1. **PRINCIPE-EXPLICATION :**

Dans ce code, nous avons commencé par charger les données du fichier CSV contenant les mesures de patients atteints de diabète. Ensuite, nous avons séparé les données en variables indépendantes (X) et dépendante (y), et normalisé les données en soustrayant la moyenne et en divisant par l'écart type.

Ensuite, nous avons appliqué l'analyse factorielle des composantes (AFC) en utilisant la méthode PCA de la bibliothèque Scikit-learn. Nous avons choisi de réduire la dimensionnalité à deux composantes principales (n\_components=2) pour faciliter la visualisation.

Ensuite, nous avons divisé les données en ensembles de formation et de test, et entraîné un modèle de régression logistique sur les composantes principales en utilisant la bibliothèque Scikit-learn. Nous avons ensuite prédit les résultats sur les données de test et évalué les performances du modèle en utilisant la précision (accuracy\_score) comme métrique.

Notez que ce code est juste un exemple et qu'il peut être adapté en fonction des besoins et des objectifs spécifiques de votre projet.

Dans ce code, nous avons ajouté une section pour afficher les données après l'analyse factorielle des composantes en utilisant la bibliothèque Matplotlib. Nous avons utilisé la fonction scatter pour afficher les données en deux dimensions, en utilisant les composantes principales comme axes x et y. Nous avons également coloré les points en fonction de la valeur de la variable dépendante (y) pour mieux visualiser la relation entre les composantes principales et la variable dépendante.

Notez que ce code est juste un exemple et qu'il peut être adapté en fonction des besoins et des objectifs spécifiques de votre projet.

1. **EXPLICATION DU GRAPHE OBTENU :**

Le graphe obtenu en sortie représente les données du dataset du diabète après l'analyse factorielle des composantes (AFC). Les données sont représentées en deux dimensions, en utilisant les deux premières composantes principales comme axes x et y. Chaque point sur le graphe représente un patient, et la couleur du point indique si le patient est atteint de diabète (en bleu) ou non (en orange).

L'interprétation du graphe dépend de l'objectif de l'analyse. Dans ce cas, l'objectif est de voir s'il y a une relation entre les variables mesurées et la présence de diabète chez les patients. Si les points bleus et oranges sont mélangés et difficilement séparables sur le graphe, cela pourrait indiquer qu'il n'y a pas de relation claire entre les variables mesurées et la présence de diabète. Si les points bleus et oranges sont bien séparés sur le graphe, cela pourrait indiquer qu'il y a une relation claire entre les variables mesurées et la présence de diabète.

Dans ce cas, on peut observer que les points bleus et oranges sont relativement mélangés sur le graphe, avec un certain chevauchement entre les deux groupes. Cela pourrait indiquer que les variables mesurées ne permettent pas de prédire avec certitude la présence ou l'absence de diabète chez les patients, ou que d'autres variables non mesurées peuvent également jouer un rôle important dans la maladie.

En résumé, le graphe obtenu en sortie de l'analyse factorielle des composantes représente les données du diabète en deux dimensions, en utilisant les deux premières composantes principales comme axes x et y, et permet d'observer la relation entre les variables mesurées et la présence de diabète chez les patients. L'interprétation du graphe dépend de l'objectif de l'analyse et de la séparabilité des groupes de données sur le graphe.

1. **EXPLICATION DE LA PROXIMITÉ ENTRE LES VARIABLES :**

La proximité ou le mélange des deux couleurs sur le graphe de l'analyse factorielle des composantes peut être interprété comme une indication de la relation entre les variables mesurées et la présence de diabète chez les patients.

Dans notre exemple, les points bleus représentent les patients atteints de diabète et les points oranges représentent les patients non atteints de diabète. Si les points bleus et oranges sont mélangés sur le graphe, cela peut indiquer que les variables mesurées ne permettent pas de séparer clairement les patients atteints de diabète des patients non atteints de diabète. Cela peut être dû à plusieurs facteurs, tels que la présence de bruit ou de données manquantes, ou le fait que les variables mesurées ne capturent pas la variabilité suffisante dans les données.

D'un autre côté, si les points bleus et oranges sont bien séparés sur le graphe, cela peut indiquer qu'il existe une relation claire entre les variables mesurées et la présence de diabète chez les patients. Cela peut être le cas si les variables mesurées sont fortement corrélées avec la présence de diabète, ou si les variables mesurées capturent une grande partie de la variabilité dans les données.

Il est important de noter que l'interprétation du graphe dépend de plusieurs facteurs, tels que le nombre de composantes principales utilisées, la qualité des données et la méthode d'analyse factorielle utilisée. En général, il est recommandé de combiner l'analyse factorielle des composantes avec d'autres méthodes d'analyse pour obtenir une vue d'ensemble complète des données.

random\_state=42

random\_state=42 est un paramètre dans la fonction train\_test\_split de la bibliothèque Scikit-learn qui permet de fixer la graine aléatoire utilisée pour la division aléatoire des données en ensembles de formation et de test.

Lorsque le paramètre random\_state est fixé à une valeur particulière (dans ce cas, 42), la division des données en ensembles de formation et de test sera toujours la même, quelle que soit la façon dont le code est exécuté. Cela assure une reproductibilité des résultats, ce qui est important pour la validation et la comparaison des modèles.

En général, il est recommandé de fixer la graine aléatoire lors de la division des données pour assurer la reproductibilité des résultats. Cependant, la valeur exacte utilisée pour la graine aléatoire n'a pas d'importance, tant qu'elle est cohérente dans les différentes exécutions du code.

1. Hu
2. uuu