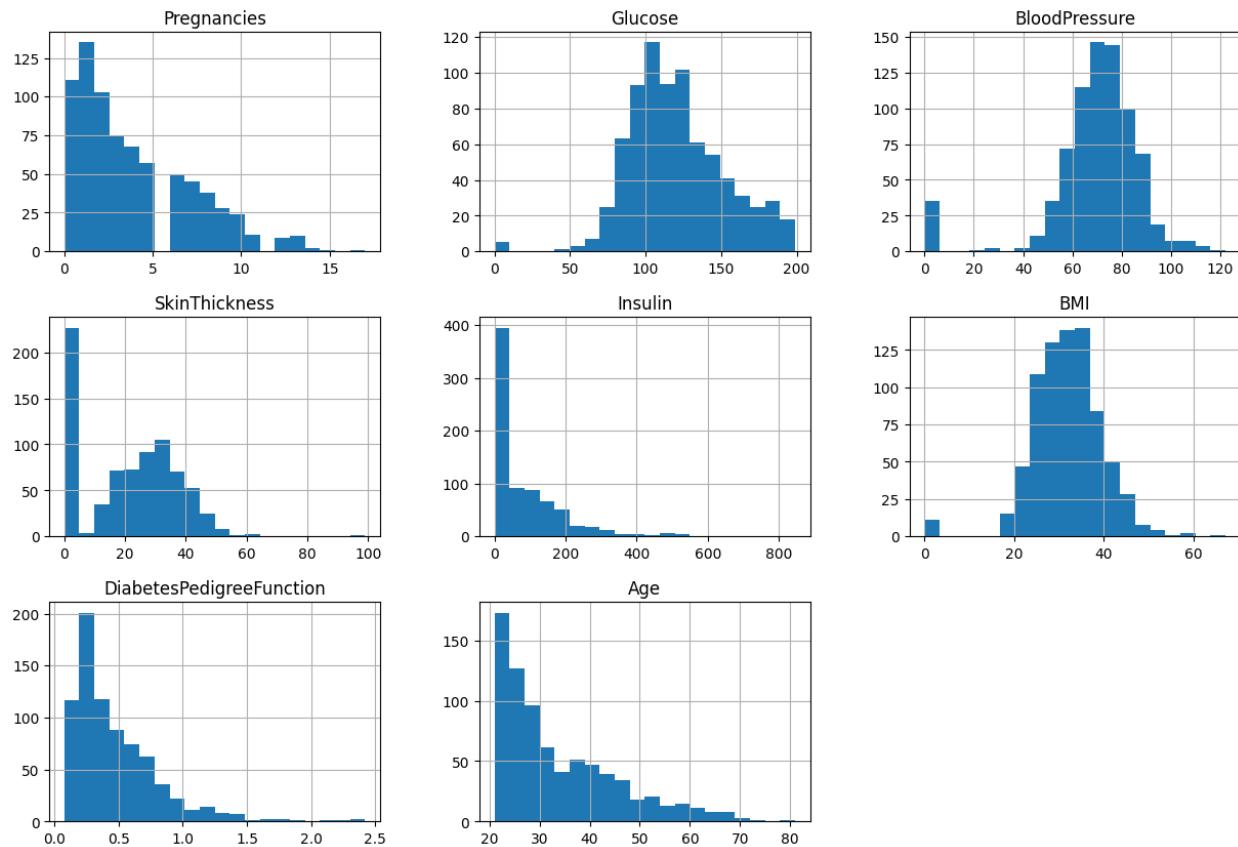


Distribution des variables



Analyse variable par variable – Distribution statistique

L'analyse de la distribution des variables met en évidence des comportements statistiques différents selon les indicateurs cliniques. Voici une description détaillée variable par variable :

1. Pregnancies (Nombre de grossesses)

La distribution est asymétrique à droite, ce qui signifie que la majorité des patientes ont un nombre relativement faible de grossesses. La plupart se situent entre 0 et 5 grossesses. Toutefois, quelques cas exceptionnels atteignent plus de 15 grossesses, ce qui constitue des valeurs extrêmes.

2. Glucose (Taux de glucose)

La distribution présente une forme de cloche, légèrement asymétrique. On observe une concentration des données entre 100 et 125 mg/dL, plage qui correspond à une zone de pré-diabète selon les standards médicaux. Certaines valeurs très basses, proches de 0, sont médicalement improbables et suggèrent des erreurs de saisie ou des valeurs manquantes codées par 0.

3. BloodPressure (Pression artérielle)

La distribution est assez proche d'une loi normale (distribution gaussienne), centrée autour de 70 à 80 mmHg, ce qui correspond aux normes physiologiques attendues. Néanmoins, quelques valeurs proches de zéro sont suspectes et irréalistes, et doivent être considérées comme manquantes.

4. SkinThickness (Épaisseur du pli cutané)

La distribution est fortement biaisée à droite. Un grand nombre de valeurs sont égales à zéro, ce qui indique probablement des données manquantes ou non mesurées. Un traitement spécifique est donc nécessaire pour corriger ou imputer ces valeurs.

5. Insulin (Insulinémie)

La distribution est extrêmement asymétrique. La majorité des enregistrements ont une valeur égale à zéro, ce qui est peu plausible sur le plan médical et suggère des valeurs manquantes. À l'inverse, quelques individus présentent des taux très élevés, allant jusqu'à 900+, ce qui correspond à des cas rares mais possibles.

6. BMI (Indice de Masse Corporelle)

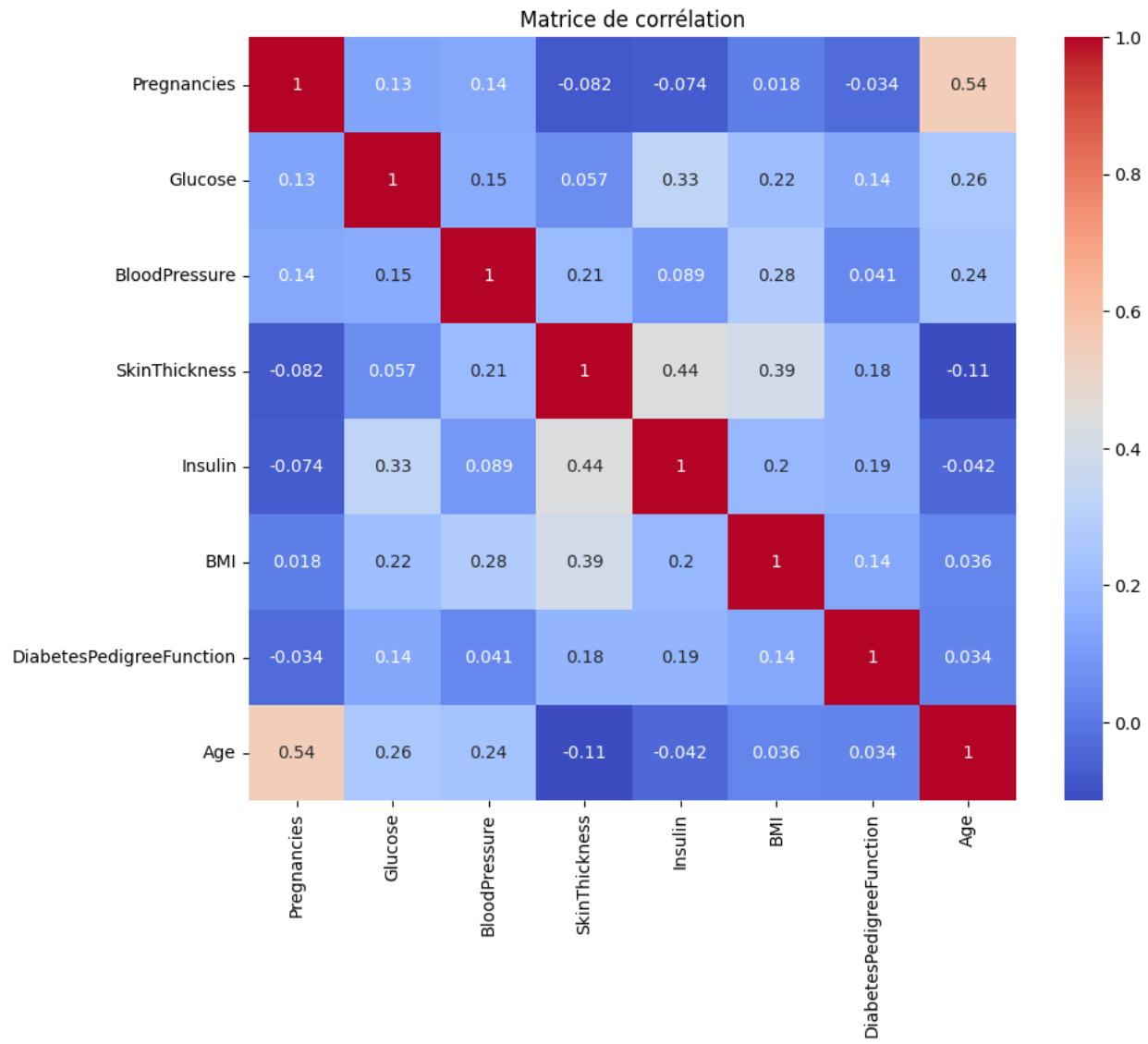
La distribution est légèrement asymétrique à gauche. La majorité des individus ont un IMC compris entre 25 et 40, ce qui correspond à des cas de surpoids ou d'obésité. Les valeurs nulles sont également à considérer comme manquantes ou invalides, car un IMC égal à zéro n'a pas de sens clinique.

7. DiabetesPedigreeFunction (Antécédents familiaux de diabète)

La distribution de cette variable est fortement biaisée à droite. Elle mesure la prédisposition génétique au diabète. De nombreuses valeurs sont proches de zéro, tandis que peu d'individus présentent des valeurs supérieures à 1.0.

8. Age (Âge des patients)

La majorité des personnes interrogées ont entre 20 et 45 ans. Cependant, quelques patients sont âgés de plus de 70 ans, avec une valeur maximale proche de 80 ans. On dispose donc d'un éventail d'âge assez large, ce qui est favorable à une modélisation plus robuste.



La **matrice de corrélation** présentée ci-dessus permet d'analyser les relations linéaires entre les différentes variables du jeu de données. Chaque cellule du tableau indique le **coefficent de corrélation** entre deux variables, avec des valeurs comprises entre -1 et +1. Une valeur proche de +1 indique une **forte corrélation positive** (les deux variables évoluent dans le même sens), une valeur proche de -1 indique une **forte corrélation négative** (elles évoluent en sens inverse), tandis qu'une valeur proche de 0 indique **peu ou pas de corrélation linéaire**.

Dans ce cas, plusieurs corrélations intéressantes sont observées :

La variable "**Pregnancies**" (nombre de grossesses) présente une corrélation modérée avec l'**âge** (0.54), ce qui est cohérent : plus une femme est âgée, plus elle a eu de grossesses.

La variable "**Glucose**" (taux de glucose) est modérément corrélée à **l'insuline** (0.33) et à **l'indice de masse corporelle (BMI)** (0.22). Cela reflète une interdépendance physiologique entre le taux de sucre dans le sang, l'obésité et l'activité pancréatique.

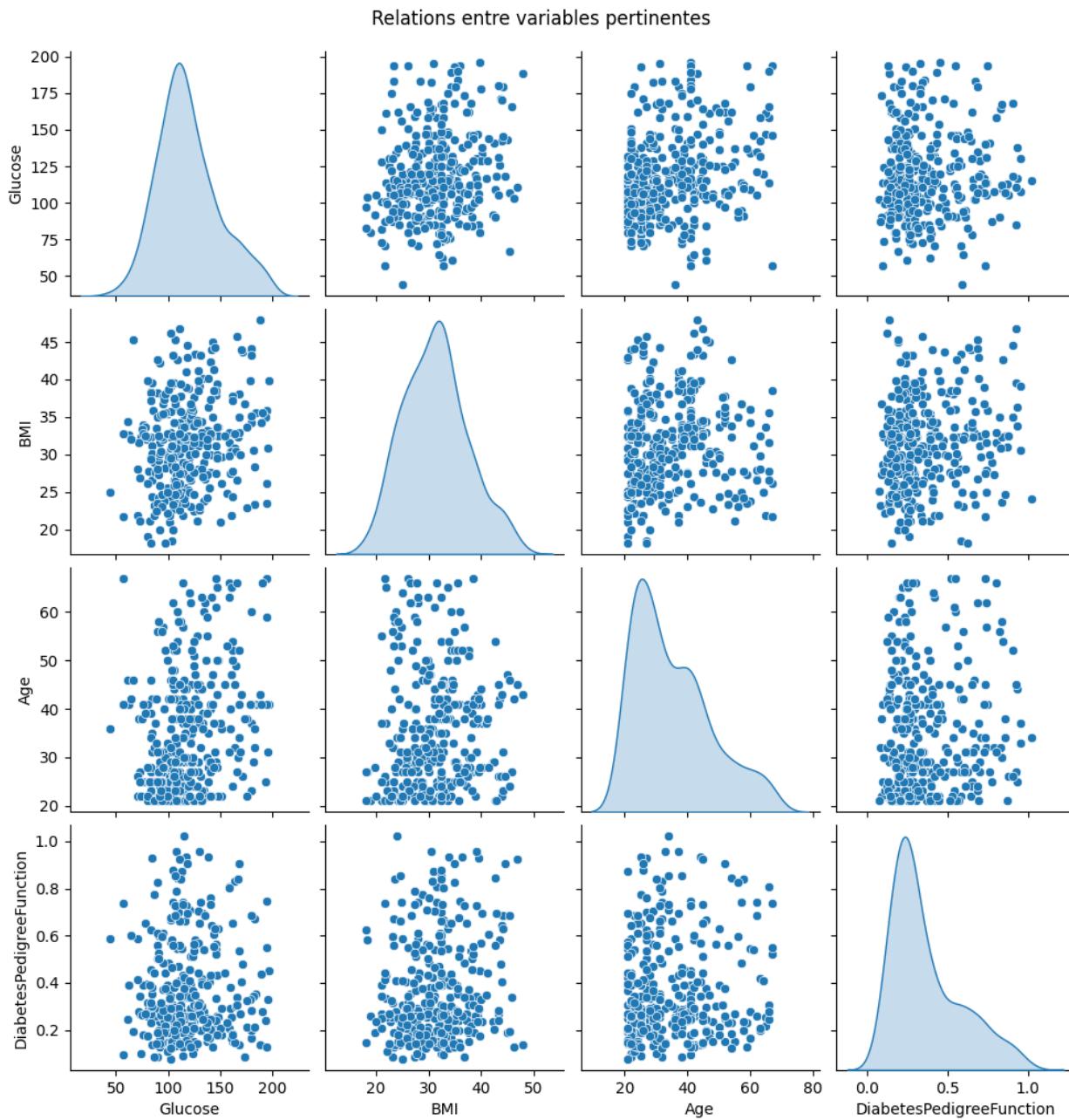
"**SkinThickness**" (épaisseur du pli cutané) est corrélée avec **l'insuline** (0.44) et **le BMI** (0.39), ce qui s'explique par le fait que ces trois indicateurs sont souvent liés au stockage des graisses et à l'obésité.

Le **BMI** est également légèrement corrélé avec la **pression artérielle** (0.28) et le **glucose** (0.22), confirmant que l'excès de poids peut impacter d'autres fonctions physiologiques.

La variable "**DiabetesPedigreeFunction**", qui mesure la prédisposition héréditaire au diabète, est faiblement corrélée avec les autres variables, ce qui suggère qu'elle apporte une information unique au modèle.

Enfin, l'**âge** est très faiblement corrélé à la plupart des variables, à l'exception des grossesses.

Globalement, cette matrice ne révèle **aucune corrélation très forte entre deux variables explicatives**, ce qui est un bon indicateur pour la modélisation, car cela signifie qu'il y a peu de redondance dans les informations. Toutefois, quelques associations modérées peuvent guider le choix des variables les plus pertinentes ou justifier des regroupements ou réductions de dimensions dans les phases suivantes (ex : PCA ou sélection de caractéristiques).



Analyse des relations entre variables pertinentes

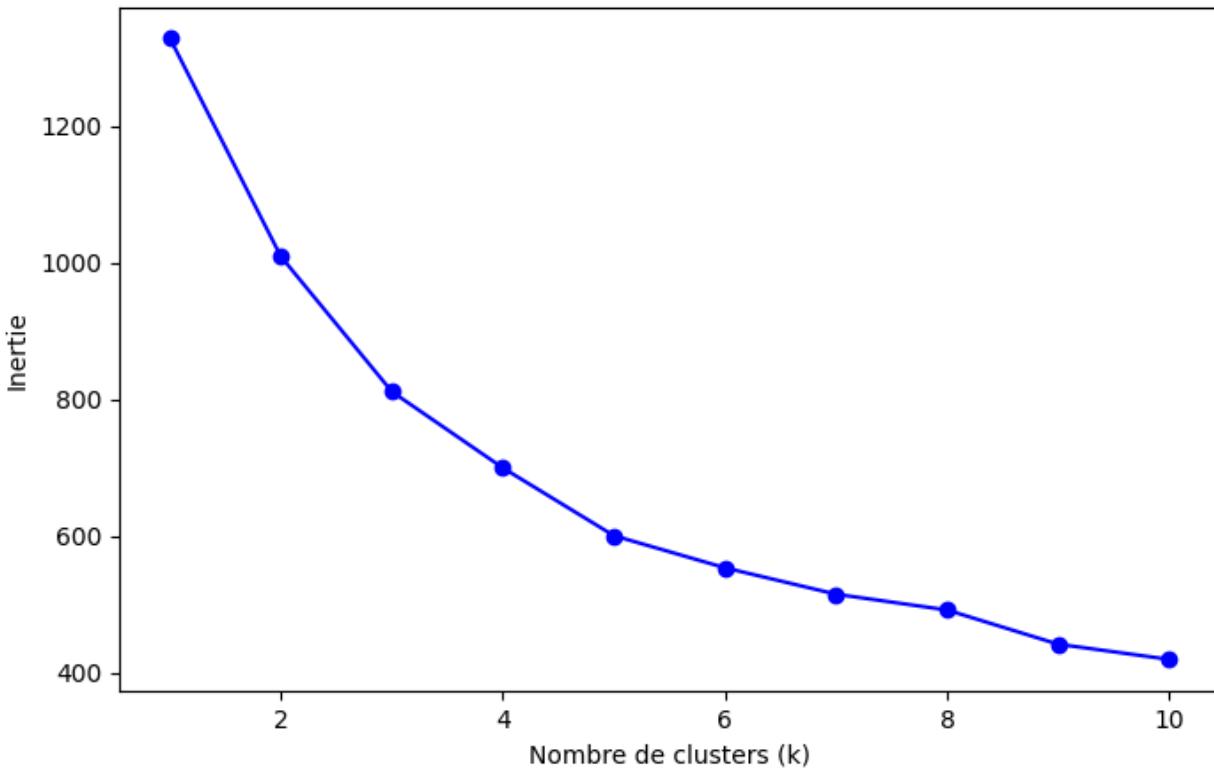
Le graphique ci-dessus représente une **matrice de dispersion** (ou *pairplot*) entre quatre variables considérées comme particulièrement pertinentes pour la prédition du diabète : **Glucose**, **BMI** (**Indice de Masse Corporelle**), **Age**, et **DiabetesPedigreeFunction**. Ce type de visualisation permet d'étudier les relations bivariées sous forme de nuages de points (*scatter plots*) tout en montrant en diagonale la distribution de chaque variable via des courbes de densité.

L'analyse de cette figure révèle plusieurs éléments clés :

- La variable **Glucose** suit une distribution unimodale légèrement asymétrique, centrée autour de 120 mg/dL. Elle semble assez dispersée mais présente une légère concentration verticale dans sa relation avec le BMI, sans lien linéaire fort apparent.
- La variable **BMI** est aussi globalement concentrée entre 25 et 40, ce qui confirme sa distribution centrée sur des cas de surpoids ou d'obésité. Sa relation avec l'âge ou les antécédents génétiques est très diffuse et ne montre pas de corrélation visuelle forte.
- La variable **Age** montre une distribution concentrée entre 20 et 50 ans, avec une faible pente ascendante dans ses relations croisées avec le glucose ou le BMI, mais sans tendance linéaire claire. Cela confirme les résultats de la matrice de corrélation précédente.
- Enfin, la variable **DiabetesPedigreeFunction** présente une distribution fortement biaisée à droite, avec la majorité des valeurs proches de 0. Son nuage de points avec les autres variables est très dispersé, ce qui suggère qu'elle n'a pas de relation linéaire forte avec les autres attributs, mais qu'elle pourrait néanmoins contribuer de façon indépendante à la prédiction.

En conclusion, cette visualisation confirme qu'il n'existe **pas de relation linéaire marquée entre ces variables**, mais qu'elles présentent toutes des comportements distincts, justifiant leur conservation dans le modèle prédictif. Elles apportent chacune une contribution potentiellement complémentaire dans la classification du risque de diabète.

Méthode du coude pour déterminer k optimal



Détermination du nombre optimal de clusters – Méthode du coude

Le graphique ci-dessus illustre l'application de la **méthode du coude (elbow method)** pour déterminer le **nombre optimal de clusters** à utiliser dans une classification non supervisée, en particulier dans un algorithme de **k-means**.

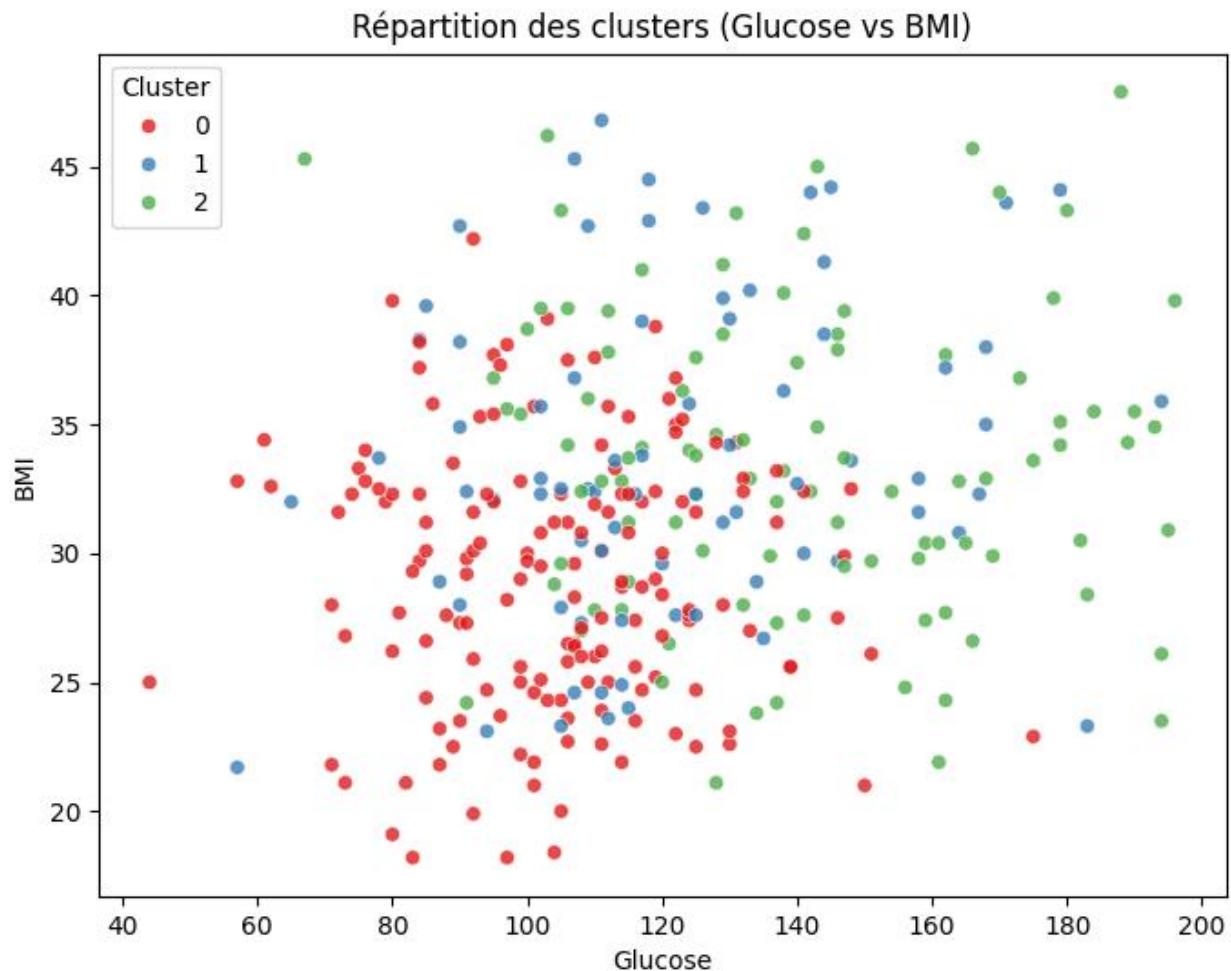
Sur l'axe horizontal figure le nombre de clusters testés (k), allant de 1 à 10, tandis que l'axe vertical représente l'**inertie intra-cluster**, c'est-à-dire la somme des distances entre les points de données et leur centre de cluster. Cette métrique mesure la **cohésion des clusters** : plus l'inertie est faible, plus les points sont proches de leur centre.

On observe que lorsque k augmente :

- L'inertie diminue rapidement entre k=1 et k=3, ce qui montre une amélioration significative de la compacité des clusters.
- À partir de k=4 ou k=5, la diminution devient plus lente et moins significative.
- La courbe forme un "coude" visible autour de **k = 4 ou k = 5**, qui correspond au point où l'ajout de nouveaux clusters n'apporte **qu'un gain marginal** en termes de cohésion.

Selon ce graphique, le **nombre optimal de clusters** semble donc être **k = 4 ou k = 5**, car c'est à ce niveau que le compromis entre complexité et qualité de regroupement est le plus pertinent.

Choisir un k plus élevé n'apporte qu'une faible réduction de l'inertie au prix d'une complexité accrue.



Répartition des clusters (Glucose vs BMI)

Le graphique présenté illustre la **répartition des individus** en trois **clusters distincts** obtenus à l'aide de l'algorithme de classification non supervisée **k-means**, en fonction de deux variables cliniquement importantes : le **taux de glucose** (en abscisse) et l'**indice de masse corporelle (BMI)** (en ordonnée).

Chaque point représente un individu, et la couleur indique le cluster auquel il a été assigné :

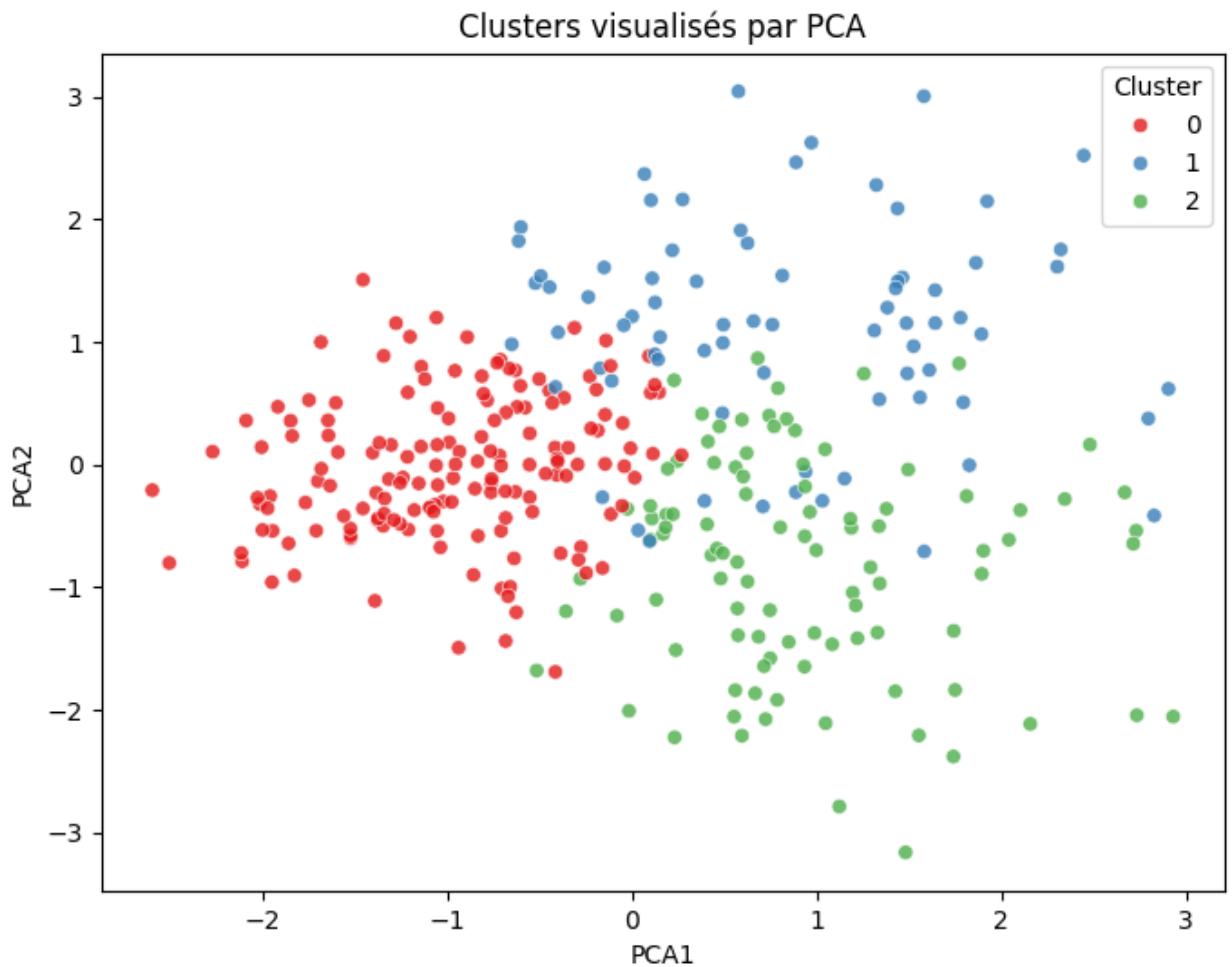
- ● Cluster 0 (rouge)
- ● Cluster 1 (bleu)
- ● Cluster 2 (vert)

L'analyse du nuage de points montre que :

- Le **cluster 0** regroupe majoritairement des individus ayant des **valeurs modérées de glucose (autour de 100–120)** et un **BMI compris entre 25 et 35**, ce qui correspond à une population au profil intermédiaire.
- Le **cluster 1** semble plus dispersé en termes de glucose, mais regroupe des individus avec un **BMI globalement plus élevé**, ce qui pourrait suggérer une population en situation de **surpoids avancé ou obésité**, indépendamment du taux de sucre.
- Le **cluster 2**, quant à lui, regroupe une majorité d'individus ayant des **valeurs élevées de glucose (souvent > 140 mg/dL)**, ce qui pourrait correspondre à une **population à risque diabétique élevé**, même si leur BMI est plus variable.

Ce graphique met donc en évidence l'existence de **groupes distincts dans la population étudiée**, selon des combinaisons particulières de glucose et d'IMC. Ces clusters peuvent être interprétés comme des **profils cliniques** différents, par exemple :

- des patients à faible risque (glucose bas, BMI modéré),
- des patients obèses mais sans hyperglycémie sévère,
- ou des patients potentiellement diabétiques ou pré-diabétiques.



Visualisation des clusters par analyse en composantes principales (PCA)

Le graphique ci-dessus représente la **projection des individus dans un plan à deux dimensions** à l'aide de la **PCA (Analyse en Composantes Principales)**, une technique de réduction de dimension. Cette méthode permet de **résumer l'information contenue dans plusieurs variables initiales** (par exemple : glucose, BMI, insuline, etc.) en deux axes principaux, nommés ici **PCA1 et PCA2**, qui capturent la plus grande part de variance du jeu de données.

Les points sont colorés selon leur **cluster d'appartenance** (résultat d'un algorithme K-Means appliqué en amont) :

- ● **Cluster 0** : groupe compact et bien séparé dans la partie gauche du graphique
- ● **Cluster 1** : groupe plus étendu au centre

- ● **Cluster 2** : réparti dans la zone droite, avec davantage de dispersion

Cette représentation est utile pour **valider visuellement la qualité de la segmentation** :

- Le **cluster 0** apparaît très dense, ce qui suggère une forte cohérence interne entre ses membres.
- Le **cluster 1** montre une répartition plus diffuse, indiquant une certaine variabilité dans les profils qu'il contient.
- Le **cluster 2** semble s'étendre dans une direction opposée à cluster 0, ce qui montre une **bonne séparation spatiale**.

Même si les composantes PCA ne sont pas interprétables directement en termes cliniques, elles permettent ici d'illustrer que les clusters identifiés sont **relativement bien séparés** dans l'espace réduit, ce qui est un bon indicateur de la **qualité du clustering** réalisé.