

# Rapport du projet d'apprentissage automatique

Wassim BEN YOUSSEF

Ahmed HERMI

Karim ASSAAD

Yoann DACRUZ

Team Data Titans

Avril 2016

# Sommaire

<b>1</b>	<b>Analyse de données</b>	<b>3</b>
1.1	Analyse par histogrammes et box-plot . . . . .	3
1.2	Etude des influences de certaines variables sur la survie des individus . . . . .	6
1.2.1	L'influence du sexe de l'individu . . . . .	6
1.2.2	L'influence de la classe . . . . .	7
1.3	Les données manquantes . . . . .	7
<b>2</b>	<b>Création de nouvelles variables pour mieux trier les données</b>	<b>8</b>
2.1	Création de la variable Titre . . . . .	8
2.2	Création de la variable LastName . . . . .	8
2.3	Création des variables FamilySize et FamilyID . . . . .	8
2.4	Reconfiguration de la variable Cabin . . . . .	9
2.5	L'importance de la variable âge . . . . .	10
<b>3</b>	<b>Recherche d'information sur les données manquantes</b>	<b>11</b>
3.1	Méthodes basiques . . . . .	11
3.2	Les arbres de décision . . . . .	11
3.2.1	Le principe . . . . .	11
3.2.2	Les méthodes utilisées sous R . . . . .	12
3.2.3	Utilisation sur les variables Age, Fare, Cab et Embarked . . . . .	12
3.3	L'éventuel utilisation de Random Forest . . . . .	16
<b>4</b>	<b>Utilisation d'algorithmes performants</b>	<b>17</b>
4.1	Arbre de décision sur Survival . . . . .	18
4.2	Le Support Vector Machine . . . . .	18
4.3	Le Random Forest . . . . .	19
4.4	Le CForest, une amélioration du Random Forest . . . . .	20

# Introduction

L'apprentissage automatique est l'utilisation de méthodes permettant aux machines d'évoluer par un processus systématique pour répondre à des problématiques qu'un simple algorithme n'aurait pas la capacité de résoudre. Aujourd'hui, l'apprentissage automatique est prépondérant dans les nouvelles technologies et représente un enjeu majeur qu'un ingénieur doit savoir appréhender. Le projet présenté ici est une application des principales méthodes de "machine-learning" utilisées pour répondre à une problématique simple en se basant sur des données du problème. L'objectif est de modéliser une solution qui permette de prédire quelles sont les individus ayant survécus ou pas au naufrage du Titanic en fonction des différentes informations que nous avons sur ces personnes. Notre travail se portera sur une base de données contenant 11 variables : tout d'abord Survival qui est la variable à prédire et qui définit si l'individu a survécu ou non, puis différentes variables qui donnent des informations sur les individus : leur âge, leur nom, la classe (première classe, deuxième classe ou troisième classe), le sexe, le nombre de frères/soeurs et époux/épouse également présent sur le bateau, le nombre de parents et enfants présents sur le bateau, le numéro de ticket, le numéro de cabine, le port d'embarcation et le tarif auquel le passager a pris son billet. A noter que nous disposons de deux tableaux de données, test et train. Le tableau test nous servira à tester notre modèle, c'est pourquoi nous supprimons la colonne Survival de ce tableau pour par la suite prédire ces valeurs à l'aide du modèle, puis nous fusionnons les deux tableaux en un seul pour y effectuer notre analyse et nos démarches. En comptant ces deux tableaux, nous étudierons une population d'environ 1300 individus. Notre démarche suivra le déroulement suivant : dans un premier temps nous ferons une rapide analyse des données, puis nous détaillerons les différentes méthodes utilisées pour pallier le manque d'information dûe aux données manquantes, ensuite nous expliquerons comment nous réajustons et trions les données pour avoir une meilleur approche du problème et enfin nous verrons les différentes méthodes de machine-learning utilisées pour obtenir un modèle statistique. Pour effectuer cette étude, nous avons choisi d'utiliser le langage R.

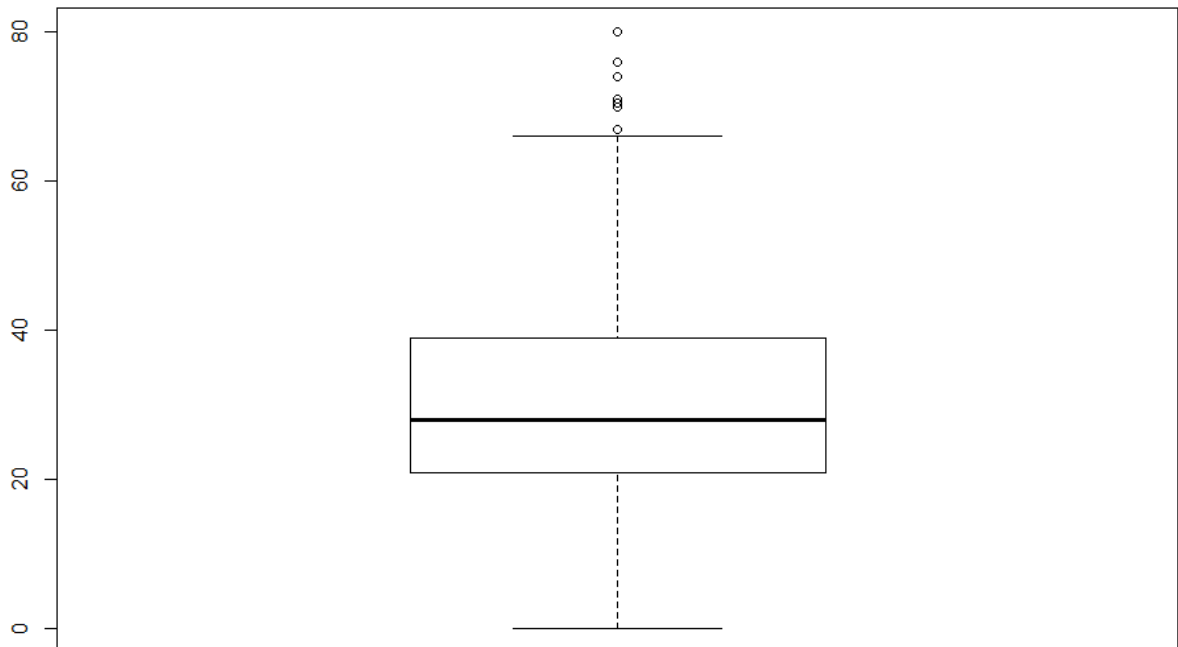
# Chapitre 1

## Analyse de données

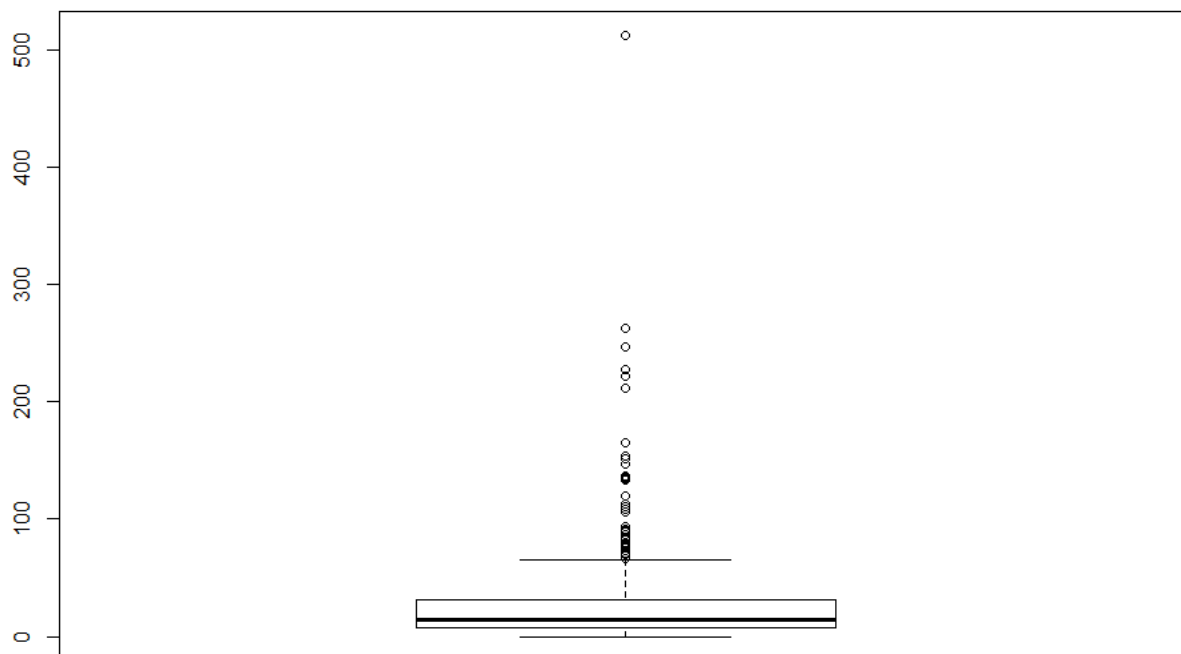
### 1.1 Analyse par histogrammes et box-plot

Dans un premier temps, nous allons effectuer une analyse exhaustive des données fournies pour bien cerner le problème.

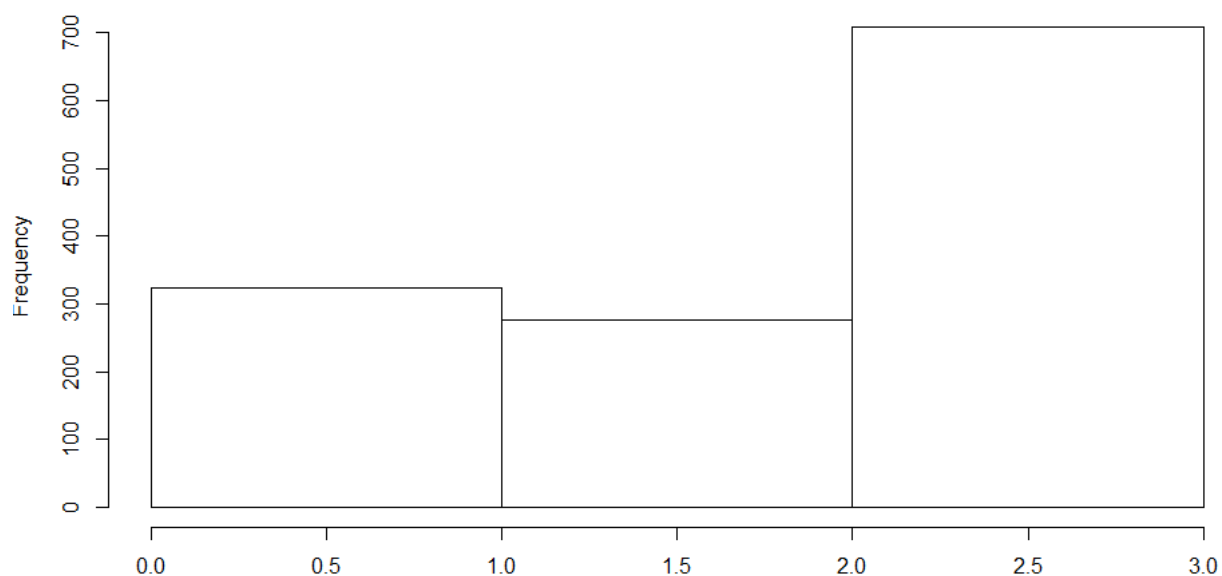
**Box-plot de l'âge des passagers**



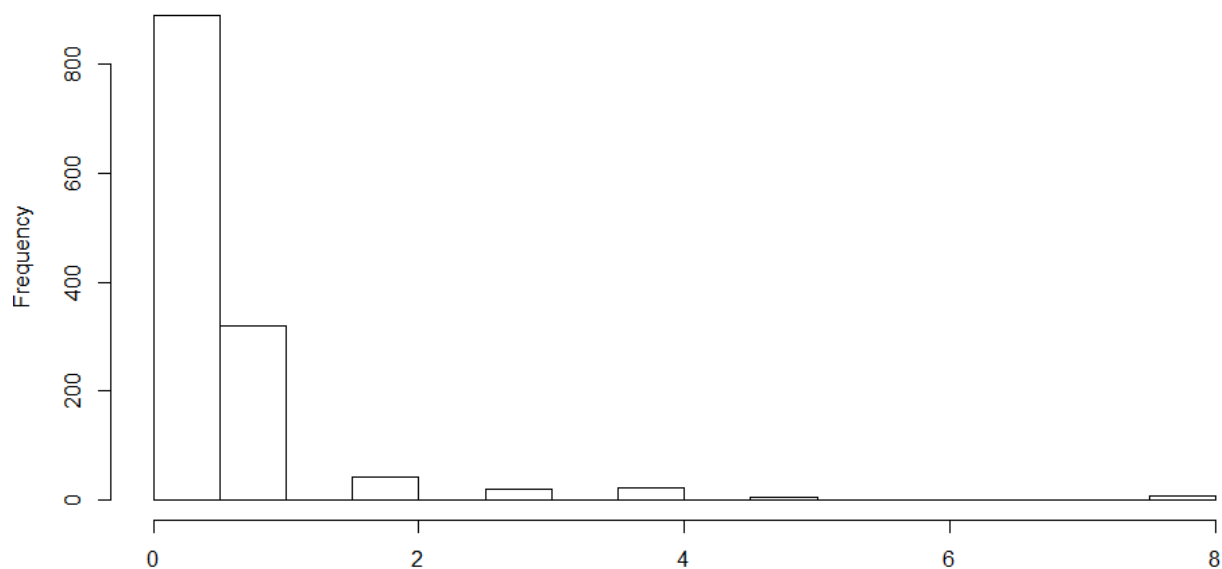
Cette box-plot indique que la majorité des passagers ont un âge compris entre 20 et 40 ans. Nous observons également qu'il y a plus de personnes qui ont un âge supérieur à 40 ans que de personnes ayant un âge inférieur à 20 ans. Cependant, seul peu de passagers sont âgés de plus de 60 ans. L'âge des passagers est ainsi très diversifié et réparti sur tout l'intervalle de 0 à 80.

**Box-plot sur le tarif payé par les passagers**

Contrairement à l'âge, le tarif payé par les passagers semble très inégal avec une grande partie des passagers ayant privilégié un billet d'une valeur inférieure à 50, et seul peu de passagers ont un tarif supérieur à 50 avec certains mais très peu qui dépassent les 100.

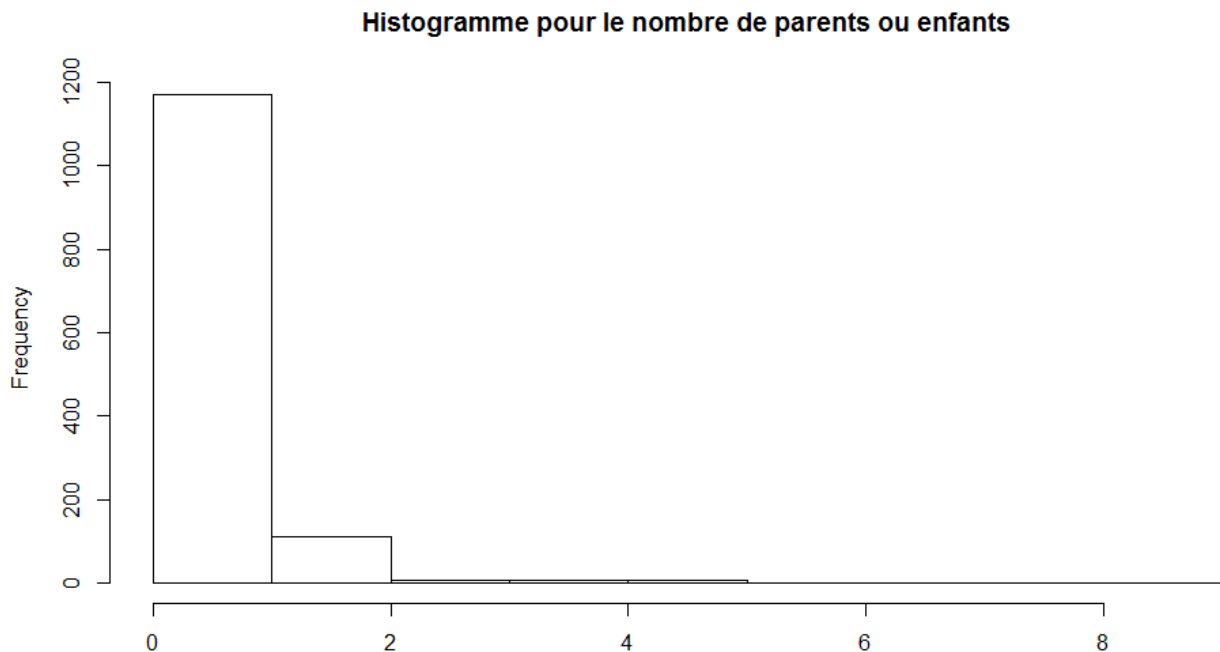
**Histogramme indiquant le nombre de passagers par classes**

Nous pouvons à nouveau sur cet histogramme apprécier les inégalités financières entre les différents passagers en observant qu'une grande partie des passagers sont en troisième classe. Nous notons cependant qu'il y a un plus grand nombre de personnes en première classe qu'en deuxième classe.

**Histogramme pour le nombre de frères et soeurs ou conjoints**

Cette histogramme indique qu'une grande majorité des passagers voyage seul ou avec des amis

et non avec des membres de leur familles, l'histogramme suivant nous le confirme également.



## 1.2 Etude des influences de certaines variables sur la survie des individus

Intuitivement, nous pouvons penser que certaines variables ont une assez grande influence sur la survie ou non de l'individu. Nous allons étudier ce phénomène sur deux variables qualitatives que sont Sex et PClass.

### 1.2.1 L'influence du sexe de l'individu

Il semble assez logique de penser que plus de femmes ont été sauvées lors du naufrage que d'hommes. Nous allons, pour prouver cela, calculer le pourcentage de morts/survivants pour chacun des deux sexes.

	0	1
female	0.2579618	0.7420382
male	0.8110919	0.1889081

Sur ce tableau, nous voyons qu'environ 74% des femmes ont été sauvées tandis que plus de 80% des hommes n'ont pas survécus au naufrage. Nous pouvons ainsi confirmer que la variable Sex est très influente sur le résultat finale.

### 1.2.2 L'influence de la classe

Nous pouvons également penser que les plus riches passagers installés en première classe ont été plus épargés que les passagers plus modestes. Nous allons donc effectuer la même étude que précédemment :

	0	1
1	0.3703704	0.6296296
2	0.5271739	0.4728261
3	0.7576375	0.2423625

A nouveau, suite à ces calculs, nous pouvons observer que plus de 75% des individus installés en troisième classe n'ont pas survécu tandis que presque 63% des individus en première classe ont survécu, enfin en deuxième classe, il semble y avoir un équilibre avec presque 50% dans les deux cas. Nous ne pouvons donc que confirmer que l'appartenance sociale a un réel effet sur le fait d'avoir survécu ou non au naufrage.

## 1.3 Les données manquantes

Pour terminer cette analyse, nous allons calculer le nombre de données manquantes dans chacune des variables :

PassengerId	Survived	Pclass	Name	Sex	Age	Sibsp	Parch
0	418	0	0	0	263	0	0
Ticket	Fare	Cabin	Embarked				
0	1	1014	2				

Nous observons des données manquantes pour quatre variables (Cabin, Age, Fare et Embarked), dont deux ont un nombre assez important de données manquantes (Age et Fare). À noter que les valeurs manquantes de Survived ne sont pas à considérer puisqu'elles proviennent du tableau test dont nous avons enlevé la colonne Survived. Dans la prochaine partie, nous verrons comment nous avons utilisé les données déjà présentes dans le tableau pour obtenir plus d'informations et pour extraire plus d'informations.



## Chapitre 2

# Création de nouvelles variables pour mieux trier les données

Dans l'analyse de données faite au chapitre précédent, nous avons vu la réelle influence du grade social des individus. Il nous a donc paru intéressant d'ajouter de nouvelles variables pour regrouper les individus selon de nouveaux critères tout en se basant sur les informations données par le tableau. Cela s'apparente à obtenir encore plus d'informations à partir des informations déjà existante dans le but d'améliorer notre modèle.

### 2.1 Création de la variable Titre

Dans un premier temps, nous avons créé une nouvelle variable "Titre". En effet, dans la variable "Name", nous avons une information que nous ne pouvions exploiter, qui est le titre de l'individu, Mrs, Lady, Master ... Or ces titres sont très importants pour déterminer la classe sociale de l'individu, une personne ayant le titre "Don" peut être considérée comme plus aisée qu'une personne ayant le simple titre de "Mr". Nous avons donc extrait les titres des individus à partir de la variable "Name" dans laquelle les titres sont indiqués.

### 2.2 Création de la variable LastName

A partir de la variable "Name", nous avons également extrait les noms de famille des individus, ceci permettant de directement identifier les personnes de la même famille.

### 2.3 Création des variables FamilySize et FamilyID

Pour synthétiser les informations contenues dans les variables Parch et SibSp, nous avons créé tout d'abord la variable FamilySize qui compte le nombre de personnes dans la famille de l'individu en fonction de Parch et de SibSp et en comptant également l'individu lui-même. Cependant, nous avons remarqué que beaucoup d'individu avait cette variable égale à 1 ou à 2, nous avons donc créé la variable FamilyID dans laquelle tout les individus ayant leur variable FamilySize égale

à 1 ou à 2 auront comme valeur "Small" pour cette variable, ce qui permettra de directement identifier les personnes qui sont venues avec peu ou pas de familles. Les autres individus auront la même valeur pour FamilySize et pour FamilyID.

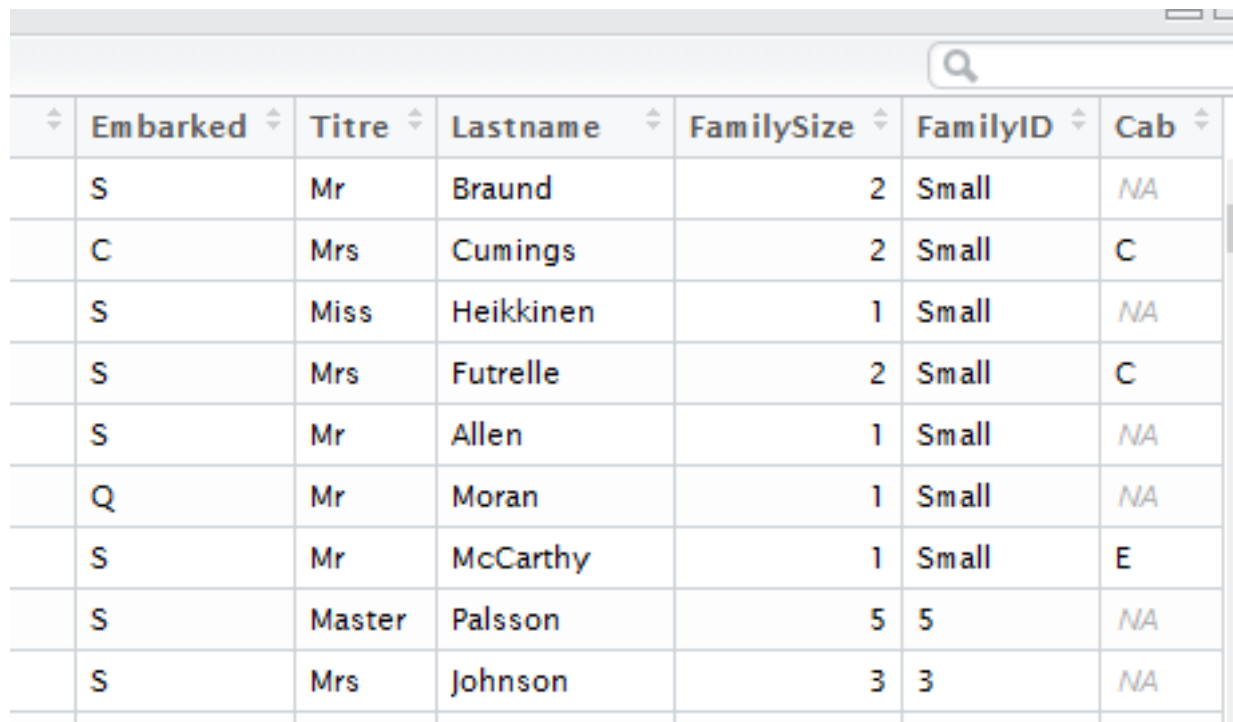
## 2.4 Reconfiguration de la variable Cabin

La variable Cabin est sûrement la plus compliquée et la plus difficile à interpréter du fait non seulement de ces nombreuses valeurs manquantes mais surtout car on ne sait pas vraiment à quoi correspond les valeurs, nous savons simplement que se sont des numéros de cabines. Nous avons donc choisi d'utiliser cette variable de deux façons : dans un premier temps nous avons extrait uniquement la lettre associée au numéro de cabine que nous avons mis dans une variable appelée "Cab". Cette nouvelle variable sera beaucoup plus simple à utiliser que la variable "Cabin".

Ensuite, nous avons essayé d'avoir un raisonnement pour interpréter cette variable. Nous avons fait comme hypothèse que chaque lettre correspond à un étage du bateau puis que le numéro 1 correspond à l'avant du bateau, alors plus le numéro est grand plus on se rapproche de l'arrière du bateau. Pour se faire, nous avons créé une nouvelle variable "Cab\_pos" qui peut prendre les valeurs "Beginning", "middle" ou "end". Cependant, nous nous sommes rendu compte plus tard que ce raisonnement n'était pas fructueux et ne permettait pas vraiment d'améliorer nos résultats. Nous avons donc abandonné cette démarche.

Voici une image de la fin du tableau considéré après les modifications expliquées précédemment

:



Embarked	Titre	Lastname	FamilySize	FamilyID	Cab
S	Mr	Braund	2	Small	NA
C	Mrs	Cumings	2	Small	C
S	Miss	Heikkinen	1	Small	NA
S	Mrs	Futrelle	2	Small	C
S	Mr	Allen	1	Small	NA
Q	Mr	Moran	1	Small	NA
S	Mr	McCarthy	1	Small	E
S	Master	Palsson	5	5	NA
S	Mrs	Johnson	3	3	NA

## 2.5 L'importance de la variable âge

Plus tard, après avoir calculé nos modèles basés sur les données expliquées précédemment, nous nous sommes rendus compte qu'il était possible d'améliorer le résultat final en créant une nouvelle variable "Child". En effet, il est logique que lors du naufrage, l'équipage a privilégié la survie des enfants ainsi que des femmes. C'est pourquoi, nous avons créé cette nouvelle variable binaire qui indique si un individu est considéré comme enfant en dessous d'un certain âge. Nous avons tout d'abord fixé le seuil à 18 ans, ce qui nous a permis d'améliorer notre résultat. Nous avons par la suite diminué cet âge à 8 ans, cela a également eu un léger impact sur le résultat en l'améliorant.

## Chapitre 3

# Recherche d'information sur les données manquantes

Précédemment, nous avons repéré de nombreuses données manquantes dans notre tableau. Celles-ci représentent un véritable problème lorsqu'il s'agit d'obtenir une modélisation statistique car elles sont un grand manque d'information. Nous ne pouvons obtenir une bonne modélisation statistique sans au préalable tenir compte de ces valeurs manquantes car sinon notre modèle serait beaucoup trop imprécis. Nous allons à présent détailler les méthodes utilisées pour remplacer ces données manquantes.

### 3.1 Méthodes basiques

Nous avons dans un premier temps utiliser des méthodes très simples pour remplacer les données manquantes du tableau. La première serait de simplement supprimer tout les individus ayant une donnée manquante. Cela semblerait totalement inefficace puisqu'une grande partie de l'information disparaîtrait, d'autant plus que nous sommes sur une population d'un peu plus de 1300 individus, nous ne pouvons nous permettre d'effacer de la liste les 1014 passagers dont nous ne connaissons le numéro de cabine, cela serait absurde. Une autre solution serait de remplacer les valeurs manquantes par la valeur de la moyenne de la variable. Cependant, cela est également inefficace car les valeurs des variables seront alors totalement biaisées. De plus, cette méthode ne peut marcher sur les variables qualitatives.

### 3.2 Les arbres de décision

#### 3.2.1 Le principe

Les méthodes énoncées précédemment étant jugées inefficaces, nous allons utiliser un outil beaucoup plus puissant et propre à l'apprentissage automatique : les arbres de décision. Le but général d'un arbre de décision est d'expliquer une valeur à partir d'une série de variables. Nous pourrions ainsi remplacer les valeurs manquantes par des valeurs "prédites" par ces arbres de décision. Le principe de l'arbre de décision se base sur une hiérarchie des variables explicatives en tenant

compte de leur capacité prédictives : on commence par utiliser la variable qui a le plus de "pouvoir de prédiction" pour séparer les individus en un nombre de classes bien défini, puis à chaque itération on utilise le même principe en prenant la variable la "plus explicative". Au terme de ce parcours, on obtient les feuilles finales de l'arbre dont chacune correspond à un chemin de l'arbre. Cela mènera à dire que chaque individu ayant ces variables identiques au chemin  $k$  aura sa valeur à expliquer égale à la valeur de la feuille  $k$ .

La principale difficulté de cette méthode repose sur le choix de variables de décision. En effet, il faut savoir choisir les bonnes variables à utiliser pour obtenir un arbre de décision optimal.

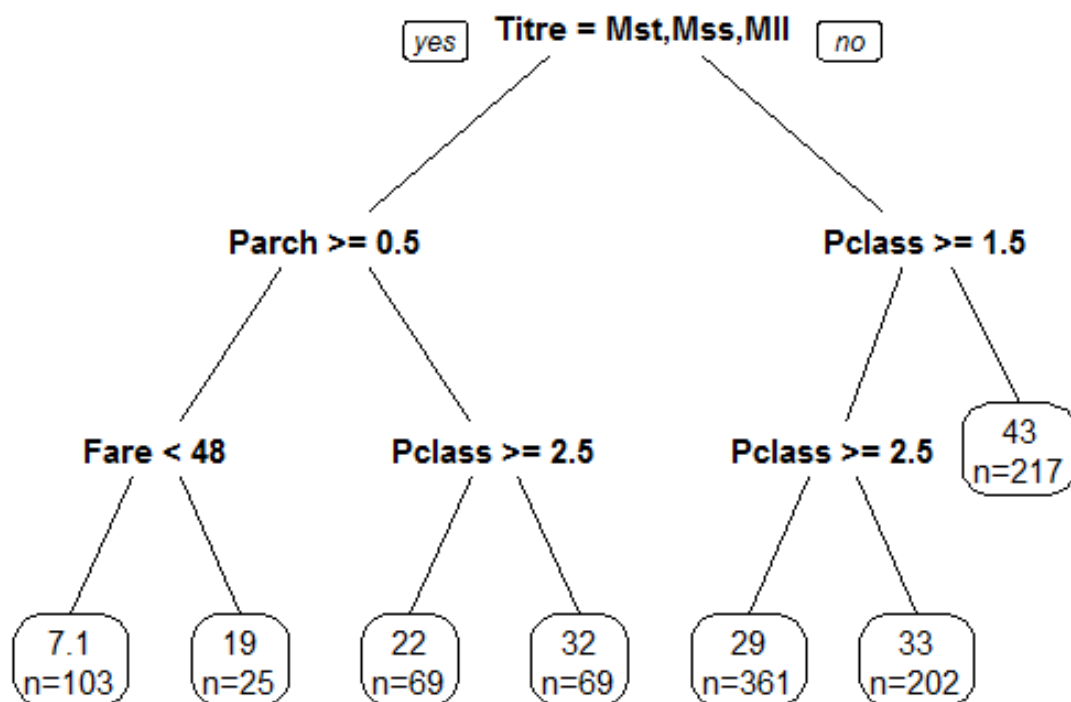
### 3.2.2 Les méthodes utilisées sous R

Pour utiliser les arbres de décision sous R, nous faisons appel aux bibliothèques "rpart", "rattle" et "rpart.plot". L'implémentation se fait en deux temps : tout d'abord on utilise la fonction "rpart" avec comme paramètre la variable à prédire, les différentes variables explicatives choisies, le tableau de données et la méthode ("class" pour les valeurs qualitatives et "anova" pour les valeurs quantitatives). Ensuite, pour obtenir optimiser le modèle obtenu, nous utilisons la fonction "prune" qui prend en paramètre le modèle obtenu par "rpart". Enfin, nous utilisons "predict" sur le résultat obtenu pour que ce modèle prédise les données du tableau.

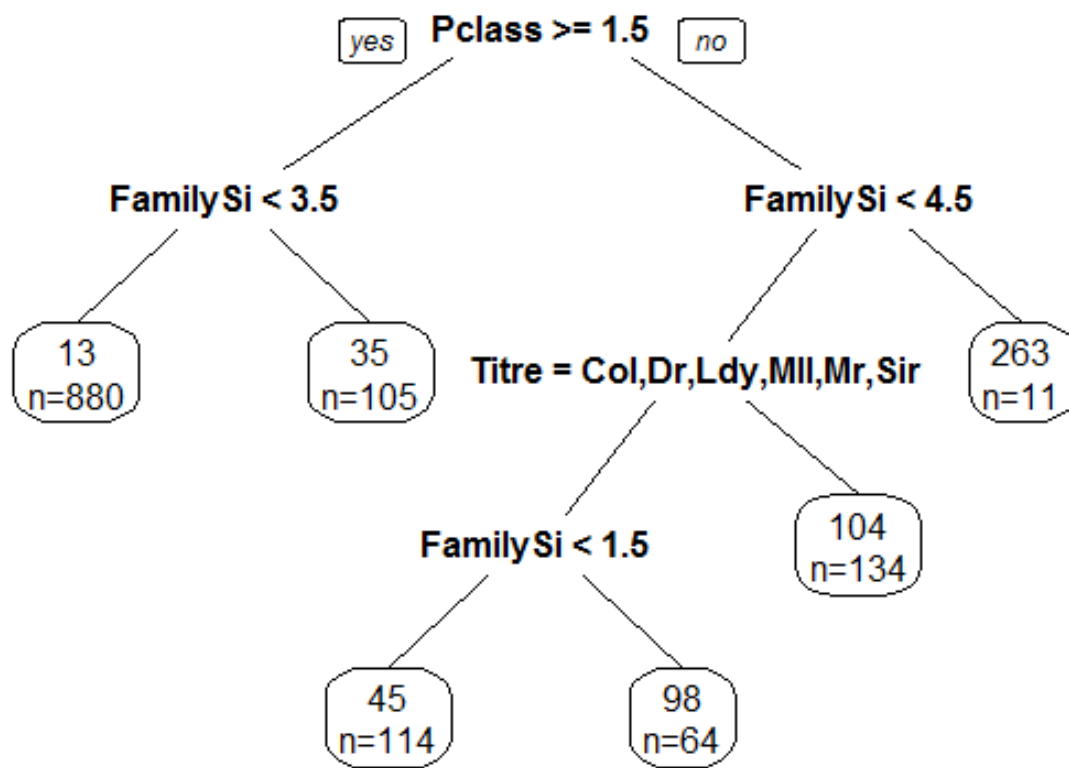
### 3.2.3 Utilisation sur les variables Age, Fare, Cab et Embarked

Finalement, nous utilisons cette méthode pour prédire les valeurs manquantes de ces quatre variables. Pour se faire, le choix des variables explicatives a été difficile à faire. En effet, pour ne pas tomber dans le piège du sur-apprentissage, il ne faut pas choisir un nombre trop élevé de variables, de même il faut éviter les modèles trop simplistes avec trop peu de variables. Il est donc inutile et non-optimale de choisir toutes les variables pour chacune des prédictions. Pour sélectionner les variables, nous avons dans un premier temps choisi celles qui nous semblaient logiques, par exemple pour la variable "Age", il est évident que les variables "Sibsp" et "Parch" sont importantes. Ne pouvant pas uniquement se baser sur une réflexion pareille, nous avons par la suite effectué de nombreux tests en vérifiant par la suite quelle est le meilleur modèle. Nous avons finalement conclu que les variables explicatives à utiliser sont les suivantes : pour "Age", nous avons pris "Pclass" "SibSp" "Parch" "Fare" "Titre" et "FamilySize", pour "Fare" nous avons "Pclass" "Sex" "SibSp" "Parch" "Age" "Embarked" "Titre" et "FamilySize", pour "Embarked" nous avons "Pclass" "Sex" "SibSp" "Parch" "Fare" "Age" "Titre" et "FamilySize" et enfin pour "Cab" nous avons pris "Pclass" "Sex" "SibSp" "Parch" "Fare" "Age" "Titre" et "FamilySize". A noter que nous n'avons pas fait de prédictions sur la variable "Cabin" qui était beaucoup trop compliquée au vu du grand nombre de valeurs manquantes ainsi que du grand nombre de valeurs que la variable peut prendre, nous avons plutôt utilisé la variable "Cab" qui a certe autant de valeurs manquantes que "Cabin" mais qui par contre ne prend pas beaucoup de valeurs (uniquement A, B et C).

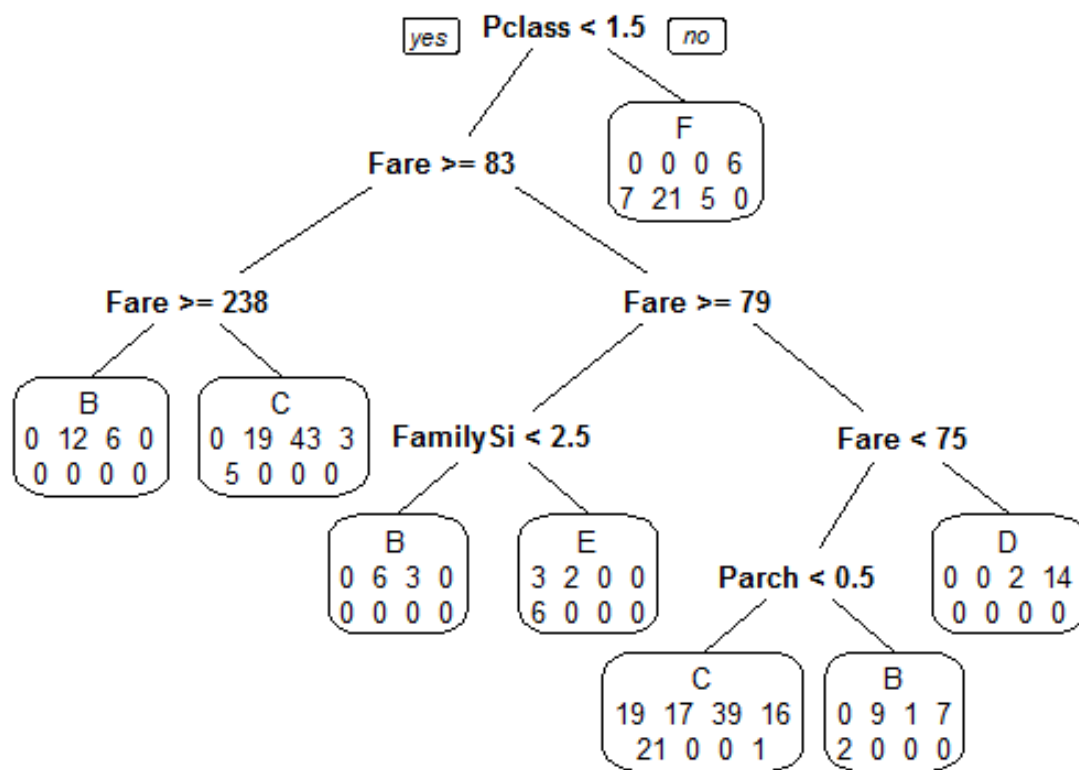
A présent, nous allons vous présenter les résultats obtenus. Nous afficherons uniquement les arbres obtenus à la suite de l'optimisation faite avec le modèle "prune". Tout d'abord voici l'arbre obtenu pour la variable âge :



Nous pouvons voir sur cet arbre que l'algorithme n'a pas vraiment utilisé toutes les variables choisies en paramètre et qu'il n'a choisi que les plus pertinentes en les classant par ordre de pertinence. Les dernières feuilles de l'arbre correspondent aux âges prédits, nous pouvons par exemple voir que les individus ayant la variable "Titre" à yes, la variable Parch supérieure à 0.5 et la variable Fare inférieure à 48 sont prédit d'avoir soit un âge de 7 soit un âge de 19. Le n correspond lui au nombre d'individus susceptibles de prendre cette valeur. Voici à présent l'arbre de décision de la variable "Fare" :

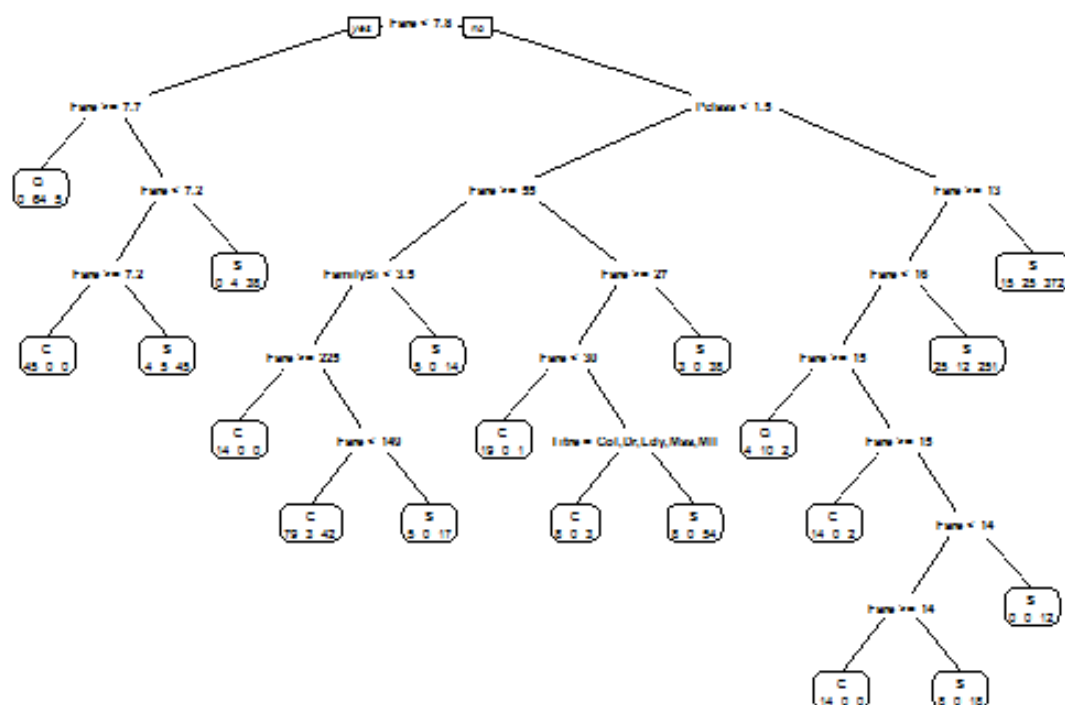


L'avantage que l'on peut observer de ces deux arbres est qu'ils ne sont pas très long, ni trop court non plus, ce qui évite le phénomène de sur-apprentissage.  
Voici l'arbre de la variable "Cab" :



Enfin, voici l'arbre de la variable "Embarked" :





La grandeur de l'arbre semblerait annoncer que nous sommes tombés dans le piège du sur-apprentissage, mais ce modèle a été le plus prometteur et a donné de bons résultats par la suite.

### 3.3 L'éventuel utilisation de Random Forest

Le modèle Random Forest est un outil plus puissant que les arbres de décision, dont nous allons expliquer le principe plus loin. Cependant, son utilisation pour prédire les données manquantes n'a pas été fructueuse et n'a pas donné de meilleurs résultats que les arbres de décision. Nous avons donc choisi de garder la méthode et les modèles précédents.

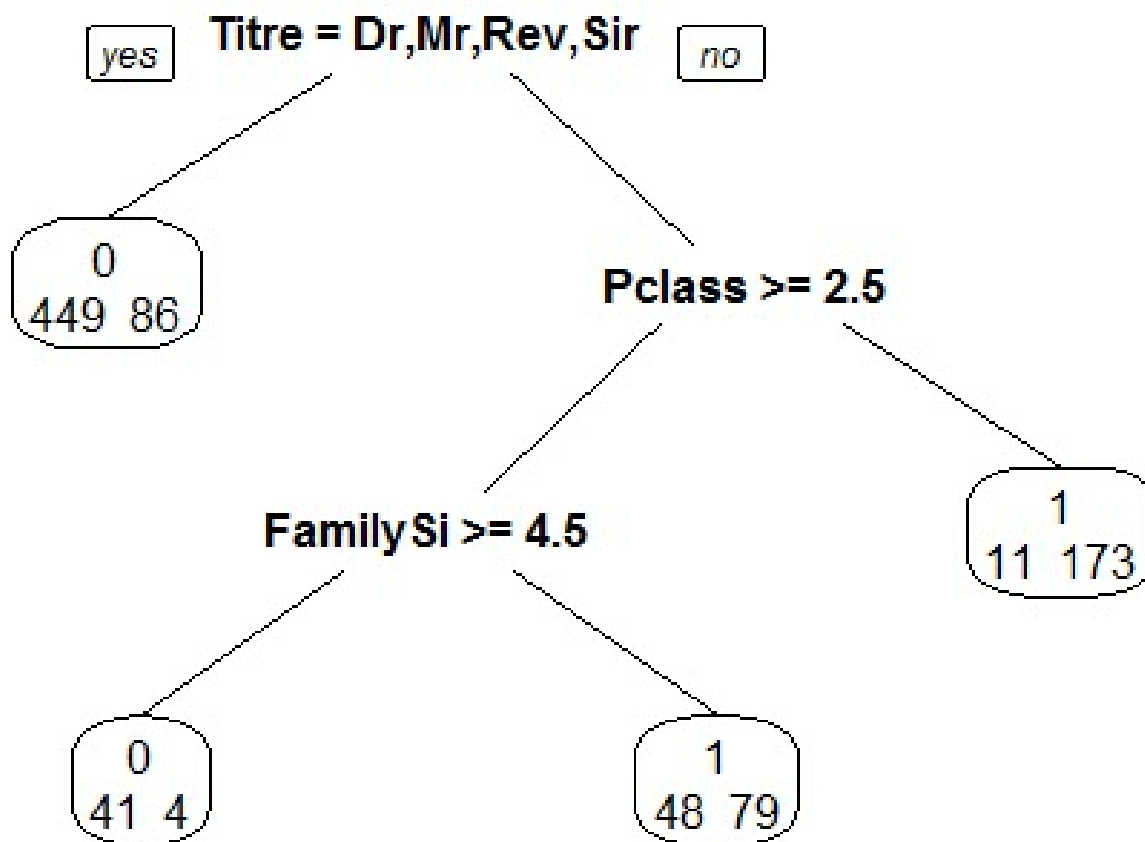
## Chapitre 4

# Utilisation d'algorithmes performants

Après avoir palié au problème des données manquantes et ajouté de nouvelles variables permettant de trier les données, nous avons pu utiliser différentes méthodes de machine learning afin d'obtenir un modèle statistique du problème.

## 4.1 Arbre de décision sur Survival

Nous avons d'abord commencé par faire un arbre de décision sur survival.



Cependant cette méthode comportait un défaut majeur dont nous nous sommes aperçus lors de l'utilisation sur la base de données train. En effet, la variation du nombre d'individus a considérablement fait varier le modèle, cette méthode ne nous a pas permis d'obtenir un modèle stable. La performance d'un arbre de décision dépend très fortement de l'échantillon utilisé. Ainsi lorsqu'on utilisera le modèle obtenu pour prédire la variable survival sur le tableau Test, le résultat ne sera pas du tout optimal.

Nous avons ensuite donc préféré une autre méthode qui donne de meilleurs résultats : la méthode SVM.

## 4.2 Le Support Vector Machine

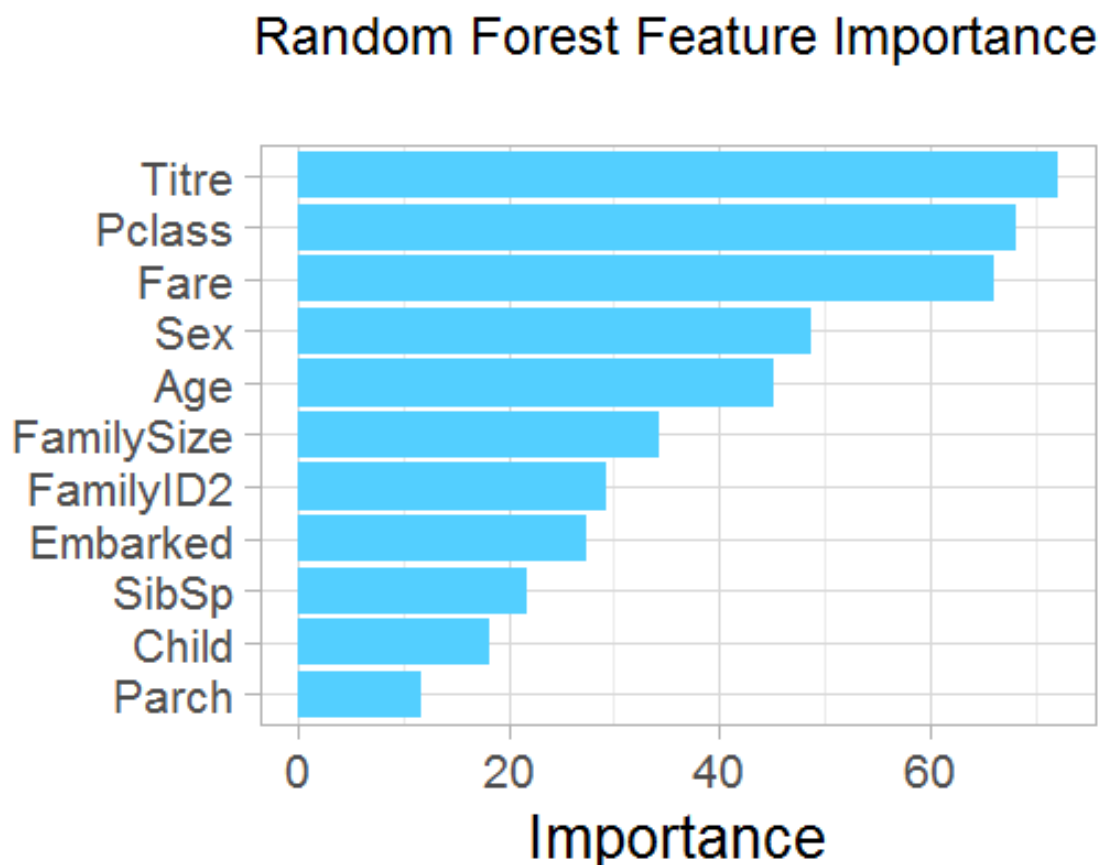
L'avantage de la méthode SVM vue en cours par rapport aux arbres de décision est sa stabilité, cet algorithme s'adapte donc très bien à la modification des variables d'entrée. Cette méthode est d'autant plus efficace sur un problème binaire tel que celui que nous étudions. L'algorithme va alors synthétiser une fonction qui va séparer les individus en deux groupes : les survivants et ceux

qui n'ont pas survécus, tout en essayant de maximiser la distance entre l'hyperplan engendré par la fonction et les individus situés des deux côtés de cet hyperplan. Le SVM va également chercher à minimiser l'erreur d'apprentissage dans l'objectif d'avoir un modèle qui pourra être utilisé avec d'autres données ayant quelques différences avec les données de base, dans l'objectif d'avoir un modèle le plus stable possible.

### 4.3 Le Random Forest

Pour obtenir un meilleur modèle que celui obtenu précédemment, nous avons un outil très puissant qu'est le random forest. Cet algorithme base sa logique sur les arbres de décision dans l'objectif de lutter contre son principal défaut qui est son manque de stabilité et sa trop forte dépendance en l'échantillon utilisé pour obtenir le modèle.

Pour pallier ce défaut, le random forest utilise une multitude d'arbres (d'où le "Forest"). Ces arbres sont effectués sur quelques observations tirés aléatoirement parmi l'échantillon utilisé, puis ces arbres sont assemblés pour obtenir un résultat meilleur. Nous avons ainsi utilisé la fonction `randomForest` avec comme paramètres notre tableau `Train` dont on a remplacé les variables manquantes grâce aux arbres de décision, nous avons mis le paramètre "importance" à `TRUE` et "ntree" à 2000 :



Ce graphe nous montre l'importance de chaque variable pour le modèle. On voit l'utilité d'avoir créer les nouvelles variables qui ont une grande importance.

Cette méthode nous a donné un résultat concluant avec un assez bon résultat sur Kaggle. Cependant nous avons découvert qu'il était possible d'améliorer le résultat en utilisant la méthode du cforest.

#### 4.4 Le CForest, une amélioration du Random Forest

Le CForest est un autre type d'implémentation du Random Forest qui va privilégier une approche par l'inférence conditionnelle. L'inférence conditionnelle va permettre à l'algorithme de s'affranchir encore plus du problème de "surapprentissage" qui est le défaut principal des arbres de décision que le Random Forest cherche à limiter.

```
> varimp(fit)
      Pclass      Sex      Child      SibSp      Fare      Embarked      Titre
0.045828746 0.072920489 0.004848624 0.005544343 0.014339450 0.004892966 0.143747706
FamilySize  FamilyID2
0.011559633 0.012013761
```

Ce tableau nous montre à nouveau l'importance des nouvelles variables, notamment la variable "Titre". Ce modèle nous a donné notre meilleur résultat, avec un score de 0.8134 sur Kaggle.

# Conclusion

Nous avons beaucoup apprécié ce projet. En effet, il nous a beaucoup appris sur l'apprentissage automatique en nous permettant d'appliquer des méthodes de machine learning sur un problème concret.

Cela nous a notamment appris à effectuer les traitements nécessaires à effectuer sur les données afin de pouvoir obtenir un modèle statistique et à valider la cohérence du choix d'un modèle.

Nous avons ainsi acquis une nouvelle démarche qui nous permet d'exploiter au mieux les données avec un objectif simple : obtenir un modèle avec la plus petite erreur. Cette mise en situation nous a permis d'appréhender les enjeux du métier de "Data scientist" qui sont la gestion et l'analyse des données à l'aide d'algorithmes qui vont véritablement "faire parler" les données pour aboutir à la réponse à un problème bien précis. Le problème étudié ici a été la survie des passagers du Titanic. Cette méthodologie peut s'exporter en entreprise pour par exemple chercher à prédire l'évolution du prix d'une action, d'une maison, d'une voiture ... La démarche acquise au cours de ce projet peut ainsi facilement s'appliquer à de nombreux domaines dans le monde de l'entreprise ou de la recherche.