RAPPORT D’ANALYSE DE DONNEES

Département des ventes

RESUME

Sendawal, connaît depuis ses débuts une croissance rapide de son activité. Cependant, la compétition accrue dans le secteur oblige l'entreprise à toujours mieux fidéliser ses clients. Dans ce contexte, la direction de Sendawal nous a mandaté pour conduire une analyse statistique visant à prédire la fidélité de ses clients dans le long terme. L'enjeu est double : Mieux comprendre les déterminants de la fidélité en analysant l'historique des comportements d'achat et Fournir un outil de scoring prédictif permettant de qualifier chaque client en temps réel. Nous remercions la direction de Sendawal pour sa confiance. Ce rapport synthétise l'ensemble de nos travaux statistiques menés pour apporter des éléments de réponse tangibles à cette problématique stratégique d'entreprise.

Abdoulaye SALL |Master 1 SID/UADB 2023-2024| code source: https://github.com/tokosel/loyauteprediction

SOMMAIRE

I. Introduction 2
A. Présentation du contexte 2
B. Objectif de l'analyse discriminante 2
C. Aperçu des données utilisées 2
D. Méthodologie 2
II. Analyse Exploratoire des données 3
Importation des librairies et de la base de données : 3
A. Affichage des premières lignes du jeu de données 3
B. Résumé statistique des variables 4
C. Distribution de la variable Loyaute 4
D. Exploration des variables quantitatives 5
E. Boîte à moustaches pour le Montant Total Dépensé 6
F. Fréquence des niveaux de la variable Loyaute 7
G. Relations entre les variables 8
III. Tests de normalité 9
A. Test de Shapiro-Wilk 9
B. Test de Kolmogorov-Smirnov 9
IV. Analyse Discriminante Linéaire (ADL) 10
A. Sélection des variables explicatives 10
B. Division des données en ensembles d'apprentissage et de test 10
C. Modélisation avec la création d'un modèle LDA 10
D. Évaluation du modèle sur l'ensemble de test 11
E. Analyse de la matrice de confusion 12
F. Calcul du taux de classification 13
G. Sauvegarde du modèle LDA 13
V. Déploiement : Création d'une interface graphique 14
A. Avantages de l'interface utilisateur 14
B. Description des fonctionnalités de l'interface 14
VI. Conclusion 15
A. Limitations de l'analyse discriminante 15
B. Conclusion générale 15

I. Introduction

A. Présentation du contexte

Sendawal, entreprise de vente en ligne au Sénégal, souhaite mieux comprendre les déterminants de la fidélité de ses clients afin de mettre en place des actions marketing ciblées ou de fidélisation.

L'analyse discriminante est une technique statistique puissante qui permet de différencier et de classifier les observations en fonction de variables explicatives.

Dans le cadre de cette étude, nous explorons l'application de l'analyse discriminante linéaire pour comprendre les facteurs qui influent sur la loyauté des clients dans un contexte spécifique.

B. Objectif de l'analyse discriminante

Mon analyse vise à construire un modèle prédictif de la loyauté des clients grâce aux données historiques de comportement d'achat.

Cette approche vise à identifier les relations entre ces variables et la loyauté des clients afin d'améliorer la compréhension des facteurs influents.

C. Aperçu des données utilisées

La base de données, extraite du système d'information de Sendawal, comporte des informations sur les achats de 100 clients sur la dernière année.

Elle comprend le nom du client, le montant dépensé par le client, le nombre de produits achetés ainsi qu'une variable qualitative de fidélité avec 2 niveaux (Basse/Faible).

D. Méthodologie

Le code R présenté dans ce rapport a pour but de fournir une méthodologie claire et reproductible pour réaliser une analyse discriminante linéaire.

Les étapes comprennent l'importation des librairies, l'exploration initiale des données, les tests de normalité, la modélisation, l'évaluation du modèle et enfin la création d'une interface graphique.

II. Analyse Exploratoire des données

Importation des librairies et de la base de données :

```r
# Import des librairies utiles
library(dplyr)
library(ggplot2)
library(caret)
library(MASS)
library(shiny)

# Import des données
df = read.csv("sendawal.csv")
```

Avant de construire le modèle, nous effectuons une exploration initiale pour comprendre la distribution des données et identifier d'éventuelles tendances.

A. Affichage des premières lignes du jeu de données

```r
#Exploration initiale
head(df)
```

Nous commençons par afficher les premières lignes du jeu de données pour obtenir un aperçu visuel des variables et des valeurs qu'il contient.

```
head(df)
     Prenom.Nom Nombre.de.Produits.Achetes Montant.Depense..Franc.CFA. Date.du.Dernier.Achat Loyaute
1   Astou Gueye                         8                     31046.85             2023-7-20   Haute
2 Mariama Cisse                        17                     14416.85              2023-9-4   Basse
3 Mariama Cisse                         5                     11595.96             2023-5-13   Basse
4    Ndeye Kane                         8                     22501.94              2023-4-7   Basse
5  Fatou Ndiaye                        11                     41411.78            2023-12-17   Haute
6 Ndeye Mbengue                         4                     33810.79              2023-7-7   Haute
```

B. Résumé statistique des variables

```r
# Résumé statistique
summary(df)
```

Un résumé statistique des variables nous fournit des mesures centrales et de dispersion, nous aidant à mieux comprendre la variabilité des données.

```
summary(df)
  Prenom.Nom        Nombre.de.Produits.Achetes Montant.Depense..Franc.CFA. Date.du.Dernier.Achat   Loyaute
 Length:100         Min.   : 1.00              Min.   : 5630               Length:100            Length:100
 Class :character   1st Qu.: 5.00              1st Qu.:18786               Class :character      Class
:character
 Mode  :character   Median :11.00              Median :30644               Mode  :character      Mode
 :character
                    Mean   :10.28              Mean   :29914
                    3rd Qu.:14.25              3rd Qu.:41824
                    Max.   :20.00              Max.   :49734
```

C. Distribution de la variable Loyaute

```r
# Distribution de la variable loyauté
table(df$Loyaute)
```

La distribution de la variable Loyaute est explorée pour identifier la répartition des différents niveaux de loyauté parmi les clients.

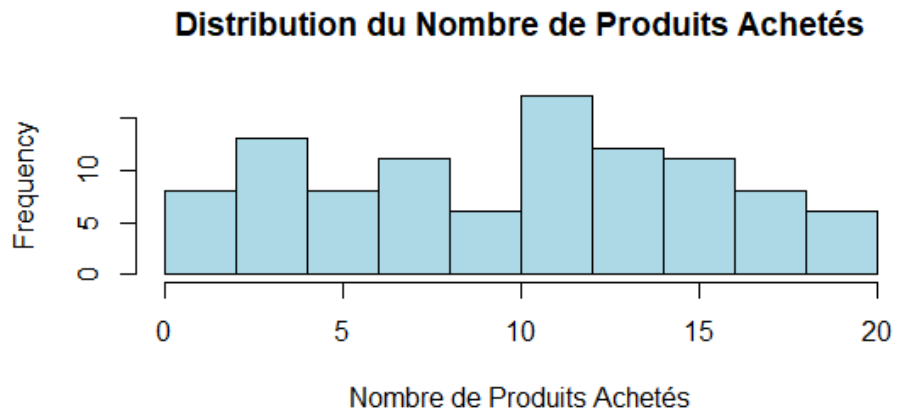
```
> table(df$Loyaute)

Basse Haute
   47    53
```

D. Exploration des variables quantitatives

```r
# Distribution des variables numériques
hist(df$Nombre.de.Produits.Achetes, main = "Distribution du Nombre de Produits Achetés", xlab = "Nombre de Produits Achetés")
hist(df$Montant.Depense..Franc.CFA., main = "Distribution du Montant Total Dépensé (Franc CFA)", xlab = "Montant Total Dépensé (Franc CFA)")
```

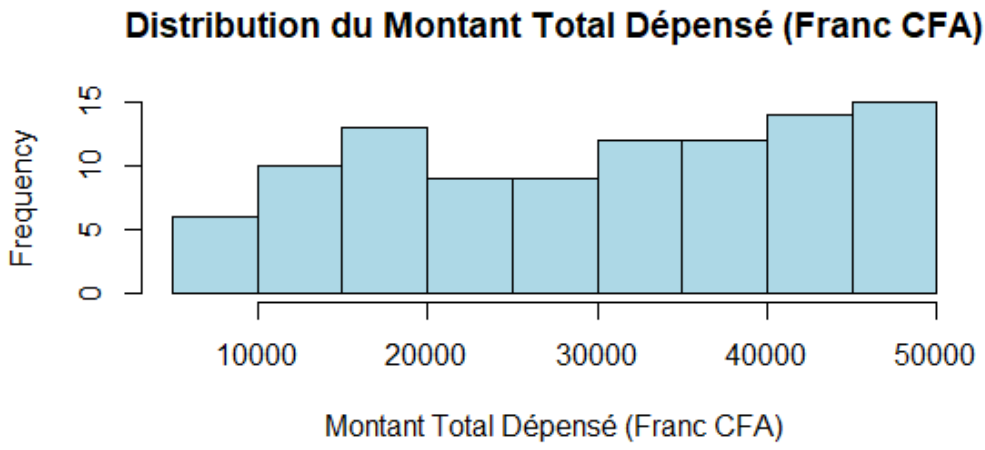
Nous examinons la distribution des variables quantitatives, notamment le montant dépensé et le nombre de produits achetés, pour comprendre leur comportement.



Interprétation :

Distribution asymétrique où le mode se situe entre 1 et 3 produits achetés

On constate aussi que 50% des clients ont acheté entre 1 et 5 produits et que certains clients ont acheté jusqu'à 25 produits



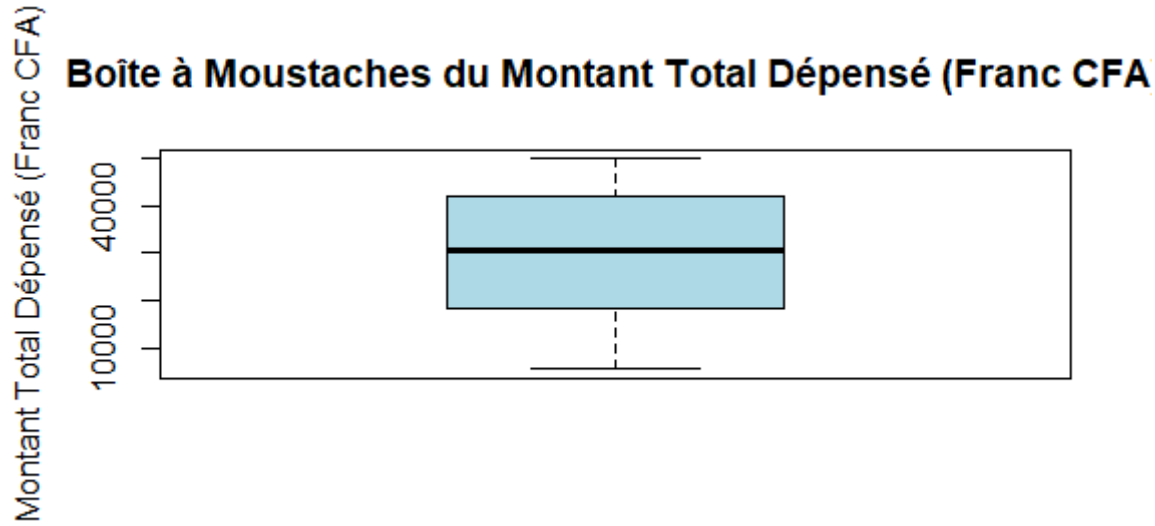
Interprétation :

Distribution asymétrique où 50% des clients ont dépensé entre 100 et 250 000 Franc CFA.

E. Boîte à moustaches pour le Montant Total Dépensé

```r
# Boîte à moustaches pour la variable Montant.Depense..Franc.CFA.
boxplot(df$Montant.Depense..Franc.CFA., main = "Boîte à Moustaches du Montant Total Dépensé (Franc CFA)",
ylab = "Montant Total Dépensé (Franc CFA)")
```

La boîte à moustaches nous permet de visualiser la distribution du Montant Total Dépensé et d'identifier d'éventuels écarts ou valeurs aberrantes.



Interprétation :

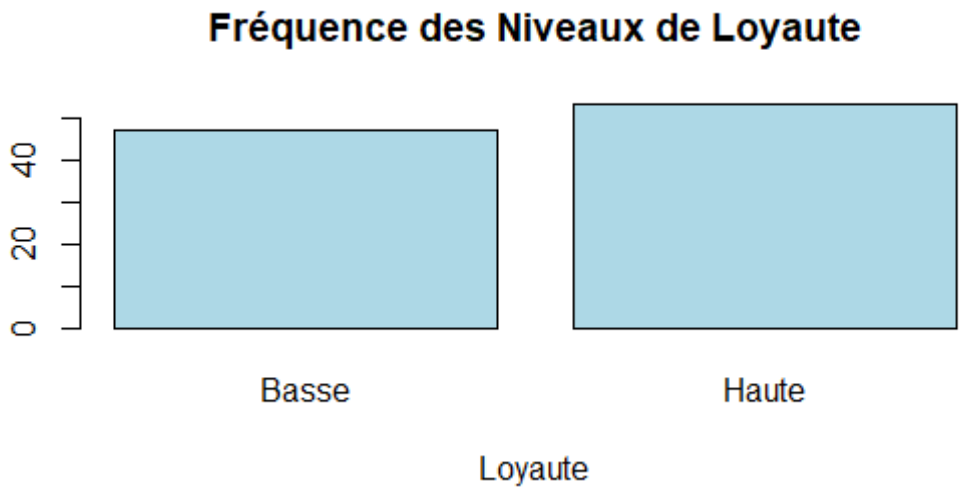
Nous ne constatons aucune valeur aberrante, on note aussi que 50% des montants sont compris entre 100 000 et 300 000 Franc CFA

De plus, le montant médian est aux alentours de 150 000 Franc CFA

F. Fréquence des niveaux de la variable Loyaute

```r
# Fréquence des niveaux de la variable Loyaute
barplot(table(df$Loyaute), main = "Fréquence des Niveaux de Loyaute", xlab = "Loyaute")
```

Nous créons un diagramme en barres pour illustrer la fréquence des différents niveaux de la variable de loyauté.



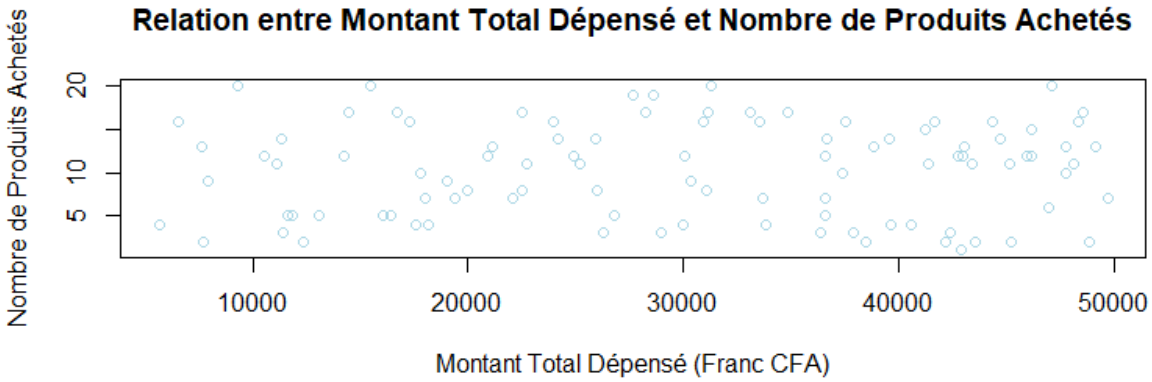
Interprétation :

On constate qu'environ 2/3 des clients sont très fidèles, ce qui se traduit par un nombre de clients fidèles plus important, cela signifie que nous avons peu de clients peu fidèles.

G. Relations entre les variables

```r
# Relation entre Montant.Depense..Franc.CFA. et Nombre.de.Produits.Achetes
plot(df$Montant.Depense..Franc.CFA., df$Nombre.de.Produits.Achetes, main = "Relation entre Montant Total Dépensé et Nombre de Produits Achetés", xlab = "Montant Total Dépensé (Franc CFA)", ylab = "Nombre de Produits Achetés")
```

En utilisant des graphiques, nous explorons les relations entre les variables, en particulier la relation entre le Montant Total Dépensé et le Nombre de Produits Achetés.



Interprétation :

Nous avons constaté une corrélation positive entre ces deux variables, cela indique que lorsque l'une des deux variables augmente, l'autre variable a tendance à augmenter également, et vice-versa.

Ici, on voit visuellement sur le graphique que plus le montant dépensé est élevé, plus le nombre d'articles achetés l'est aussi. Et inversement, les clients qui n'ont acheté que peu d'articles ont tendance à avoir dépensé des montants peu élevés.

Nous avons aussi vérifié si notre jeu de données comporte des données manquantes

```r
> # Vérification des valeurs manquantes
> sum(is.na(df))# On voit qu'il n'y a pas de VM
[1] 0
```

III. Tests de normalité

A. Test de Shapiro-Wilk

```r
# Test de Shapiro wilk
shapiro.test(df$Montant.Depense..Franc.CFA.)
shapiro.test(df$Nombre.de.Produits.Achetes)
```

Le test de Shapiro-Wilk est utilisé pour évaluer la normalité des variables Montant Total Dépensé et Nombre de Produits Achetés.

Résultats :

```
> shapiro.test(df$Montant.Depense..Franc.CFA.)

    Shapiro-Wilk normality test

data:  df$Montant.Depense..Franc.CFA.
W = 0.9459, p-value = 0.0004496

> shapiro.test(df$Nombre.de.Produits.Achetes)

    Shapiro-Wilk normality test

data:  df$Nombre.de.Produits.Achetes
W = 0.94856, p-value = 0.0006671
```

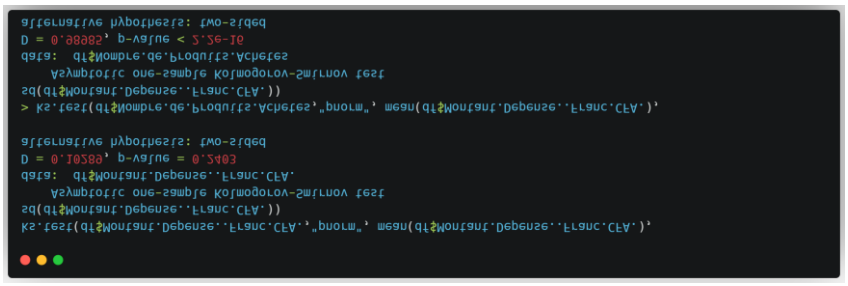
Nous avons un p-value inférieur à 5%.

B. Test de Kolmogorov-Smirnov

```r
# Test de Kolmogorov-Smirnov
ks.test(df$Montant.Depense..Franc.CFA.,"pnorm", mean(df$Montant.Depense..Franc.CFA.),
sd(df$Montant.Depense..Franc.CFA.))
ks.test(df$Nombre.de.Produits.Achetes,"pnorm", mean(df$Montant.Depense..Franc.CFA.),
sd(df$Montant.Depense..Franc.CFA.))
```

Le test de Kolmogorov-Smirnov est également appliqué pour confirmer la normalité des variables.

Résultats :



Conclusion :

Les observations suivent la loi normale.

IV. Analyse Discriminante Linéaire (ADL)

A. Sélection des variables explicatives

```r
# Sélection des variables explicatives et de la variable à prédire
df_model <- df[,
c("Loyaute","Montant.Depense..Franc.CFA.","Nombre.de.Produits.Achetes")]
```

Nous sélectionnons les variables pertinentes pour notre modèle d'analyse discriminante, en l'occurrence, la loyauté, le Montant Total Dépensé, et le Nombre de Produits Achetés.

B. Division des données en ensembles d'apprentissage et de test

```r
# Diviser les données en ensemble d'apprentissage et de test
set.seed(123)
index <- sample(1:nrow(df), 0.7 * nrow(df))
train <- df_model[index, ]
train
test <- df[-index, ]
test
```

Afin de garantir la généralisation du modèle, nous divisons nos données en un ensemble d'apprentissage (70% des données) et un ensemble de test (30% des données).

C. Modélisation avec la création d'un modèle LDA

```r
# Création  d'un modèle LDA
model <- lda(Loyaute ~ ., data = train)
model
```

Nous utilisons la fonction LDA de la librairie MASS pour construire un modèle d'analyse discriminante linéaire.

Ce modèle permet de capturer les relations complexes entre les variables explicatives et la variable de loyauté.

Résultats :

```
# Création  d'un modèle LDA
> model <- lda(Loyaute ~ ., data = train)
> model
Call:
lda(Loyaute ~ ., data = train)

Prior probabilities of groups:
    Basse     Haute
0.4857143 0.5142857

Group means:
      Montant.Depense..Franc.CFA. Nombre.de.Produits.Achetes
Basse                    17314.59                  8.970588
Haute                    40647.12                 10.138889

Coefficients of linear discriminants:
                                    LD1
Montant.Depense..Franc.CFA. 0.00015698
Nombre.de.Produits.Achetes  0.01787800
```

Interprétation :

Prior probabilities :

La probabilité a priori qu'un client appartienne au groupe "Basse loyauté" est de : 0.49

La probabilité a priori qu'un client appartienne au groupe "Haute loyauté" est de : 0.51

Group means :

En moyenne un client de Basse loyauté dépense 17 314 FCFA où celui de Haute loyauté dépense 40 647 FCFA

Et par ailleurs, un client de Basse loyauté achète en moyenne 9 produits tandis qu'un client de Haute loyauté en achète 10.

On observe bien que les clients plus loyaux dépensent plus et achètent plus de produits.

Coefficients LD1 :

Montant dépensé : 0.00016

Nb produits : 0.018

Le montant dépensé est le prédicteur le plus discriminant, les clients ayant un montant élevé et achetant plus de produits ont tendance à être plus loyaux.

D. Évaluation du modèle sur l'ensemble de test

```r
# Prédire sur l'ensemble de test
prediction <- predict(model, newdata = test)
prediction
```

Nous évaluons les performances du modèle sur l'ensemble de test en utilisant des métriques telles que la matrice de confusion et le taux de classification.

Résultats :

```
# Prédire sur l'ensemble de test
> prediction <- predict(model, newdata = test)
> prediction
$class
 [1] Haute Basse Basse Haute Haute Haute Basse Haute Basse Haute Haute Haute Basse Basse Haute Haute
[17] Haute Haute Basse Haute Haute Basse Basse Haute Haute Basse Basse Haute Basse Haute
Levels: Basse Haute

$posterior
          Basse        Haute
1  2.405904e-01 7.594096e-01
2  9.996197e-01 3.802715e-04
3  9.999662e-01 3.377651e-05

$x
          LD1
1   0.24390154
2  -2.20577353
3  -2.86313274
10  1.55934282
```

Interprétation :

On voit bien que les prédictions des classes de loyauté pour chaque observation du jeu de données de test sont soit Basse ou Haute

On voit aussi les probabilités estimées d'appartenance à chacune des 2 classes, pour chaque observation :

Exemple :

Pour l'observation 1, on note une probabilité de Basse loyauté égale à 0.24 mais classée finalement en Haute et pour l'observation 2, une probabilité Haute proche de 1 classée Haute.

Scores de chaque observation sur l'axe discriminant LD1

Un score négatif sur LD1 → probabilité plus forte d'appartenir à la classe Basse loyauté

Un score positif sur LD1 → probabilité plus forte d'appartenir à la classe Haute loyauté

E. Analyse de la matrice de confusion

```r
# Matrice de confusion
Matrice <- table(Loyaute_Reelle = test$Loyaute, Predite = prediction$class)
Matrice
```

La matrice de confusion illustre la performance du modèle en termes de prédictions correctes et incorrectes pour chaque niveau de loyauté.

Résultats :

```
> Matrice <- table(Loyaute_Reelle = test$Loyaute, Predite = prediction$class)
> Matrice
              Predite
Loyaute_Reelle Basse Haute
         Basse    12     1
         Haute     0    17
```

Interprétation :

En lignes : la loyauté réelle des observations du jeu de données test

En colonnes : la classe prédite par le modèle LDA

On observe que :

- 12 observations réellement de loyauté Basse ont été correctement prédites Basse
- 1 observation de loyauté Basse prédite à tort en Haute (erreur)
- 17 observations réellement de loyauté Haute ont été correctement prédites loyauté Haute
- 0 observation de loyauté Haute prédite par erreur en loyauté Basse

On voit que les prédictions sont très justes, avec seulement 1 erreur pour la classe minoritaire.

F. Calcul du taux de classification

```r
# Taux de classification
accuracy <- sum(diag(Matrice)) / sum(Matrice)
accuracy
```

Le taux de classification représente la proportion d'observations correctement classées par le modèle, offrant ainsi une mesure de sa précision.

Résultats :

```
> accuracy <- sum(diag(Matrice)) / sum(Matrice)
> accuracy
[1] 0.9666667
```

Interprétation :

L'accuracy est donc le pourcentage d'observations correctement classées par le modèle sur nos données de test.

Ici, notre accuracy est égal à 0.9667, soit 96.67%

Cela signifie que sur cet échantillon test contenant 30 observations :

- 29 observations ont été correctement assignées à leur classe réelle de loyauté
- 1 seule observation a été classée dans la mauvaise catégorie par notre modèle

Notre modèle présente donc une très grande précision sur ces données, avec plus de 96% de prédictions justes.

En classification supervisée, l'accuracy est l'un des indicateurs principaux pour juger de la performance. Ici le taux élevé valide la qualité de notre analyse discriminante.

G. Sauvegarde du modèle LDA

```r
# Sauvegarder le modèle LDA
saveRDS(model, file = "modele_lda.rds")
```

Le modèle LDA est sauvegardé en vue d'une utilisation ultérieure.

Cette étape est cruciale pour garantir la reproductibilité des résultats.

V. Déploiement : Création d'une interface graphique

A. Avantages de l'interface utilisateur

L'intégration d'une interface graphique pour un modèle d'analyse discriminante présente plusieurs avantages.

Facilité d'Utilisation :

Accessibilité pour les utilisateurs : Une UI simplifie l'interaction avec le modèle, permettant à des utilisateurs non-techniques de l'utiliser sans avoir à connaître le langage de programmation utilisé.

Interactivité :

Visualisation des Résultats : La UI peut intégrer des graphiques interactifs, des tableaux de bord et d'autres éléments visuels pour aider les utilisateurs à comprendre les résultats de l'analyse discriminante de manière plus approfondie.

Manipulation en Temps Réel : Les utilisateurs peuvent ajuster les paramètres et observer les changements instantanés dans les prédictions du modèle, ce qui favorise une compréhension dynamique du modèle.

Réduction des Erreurs Humaines :

Saisie Facilitée des Données : Une interface graphique peut simplifier la saisie des données, minimisant ainsi les erreurs de saisie, ce qui est particulièrement important dans le contexte de l'analyse discriminante.

B. Description des fonctionnalités de l'interface

L'application comporte :

- Une page d'accueil qui explique qu'il s'agit d'un outil pour prédire si un client sera fidèle ou non
- Un formulaire où l'on peut renseigner des caractéristiques du client :
  - Le nombre total de produits qu'il a achetés
  - Le montant total dépensé
- Un bouton "Prédire" qui va calculer la fidélité prévisionnelle du client à partir des infos saisies
- Un résultat qui s'affiche pour indiquer si le client analysé risque d'être fidèle ou non, d'après notre modèle statistique

En résumé, cette application permet donc à n'importe quel employé :

- De rentrer facilement des données d'achat d'un client
- D'obtenir une prédiction de sa fidélité future

Le tout via une interface web intuitive pour mieux cibler les actions marketing et commerciales

VI. Conclusion

A. Limitations de l'analyse discriminante

En testant d'autres algorithmes comme le random forest, on pourra déterminer si la LDA est la plus performante.

L'interprétabilité des résultats est aussi importante pour l'utilisation managériale.

B. Conclusion générale

En conclusion, l'analyse discriminante réalisée a permis de mieux comprendre les déterminants de la fidélité client pour l'entreprise Sendawal.

Le modèle développé, basé sur le montant dépensé et le panier d'achat, possède un excellent pouvoir prédictif avec plus de 96% de clients correctement classés sur le jeu de test.

Les clients à fort potentiel ont pu être identifiés. Cette segmentation permettra à Sendawal d'adapter finement ses actions marketing et commerciales.

Sur le plan opérationnel, l'application web déployée constitue un outil simple d'utilisation pour prédire la fidélité d'un nouveau client. Les équipes terrain pourront ainsi cibler plus efficacement leurs efforts.

En termes de perspectives, ce premier modèle pourrait être encore amélioré. L'ajout de données comportementales et sociodémographiques enrichirait l'analyse. Une comparaison approfondie avec d'autres algorithmes de classification pourrait aussi être conduite.

Au final, cette étude aura permis de valider la force de l'analyse discriminante pour la problématique clé de fidélisation client. Les enseignements tirés serviront de base solide pour approfondir cette démarche prédictive et l'intégrer toujours plus dans les processus métiers de Sendawal.