

## Universite Sultan Moulay Slimane, École Nationale des Sciences Appliquees – Beni Mellal



FILIERE: TRANSFORMATION DIGITALE INDUSTRIELLE
MODULE: ANALYSE DE DONNEES ET BIG DATA
SEMESTRE:4

## Rapport mini-projet

# **Analyse discriminatoire**



Réalisé par :

LAMHAMDI Anouar
OUHAGUA Omar

**Prof** 

**DR. Mohamed GOUSKIR** 

## Table des matières

I. I	ntroduction:	. 2
II.	L'analyse discriminatoire	. 2
1.	Définition Objectifs de l'analyse discriminatoire	. 2
2.	Les différentes techniques d'analyse discriminatoire	. 2
3.	Les hypothèses et les prérequis pour réaliser une analyse discriminatoire	. 2
III.	Etude de cas [Entreprise BETA]	. 2
1.	contexte:	. 2
2.	Problématique à résoudre :	. 2
3.	Méthodologie :	. 2
4.	Description des données utilisées dans le mini-projet	. 2
5.	Préparation des données :	. 4
a	) Importation des données	. 4
b	Transformation des variables qualitatives en variables indicatrices	. 6
c	Transformer la variable « satisfaisant » en facteur	. 9
d	l) Diviser les données	10
6.	APPLIQUER L'ANALYSE DISCRIMINANTE	. 0
1	) Choix et justification de la méthode ADL pas à pas :	. 0
2	Les résultats	. 0
7.	Interprétation des résultats :	. 1
Q	Conclusion	1

## I. INTRODUCTION:

## II. L'ANALYSE DISCRIMINATOIRE

- 1. DEFINITION OBJECTIFS DE L'ANALYSE DISCRIMINATOIRE
- 2. LES DIFFERENTES TECHNIQUES D'ANALYSE DISCRIMINATOIRE
- 3. LES HYPOTHESES ET LES PREREQUIS POUR REALISER UNE ANALYSE DISCRIMINATOIRE

.

## III. ETUDE DE CAS [ENTREPRISE BETA]

#### 1. CONTEXTE:

BETA est une entreprise de restauration qui fournit des services de restauration dans la cité universitaire de Béni Mellal. Les étudiants ont la possibilité de réserver des repas à partir d'un catalogue fixe pour toute la semaine. Les repas sont ensuite distribués par des agents après vérification de la réservation. L'entreprise rencontre des problèmes de qualité des repas et de service, ce qui conduit à des grèves de la part des étudiants insatisfaits.

#### 2. PROBLEMATIQUE A RESOUDRE:

L'entreprise BETA souhaite améliorer la qualité de son service de restauration pour réduire les grèves. Pour atteindre cet objectif, BETA se demande quels sont les critères qui influencent la satisfaction des étudiants et comment elle peut les améliorer.

#### 3. METHODOLOGIE:

Pour répondre à cette problématique, une analyse discriminatoire sera utilisée pour identifier les critères qui ont le plus d'impact sur la satisfaction des étudiants et pour déterminer les caractéristiques des étudiants les plus satisfaits.

#### 4. DESCRIPTION DES DONNEES UTILISEES DANS LE MINI-PROJET

Dans cette étude de cas sur l'entreprise de restauration BETA dans la cité universitaire de Béni Mellal, nous avons utilisé les variables suivantes :

Qualité du service, Qualité des repas : des variables quantitatives continues mesurant respectivement la qualité globale du service, la qualité des repas fournissent par l'entreprise

BETA sur une échelle de 1 à 10, où 1 représente une très mauvaise qualité et 10 représente une excellente qualité.

Le temps d'attente (TA) : correspond au temps que les étudiants doivent attendre avant de recevoir leur repas sur une échelle de 1 à 10, où 1 représente une grande durée d'attente et 10 représente que l'étudiant reçoit son repas vitement.

Sexe : une variable catégorielle nominale indiquant le sexe des étudiants.

Programme d'étude : une variable catégorielle nominale indiquant le programme d'étude des étudiants.

Age correspond à l'âge des étudiants.

La propreté du restaurant (PR) : correspond à une évaluation de la propreté du restaurant de l'entreprise BETA. Cette évaluation peut inclure des critères tels que la propreté des tables, des chaises, des sols, des toilettes, etc. sur une échelle de 1 à 10, où 1 représente une très mauvaise qualité et 10 représente une excellente qualité.

Nombre d'années de résidence : qui correspond au nombres d'années que l'étudiant a passé dans la cité universitaire qui est un nombre entier. Les données nécessaires pour cette analyse sont collectées à travers des enquêtes auprès des étudiants. Les enquêtes sont diffusées en ligne à travers les groupes Facebook et WhatsApp en demandant aux étudiants de remplir un questionnaire.



Figure 1 Capture de questionnaire diffusé

## 5. PREPARATION DES DONNEES:

Avant de procéder à l'analyse des données de l'entreprise BETA, nous avons dû préparer les données en effectuant des étapes de nettoyage et de transformation. Les données ont été collectées à l'aide d'un formulaire Google et stockées dans un fichier Excel. Un échantillon des données préparées est présenté ci-dessous.

4	A	В	С	D	Е	F	G	Н	1
1	La qualité des repas (QR)	Qualité de service des agents (QSA)	Le temps d'attente (TA)	La propreté du restaurt (PR)	age	sexe	Etablissemnt	Satisfait	durée résidance
2	6	7	5-10 minutes	7	22	Masculin	ENSA	0	4
3	1	1	15-20 minutes	1	21	Masculin	ENSA	0	4
4	6	4	5-10 minutes	8	22	Masculin	ENSA	1	4
5	7	8	5-10 minutes	8	22	Masculin	ENSA	1	3
6	6	5	Moins de 5 minutes	6	21	Masculin	ENSA	0	1
7	5	6	10-15 minutes	4	23	Masculin	ENSA	0	1
8	4	9	plus de 20 min	6	24	Masculin	FP	0	5
9	6	6	10-15 minutes	5	23	Masculin	FST	0	3
10	3	3	15-20 minutes	8	21	Masculin	EST	0	1
11	7	5	10-15 minutes	7	21	Masculin	ENSA	1	3
12	1	5	Moins de 5 minutes	7	21	Masculin	ENSA	0	1
13	5	5	5-10 minutes	5	23	Masculin	ENSA	1	4
14	2	2	10-15 minutes	1	22	Féminin	ENSA	0	4
15	6	7	10-15 minutes	7	21	Féminin	ENSA	0	4
16	4	5	5-10 minutes	7	24	Masculin	FST	0	5
17	4	6	Moins de 5 minutes	2	21	Masculin	ENSA	1	1
18	5	7	Moins de 5 minutes	7	21	Masculin	ENSA	1	3
19	1	1	5-10 minutes	1	24	Masculin	FST	0	5
20	5	4	15-20 minutes	6	19	Féminin	ESEF	0	1
21	5	5	15-20 minutes	5	20	Féminin	FP	1	2
22	4	5	10-15 minutes	5	22	Masculin	ENSA	0	4
23	6	9	Moins de 5 minutes	5	21	Masculin	ENSA	1	2
24	6	8	Moins de 5 minutes	7	23	Masculin	ENCG	0	4

Figure 2 échantillon des données collectées

## a) Importation des données

```
# Obtenir le chemin du répertoire de travail actuel
getwd()

# Changer le répertoire de travail pour le dossier contenant les données
setwd("D:/TDI/s4/Analyse de données/analyse discriminatoire")

# Charger les bibliothègues nécessaires
library(readxl)
library(knitr)

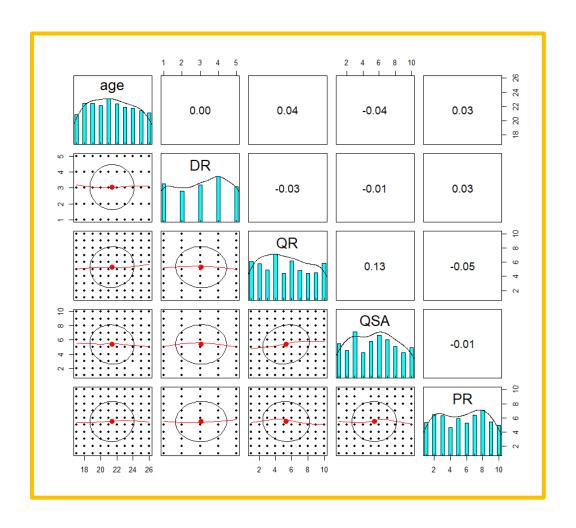
# Importer les données depuis le fichier Excel
data = read_excel("donnee_AD.xlsx", range ="A1:I493", col_names = TRUE)

# Afficher les premières lignes des données pour s'assurer de l'importation correcte
head(data)

# Renommer les noms de colonnes
names(data) = c("QR", "QSA", "TA", "PR", "age", "sexe", "ETAB", "satisfaction", "DR")

# Réorganiser les colonnes pour avoir la variable "satisfaction" en premier
data = data[, c("satisfaction", "age", "sexe", "ETAB", "DR", "QR", "QSA", "PR", "TA")]

# Afficher les données dans un tableau formaté
knitr::kable(data, align = "c")
```



satisfaction								
		::				::		· ·
0 1	22	Masculin		4	6	/	/	5-10 minutes
0 1	21	Masculin		4	1	1	1	15-20 minutes
	22	Masculin		4	6		8	5-10 minutes
		Masculin		3	7		8	
0	21	Masculin	ENSA	1	6	5	6	Moins de 5 minutes
0	23	Masculin	ENSA	1	5		4	10-15 minutes
0	24	Masculin	FP	5	4	9	6	
0	23	Masculin	FST	3	6	6	5	10-15 minutes
0	21	Masculin	EST	1	3		8	15-20 minutes
1	21	Masculin	ENSA	3	7	5		10-15 minutes
0	21	Masculin	ENSA	1	1	5	7	Moins de 5 minutes
1	23	Masculin	ENSA	4	5	5	5	5-10 minutes
0	22	Féminin	ENSA	4	2	2	1	10-15 minutes
0	21	Féminin	ENSA	4	6	7	7	10-15 minutes
0	24	Masculin	FST	5	4	5	7	5-10 minutes
1	21	Masculin	ENSA	1	4	6	2	Moins de 5 minutes
1	21	Masculin	ENSA	3	5	7	7	Moins de 5 minutes
0	24	Masculin	FST	5	1	1	1	5-10 minutes
0	19	Féminin	ESEF	1	5	4	6	15-20 minutes
1	20	Féminin	FP	2	5	5	5	15-20 minutes
j o j	22	Masculin	ENSA	4	4		5	
1 1	21	Masculin	ENSA	2	6	9	5	Moins de 5 minutes
j o j	23	Masculin	ENCG	4	6	8	7	Moins de 5 minutes
j o j	21	Masculin	FST	1	4		7	10-15 minutes
j 1 j	23	Masculin	ENSA	2	6	7	9	5-10 minutes

Cette partie de code importe les données à partir d'un fichier Excel, les réorganise et renomme les colonnes. Il affiche ensuite les données dans un tableau formaté en utilisant la fonction **kable**() de la bibliothèque **knitr** 

## b) Transformation des variables qualitatives en variables indicatrices

La variable « sexe »

```
# Transformer les variables catégorielles en variables indicatrices

# La variable "sexe"

# Obtenir l'indicateur de sexe en utilisant la fonction model.matrix(),
# uui transforme la variable en variables indicatrices
sexe_indicatrice = model.matrix(> sexe - 1, data = data)

# Afficher l'indicateur de sexe
sexe_indicatrice

# Renommer les colonnes de l'indicateur de sexe pour correspondre aux catégories
colonames(sexe_indicatrice)[1] = "sexef"
colonames(sexe_indicatrice)[2] = "sexef"

# Afficher l'indicateur de sexe avec les colonnes renommées
sexe_indicatrice

# Supprimer la colonne "sexe" qui contient "masculin" et "féminin"
data < data[, "which(names(data) == "sexe")]

# Afficher les données sans la colonne "sexe"
knitr::kable(data, align = "c")

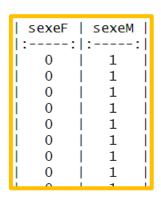
# Ajouter les deux colonnes "sexef" et "sexeM" à l'ensemble de données
data < cbind(data,sexe_indicatrice)

# Afficher les données avec les colonnes "sexef" et "sexeM"
knitr::kable(data, align = "c")

# Récorganiser les colonnes pour avoir la variable "satisfaction" en premier
data < data[, c("satisfaction", "age", "sexef", "sexeM", "ETAB", "DR", "QR", "QSA", "PR", "TA")]

# Afficher les données dans un tableau formaté
knitr::kable(data, align = "c")
```

#### Commentaire



Le code utilise la fonction model.matrix pour transformer la variable catégorielle "sexe" en variables indicatrices et les ajouter aux données. Les colonnes de l'indicateur de sexe sont renommées pour correspondre aux catégories et la colonne "sexe" est supprimée des données. Les nouvelles colonnes indicatrices "sexeF" et "sexeM" sont ajoutées aux données et les colonnes sont réorganisées pour avoir la variable "satisfaction" en premier. Le tableau final est affiché avec la fonction kable.

#### Pourquoi transformer les données

La transformation des variables qualitatives en variables indicatrices est importante pour l'analyse discriminante car elle permet de traiter les variables qualitatives comme des variables numériques, ce qui est nécessaire pour effectuer des calculs mathématiques sur ces variables. Dans l'analyse discriminante, les variables indicatrices sont utilisées pour construire des modèles de discrimination qui permettent de distinguer des groupes en fonction de ces variables. Les variables indicatrices sont également utiles pour effectuer des tests statistiques et des analyses de régression sur des données qualitatives.

#### La variable « ETAB »

```
### La variable "ETAB" ###
ETAB_indicatrice <- model.matrix(~ ETAB - 1, data = data)
ETAB_indicatrice
colnames(ETAB_indicatrice)[1] <- "ENCG"
colnames(ETAB_indicatrice)[2] <- "ENSA"
colnames(ETAB_indicatrice)[3] <- "EST"
colnames(ETAB_indicatrice)[4] <- "FEG"
colnames(ETAB_indicatrice)[5] <- "FL"
colnames(ETAB_indicatrice)[6] <- "FP"
colnames(ETAB_indicatrice)[6] <- "FP"
colnames(ETAB_indicatrice)[7] <- "FST"

#supprimmer la colonne "ETAB"
data <- data[, -which(names(data) == "ETAB")]
#ajouter les nouvelles colonnes
data=cbind(data,ETAB_indicatrice)
knitr::kable(data, align = "c")
data <- data[, c("satisfaction", "age", "sexeF", "sexeM", "ENCG", "ENSA", "EST", "FEG", "FL", "FP", "FST", "DR", "QSA", "PR", "TA")]
knitr::kable(data, align = "c")</pre>
```

L ENGG L	ENGA	LECT	1 55	c   E	LED	LECT	L ETABLET I
ENCG	ENSA			_			ETABFST
		1	: :	-: :	: ::	1	
0	T	0	0	0	1 0	1 0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0

Cette partie du code sert à transformer la variable qualitative "ETAB" en une série de variables indicatrices. Elle utilise la fonction "model.matrix" pour créer une matrice d'indicateurs qui représentent chaque niveau unique de la variable "ETAB". Les nouvelles variables indicatrices sont ensuite ajoutées au jeu de données d'origine, puis la variable "ETAB" est supprimée car elle n'est plus nécessaire. Les noms de colonnes sont également renommés pour correspondre aux différentes catégories de "ETAB". Enfin, les colonnes sont réorganisées pour avoir la variable "satisfaction" en premier.

#### La variable « TA »

TA10-15 minutes	TA15-20 minutes	TA5-10 minutes	TAMoins de 5 minutes	TAplus de 20 min l
			::	
i 0	0	1	0	0 i
i o	1	0	0	o i
i o	0	1	0	i o i
i o	0	1	0	i o i
i o	0	0	1	o i
1	0	0	0	i o i
i o	0	0	0	1
1	0	0	0	0 j
0	1	0	0	0 [
1	0	0	0	0
0	0	0	1	0
0	0	1	0	0
1	0	0	0	0
1	0	0	0	0
0	0	1	0	0
1 0	0	0	1	0
0	0	0	1	0
0	0	1	0	l 0 İ
0	1	0	0	0
I 0	1	0	0	0
1	0	0	0	0
1 0	0	0	1	. o i

```
### La variable "TA" ###
TA_indicatrice <- model.matrix(\sim TA - 1, data = data)
TA_indicatrice
# ajouter les colonnes à data
data=cbind(data,TA_indicatrice)
knitr::kable(data, align = "c")
#supprimmer la colonne de "TA"
data <- data[, -which(names(data) == "TA")]</pre>
knitr::kable(data, align = "c")
View(data)
# connaitre le type des données
str(data)
#La variable "PR" est de type Chr, il faut le convertir en num
str(data$PR)
data$PR <- as.numeric(data$PR)</pre>
str(data$PR)
```

La première partie du code utilise la fonction model.matrix pour créer les nouvelles variables indicatrices, renommer les colonnes, supprimer la colonne originale et ajouter les nouvelles colonnes à la dataframe "data".

La deuxième partie permet de convertir la variable "PR" qui est de type caractère en une variable numérique.

c) Transformer la variable « satisfaisant » en facteur

```
#### transformer la variable "satsfaction" en factor
for (i in seq_along(data$satisfaction)){

if(data$satisfaction[i]==1){
    data$satisfaction[i]<-"s"
}else{data$satisfaction[i]<-"ps"}
}
data$satisfaction=as.factor(data$satisfaction)
str(data$satisfaction)
str(data$satisfaction)
str(data)

#données actifs
data.actif <- subset(data, select = c(1, 2, 12,13,14,15))
#Now our data is ready to be treated like she deserve
# exploration des données
install.packages("psych")
library(psych)</pre>
```

satisfaction	ممد ا	l cavaE	Lsav	ωM   ENC	C I ENS	A LEST	l EEG	l EL I
_	age  ::		: :			-: :		
			.	I 0		111	.	. [ ]
l ps	22	0	1		1	0	1 0	101
l ps	21	0	1	.   0	1	0	0	101
l s	22	0	1	.   0	1	0	0	0
l s	22	0	1	.   0	1	0	0	0
ps	21	0	1	.   0	1	0	0	0
ps	23	0	1	.   0	1	0	0	0
ps	24	0	1	.   0	0	0	0	0
ps	23	0	1	.   0	0	0	0	0
ps	21	0	1	.   0	0	0	1	0
s	21	0	1	.   0	1	0	0	0
ps	21	0	1	.   0	1	0	0	0
s	23	0	1	.   0	1	0	0	0
ps	22	1	0	)   0	1	0	0	0
ps	21	1	0	)   0	1	0	0	0
ps ps	24	0	1	.   0	0	0	0	0
j s	21	0	1	.   0	1	0	0	0
j s	21	0	1	.   0	1	0	0	0
l nc	21	i n	1 1	ĺΛ	iΛ	ĺΛ	ĺΛ	i n i

Ce code transforme la variable "satisfaction" en une variable factorielle avec deux niveaux ("s" pour satisfait et "ps" pour pas satisfait). Il utilise une boucle pour parcourir chaque élément de la colonne "satisfaction" et applique la transformation en fonction de la valeur de la satisfaction. La fonction as.factor() est ensuite utilisée pour convertir la variable en un facteur.

Ensuite, la fonction subset() est utilisée pour sélectionner certaines colonnes spécifiques du dataframe "data" et les stocker dans un nouveau dataframe "data.actif" pour une analyse

#### d) Diviser les données

```
#données actifs
data.actif \leftarrow subset(data, select = c(1, 2, 12,13,14,15))
#Now our data is ready to be treated like she deserve
# exploration des données
install.packages("psych")
library(psych)
# présentation graphique des données
pairs.panels(data.actif[,-1])
# <u>Diviser les données</u> en un ensemble <u>d'apprentissage</u> et un ensemble de test (70-30)
library(caTools)
set.seed(123)
split = sample.split(data.actif$satisfaction, SplitRatio = 0.7)
train = subset(data.actif, split == TRUE)
train
str(train)
test = subset(data.actif, split == FALSE)
```

```
> head(test)
   satisfaction age DR QR QSA PR
2
              ps
                   21
                       4
                          1
                               1
                                   1
6
                   23
                       1
                           5
                               6
                                   4
              ps
7
              ps
                       5
                          4
                               9
                   24
                                   6
                         1
11
                   21
                       1
                               5
                                   7
              ps
                       5
                               5
                                   7
15
                   24
                          4
              ps
                       2
                           5
                               5
                                   5
20
                   20
               S
```

```
train = subset(data.actif, split == TRUE)
> head(train)
  satisfaction age DR QR QSA PR
                            4 6 4 6
                   ps 22
s 22
                 ps 21 1 6
ps 23 3 6
                                              6
9
> str(train)
'data.frame':
 'data.frame': 345 obs. of 6 variables:

$ satisfaction: Factor w/ 2 levels "ps","s": 1 2 2 1 1 1 2 2 1 1 ...

$ age : num 22 22 22 21 23 21 21 23 22 21 ...
 $ DR
$ QR
                       : num 4 4 3 1 3 1 3 4 4 4 ...
: num 6 6 7 6 6 3 7 5 2 6 ...
                                7 4 8 5 6 3 5 5 2 7 ...
"7" "8" "8" "6" ...
 $ QSA
 $ PR
                       : chr
> test = subset(data.actif, split == FALSE)
```

Ce code utilise la fonction sample.split du package caTools pour diviser le jeu de données data.actif en deux sous-ensembles : un ensemble d'apprentissage train et un ensemble de test test. La proportion des données dans l'ensemble de formation est fixée à 70% à l'aide du paramètre SplitRatio = 0.7. La fonction set.seed(123) est utilisée pour reproduire les mêmes résultats à chaque exécution du code. La fonction subset() est utilisée pour extraire les données correspondant à l'ensemble d'apprentissage et à l'ensemble de test. La fonction head() est utilisée pour afficher les 6 premières lignes de chaque ensemble de données, et la fonction str() est utilisée pour afficher la structure de l'ensemble d'apprentissage train.

## 6. APPLIQUER L'ANALYSE DISCRIMINANTE

1) Choix et justification de la méthode ADL pas à pas :

Nous avons choisi la méthode ADL (Analyse discriminante linéaire) pour notre analyse discriminante car elle est l'une des méthodes les plus couramment utilisées. Elle permet de séparer les groupes de données en utilisant des combinaisons linéaires des variables explicatives.

Pour analyser les données, nous avons utilisé la méthode ADL pas à pas, qui est une approche itérative qui sélectionne de manière séquentielle les variables explicatives les plus importantes pour construire le modèle final.

Dans le cadre de notre analyse de données de restauration, nous avons voulu comprendre les facteurs qui impactent plus sur la satisfaction des étudiants.

En résumé, la méthode ADL pas à pas de l'analyse discriminante s'est avérée être une méthode efficace pour identifier les variables les plus importantes pour discriminer entre les groupes et elle a été particulièrement utile pour notre cas.

### 2) Les résultats

```
# Appliquer l'analyse discriminante
library(klaR)
# Exécuter la sélection de variables stepwise
model <- stepclass(satisfaction ~ ., data = na.omit(train), direction = "backward", method = "lda")
# Afficher les var|iables sélectionnées
model
```

```
> # Afficher les variables sélectionnées
> model
method : lda
final model : satisfaction ~ age + DR + QR + QSA
<environment: 0x00000131486ee7b0>

correctness rate = 0.5306
> |
```

Ce code utilise la méthode de sélection de variables stepwise backward pour choisir les variables les plus importantes pour prédire la variable de réponse "satisfaction" à partir des autres variables du jeu de données d'entraînement. La fonction stepclass() de la bibliothèque MASS est utilisée pour effectuer cette sélection, en utilisant le modèle d'analyse discriminante linéaire (LDA). La fonction renvoie les variables sélectionnées pour le modèle final.

#### 7. INTERPRETATION DES RESULTATS:

Le résultat de la fonction montre que la méthode a été exécutée avec succès avec une amélioration de performance supérieure à 5% à chaque étape de la sélection. Le modèle final sélectionné contient 4 variables explicative, "age", "DR", "QR" et "QSA". Cela veut dire que l'entreprise BETA doit agir sur La qualité des repas et la qualité de services des agents pour améliorer la satisfaction des étudiants. D'autre part elle doit comprendre que les étudiants anciens sont moins satisfaits que les autre donc elle doit penser à donner des distinctions pour améliorer leurs satisfactions. La performance de ce modèle final est estimée à 53,06% de taux de classification correcte sur les données d'entraînement.

#### 8. CONCLUSION:

Cependant, il est important de souligner que notre modèle n'est qu'une approximation de la réalité, et que d'autres facteurs peuvent influencer la satisfaction des étudiants qui n'ont pas été pris en compte dans cette analyse. Par conséquent, nous recommandons une approche plus globale pour comprendre les attentes et les besoins des étudiants, en incluant des enquêtes de satisfaction et d'autres formes de feedback des étudiants.

En fin de compte, cette analyse discriminante peut être un outil utile pour BETA Restauration pour prédire la satisfaction des étudiants et améliorer leur expérience. Cependant, il ne doit pas être utilisé comme la seule méthode pour évaluer la satisfaction des étudiants et des méthodes supplémentaires doivent être utilisées pour une meilleure compréhension des attentes des étudiants