 ****

**Université Mohammed V de Rabat**

**École nationale supérieure d’informatique et d’analyse des systèmes**

**Filière : Ingénierie Digitale pour la Finance**

Encadré par :

* **Mme Lanbouri Zineb**

Réalisé par :

* **Hibatallah Amine**
* **Abouhadid Yassine**
* **Boutouar Yassine**
* **Akhsas Othmane**

**Introduction :**

Le présent travail s’inscrit dans le projet du module « Réseaux de neuronnes » qui a pour but d’assimiler les compétences acquises du module et de les mettre à profit via l’implémentation d’un réseau neuronal multicouche. Réciproquement, cette même mise en pratique des Réseaux de Neurones est le moyen le plus sûr de maîtriser tous les aspects liés au fonctionnement de l’algorithme.

Ce projet suivra le plan suivant :

* Identification et compréhension du problème,
* Acquisition des données,
* Exploration des données,
* Prétraitement des données,
* Paramétrage,
* Modélisation.

1. **Concepts théoriques :**
2. **Réduction de dimension :**
3. *Définition :*

La réduction de la dimensionnalité est un processus étudié en analyse de données, qui consiste à prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension.

L'analyse en composantes principales (ACP) est l'une des méthodes d'analyse de données multivariées les plus utilisées. Il permet d'explorer des ensembles de données multidimensionnels composés de variables quantitatives. Il est largement utilisé en biostatistique, en marketing, en sciences sociales et dans de nombreux autres domaines.

1. *Pseudocode :*

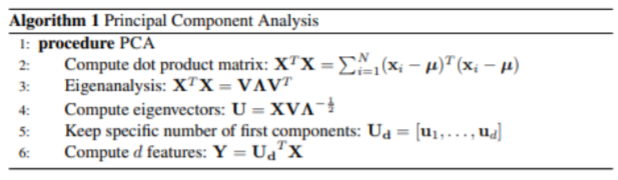


Figure 1 : pseudocode ACP

1. Perceptron multicouche :
2. *Définition :*

Le perceptron multicouche est un type de réseau neuronal artificiel organisé en plusieurs couches au sein duquel les informations circulent de la couche d'entrée à la couche de sortie en passant par des couches cachées; il s'agit donc d'un réseau de propagation directe.

1. *Pseudocode :*



Figure 2: pseudocode Perceptron Multicouche

1. **Compréhension du problème et acquisition des données :**
2. **Problématique :**

Selon l'Organisation mondiale de la santé, le cancer du col de l'utérus est le quatrième Cancer le plus fréquent qui a un taux de mortalité élevé qui affecte les femmes partout dans le monde en particulier dans les pays à revenu faible et intermédiaire. L’évolution qu’ont connu les domaines de la science d’informatique et de la technologie ont abouti ç l’apparition de plusieurs recherches sur ce cancer. Ce présent rapport est destiné à étudier la technique de classification « Perceptron Multicouche » dans l'exploration de données sur les facteurs de risque des ensembles de données sur le cancer du col de l'utérus.

1. **Dataset**
2. *Description :*

L'ensemble de données utilisé dans cette recherche est le facteur de risque de cancer du col de l'utérus extrait de UCI Machine Learning Repository. Cette dataset a été collectée à ‘’ Hospital Universitario de Caracas’’ au Caracas, Venezuela.

Les données sur le cancer du col utérin concernent 858 échantillons et 32 ​​caractéristiques ainsi que 4 classes (Hinselmann, Schiller, Cytology and Biopsy). Ces attributs comprennent des informations démographiques, des habitudes comme le tabagisme et d’anciens records médicaux.

1. *Nécessité du prétraitement :*

L'ensemble de données devrait subir une phase de prétraitement intense des données en raison d'un déséquilibre et de la présence de beaucoup de valeurs manquantes. De plus, des techniques de réduction de dimensionnalité sont nécessaires pour améliorer la précision du classificateur.

La complexité de cette dataset réside dans les multiples approches de dépistage et de diagnostic de ce type de cancer, qui conduisent à un écosystème complexe. Par conséquent, la prédiction du risque factoriel du patient et la meilleure stratégie de dépistage constituent un problème majeur.

1. *Variables :*

1-(int) Age  
2-(int) Number of sexual partners  
3-(int) First sexual intercourse (age)  
4-(int) Num of pregnancies  
5-(bool) Smokes  
6-(bool) Smokes (years)  
7-(bool) Smokes (packs/year)  
8-(bool) Hormonal Contraceptives  
9-(int) Hormonal Contraceptives (years)  
10-(bool) IUD  
11-(int) IUD (years)  
12-(bool) STDs  
13-(int) STDs (number)  
14-(bool) STDs:condylomatosis  
15-(bool) STDs:cervical condylomatosis  
16-(bool) STDs:vaginal condylomatosis  
17-(bool) STDs:vulvo-perineal condylomatosis  
18-(bool) STDs:syphilis  
19-(bool) STDs:pelvic inflammatory disease  
20-(bool) STDs:genital herpes  
21-(bool) STDs:molluscum contagiosum  
22-(bool) STDs:AIDS  
23-(bool) STDs:HIV  
24-(bool) STDs:Hepatitis B   
25-(bool) STDs:HPV  
26-(int) STDs: Number of diagnosis  
27-(int) STDs: Time since first diagnosis  
28-(int) STDs: Time since last diagnosis  
29-(bool) Dx:Cancer  
30-(bool) Dx:CIN  
31-(bool) Dx:HPV  
32-(bool) Dx  
33-(bool) Hinselmann: target variable  
34-(bool) Schiller: target variable  
35-(bool) Cytology: target variable  
36-(bool) Biopsy: target variable

1. *Visualisation et exploration des données:*

Pour les 5 premières lignes:

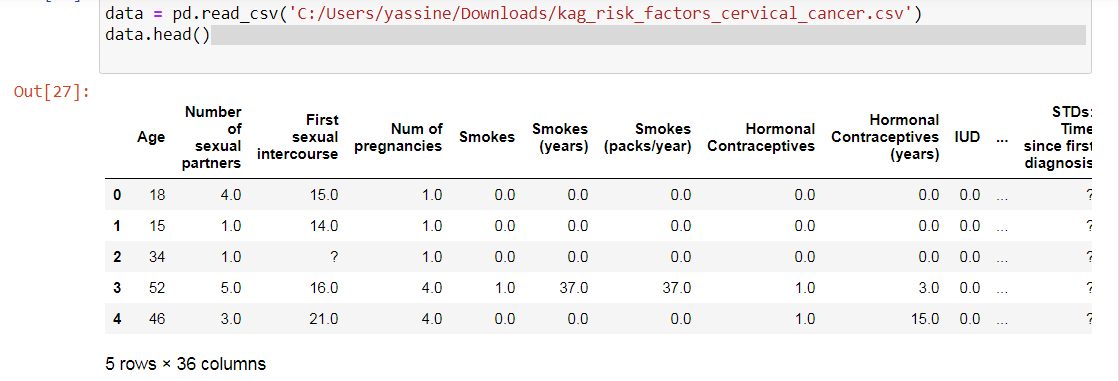


Figure 3: Visualisation des données

La taille de la dataset:

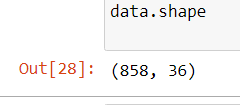


Figure 4: Taille de la dataset

**Biopsy** nous servira de prédicteur pour diagnostiquer le cancer du col de l'utérus. Pour les exemples de ce livre, le résultat de **Biopsy** a été utilisé comme variable target (à prédire). Elle montre si le test du patient est positif ou non.

1. **Résultats expérimentaux :**
2. **Gestion des données manquantes :**

La dataset comprend 3 622 valeurs manquantes sur 27 456 observations, ce qui constitue 13,2% des données. La raison est que certains patients ont décidé de ne pas répondre à quelques questions en raison de problèmes de confidentialité.

Le tableau suivant montre la répartition de ces données manquantes :

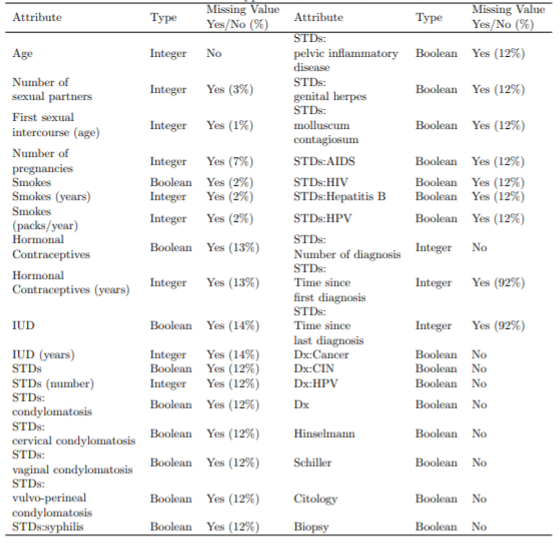


Figure 5: Répartition des valeurs manquantes

Pour régler ce problème :

* Dans le cas où les valeurs manquantes ne dépassent pas 20%, on les a remplacé par le mode.
* Dans le cas où les valeurs manquantes dépassent 20%, on a éliminé la colonne correspondante.

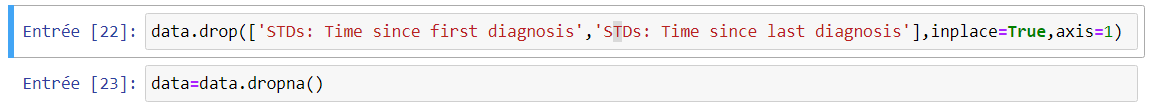


Figure 6: gestion des valeurs manquantes

La dataset devient sous la forme suivante :

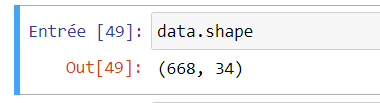


Figure 7: Nouvelle taille de la dataset

1. **Réduction de dimension :**

Après avoir appliqué la réduction de dimension par ACP sur nos données, nous avons choisi de prendre la plus petite dimension qui conserve 95% de la variance :

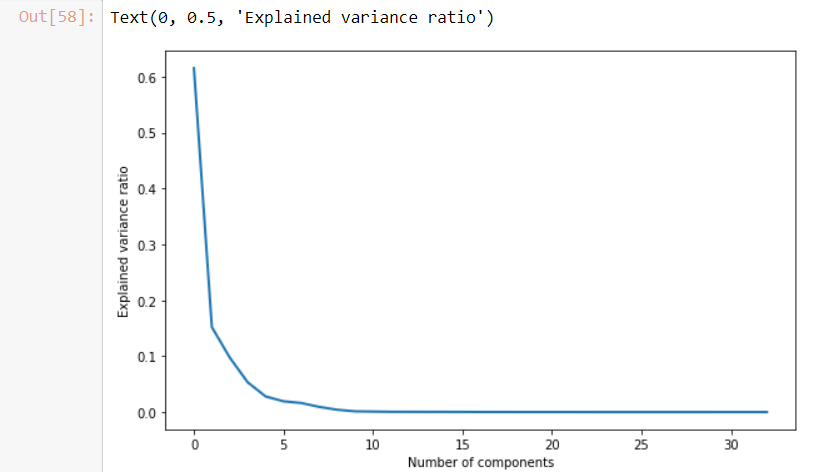


Figure 8: Résultat ACP

On obtient le résultat suivant :

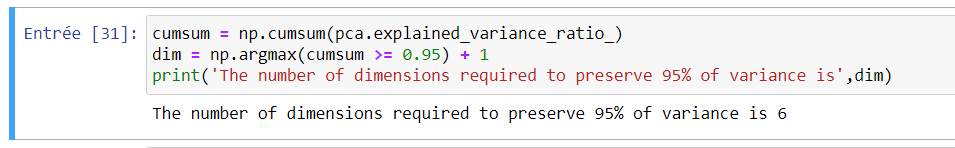


Figure 9: Résultat ACP

Après cette étape, on a réparti nos données en deux parties : une première de test (30%) et une deuxième d’entrainement (70%).

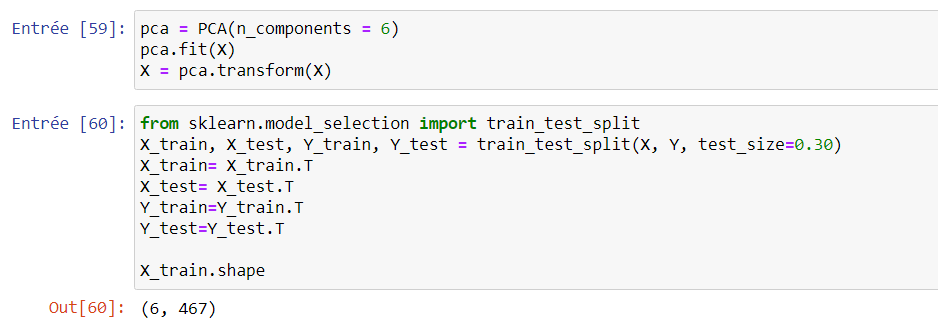


Figure 10: Répartition des données : test et entraînement

1. **Perceptron Multicouche :**

On a utilisé une architecture neuronale multicouche de dimension [6 ;3 ;3 ;1] c'est-à-dire de 4 couches :

* Une couche d’entrée à 6 nœuds
* Deux couches cachées à 3 nœuds chacune
* Une couche de sortie d’un seul nœud

Le taux d’apprentissage est 5% et le nombre d’itérations est 5000 :

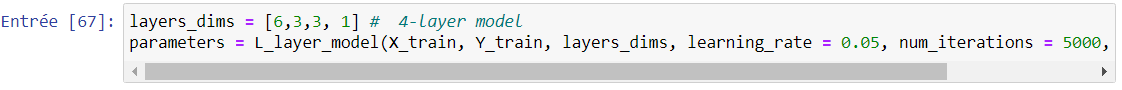


Figure 11: initialisation Perceptron Multicouche

Nous avons obtenu la courbe de coût après itération suivante qui décrit la variation de la fonction de coût après chaque 100 itérations en utilisant un taux d'apprentissage de 0,05 :

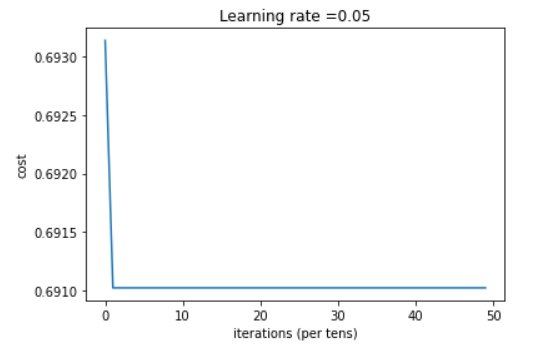


Figure 12: résultat Perceptron Multicouche

Pour les précisions :

rn15.PNG

Figure 13: Précision de l’algorithme

**CONCLUSION :**

Dans ce projet, nous avons parcouru la démarche classique de dataprocessing : Acquisition des données – Prétraitement – Application d’algorithme. Après avoir appliqué la réduction de dimension avec la méthode factorielle ACP, on a appliqué l’algorithme de classification vu en cours : Perceptron Multicouche pour analyser des facteurs de risque du cancer du col utérin. Nous avons traité les données manquantes manuellement. Les précisions de test et d’entrainement ont été satisfaisantes.