralisation qui possède le meilleur score abstrait grandement l'âge. Le constater sur la figure représentant la distribution de cet attribut. Appliquer la transformation (2,2). Combien de données on été supprimées ?*
5. *Exporter les données dans un fichier nommé* `diabete_k_5_b.csv`.

4 Apprentissage sur des données k-anonymisées

Exercice 4.1. *Prédiction sur des données anonymisées*

1. *Reprendre l'approche de prédiction sur les trois fichiers générés à la section précédente.*
2. *U_{k_2}, U_{k_5} et $U_{k_{5_b}}$ seront les valeurs de F-mesure obtenue pour ces trois fichiers.*
3. *Comparer U_{k_2}, U_{k_5}, $U_{k_{5_b}}$ et U_{max}. La qualité des prédictions a-t-elle souffert de la mise en oeuvre de la k-anonymisation ?*
4. *Comment interpréter ceci ?*