Workshop on privacy for IOT

Héber HWANG ARCOLEZI
heber.hwang_arcolezi[at]univ-fcomte[point]fr

November 5, 2020

Contents

Chapter 1

Learning and Anonymization

This practical work (lab) is largely inspired by the tutorial of B. NGUYEN et P. CLEMENTE[1]. This lab is based on a dataset[2] of the National Institute of Diabetes and Digestive and Kidney Diseases (USA). The objective of the dataset is to predict (algorithmically) whether or not a patient of Pima Indian descent has diabetes, based on certain diagnostic measures.

In this lab, we will make predictions on the so-called raw data, the 2-anonymized data and those that will be 5-anonymized. We will compare the prediction results to see how this anonymization disturbs learning.

We will preprocess this data classifier to replace the value 1 (resp. 0) by Yes (resp. No) in the Outcome column.

1.1 The tools

Weka[3] (Waikato Environment for Knowledge Analysis) is an open source machine learning software suite developed at the University of Waikato (New Zealand). It is notably capable of preprocessing data, grouping them (data clustering), performing statistical classification, ...

ARX[4] is comprehensive open source software for protecting sensitive personal data in a dataset. It supports a wide variety of privacy and risk models, data transformation methods, and output data utility analysis methods. Download it in one of the versions usable on your OS.

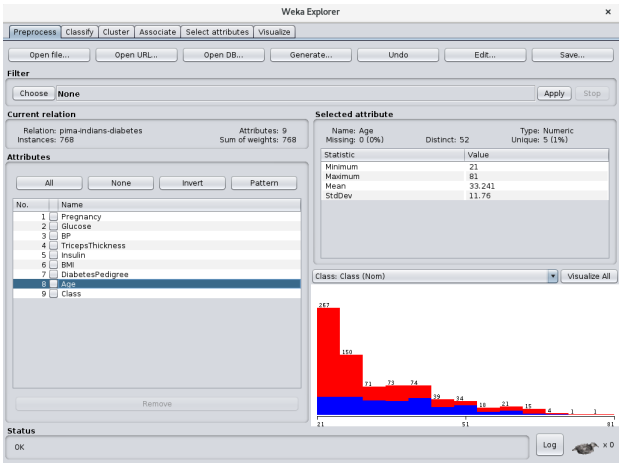

Figure 1.1: Explore Weka on PIMA dataset

1.2 Learning on raw data

1.2.1 Project initialization

After launching Weka, choose the explore button to load the file `diabetes.csv`. By clicking on one of the attributes (as in figure 1.1), you can access its histogram. The data set includes 768 patients characterized by 9 attributes including *Outcome* that we want to predict. The other 8 are:

- the number of times the patient has been pregnant (Pregnancy);
- its glucose level after ingestion after 2 hours (Glucose);
- his blood pressure (BP in mm Hg);
- the thickness of the skin of his triceps (TricepsThickness in mm);
- taking insulin after 2 hours (Insulin in mu U / ml);
- body mass index (BMI in (kg / m) ^{2});
- the diabete pedigree function (DiabetesPedigree);
- its age in years (AGE).

According to the literature[5], the most "efficient" predictors for this dataset are random forests and the naive Bayesian classification.

To measure the effectiveness of a classification system, we will evaluate the elements of the 2x2 confusion matrix. In this matrix, each row L corresponds to a real class, each column C corresponds to an estimated class. The row cell L, column C contains the number of elements of the real class L which were estimated to belong to the class C.

		Estimated class	
		P	N
Real class	P	TP	FN
	N	FP	TN

Some indicators to measure the models' performance are generally used. In the case of a binary prediction:

- Recall: $\frac{TP}{TP+FN}$;
- Specificity: $\frac{TN}{TN+FP}$;
- Accuracy: $\frac{TN+TP}{TN+TP+FN+FP}$.
- Precision: $\frac{TP}{TP+FP}$.
- F-Score: $\frac{2 \times \text{Recall} \times \text{Precision}}{\text{Precision} + \text{Recall}}$

We will focus in this lab mainly on F-score, as an aggregate measure.

[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.

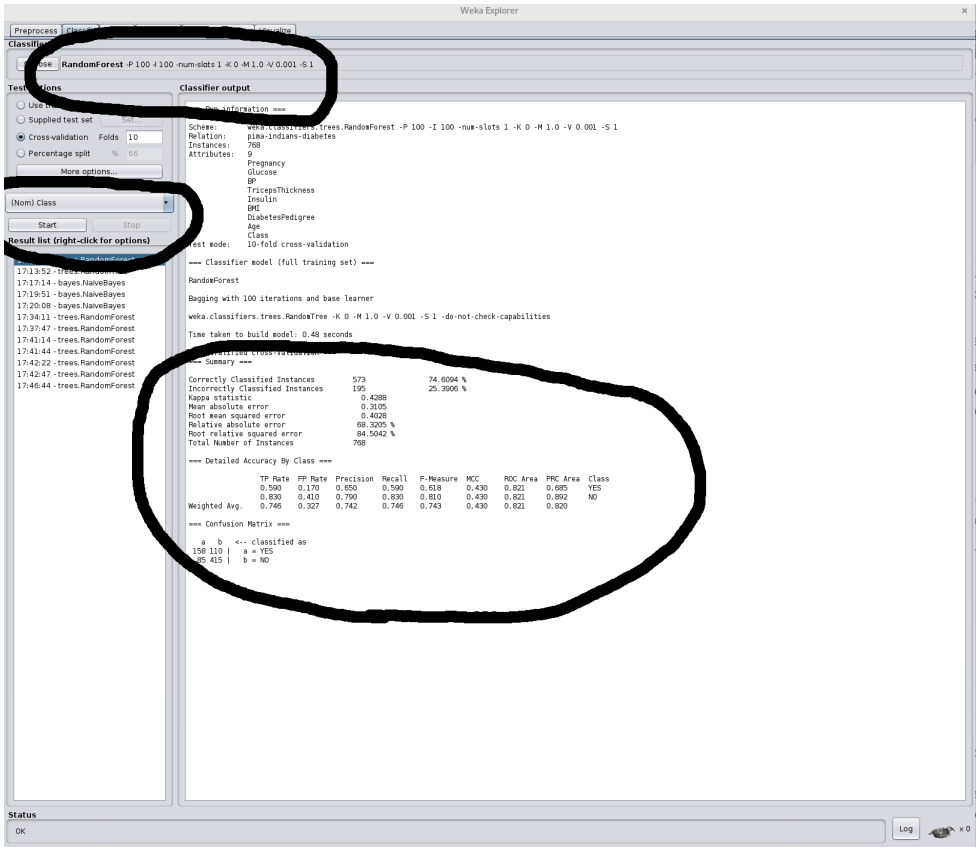

Figure 1.2: Random Forest Outcome Prediction

1.2.2 Random forests

Start by understanding the general principle of a random forest prediction.

Then, in Weka, choose the "classify" tab, then choose (Choose) the random forest algorithm classified in the family of trees (tree). Check that it is indeed the Outcome attribute that we are trying to predict (see center-left part of the figure 1.2). Finally, launch the learning algorithm (Start) and interpret the result displayed on the right. We will be particularly interested in F-measure.

With the parameters whose values are fixed by default, we have an average F-measure equal to 0.743 (see lower part of the figure 1.2).

We will modify these parameter values to increase this F-measure. To do this, just double-click on the settings to the right of the Choose button (cf. upper part of the figure 1.2). The number of trees used is the parameter that we are going to modify. Unfortunately, in the GUI it's called iteration count. Vary this number of trees between 1 and 15 to find the number of trees that maximizes the value of F-measure.

1.2.3 Bayesian Network

Understand what the general naive Bayesian classification method is and its particularization here.

In Weka, choose the " classify " tab, then choose (Choose) the naive classification algorithm of Bayes.

Compare the prediction results with the random forest method.

1.3 2-anonymization and 5-anonymization by generalization

Import the dataset into the ARX tool.

1.3.1 Define generalization hierarchies for Quasi IDentifiers

In this dataset, two attributes can be Quasi IDentifiers: Pregnancy and Age. The data will be aggregated according to the following generalization rules:

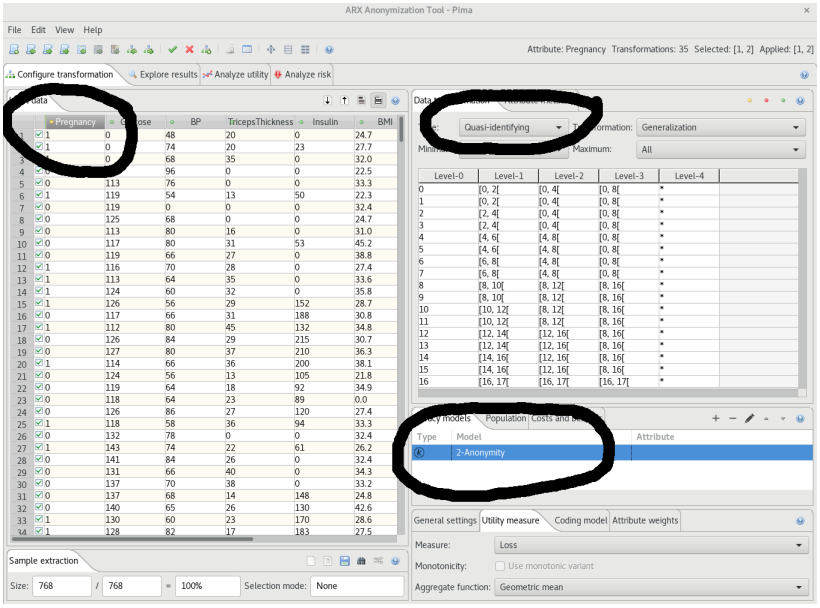

Figure 1.3: Define a Quasi IDentifier, and 2-anonymity

- Pregnancy, with 4 levels : classes with amplitude 2 ([0,2[,. . . , [16,17[), with amplitude 4, ([0,4[,. . . ,[16,17[), with amplitude 8, ([0,8[,[8,16[, [16,17[),*
- Age, with 6 levels : classes with amplitude 2, with amplitude 4, with amplitude 8, with amplitude 16, with amplitude 32,*

In total, there are a priori 24 possibilities of generalization.

In ARX, we specify that the type of Pregnancy is Quasi-identifying, as specified in the figure 1.3.

To define the generalization hierarchy for this attribute,

1. we select the Pregnancy attribute,
2. we select the menu Edit> Create hierachy,
3. we specify that we will reason by intervals,
4. we specify that the first interval is [0,2[
5. we add a new level of size 2, by clicking with the right mouse button, as specified in the figure 1.4, until you have 3 levels of generalization (which will be increased by the global generalization *).

In the same way, define in ARX the generalization corresponding to the Age attribute.

1.3.2 2-anonymity and 5-anonymity

We first choose the privacy model corresponding to k-anonymity and we set $k = 2$, as shown at the bottom of the figure 1.3. This step is performed instantly.

There were 24 possibilities of generalization, organized according to a lattice. When we analyze the lattice of solutions, we see that there are only 2 of them allowing to achieve a 2-anonymity:

4,4 : what generalization does this correspond to? What's the score? What does this mean?

4,6 : same question.

We will allow a few data suppression (5%) to end up with k-anonymity. Set this parameter in the General settings of ARX (just below k-anonymity. Then restart the anonymization request.

Figure 1.5 (top) shows an extract of this lattice by putting in yellow (1,2) the one with the best score (bottom) in terms of information loss. A priori, it is this generalization which is the most interesting. What is the amplitude of the classes of pregnancy and age? Check this by looking at the distribution of these two attributes in the Analysis utility tab, as shown in the figure 1.6

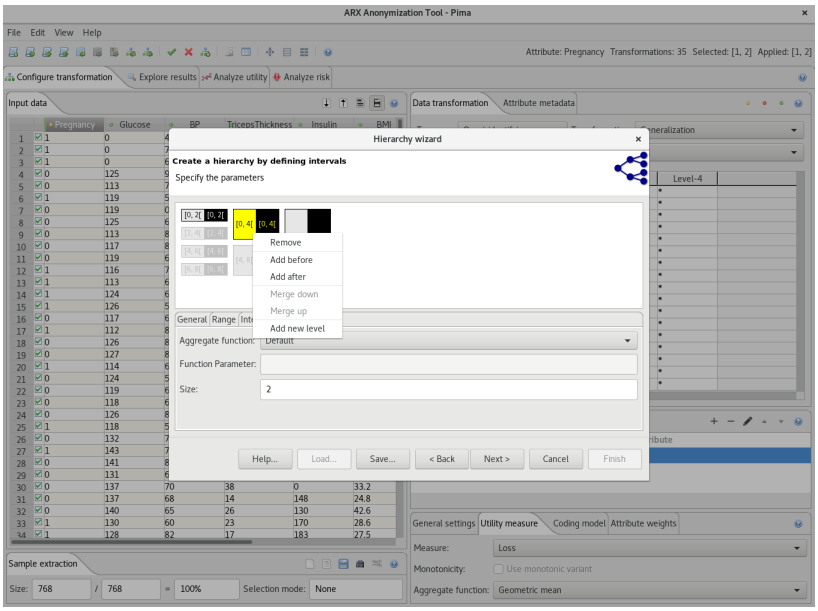

Figure 1.4: Define a generalization hierarchy

Figure 1.5: Generalization lattice

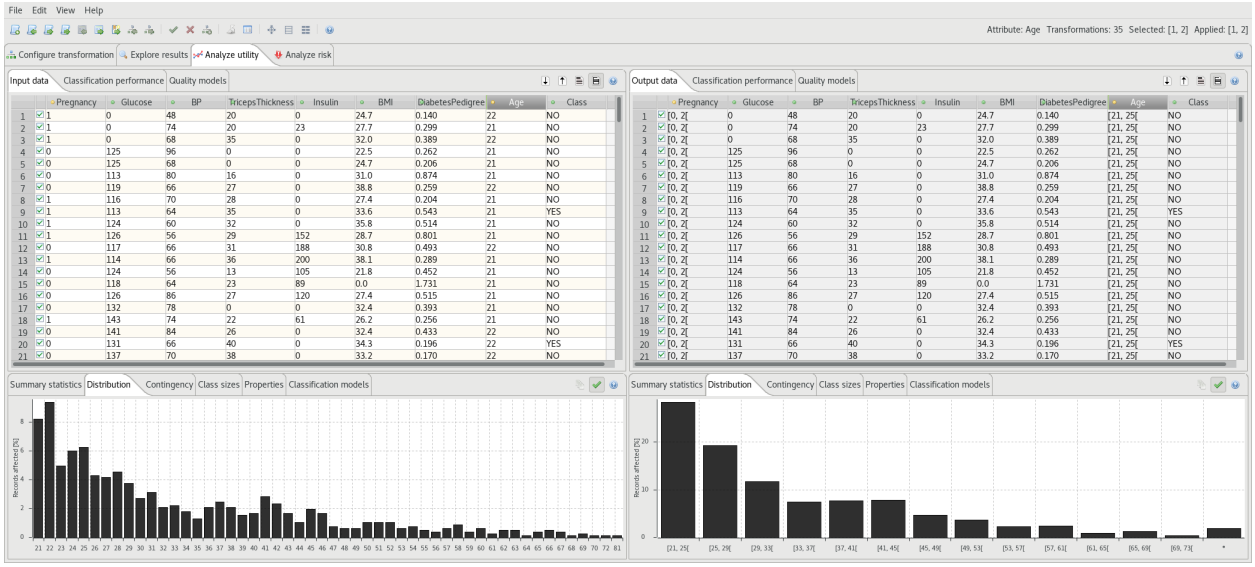

Figure 1.6: Age attribute distributions

We also note that to achieve this 2-anonymity, 15 lines have been deleted (see the Class sizes tab, to the right of Distribution).

It remains to export the anonymized 2 data to be able to analyze them later. To do this, File> Export Data in a file named `diabete_k_2.csv`.

Implement 5-anonymity. How much data has been suppressed?

Export the data to a named file `diabete_k_5.csv`.

The generalization strategy with the best score greatly abstracts age. Note this in the figure representing the distribution of this attribute. Apply the transformation (2,2). How much data has been deleted? Export the data to a named file `diabete_k_5_b.csv`.

1.4 Learning on k-anonymous data

- Perform the two prediction approaches on the three files generated in the previous section.

  1. Has the quality of predictions suffered from the implementation of k-anonymization?
  2. Conclude.