UNIVERSITÉ PARIS DESCARTES | USPC Université Sorbonne Paris Cité

Apprentissage Supervisé

Bayesien Naïf, KNN, LDA, QDA, Linear SVM, Non Linear SVM, Régression logistique, CART et Random Forest

M2 MLDS 2017/2018

Réalisé par :

- Ngoc Tu
- Kalidou BA

1. Introduction

L'apprentissage supervisé est le concept derrière plusieurs applications sympas de nos jours : reconnaissance faciale de nos photos par les smartphones, filtres anti-spam des emails, etc. Quand la variable à prédire prend une valeur discrète, on parle d'un problème de classification. Parmi les algorithmes de classification, on retrouve : Support Vector Machine (SVM), Réseaux de neurones, Naïve Bayes, Logistic Regression…
Chacun de ses algorithmes a ses propres propriétés mathématiques et statistiques. En fonction des données d'entrainement (Training set), et nos features, on optera pour l'un ou l'autre de ces algorithmes. Toutefois, la finalité est la même : pouvoir prédire à quelle classe appartient une donnée.
L'objectif de ce projet consiste à appliquer un certain nombre de techniques d'apprentissage supervisé (Bayesien Naïf, KNN, LDA, QDA, Linear SVM, Non Linear SVM, Régression logistique, CART et Random Forest), à travers l'étude de données synthétiques et deux cas pratiques nécessitant l'utilisation de logiciels de traitement statistique de données R.

2. Etude exploratoire préliminaire

L'étude exploratoire a pour objet d'éclairer la complexité d'un sujet, d'en montrer les différents aspects et d'en permettre une meilleure compréhension. Elles comprennent les études qualitatives et documentaires et se distinguent des études quantitatives qui ont pour vocation de donner des résultats représentatifs.

2.1. Données Synthétiques:

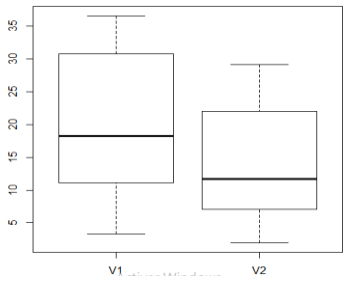
Base de données « aggregation »

Il s'agit de 3 bases de données synthétiques possédant des caractéristiques différentes, en termes de nombre de classes et de la structure des classes.

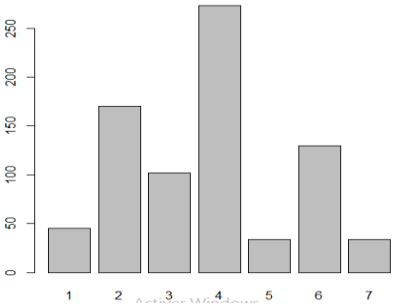
Tables	# d'observations	# de variables	# nombre de classes
Flame	240	2	2
Spiral	312	2	3
Aggregation	788	2	7

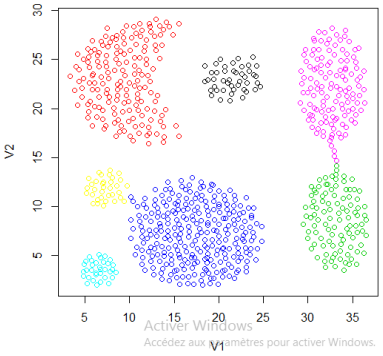
Dans la base de données «Aggregation» la boîte à moustaches des fréquences indique que la valeur médiane est de 18 pour la première variable et 11 pour la deuxième. La valeur prise par la plupart des observations est située entre 11 et 31 pour V1 entre 7 et 22 pour V2, mais la valeur de certaines observations peut baisser jusqu'à 3 pour V1, 2 pour V2 ou atteindre 37 pour V1, 29 pour la variable V2.

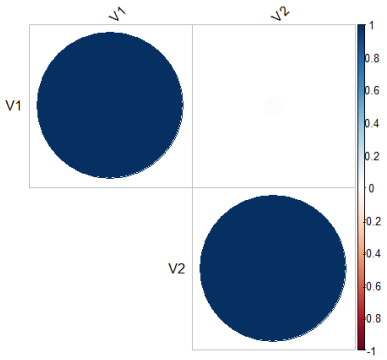
```
       V1                V2
Min.   : 3.35   Min.   : 1.950
1st Qu.:11.15   1st Qu.: 7.037
Median :18.23   Median :11.725
Mean   :19.57   Mean   :14.172
3rd Qu.:30.70   3rd Qu.:21.962
Max.   :36.55   Max.   :29.150
```



La classe 4 est la classe la plus dense avec un effectif de 273 observations tandis que la classe 5 est la moins représentative avec un effectif de 34. Visuellement les classes sont bien séparées et ont pour la plupart la même forme mais des densités qui varient selon la classe.

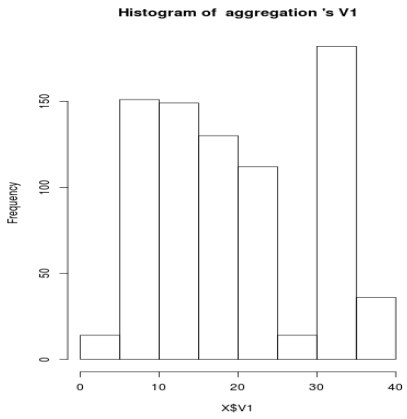


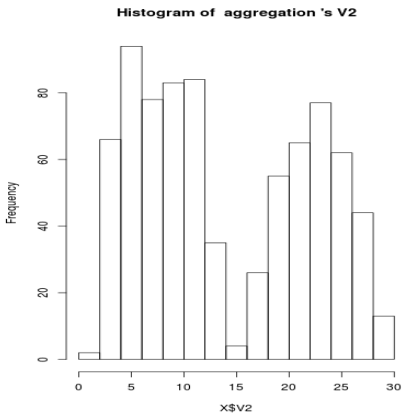





```
             V1          V2
V1 1.00000000 0.01540371
V2 0.01540371 1.00000000
```

On constate une indépendance linéaire presque totale entre les deux variables.





Nous utilisons la fonction rcorr() du package Hmisc pour calculer le niveau de significativité pour les corrélations de pearson . En utilisant cette fonction le coefficient de corrélation r de Pearson est calculé pour toutes les paires de variables possibles dans la table de données.

Comme résultat, la fonction rcorr() renvoie une liste avec les éléments suivants :

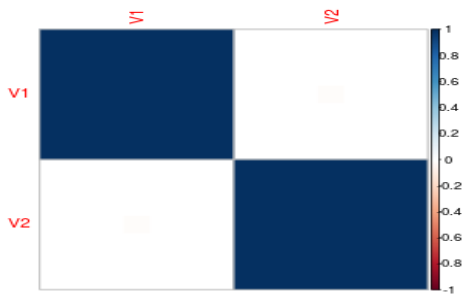
- La matrice de corrélation. La matrice du nombre d'observations utilisé dans l'analyse de chaque paire de variables.

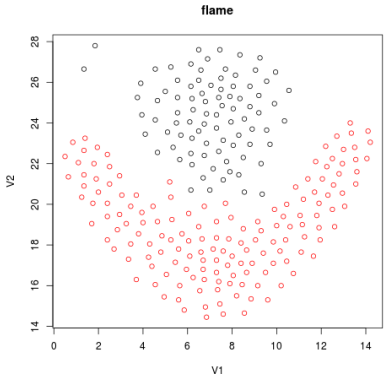
-Le p-values correspondant aux niveaux de significativité des corrélations.

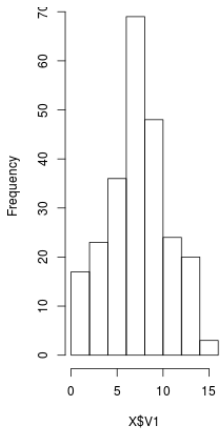
```
      V1   V2
V1 1.00 0.02
V2 0.02 1.00

n= 788


P
   V1     V2
V1        0.6659
V2 0.6659
```

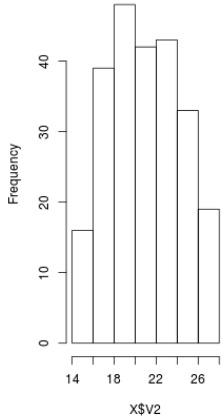
Du fait que le jeu de données « aggregation » soit pas binomiale il est impossible de construire la courbe de roc pour voir le comportement des différents algorithmes de classification selon leur taux de bien classement sur l'échantillon de teste

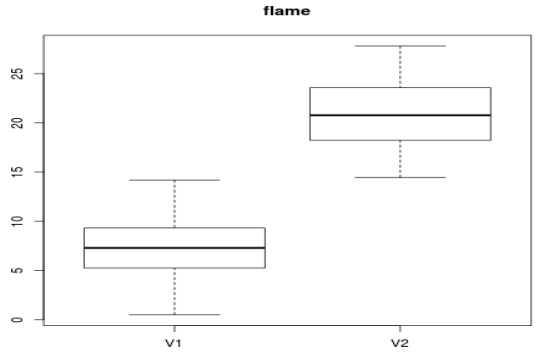
Base de données « Flam »











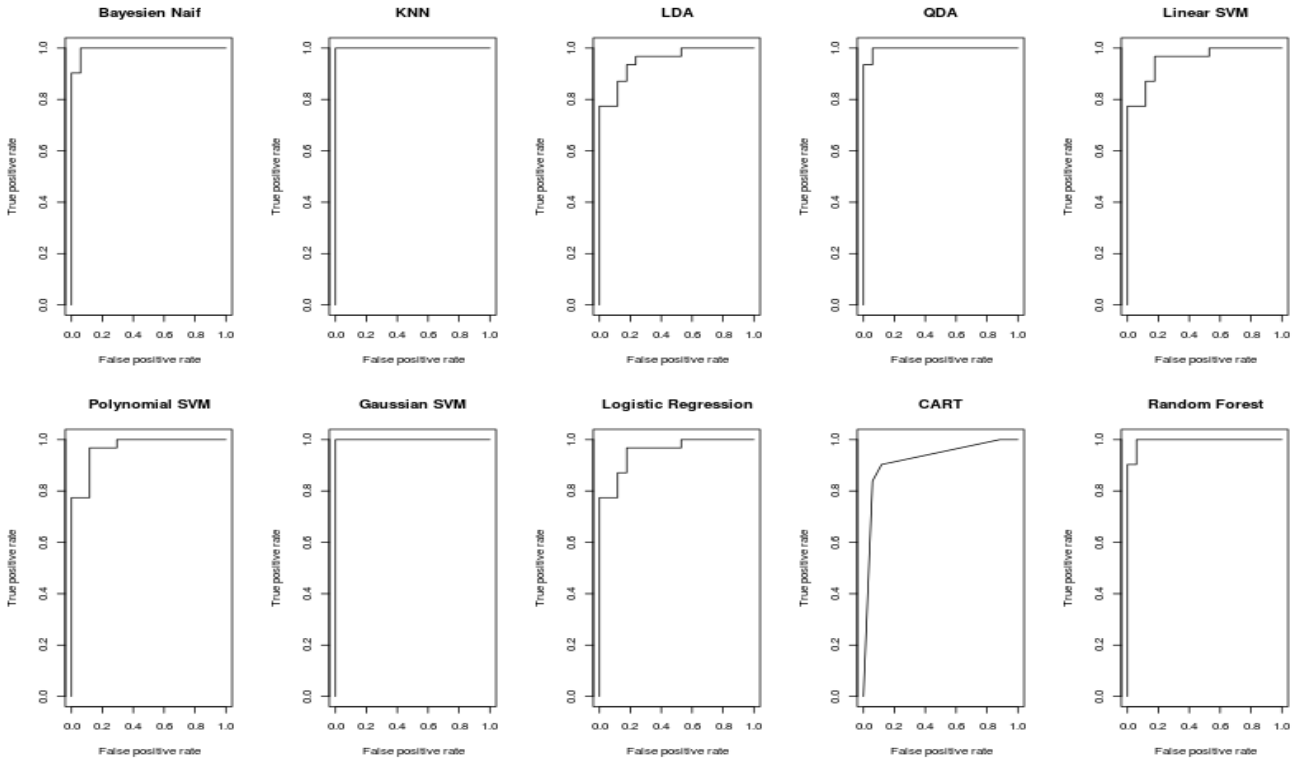
Ce jeu de données est composé d'observations classées en deux catégories ayant différentes formes et densité différentes. Ces deux catégories de données sont visuellement bien séparées

La boîte à moustaches des fréquences indique que la valeur médiane est de 6.5 pour la première variable et 21 pour la deuxième. La valeur prise par la plupart des observations est située entre 5 et 10 pour V1 entre 18 et 24 pour V2, mais la valeur de certaines observations peut baisser jusqu'à 0 pour V1, 14 pour V2 ou atteindre 14 pour V1, 27 pour la variable V2.

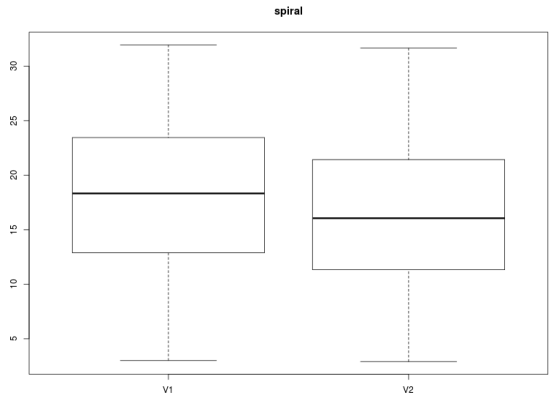
L'application des différents algorithmes de classification donne des taux de performance variables selon l'algorithme considéré.

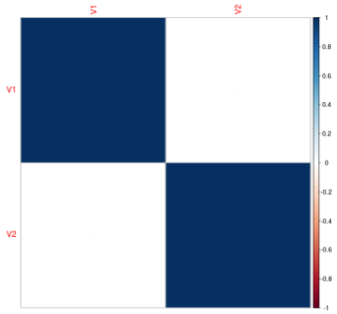
Si la courbe de roc a une forme demie rectangulaire on peut dire que l'algorithme classe parfaitement les données dans leurs classes exactes c'est le cas du KNN qui a un taux de bien classé égale à 1 et l'aire qui sépare la courbe de l'orthogonal est maximale.

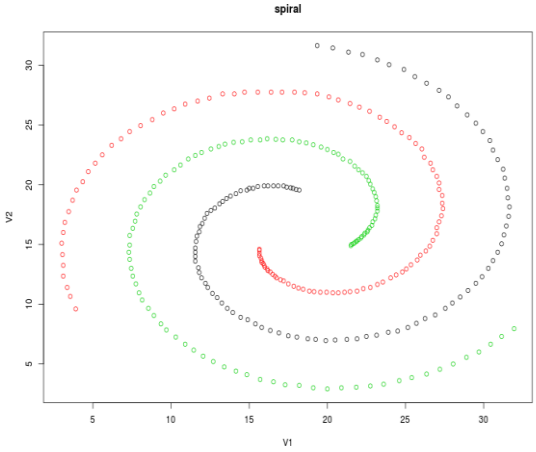
Par contre si la courbe n'est stable en ses extrémités on peut dire que l'algorithme classifie en commettant des erreurs c'est le cas de Linéaire SVM et dans ce cas la courbe se rapproche de l'orthogonal diminuant ainsi l'air.

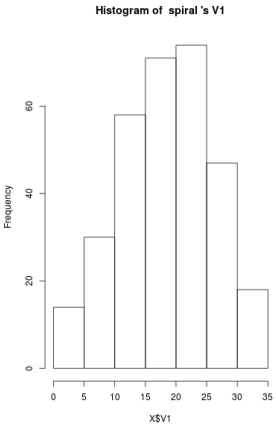


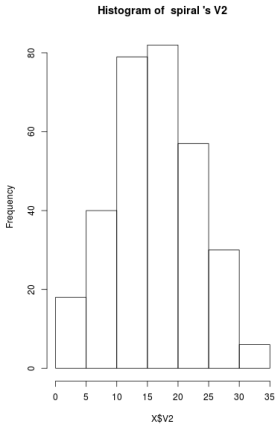
Base de données « Spiral »











Dans ce jeu de donné les variables ont presque la même distribution avec des comprise entre 2 et 35. Les différentes classe sont bien séparées et prennent des formes identiques spirales.

Puisque le jeu de données ne suit pas une distribution binomiale on ne peut donc représenté la courbe de roc.

Résumé des taux de performances obtenus en appliquant

Méthodes / Data	KNN	Naïf Bays	LDA	QDA	LSVM	PSVM	GSVM	LR	CART	RF
Flam	**1**	**0.958**	**0.875**	**0.958**	**0.875**	**0.937**	**0.979**	**0.875**	**0.895**	**0.895**
Spiral	**1**	**0.317**	**0.317**	**0.333**	**0.381**	**0.508**	**1**	**0.317**	**0.857**	**0.857**
Aggregation	**1**	**0.993**	**0.993**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

L'algorithme de KNN classifie parfaitement le différent jeu de données ce qui veut dire qu'il n'est pas sensible à la forme de distribution des classes.

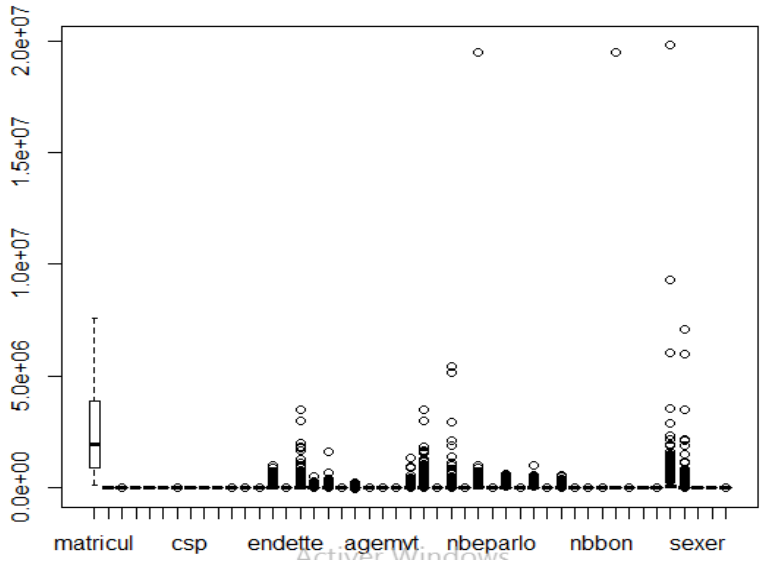
Par contre le Linéaire SVM est très sensible à la forme, si c'est spiral il trouve des difficultés de faire correctement la classification. Mais il se débrouille lorsque la forme des classes est circulaire en ayant un taux de bien classé égale à 1.

2.2. Données réelles:

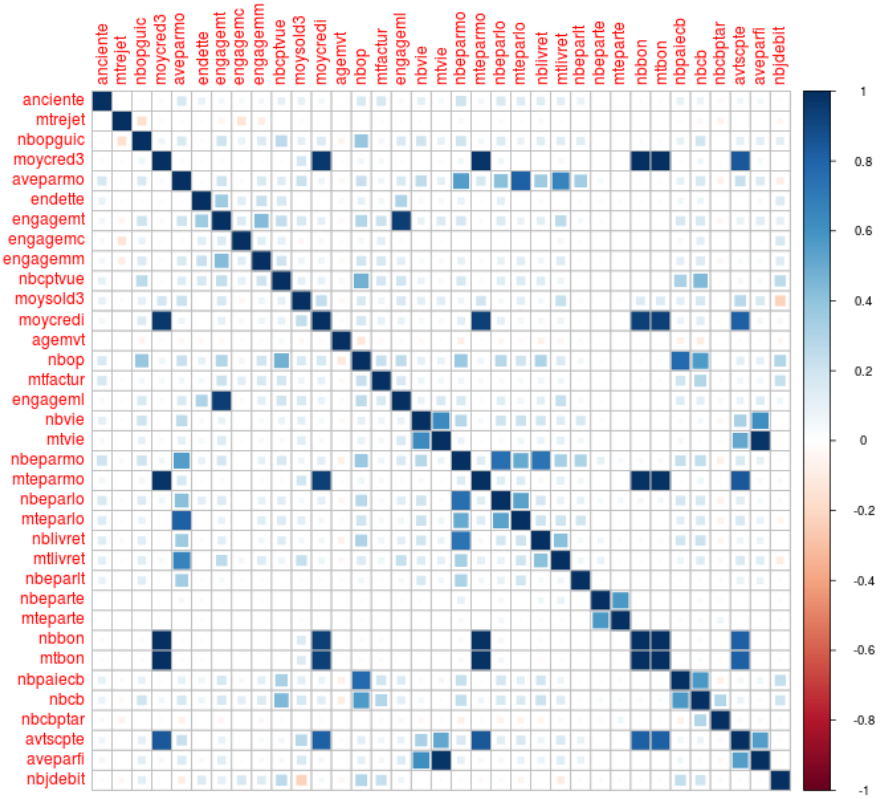
Cette partie s'intéresse à deux cas pratiques (clients d'une banque, transactions bancaires), l'objectif est d'appliquer les différentes approches vues en cours, choisir pour chaque méthode le meilleur modèle et ensuite comparer ces modèles sur un ensemble de test qui n'a pas été utilisé dans les phases d'apprentissage et de validation des modèles en concurrence.

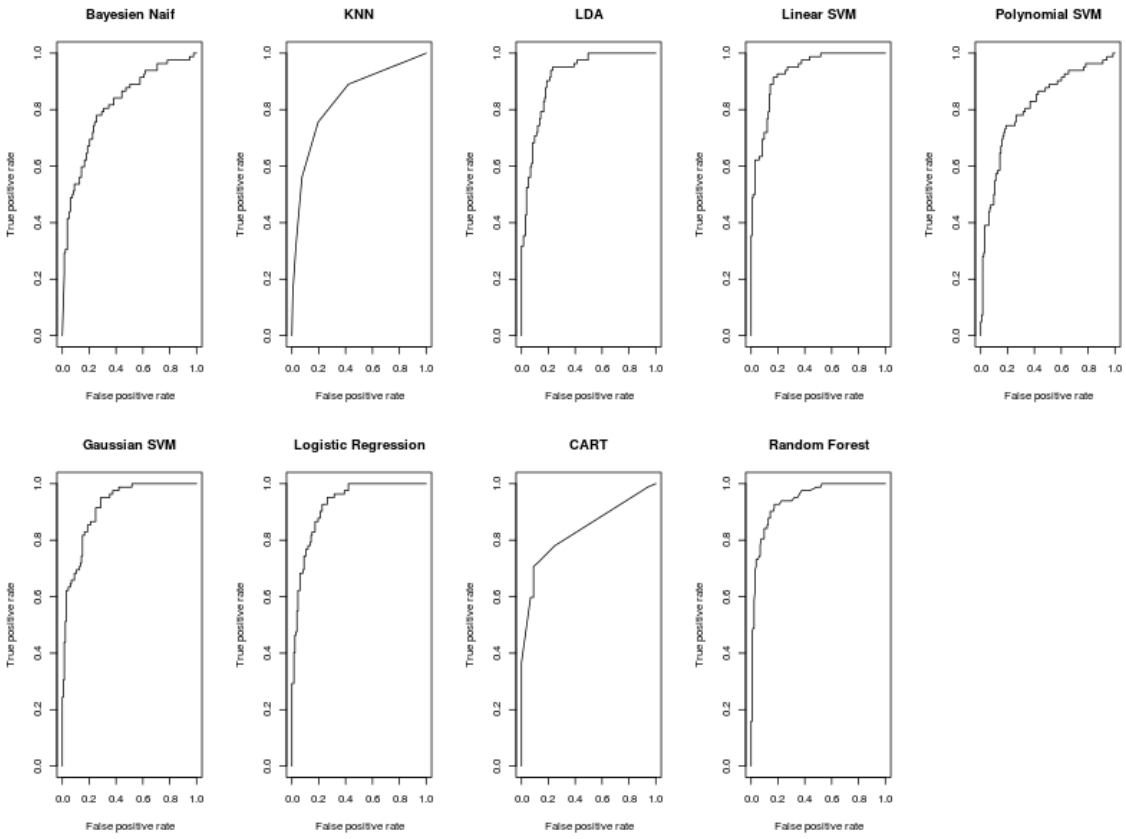
-Visa Premier

Il s'agit d'une base de données décrivant les clients d'une banque et leurs comportements (mouvements, soldes des différents comptes). La variable à expliquer Y est la variable binaire « Possession de la carte Visa Premier ». Dans cette base 714 des clients ne possèdent pas la carte Visa Premier sur un total de 1073.

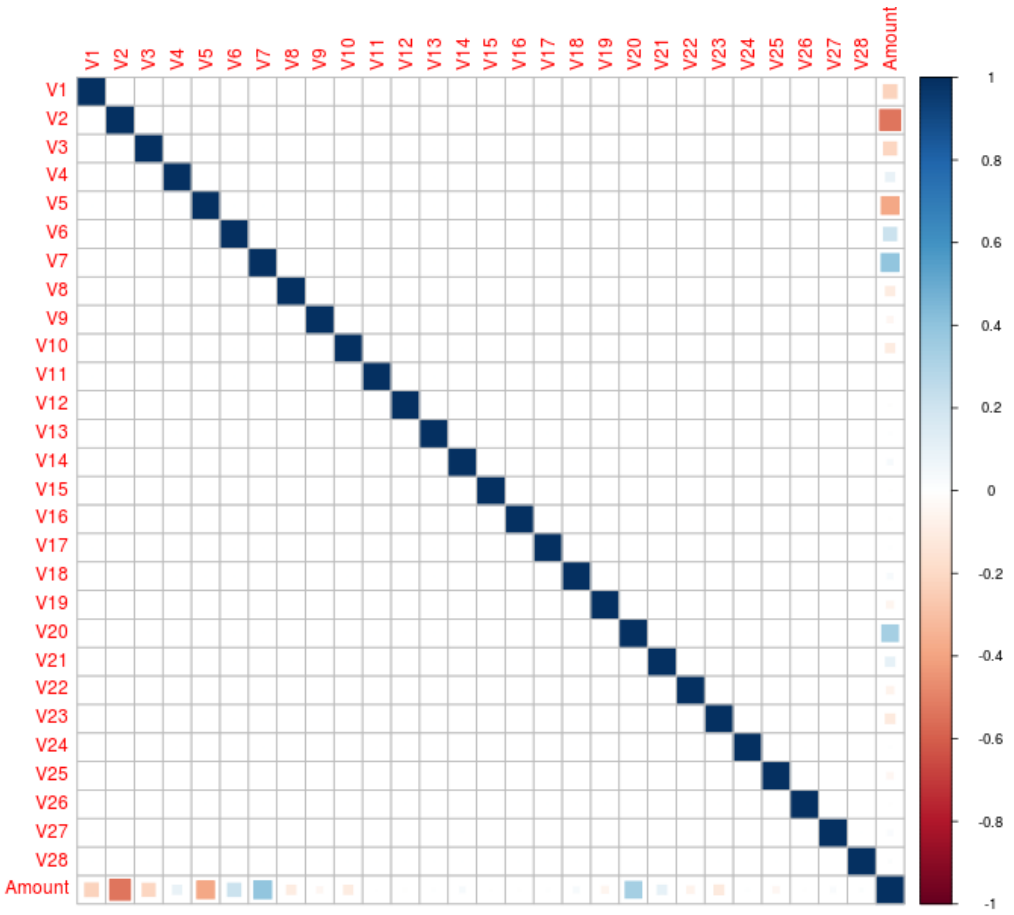


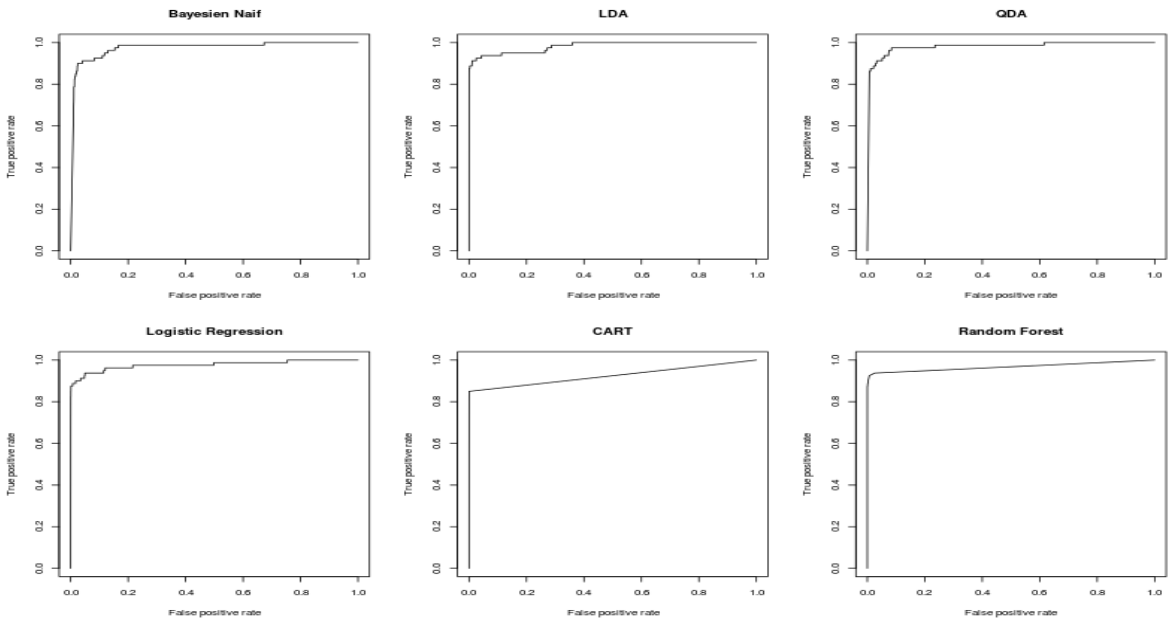
Ce boxplot donne une représentation graphique des résumés du jeu de données. Nous observons la présence de valeurs extrêmes qui méritent d'être traitées afin que nos résultats ne soient pas biaisés


anciente
mtrejet
nbopguic
moycred3
aveparmo
endette
engagemt
engagemc
engagemm
nbcptvue
moysold3
moycredi
agemvt
nbop
mtfactur
engageml
nbvie
mtvie
nbeparmo
mteparmo
nbeparlo
mteparlo
nblivret
mtlivret
nbeparlt
nbeparte
mteparte
nbbon
mtbon
nbpaiecb
nbcb
nbcbptar
avtscpte
aveparfi
nbjdebit
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Bayesien Naif
KNN
LDA
Linear SVM
Polynomial SVM
Gaussian SVM
Logistic Regression
CART
Random Forest
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

-Credit Cart


V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 Amount
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Bayesien Naif
LDA
QDA
Logistic Regression
CART
Random Forest
True positive rate
False positive rate

Résumé

Méthodes / Data	BN	KN	LDA	QDA	LSVM	PSVM	GSVM	LR	CART	RF
VisaPremier	**0.716**	**0.786**	**0.790**	**NA**	**0.827**	**0.739**	**0.832**	**0.823**	**0.832**	**0.832**
CreditCart	**0.978**	**NA**	**0.999**	**0.976**	**NA**	**NA**	**NA**	**0.999**	**0.999**	**0.99**

L'exécution du KNN es SVN sur le jeu de données CréditCart nécessite un temps d'exécution élevé.

Le bayésien naïf se comporte bien lors de la classification des observations avec un taux de 0.716 pour le VisaPremier et 0.978 pour le CreditCart
Encore mieux avec Randon Forest RF qui un taux de bien classé considérable pour les deux jeux de données.

Conclusion :

Ces différents algorithmes de classifications dépendent de plusieurs phénomènes.
Selon la forme des classes

BN bayésien Naïf

KNN K plus proche voisins

RF Randon Forest

LR Logistic Regression

LSVM linéaire SVM

PSVM Polynomial SVM

GSVM Gaussien SVM