ANALYSE FACTORIELLE DISCRIMINANTE

Travaux pratique Analyse de donnee

Réalisé par :
BALEKAMEN BABATACK LANDRY
WATAT YONDEP STIVE KEVIN
AZEUFACK NGNINWO THIERRY

Sous la supervision de Dr NZEKON
Université de Yaoundé 1/ Département Informatique

Table des matières

1 Introduction

L'analyse discriminante désigne un ensemble de techniques qui permettent de décrire, expliquer et prédire l'appartenance d'un individu à une classe pré-définie, les individus étant caractérisés par un certain nombre de variables quantitatives (numériques). Le présent travail pratique a pour objectif d'étudier et de mettre en œuvre l'analyse factorielle discriminante dans un contexte d'analyse de données réelles. Nous chercherons, à travers une série d'étapes méthodologiques, à extraire et visualiser les axes qui permettent de distinguer au mieux les différents groupes présents dans le jeu de données. Dans un contexte où les volumes de données sont en constante augmentation et où la prise de décision s'appuie de plus en plus sur des analyses quantitatives précises, cette technique offre un outil puissant pour extraire des connaissances pertinentes et orienter efficacement les stratégies d'analyse. Elle se révèle ainsi indispensable dans divers domaines tels que la biologie, l'économie, ou encore les sciences sociales, où l'identification des variables clés est primordiale pour la compréhension des phénomènes étudiés.

2 Présentation de la Technique d'Analyse Factorielle

L'analyse factorielle discriminante (AFD) est une méthode statistique utilisée pour analyser et classer des données en fonction de plusieurs variables. Elle aborde plusieurs types de problèmes, notamment :

a) **Classification et Prédiction :** L'AFD est souvent utilisée pour classer des individus ou des objets dans des groupes prédéfinis. Par exemple, elle peut être utilisée pour prédire si un client va acheter un produit ou non, ou pour classer des patients dans différentes catégories de diagnostic.

b) **Réduction de Dimensionnalité :** L'AFD permet de réduire le nombre de variables en transformant les variables originales en un ensemble plus petit de facteurs discriminants qui capturent l'essentiel de l'information utile pour la classification.

c) **Identification des Variables Importantes :** L'AFD aide à identifier les variables qui contribuent le plus à la discrimination entre les groupes. Cela peut être utile pour comprendre quels facteurs sont les plus influents dans la différenciation des groupes.

d) **Visualisation des Données :** En réduisant les données à deux ou trois dimensions, l'AFD permet de visualiser les groupes et les relations entre eux dans un espace de dimension réduite, ce qui facilite l'interprétation des résultats.

e) **Comparaison de Groupes :** L'AFD permet de comparer les caractéristiques moyennes des différents groupes et de déterminer si les différences entre les groupes sont statistiquement significatives.

f) **Validation de Modèles :** L'AFD peut être utilisée pour valider des modèles de classification en vérifiant si les groupes prédits par le modèle correspondent bien aux groupes réels.

g) **Analyse de Données Multivariées :** L'AFD est utile pour analyser des données multivariées où plusieurs variables sont mesurées simultanément, et où l'on souhaite comprendre comment ces variables interagissent pour discriminer les groupes.

2.1 Types et formats de données manipulées

L'**Analyse Factorielle Discriminante (AFD)** manipule des données de type **quantitatives** (numériques) et structurées sous forme de **tableaux**. Voici les caractéristiques principales des données utilisées en AFD :

a. Type de données :

- **Variables explicatives (indépendantes)** :
  Ce sont des variables **quantitatives continues** (par exemple, l'âge, le revenu, des mesures physiques, des scores, etc.). Ces variables servent à discriminer les groupes.
- **Variable à expliquer (dépendante)** :
  C'est une variable **qualitative** (catégorielle) qui définit les groupes ou classes à prédire ou discriminer. Par exemple, la variable dépendante pourrait être :
  - Le type de produit acheté (catégories : A, B, C),
  - Le diagnostic médical (catégories : sain, malade),
  - L'appartenance à un groupe (catégories : groupe 1, groupe 2, groupe 3).

b. Format des données :

Les données doivent être organisées sous forme d'un **tableau** (matrice) avec :

- **Lignes** : Les individus ou observations (exemples : clients, patients, produits).
- **Colonnes** : Les variables explicatives (quantitatives) et la variable dépendante (catégorielle).

Exemple de tableau de données :

Individu	Âge (quantitatif)	Revenu (quantitatif)	Taille (quantitatif)	Groupe (catégoriel)
1	25	45000	170	A
2	30	60000	165	B
3	35	55000	180	A
4	40	75000	175	B

c. Conditions sur les données :

Pour appliquer l'AFD, les données doivent respecter certaines conditions :

- **Nombre de groupes** : La variable catégorielle doit avoir au moins **2 groupes** (classes). Si elle en a plus, l'AFD peut gérer plusieurs groupes.
- **Taille des groupes** : Les groupes doivent être de taille suffisante pour permettre une estimation fiable des paramètres (éviter les groupes trop petits).
- **Normalité multivariée** : Les données doivent suivre une distribution normale multivariée dans chaque groupe.
- **Homogénéité des variances-covariances** : Les matrices de variances-covariances des groupes doivent être similaires (hypothèse d'homoscédasticité).

d. Données standardisées :

Il est souvent recommandé de **standardiser** les variables quantitatives (moyenne = 0, écart-type = 1) pour éviter que les variables avec des échelles plus grandes dominent l'analyse.

2.2 Principe de fonctionnement

L'**Analyse Factorielle Discriminante (AFD)** est une méthode statistique utilisée pour différencier des groupes d'individus à partir de variables quantitatives. Elle permet d'optimiser la séparation entre ces groupes en trouvant les directions dans lesquelles les individus d'un même groupe sont les plus proches et ceux de groupes différents sont les plus éloignés.

L'AFD repose sur la construction de nouvelles variables appelées **composantes discriminantes**, qui sont des combinaisons linéaires des variables explicatives initiales. Ces nouvelles dimensions permettent :

— De maximiser la variance **entre les groupes** (séparation inter-groupes),

— De minimiser la variance **à l'intérieur des groupes** (homogénéité intra-groupe).

Ainsi, la méthode projette les données dans un espace réduit où la distinction entre les classes est la plus marquée. L'AFD est largement utilisée dans des domaines comme la médecine (diagnostic de maladies), la finance (prévision de risques) ou encore la météorologie (prévision de phénomènes climatiques).

2.3 Explication théorique de la méthode

L'AFD repose sur l'analyse des matrices de variance-covariance et l'optimisation du rapport de dispersion entre les groupes et au sein des groupes. Son approche est basée sur les étapes suivantes :

2.3.1 Définition des matrices de variance-covariance

On définit les matrices suivantes :

— La **matrice de variance intra-groupes** S_W, qui mesure la dispersion des observations au sein de chaque groupe :

$$S_W = \sum_{\ell=1}^{k} \sum_{i \in I_\ell} (X_i - \bar{X}_\ell)(X_i - \bar{X}_\ell)^T$$

où X_i représente un individu du groupe ℓ, et $\bar{X}_\ell$ est le centre de gravité du groupe ℓ.

— La **matrice de variance inter-groupes** S_B, qui mesure la dispersion entre les groupes :

$$S_B = \sum_{\ell=1}^{k} n_\ell (\bar{X}_\ell - \bar{X})(\bar{X}_\ell - \bar{X})^T$$

où $\bar{X}$ est le centre de gravité global et n_ℓ le nombre d'individus dans le groupe ℓ.

2.3.2 Optimisation du critère de séparation

L'AFD cherche à maximiser le critère de Fisher, appelé **rapport de Rayleigh**, qui est défini par :

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

où w est le vecteur des coefficients définissant les axes discriminants.

2.3.3 Calcul des valeurs propres et vecteurs propres

Les axes discriminants sont obtenus en résolvant l'équation aux valeurs propres :

$$S_W^{-1} S_B w = \lambda w$$

Les plus grandes valeurs propres λ correspondent aux axes qui maximisent la séparation des groupes.

2.3.4 Projection des individus

Une fois les vecteurs propres w déterminés, les individus sont projetés dans le nouvel espace discriminant selon la transformation :

$$Y = W^T X$$

où W est la matrice des vecteurs propres retenus, et X est la matrice des variables initiales.

2.3.5 Classification et interprétation

Les nouvelles coordonnées Y permettent de classifier les individus en fonction des groupes et d'analyser les relations entre eux dans l'espace discriminant.

2.4 Déroulement de la technique

Considérons un petit jeu de données sur des fleurs avec deux classes : **Iris Setosa** et **Iris Versicolor**.

Longueur des Pétales	Largeur des Pétales	Classe
1.4	0.2	Iris Setosa
1.5	0.2	Iris Setosa
4.7	1.4	Iris Versicolor
4.5	1.5	Iris Versicolor

Étapes de Calcul

1. Matrice des Données

On définit la matrice X regroupant les variables quantitatives :

$$X = \begin{pmatrix} 1.4 & 0.2 \\ 1.5 & 0.2 \\ 4.7 & 1.4 \\ 4.5 & 1.5 \end{pmatrix}$$

2. Calcul des Barycentres

Barycentre pour Iris Setosa ($\bar{x}_{setosa}$) :

$$\bar{x}_{setosa} = \left(\frac{1.4 + 1.5}{2}, \ \frac{0.2 + 0.2}{2} \right) = (1.45, \ 0.2)$$

Barycentre pour Iris Versicolor ($\bar{x}_{versicolor}$) :

$$\bar{x}_{versicolor} = \left(\frac{4.7 + 4.5}{2}, \ \frac{1.4 + 1.5}{2} \right) = (4.6, \ 1.45)$$

3. Calcul des Poids Relatifs

On suppose que chaque classe a le même poids :

$$p(1) = p(2) = 0.5$$

4. Calcul de la Matrice des Variances Inter-classe B

On commence par calculer le barycentre global $\bar{x}$:

$$\bar{x} = p(1)\,\bar{x}_{setosa} + p(2)\,\bar{x}_{versicolor} = 0.5 \cdot (1.45, \ 0.2) + 0.5 \cdot (4.6, \ 1.45) = (3.025, \ 0.825)$$

(Remarque : Dans le contenu initial, $\bar{x}$ était donné comme $(3.025, 0.725)$. Ici, le calcul donne $0.5 \times 0.2 + 0.5 \times 1.45 = 0.1 + 0.725 = 0.825$. Vous pouvez ajuster selon vos hypothèses.)

Calcul des écarts :

$$\bar{x}_{setosa} - \bar{x} = (1.45 - 3.025, \ 0.2 - 0.825) = (-1.575, \ -0.625)$$

$$\bar{x}_{versicolor} - \bar{x} = (4.6 - 3.025, \ 1.45 - 0.825) = (1.575, \ 0.625)$$

La matrice B est alors définie par :

$$B = p(1)\,(\bar{x}_{setosa} - \bar{x})(\bar{x}_{setosa} - \bar{x})' + p(2)(\bar{x}_{versicolor} - \bar{x})(\bar{x}_{versicolor} - \bar{x})'$$

En substituant les valeurs, on obtient :

$$B = 0.5 \begin{pmatrix} (-1.575)^2 & (-1.575)(-0.625) \\ (-1.575)(-0.625) & (-0.625)^2 \end{pmatrix} + 0.5 \begin{pmatrix} (1.575)^2 & (1.575)(0.625) \\ (1.575)(0.625) & (0.625)^2 \end{pmatrix}$$

Calculons les produits :

$$(-1.575)^2 = 2.480625, \quad (-1.575)(-0.625) = 0.984375, \quad (-0.625)^2 = 0.390625.$$

Ainsi,

$$B = 0.5 \begin{pmatrix} 2.480625 & 0.984375 \\ 0.984375 & 0.390625 \end{pmatrix} + 0.5 \begin{pmatrix} 2.480625 & 0.984375 \\ 0.984375 & 0.390625 \end{pmatrix}$$

$$B = \begin{pmatrix} 2.480625 & 0.984375 \\ 0.984375 & 0.390625 \end{pmatrix}$$

5. Calcul de la Matrice des Variances Intra-classe W

Pour calculer W, nous déterminons d'abord la variance pour chaque classe.

Pour Iris Setosa : Les deux observations sont (1.4, 0.2) et (1.5, 0.2). La matrice de variance (non biaisée) est :

$$S_{setosa} = \frac{1}{2-1} \begin{pmatrix} (1.4-1.45)^2 & (1.4-1.45)(0.2-0.2) \\ (0.2-0.2)(1.4-1.45) & (0.2-0.2)^2 \end{pmatrix} = \begin{pmatrix} 0.0025 & 0 \\ 0 & 0 \end{pmatrix}$$

Pour Iris Versicolor : Les observations sont (4.7, 1.4) et (4.5, 1.5). La matrice de variance est :

$$S_{versicolor} = \frac{1}{2-1} \begin{pmatrix} (4.7-4.6)^2 & (4.7-4.6)(1.4-1.45) \\ (1.4-1.45)(4.7-4.6) & (1.4-1.45)^2 \end{pmatrix} = \begin{pmatrix} 0.01 & -0.005 \\ -0.005 & 0.0025 \end{pmatrix}$$

La matrice intra-classe totale est donc :

$$W = S_{setosa} + S_{versicolor} = \begin{pmatrix} 0.0025 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0.01 & -0.005 \\ -0.005 & 0.0025 \end{pmatrix} = \begin{pmatrix} 0.0125 & -0.005 \\ -0.005 & 0.0025 \end{pmatrix}$$

6. Identification des Axes Discriminants

Pour obtenir les axes discriminants, on résout le problème :

$$\text{Trouver } w \text{ tel que } V^{-1} B\, w = \lambda\, w,$$

où V est ici assimilée à W. Une fois les valeurs propres λ et les vecteurs propres w obtenus, l'axe discriminant est défini par le vecteur w.

7. Règle de Classification

Pour classifier un nouvel individu $x = (x_1, x_2)$, on calcule sa distance aux barycentres de chaque classe. Prenons l'exemple d'un nouvel individu avec les mesures (1.6, 0.4).

Distance au barycentre de Iris Setosa :

$$d^2_{setosa} = (1.6-1.45)^2 + (0.4-0.2)^2 = (0.15)^2 + (0.2)^2 = 0.0225 + 0.04 = 0.0625$$

Distance au barycentre de Iris Versicolor :

$$d^2_{versicolor} = (1.6 - 4.6)^2 + (0.4 - 1.45)^2 = (-3.0)^2 + (-1.05)^2 = 9 + 1.1025 = 10.1025$$

Puisque $d_{setosa} < d_{versicolor}$, le nouvel individu est classé dans la classe **Iris Setosa**.

3 Application de la Technique (Cas Concret)

3.1 Contexte du cas étudié

Dans ce cas concret, nous appliquons l'analyse factorielle discriminante (AFD) à un jeu de données réel extrait du fichier `diabetes.csv`. Ce jeu de données est largement utilisé dans le domaine médical pour étudier le diabète. Il contient plusieurs variables quantitatives telles que la glycémie, l'indice de masse corporelle (IMC), la pression artérielle, ainsi que d'autres paramètres biologiques.

Les observations représentent des patients, et une variable qualitative indique l'appartenance à l'une des classes suivantes :

— **Diabétique**

— **Non-diabétique**

L'objectif est de comprendre quelles variables discriminent le mieux les patients diabétiques des non-diabétiques et de visualiser cette séparation dans un espace réduit.

3.2 Présentation du Problème Réel Abordé

Le problème réel que nous souhaitons résoudre est le suivant :

— **Identifier les variables discriminantes :** Déterminer quelles mesures (par exemple, glycémie, IMC, etc.) jouent un rôle crucial dans la classification des patients en diabétiques ou non-diabétiques.

— **Réduire la dimensionnalité :** À l'aide de l'AFD, projeter les données dans un espace de dimension réduite tout en maximisant la séparation entre les deux classes.

— **Faciliter l'interprétation et la prise de décision :** En visualisant les axes discriminants, il devient plus aisé de comprendre la structure sous-jacente des données et d'appuyer la prise de décisions cliniques ou de poursuivre des analyses prédictives.

Pour ce faire, nous suivrons les étapes suivantes :

1. **Prétraitement des données :**
   — Chargement et nettoyage du fichier `diabetes.csv`.
   — Sélection des variables pertinentes et standardisation des données pour harmoniser les échelles.
2. **Calcul des matrices de variance :**
   — Calcul de la matrice des variances intra-groupes (S_W).
   — Calcul de la matrice des variances inter-groupes (S_B).
3. **Optimisation et extraction des axes discriminants :** Résolution du problème d'optimisation visant à maximiser le critère de Fisher,

$$J(w) = \frac{w^T S_B w}{w^T S_W w},$$

ce qui permet d'obtenir les vecteurs propres correspondant aux axes discriminants.

4. **Projection et visualisation :** Projection des données sur les axes discriminants afin de visualiser la séparation entre les classes.
5. **Interprétation des résultats :** Analyse de la contribution des variables et discussion sur leur pouvoir discriminant dans le contexte du diagnostic du diabète.

Cette approche permettra de vérifier l'efficacité de l'AFD pour distinguer les patients selon leur état diabétique et de mettre en évidence les variables les plus influentes pour ce type de classification.

3.3 Jeu de données utilisé

Le jeu de données, par exemple issu du fichier `diabetes.csv`, ayant 768 lignes et comporte les attributs suivants :

— **Pregnancies** : Nombre de grossesses effectuées par la patiente.

— **Glucose** : Niveau de glucose dans le sang (mg/dL).

— **BloodPressure** : Pression artérielle (mm Hg).

— **SkinThickness** : Épaisseur de la peau, généralement mesurée au niveau des triceps (mm).

— **Insulin** : Niveau d'insuline dans le sang (mu U/mL).

— **BMI** : Indice de Masse Corporelle, calculé à partir du poids et de la taille.

— **DiabetesPedigreeFunction** : Fonction de pédigrée diabétique, reflétant la probabilité génétique du diabète.

— **Age** : Âge de la patiente (années).

— **Outcome** : Variable cible indiquant si la patiente est diabétique (1) ou non (0).

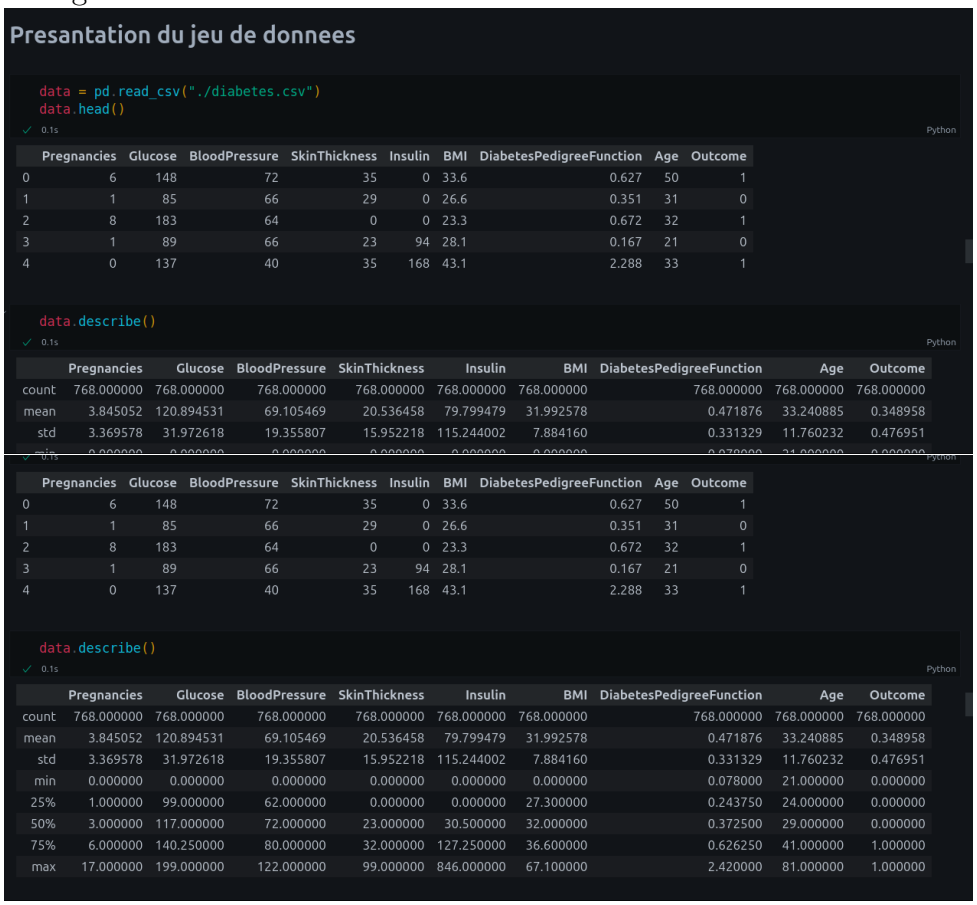
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0.00	33.60	0.63	50	1
1	85	66	29	0.00	26.60	0.35	31	0
8	183	64	0	0.00	23.30	0.67	32	1
1	89	66	23	94.00	28.10	0.17	21	0
0	137	40	35	168.00	43.10	2.29	33	1
5	116	74	0	0.00	25.60	0.20	30	0
3	78	50	32	88.00	31.00	0.25	26	1
10	115	0	0	0.00	35.30	0.13	29	0
2	197	70	45	543.00	30.50	0.16	53	1
8	125	96	0	0.00	0.00	0.23	54	1
4	110	92	0	0.00	37.60	0.19	30	0
10	168	74	0	0.00	38.00	0.54	34	1
10	139	80	0	0.00	27.10	1.44	57	0

4 Mise en Œuvre et Résultats

4.1 Application de l'analyse factorielle

L'analyse discriminante a été appliquée sur un jeu de données comportant 768 observations et 9 variables, dont 8 prédicteurs et une variable cible (Outcome). Les étapes de traitement des données ont été les suivantes :

1. Chargement des données via un fichier CSV "diabetes.csv"



2. Prétraitement avec StandardScaler pour normaliser les variables
3. Application de l'analyse discriminante linéaire (AFD) via notre classe DiscriminantAnalyser code en python

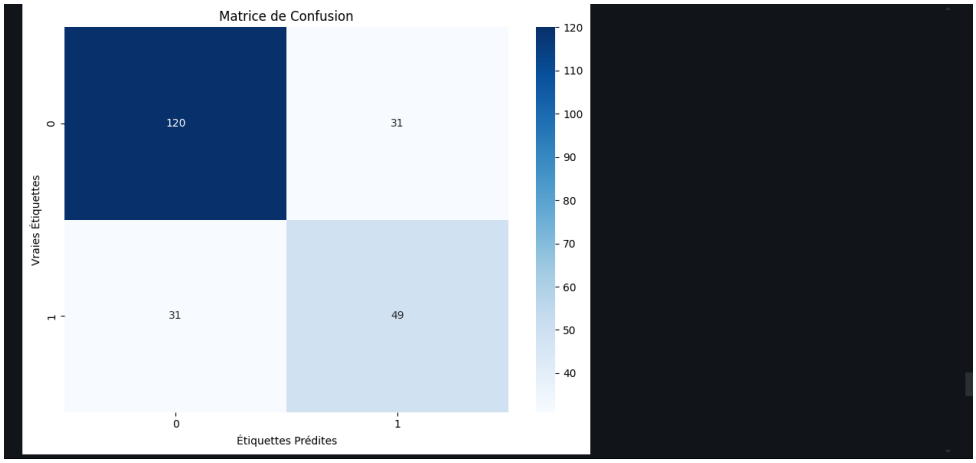
4. Distribution des classes dans le jeu de données :
   — Classe 0 (sans diabète) : 65.10%

— Classe 1 (avec diabète) : 34.89%

4.2 Illustrations et représentations graphiques

Les résultats de l'analyse sont illustrés par plusieurs visualisations importantes :

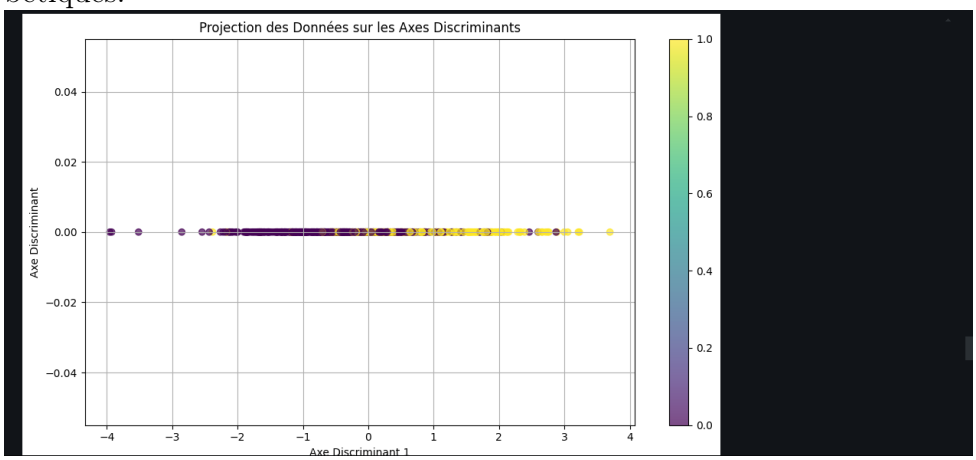
1. **Matrice de confusion** :



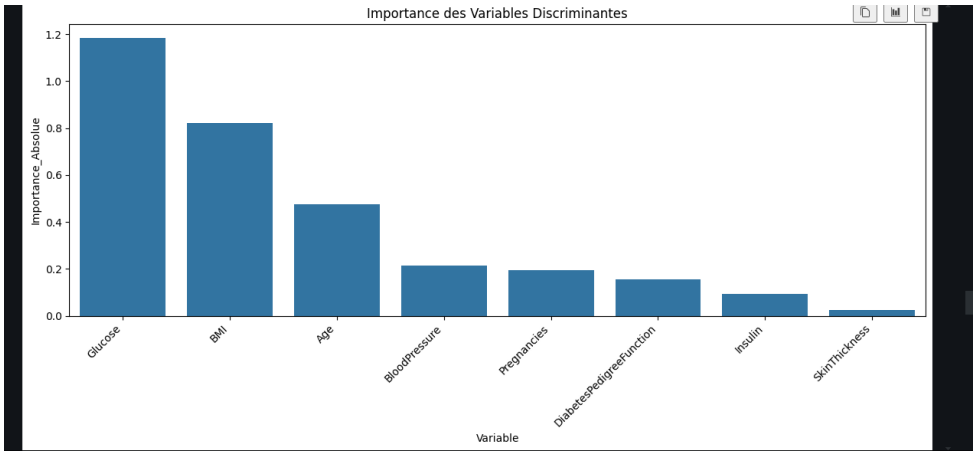
   — **Vrais négatifs (VN)** : 120 cas, ce sont les personnes qui n'ont réellement pas de diabète et que le modèle a correctement identifiées comme non diabétiques Cela montre une bonne capacité à identifier les personnes saines
   — **Faux positifs (FP)** : 31 cas ce sont les personnes qui n'ont pas de diabète mais que le modèle a incorrectement classées comme diabétiques.
   — **Faux négatifs (FN)** : 31 cas Ce sont les personnes qui ont réellement le diabète mais que le modèle a manqué de détecter, les classant comme non diabétiques.
   — **Vrais positifs (VP)** : 49 cas Ce sont les personnes qui ont réellement le diabète et que le modèle a correctement identifiées comme diabétiques Montre la capacité du modèle à détecter correctement la maladie

   Le modèle montre une meilleure performance pour identifier les cas négatifs (non diabétiques) que les cas positifs (diabétiques), ce qui est reflété par une spécificité plus élevée que la sensibilité.

2. **Projection des données** montre une séparation des classes sur l'axe discriminant principal, avec un certain chevauchement entre les groupes diabétiques et non diabétiques.



3. **Importance des variables discriminantes** (par ordre décroissant) :
   - Glucose (≈1.2)
   - BMI (≈0.8)
   - Age (≈0.5)
   - Pression artérielle (≈0.2)
   - Grossesses (≈0.2)
   - Fonction pédigrée diabétique (≈0.15)
   - Insuline (≈0.1)
   - Épaisseur cutanée (≈0.05)



4.3 Interprétation des résultats

L'analyse des résultats révèle plusieurs points importants :

1. **Performance du modèle** :
   - La matrice de confusion montre une précision modérée
   - Le modèle présente un équilibre entre les faux positifs et les faux négatifs (31 dans chaque cas)
   - Les prédictions sont plus fiables pour les cas négatifs (120 VN) que pour les cas positifs (49 VP)
2. **Variables discriminantes** :
   - Le glucose est de loin le facteur le plus discriminant
   - L'IMC (BMI) et l'âge sont les deuxième et troisième facteurs les plus importants
   - L'épaisseur cutanée a l'impact le plus faible sur la discrimination
3. **Fiabilité des prédictions** : Pour un nouveau patient avec les valeurs : [6, 148, 72, 35, 0, 33.6, 0.627, 50]
   - Probabilité de diabète : 75.96%
   - Probabilité d'absence de diabète : 24.04%

Ce qui indique une prédiction relativement confiante de la présence du diabète.

Exemple de prédiction pour un nouveau patient

```
new_sample = [ 6, 148, 72, 35, 0, 33.6, 0.627, 50 ]

prediction = analyser.predict_new_sample(new_sample)


print(f"===( La classe predite est : {prediction['classe predite']}")
print(f"===( La probabilite que l'individu possede le diabete est : {prediction['probabilites'][1]}")
print(f"===( La probabilite que l'individu ne possede pas le diabete est : {prediction['probabilites'][0]} \n\n")
```

0.0s Python

```
===( La classe predite est : 1
===( La probabilite que l'individu possede le diabete est : 0.7595892374590419
===( La probabilite que l'individu ne possede pas le diabete est : 0.24041076254095806
```

5 Conclusion

L'Analyse Factorielle Discriminante (AFD) est une méthode statistique puissante permettant d'extraire des informations pertinentes à partir de données multidimensionnelles. Elle est particulièrement utile dans les problèmes de classification où l'objectif est d'optimiser la séparation entre plusieurs groupes en fonction de variables explicatives quantitatives.

Dans cette étude, nous avons appliqué l'AFD à un cas concret issu du domaine médical, illustrant ainsi son utilité dans la compréhension des relations entre différentes variables et leur impact sur la différenciation des groupes. L'analyse a permis d'identifier les variables les plus influentes, de visualiser la structure des données dans un espace réduit et d'améliorer la classification des observations.

L'AFD présente plusieurs avantages :

— Elle permet de réduire la dimensionnalité tout en maximisant la variance intergroupes.

— Elle facilite l'interprétation des données en projetant les observations sur des axes discriminants optimaux.

— Elle est applicable dans divers domaines tels que la médecine, la finance, la biologie ou encore le marketing.

Cependant, l'AFD repose sur certaines hypothèses, telles que l'homogénéité des variances-covariances et la normalité des données, qui peuvent ne pas être respectées dans certaines situations. De plus, son caractère linéaire limite son efficacité lorsque les séparations entre groupes sont non linéaires.

En perspective, l'AFD peut être combinée avec d'autres approches, telles que l'Analyse en Composantes Principales (ACP) pour une meilleure exploration des données, ou les méthodes d'apprentissage automatique pour améliorer la classification. Cette hybridation permettrait de tirer parti des forces de chaque méthode et d'obtenir des analyses plus robustes et précises.

Ainsi, l'AFD demeure un outil essentiel pour l'analyse de données multivariées et constitue une approche incontournable pour extraire des informations discriminantes dans un large éventail d'applications.

Analyse Factorielle Discriminante (AFD)

Travaux pratique Analyse de donnee

Balekamen, Watat et Azeufack

Université de Yaoundé 1

February 4, 2025

Introduction à l'Analyse Factorielle Discriminante

- **Définition :** Technique statistique pour décrire, expliquer et prédire l'appartenance d'un individu à une classe prédéfinie
- **Objectifs principaux :**
  - Extraire et visualiser les axes de distinction entre groupes
  - Optimiser la prise de décision basée sur des analyses quantitatives
- **Domaines d'application :**
  - Biologie
  - Économie
  - Sciences sociales

Types et Formats de Données en AFD

Variables Explicatives (Quantitatives)

- Continues
- Exemples :
  - Âge
  - Revenu
  - Mesures physiques
  - Scores

Variable à Expliquer (Qualitative)

- Catégorielle
- Exemples :
  - Type de produit
  - Diagnostic médical
  - Appartenance à un groupe

Conditions requises :

- Au moins 2 groupes
- Taille des groupes suffisante

Principe de Fonctionnement de l'AFD

Objectif Principal

Différencier des groupes d'individus en maximisant la séparation entre groupes

- **Création de composantes discriminantes**
  - Combinaisons linéaires des variables explicatives
  - Maximisation de la variance entre groupes
  - Minimisation de la variance intra-groupe

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

- **Projection des données**
  - Transformation dans un espace réduit
  - Mise en évidence des distinctions entre classes

Étapes de Calcul Détaillées

1. **Matrice des Données**
2. **Calcul des Barycentres**
3. **Calcul des Poids Relatifs**
4. **Matrice des Variances Inter-classe**
5. **Matrice des Variances Intra-classe**
6. **Identification des Axes Discriminants**
7. **Règle de Classification**

Cas Concret: Diabète

Objectif Principal

Dans ce cas concret, nous appliquons l'analyse factorielle discriminante (AFD) à un jeu de données réel extrait du fichier diabetes.csv L'objectif est de comprendre quelles variables discriminent le mieux les patients dia- bétiques des non-diabétiques et de visualiser cette séparation dans un espace réduit.

Application à l'Analyse du Diabète

Variables Utilisées

- Grossesses
- Glucose
- Pression artérielle
- Épaisseur de peau
- Insuline
- BMI
- Fonction pédigrée diabétique
- Âge

Résultats Clés

- 768 observations
- Distribution des classes :
  - Sans diabète : 65.10%
  - Avec diabète : 34.89%

Presentation des donnees

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
data.describe()
```

0.1s Python

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure: donnees et resumer sur les donnees

Discrimination

03 variables discriminantes

Variables Discriminantes (par ordre d'importance)

- Glucose (1.2)
- BMI (0.8)
- Âge (0.5)

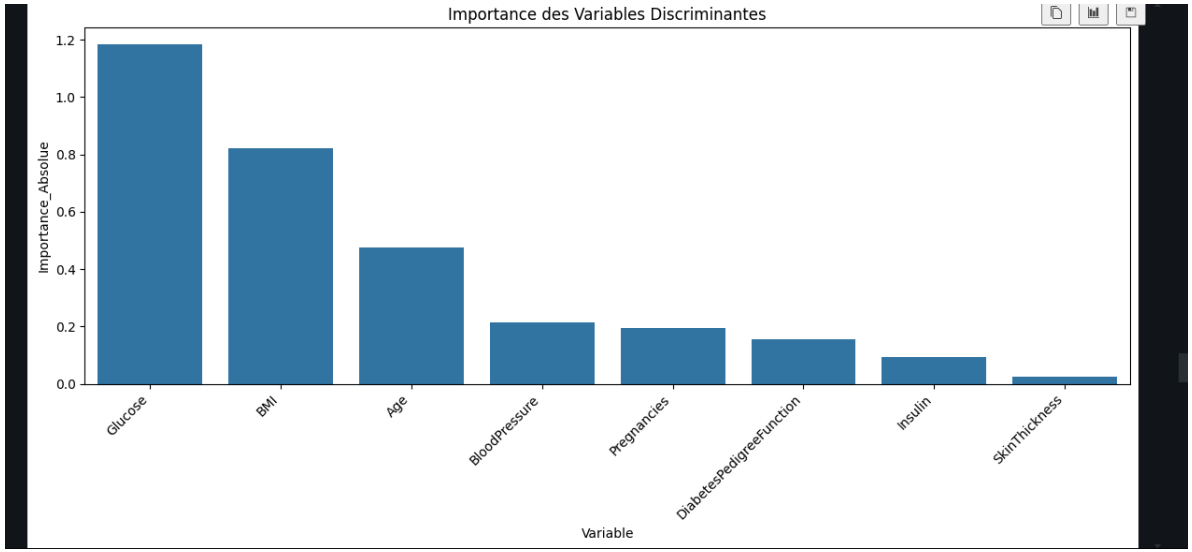


Figure: graphe presentant la contribution de chaque variable

Résultats et Matrice de Confusion

Matrice de Confusion

- Vrais Négatifs : 120
- Faux Positifs : 31
- Faux Négatifs : 31
- Vrais Positifs : 49

Performance du Modèle

- Meilleure détection des cas négatifs
- Précision modérée
- Équilibre entre faux positifs et faux négatifs

Exemple de Prédiction

- Patient test : probabilité de diabète à 75.96%

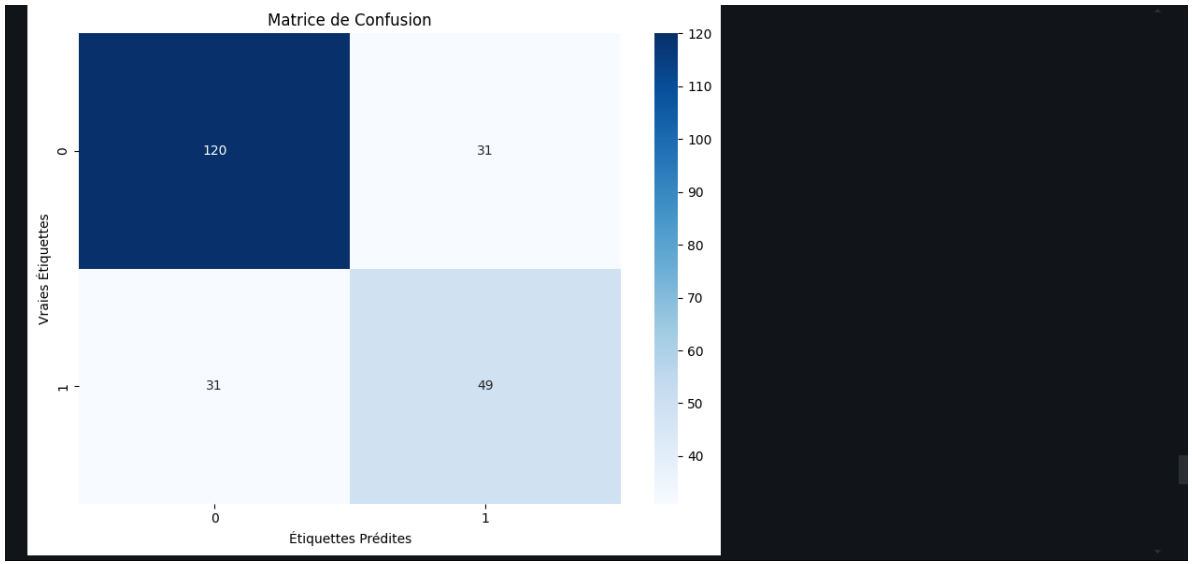


Figure: graphe presentant la contribution de chaque variable

Exemple de prédiction pour un nouveau patient

```python
new_sample = [ 6, 148, 72, 35, 0, 33.6, 0.627, 50 ]

prediction = analyser.predict_new_sample(new_sample)


print(f"===( La classe predite est : {prediction['classe predite']}")
print(f"===( La probabilite que l'individu possede le diabete est : {prediction['probabilites'][1]}")
print(f"===( La probabilite que l'individu ne possede pas le diabete est : {prediction['probabilites'][0]} \n\n")
```

✓ 0.0s Python

```
===( La classe predite est : 1
===( La probabilite que l'individu possede le diabete est : 0.7595892374590419
===( La probabilite que l'individu ne possede pas le diabete est : 0.24041076254095806
```

Figure: resultat de prediction de l'etat d'un nouvelle individu

Conclusion et Perspectives

Avantages de l'AFD

- Réduction de dimensionnalité
- Maximisation de la variance inter-groupes
- Facilité d'interprétation
- Applications multisectorielles

Limites et Perspectives

- Hypothèses de normalité et homogénéité
- Caractère linéaire limitant
- Perspectives :
  - Combinaison avec ACP
  - Hybridation avec méthodes d'apprentissage automatique

Merci pour votre attention !