COURS DE FOUILLE DE DONNÉES

M1, P-20- IFI 2016

Enseignante: **NGUYỄN Thị Minh Huyền**

TP2 – G2

Description du jeu de données

Données Utilisées: **Mammographie Mass Data Set**

Étudiant: **Ginel Dorleon**

Gervais Fotsing Sikadie S.

Plan de ce Document

I. INTRODUCTION

De nos jours toutes les entreprises collectent et stockent de grandes quantités de données. Ces mégabases de données, qui ne cessent d'augmenter jour après jour, sont peu exploitées, alors qu'elles cachent de connaissances décisives face au marché et à la concurrence. Pour combler ce besoin, une nouvelle industrie est née : la Fouille de Données ou la Science des Données.
La Fouille de Données a ainsi pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances. [1]

Pour appliquer ainsi les connaissances acquises durant l'introduction du cours de fouille de données, nous allons ainsi étudier et faire une description détaillée d'un jeu de données en utilisant un logiciel spécialisé, *tanagra*

II. CONTEXTE

a- Résumé

Discrimination des masses mammographiques bénignes et malignes basé sur les attributs BI-RADS et l'âge du patient. [2]

Jeu de données considéré: Mammographie Mass Data Set

III. DESCRIPTION

La mammographie est la méthode la plus efficace pour le dépistage du cancer du sein disponible aujourd'hui. Cependant, la faible valeur prédictive positive de la biopsie mammaire résultant de l'interprétation des mammographies conduit à environ 70% de biopsies inutiles avec des résultats bénins. Pour réduire le nombre élevé de biopsies de sein inutiles, plusieurs diagnostics assistés par ordinateur (DAO) ont été proposés au cours des dernières années. Ces systèmes aident les médecins à décider d'effectuer une biopsie du sein sur une lésion observée lors d'une mammographie ou pour effectuer un suivi à court terme. Cet ensemble de données peut être utilisé pour prédire la gravité

(bénigne ou maligne) d'une lésion de masse mammograhique de BI-RADS, les attributs et l'âge du patient. Il contient une évaluation BI-RADS, l'âge du patient et trois attributs BI-RADS en même temps que la vérité au sol (le champ de la gravité) pour 516 bénigne et 445 masses malignes qui ont été identifiés sur la mammographie numérique recueillie à l'Institut de radiologie de l'Université d'Erlangen-Nuremberg entre 2003 et 2006. Chaque instance a une évaluation BI-RADS associé allant de 1 (certainement bénigne) à 5 (très évocatrice de malignité) attribué à un processus à double examen par les médecins. En supposant que tous les cas avec l'évaluation BI-RADS supérieure ou égale à une valeur donnée (variant de 1 à 5), sont malignes et les autres cas bénins, sensibilités et spécificités associés peuvent être calculées. Ceux-ci peuvent être une indication de la façon dont un système de DAO effectue par rapport aux radiologues. [2]

Généralement, le radiologue analyse la radiographie pour décider s'il y a une tumeur ou juste des tissus normaux et si la tumeur existante est maligne (cancéreuse) ou bénigne (doux). En raison des variations des interprétations de la mammographie, le problème est pathologiste. Ainsi le pathologiste analyse les cellules et les tissus sous un microscope pour déterminer s'ils sont malins ou bénins. Les pathologistes aident à caractériser les spécimens prélevés lors de la biopsie ou d'autres procédures chirurgicales et aide à déterminer traitement. Pour déterminer la qualité histologique d'une tumeur, une échantillon de cellule mammaire doit être prélevée sur un sein biopsie, tumorectomie ou mastectomie. Ainsi, cette étude a été effectuée dans le but d'accroître la capacité des médecins à détecter la gravité (bénigne ou maligne) d'une mammographie des lésions de masse à partir des attributs BI-RADS et des âges. L'objectif est de réduire le nombre élevé de biopsies inutiles du sein. [2]

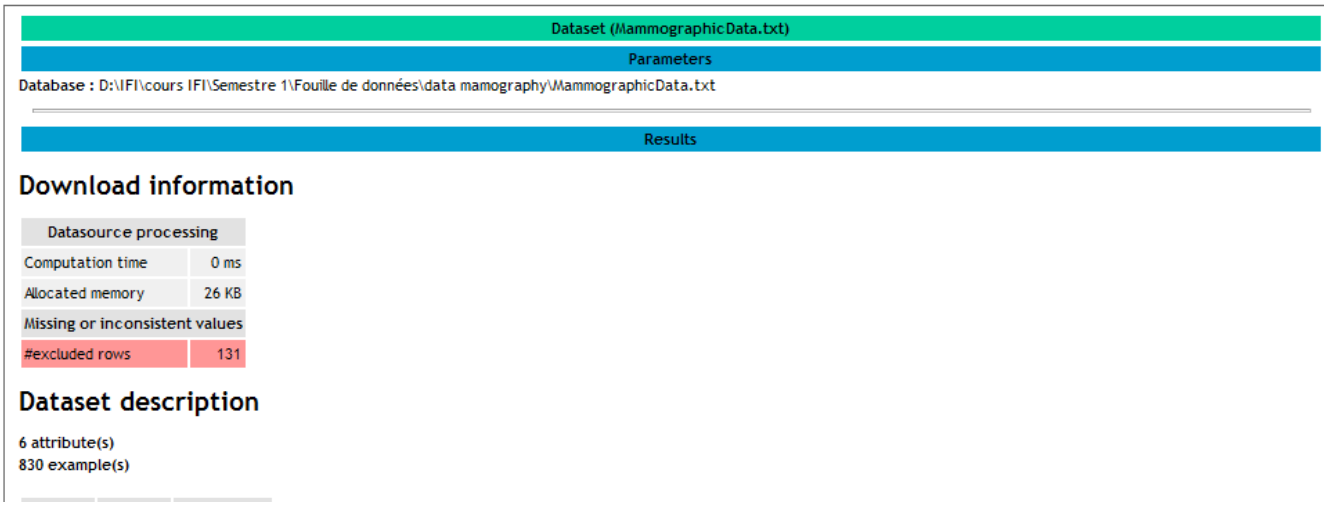
Les données de masse mammographique utilisées ici ont été recueillies à l'Institut de radiologie de l'Université Erlangen-Nuremberg entre 2003 et 2006. BIRADS représente les données sur l'imagerie du sein et les rapports Système et a été développé par l'American College of Radiologie (ACR), en collaboration avec plusieurs Organisations en 1991 à présenter des réponses Rapports de mammographie ambigus avec indécis Conclusions des radiologues . [2]

a- Caractéristique du jeu de données considéré

Ensemble de données Caractéristiques:	multivariée	**Nombre d'instances:**	961	**Région:**	La vie
Attribut Caractéristiques:	Entier	**Nombre d'attributs:**	6	**Date**	2007-10-29

b- Instance

Notre jeu de donnée contient un nombre de 961 instances dont 131 exclues par Tanagra, 830 considérées.


Dataset (MammographicData.txt)

Parameters

Database : D:\IFI\cours IFI\Semestre 1\Fouille de données\data mamography\MammographicData.txt

Results

Download information

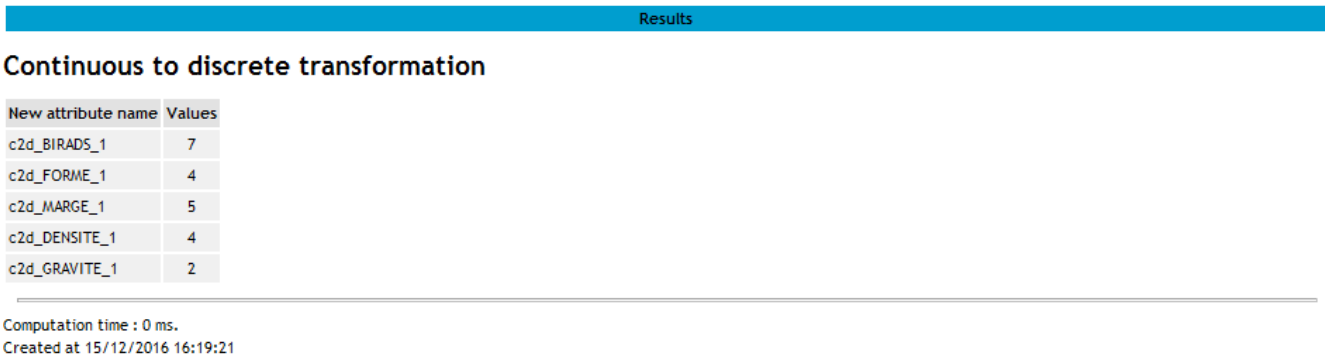
Datasource processing	
Computation time	0 ms
Allocated memory	26 KB
Missing or inconsistent values	
#excluded rows	131

Dataset description

6 attribute(s)
830 example(s)

c- Attribut

Le jeu de donnée considéré contient un total de 6 attributs. Parmi les attributs, on a un (1) attribut pour le champ de l'objectif c'est-a-dire l'attribut Gravité, un (1) attribut non prédictif et enfin quatre (4) attributs prédictif.


Results

Continuous to discrete transformation

New attribute name	Values
c2d_BIRADS_1	7
c2d_FORME_1	4
c2d_MARGE_1	5
c2d_DENSITE_1	4
c2d_GRAVITE_1	2

Computation time : 0 ms.
Created at 15/12/2016 16:19:21

d- Information sur les Attributs

1- *BI-RADS*

Acronyme de *Breast Imaging-Reporting and Data System*, c'est un outil d'assurance de qualité conçu à l'origine pour une utilisation avec la mammographie. Donc, l'attribut BIRADS considère l'évaluation de BIRADS sur un 1 à 5 (ordinal) . C'est un attribut de type discret/discontinu.

2- *Âge*

L'attribut âge représente l'age du patient en années (entier), c'est un attribut de type continu.

3- *Forme*

Attribut de type discret, l"attribut forme caractérise la forme de masse et peut prendre les valeurs suivantes: rond = 1, ovale = 2, lobulaire = 3, irrégulière = 4. Ces valeurs numériques attribuées sont à titre indicatif et peuvent être remplacées par d'autres valeurs dépendamment de l'analyse qu'on veut réaliser.

4- *Marge*

Attribut de type discret, l'attribut décrit la marge de masse et peut prendre les valeurs suivantes: circonscrite = 1, 2=microbullo, obscurcie = 3, mal défini = 4, spiculée = 5.

5- *Densité*

L'attribut densité détermine la densité de masse détectée, cet attribut est de type discret/discontinu et les valeurs que prend cet attribut sont:
élevée = 1, iso = 2, bas = 3, contenant des matières grasses = 4

6. *Gravité*

Cet attribut est le résultat qu'on attend. cet attribut traduit le niveau de gravité de la masse détectée. Cet attribut de type discret prend 2 valeurs qui sont bénigne, désigné par 0 ou maligne désigné par1 (binominal)

IV. CONCLUSION

Dans ce premier travail réalisé nous permet de prendre connaissance de l'ensemble de notre jeu de données. Cela nous permet en autre de bien maîtriser les premiers pas du processus d'exploration des données. Un début très promettant car non seulement nous pouvons faire la description de nos données, mais aussi nous pouvons les utiliser sur notre logiciel d'analyse de données, *tanaga*. Dans un prochain travail, nous présenterons l'analyse descriptive de nos données, la description des données et méthodes statistiques de notre jeu de données.

V. Référence et papiers

[1] https://moodle.insa-rouen.fr/course/view.php?id=92&lang=fr
[2] http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass
[3] https://www.iumsp.ch/sites/default/files/Colloque_IUMSP_20160412.pdf