Sécurité Appliquée-TP
Protection de la vie privée-PVP

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

13 octobre 2020

Table des matières

Chapitre 1

Apprentissage et Anonymisation

Ce TP s'inspire largement du TD de B. NGUYEN et P. CLEMENTE [1].

Ce TP 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 brutes, 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.

On prétraitera ce classeur de données pour remplacer la valeur 1 (resp. 0) par Yes (resp. No) dans la colonne Outcome.

1.1 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. Le télécharger dans une des versions utilisables sur votre OS.

1.2 Apprentissage sur des données brutes

1.2.1 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 1.1), 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) ;
- 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.

1. http://benjamin-nguyen.fr/ENS/4ASTI-EA-BIGDATA-SECU/TD_ARX_WEKA.pdf
2. https://www.kaggle.com/uciml/pima-indians-diabetes-database
3. WEKA :https://waikato.github.io/weka-wiki/downloading_weka/
4. ARX : https://arx.deidentifier.org/downloads/
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.

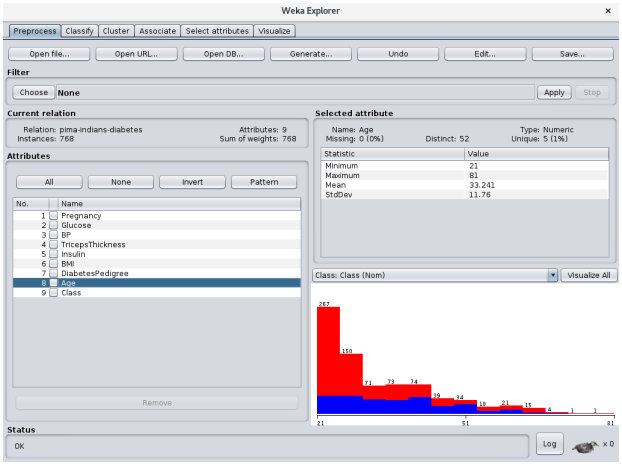

FIGURE 1.1 – Explorer Weka sur les données PIMA

Pour mesurer l'efficacité d'un système de classification, on va évaluer les éléments de la matrice 2x2 de confusion. 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.

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

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.

1.2.2 Forêt aléatoires

Commencer par comprendre le principe général d'une prédiction par forêt aléatoire.

Ensuite, dans Weka, choisir l'onglet "classify", puis choisir (Choose) l'algorithme de forêt aléatoire (random forest) classé dans la famille des arbres (tree). Vérifier que c'est bien l'attribut Outcome que l'on cherche à prédire (cf. partie centre-gauche de la figure 1.2). Enfin lancer l'algorithme d'apprentissage (Start) et interpréter le résultat affiché dans la partie droite. On s'intéressera particulièrement à la F-mesure.

Avec les paramètres dont les valeurs sont fixées par défaut, on a une F-mesure en moyenne égale à 0.743. (cf. partie basse de la figure 1.2).

On va modifier ces valeurs de paramètre pour accroître cette F-mesure. Pour ce faire, il suffit de double-cliquer sur les paramètres à droite du bouton Choose (cf. partie haute de la figure 1.2). Le nombre d'arbres utilisés est le paramètre que l'on va modifier. Malheureusement, dans l'interface graphique, il se nomme nombre d'itérations. Faire varier ce nombre d'arbres entre 1 et 15 pour trouver le nombre d'arbres qui maximise la valeur de F-mesure.

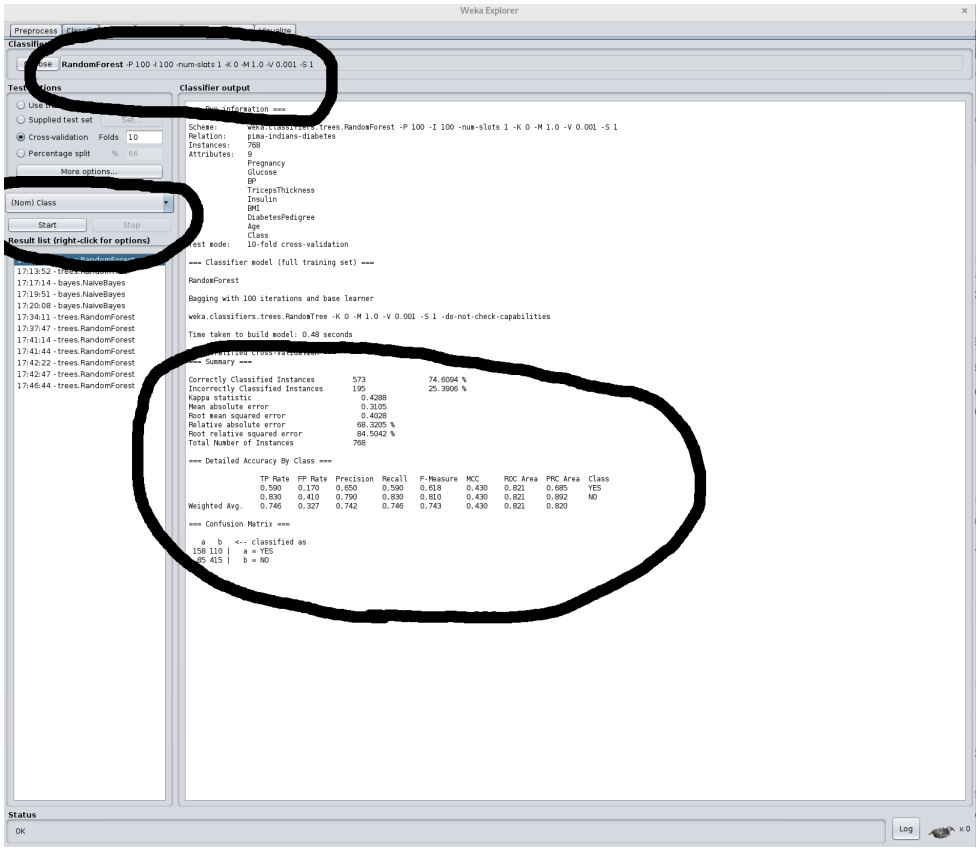

FIGURE 1.2 – Prédiction de Outcome par foret aléatoire

1.2.3 Réseau Bayésien

Comprendre ce qu'est la méthode générale de classification naïve bayésienne et sa particularisation ici.
Dans Weka, choisir l'onglet "classify", puis choisir (Choose) l'algorithme de classification naïve de Bayes.
Comparer les résultats de prédiction avec la méthode à base de forêt aléatoire.

1.3 2-anonymisation et 5-anonymisation par généralisation

Importer le jeu de données dans l'outil ARX.

1.3.1 Définir des hiérarchies de généralisation pour les Quasi IDentifiants

Dans ce jeu de données, deux attributs peuvent être des Quasi IDentifiants : Pregnancy et Age. Les données vont être agrégées selon les règles de généralisation suivante :

— Pregnancy, avec 4 niveaux : classes d'amplitude 2 ([0,2[,. . ., [16,17[), d'amplitude 4, ([0,4[,. . .,[16,17[), d'amplitude 8, ([0,8[,[8,16[, [16,17[),*

— Age, avec 6 niveaux : classes d'amplitude 2, d'amplitude 4, d'amplitude 8, d'amplitude 16, d'amplitude 32,*

Au total, il y a a priori 24 possibilités de généralisation.

Dans ARX, on spécifie que le type de Pregnancy est Quasi-identifying, comme cela l'est précisé à la figure 1.3.
Pour définir la hiérarchie de généralisation de cet attribut,

1. on sélectionne l'attribut Pregnancy,
2. on sélectionne le menu Edit > Create hierachy,
3. on précise que l'on va raisonner par intervalles,
4. on précise que le premier intervalle est [0,2[
5. 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 1.4, jusqu'à avoir 3 niveaux de généralisation (qui va être augmenté de la généralisation globale *).

Définir de même dans ARX la généralisation correspondant à l'attribut Age.

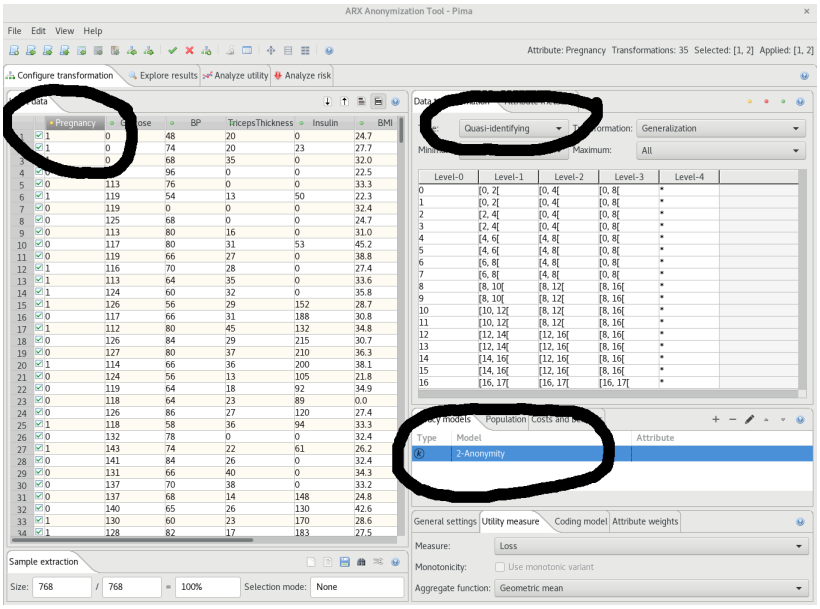

FIGURE 1.3 – Définir un Quasi IDentifiant, et le 2-anonymat

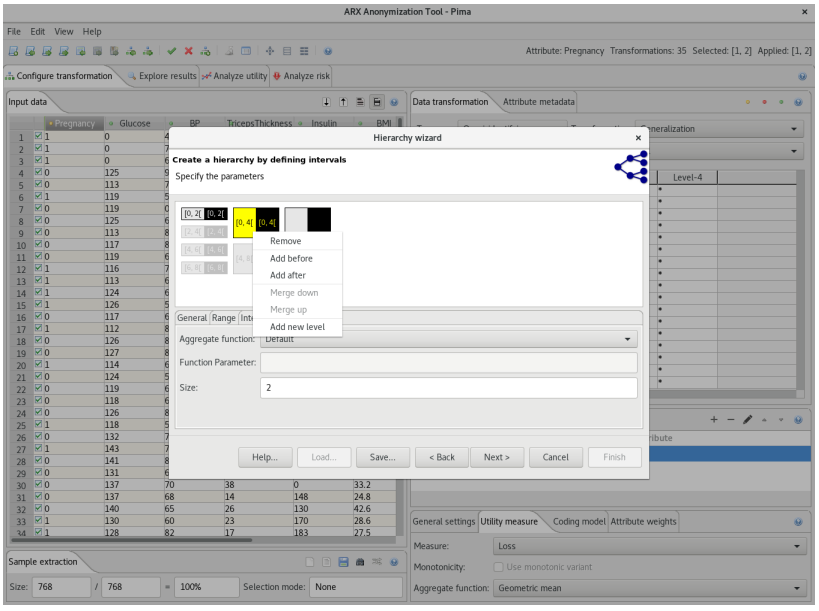

FIGURE 1.4 – Définir une hiérarchie de généralisation

FIGURE 1.5 – Treillis de généralisation

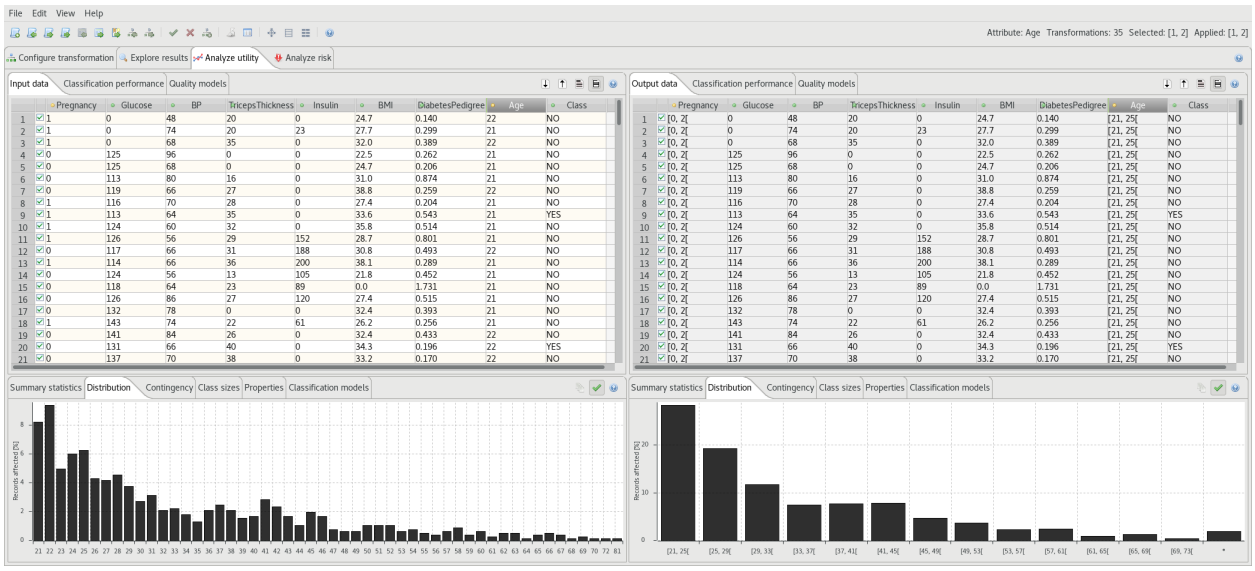

FIGURE 1.6 – Distributions de l'attribut Age

1.3.2 2-anonymat et 5-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 1.3. Cette étape est réalisée instantanément.

Il y avait 24 possibilités de généralisation, organisées selon un treillis. Lorsqu'on analyse le treillis des solutions, on constate qu'il n'y en a que 2 permettent d'aboutir à un 2-anonymat :

4,4 : à quelle généralisation cela correspond-il ? Quel est le score ? Que siginifie celui-ci ?

4,6 : même question.

On va autoriser quelques suppressions de données (5%) pour aboutir au k-anonymat. Fixer ce paramètre dans les paramètres généraux (General settings) d'ARX (juste en dessous de k-anonymity. Relancer alors la demande d'anonymisation.

La figure 1.5 (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. 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 1.6

On constate aussi que pour réaliser ce 2-anonymat, 15 lignes ont été supprimées (voir l'onglet Class sizes, à droite de Distribution).

Il reste à exporter les données 2 anonymisées pour pouvoir les analyser ultérieurement. Pour cela, File > Export Data dans un fichier nommé `diabete_k_2.csv`.

Mettre en oeuvre du 5-anonymat. Combien de données on été supprimées ?

Exporter les données dans un fichier nommé `diabete_k_5.csv`.

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 ? Exporter les données dans un fichier nommé `diabete_k_5_b.csv`.

1.4 Apprentissage sur des données k-anonymisées

Reprendre les deux approches de prédiction sur les trois fichiers générés à la section précédente.

La qualité des prédictions a-t-elle souffert de la mise en oeuvre de la k-anonymisation ?

Conclure.