`

Conservatoire national des arts et métiers (CNAM)

Plan du sujet du

Projet STA211-Entreposage et fouille de données

Nom du projet

---« Bank-Marketing » ---

Présenté par :

----« Elio-Bou-Serhal » ----

Professeur responsable :

---« Dr-Jihanne-Karameh » ---

Contents :

Figure :

Tableaux:

« La prédiction de campagnes de marketing téléphonique de la banque portugaise »

Introduction :

L'apprentissage automatique (Machine Learning : ML) est un sous-domaine de la fouille de données qui étudie les techniques automatiques pour apprendre à faire des prévisions précises en se basant sur des observations passées. Il utilise deux types de techniques :

1. L'apprentissage supervisé : (classification et régression), qui entraîne un modèle sur des données d'entrée et de sortie connues pour qu'il puisse prédire les sorties.
2. L'apprentissage non supervisé : (clustering), qui trouve des motifs cachés ou des structures intrinsèques dans les données d'entrée.

Ce qui nous intéresse dans ce projet, une base de donnée supervisée qui contient des informations sur **45 211 clients** potentiels d'une banque et **17 variables** enregistrées pour chaque observation qui comprennent des informations démographiques sur les clients, telles que leur âge, leur situation maritale, leur niveau d'éducation, leur emploi, leur situation de crédit (par exemple, s'ils ont des retards de paiement), leur situation de logement (s'ils ont un prêt hypothécaire ou non), ainsi que des informations sur la façon dont la banque a contacté les clients, la durée de la conversation, la fréquence des appels précédents, etc...
La variable cible binaire "y" indique si un client potentiel a souscrit à un dépôt à terme ou non. (Y=0 → client qui a souscrit à un dépôt, Y=1→ client n'a pas souscrit à un dépôt)

On peut remarquer que l'importance des caractéristiques et les conséquences peuvent varier d'un modèle à un autre. Ce projet vise à analyser la performance de 8 modèles en choisissant les paramètres optimaux afin d'évaluer l'approche proposée.

1. Naïve Bayes (NB),
2. k-Nearest Neighbors (KNN),
3. arbres de décision (DT),
4. réseaux de neurones artificiels (ANN),
5. machines à vecteurs de support (SVM),
6. Régression Linéaire (LR),
7. Foret Aléatoire (RF),
8. Règle de Décision unique (One-R algorithme)

Ces modèles ont été utilisés par les professeurs :

1. « Stéphane Cédric Koumetio », « Walid Cherif », « Hassan Silkan ». (reference article 1&2)
2. « Ianguo Che », «Sai Zhao », « Yongfan Li », « Kai Li ». (reference article 3)
3. « Yasemin Gultepe» , « Wisam Gwad», « Yuosra Aljamel », « Yossf Ahmed4».(refrence article 4)
4. « Mohammad Abu Tareq Rony », « Md. Mehedi Hassan », « Eshtiak Ahmed », « Asif Karim », « Sami Azam », « D. S. A. Aashiqur Reza » (reference article 5)

Ces modèles ont été utilisés par les professeurs suivants dans le cadre de leurs travaux précédents sur la prédiction de campagnes de marketing téléphonique de la banque portugaise.

Mots clés : Apprentissage supervisée et non supervisée, DT, RF, LR, NB, One-R-algorithm, KNN, ANN, SVM, Accuracy, AUC, prédiction des compagnes, performance.

Objectif :

Notre objectif est de construire 8 modèles prédictifs dans le but de déterminer les caractéristiques des clients les plus susceptibles de prendre un dépôt à terme, en utilisant les autres variables comme prédicteurs.

En se basant sur 5 différentes articles nous allons :

1. Synthétiser ces recherches et analyser les résultats obtenus dans chaque article.
2. Tirer des conclusions globales qui peuvent guider notre propre approche pour prédire les résultats des campagnes de marketing téléphonique de la banque portugaise en utilisant plusieurs mesures métriques :

La matrice de confusion qui fournit les informations suivantes :

1. Le nombre de vrais positifs (VP) : observations positives correctement classées
2. Le nombre de faux positifs (FP) : observations négatives incorrectement classées comme positives
3. Le nombre de vrais négatifs (VN) : observations négatives correctement classées
4. Le nombre de faux négatifs (FN) : observations positives incorrectement classées comme négatives.

Ces informations sont utilisées pour calculer diverses mesures de performance du modèle, telles que la précision, le rappel et le score F1.

La courbe de Roc qui mesure la capacité du modèle à distinguer les exemples positifs des exemples négatifs en traçant une courbe ROC représentant la sensibilité du modèle (vrais positifs) en fonction de la spécificité (faux positifs) pour différents seuils de classification.

Étape à faire :

ÉTAPE 1 :

1. Importer la base de données sur R-studio.

ÉTAPE 2 :

1. Faire une analyse descriptive de la base de données
2. Représenter graphiquement chaque variable

ÉTAPE 3 :

1. Etudier la corrélation entre toutes les variables en utilisant la matrice de corrélation pour les variables quantitatives et le t-test pour les variables qualitatives.
2. Tester la linéarité de la base de données.
3. Effectuer une analyse factorielle pour réduire la dimension des variables en utilisant plusieurs méthodes (kmeans, Boruta, Hcluster, PCA..).

ÉTAPE 4 :

1. Détecter s'il existe des valeurs manquantes et si non, on doit remplacer quelques valeurs par « NA » pour les visualiser et les imputer en utilisant différentes méthodes de visualisation (histogramme, pattern…) et d'imputation (simple, multiple…) pour pouvoir comparer la distribution des variables contenant ces valeurs manquantes avant et après imputation.
2. Eliminer toutes les valeurs manquantes.

ÉTAPE 5 :
1. Diviser la base de données en ensemble d'entraînement et ensemble de test pour tester les performances de chaque modèle.
2. Transformer la base de données mixte en base de données numérique.
3. Normaliser chaque partie individuellement (séparément).
4. Réduire le nombre des modalités des variables qualitatives.

ÉTAPE 6 :
1. Entraîner les 8 modèles différents en utilisant les données d'entraînement sur une base de données normalisée et sur une base de données non-normalisée.
2. Évaluer leur performance en utilisant la matrice de confusion et la courbe ROC.

ÉTAPE 7 :
1. Créer un tableau contenant tous les résultats pour comparer ces 8 modèles
2. Représenter ces 8 courbes dans un seul graphe.
3. Interpréter les résultats obtenus.
4. Indiquer les hyper paramètres du modèle le plus performant.
5. Tirer des conclusions sur les caractéristiques les plus importantes pour prédire si un client potentiel souscrira à un dépôt à terme.

Résumé des articles :

ARTICLE 1 :
Dans le premier article, **cinq** modèles différents ont été utilisés pour prédire les résultats des campagnes de marketing téléphonique de la banque portugaise **NB, LR, DT, ANN et SVM**. Les résultats ont montré que le meilleur modèle pour prédire ces campagnes était l'arbre de décision DT, suivi de près par le réseau de neurones artificiels ANN.

ARTICLE 2 :
Dans le deuxième article, **quatre** modèles différents ont été utilisés : **NB, DT, ANN et SVM**. Les résultats ont montré que le meilleur modèle pour prédire ces campagnes était la machine à vecteurs de support SVM, suivi de près par l'arbre de décision DT.

ARTICLE 3 :
Le troisième article a utilisé **un** modèle spécifique appelé : **t-SNE-SVM**, la technique de réduction de dimensionnalité t-SNE. Une fois les données réduites à une dimension inférieure à l'aide du t-SNE, le modèle SVM sera appliqué pour la classification des résultats des campagnes de marketing.

ARTICLE 4
Dans le quatrième article, **deux** modèles ont été utilisés : l'algorithme **One-R** et le **NB**. Les résultats ont indiqué que le meilleur modèle pour prédire les campagnes de marketing téléphonique était l'algorithme One-R.

ARTICLE 5 :

Dans le cinquième article, **cinq** modèles ont été utilisés : **LR, RF, SVM, KNN et ANN.** Les résultats ont montré que le meilleur modèle pour prédire les campagnes de marketing téléphonique était la régression logistique LR, suivie de près par l'arbre de décision DT.

Ces différents articles mettent en évidence l'utilisation de divers modèles pour prédire les résultats des campagnes de marketing téléphonique de la banque portugaise, c'est pour cela nous allons joindre toutes les méthodes dans un seul article pour comparer ces résultats obtenus et montrer l'importance des plusieurs approches et techniques pour améliorer l'efficacité des campagnes de marketing.

Explication Théorique :

Nous allons voir une petite explication sur les 8 modèles utilises :

Naïve Bayes (NB) :

Le modèle Naïve Bayes est basé sur le théorème de Bayes et suppose une indépendance conditionnelle entre les caractéristiques. Il calcule la probabilité qu'une instance appartienne à une classe en utilisant les probabilités a priori des classes et les probabilités conditionnelles des caractéristiques.

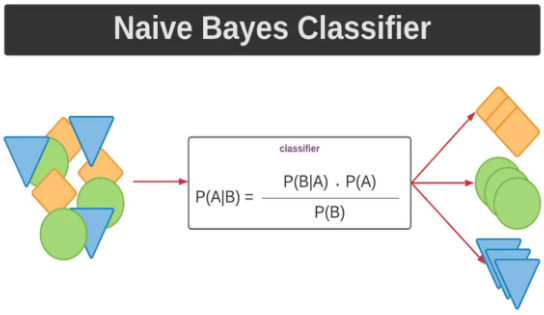


Figure 1: Naive Bayes theorem (source: By Bashir Alam)

k-Nearest Neighbors (KNN) :

L'algorithme k-Nearest Neighbors attribue une classe à une instance en fonction des classes des k voisins les plus proches. Il est utilisé pour une base de donnée non-supervisée mais ici on va l'utilisée pour une base de donnée supervisée. La distance entre les instances est généralement calculée à l'aide de la distance euclidienne. KNN est un modèle non paramétrique, ce qui signifie qu'il n'apprend pas explicitement de paramètres, mais mémorise simplement les données d'entraînement.

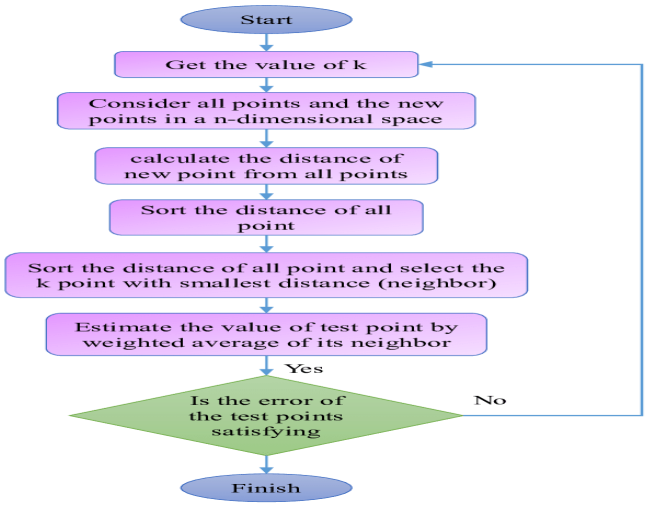


Figure 2: KNN algorithm (source: ResearchGate)

Arbres de décision (DT) :

Les arbres de décision sont des modèles représentés sous forme d'arbre, où chaque nœud représente une caractéristique et chaque feuille représente une classe. Le processus de construction d'un arbre de décision implique de sélectionner la caractéristique la plus informative à chaque nœud, afin de maximiser la séparation entre les classes. Les arbres de décision sont faciles à interpréter et peuvent gérer des données qualitatives et quantitatives.

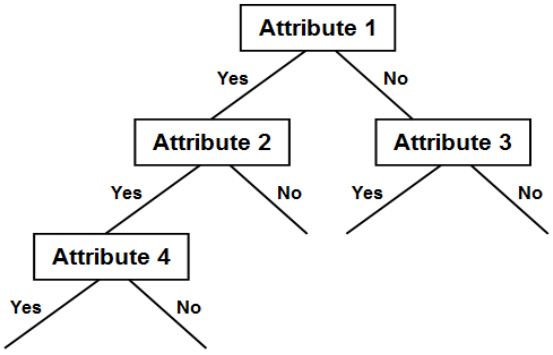


Figure 3: DT algorithm (source: ResearchGate)

Réseaux de neurones artificiels (ANN) :

Les réseaux de neurones artificiels sont des modèles qui imitent le fonctionnement du cerveau humain. Ils sont composés de couches de neurones interconnectés. Chaque neurone applique une fonction d'activation aux entrées pondérées pour générer une sortie. Les réseaux de neurones peuvent apprendre des représentations complexes des données en ajustant les poids des connexions lors de la phase d'apprentissage.

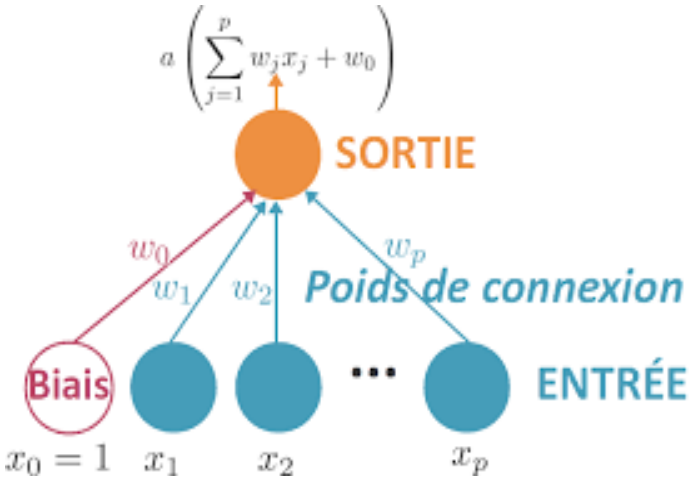


Figure 4: ANN algorithm (source : open classrooms)

Machines à vecteurs de support (SVM) :

Les machines à vecteurs de support sont des modèles qui trouvent un hyperplan optimal pour séparer les données en différentes classes. L'objectif est de maximiser la marge entre les classes les plus proches. SVM peut également utiliser des noyaux pour traiter des données non linéairement séparables en les transformant dans un espace de dimension supérieure.

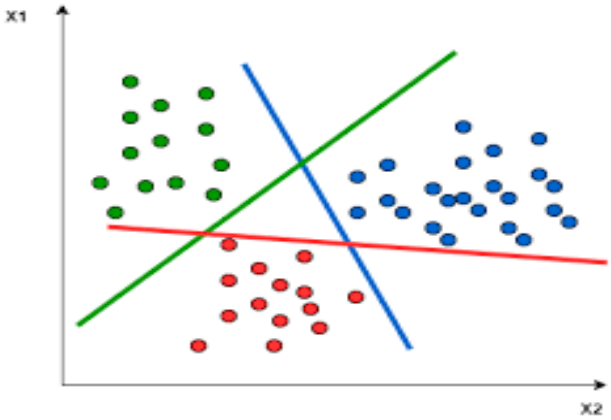


Figure 5: SVM theorem (source: Nada Belaidi)

Régression linéaire (LR) :

La régression linéaire est un modèle utilisé pour prédire une variable continue en fonction des caractéristiques. Elle cherche à trouver la meilleure relation linéaire entre les caractéristiques et la variable cible. La régression linéaire est basée sur l'hypothèse que la variable cible peut être approximée par une combinaison linéaire des caractéristiques.

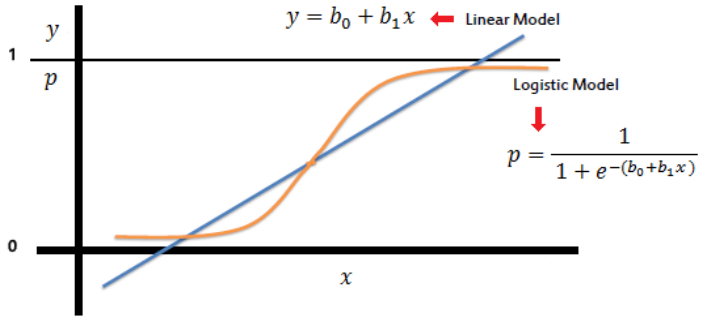


Figure 6: LR theorem (source: Dr said Sayyad)

Arbre aléatoire (RF) :

Les arbres aléatoires sont des ensembles d'arbres de décision construits à partir de sous-ensembles aléatoires des données d'entraînement et des caractéristiques. Chaque arbre est construit indépendamment, et la décision finale est prise par vote majoritaire. Les arbres aléatoires sont robustes et efficaces pour la classification et la régression.

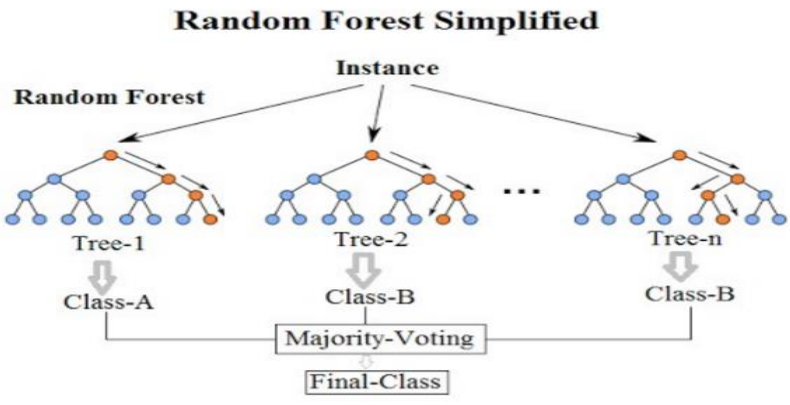


Figure 7: RF algorithm (source: Wikipedia)

Règle de décision unique (One-R algorithme) :

La règle de décision unique est un algorithme simple qui sélectionne la caractéristique la plus informative pour la classification. Il crée une règle de décision basée uniquement sur cette caractéristique. La règle de décision unique est facile à comprendre et peut être utilisée comme un point de référence rapide pour évaluer les performances des autres modèles.

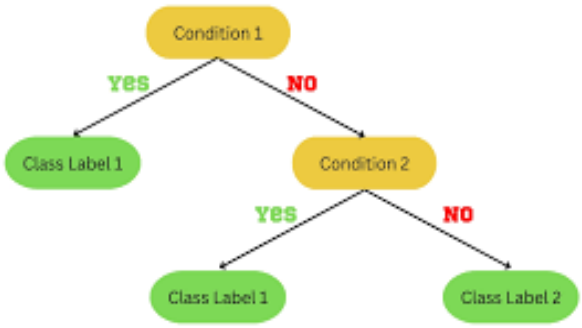


Figure 8: One-R algorithm (source: coding ninjas)

Nous allons voir une petite explication sur tous les modèles qu'on a déjà vu cette année pour traiter une base de données supervisée et non-supervisée :

Bagging (Bootstrap Aggregating) :

Le Bagging est une méthode d'ensemble qui vise à réduire la variance des modèles individuels en combinant les prédictions de plusieurs modèles. Il utilise des échantillons Bootstrap de l'ensemble de données d'entraînement pour entraîner des modèles similaires, mais légèrement différents les uns des autres. En agrégeant les prédictions de ces modèles, le Bagging améliore la stabilité et la performance globale par exemple l'algorithme des forêts aléatoires.

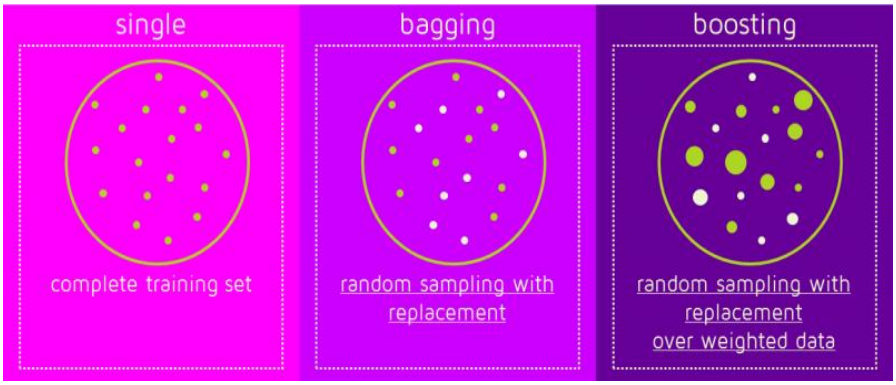


Figure 9: Bagging & Boosting theorems (source: QuantDare)

Boosting :

Le Boosting est une autre méthode d'ensemble qui combine plusieurs modèles faibles pour créer un modèle fort. Contrairement au Bagging, le Boosting se concentre sur la réduction de l'erreur de biais. Il entraîne une série de modèles itérativement, en accordant plus de poids aux instances mal prédites par les modèles précédents. Le Boosting crée un modèle final qui est une combinaison pondérée des modèles individuels, où les modèles les plus performants ont un poids plus élevé par exemple le Gradient Boosting.

L'analyse des données multivues :

C'est une approche qui vise à exploiter plusieurs sources de données hétérogènes et à les combiner pour obtenir une vision plus complète et plus précise d'un phénomène ou d'un système. L'objectif est d'extraire des connaissances et des modèles à partir de ces données multivues, en prenant en compte les interactions et les dépendances entre les différentes variables. L'analyse des données multivues peut être réalisée à l'aide de techniques telles que : ACP, ANN ou les modèles de graphe, permettant ainsi d'explorer les relations complexes et les interactions entre les différentes variables.

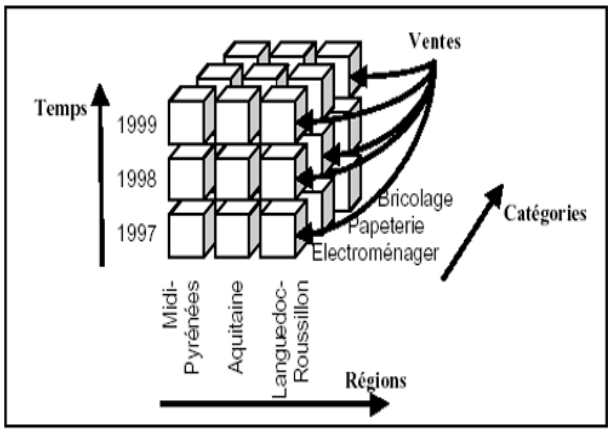


Figure 10: Principe d'Analyse des données multivues (source : univ-mlv.fr)

Le modèle de cartes auto-organisatrices (SOM) :

C'est une technique d'apprentissage non supervisé largement utilisée pour analyser des données multivues. Il s'agit d'un algorithme de réduction de dimensionnalité qui permet de représenter des données complexes dans un espace de dimension réduite, Le SOM est basé sur les neurones artificiels, où chaque neurone représente un point dans l'espace réduit. Cette technique permet de visualiser et d'explorer facilement les structures cachées dans les données multivues, ce qui peut être extrêmement utile pour la compréhension et l'interprétation des données.

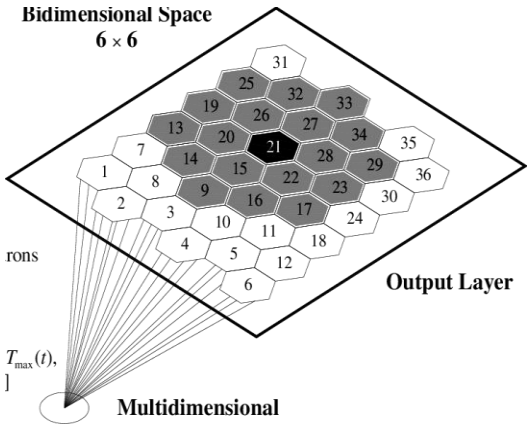


Figure 11: SOM algorithm (source: ResearchGate)

La règle d'association :

C'est un outil d'exploration de données qui vise à découvrir des relations intéressantes et fréquentes entre les différents éléments d'un ensemble de données. Elle est largement utilisée dans le domaine de l'apprentissage automatique et du data mining pour trouver des corrélations entre les items. Ces règles peuvent être utilisées pour prendre des décisions, recommander des produits ou des services. Il existe plusieurs algorithmes mais le plus couramment utilisé est l'algorithme Apriori, qui recherche les items fréquents dans les données et génère des règles à partir de ces items.

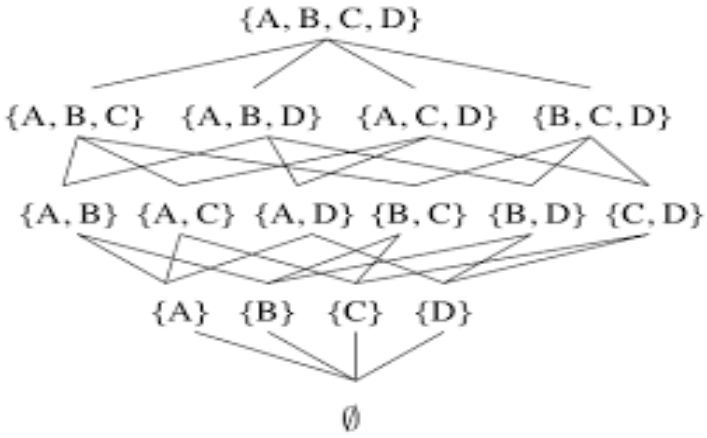


Figure 12: Principe de la Règle d'association (source : univ-mlv.fr)

Analyse descriptive :

La base de données est liée aux campagnes de marketing direct d'une institution bancaire portugaise, en se basant sur des appels téléphoniques afin de déterminer si le client souscrirait ou non à un dépôt. Voici un tableau montrant le type, le nom et la description de chaque variable :

Tableau 1 : Analyse Descriptive

	Variable	**Type de variable**	**Description de la variable**
Bank client data	age	Numérique	De 17 à 98
	job	Catégorique	"admin.","unknown","unemployed","management","housemaid", "entrepreneur","student","bluecollar","selfemployed","retired", "technician","services"
	marital	Catégorique	"married","divorced","single"
	education	Catégorique	"unknown","secondary","primary","tertiary"
	default	Catégorique	has personal loan? (binary: "yes","no")
	balance	Numérique	average yearly balance, in euros (numeric)
	housing	Catégorique	has housing loan? (binary: "yes","no")
	loan	Catégorique	has personal loan? (binary: "yes","no")
	contact	Catégorique	"unknown","telephone","cellular"

Related with the last contact of the current campaign	day	Numérique	last contact day of the month
	month	Catégorique	last contact month of year: "jan", "feb", "mar", ..., "nov", "dec"
	duration	Numérique	last contact duration, in seconds
Other attributes	campaign	Numérique	number of contacts performed during this campaign and for this client
	pdays	Numérique	number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)
	previous	Numérique	number of contacts performed before this campaign and for this client
	poutcome	Catégorique	outcome of the previous marketing campaign: "unknown","other","failure","success"
	y	Binaire	Has the client subscribed a term deposit? (binary: "yes","no")

En utilisant la fonction « describe» on peut prendre une idée générale sur les outils statistiques comme moyenne, fréquence et distribution:

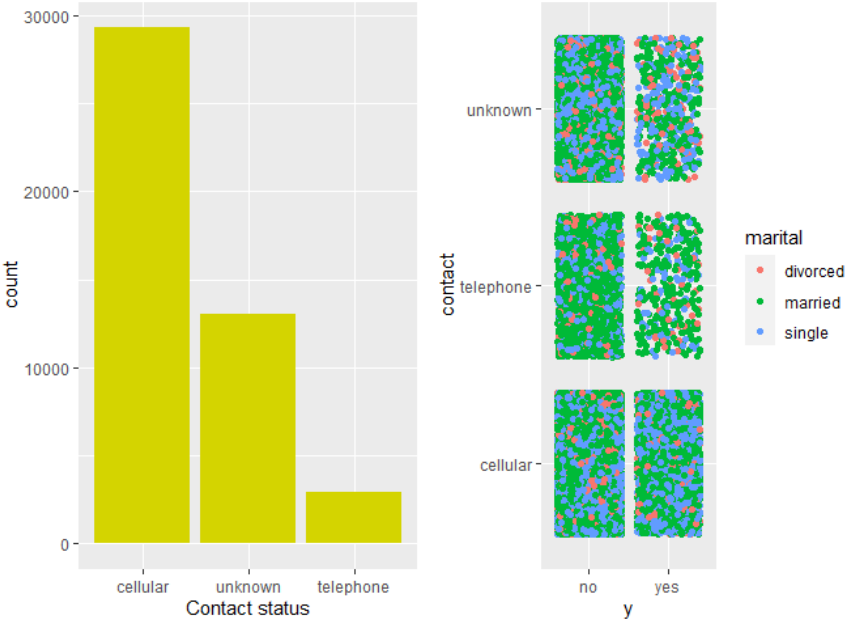
1. Âge : L'âge varie de 18 à 95 ans, avec une moyenne d'âge de 40,94 ans. La distribution est légèrement asymétrique positive, avec une différence moyenne de Gini (Gmd) de 11,87.
2. Job: Il y a 12 catégories d'emplois distinctes. Les emplois les plus courants sont "admin." (5 171 observations) et "ouvrier qualifié" (9 732 observations).
3. Marital: Il y a trois catégories de situation matrimoniale : "divorcé(e)", "marié(e)" et "célibataire".
4. La majorité des individus sont mariés (27 214 observations), suivis des célibataires (12 790 observations).
5. Education: Il y a quatre catégories d'éducation : "primaire", "secondaire", "tertiaire" et "inconnue", tel que Le niveau d'éducation le plus courant est le secondaire (23 202 observations).
6. Default: La majorité des clients (44 396 observations) n'ont pas de crédit en défaut.
7. Balance (Solde) : La Balance varie de -8 019 à 102 127, avec une moyenne de 1 362 tel que La distribution est asymétrique positive, avec une différence moyenne de Gini de 2 054.
8. Housing (Logement) : Environ 55,6 % des clients (25 130 observations) ont un prêt immobilier.
9. Loan (Prêt personnel) : Environ 16 % des clients (7 244 observations) ont un prêt personnel. Ce qui montre une base de donnée déséquilibrée (unbalanced data) puisque c'est la variable cible
10. Contact : Représente le type de communication utilisé, tel que la méthode de contact la plus courante est le téléphone portable (29 285 observations).
11. Day (Jour) : Le jour du mois (de 1 à 31) où le dernier contact a été effectué, tel que La valeur moyenne est de 15,81, ce qui indique une distribution approximativement uniforme.
12. Mois : Indique le mois où le dernier contact a été effectué, tel que le mois le plus courant est mai (13 766 observations).
13. Duration (Durée) : La durée du dernier contact en secondes, tel que la durée moyenne est de 258,2 secondes.
14. Campaign (Campagne) : Le nombre de contacts effectués lors de cette campagne pour le client, tel que le nombre moyen de contacts est de 2,764.
15. Pdays : Le nombre de jours écoulés depuis le dernier contact avec le client, tel que la valeur moyenne est de 40,2 mais une grande proportion de -1 indique que certains clients n'ont pas été contactés précédemment.
16. Previous : Le nombre de contacts effectués avant cette campagne pour le client, tel que le nombre moyen de contacts est de 0,5803

Représentation Graphiques :

L'utilisation combinée de ces différentes représentations graphiques offre une vision plus complète de la distribution des données et permet d'identifier les caractéristiques importantes de chaque variable. Notons que notre base de donnée ne contienne pas ni valeurs manquantes ni duplications.

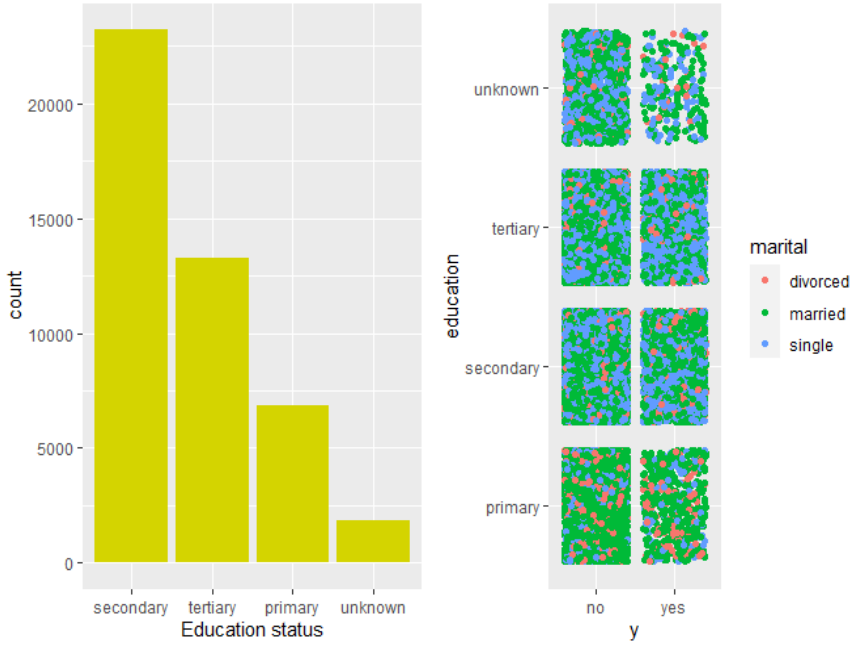
POUR LES VARIABLES QUALITATIVES : nous allons voir un diagramme en bar représentant la fréquence des différentes modalités, et un graphe de dispersion de la variable explicative qualitative par rapport à la variable cible.

POUR LES VARIABLES QUANTITATIVES : nous allons voir deux histogrammes représentants respectivement les effectifs et la densité de chaque variable quantitative, de plus nous allons voir la boîte à moustache de chacune.



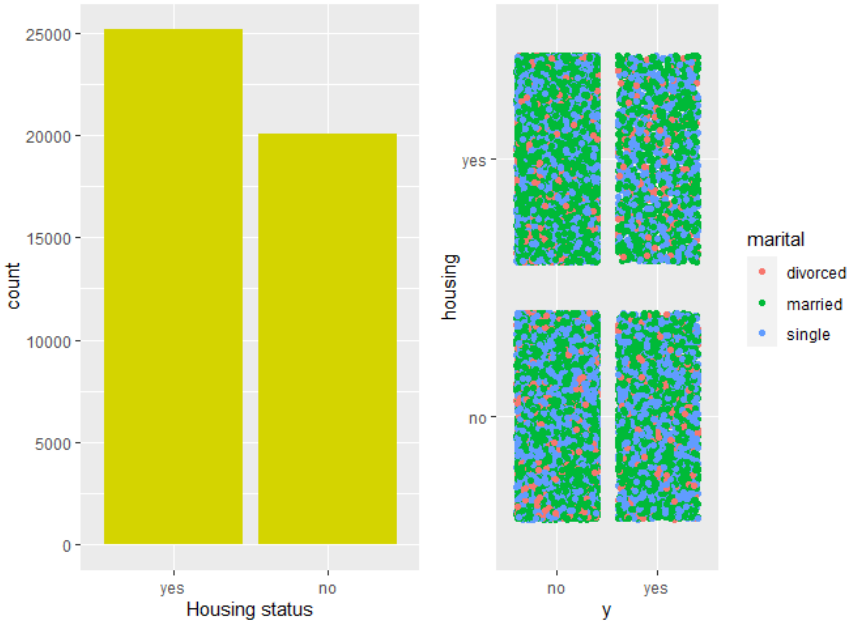
C'est un graphe montrant la fréquence par rapport aux 3 modalités de variable « contact » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 3 modalisées "unknown", "telephone""cellular" selon la variable « marital »

Figure 13: Représentation Graphique de la variable "count"



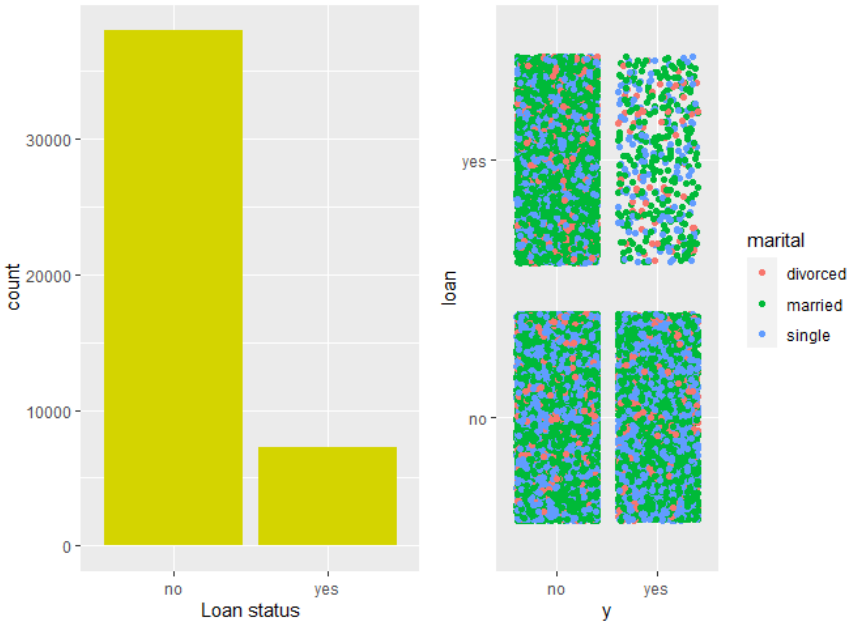
C'est un graphe montrant la fréquence par rapport aux 4 modalités de variable « éducation » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 4 modalisées : "unknown", "secondary", "primary" et "tertiary" selon la variable « marital »

Figure 14: Representation Graphique de la variable "education"



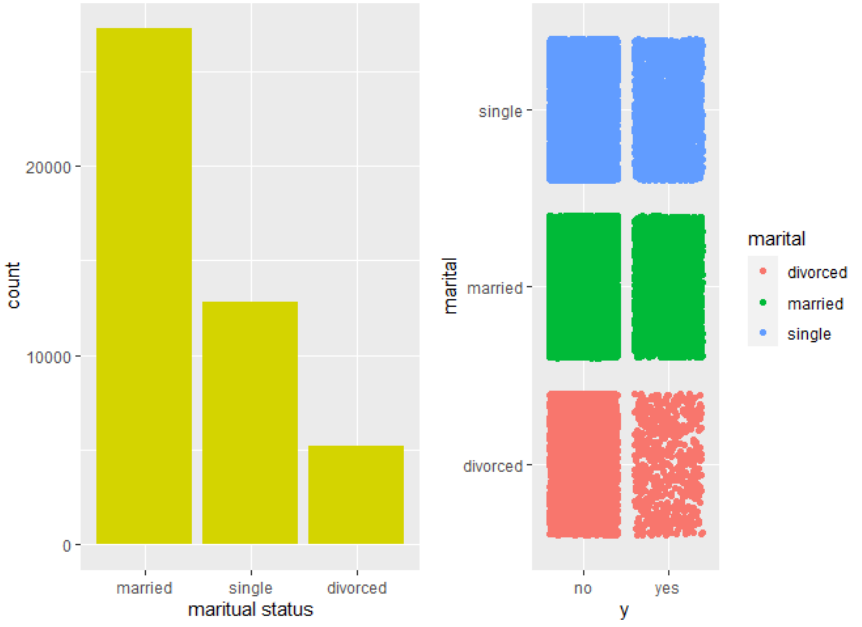
C’est un graphe montrant la fréquence par rapport aux 2 modalités de variable binaire « status » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 2 modalisées selon la variable « marital ».

Figure 15: Représentation Graphique de la variable "status"



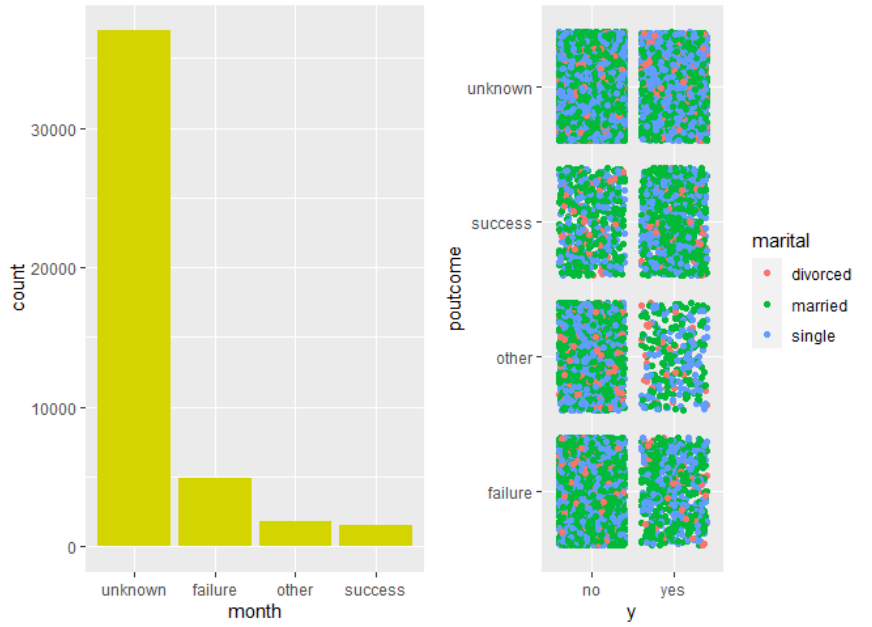
C’est un graphe montrant la fréquence par rapport aux 2 modalités de variable binaire « Loan» qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 2 modalisées selon la variable « marital »

Figure 16: Représentation Graphique de la variable "Loan"



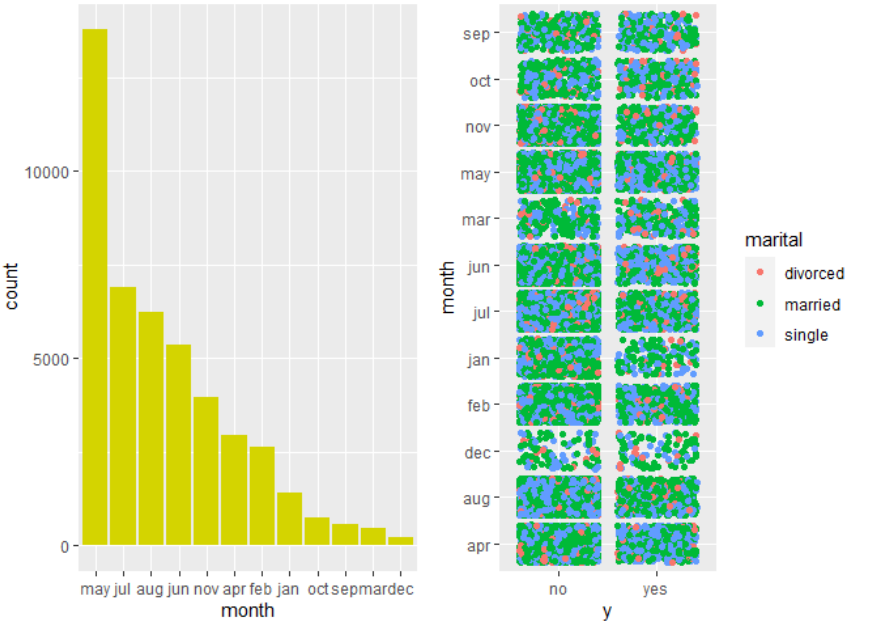
C’est un graphe montrant la fréquence par rapport aux 3 modalités de variable « marital » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 3 modalitées : "married", "divorced" et "single" selon la variable « marital » et « y »

Figure 17: Représentation graphique de la variable "marital"



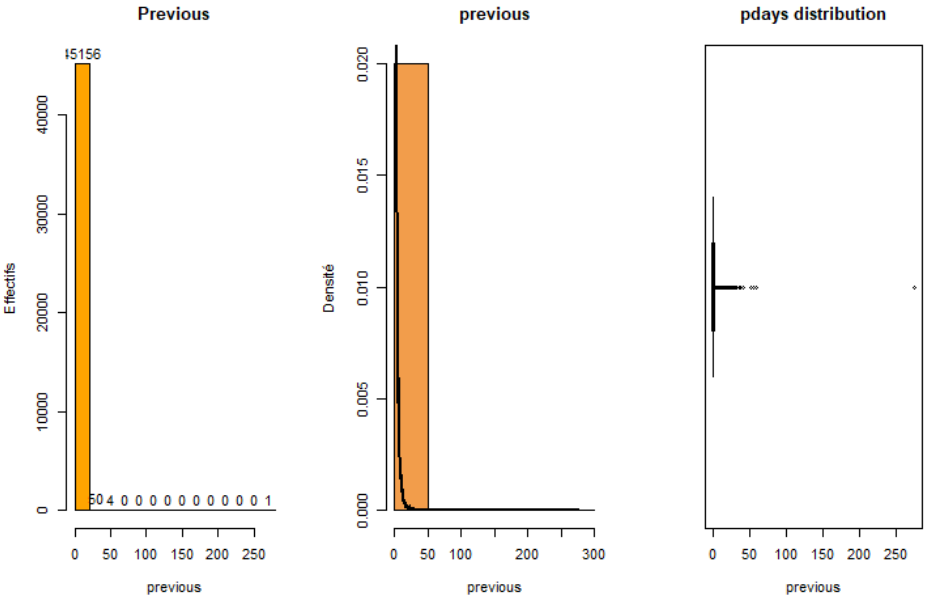
C’est un graphe montrant la fréquence par rapport aux 4 modalitées de variable « poutcome » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 4 modalitées : "unknown", "other", "failure"et "success" selon la variable « marital »

Figure 18: Représentation Graphique de la variable "poutcome"



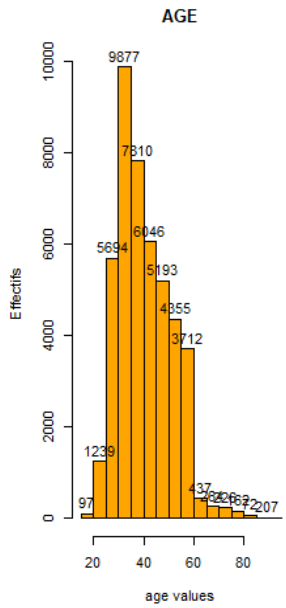
C’est un graphe montrant la fréquence par rapport aux 12 modalitées de variable « month» qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de ces 12 mois selon la variable « marital »

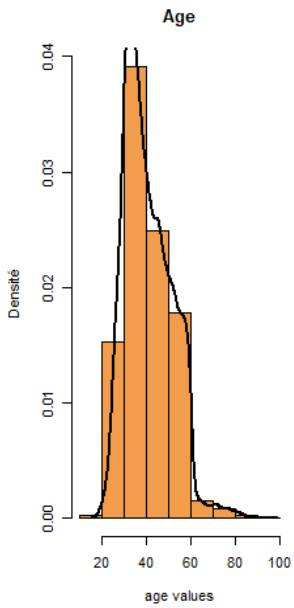
Figure 19: Représentation Graphique de la variable "month"

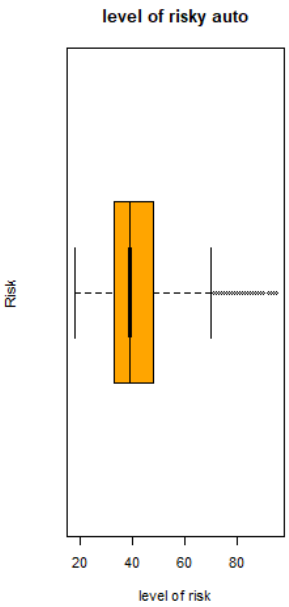


La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « previous », de plus une boite a moustache qui représente certains points aberrantes

Figure 20: Représentation Graphique de la variable "previous"

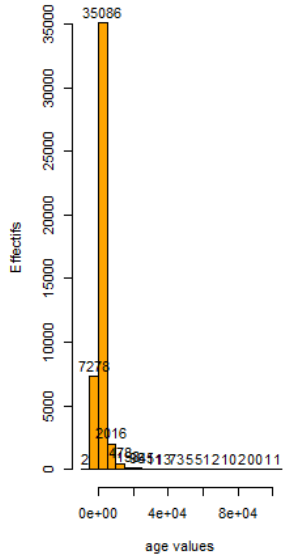


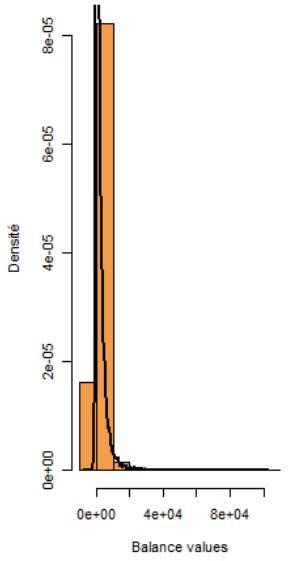


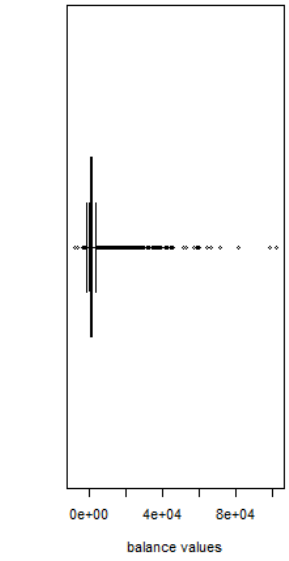


La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « Age » qui a une distribution asymétrique positive, de plus une boite a moustache qui représente quelques points aberrantes.

Figure 21: Représentation Graphique de la variable "Age"

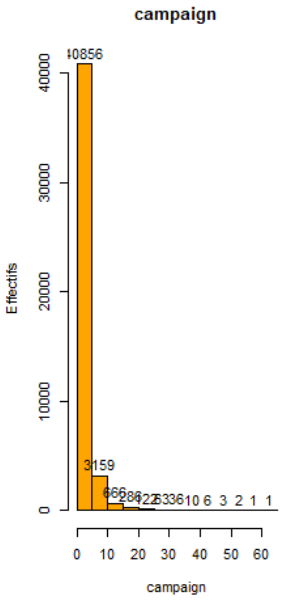


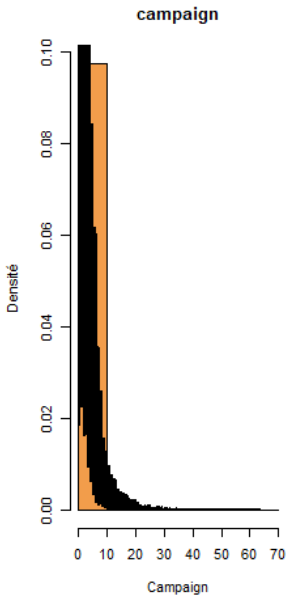


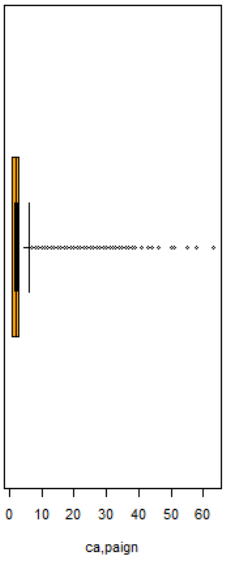


La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « Balance » qui a une distribution asymétrique positive, de plus une boite a moustache qui représente plusieurs points aberrantes.

Figure 22: Représentation Graphique de la variable "Balance"

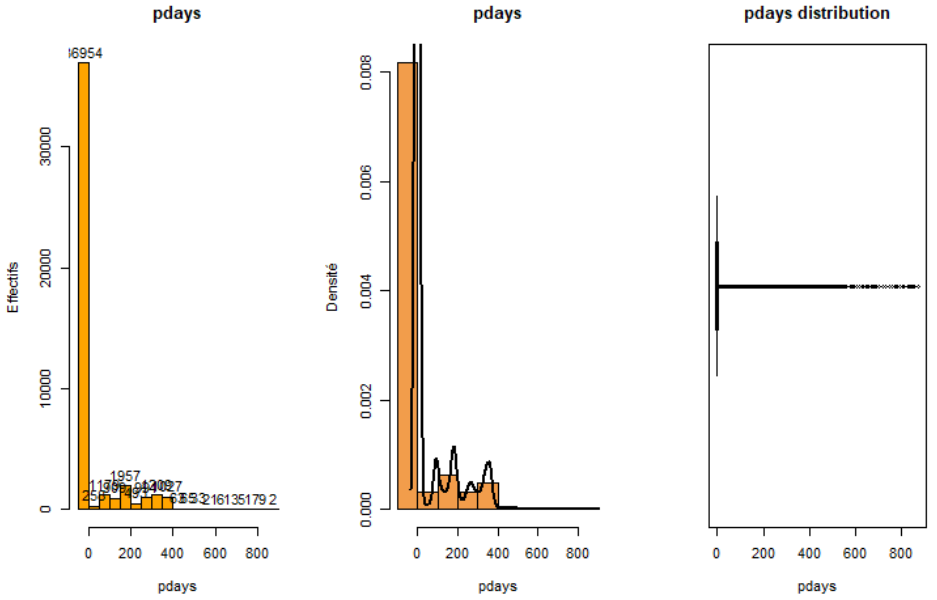






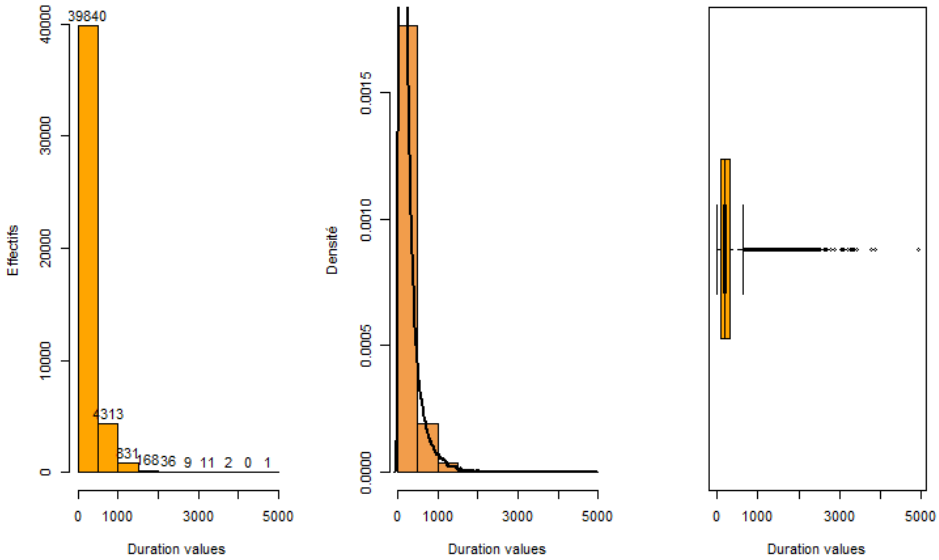
La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « Campaign », de plus une boite a moustache qui représente plusieurs points aberrantes.

Figure 23: Représentation Graphique de la variable "Campaign"



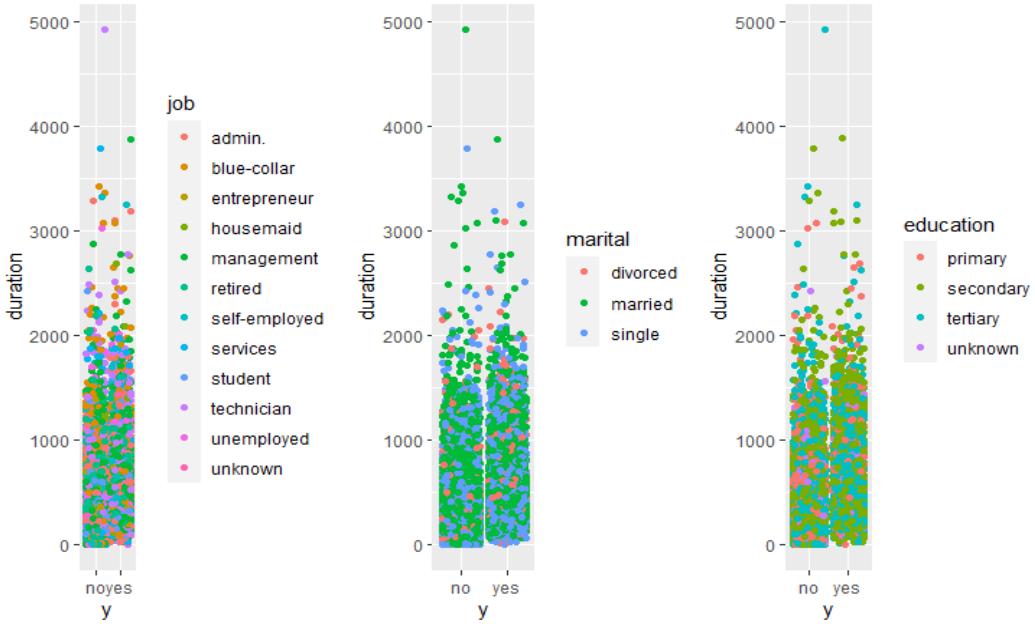
La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « pdays », de plus une boite a moustache qui représente plusieurs points aberrantes.

Figure 24: Représentation Graphique de la variable "pdays"



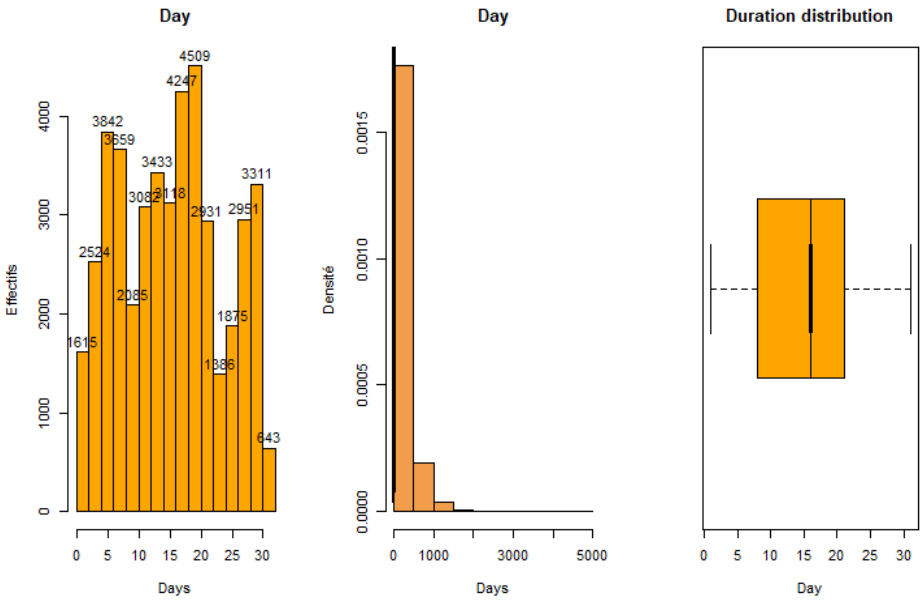
La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « Duration » qui a une distribution asymétrique, de plus une boite a moustache qui représente plusieurs points aberrants.

Figure 25: Représentation Graphique de la variable "Duration"



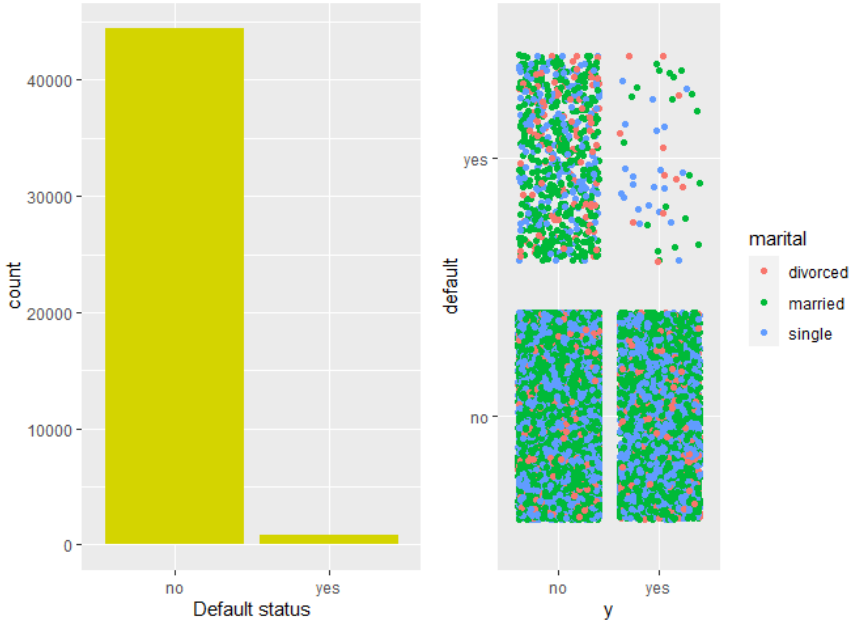
La figure présente la dispersion de la variable « duration » selon 3 variables : job, marital et éducation, qui est fortement liée à la variable cible « y »

Figure 26: Dispersion e la variable "Duration"



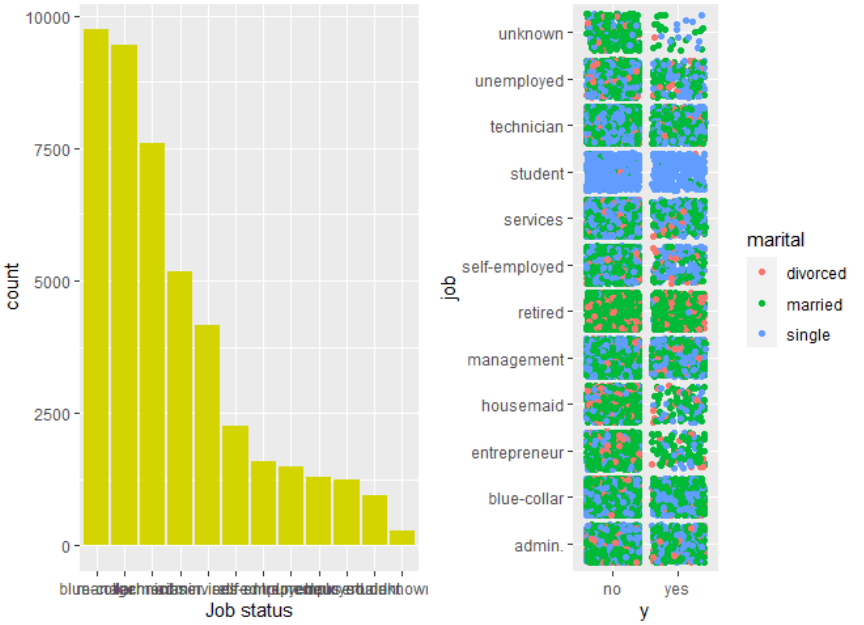
La figure présente deux histogrammes, l'un représentant la distribution des effectifs et l'autre la densité de la variable « Day » qui a une distribution une distribution approximativement uniforme, de plus une boite a moustache qui ne représente pas aucun points aberrantes.

Figure 27: Représentation Graphique de la variable "Days"



C'est un graphe montrant la fréquence par rapport aux 2 modalités de variable « default » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de cette variable binaire mois selon la variable « marital »

Figure 28: Représentation Graphique de la variable "Default"



C'est un graphe montrant la fréquence par rapport aux 12 modalités de variable « job » qui sont classées par un ordre croissant du plus fréquents au moins fréquents, de plus on peut voir la dispersion de cette variable selon la variable « marital »

Figure 29: Représentation Graphique de la variable "Job"

Analyse de corrélation et réduction de dimension

Relation entre les variables qualitative et la variable cible :

On a effectué le test de khi-deux entre la variable qualitative et la variable cible :
Les résultats montrent que la valeur de p (p-value) est inférieure à 0,05 alors il y a une association significative entre les variables étudiées, c'est-à-dire on rejette l'hypothèse nulle (H0 : les variables ne sont pas significatives vs H1 : les variables sont significatives)

Tableau 2 : Tableau de Correlation en utilisant khi-deux test

	PEARSON'S CORRELTION TEST			
	Test	**X.squared**	**df**	**p.value**
1	t1: job & y	836.10549	11	3.34E-172
2	t2: marital & y	196.49595	2	2.15E-43
3	t3: education & y	238.92351	3	1.63E-51
4	t4: default & y	22.20225	1	2.45E-06
5	t5: hosing & y	874.82245	1	2.92E-192
6	t6: loan & y	209.61698	1	1.67E-47
7	t7: contact & y	1035.71423	2	1.25E-225
8	t8: month & y	3061.83894	11	0.00E+00
9	t9: poutcome & y	4391.50659	3	0.00E+00

Relation entre les variables quantitatives :

On a utilisé plusieurs méthodes:

Méthode 1 : matrice de corrélation

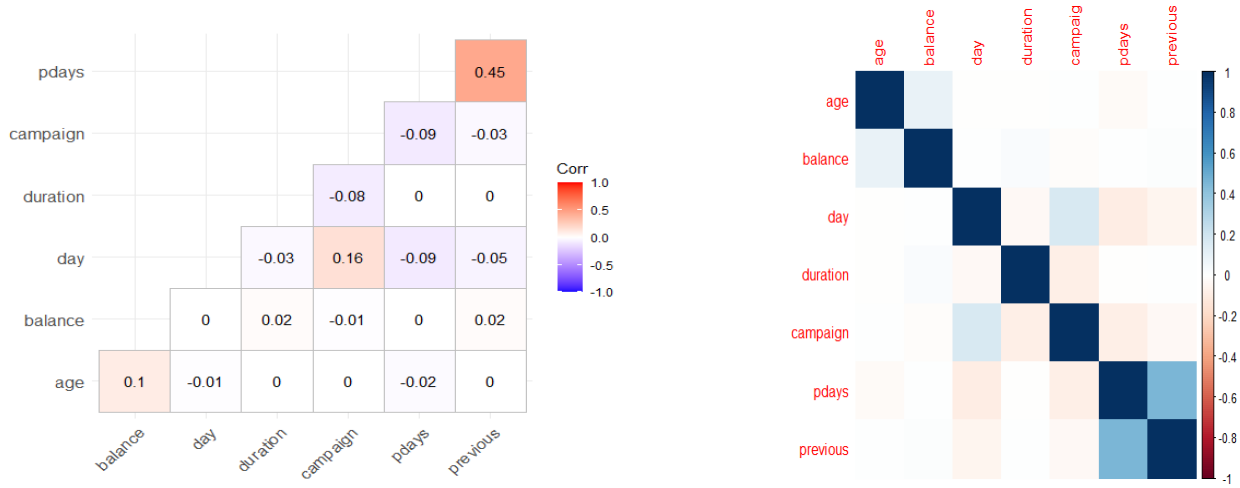


Figure 30: Matrice de correlation

C'est deux matrice de corrélation ne montrent pas une forte relation entre les variables quantitative (variables explicatives), la plus forte corrélation c'est entre : « pdays » et « previous » qui vaut 0.45 ce qui est logique entre le nombre de jours écoulés depuis le dernier contact avec le client et le nombre de contacts effectués avant cette campagne pour le client.

Méthode 2 : Analyse des composantes principales (PCA)

C'est une méthode appliquée pour réduire la dimension de la base de données.
Notons qu'on normalise la base de données puisque les variables ne sont pas des mêmes unités.
En appliquant cette méthode, on a réduit la dimension de 17 à 7 variables, mais les proportions de variance expliquée par les 7 composantes principales ne sont pas élevées : 21,56% ; 16,50% ; 15,67 ; 13,93% ; 12,82% ; 11,80% ; 7,72%

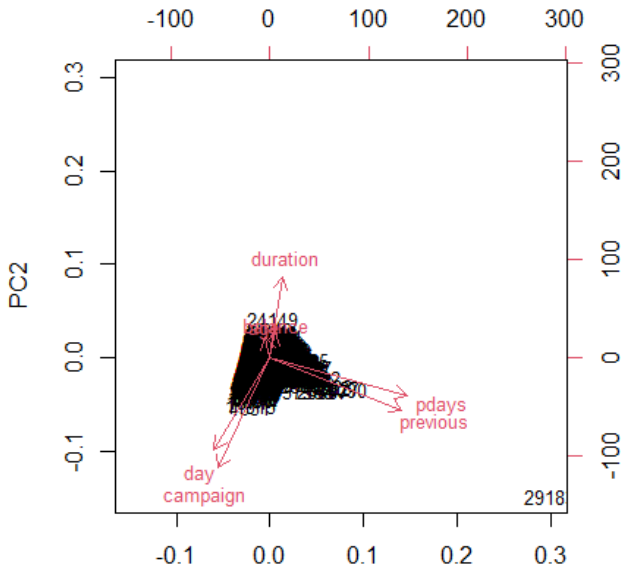


Figure 31: Biplot représentant les relations entre les individus et les variables

C'est un graphe à 2 dimensions montrant la relation entre :

1. les observations: sous forme de points regroupes.
2. les variables: sous forme de flèches éloignées qui suggèrent une faible corrélation entre eux.

Mais d'après les résultats on constate que l'ACP n'a pas donné des résultats efficaces puisque les proportions de variance expliquée par PC 1 et PC2 sont faibles : 21,56% et 16,50%

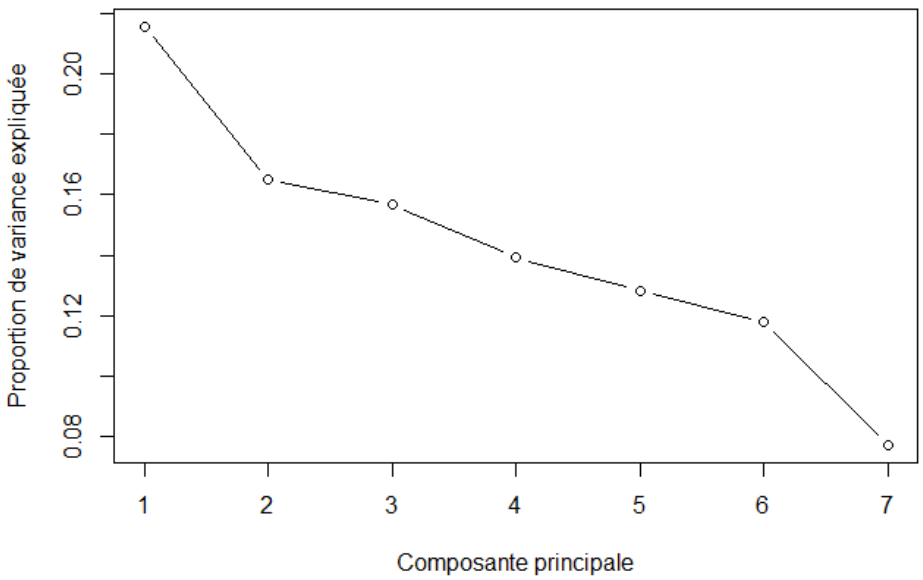


Figure 32: Variation des proportions de variance selon les composantes principales

Puisque la proportion de variance expliquée par les composantes n'est pas élevée, il est possible que cette réduction de dimensionnalité entraîne une perte d'informations et affecte la précision des prédictions. C'est pour cela on ne va pas réduire la dimension comme l'article 5 (ils ont réduit la dimension parce qu'ils ont obtenu des proportions des variances élevés).

Réduire la dimension de la base de données :

Nous allons utiliser 3 méthodes qui s'appliquent sur une base de donnée quantitatives, et une 4eme sur une base de données mixte.

Méthode 1 : Analyse des composantes mixte (PCAmix)

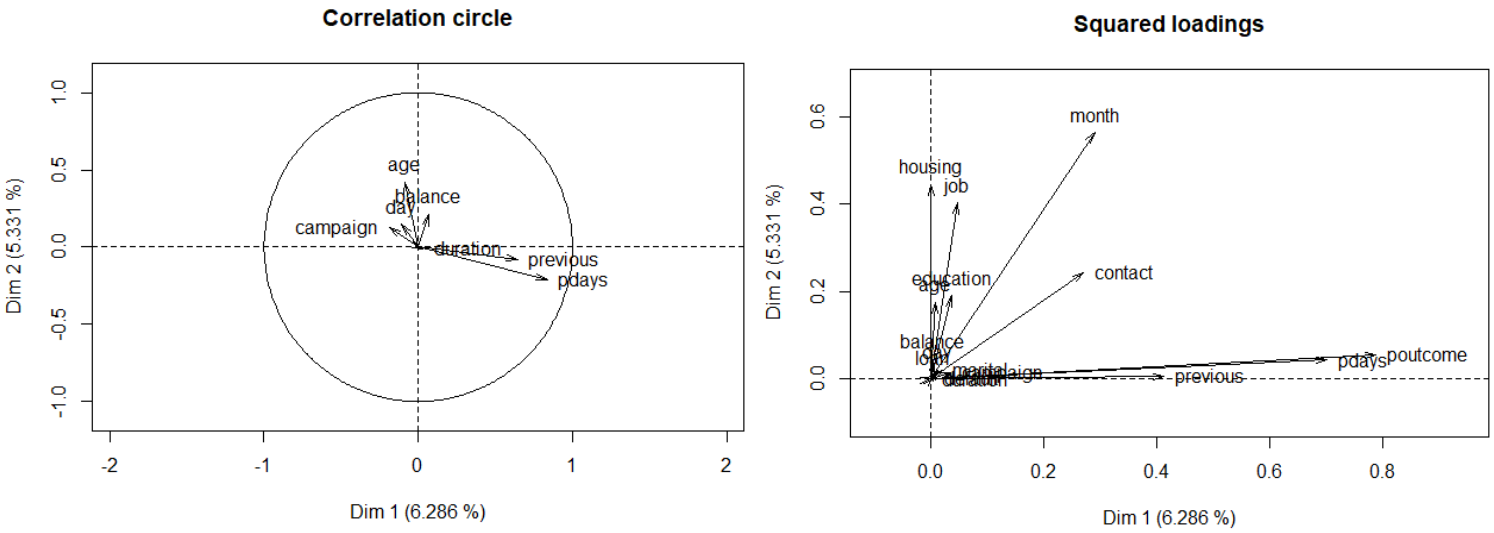


Figure 33: Représentation de la corrélation des variables en utilisant PCAmix

Les deux figures montrent des résultats similaires à celui de l'ACP, alors on ne va pas réduire la dimension de la base de données pour ne pas perdre des informations et affecter la prédiction

Méthode 2 : Hierarchical clustering algorithm (Hcluster)

Apres avoir testé les stabilités de 100 Boostrap, on a divise la base de donnée en 3 classes (k=3)
Cluster 1 : « Age et balance » avec des corrélations de 0.55 et -0.74 respectivement
Cluster 2 : « day, campaign et duration » avec des corrélations de 0.55, 0.46 et 0.18
Cluster 3 : « pdays et previous » avec des corrélations de 0.73 et 0.85 respectivement.
Donc on conclut que la méthode hclust a réduit notre base de données de 17 à 7 variables synthétiques avec une qualité de classification de 40.8% (gain en cohésion).

Méthode 3 : Kmeans classification

On utilisant l'algorithme de Kmeans nous avons aussi divisé la base de données en 3 groupes et n ous avons obtenus les mêmes variables dans les 3 groupes de la méthode 2 avec une qualité de cl assification de 40.79% (gain en cohésion)

Méthode 4 : Algorithme du Boruta

Boruta est une méthode de sélection de variables qui utilise l'algorithme Boruta pour évaluer l'importance des variables indépendantes par rapport à une variable cible (variable dépendante), cette méthode est divise en 3 étapes :
Etape 1: il commence à créer des copies aléatoires appelées "ombres" des variables indépendantes, par exemple la variable "Age" sera créé "Age copie" (de même pour les autres)
Etape 2: On mélange les valeurs des ombres d'une manière aléatoire, dans ce cas les valeurs des ombres ne sont pas dans le même ordre que l'initial et peuvent être différentes de celles de la variable réelle par exemple si l'Age=12, 34,56 (on peut les permutées ou changées)

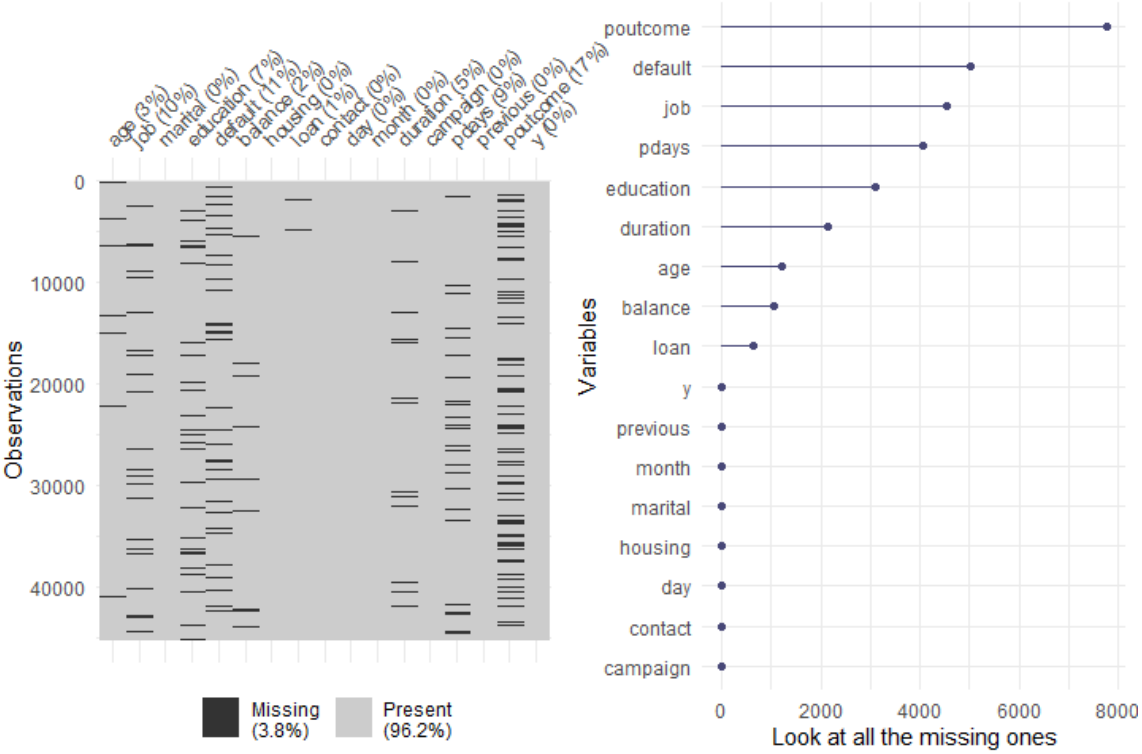
➔ l'Age ombre=34, 56,12 ou → l'Age ombre=51, 39,55

Etape 3: On compare les copie des ombres avec les valeurs réelles pou choisir les variables les plus significatives pour la prédiction de la variable cible.
Conclusion : après l'application de 199 itérations en 1.241218 mins nous avons obtenus 14 de 17 variables explicatives, autrement dit Boruta a considère que la variable « éducation » et « job » n'affectent pas la variables comme les autres 14. Mais dans notre cas on ne va pas éliminer aucunes variables pour obtenir des meilleures prédictions

Manipulation des valeurs manquantes

Puisque notre base de donnée ne contienne pas des valeurs manquantes alors nous allons remplacer aléatoirement des valeurs manquantes dans le but de les traitées :

Visualiser les valeurs manquantes :



C'est deux représentations graphiques montrent les différentes pourcentages des valeurs manquantes dans chaque variables explicatives telle que « poutcome » a le plus grand pourcentage (17%) de toutes les NA

Figure 34: Représentation Graphique des NA en % dans chaque variable

Imputer les valeurs manquantes :

Maintenant nous allons imputer ces NA par deux méthodes :

Méthode 1 : Imputation multiple

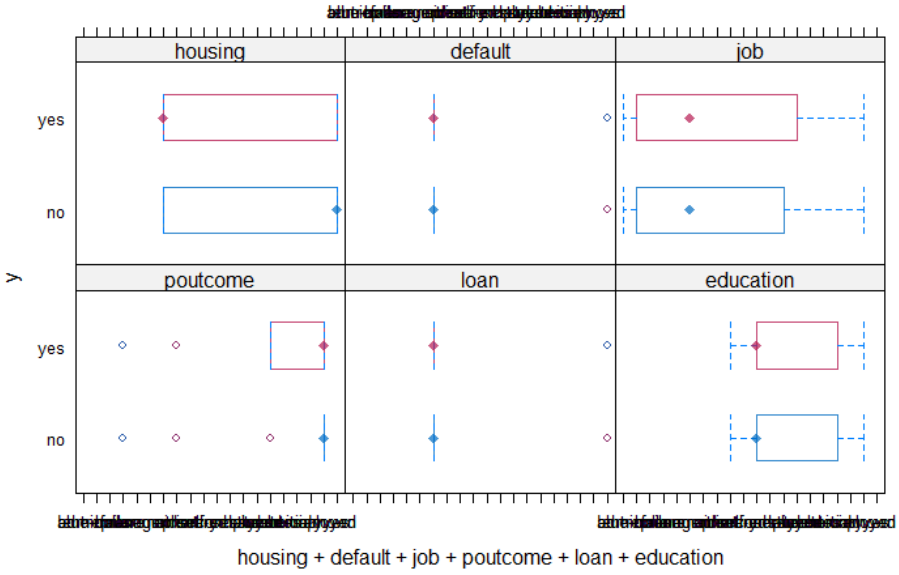
En utilisant l'imputation multiple on impute toutes les valeurs manquantes par 5 itérations, en choisissant une méthode (GLM, Markov Chain..) c'est-à-dire on répète l'imputation 5 fois mais à chaque fois on change les modalitées.

Méthode 2 : Imputation simple

En utilisant l'imputation simple, on impute chaque variable séparément en choisissant une méthode (imputation par moyenne, médiane ou aléatoire).

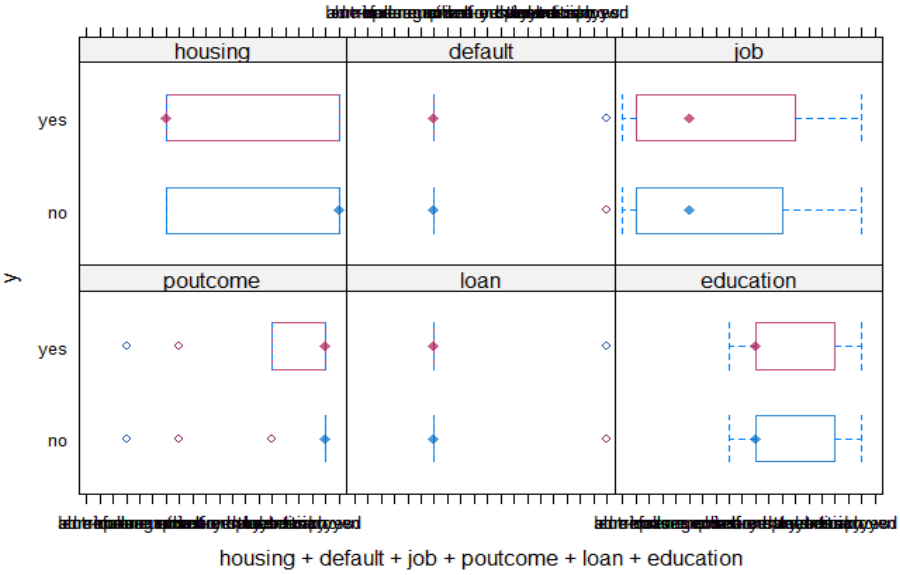
Représentation graphique Avant/Apres imputation :

On a appliqué une imputation multiple, et maintenant nous allons voir une représentation graphique des valeurs théoriques (avant imputation) et des valeurs prédites (après imputation) des 2 premières itérations d'imputation multiple. (t1 et t2)



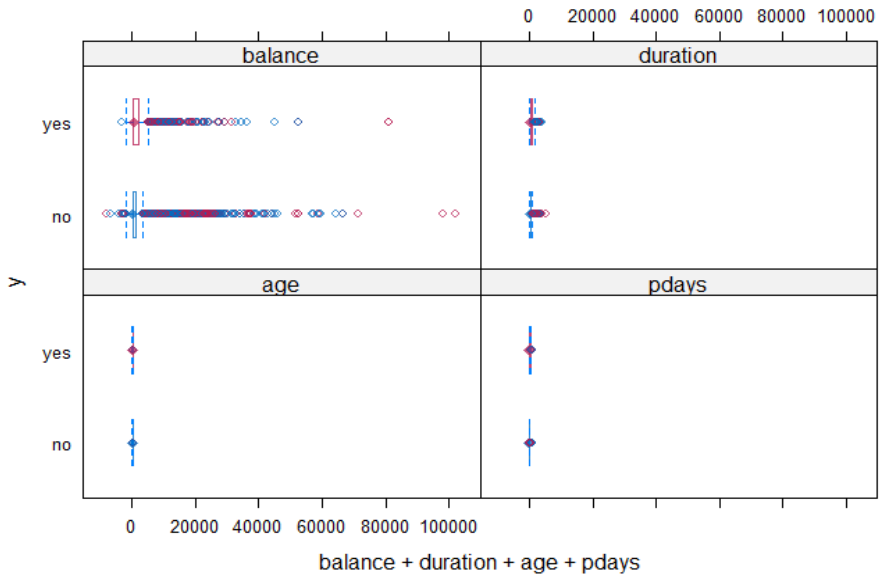
Ce graphe montre que la forme des points magenta (imputés) corresponde à la forme des points bleus (observés) des 6 variables qualitatives de la **1ere itération** d'où cette correspondance de forme indique que les valeurs imputées sont effectivement des "valeurs plausibles".

Figure 35: Représentation Graphique avant/Apres imputation des variables qualitatives de la 1ere itération



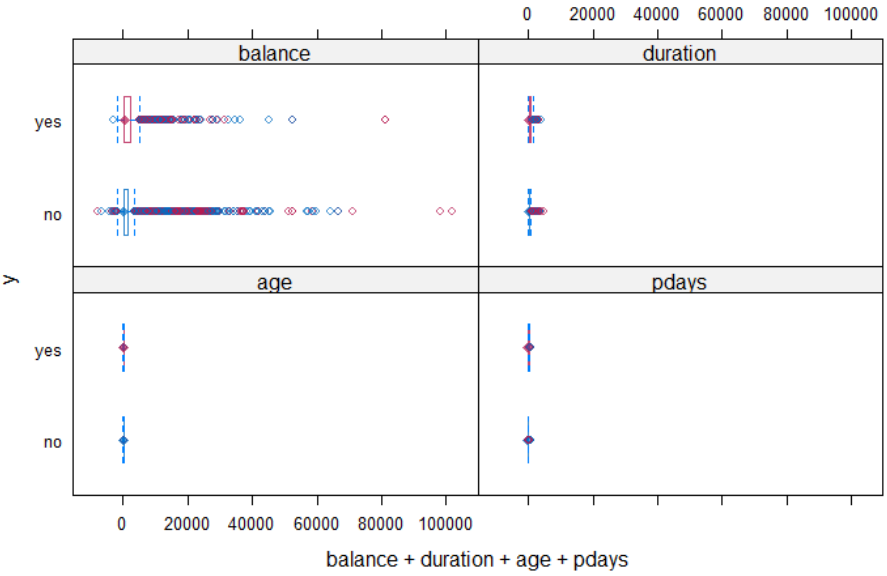
Ce graphe montre que la forme des points magenta (imputés) corresponde à la forme des points bleus (observés) des 6 variables qualitatives de la **2eme itération** d'où cette correspondance de forme indique que les valeurs imputées sont effectivement des "valeurs plausibles".

Figure 36: Représentation Graphique avant/Apres imputation des variables qualitatives de la 2eme itération



Ce graphe montre que la forme des points magenta (imputés) corresponde à la forme des points bleus (observés) des 4 variables quantitatives de la **1ere itération** d'où cette correspondance de forme indique que les valeurs imputées sont effectivement des "valeurs plausibles".

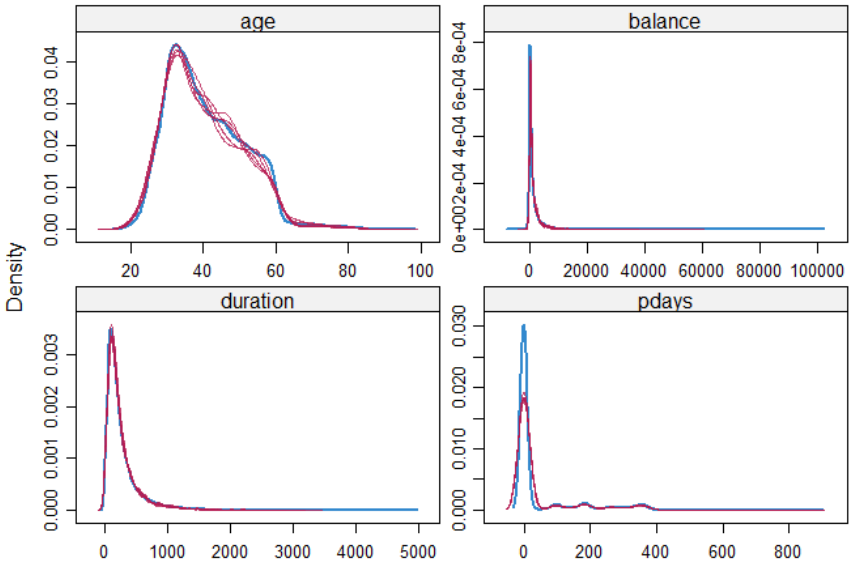
Figure 37: Représentation Graphique avant/Apres imputation des variables quantitatives de la 1ere itération



Ce graphe montre que la forme des points magenta (imputés) corresponde à la forme des points bleus (observés) des 4 variables quantitatives de la **2eme itération** d'où cette correspondance de forme indique que les valeurs imputées sont effectivement des "valeurs plausibles".

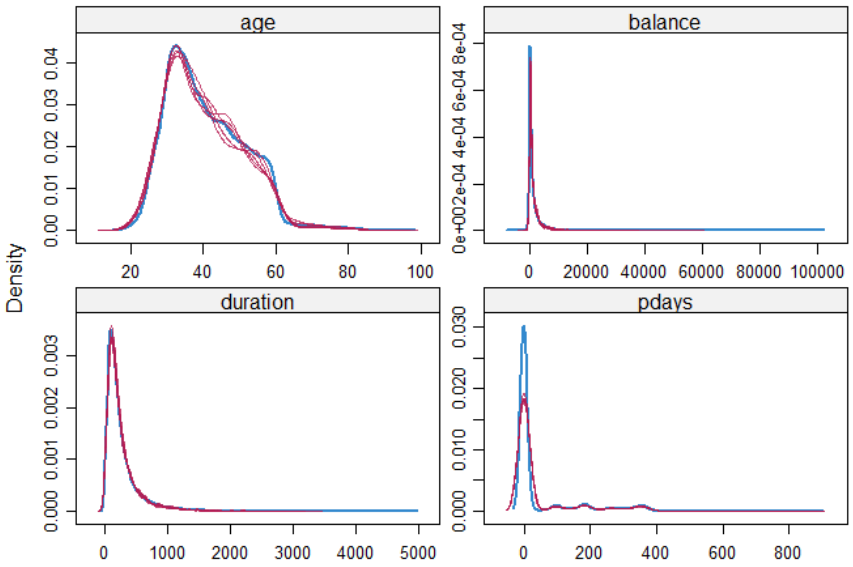
Figure 38: Représentation Graphique avant/Apres imputation des variables quantitatives de la 2eme itération

Dans ces 4 figures on a représentés les variations des NA avant et après imputation de la variables cible y en fonction des variables explicatives (quantitatives et qualitatives) qui contiennent des NA. Maintenant on va voir la densité des variables quantitatives : des valeurs théoriques (avant imputation) et des valeurs prédites (après imputation) des 2 premières itérations d'imputation multiple. (t1 et t2)



Ce graphe montre que la densité des données imputées (magenta) est similaire aux densités des données observées (bleus) des 4 variables quantitatives de la **1ere itération**

Figure 39: La Densité Avant/Apres imputation des variables quantitatives de la 1ere itération



Ce graphe montre que la densité des données imputées (magenta) est similaire aux densités des données observées (bleus) des 4 variables quantitatives de la **2eme itération**

Figure 40: La Densité Avant/Apres imputation des variables quantitatives de la 2eme itération

Transformation et Réduction de la base de données :

Notre base de données est mixte, c'est pour cela nous somme besoin de réduire le nombre des modalités de chaque variable qualitative, et de transformer la base de données mixte en une base de données numérique. On va travailler par 4 bases de données :

1. **Data** : sans transformation sans réduction (base de données initiale mixte)
2. **Data_reduced** : la base initiale avec une réduction de nombre de modalités
3. **Data_numeric** : la base initiale mais transformer en nombre (base de données numériques)
4. **Data_numeric_scaled** : la base numérique avec une normalisation

Transformer la base de données :

Transformer la variable qualitative Job en quantitative :
"admin."=1, "management"=2, "technician"=3, "blue-collar"=4, "entrepreneur"=5, "housemaid"=6, "retired"=7, "self-employed"=8, "services"=9, "student"=10, "unemployed"=11, "unknown"=12

Transformer la variable qualitative Marital en quantitative :
"married"=1, "divorced"=2, "single"=3

Transformer la variable qualitative Education en quantitative :
"primary"=1, "secondary"=2, "tertiary"=3, "unknown"=4

Transformer la variable qualitative Contact en variable quantitative :
"Known Contact"=1, "cellular"=2, "telephone"=3, "Unknown Contact"=4

Transformer la variable qualitative Month en variable quantitative :
"apr"=4, "aug"=8, "dec"=12, "feb"=2, "jan"=1, "jul"=7, "jun"=6, "mar"=3, "may"=5, "nov"=11, "oct"=10, "sep'=9

Réduire la base de données :

Transformer la variable Job en 2 catégories au lieu de 12 :

1. "admin.", "management", "technician" → **"high quality"**
2. "blue-collar","entrepreneur","housemaid","retired","self-employed","services", "student", "unemployed", "unknown" → **"average quality".**

Transformer la variable Marital en 2 catégories au lieu de 3 :

1. "married"→ **"Married"**
2. "divorced", "single"→ **"Non Married"**

Transformer la variable Education en 2 catégories au lieu de 4 :

1. "primary", "secondary" → **"Lower education"**
2. "tertiary", "unknown" → **"Higher education"**

Transformer la variable contact en 2 catégories au lieu de 4 :

1. "cellular", "telephone"→ **"Known Contact"**
2. "unknown" → **"Unknown Contact"**

Transformer la variable month en 2 catégories au lieu de 12 :

1. "jun", "jul", "aug" → **"Summer Months"**
2. Toutes les autres modalités → **"Non-Summer Months**"

Transformer la variable poutcome en 2 categories au lieu de 4 :

1. "failure", "unknown" → **"Unsuccessful Outcome"**
2. "other", "success" → **"Successful Outcome"**

Normaliser la base de données:

En utilisant la fonction « scale » on peut normaliser la base de données puisque les variables n'ont pas les mêmes unités de mesures.

Partition et suréchantillonnage de la base de données :

Partition de la base de données :

Avant d'appliquer les modèles d'apprentissage supervisée sur la base de données, il faut la diviser en deux ensembles : ensemble d'apprentissage (contenant 70% des clients) et ensemble de validation (contenant 30% des clients), dans le but d'appliquer les 8 modèles sur l'ensemble d'apprentissage et ensuite tester ces performances sur l'ensemble de validation.

Effectuer un suréchantillonnage :

Notre base de données est déséquilibrée, comme nous avons vu dans la représentation graphique de la variable cible que seulement 16% de la population s'inscrivent a un dépôt alors que les autres ne s'inscrivent pas, ce qui affecte négativement sur la performance des modèles, c'est pour cela après la partition de notre base nous allons effectuer un suréchantillonnage sur la partie « apprentissage » pour augmenter le nombre des clients qui n'ont pas inscrit à un dépôt en utilisant la fonction "ovun.sample"

Application des 8 modèles :

Explication théorique sur les mesures métriques :

Avant d'appliquer ces 8 modèles nous allons voir théoriquement les mesures métriques qui aident à comparer la prédiction de chaque modèle en utilisant les Symboles :
VP= Vrai positifs ; VN= Vrai négatifs ; FP=Faux positifs et FN=Faux négatifs.

1. Accuracy (Exactitude): **Accuracy = (VP + VN) / (Total)**
2. Sensitivity (Sensibilité): **Sensitivity = VP / (VP + FN)**
3. Specificity (Spécificité): **Specificity = VN / (VN + FP)**
4. Pos Pred Value (Valeur prédictive positive): **Pos Pred Value = VP / (VP + FP**)
5. Neg Pred Value (Valeur prédictive négative): **Neg Pred Value = VN / (VN + FN)**
6. Prévalence (Prévalence) : **Prevalence = (VP + FN) / (Total)**
7. Detection Rate (Taux de détection) : **Detection Rate = VP / (VP + FN)**
8. Detection Prevalence (Prévalence de détection) : **Detection Prevalence = (VP + FP) / (Total)**
9. Balanced Accuracy (Exactitude équilibrée): **Balanced Accuracy = (Sensitivity + Specificity) / 2**

Accuracy (Exactitude) : L'accuracy représente la proportion globale de prédictions correctes par rapport au total des prédictions. C'est une mesure générale de performance qui convient lorsque les classes sont équilibrées.

Sensitivity (Sensibilité) : La sensibilité, également appelée taux de vrais positifs, mesure la capacité du modèle à détecter les vrais positifs (événements réellement positifs) parmi tous les cas positifs réels. Elle est particulièrement utile lorsque la détection des positifs est cruciale, par exemple dans les tests de dépistage des maladies.

Specificity (Spécificité) : La spécificité, également appelée taux de vrais négatifs, mesure la capacité du modèle à détecter les vrais négatifs (événements réellement négatifs) parmi tous les cas négatifs réels. Elle est importante lorsque la minimisation des faux positifs est critique, par exemple dans les tests de sécurité où les faux positifs peuvent entraîner des erreurs coûteuses.

Pos Pred Value (Valeur prédictive positive) : La valeur prédictive positive mesure la proportion de vrais positifs parmi toutes les prédictions positives du modèle. Elle permet d'évaluer la fiabilité des prédictions positives.

Neg Pred Value (Valeur prédictive négative) : La valeur prédictive négative mesure la proportion de vrais négatifs parmi toutes les prédictions négatives du modèle. Elle évalue la fiabilité des prédictions négatives.

Prevalence (Prévalence) : La prévalence représente la proportion de cas positifs réels parmi l'ensemble des observations. Elle peut aider à interpréter les autres mesures en fonction de la distribution des classes dans les données.

Detection Rate (Taux de détection) : Le taux de détection mesure la proportion de vrais positifs parmi tous les cas positifs réels. Il est utile pour évaluer la capacité du modèle à détecter les cas positifs réels.

Detection Prevalence (Prévalence de détection) : La prévalence de détection mesure la proportion de cas prédits positifs (vrais positifs et faux positifs) parmi l'ensemble des observations. Elle fournit des informations sur la fréquence des prédictions positives du modèle.

Balanced Accuracy (Exactitude équilibrée) : L'exactitude équilibrée est la moyenne de la sensibilité et de la spécificité. Elle fournit une mesure globale de la performance du modèle en tenant compte à la fois des vrais positifs et des vrais négatifs.

Modèle 1 : Arbre de décision (DT)

Le modèle DT nécessite pas une base de données numérique c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer les performances de la prédiction sur les ensembles de tests de : **Data** et **Data_reduced**.

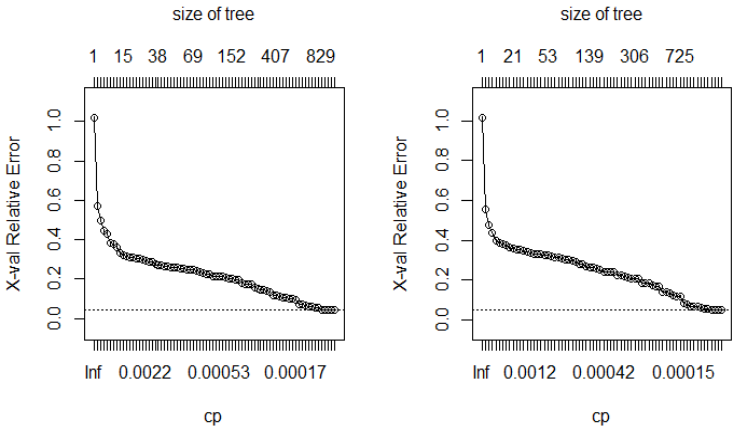


Figure 41:Representation Graphique de l'erreur en fonction du "cp" pour le modèle DT

C'est un graphe montrant l'erreur relative en fonction des différentes possibilités de cp (nbr d'élagage de l'arbre de décision). On remarque qu'il n'existe pas une grande différence pour choisir le cp entre les 2 base de données data et Data_reduced alors la meilleure valeur de cp est de 0.000100391

Tableau 3 : Les variables importantes selon le modèle DT

Variables	Important Variables of Data	Important Variables of Data reduced
duration	5721.9006	6311.15675
month	2545.50278	2367.18547
balance	1961.16636	2854.36018
age	1886.61324	2367.18547
day	1733.45783	2343.77924
job	1449.48427	347.79999
poutcome	1107.47612	1189.99586
contact	1050.2	833.88913
campaign	991.37172	1146.74904
pdays	822.70506	1389.14798
previous	623.86954	944.33625
housing	529.94752	465.78409
education	508.6554	369.93917
marital	360.64076	372.2222
loan	222.92748	256.16772
default	46.44389	66.90226

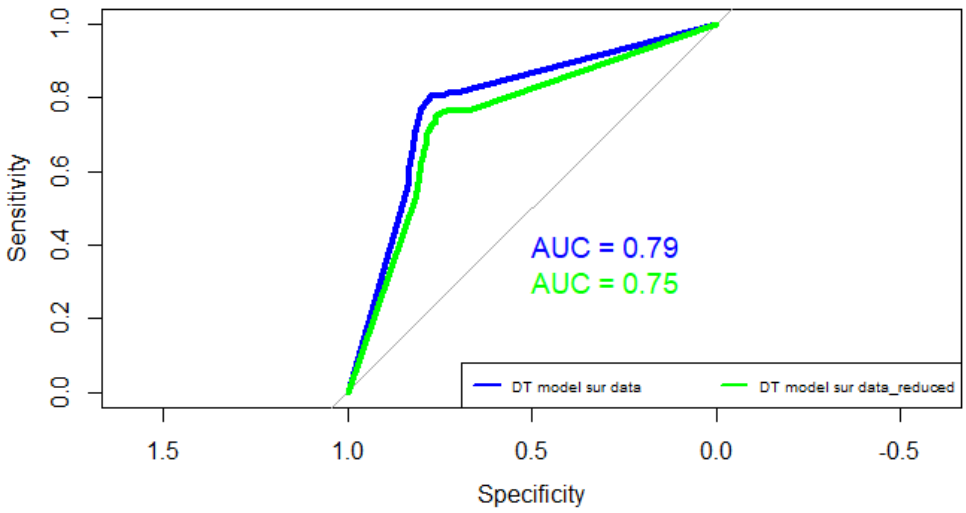
En utilisant ce modèle on peut tirer dans un tableau par un ordre croissant l'importance des variables explicatives pour construire l'arbre de décision de chaque base de données :

On remarque que la variable « duration » est le plus important suivi par « month et balance » pour construire le modèle d'arbre de décision.

Maintenant nous allons voir la performance de ce modèle en utilisant la courbe de Roc et des autres mesures métriques comme : Accuracy, sensibilités, spécificités... qui sont tirées en utilisant la fonction « Confusion Matrix » de la Library « caret ».

La courbe de Roc :

La courbe de roc de la base de données data a une valeur AUC=79% plus grande que 75%, celle de Data_reduced cela est dû à la réduction des modalités qui a affecté négativement sur la prédiction.



C'est un graphe montrant les deux courbes de Roc du modèle DT des deux bases de données :
Data et Data_reduced

Figure 42: Courbe de Roc du modèle DT

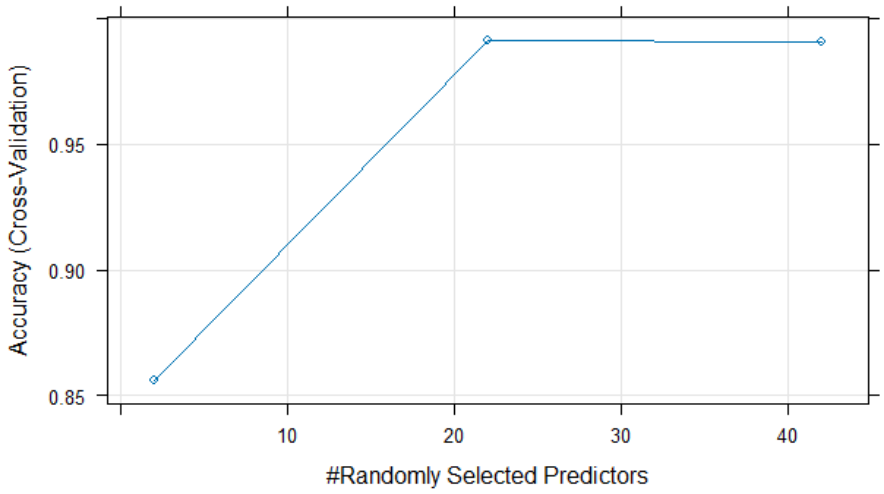
Autres mesures métriques :

1. **Accuracy :** Le modèle DT sur la base de données Data a une valeur de **0.6266** avec un intervalle de confiance **(0.6209, 0.6323)**, tandis la base de données Data_reduced a donné respectivement **0.6296** et **(0.6239, 0.6353)**. On constate qu'il n'existe pas une grande différence entre les valeurs. Mais l'accuracy de l'article 1 est 98.98% plus grand que 62.66%

2. **Sensibilité et Sensitivité :** le modèle DT sur la base de données Data nous a donnée **0.3360** comme sensibilité et **0.9168** comme spécificité alors que la base de données Data_reduced nous a donnée respectivement **0.3402** et **0.9186**. On constate qu'il n'existe pas une grande différence entre ces valeurs

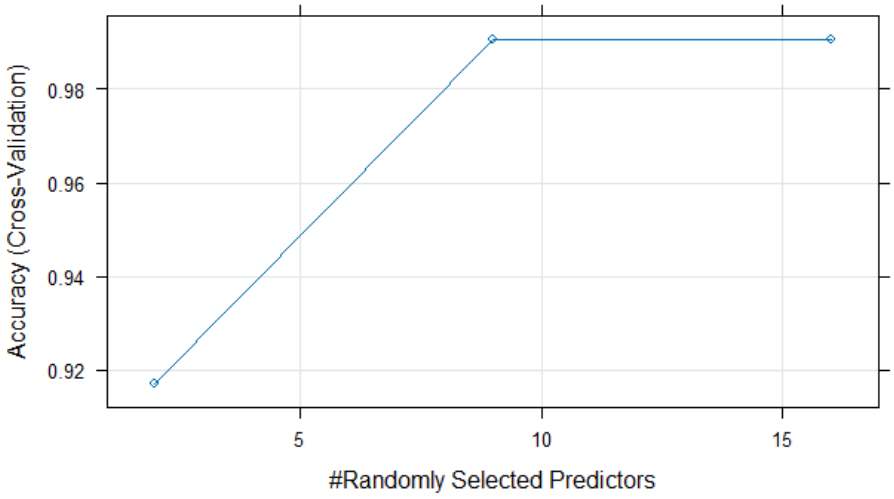
Modèle 2 : Foret aléatoire (RF)

Le modèle RF ne nécessite pas une base de données numérique comme le modèle DT, c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer les performances de la prédiction sur les ensembles de tests de : **Data** et **Data_reduced**.



Dans ce graphe : on peut constater que l'exactitude (Accuracy) augmente avec le nombre d'arbres, mais elle stabilise à un certain point. Cela suggère qu'ajouter plus d'arbres n'améliorera pas beaucoup la performance du modèle en termes d'exactitude.

Figure 43: La variation de l'exactitude en fonction de nbr d'arbre de la base de données Data



Dans ce graphe : on observe une tendance similaire au graphe précédant où l'exactitude augmente avec le nombre d'arbres. Mais l'exactitude semble légèrement inférieure par rapport au 1er graphe, Cela peut suggérer que la réduction de l'ensemble de données a entraîné une légère diminution de la performance du modèle.

Figure 44: La variation de l'exactitude en fonction de nbr d'arbre de la base de données Data_reduced

Tableau 4 : Importance des variables selon le modèle RF

Variable	Importance
age	1008.3154
jobblue-collar	103.88051
jobentrepreneur	39.06643
jobhousemaid	27.29231
jobmanagement	94.23078
jobretired	36.72991
jobself-employed	45.75519
jobservices	73.4051
jobstudent	63.75433
jobtechnician	88.87308
jobunemployed	47.20258
jobunknown	14.4647
maritalmarried	109.71526
maritalsingle	86.23082
educationsecondary	94.85273
educationtertiary	111.06387
educationunknown	43.46462
defaultyes	13.27096
balance	1176.52163
housingyes	435.28147
loanyes	118.77577
contacttelephone	61.90489
contactunknown	566.66217
day	955.70311
monthaug	197.33534
monthdec	11.48928
monthfeb	150.00675
monthjan	88.79404
monthjul	172.20043
monthjun	144.27802
monthmar	209.13141
monthmay	156.82511
monthnov	136.08993
monthoct	156.10046
monthsep	65.64133
duration	5049.36372
campaign	398.84078
pdays	460.82754
previous	228.69669
poutcomeother	33.52715
poutcomesuccess	810.87813
poutcomeunknown	110.61971

Variable	Importance
age	1256.59551
jobaverage quality	190.5129
maritalmarried	171.57266
educationhigher education	199.98447
default1	20.08152
balance	1542.65588
housing1	492.25644
loan1	150.75072
contactunknown	563.95324
day	1213.36666
monthsummer-months	425.21178
duration	5507.12015
campaign	499.5191
pdays	661.85399
previous	363.84666
poutcomeSuccessful Outcome	737.17482

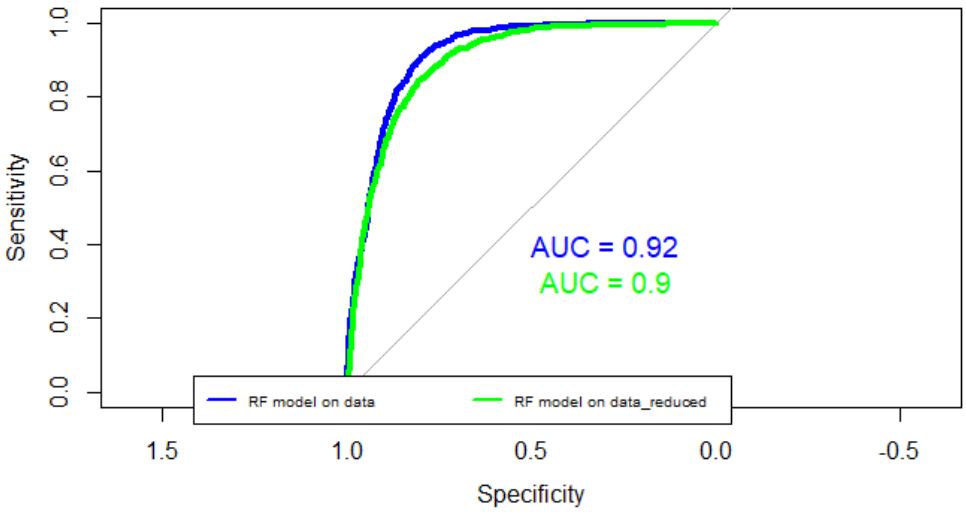
En utilisant ce modèle on peut tirer dans un tableau par un ordre croissant l'importance des variables explicatives (comme le modèle DT) pour construire les différents arbres de décision de chaque base de données :

On remarque que la variable « Age » est le plus important suivi par « éducation » dans la base de données Data, alors qu'il est suivi par « job » dans la base de données Data_reduced. Ceci montre que la réduction de la base de données à changer la division des différents arbres dans ce modèle.

Maintenant nous allons voir la performance de ce modèle en utilisant la courbe de Roc et des autres mesures métriques comme : Accuracy, sensibilités, spécificités... qui sont tirées en utilisant la fonction « Confusion Matrix » de la Library « caret ».

La courbe de Roc :

La courbe de roc de la base de données data à une valeur AUC=92% plus grande que 90%, celle de Data_reduced cela est dû à la réduction des modalités qui a affecté négativement sur la prédiction.



C'est un graphe montrant les deux courbes du modèle RF de Roc des deux bases de données : Data et Data_reduced

Figure 45: Courbe de Roc du modèle RF

Autres mesures métriques :

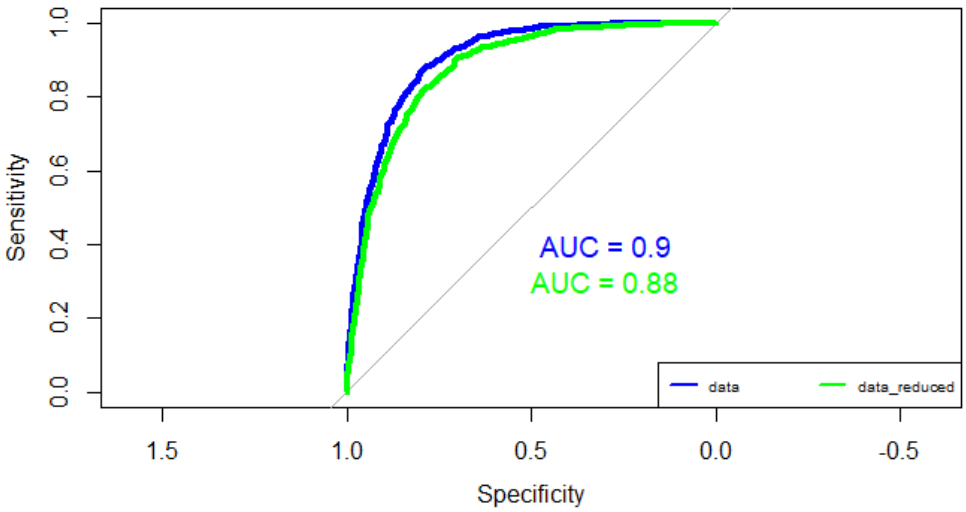
1. **Accuracy :** Le modèle RF sur la base de données data a une valeur de **0.6299** avec un intervalle de confiance **(0.6242, 0.6356)** tandis la base de données Data_reduced a donné respectivement **0.6422** et (0.6366, 0.6479). On constate qu'il n'existe pas une grande différence entre les valeurs.
2. **Sensibilité et Sensitivité :** le modèle DT sur la base de données Data nous a donnée **0.2906** comme sensibilité et **0.9688** comme spécificité alors que la base de données Data_reduced nous a donnée respectivement **0.3094** et **0.9747**. On constate qu'il n'existe pas une grande différence entre ces valeurs

Modèle 3 : Régression Logistique (LR)

Le modèle LR ne nécessite pas une base de données numérique comme le modèle DT et RF, c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer les performances de la prédiction sur les ensembles de tests de : **Data** et **Data_reduced**.

La courbe de Roc :

La courbe de roc de la base de données data à une valeur AUC=90% plus grande que 88%, celle de Data_reduced cela est dû à la réduction des modalités qui a affecté négativement sur la prédiction.



C'est un graphe montrant les deux courbes de Roc du modèle LR des deux bases de données : Data et Data_reduced

Figure 46: Courbe de Roc du modèle LR

Autres mesures métriques :

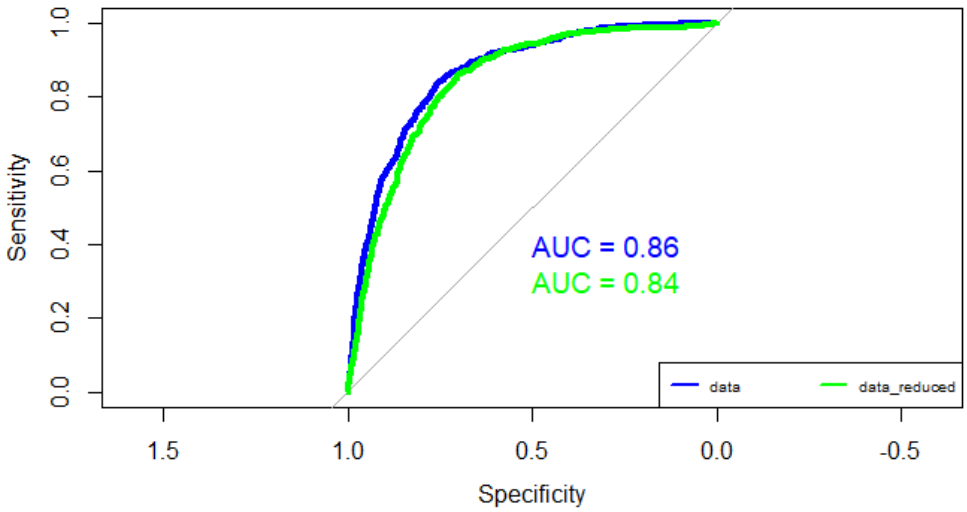
1. **Accuracy :** Le modèle RF sur la base de données data a une valeur de **0.8265** avec un intervalle de confiance **(0.822, 0.8309)** tandis la base de données Data_reduced a donné respectivement **0.7668** et **(0.7618, 0.7718).** On constate qu'il existe une petite différence entre les valeurs cela est dû à la réduction des modalités qui a affecté sur la prédiction du modèle LR plus que celle des modèles précédents DT et RF. Mais dans l'Article 5 l'AUC vaut 90.16% plus grand que 82.62%
2. **Sensibilité et Sensitivité :** le modèle DT sur la base de données Data nous a donnée **0.8359** comme sensibilité et **0.8170** comme spécificité alors que la base de données Data_reduced nous a donnée respectivement **0.7018** et **0.8318**. On constate aussi qu'il existe une petite différence entre ces valeurs est cela est due à la même raison.

Modèle 4 : Théorème de Bayes naïf (NB)

Le modèle NB ne nécessite pas une base de données numérique comme le modèle DT, RF et LR c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer les performances de la prédiction sur les ensembles de tests de : **Data** et **Data_reduced**.

La courbe de Roc :

La courbe de roc de la base de données data à une valeur AUC=92% plus grande que 90%, celle de Data_reduced cela est dû à la réduction des modalités qui a affecté négativement sur la prédiction.



C'est un graphe montrant les deux courbes de Roc du modèle NB des deux bases de données : Data et Data_reduced

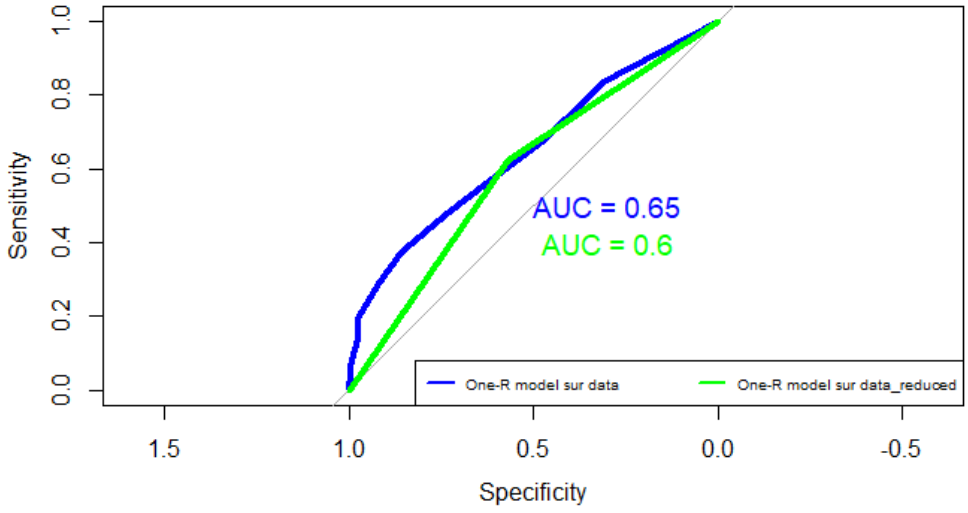
Figure 47: Courbe de Roc du modèle NB

Autres mesures métriques :

1. **Accuracy :** Le modèle NB sur la base de données data a une valeur de **0.7831** avec un intervalle de confiance **(0.7782, 0.7879)** tandis la base de données Data_reduced a donné respectivement **0.7431** et **(0.7379, 0.7482).** On constate qu'il n'existe pas une grande différence entre les valeurs. Mais l'accuracy de l'arcticle vaut 59.49% plus petite que 78.31%
2. **Sensibilité et Sensitivité :** le modèle DT sur la base de données Data nous a donnée **0.7694** comme sensibilité et **0.7968** comme spécificité alors que la base de données Data_reduced nous a donnée respectivement **0.6613** et **0.8248**. On constate qu'il n'existe pas une grande différence entre ces valeurs

Modèle 5 : Algorithme One-R (One-R)

Le modèle NB ne nécessite pas une base de données numérique comme le modèle DT, RF, LR et NB c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer les performances de la prédiction sur les ensembles de tests de : **Data** et **Data_reduced**.



C'est un graphe montrant les deux courbes de Roc du modèle One-R algorithme des deux bases de données : Data et Data_reduced

Figure 48: Courbe de Roc du modèle One-R

La courbe de roc de la base de données data à une valeur AUC=65% plus grande que 60%, celle de Data_reduced cela est dû à la réduction des modalités qui a affecté négativement sur la prédiction.

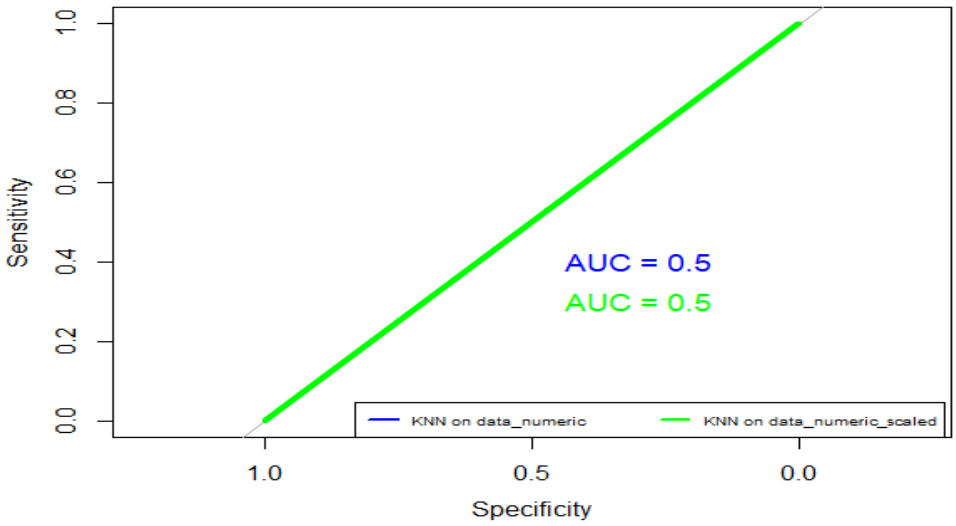
Autres mesures métriques :

1. **Accuracy :** Le modèle sur la base de données data à une valeur de **0.6162** avec un intervalle de confiance **(0.6105, 0.622)** tandis la base de données Data_reduced a donné respectivement **0.596** et **(0.5902, 0.6018).** On constate qu'il n'existe pas une grande différence entre les valeurs. Mais dans l'article 4 l'accuracy vaut 89.38% plus grand que 65%
2. **Sensibilité et Sensitivité :** le modèle One-R sur la base de données Data nous a donnée **0.3747** comme sensibilité et **0.8575** comme spécificité alors que la base de données Data_reduced nous a donnée respectivement **0.6233** et 0.5687. On constate qu'il n'existe pas une grande différence entre ces valeurs

Modèle 6 : K plus proche voisins (KNN)

Le modèle KNN nécessite une base de données numérique car elle travaille en utilisant la distance euclidienne c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer les performances de la prédiction sur les ensembles de tests de :
Data_numeric et **Data_numeric_scaled**.

Notons que ce modèle est applicable sur une base de données non-supervisée mais en se basant sur l'article 5, ils ont utilisé ce modèle pour faire une prédiction



C'est un graphe montrant les deux courbes de Roc du modèle KNN des deux bases de données : Data_numeric et Data_numeric_scaled

Figure 49: Courbe de Roc du modèle KNN

La courbe de roc de la base de données Data_numeric et Data_numeric_scaled ont la même valeur d'AUC qui est de 50%, c'est une mauvaise prédiction qui montre que le modèle KNN est efficace dans une base de donnée non-supervisée (pas dans notre cas)

Autres mesures métriques :

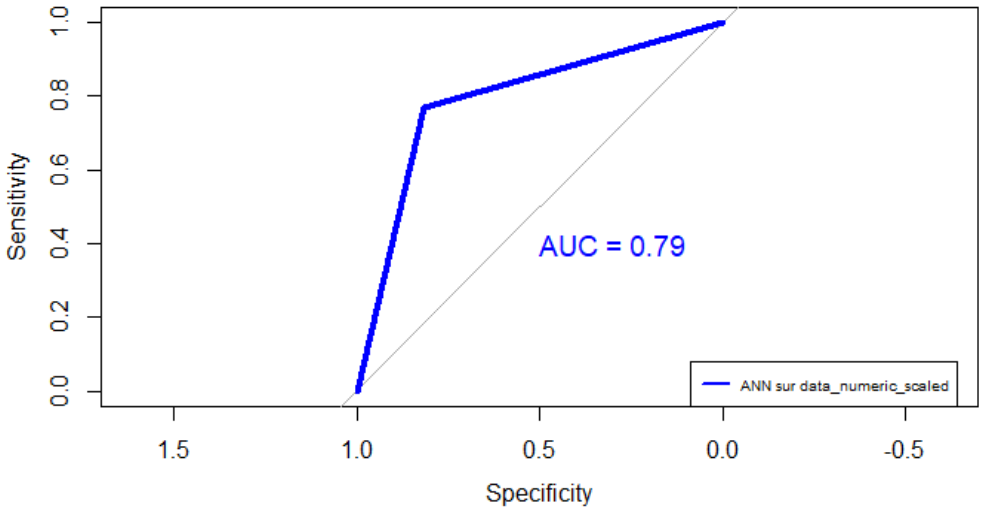
1. **Accuracy :** Le modèle sur la base de données numérique Data_numeric à une valeur de **0.6446** avec un intervalle de confiance **(0.639, 0.6502)** tandis la base de données

Data_numeric_scaled a donné respectivement **0.596** et **(0.5902, 0.6018).** On constate qu'il n'existe pas une différence entre les valeurs. Mais l'accuracy dans l'article 5 vaut 89.91% plus grand que 50%

2. **Sensibilité et Sensitivité :** le modèle KNN sur la base de données Data_numeric et Data_numeric_scaled nous a donnée **0.8954** comme sensibilité et **0.3935** comme spécificité (même mesure métriques)

Modèle 7 : Réseaux de neurones artificiels (ANN)

Le modèle ANN nécessite une base de données numérique normalisée c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer la performance de la prédiction sur les ensembles de test de **Data_numeric_scaled**.



C'est un graphe montrant les deux courbes de Roc du modèle ANN sur la base de donnes : Data_numeric_scaled

Figure 50 : Courbe de Roc du modèle ANN

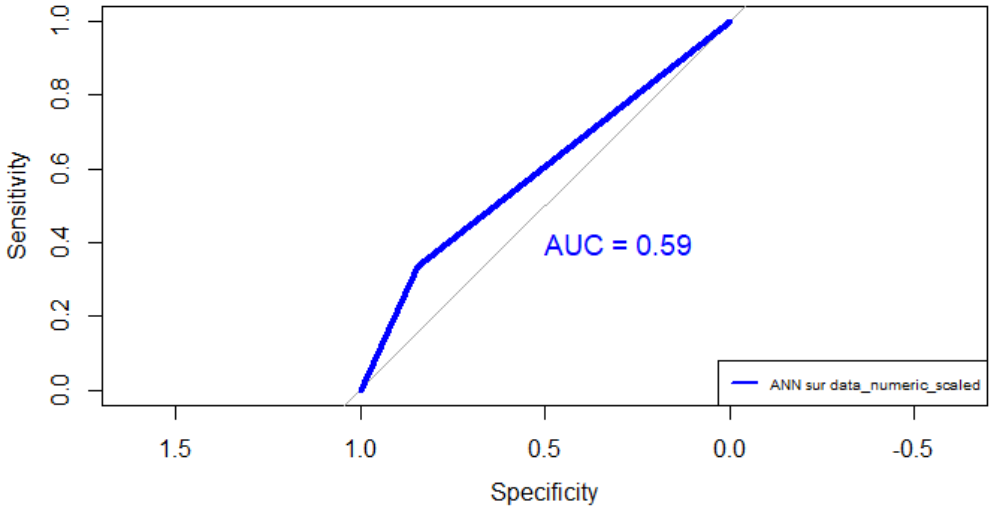
La courbe de roc de la base de données Data_numeric_scaled a une valeur d'AUC=79% ,on peut dire que c'est une bonne prédiction mais pas la meilleurs, mais il faut prendre en considération le nombre de couches cachées qui affectent aussi a cette prédiction.

Autres mesures métriques :

1. **Accuracy :** Le modèle sur la base de données numérique Data_numeric_scaled à une valeur de **0.7871** avec un intervalle de confiance **(0.7822, 0.7919).**
2. **Sensibilité et Sensitivité :** le modèle ANN sur la base de données Data_numeric_scaled a donnée **0.8099** comme sensibilité et **0.7642** comme spécificité.

Modèle 8 : Machine a vecteur de support (SVM)

Le modèle SVM nécessite une base de données numérique normalisée c'est pour cela on va l'appliquer sur les ensembles d'entrainement pour comparer la performance de la prédiction sur les ensembles de test de **Data_numeric_scaled**.



C'est un graphe montrant les deux courbes de Roc du modèle SVN sur la base de donnes : Data_numeric_scaled

Figure 51: Courbe de Roc du modèle ANN

La courbe de roc de la base de données Data_numeric_scaled a une valeur d'AUC=59%. On peut dire que ce n'est pas une bonne prédiction mais pas la mauvaise car le modèle KNN a une valeur de 50% plus petite que 59%

Autres mesures métriques :

1. **Accuracy :** Le modèle sur la base de données numérique Data_numeric_scaled à une valeur de **0.8847** avec un intervalle de confiance **(0.873, 0.8956).** Mais dans l'article 5 l'accuracy vaut 87.32% plus grand que 59%
2. **Sensibilité et Sensitivité :** le modèle ANN sur la base de données Data_numeric_scaled a donnée **1** comme sensibilité et **0** comme spécificité.

Comparaison des 8 modèles :

Tableau de comparaison :

Tableau 5 : Tableau de comparaison des résultats (1)

		Different Models			
		DT	**RF**	**LR**	**NB**
Metrics mesures	**Accuracy**	0.6266	0.6299	0.8265	0.7831
	95% CI	(0.6209, 0.6323)	(0.6242, 0.6356)	(0.822, 0.8309)	(0.7782, 0.7879)
	No Info Rate	0.5004	0.5004	0.5004	0.5004
	P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
	Kappa	0.2529	0.2595	0.6529	0.5662
	Mcnemar's Test P-Value	< 2.2e-16	< 2.2e-16	0.0001915	6.85E-07
	Sensitivity	0.336	0.2906	0.8359	0.7694
	Specificity	0.9168	0.9688	0.817	0.7968
	Positive Pred Value	0.8014	0.9029	0.8206	0.7913
	Negative Pred Value	0.5803	0.5776	0.8325	0.7753
	Prevalence	0.4996	0.4996	0.5004	0.5004
	Detection Rate	0.1679	0.1452	0.4183	0.385
	Detection Prevalence	0.2095	0.1608	0.5097	0.4865
	Balanced Accuracy	0.6264	0.6297	0.8265	0.78
	AUC	0.63	0.89	0.86	0.88
TIME EXECUTION	**min**	617.8131	464071.5526	124.6368	17.2113
	lq	619.7488	475315.2889	143.3551	17.8563
	mean	635.01084	482682.875	157.18664	22.89908
	median	626.7372	480603.6535	159.5029	18.36105
	uq	643.3902	491769.5943	170.1162	18.7216
	max neval	685.695	503334.1681	196.3343	64.2936
	cl	10	10	10	10
	d	ab	c	b	b

1. En termes de précision **Accuracy**, les résultats varient de 0,6266 pour le modèle **Decision Tree** (DT) à 0,8265 pour le modèle **Logistic Regression** (LR). Cela indique que le modèle LR a la plus haute précision globale parmi les quatre modèles.
2. En ce qui concerne la sensibilité : **Sensitivity**, le modèle **Logistic Regression** (LR) se distingue également avec une valeur de 0,8359, indiquant sa capacité à bien détecter les clients qui feront un dépôt. Le modèle **Naive Bayes** (NB) suit de près avec une sensibilité de 0,7694.
3. Du côté de la spécificité : **Specificity**, qui mesure la capacité à bien identifier les clients qui ne feront pas de dépôt, le modèle **Random Forest** (RF) obtient la meilleure performance avec une valeur de 0,9688. Suivi par le modèle **Logistic Regression** (LR) avec une spécificité de 0,817.
4. En termes de **valeur prédictive positive : Positive Pred Value**, le modèle **Random Forest** (RF) affiche la valeur prédictive positive la plus élevée (0,9029), tandis que le modèle **Decision**

Tree (DT) présente la **valeur prédictive négative : Negative Pred Value** la plus élevée (0,5803).

5. En ce qui concerne **l'AUC** : une mesure de la capacité globale de discrimination du modèle, le modèle **Logistic Regression** (LR) obtient la valeur la plus élevée (0,86), suivi du modèle **Naive Bayes** (NB) avec 0,88.
6. Enfin, en termes de **temps d'exécution,** le modèle **Decision Tree** (DT) est le plus rapide, avec un temps d'exécution minimum de 617,8131 secondes, tandis que le modèle **Random Forest** (RF) est le plus lent, avec un temps d'exécution maximum de 503334,1681 secondes.

En conclusion, parmi les quatre modèles évalués, le modèle ***Logistic Regression*** **(LR)** se démarque avec une précision élevée, une sensibilité solide et une spécificité raisonnable.

Tableau 6 : Tableau de comparaison des résultats (2)

		Different Models			
		ONE-R	**KNN**	**ANN**	**SVM**
Metrics mesures	**Accuracy**	0.7144	0.6446	0.7871	0.8847
	95% CI	(0.7091, 0.7197)	(0.639, 0.6502)	(0.7822, 0.7919)	(0.873, 0.8956)
	No Info Rate	0.5001	0.5004	0.5004	0.8847
	P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16	0.5139
	Kappa	0.4288	0.289	0.5741	0
	Mcnemar's Test P-Value	< 2.2e-16	< 2.2e-16	< 2.2e-16	<2e-16
	Sensitivity	0.4991	0.8954	0.8099	1
	Specificity	0.9296	0.3935	0.7642	0
	Positive Pred Value	0.8764	0.5965	0.7748	0.8847
	Negative Pred Value	0.65	0.7897	0.8006	NaN
	Prevalence	0.4999	0.5004	0.5004	0.8847
Metrics mesures	**Detection Rate**	0.2495	0.448	0.4052	0.8847
	Detection Prevalence	0.2847	0.751	0.5231	1
	Balanced Accuracy	0.7144	0.6444	0.7871	0.5
	AUC	0.72	0.5	0.79	0.5
TIME EXECUTION	**min**	58.2629	6372.2418	281.4024	45.955
	lq	65.4899	6379.3752	300.9274	46.1978
	mean	71.25973	6528.52347	314.75139	47.10383
	median	68.49065	6568.8896	306.91515	46.45085
	uq	71.1151	6626.5168	340.9251	46.9497
	max neval	107.2104	6709.4158	356.8964	52.2234
	cl	10	10	10	10
	d	b	a	b	b

1. En termes de précision **Accuracy**, les résultats varient de 0,6446 pour le modèle **KNN** à 0,8847 pour le modèle **SVM**. Cela indique que le modèle SVM atteint la plus haute précision globale parmi les quatre modèles.
2. En ce qui concerne la sensibilité : **Sensitivity,** qui mesure la capacité à détecter les clients qui feront un dépôt, le modèle **SVM** se distingue avec une valeur de 1, suivi du modèle **ANN** avec une sensibilité de 0,8099.
3. Du côté de la spécificité : **Specificity**, qui mesure la capacité à identifier correctement les clients qui ne feront pas de dépôt, le modèle **ONE-R** obtient la meilleure performance avec une valeur de 0,9296.
4. En termes de valeur prédictive positive : **Positive Pred Value**, le modèle **SVM** affiche la valeur la plus élevée (0,8847), ce qui indique qu'il prédit avec précision les clients qui feront un dépôt. Et en termes de valeur prédictive négative : **Negative Pred Value,** le modèle **ANN** affiche la valeur la plus élevée (0.8006)
5. En ce qui concerne **l'AUC**, qui mesure la capacité globale de discrimination du modèle, le modèle **ANN** obtient la valeur la plus élevée (0,79), suivi du modèle **ONE-R** avec une AUC de 0,72.
6. En termes de **temps d'exécution**, le modèle **SVM** est le plus rapide, avec un temps d'exécution minimum de 45,955 secondes et un temps d'exécution maximum de 52,2234 secondes. Le modèle **KNN** est le plus lent, avec un temps d'exécution minimum de 6372,2418 secondes et un temps d'exécution maximum de 6709,4158 secondes.

En conclusion, parmi les quatre modèles, le modèle **Support vector Machine (SVM)** se démarque avec une précision élevée, une sensibilité parfaite et une valeur prédictive positive élevée.

Finalement comparant SVM et LR :
La régression logistique a montré une précision (Accuracy) de 0,8265 ce qui est légèrement supérieur à la précision de 0,7871 obtenue par le modèle SVM. Cela indique que la régression logistique a une meilleure performance globale en termes de prédiction des dépôts des clients. (même résultat obtenus dans l'article 5 : LR le meilleur modèle)

Cependant, il est important de noter que le modèle SVM a atteint une sensibilité (Sensitivity) parfaite de 1, ce qui signifie qu'il a réussi à détecter tous les clients qui feront un dépôt. En revanche, la régression logistique a une sensibilité de 0,8359, ce qui indique qu'elle a manqué de prédire certains clients qui effectueront un dépôt.

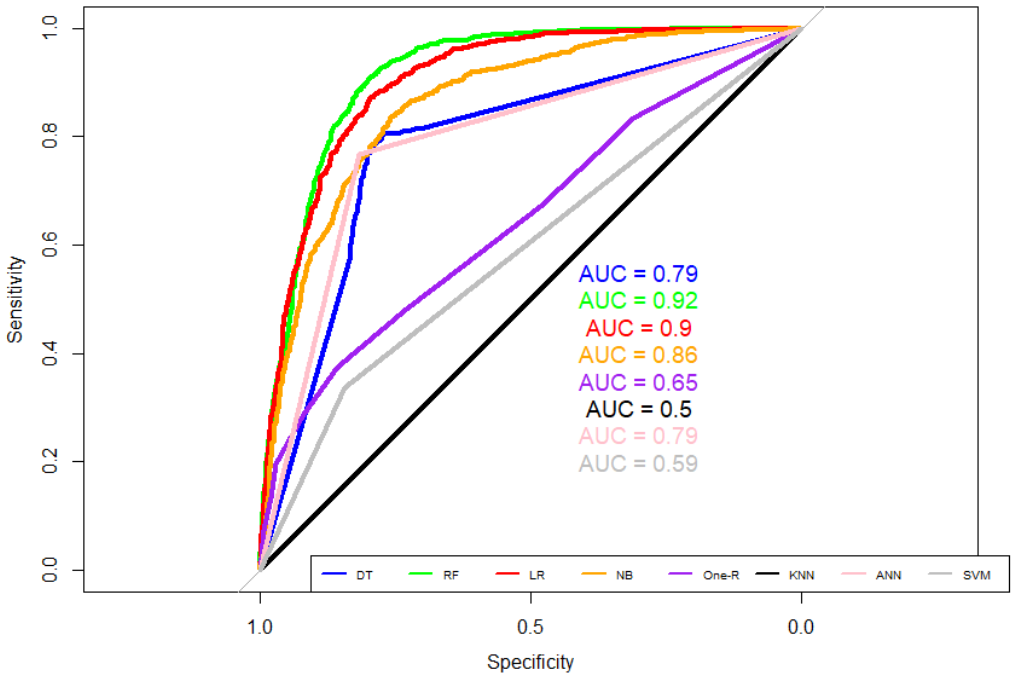
Du côté de la spécificité (Specificity), la régression logistique a obtenu un score de 0,817, tandis que le modèle SVM a une spécificité de 0. Cela signifie que la régression logistique a mieux réussi à identifier correctement les clients qui ne feront pas de dépôt, par rapport au modèle SVM.

En considérant ces éléments, la régression logistique semble offrir un meilleur équilibre entre la précision globale et la capacité à identifier les clients qui ne feront pas de dépôt. Mais, si on veut détecter les clients qui vont effectuer un dépôt alors le modèle SVM pourrait être plus approprié avec sa sensibilité parfaite de 1.

En résumé, la régression logistique est recommandée si on recherche une prédiction globale précise avec une bonne capacité à identifier les clients sans dépôt. Mais Si l'objectif principal est de détecter tous les clients qui feront un dépôt, le modèle SVM peut être plus adapté.

Représentation graphique :

Courbe de Roc des 8 modèles

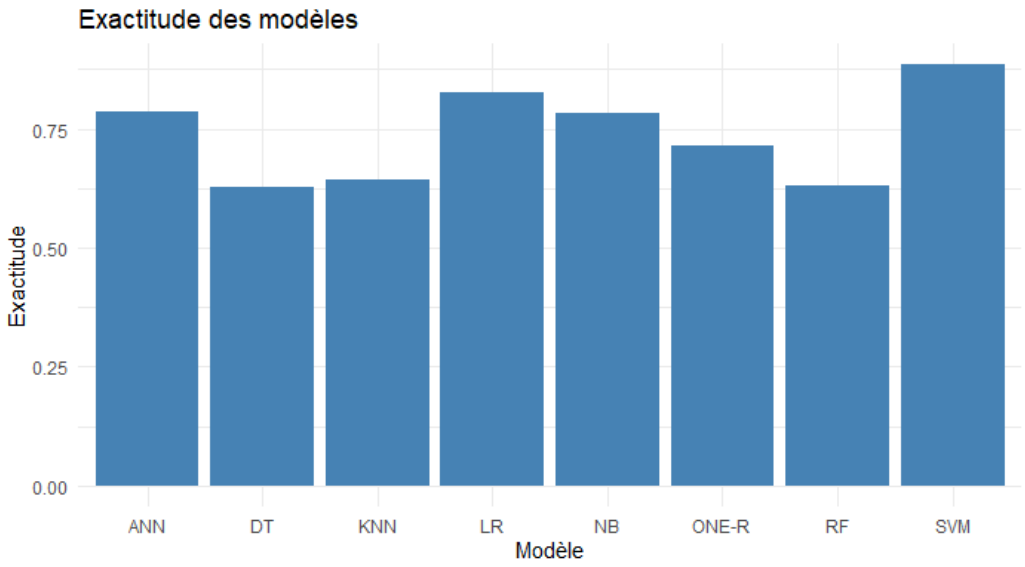


Dans cette figure on peut voir les 8 courbes de Roc avec la valeur d'AUC. Comme nous avons vu que les 3 meilleurs courbes correspondent aux : RF de 92%, LR de 90% et NB modèle 86% . Mais bien sûr qu'il faut prendre en considération les autres mesures métriques.

Figure 52: Courbe de Roc des 8 modèles

Exactitude des 8 modèles :

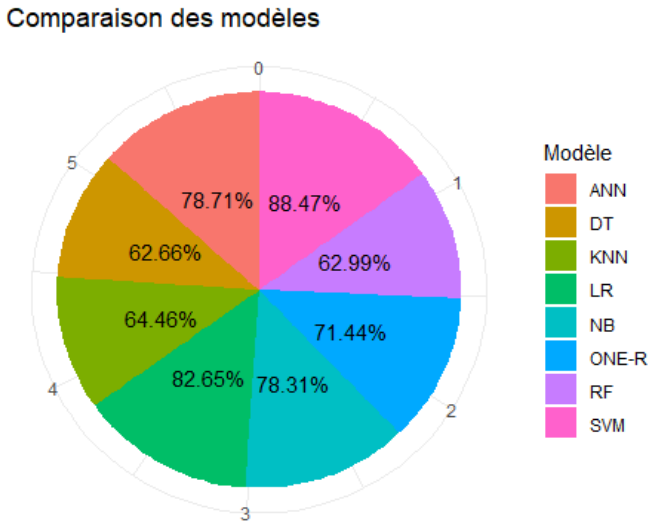
*D*iagramme en bar :



Représentation Graphique de l'Accuracy des 8 modèles. Comme nous avons vu que les 3 meilleurs accuracy correspondent aux modèles : SVM de 88,47%, LR de 82,65% et ANN de 78,71%

Figure 53: Diagramme en bar représentant l'exactitude des 8 modèles

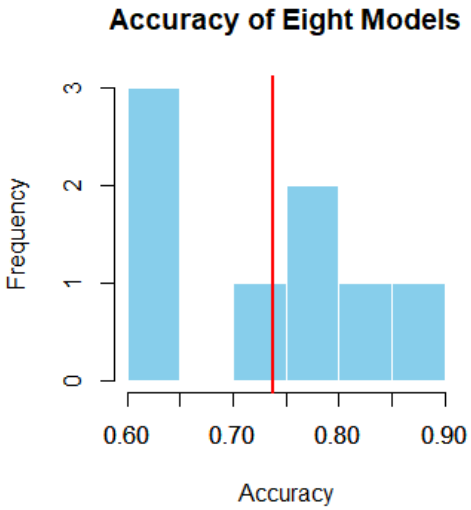
Diagramme circulaire :

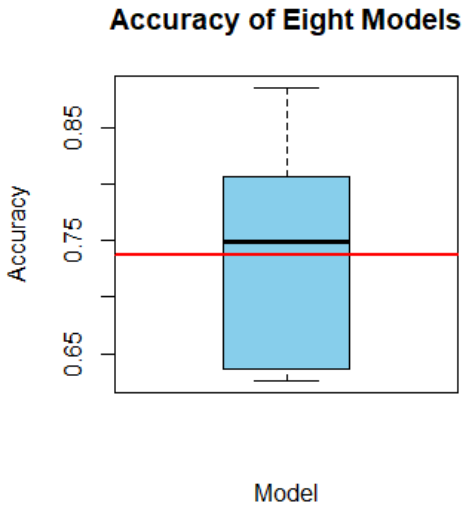


On peut visualiser l'exactitude dans une autre representation graphique

Figure 54 : Diagramme circulaire représentant l'exactitude des 8 modèles

Histogramme et boite a moustache :





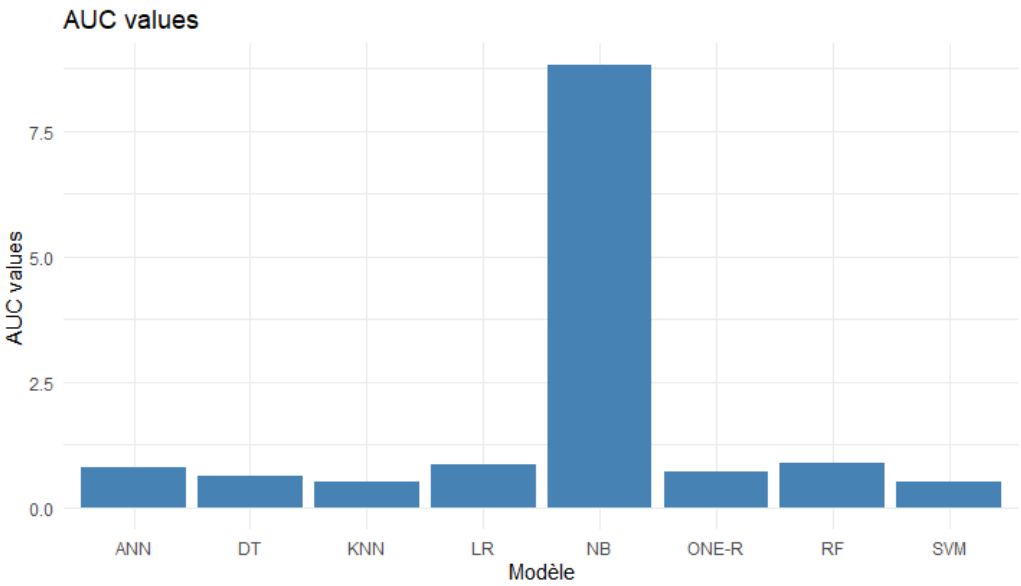
C'est une representation qui permette de combiner l'exactitude de ces 8 modeles pour montrer que la moyenne d'exactitude de ces 8 models est inferieur a 75%.

Figure 55 : Représentation Graphique de la moyenne des valeurs d'exactitude

Donc d'apres ces 3 representation des differentes valeurs d'exactitude, on constate que les modeles SVM, ANN et LR ont les valeurs les plus elevees et que la mayenne de ces valeurs vaut a peu pres 75% ce qui montre que ces 8 modeles sont bien choisits.

AUC des 8 modèles :

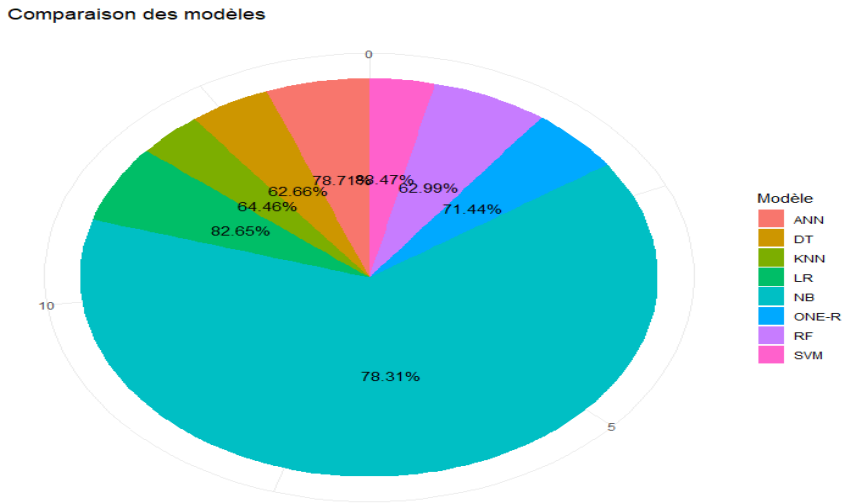
Diagramme en bar :



Représentation Graphique de l'Accuracy des 8 modèles. Comme nous avons vu que les 3 meilleurs accuracy correspondent aux modèles NB, RF et LR

Figure 56: Diagramme en bar représentant l'AUC des 8 modèles

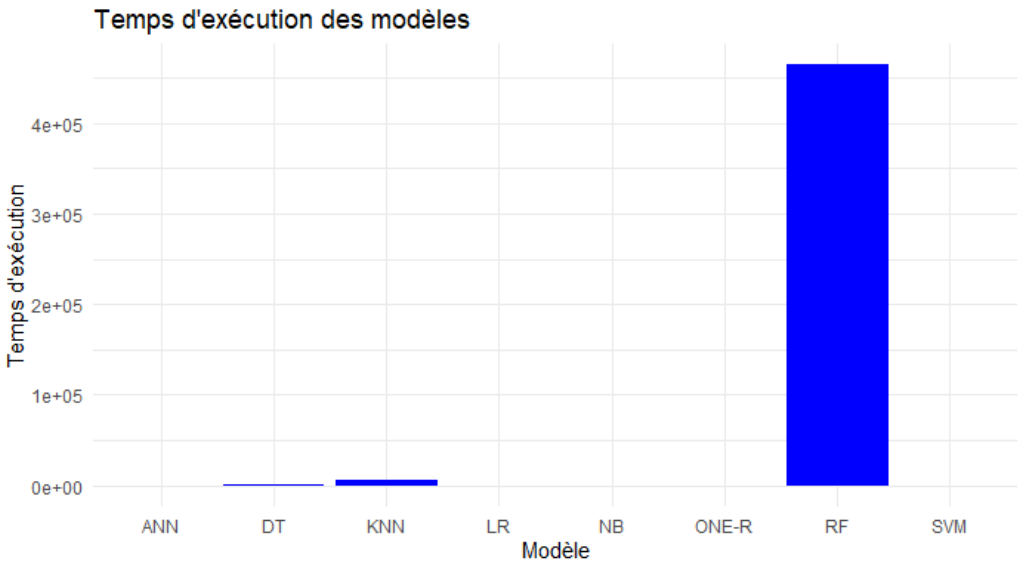
Diagramme circulaire :



On peut visualiser l'AUC dans une autre representation graphique

Figure 57: Diagramme circulaire représentant l'AUC des 8 modèles

Temps d'exécution des modèles :



C'est une représentation graphique du temps d'exécution de chaque modèle. Comme nous avons vu dans le tableau que le modèle le plus durable est RF suivi par KNN et DT.

Figure 58: Représentation Graphique du temps d'exécution des 8 modèles

Conclusion :

On a analysé la base de données et essayé de réduire la dimension en utilisant différente méthodes, ensuite on a appliqué 8 modèles pour comparer les performances, évalué leurs résultats à l'aide de mesures appropriées telles que l'exactitude, la précision, le rappel et la courbe ROC...

En conclusion, pour choisir un meilleur modèle sa dépend du l'objective de l'étude car chaque modèle offre une approche unique et chacun a ces forces et ces faiblesses.

Référence :

1. (PDF) A data modeling approach for classification problems: application to bank telemarketing prediction (researchgate.net) → (ARTICLE 1)
2. (PDF) Optimizing the prediction of telemarketing target calls by a classification technique (researchgate.net) → (ARTICLE 2)
3. Journal of Service Science and Management - SCIRP → (ARTICLE 3)
4. (PDF) Mining a Marketing Campaigns Data of Bank (researchgate.net) → (ARTICLE 4)
5. (PDF) Identifying Long-Term Deposit Customers: A Machine Learning Approach (researchgate.net) → (ARTICLE 5)
6. https://www.r-bloggers.com/2021/08/how-to-plot-categorical-data-in-r-quick-guide/
7. https://rstudio-pubs-static.s3.amazonaws.com/224337_f0de438bd82e4a769e55e039e33b6a0a.html
8. https://odr.inrae.fr/intranet/carto/cartowiki/index.php/Statistiques_descriptives_avec_R
9. https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
10. https://www.r-bloggers.com/2015/10/imputing-missing-data-with-r-mice-package/
11. https://jtr13.github.io/spring19/Community_Contribution_Group16.html
12. http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs
13. https://www.geeksforgeeks.org/create-a-plot-matrix-of-scatterplots-in-r-programming-pairs-function/
14. http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/85-afdm-dans-r-avec-factominer-scripts-faciles-et-cours/
15. https://mate-shs.cnrs.fr/actions/tutomate/tuto32-les-analyses-factorielles-multiples-afm-amand/analyse-factorielle-multiple-avec-r/
16. http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/76-afdm-analyse-factorielle-des-donnees-mixtes-avec-r-l-essentiel/
17. https://www.datacamp.com/tutorial/decision-trees-R
18. https://www.r-bloggers.com/2021/04/random-forest-in-r/
19. https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/
20. https://www.r-bloggers.com/2021/04/naive-bayes-classification-in-r/
21. https://www.r-bloggers.com/2013/04/stock-market-predictions-with-artificial-neural-networks/
22. https://www.r-bloggers.com/2023/03/one-class-svm/
23. https://www.r-bloggers.com/2015/10/using-knn-classifier-to-predict-whether-the-price-of-stock-will-increase/
24. https://www.r-bloggers.com/2023/03/implementing-a-one-step-gee-algorithm-for-very-large-cluster-sizes-in-r/

25. https://www.google.com/url?sa=i&url=https%3A%2F%2Fhands-on.cloud%2Fnaive-bayes-classifier-python-tutorial%2F&psig=AOvVaw3VPLSaiDlozZFKoNW2S_cB&ust=1685710052391000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCNjQ8PmNov8CFQAAAAAdAAAAABAO
26. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FA-simple-flowchart-for-the-k-nearest-neighbor-modeling_fig1_346429285&psig=AOvVaw1CeeGi3uyT_z1byH2UuLtz&ust=1685711024586000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCKiAgsmRov8CFQAAAAAdAAAAABAR
27. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FA-Sample-Decision-Tree-Model-Figure-2-Decision-Stump-Model_fig1_275220711&psig=AOvVaw3E-nqMhQ6_FZuj9XqlTAkT&ust=1685711224536000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCNi2hq-Sov8CFQAAAAAdAAAAABAo
28. https://www.google.com/url?sa=i&url=https%3A%2F%2Fopenclassrooms.com%2Ffr%2Fcourses%2F4470406-utilisez-des-modeles-supervises-non-lineaires%2F4730716-entrainez-un-reseau-de-neurones-simple&psig=AOvVaw2FZjCuduPKlq3A48lt3-dM&ust=1685711546783000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCNCSusSTov8CFQAAAAAdAAAAABAJ
29. https://www.google.com/url?sa=i&url=https%3A%2F%2Fblent.ai%2Fblog%2Fa%2Fsvm-support-vector-machine&psig=AOvVaw0d-moLAJHO4rIE10UBpeFF&ust=1685711745923000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCPirpqOUov8CFQAAAAAdAAAAABAI
30. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.saedsayad.com%2Flogistic_regression.htm&psig=AOvVaw3xWI3iByJbwSFauvomj4vd&ust=1685711903113000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCIiC1fSUov8CFQAAAAAdAAAAABAS
31. https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FRandom_forest&psig=AOvVaw3VPEoVIsyUV5LQtSIgZx75&ust=1685712186642000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCKDlyfeVov8CFQAAAAAdAAAAABAD
32. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.codingninjas.com%2Fcodestudio%2Flibrary%2Frule-based-classification-in-data-mining&psig=AOvVaw0WGs_-JpWXT8TGV9tYeH3u&ust=1685712991872000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCLCg6PSYov8CFQAAAAAdAAAAABAf
33. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FComparison-of-three-methods-single-bagging-and-boosting-7_fig4_338362989&psig=AOvVaw08_CfQp0rYfH9fKWziaR37&ust=1685717716454000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCPiu-cKqov8CFQAAAAAdAAAAABA5
34. https://www.google.com/url?sa=i&url=https%3A%2F%2Fquantdare.com%2Fwhat-is-the-difference-between-bagging-and-boosting%2F&psig=AOvVaw0lJPecpVOp6BLek_H8Akbb&ust=1685785687442000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCOCYheOnpP8CFQAAAAAdAAAAABAw
35. https://www.google.com/url?sa=i&url=http%3A%2F%2Fwww-igm.univ-mlv.fr%2F~dr%2FXPOSE2005%2Fentrepot%2Fmultidim.html&psig=AOvVaw1WXyIzGsPRPWFePUY_XojO&ust=1685787791127000&source=images&cd=vfe&ved=0CBEQjRxqFwoTCID3-cevpP8CFQAAAAAdAAAAABAD