Utilisation de la Machine Learning pour le diagnostic des patients

T.I.P.E.

Réalisé par

Hamza ABABOU

Numéro d'inscription au concours : FS018M

2021 - 2022

Plan

I. Introduction

II. Machine Learning : définition et fonctionnement

III. Modèle mathématique

IV. Etude de la corrélation

V. Conclusion

I. Introduction

En 2019 : ~ 1,4 % de bébés morts-nés (Rapport ONU)

Causes :

- **Problémes survenus pendant la grossesse**
- **Mauvais suivi médical**

Solution: **Dignostic de l'état sanitaire du fœtus avec précision**

Outil utilisé : **Algorithme de classification basé sur Machine Learning**

En très bon accord avec le thème :

Santé Prévention

II. Machine Learning : Définition et fonctionnement

Machine Learning : **Domaine scientifique qui consiste à remplir des tâches en se basant sur des motifs récurrents dans des bases de données.**

Algorithmes de classification :

- **Arbre de décision**
- **Forêt aléatoire**
- **Régression logistique**

II. Machine Learning : Définition et fonctionnement

Input:

- ✓ **Données relatives au cas à étudier :**
  Age, Groupe sanguin …

- ✓ **Données relatives à des tests conduits sur le cas à étudier :**
  Fréquence cardiaque, Vitesse de mouvements dans l'utérus, Accélérations brusques, nombre de pics …

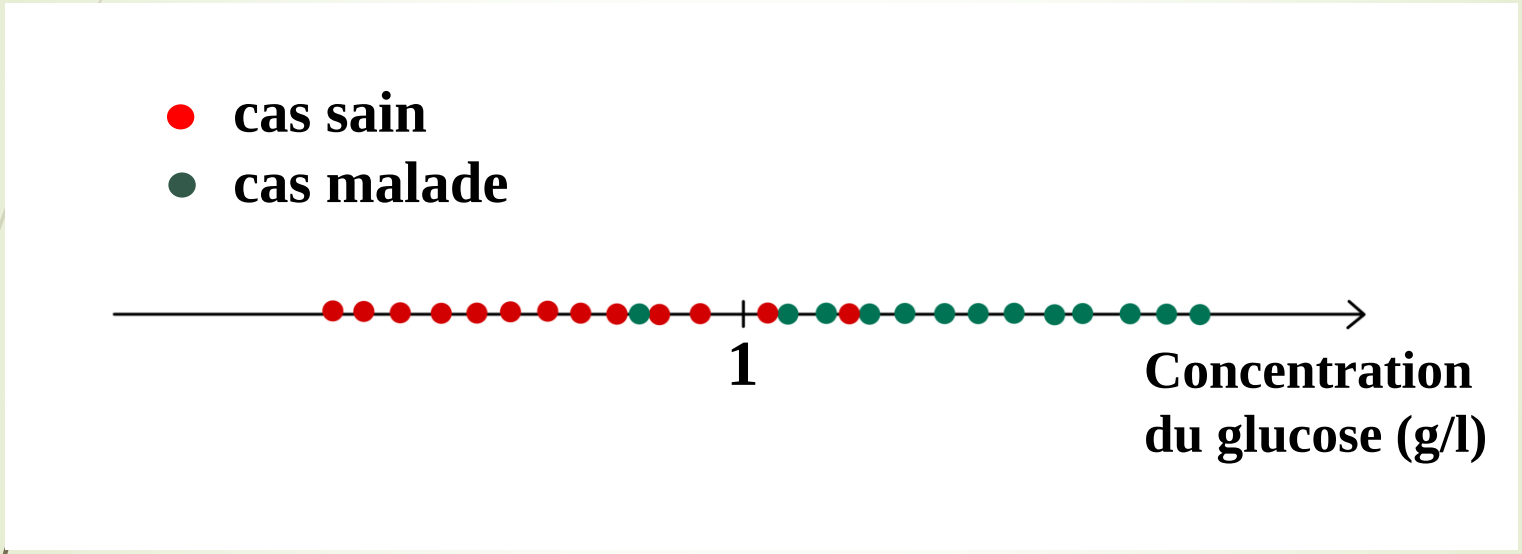
Output :

Classifier le cas étudié :

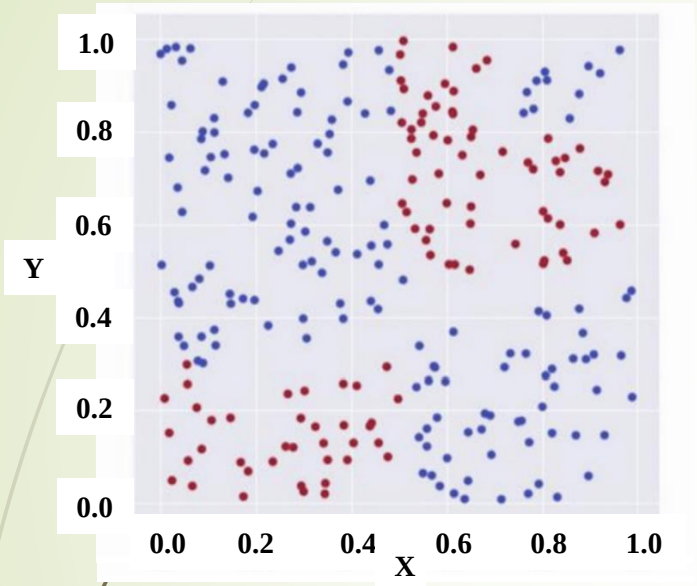
- ✓Fœtus en bonne santé
- ✓Intervention médicale d'urgence
- ✓Cas suspect : Attention particulière

II. Machine Learning : Définition et fonctionnement

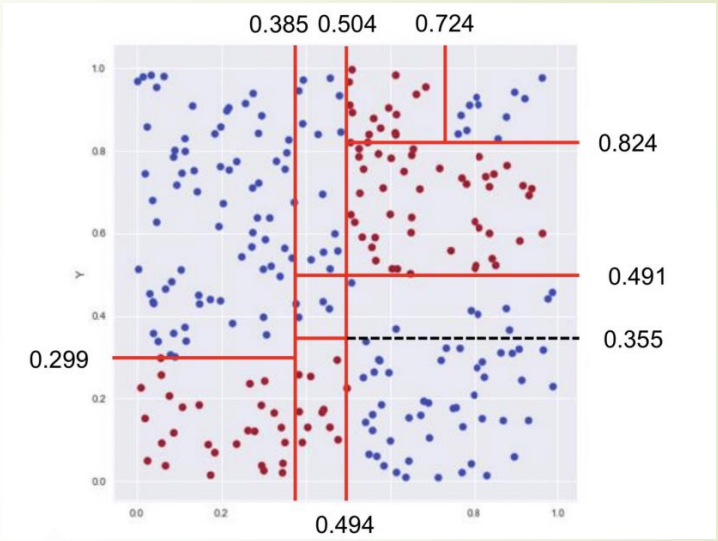
Exemple de classification



II. Machine Learning : Définition et fonctionnement



Distribution de données avec 2 catégories (Rouge et Bleu)



Division en sous-domaines à une seule catégorie

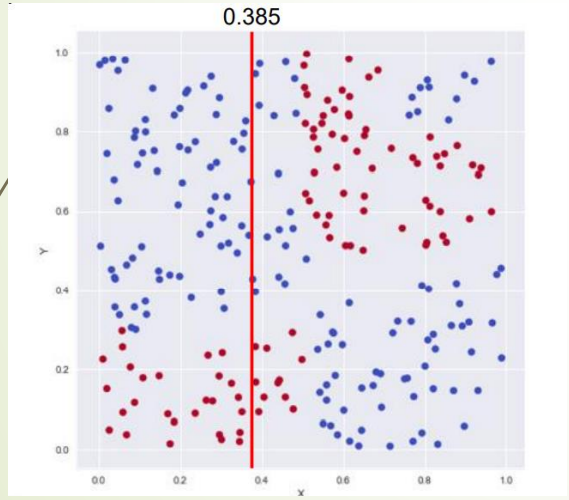
Identification de la catégorie du cas à étudier

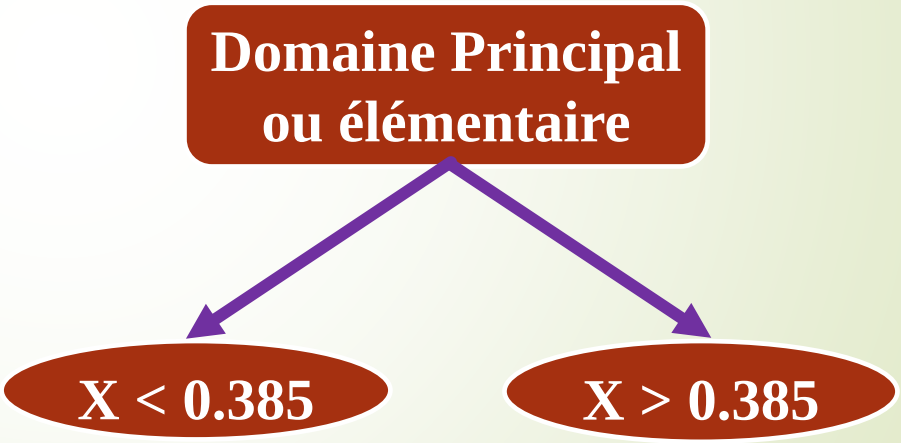
II. Machine Learning : Définition et fonctionnement

Réaliser les divisions de manière adéquate

⇩

Structure de l'arbre





Sous-domaines

III. Modèle mathématique

Détermination de la pureté du domaine

⇩

Fonction Entropie : mesure l'impureté du domaine :

$$\mathbf{E(S) = -\sum_{i=1}^{c} p_i \log_2 (p_i)}$$

S: Domaine en question

c: Nombre de catégories présentes dans le domaine
(dans notre exemple c = 2 : catégories bleue et rouge)

p_i : Proportion des éléments de la catégorie i dans le nœud S

Test d'arrêt : E(S) = 0 (Domaine pur)

III. Modèle mathématique

Entropie relative à une division X **:**

$$E(S,X) = - P(d)E(d) - P(g)E(g)$$

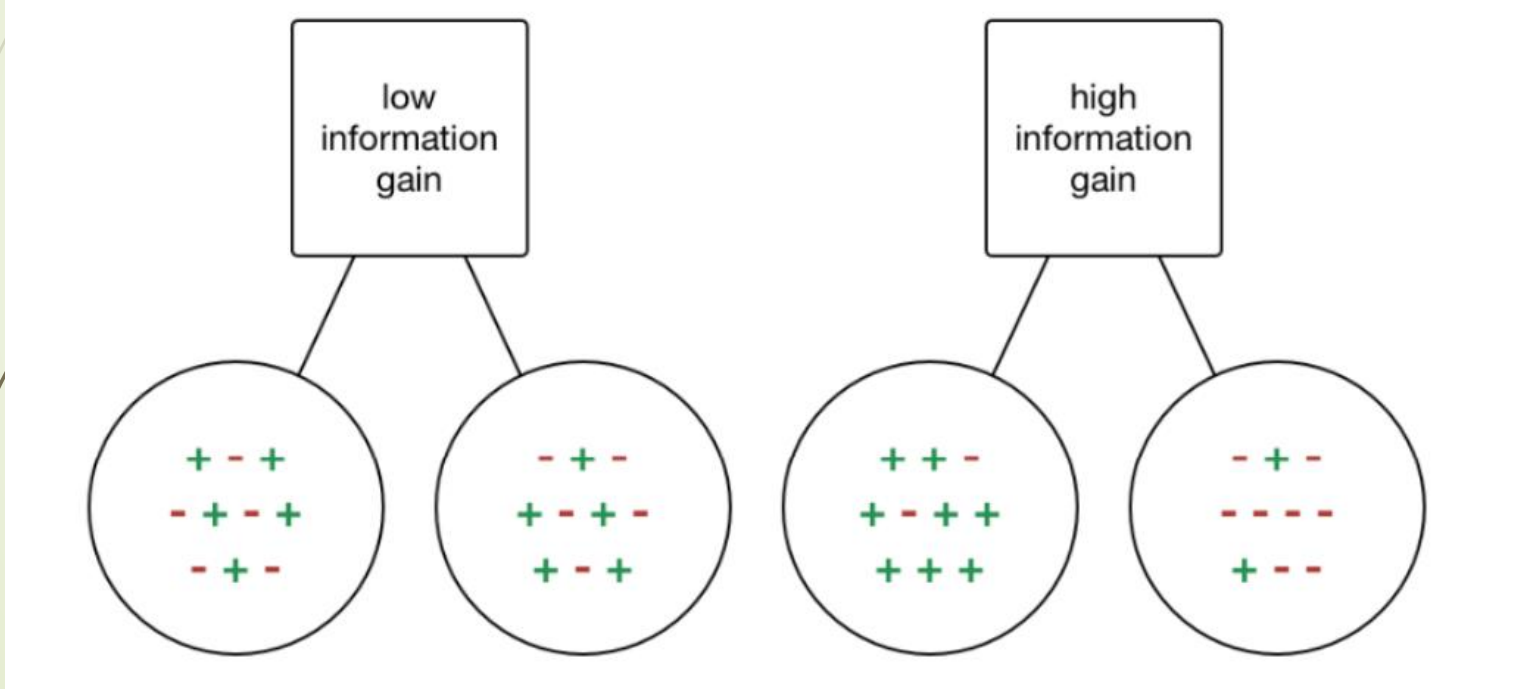
S: le domaine qu'on divise.
d, g : les sous-domaines droit et gauche
P(i) : proportion des éléments du sous-noeud i dans le noeud S

Fonction Information Gain (IG) :

mesure le degré de séparation des éléments des catégories après division

III. Modèle mathématique

Information Gain(T,X) = Entropy(T) - Entropy(T, X)



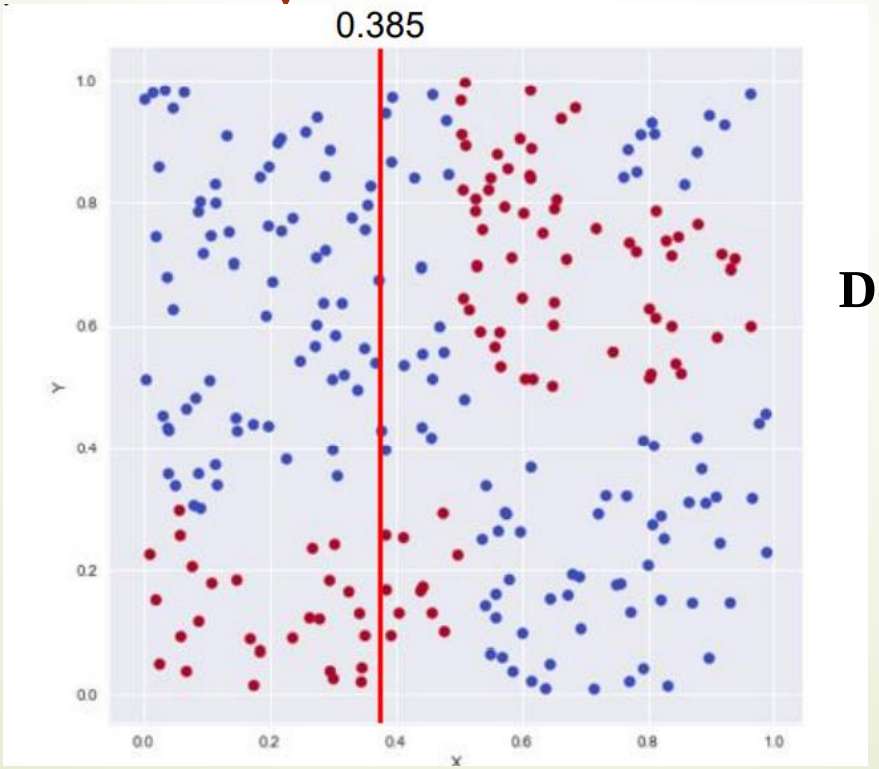
III. Modèle mathématique

Etapes exécutées par l'algorithme pour aboutir au découpage final

Etape 1: **Test d'arrêt : E(S) = 0** → Oui → **Arrêt**

↓ Non

Etape 2:



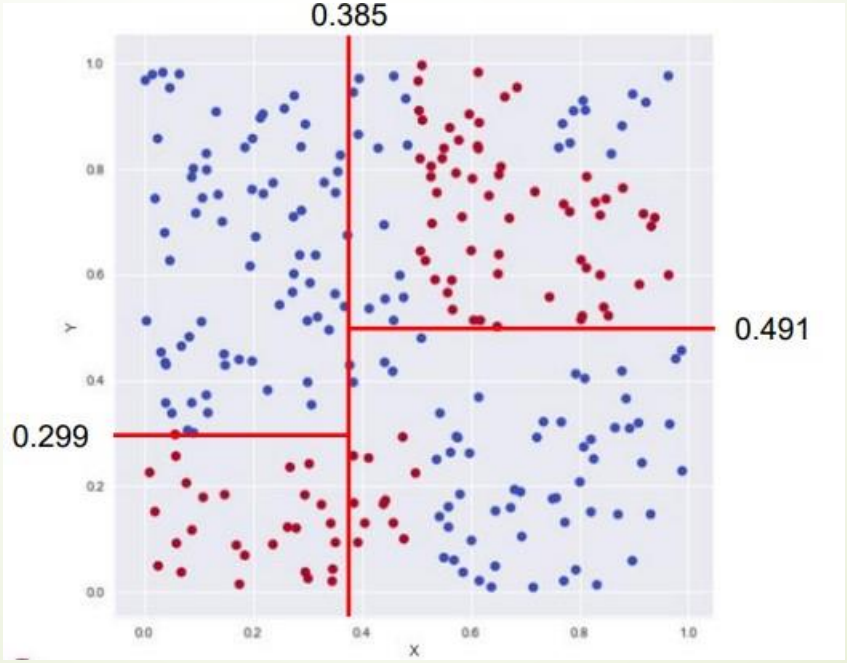
Déterminer la division à IG maximal

III. Modèle mathématique

Etapes exécutées par l'algorithme pour aboutir au découpage final

Etape 3:

Répéter les étapes 1 et 2 pour chaque sous-domaine considéré comme domaine principal

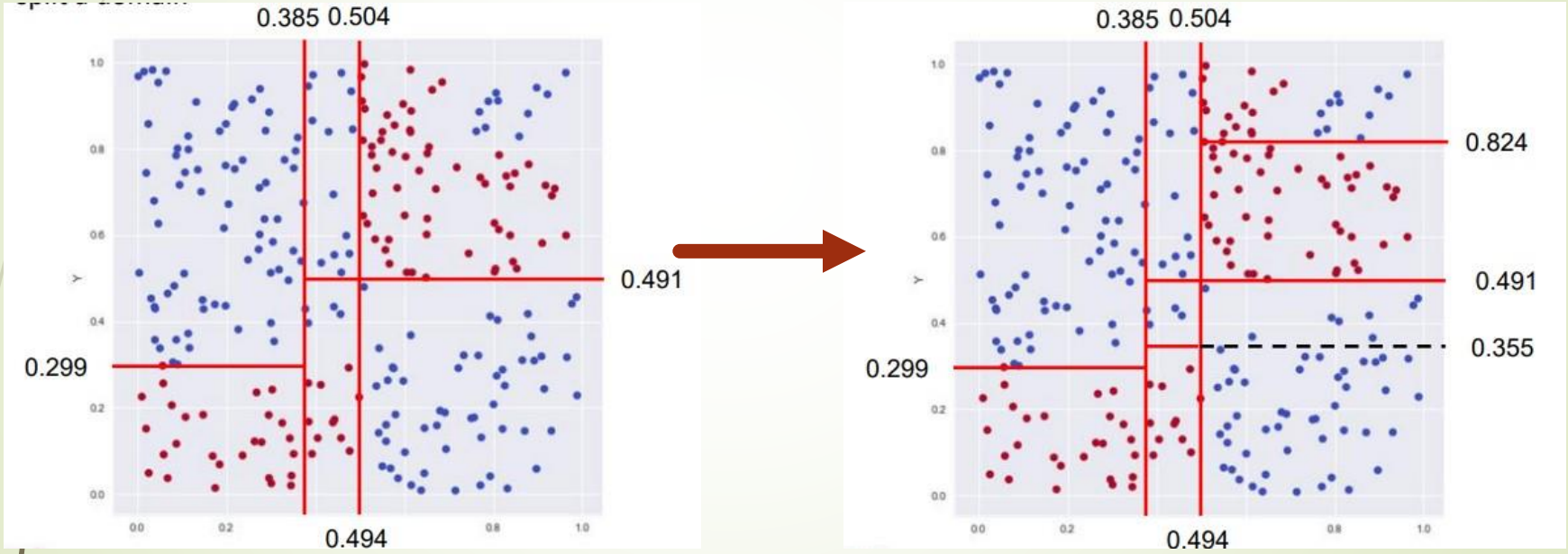


III. Modèle mathématique

Etapes exécutées par l'algorithme pour aboutir au découpage final

Etape 3:

Iteration des étapes 1 et 2



III. Modèle mathématique

Etapes exécutées par l'algorithme pour aboutir au découpage final

Résultat :

Obtenir des domaines contenant des éléments de la même catégorie

Identifier la catégorie (diagnostic) associée au cas examiné

III. Modèle mathématique

Matrice de représentation de la base de données

$$A = (a_{k,i})_{(k,i)\in[\![1,N]\!]\times[\![1,p]\!]}$$

N : nombre de cas traités

p : nombre de paramètres

	1-er paramètre	…	J-ème paramètre	…	p-ième paramètre
1-er cas traité	$a_{1,1}$		$a_{1,j}$		$a_{1,p}$
…					
k-ème cas traité	$a_{k,1}$		$a_{k,j}$		$a_{k,p}$
…					
N-ième cas traité	$a_{k,N}$		$a_{N,j}$		$a_{N,p}$

III. Modèle mathématique

Choix de la division

$(a_{k,i})_{1\leq k\leq N}$ = toutes les valeurs associées au paramètre x_i

- Faire un classement ordonné des $(a_{k,i})_{1\leq k\leq N}$ dans une liste $(L_{k,i})_{1\leq k\leq N}$
- Réaliser la division du domaine considéré en sous-domaines $x_i > L_{k,i}$ et $x_i \leq L_{k,i}$ pour tout $1 \leq k \leq N-1$

  Opération effectuée pour tous les paramètres d'étude x_i $(1 \leq i \leq p)$
- Choisir la division correspondant au IG maximal

IV. Etude de la corrélation

Grand nombre de paramètres ⟹ **Temps de traitement très long**

⇩

Corrélation entre les paramètres

$(\Omega , P(\Omega) , P)$: Espace probabilisé

- **Ω : Ensemble des cas traités, peut être identifié avec $[\![1, N]\!]$**
- **$P(\Omega)$: Ensemble des parties de Ω**
- **P : Probabilité, P: $P(\Omega) \rightarrow [0, 1]$**

$$A \mapsto P(A) = \frac{card(A)}{card(\Omega)} = \frac{card(A)}{N}$$

IV. Etude de la corrélation

- Y_i : Variable aléatoire, $i \in [\![1, P]\!]$

  $\forall\ k \in [\![1, N]\!]$: $Y_i(k) = a_{k,i}$

- $cov(A,B) = E(AB) - E(A) \,.\, E(B)$

  A , B : v-a discrétes définies d'un espace fini

- $cor(A,B) = \frac{cov(A,B)}{\sqrt{V(A)V(B)}}$

  V(A) : variance de la v-a A

IV. Etude de la corrélation

$\mathbf{cor(A,B) \in [-1,1]}$ **et** $\mathbf{cor(A,B) = \pm 1 \Leftrightarrow "\exists\ (\lambda, \beta) \in \mathbb{R}^2 tq\ B = \lambda.A + \beta"}$

$\forall\ \lambda \in R,$

$$V(\lambda A - B) = E((\lambda A - Y)^2) - (E(\lambda A - B))^2$$

$$= E(\lambda^2 A^2 - 2\lambda AB + Y^2) - (\lambda E(A) - E(B))^2$$

$$= \lambda^2(E(A^2) - E(B)^2) - 2\lambda(E(AB) - E(A)\,E(B)) + E(B^2) - E(B)^2$$

$$= \lambda^2 V(A) - 2\lambda\ Cov(A, B) + V(B).$$

$$\Delta = 4Cov(A, B)^2 - 4V(A)\ V(B) \leq 0$$

IV. Etude de la corrélation

$$Cor(A,B) = \pm 1 \Longrightarrow \Delta = 0$$

$$\text{et } \exists\, \lambda \in R \text{ tq } V(\lambda A - B) = 0$$

$$\Longrightarrow \exists\, \beta \in R \text{ tq } P(B = \lambda A + \beta) = 1 \Longleftrightarrow B = \lambda A + \beta$$

Pour deux paramètres x_i et x_j :

$$Cor\left(Y_i, Y_j\right) = \frac{cov\left(Y_i, Y_j\right)}{\sqrt{V(Y_i)\, V(Y_j)}} = \frac{E\left(Y_i, Y_j\right) - E(Y_i)\, E(Y_j)}{\sqrt{\left[E\left(Y_i^2\right) - E(Y_i)^2\right] . \left[E\left(Y_j^2\right) - E(Y_j)^2\right]}}$$

$$= \frac{N^{-1} \sum_{k=1}^{N} a_{k,i}\, a_{k,j} - N^{-1} \sum_{k=1}^{N} a_{k,i} \,.\, N^{-1} \sum_{k=1}^{N} a_{k,j}}{\sqrt{\left[N^{-1} \sum_{k=1}^{N} {a_{k,i}}^2 - \left(N^{-1} \sum_{k=1}^{N} a_{k,i}\right)^2\right] . \left[N^{-1} \sum_{k=1}^{N} {a_{k,j}}^2 - \left(N^{-1} \sum_{k=1}^{N} a_{k,j}\right)^2\right]}}$$

IV. Etude de la corrélation

$$\text{cor}(Y_i, Y_j) \approx \pm 1 \Rightarrow \forall\, k \in [\![1,N]\!] \quad a_{k,i} \approx \lambda\, a_{k,j} + \beta$$

Pour chaque division suivant le paramètre x_i, il y a une division équivalente suivant le parameter x_j

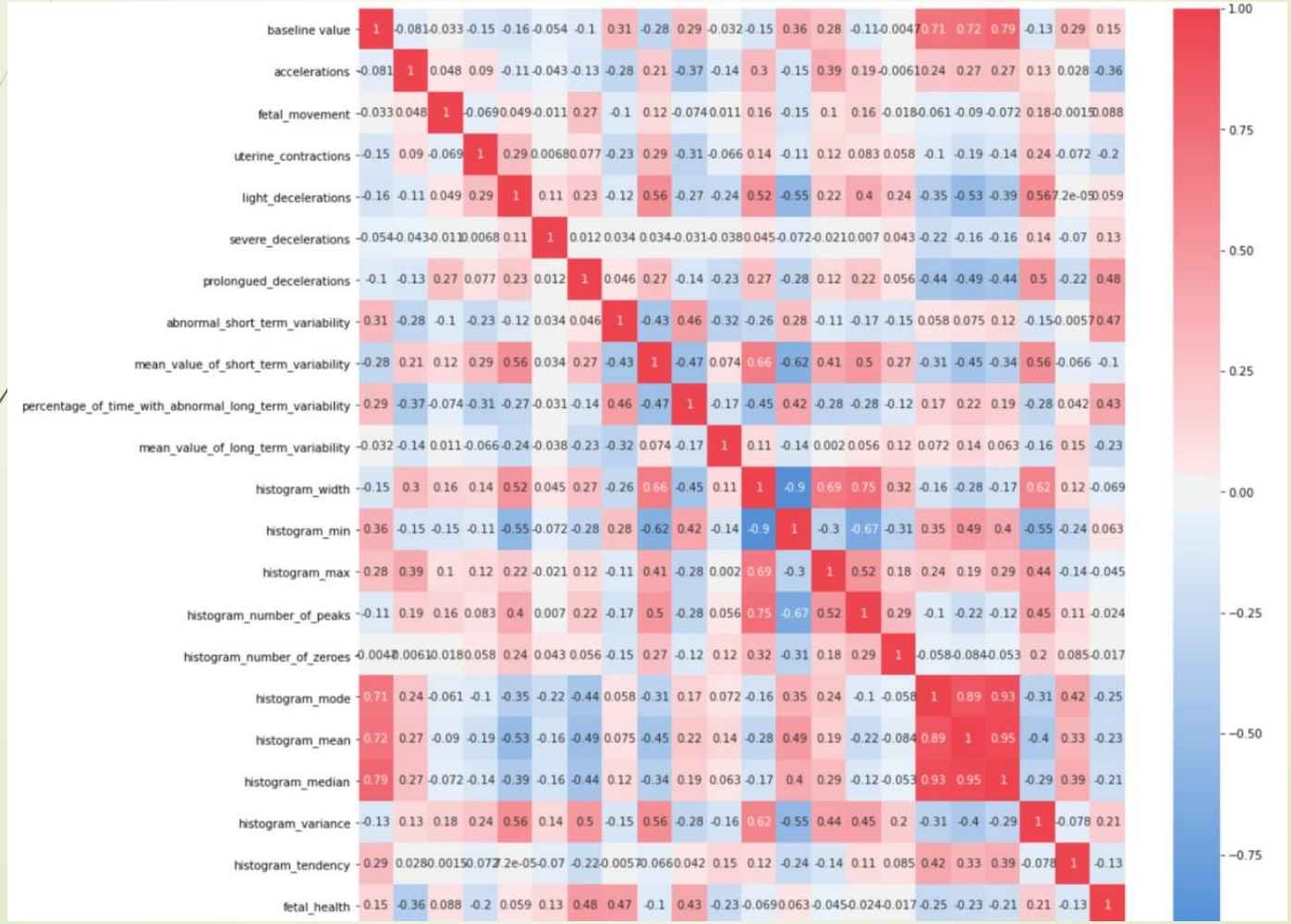
IV. Etude de la corrélation

Portion de la base de données étudiée pour le diagnostic des foetus

	baseline value	accelerations	fetal_movement	uterine_contractions	light_decelerations	severe_decelerations
1539	136.0	0.007	0.000	0.007	0.001	0.0
1826	140.0	0.000	0.001	0.003	0.007	0.0
1960	133.0	0.001	0.000	0.009	0.004	0.0
1059	127.0	0.000	0.000	0.009	0.011	0.0
1318	125.0	0.002	0.014	0.006	0.008	0.0
1758	120.0	0.002	0.000	0.001	0.009	0.0
1617	142.0	0.004	0.041	0.003	0.004	0.0
236	121.0	0.003	0.008	0.000	0.000	0.0
228	127.0	0.006	0.003	0.000	0.000	0.0
192	144.0	0.002	0.000	0.003	0.000	0.0

IV. Etude de la corrélation

Matrice de corrélation : $C=(cor(Y_i, Y_j))_{1\leq i,j\leq p}$



IV. Etude de la corrélation

Tableau comparatif entre algorithmes

Algorithme	Précision
Arbre de décision	91,60 %
Forêt aléatoire	94,00 %
Régression logistique	89,70 %

V. Conclusion

- **Présentation du modèle mathématique adopté pour développer un algorithme de classification des données**
- **Rôle important du Machine Learning à établir un diagnostic précis de l'état de santé du fœtus à partir de l'analyse de ses données en relation avec une base de données établie.**
- **Généralisation du modèle mathématique pour d'autres types de maladies**

Merci pour votre attention

Matrice de corrélation (agrandie)