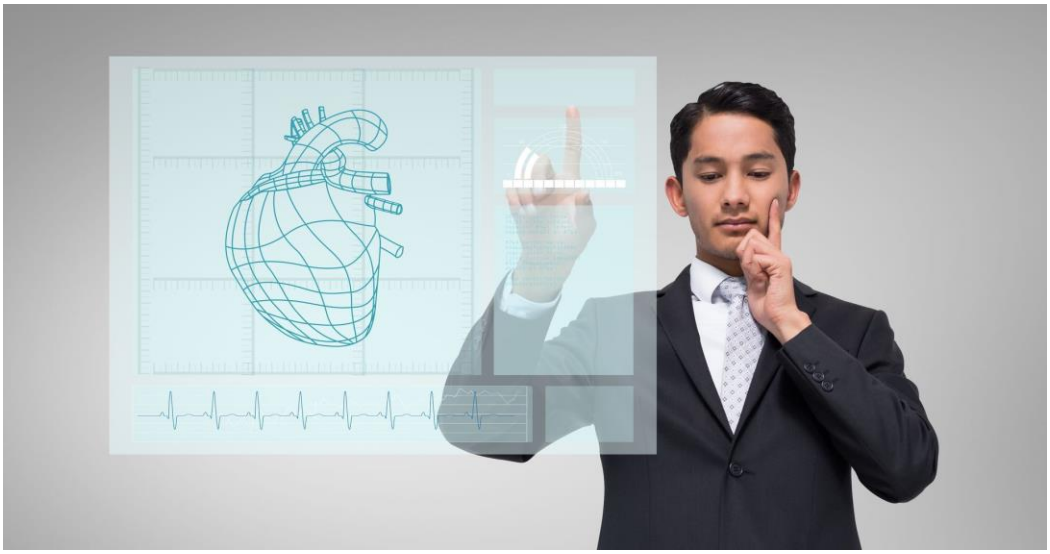
Université Ibn Tofaïl

Master Intelligence Artificielle et Réalité Virtuelle

Analyse en Composantes Principales

Réalisé par :

Ait Ouamer Ouafa



Année universitaire : 2023-2024

Table des Matières

Introduction

L'Analyse en Composantes Principales (ACP) s'impose comme un outil essentiel pour comprendre des données complexes. Elle permet de simplifier ces données tout en préservant les informations importantes. Ce document explore l'ACP, ses étapes, ses principes, et son utilité à travers l'étude des maladies cardiaques.

Comprendre les données sur les maladies cardiaques revêt une importance cruciale en raison de l'impact considérable de ces affections sur la santé mondiale. Les maladies cardiaques, telles que la maladie coronarienne et l'insuffisance cardiaque, demeurent des causes majeures de problèmes de santé et de décès. Analyser ces données n'est pas seulement un exercice académique ; c'est une étape essentielle pour mieux comprendre les facteurs influençant la santé cardiaque, identifier les tendances émergentes, et développer des stratégies de prévention et de traitement plus efficaces.

Dans ce contexte, l'ACP se révèle être un outil précieux. En simplifiant la compréhension de données multidimensionnelles relatives aux maladies cardiaques, elle offre une vision claire des relations complexes entre différentes variables. Cela permet aux chercheurs, aux médecins, et aux décideurs de prendre des décisions éclairées basées sur des données fiables. De plus, la visualisation simplifiée des données cardiaques facilite la communication des résultats aux professionnels de la santé et au grand public, favorisant ainsi une sensibilisation accrue et des actions préventives.

Nous détaillerons chaque étape de l'ACP, de la normalisation des données à la représentation graphique. En exposant des méthodes clés comme le calcul de la matrice de covariance, nous montrerons comment l'ACP simplifie la compréhension de données complexes.

À travers l'analyse de données sur les maladies cardiaques, nous illustrerons comment l'ACP simplifie la visualisation sans perdre d'information. L'objectif est de simplifier la compréhension de données complexes, créer des outils d'analyse efficaces, et réduire l'information sans la perdre.

Nous explorerons également des méthodes de calcul des valeurs propres, essentielles pour définir les composantes principales. Leur importance sera soulignée dans l'application de l'ACP à des données réelles.

Ensuite, nous aborderons l'application de différents algorithmes d'apprentissage machine pour classer les maladies cardiaques. Cette comparaison vise à évaluer la performance de l'ACP.

En résumé, ce document plonge dans l'univers de l'ACP, mettant en lumière son application dans le domaine médical via l'étude des maladies cardiaques. Nous conclurons en soulignant l'importance de choisir judicieusement les algorithmes d'apprentissage machine en fonction des contextes spécifiques. L'ACP demeure une méthodologie cruciale pour explorer des données complexes.

Principe et méthodes associées à l'ACP

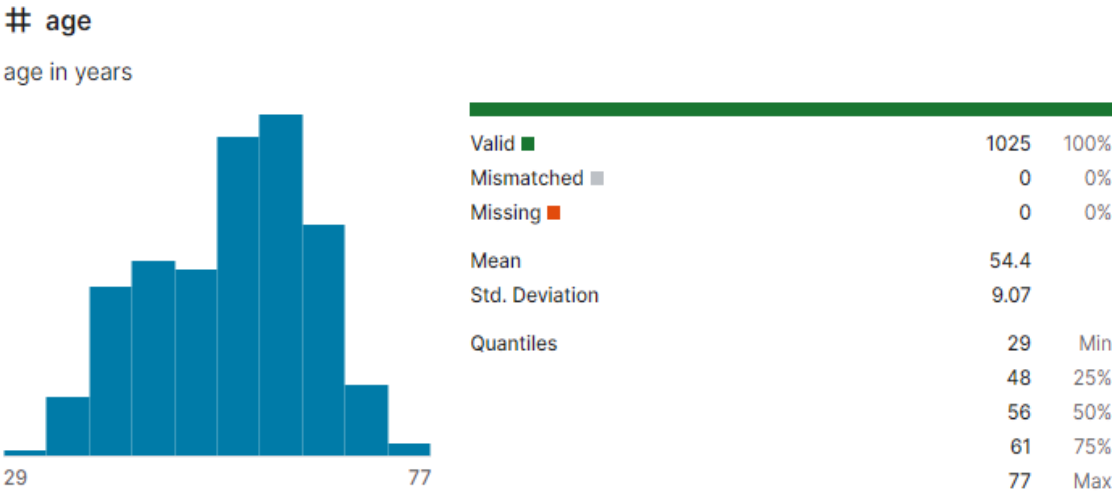
1. Introduction :

L'Analyse en Composantes Principales (ACP), est une méthode basée sur des statistiques descriptives multidimensionnelles permettant de traiter simultanément un grand nombre de variables quantitatives. Elle s'applique lorsque de nombreux individus (n individus) sont mesurés par rapport à un grand nombre de variables numériques. Ces variables sont généralement corrélées entre elles. L'objectif de l'ACP est de trouver un petit nombre de facteurs qui résumerait au mieux les données sous-jacentes. Cela se traduit par des représentations graphiques des données, tant pour les individus que pour les variables, en fonction de ces facteurs qui sont représentés sous forme d'axes. Ces représentations graphiques prennent la forme de nuages de points.

2. Description de la base de donnée :

Heart Disease est un ensemble de données qui date de 1988 et se compose de quatre bases de données : Cleveland, Hongrie, Suisse et Long Beach V. Il contient 76 attributs, y compris l'attribut prédit, mais toutes les expériences publiées font référence à l'utilisation d'un sous-ensemble de 14 d'entre eux. Le champ « cible » fait référence à la présence d'une maladie cardiaque chez le patient. Il s'agit d'un nombre entier valant 0 = pas de maladie et 1 = maladie.

- Les colonnes de notre base de donnée:



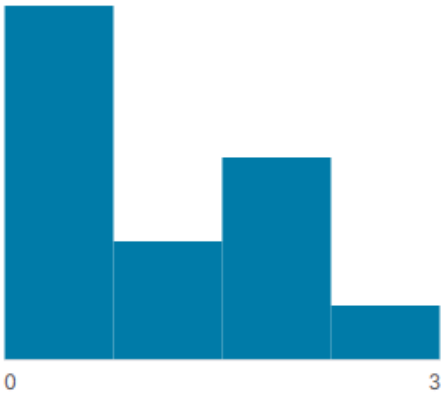
sex

(1 = male; 0 = female)

Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.7	
Std. Deviation	0.46	
Quantiles	0	Min
	0	25%
	1	50%
	1	75%
	1	Max

cp

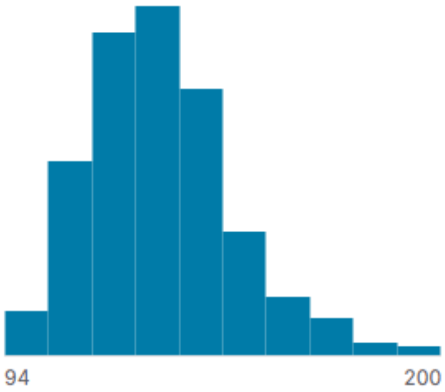
chest pain type



Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.94	
Std. Deviation	1.03	
Quantiles	0	Min
	0	25%
	1	50%
	2	75%
	3	Max

trestbps

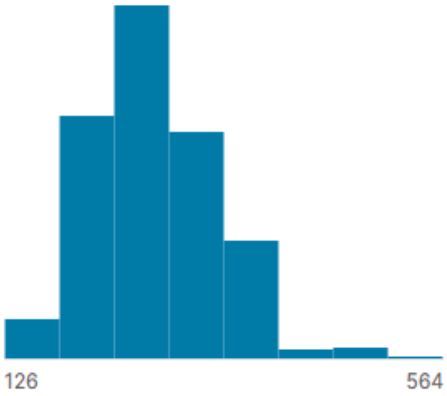
resting blood pressure (in mm Hg on admission to the hospital)



Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	132	
Std. Deviation	17.5	
Quantiles	94	Min
	120	25%
	130	50%
	140	75%
	200	Max

chol

serum cholestoral in mg/dl



Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	246	
Std. Deviation	51.6	
Quantiles	126	Min
	211	25%
	240	50%
	275	75%
	564	Max

fbs

(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.15	
Std. Deviation	0.36	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

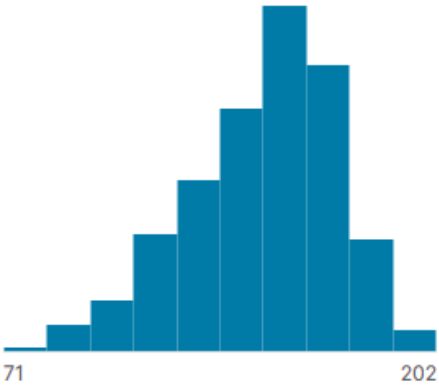
restecg

resting electrocardiographic results

Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.53	
Std. Deviation	0.53	
Quantiles	0	Min
	0	25%
	1	50%
	1	75%
	2	Max

thalach

maximum heart rate achieved



Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	149	
Std. Deviation	23	
Quantiles	71	Min
	132	25%
	152	50%
	166	75%
	202	Max

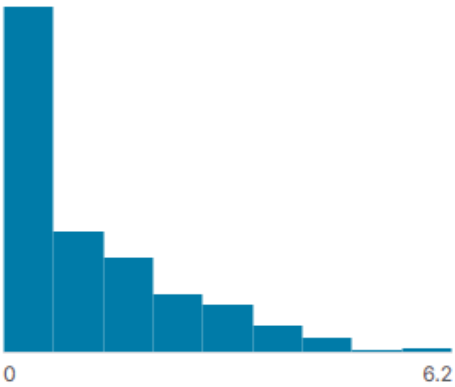
exang

exercise induced angina (1 = yes; 0 = no)

Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.34	
Std. Deviation	0.47	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

oldpeak

ST depression induced by exercise relative to rest



Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.07	
Std. Deviation	1.17	
Quantiles	0	Min
	0	25%
	0.8	50%
	1.8	75%
	6.2	Max

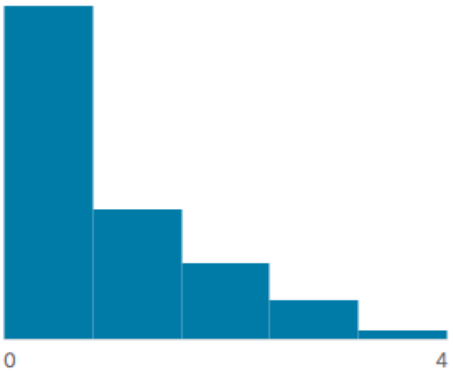
slope

the slope of the peak exercise ST segment

Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.39	
Std. Deviation	0.62	
Quantiles	0	Min
	1	25%
	1	50%
	2	75%
	2	Max

ca

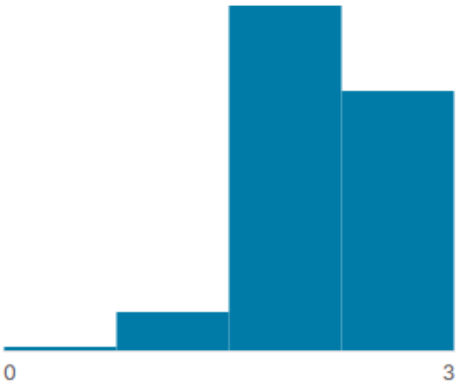
number of major vessels (0-3) colored by flourosopy



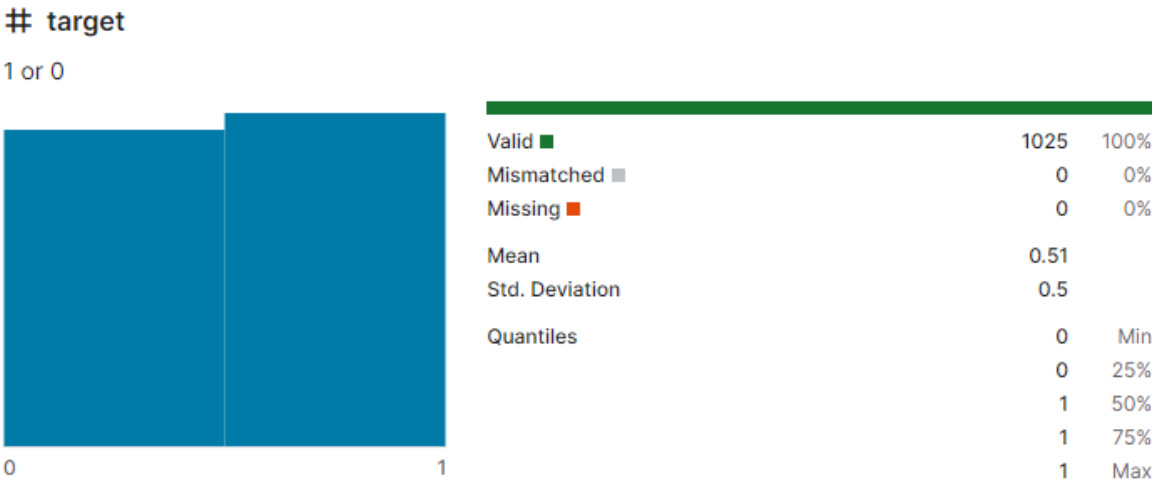
Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.75	
Std. Deviation	1.03	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	4	Max

thal

1 = normal; 2 = fixed defect; 3 = reversable defect



Valid	1025	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.32	
Std. Deviation	0.62	
Quantiles	0	Min
	2	25%
	2	50%
	3	75%
	3	Max



3. Objectif de l'ACP :

En combinant une approche géométrique, qui représente les liens entre les variables et les individus dans un espace rectangulaire, et une approche statistique, qui cherche des axes indépendants pour décrire la variance, l'Analyse en Composantes Principales (ACP) poursuit trois principaux objectifs :

- Comprendre la structure d'un ensemble de variables,
- Créer des outils pour analyser des éléments qui ne peuvent pas être directement mesurés,
- Réduire les informations provenant d'un grand nombre de variables en un ensemble plus restreint tout en minimisant la perte d'information.

4. Principe de l'ACP :

L'Analyse en Composantes Principales (ACP) consiste à substituer une série de variables par de nouvelles variables qui présentent une variance maximale, ne sont pas corrélées les unes aux autres, et sont des combinaisons linéaires des variables originales. Ces nouvelles variables, appelées composantes principales, définissent des plans factoriels qui servent de base à une représentation graphique plane des variables initiales. L'interprétation des résultats se limite généralement aux deux premiers plans factoriels, à condition que ceux-ci expliquent la majeure partie de la variance présente dans le nuage des variables d'origine.

5. Méthodes associées à l'ACP :

5.1 Données à analyser :

L'ACP est appliquée sur p variables quantitatives notées X1…..Xj……XP observées sur n individus notés 1….i….n .L'observation de la variable Xj observées sur l'individu i est xj.i

Donc l'ensemble des informations se représente de la manière suivante :

	X_1	…… ….	X_j	… …. .	X_p
1	$x_{1.1}$	…… ….	$x_{j.1}$	… …..	$x_{p.1}$
…… .	……		… …..		… …..
i	$x_{1.i}$	…… ….	$x_{j.i}$	… …	$x_{p.i}$
…… .	……		… …		… …
n	$x_{1.n}$	…… ….	$x_{j.n}$	… …	$x_{p.n}$

On obtient donc en ligne, les observations de chaque individu, et en colonne, les observations de chaque variable.

5.2 Standardisation des données :

Avant de commencer l'ACP, il est courant de standardiser les données pour que toutes les variables aient la même échelle. Cela évite que les variables ayant des unités de mesure différentes ne dominent l'analyse. Pour chaque variable Xi, on soustrait sa moyenne (μ) et on divise par son écart-type (σ) pour obtenir une nouvelle variable standardisée Zi. La formule est la suivante :

Standardisation de Xi : $$Z_i = (X_i - \mu) / \sigma$$

5.3 Calcul de la matrice de covariance :

La matrice de covariance S est essentielle pour l'ACP. Elle indique les covariances entre chaque paire de variables et permet de mesurer les relations entre les variables. Elle est calculée à partir des données standardisées. Pour un ensemble de p variables, la matrice de

covariance S est définie comme :

$$S = (1 / (n - 1)) * \Sigma (Z_i * Z_j)$$

Où n est le nombre d'individus, Σ représente la somme sur l'ensemble des individus, et Z_i et Z_j sont les variables standardisées.

Nous divisons par n−1 pour la même raison que nous le faisons lors du calcul d'écarts-types d'échantillons - cela nous donne une meilleure estimation de l'équivalent population.

5.4 Calcul des vecteurs propres et valeurs propres :

On résout l'équation caractéristique de la matrice de covariance S pour obtenir les vecteurs propres (composantes principales) v et les valeurs propres λ. Les vecteurs propres indiquent les directions dans lesquelles la variance des données est maximale, tandis que les valeurs propres indiquent l'importance de chaque composante. Plus la valeur propre est élevée, plus la composante est importante.

L'équation caractéristique est la suivante :

$$S * v = \lambda * v$$

5.5 Sélection des composantes principales :

Les vecteurs propres v sont ordonnés en fonction des valeurs propres λ. On choisit les k premières composantes principales qui expliquent la majorité de la variance des données. Généralement, on choisit un nombre de composantes principales qui cumulent une grande proportion de la variance totale, par exemple, 95% de la variance.

5.6 Calcul des scores des individus :

Les scores des individus y sont obtenus en projetant les données standardisées Z sur les k premières composantes principales. Ces scores représentent la position de chaque individu dans l'espace des composantes principales. Les scores des individus permettent de représenter chaque individu dans le nouvel espace des composantes principales.

$$y_i = [v_1 * Z_i, v_2 * Zi, ..., v_k * Z_i]$$

5.7 Interprétation des composantes principales :

Pour interpréter les composantes principales, on calcule les charges c_i des variables avec chaque composante. Les charges indiquent les corrélations entre les variables et les composantes :

$$c_i = \Sigma (Z_j * v_i) / (\sqrt{(\lambda_i * \sigma_i \text{^} 2)})$$

Une charge élevée pour une variable dans une composante signifie que cette variable est fortement corrélée avec cette composante. Les charges permettent de donner un sens aux composantes principales.

5.8 Représentation graphique :

Les individus peuvent être représentés graphiquement dans l'espace des composantes principales en utilisant les scores y. Les graphiques en deux ou trois dimensions montrent comment les individus sont répartis par rapport aux composantes principales, permettant ainsi de visualiser les relations entre eux.

6. Les méthodes utilisées pour calculer les valeurs propres :

6.1 Méthode de Jacobi :

La méthode de Jacobi est une méthode itérative de résolution de système linéaire de la forme

$$Ax = b$$

Pour cela, on utilise $x^{(k)}$ une suite qui converge vers un point fixe x, solution du système d'équations linéaires.

On cherche à construire l'algorithme pour $x^{(0)}$ donné, la suite $x^{(k+1)} = F(x^{(k)})$ avec $k \in \mathbf{N}$ $x^{(k)}$ une suite qui converge vers un point fixe x , solution du système d'équations linéaires.

$A = M - N$ où M est une matrice inversible.

$$Ax = b \Leftrightarrow Mx = Nx + b \Leftrightarrow \begin{array}{rcl} x & = & M^{-1}Nx + M^{-1}b \\ & = & F(x) \end{array}$$

où F est une fonction affine.

On décompose la matrice A de la façon suivante $A = D - E - F$ avec:

- D la diagonale
- $-E$ la partie en dessous de la diagonale
- $-F$ la partie au-dessus.

Dans la méthode de Jacobi, on choisit: $M = D$ et $N = E + F$

$$x^{(k+1)} = D^{-1}(E+F)x^{(k)} + D^{-1}b$$

avec pour la ligne *i* de $D^{-1}(E+F)$: $-(\frac{a_{i,1}}{a_{i,i}}, \cdots, \frac{a_{i,i-1}}{a_{i,i}}, 0, \frac{a_{i,i+1}}{a_{i,i}}, \cdots, \frac{a_{i,n}}{a_{i,i}})$

on a alors:

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} + \frac{b_i}{a_{ii}}$$

6.2 Méthode de la Puissance Itérée :

Principe:

La méthode de la puissance itérée est utilisée pour calculer la plus grande valeur propre et le vecteur propre associé d'une matrice A.

Pour appliquer cette méthode la matrice A doit répondre à certaines contraintes :

- A doit être symétrique:

  La symétrie de la matrice A est une condition essentielle pour appliquer la décomposition spectrale. Une matrice symétrique est une matrice carrée pour laquelle les éléments sont symétriques par rapport à sa diagonale principale (c'est-à-dire A[i][j] = A[j][i] pour tout i et j)

- A doit être positive => A doit être diagonalisable

- A doit être définie positive:

Une matrice définie positive est une matrice symétrique A telle que tous ses vecteurs propres aient des valeurs propres strictement positives. Cette condition garantit que la matrice A est non singulière (c'est-à-dire que son déterminant est non nul) et que sa racine carrée est bien définie. Les matrices définies positives sont couramment utilisées en optimisation, analyse numérique et d'autres domaines.

On prendra la matrice A de dimension A(n*n). Et on suppose qu'elle répond à toutes

ces conditions et que ses valeurs propres sont toutes différentes.

Ainsi A possède n valeurs propres : λ1, …, λn telles que $|\lambda 1| > |\lambda 2| \geq \ldots \geq |\lambda n|$.

Soient U1 …, Un. les vecteurs propres associés.

On pose x0 un vecteur de IR^n de norme égale a 1 (ou quelconque) que l'on décompose de la façon suivante :

$$x_0 = \sum_{i=1}^{n} x_i . u_i$$

On calcul la suite $x_{k+1} = A.x_k$ de la façon suivante :

$$x_k = A^k . x_0$$

$$x_k = \sum_{i=1}^{n} \lambda_1^k . a_i . u_i$$

En mettant λ_n^k en facteur on obtient : $x_k = \lambda_n^k . \left(\sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_n} \right)^k . a_i . u_i \right)$

Or $\left(\frac{\lambda_i}{\lambda_n} \right)^k$ tend vers 0 donc le terme prépondérant est : $\lambda_n^k . a_n . u_n$. Ainsi au bout de quelques itérations on obtient la plus grande valeur propre: $|\lambda_n| \approx \frac{||x_{k+1}||}{||x_k||}$

Et le vecteur propre associé : $u_n \approx x_n$

Si on se retrouve dans le cas où $|\lambda_n|$ est grand, il est nécessaire de normaliser la suite: $x_k = \frac{x_k}{||x_k||}$ (Ou alors en posant un nouveau vecteur). Le principe reste le même par la suite.

6.3 Méthode de QR :

La méthode QR est l'outil le plus utilisé pour calculer les valeurs propres d'une matrice quelconque (pas forcément symétrique). Pour les matrices symétriques, cette méthode est aussi efficace que la méthode de Jacobi.

- La décomposition QR est un produit d'une matrice unitaire (i.e Q*Q = In) et d'une matrice triangulaire supérieure dont les coefficients diagonaux sont réels et strictement positifs

- La décomposition LU est un produit d'une matrice inférieure avec des 1 sur la diagonale et d'une matrice triangulaire supérieure.

- L'application $f : (\mathrm{Un}(C) \times \mathbb{T}_n(C)) \to \mathrm{GL}_n(C), \quad (Q, R) \mapsto QR$

est un homéomorphisme.

Pour tout A ∈ GLn(C), on va faire converger une suite de matrices vers une matrice triangulaire ayant sur sa diagonale les valeurs propres de A.

Soient A ∈ GLn(C) et λ1, . . . , λn ∈ C vérifiant

|λ1| > |λ2| > · · · > |λn| > 0

tels qu'il existe une matrice inversible P d'inverse admettant une décomposition LU

$$P^{-1}AP = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

et que l'on ait

Alors la suite définie par

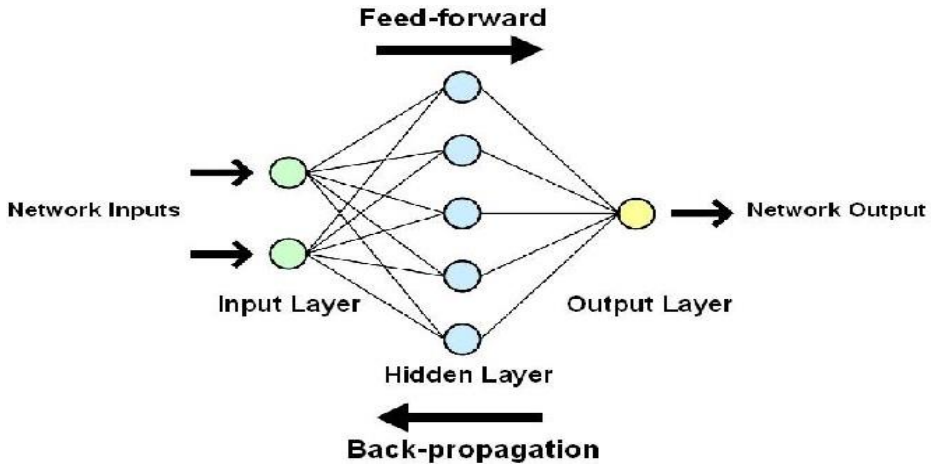
$A_1 = A, \quad A_{k+1} = R_k Q_k$ où $Q_k R_k$ est la décomposition QR de A_k

converge vers une matrice triangulaire supérieure dont la diagonale est $diag(\lambda_1, \ldots, \lambda_n)$

7. Les algorithmes appliqués :

7.1 ANN :

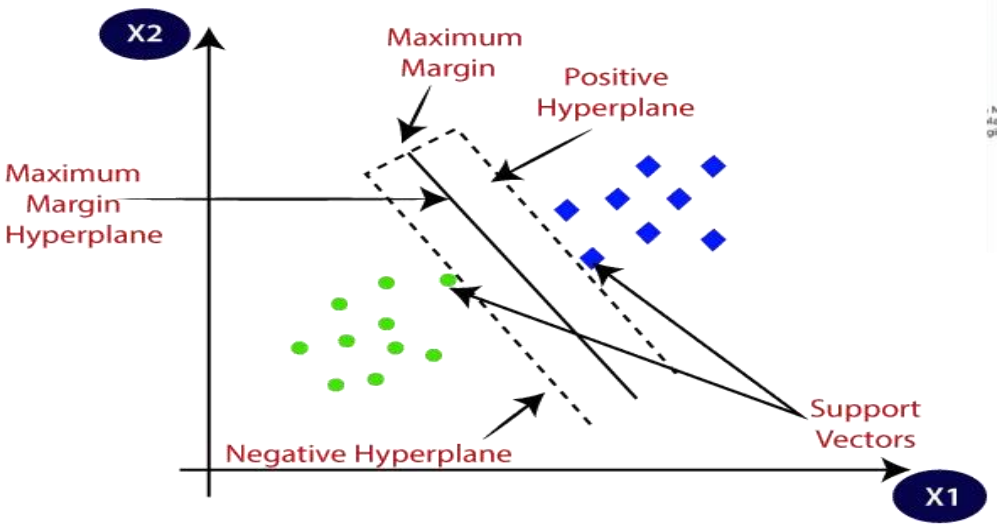
Un réseau neuronal artificiel (ANN) est un modèle informatique inspiré de la structure neuronale du cerveau humain. Il est constitué de nœuds interconnectés (neurones) organisés en couches. Les informations circulent via ces nœuds et le réseau ajuste les forces de connexion (poids) pendant la formation pour apprendre des données, ce qui lui permet de reconnaître des modèles, de faire des prédictions et de résoudre diverses tâches d'apprentissage automatique et d'intelligence artificielle.



7.2 SVM :

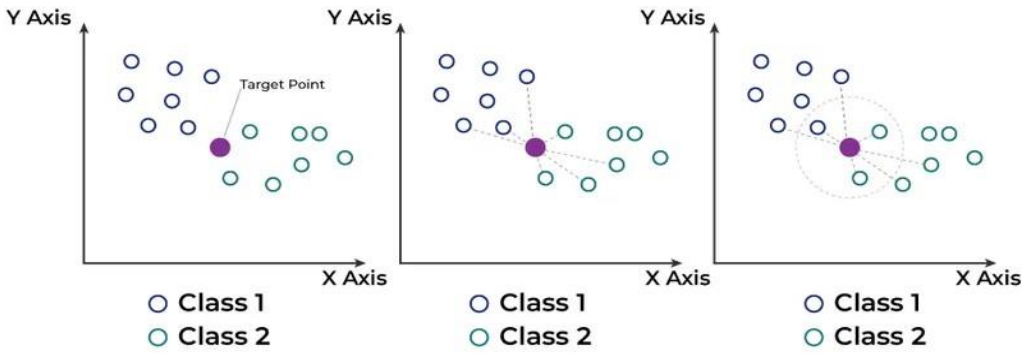
Support Vector Machine (SVM) est un puissant algorithme d'apprentissage automatique utilisé pour la classification linéaire ou non linéaire, la régression et même les tâches de détection des valeurs aberrantes. Les SVM peuvent être utilisés pour diverses tâches, telles que la classification de texte, la classification d'images, la détection de spam, l'identification d'écriture manuscrite, l'analyse d'expression génique, la détection de visage et la détection d'anomalies. Les SVM sont adaptables et efficaces dans une variété d'applications, car ils peuvent gérer des données à haute dimension et des relations non linéaires.

Les algorithmes de SVM sont très efficaces car ils essaient de trouver le maximum de séparation entre les différentes classes disponibles dans la fonctionnalité cible.



7.3 KNN :

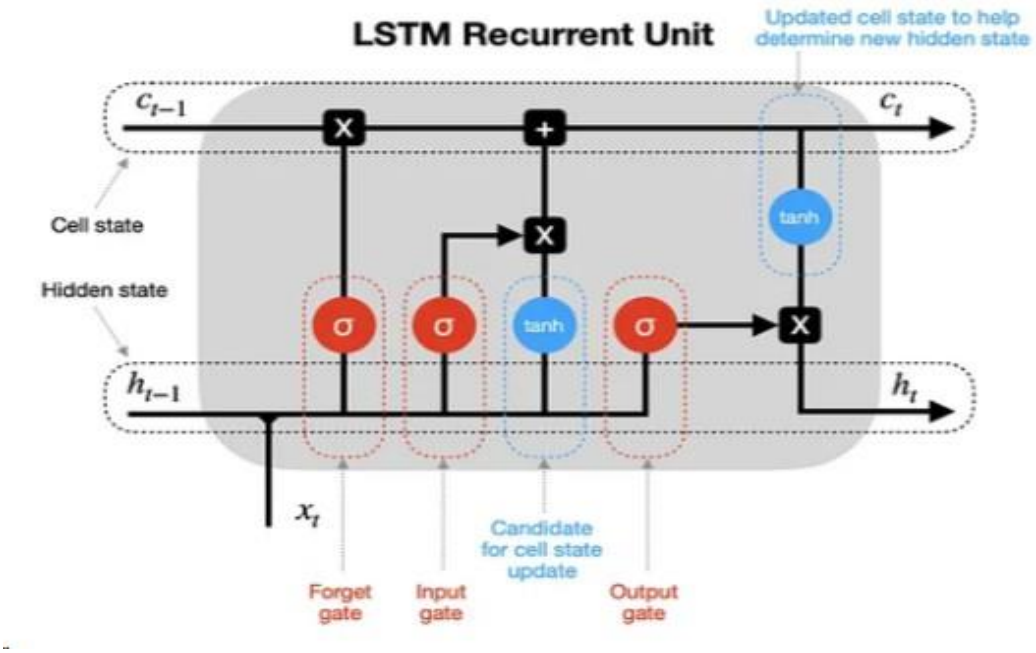
L'algorithme k-nearest neighbors, également connu sous le nom de KNN ou k-NN, est un classificateur d'apprentissage supervisé non paramétrique, qui utilise la proximité pour faire des classifications ou des prédictions sur le regroupement d'un point de données individuel. Bien qu'il puisse être utilisé pour des problèmes de régression ou de classification, il est généralement utilisé comme algorithme de classification, en partant de l'hypothèse que des points similaires peuvent être trouvés les uns près des autres.



7.4 LSTM :

LSTM (Long Short-Term Memory) est une architecture de réseau neuronal récurrent (RNN) largement utilisée dans le Deep Learning. Il excelle dans la capture de dépendances à long terme, ce qui le rend idéal pour les tâches de prédiction de séquence.

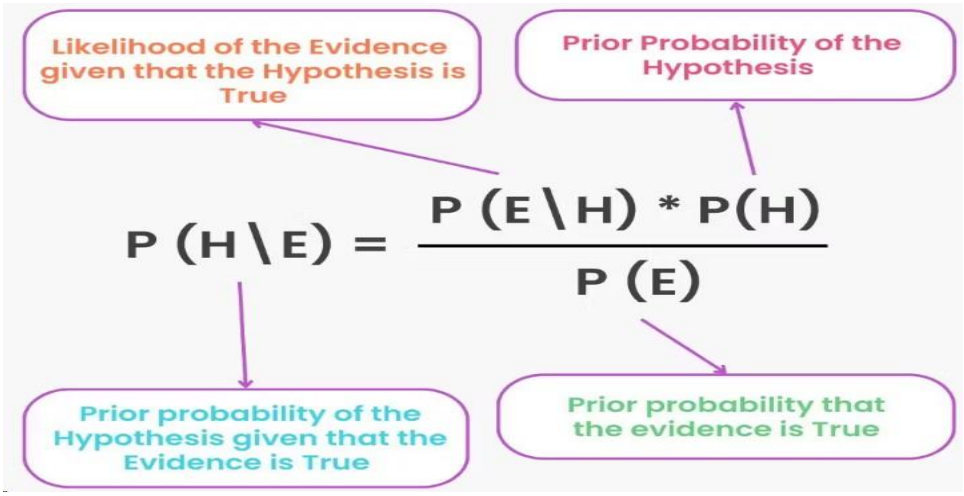
Contrairement aux réseaux de neurones traditionnels, LSTM intègre des connexions de rétroaction, ce qui lui permet de traiter des séquences entières de données, pas seulement des points de données individuels. Cela le rend très efficace dans la compréhension et la prédiction des modèles dans les données séquentielles comme les séries chronologiques, le texte et la parole.



7.5 Naive Bayes :

L'algorithme de Naive Bayes est l'un des algorithmes cruciaux de l'apprentissage automatique qui aide à résoudre les problèmes de classification. Il est dérivé de la théorie des probabilités de Bayes et est utilisé pour la classification de texte, où des ensembles de données de grande dimension sont entraînés.

C'est un algorithme qui apprend la probabilité de chaque objet, ses caractéristiques et les groupes auxquels ils appartiennent. Il est également connu comme un classificateur probabiliste. L'algorithme de Naive Bayes fait l'objet d'un apprentissage supervisé et est principalement utilisé pour résoudre des problèmes de classification.



7.6 Fuzzy c-means :

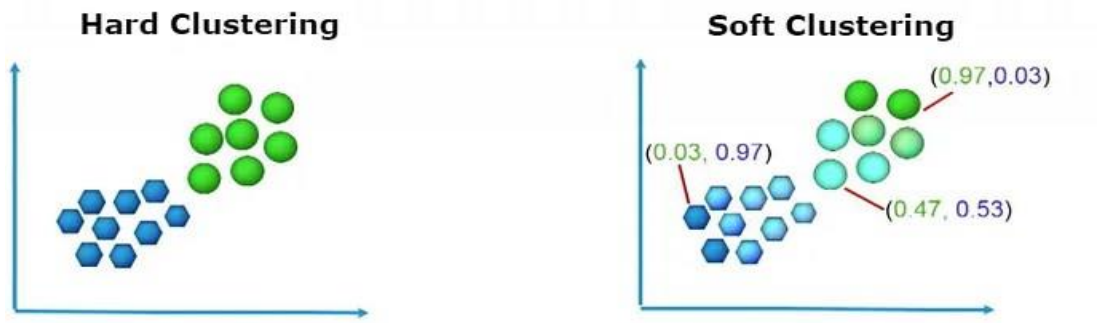
Les principes de logique fuzzy peuvent être utilisés pour regrouper des données multidimensionnelles, en attribuant à chaque point un adhésion dans chaque centre de cluster de 0 à 100 pour cent. Cet algorithme fonctionne en attribuant l'appartenance à chaque point de données correspondant à chaque centre de cluster sur la base de la distance entre le centre de cluster et le point de données. Plus les données sont proches du centre de cluster, plus son adhésion au centre de cluster particulier est importante.

Le flux de processus de c-moyens flous est énuméré ci-dessous:

1. **Supposer** un nombre fixe de clusters k.
2. **Initialisation:** Initialiser aléatoirement les k-moyennes uk associées aux clusters et calculer la probabilité que chaque point de données xi soit membre d'un cluster donné k, P(point xi a l'étiquette k|xi, k).
3. **Itération:** Recalculer le centroïde du cluster comme le centroïde pondéré étant donné les probabilités d'appartenance de tous les points de données xi:

$$\mu_k(n+1) = \frac{\sum_{x_i \epsilon k} x_i * P(\mu_k|x_i)^b}{\sum_{x_i \epsilon k} P(\mu_k|x_i)^b}$$

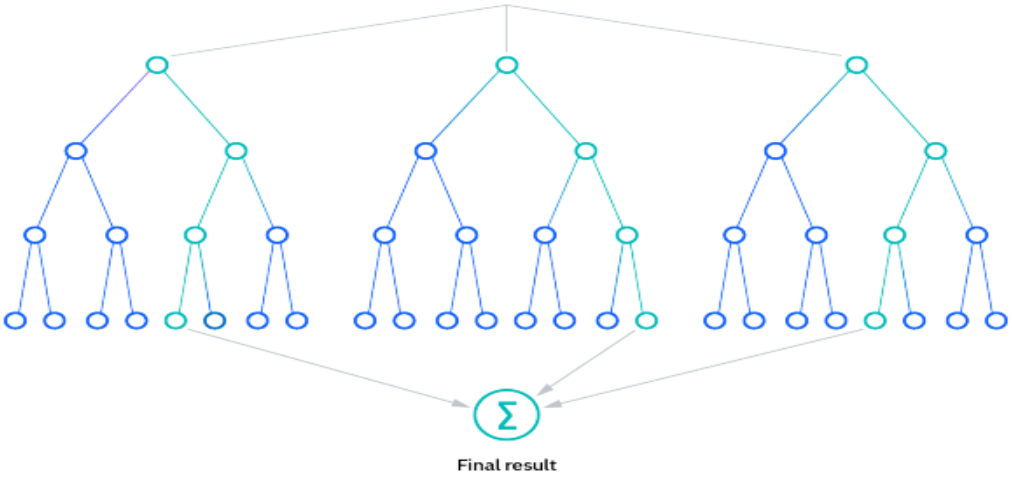
4. **Résiliation:** Itérer jusqu'à la convergence ou jusqu'à ce qu'un nombre d'itérations spécifié par l'utilisateur ait été atteint (l'itération peut être piégée à certains maxima ou minima locaux).



7.7 Random Forest :

Random Forest Algorithm est un algorithme d'apprentissage automatique supervisé qui est extrêmement populaire et est utilisé pour les problèmes de Classification et de Régression dans le Machine Learning.Plus le nombre d'arbres dans un Algorithme Random Forest est élevé, plus sa précision et sa capacité de résolution de problèmes sont élevées. Random Forest est un classificateur qui contient plusieurs arbres de décision sur divers

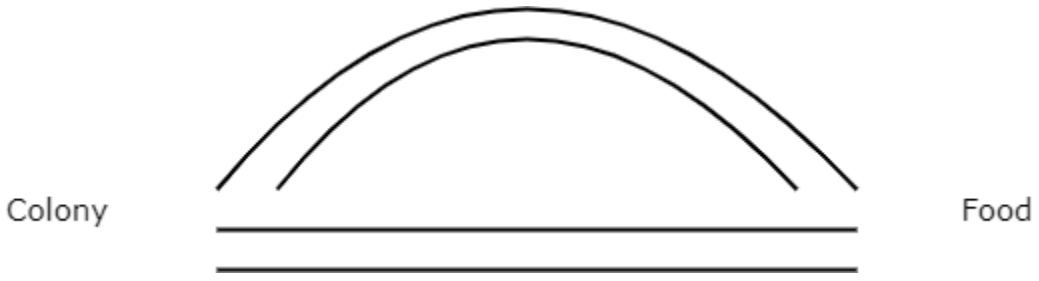
sous-ensembles de l'ensemble de données donné et prend la moyenne pour améliorer la précision prédictive de cet ensemble de données. Il est basé sur le concept d'apprentissage d'ensemble qui est un processus de combinaison de plusieurs classificateurs pour résoudre un problème complexe et améliorer les performances du modèle.

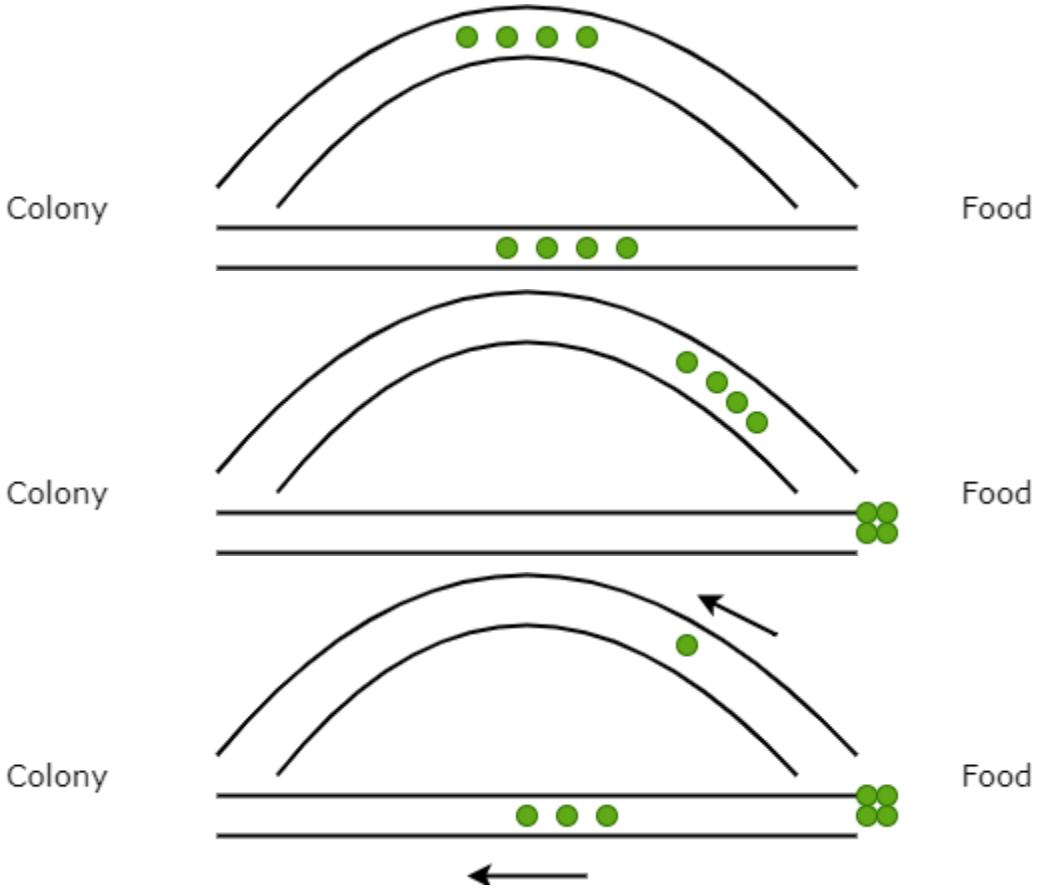


7.8 Optimisation des colonies de fourmis (ACO):

Optimisation des colonies de fourmis (ACO) est une métaheuristique basée sur la population qui peut être utilisée pour trouver des solutions approximatives à des problèmes d'optimisation difficiles.

Dans ACO, un ensemble d'agents logiciels appelé fourmis artificielles rechercher de bonnes solutions à un problème d'optimisation donné. Pour appliquer ACO, le problème d'optimisation est transformé en problème de trouver le meilleur chemin sur un graphique pondéré. Les fourmis artificielles construisent progressivement des solutions en se déplaçant sur le graphique. Le processus de construction de la solution est stochastique et est biaisé par un modèle de phéromone, c'est-à-dire un ensemble de paramètres associés à des composants de graphe (nœuds ou arêtes) dont les valeurs sont modifiées à l'exécution par les fourmis.



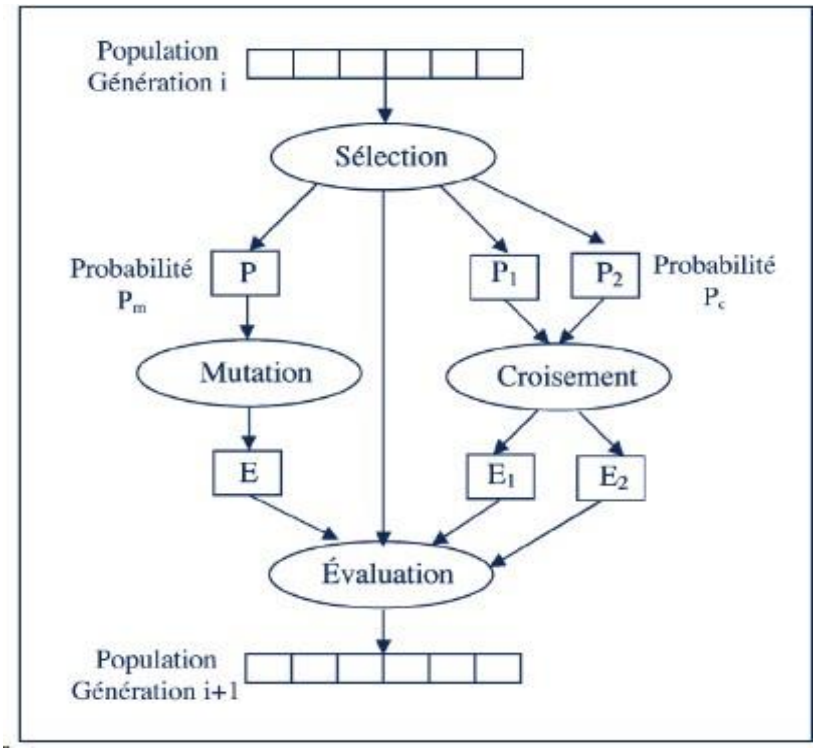


7.9 Algorithmes Génétiques:

L'Algorithme Génétique est un algorithme d'optimisation qui s'inspire du processus d'évolution des êtres vivants.

En recherche opérationnelle, l'Algorithme Génétique est une métaheuristique de la grande famille des algorithmes d'évolution qui offrent l'avantage de fournir des solutions de très grande qualité en un temps raisonnable ; l'inconvénient est qu'il n'y a aucune garantie que la solution soit l'optimum global.

- L'Algorithme Génétique se base au départ sur une population de solutions candidates appelées parfois individus, créatures, phénotypes qui va évoluer de génération en génération jusqu'à la génération qui contient les meilleures solutions.
- Chaque individu comprend des propriétés et il peut être sujet à des transformations génétiques (mutation, croisement par exemple).
- Chaque individu est évalué et cette valeur d'aptitude (fitness value) est un critère pour sa survie d'une génération à une autre.



8. Les métriques :

8.1 Matrice de confusion :

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

Matrice de confusion

8.2 Accuracy (Justesse):

La justesse est le taux de réussite ou encore le taux de prédiction.
La justesse se calcule comme ceci:

$$\text{Justesse} = \frac{VP + VN}{VP + VN + FP + FN}$$

8.3 Precision :

La précision permet de répondre à la question: “Quelle proportion d’identifications positives était effectivement correcte ?”

La précision se calcule comme ceci:

$$\text{Précision} = \frac{V_p}{V_p + F_p}$$

8.4 Recall (Rappel) :

Le rappel (recall en anglais ou aussi sensibilité ou encore taux de vrais positifs (TVP)) permet de répondre à la question suivante: “Quelle proportion de résultats positifs réels a été identifiée correctement ?”

Le rappel se calcule comme ceci:

$$\text{Rappel} = \frac{V_p}{V_p + F_n}$$

8.5 F1-Score :

Le score F1 permet de traduire l’équilibre entre la précision et le rappel. Attention, le problème de cette métrique est qu’elle ne tient pas compte de l’éventuel déséquilibre entre les classes.

Il se calcule comme ceci:

$$\text{F-score} = 2\,\frac{précision * rappel}{précision + rappel}$$

8.6 AUC (Area Under the Curve ROC) :

Un excellent modèle a une AUC proche de 1, ce qui signifie qu'il a une bonne mesure de séparabilité. Un mauvais modèle a une AUC proche de 0, ce qui signifie qu'il a la pire

mesure de séparabilité. Et lorsque l'AUC est de 0,5, cela signifie que le modèle n'a aucune capacité de séparation des classes. Ce serait un modèle naïf.

En gros l'AUC correspond à l'intégrale de la fonction ROC.

La courbe ROC (Receiver Operating Characteristic) trace le taux de vrais positifs en fonction du taux de faux positifs.

Le taux de vrais positifs (TVP) est l'équivalent du rappel.

Le taux de faux positifs (TFP) se calcule comme ceci:

$$TFP = \frac{FP}{FP + VN}$$

9. Comparaison entre les techniques de l'ACP :

- ACP en utilisant la méthode de Jacobi:

	Accuracy	Precision	F1-score	Recall	AUC
ACP	76%	77%	76%	85%	0.827
ACP avec SVM	85.3%	86%	86.2%	**94.1%**	0.850
ACP avec KNN	**89.27%**	**90%**	**89.21%**	88.34%	**0.893**
ACP avec LSTM	66.34%	65.45%	67.60%	69.90	0.663
ACP avec Fuzzy	74.63%	71.79%	76.63%	81.55%	0.791
ACP avec Fuzzy et ACO	76.59%	72.73%	78.57%	85.44%	0.841
ACP avec Naive Bayes	76.10%	72.13%	78.22%	85.44%	0.838

❖ ACP en utilisant la méthode de QR:

	Accuracy	Precision	F1-score	Recall	AUC
ACP	83.41%	83%	83.92%	**91.26%**	0.708
ACP avec SVM	83.85%	83%	83.92%	82%	0.547
ACP avec KNN	**89.26%**	**90%**	**89.21%**	88.34%	**0.893**
ACP avec LSTM	66.82%	65.76%	68.22%	70.87%	0.668
ACP avec Fuzzy	76.41%	73.31%	78.20%	83.78%	0.798
ACP avec Fuzzy et ACO	77.19%	76.89%	77.63%	78.38%	0.838
ACP avec Naive Bayes	74.27%	71.97%	75.91%	80.31%	0.825

❖ ACP en utilisant la méthode de Puissance Itérée:

	Accuracy	Precision	F1-score	Recall	AUC
ACP	82.93%	79.31%	84.02%	**89.32%**	0.680
ACP avec SVM	67.32%	67%	69%	71%	0.733
ACP avec KNN	**86.83%**	**88.77%**	**86.56%**	84.46%	**0.868**
ACP avec LSTM	67.31%	68.36%	66.66%	65.04%	0.673

ACP avec Fuzzy	79.27%	76.17%	80.81%	86.06%	0.820
ACP avec Fuzzy et ACO	76.59%	76.17	77.25%	78.37%	0.849
ACP avec Naive Bayes	78.05%	77.06%	78.87%	80.77%	0.857

- ACP avec Algorithme Génétique (random forest)

	Accuracy	Precision	F1-score	Recall	AUC
ACP - AG	**85%**	**83%**	**89%**	**99%**	0.770
AG - ACP	75%	71%	83%	**99%**	**0.790**

- Fuzzy seulement:

	Accuracy	Precision	F1-score	Recall	AUC
Fuzzy	73.48%	71.28%	75.18%	79.53%	0.734

Conclusion

En conclusion, l'analyse en Composantes Principales (ACP) est une méthode puissante et polyvalente pour explorer et comprendre la structure des données de dimensions élevées.

Dans ce projet, nous avons exploré en détail le processus de l'ACP, de la standardisation des données à la représentation graphique des individus dans l'espace des composantes principales.

Nous avons également discuté des principes fondamentaux de l'ACP, de ses objectifs et des différentes méthodes associées, notamment la standardisation des données, le calcul de la matrice de covariance, la sélection des composantes principales, et la représentation graphique.

En utilisant un exemple concret de données sur les maladies cardiaques, nous avons démontré comment l'ACP peut être appliquée pour réduire la dimensionnalité des données tout en préservant l'information importante.

L'objectif principal de l'ACP est de simplifier la compréhension d'un grand ensemble de variables, de créer des outils d'analyse, et de réduire l'information tout en minimisant la perte d'information.

Nous avons également examiné plusieurs méthodes pour calculer les valeurs propres, telles que la méthode Jacobi, la méthode de la Puissance Itérée, et la méthode de QR. Ces méthodes sont cruciales pour obtenir les composantes principales qui définissent l'espace des variables originales.

Par la suite, nous avons abordé l'application de divers algorithmes d'apprentissage machine tels que les réseaux de neurones artificiels (ANN), les machines à vecteurs de support (SVM), les k-plus proches voisins (KNN), les réseaux de neurones récurrents à longue mémoire (LSTM), Naive Bayes, Fuzzy c-means, Random Forest, Optimisation des colonies de fourmis (ACO), et les algorithmes génétiques. Ces algorithmes ont été utilisés pour comparer les performances de l'ACP dans le cadre de la classification des maladies cardiaques.

Enfin, une comparaison détaillée des résultats a été effectuée, mettant en évidence les performances relatives de chaque algorithme en termes d'accuracy, de precision, de recall, de F1-score, et d'Area Under the Curve (AUC).

Les résultats ont montré que la performance de l'ACP peut varier en fonction de l'algorithme utilisé, soulignant l'importance de choisir le bon algorithme en fonction du contexte spécifique de l'analyse.

En somme,ce qu’on a fait fournit une introduction approfondie à l'ACP, explore son application dans le domaine médical en utilisant des données sur les maladies cardiaques, et propose une évaluation comparative des performances avec différents algorithmes d'apprentissage machine.

L'ACP se révèle être une méthode essentielle pour explorer et interpréter efficacement des ensembles de données complexes et multidimensionnels.

Webographie

- Heart Disease Dataset
- ACP
- Méthode de Jacobi
- Méthode de la Puissance Itérée
- Méthode de QR
- ANN
- SVM
- KNN
- LSTM
- Naive Bayes
- Fuzzy
- Random Forest
- Colonie de Fourmis
- Algorithmes Génétiques
- Métriques