PROJET: Classification naïve bayésienne

Mélissa EVEILLARD - Kenzo RAMDANI

6 novembre 2019

Contents

1. Introduction

Le naufrage du RMS Titanic est sûrement l'un des désastres le plus tristement célèbre de l'histoire. Le 15 avril 1912, durant son voyage inaugural reliant Southampton (Royaume-Uni) à New-York, le paquebot coula après avoir percuté un iceberg, ce qui entraîna la mort de 1502 personnes sur les 2224 passagers et membres d'équipages qui faisaient partie du voyage.

Grâce à une base de données contenant les informations des passagers du bateau, nous essayerons de construire un modèle qui permet d'identifier les personnes qui ont le plus de chance de survie.

Pour construire ce modèle nous utiliserons la base de données nommé *train* qui contient un échantillon des passagers du Titanic. Ci-dessous se trouve un aperçu de la base de données

	PassengerId	Survived	Pclass	Name	Sex	Age
707	707	1	2	Kelly, Mrs. Florence "Fannie"	female	45
706	706	0	2	Morley, Mr. Henry Samuel ("Mr Henry Marshall")	male	39
566	566	0	3	Davies, Mr. Alfred J	male	24
244	244	0	3	Maenpaa, Mr. Matti Alexanteri	male	22
825	825	0	3	Panula, Master. Urho Abraham	male	2
754	754	0	3	Jonkoff, Mr. Lalio	male	23

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
707	0	0	223596	13.5000	NA	S
706	0	0	250655	26.0000	NA	S
566	2	0	A/4 48871	24.1500	NA	S
244	0	0	STON/O 2. 3101275	7.1250	NA	S
825	4	1	3101295	39.6875	NA	S
754	0	0	349204	7.8958	NA	S

2. Description des données

Cette base de données contient 594 individus et 12 variables (PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked) contenant de nombreuses informations concernant les passagers.

Les variables Survived, Pclass, Name, Sex, SibSp, Parch, Cabin et Embarked sont des variables qualitatives.

- ***Survived*** est une variable binaire désignant la survie du passager ou non (0 = Mort et 1 = Survie)
- ***Pclass*** est une variable catégorielle variant de 1 à 3 désignant le numéro de classe de voyage du passager dans le naivre
- ***Name*** contient les noms et titre des passagers
- ***Sex*** est une variable binaire codée ainsi: 1 = Femme et 2 = Homme
- ***Ticket*** correpsond au numéro du ticket
- ***Cabin*** est le numéro de cabine du passager
- ***Embarked*** correspond au port d'embarcation: C = Cherbourg (France), Q = Queenstown (Irlande) et S = Southampton (Angleterre)

Les variables Age, Fare, SibSP, Parch sont quant à elles quantitatives:

- ***PassengerId***, variable de type quantitative discrète mais on la considérera comme qualitative ordinale sur laquelle on ne peut faire aucun calcul car elle permet d'identifier chaque passager de manière unique.
- ***Age*** correspond à l'âge des passagers en années
- ***Parch***, une variable discrète correspondant au nombre de parents et/ou enfants présents à bord
- ***SibSP*** correspond au nombre d'époux, de frères et /ou soeurs présents à bord
- ***Fare*** est une variable continue correspondant au prix du ticket (en livre sterling).

Dans ce jeu de données, nous n'avons pas à notre disposition l'ensemble de ces informations pour tous les passagers; 585 observations sont incomplète. Le nombre d'observations manquantes selon les variables est le suivant:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	0	0	0	0	121	0	0	0	0	463	1

On constate que c'est pour la variable *Cabin* que nous avons le plus d'informations manquantes puisqu'ils manquent 463 valeurs, ce qui représente 80% des valeurs manquantes.

Afin de construire notre modèle, nous allons devoir nous intéresser plus particulièrement aux variables *Sex*, *Pclass* et *Age* et étudier selon ces variables la chance de survie des passagers.

Afin de pouvoir décrire et interpréter les variables qualitatives de façon plus efficace et claire, et notamment et *Pclass* et *Survived* nous la transformons en format factor:

```
train$Survived = factor(train$Survived, labels = c("Mort", "Survecu"),
    levels = c(0, 1))

train$Pclass = factor(train$Pclass, levels = c("1", "2", "3"))
```

```
ggplot(data = train, aes(fill = Survived)) + geom_bar(mapping = aes(x = Survived,
    y = (..count..)/sum(..count..))) + scale_y_continuous(labels = percent,
    limits = c(0, 0.7)) + ylab("") + xlab("Survie") + ggtitle("Proportion des morts et des survivants")
    theme(legend.position = "none", plot.title = element_text(face = "bold",
        hjust = 0.5, size = 16)) + scale_fill_manual(values = c("#717171",
    "#FFAA00")) + theme_bw()
```

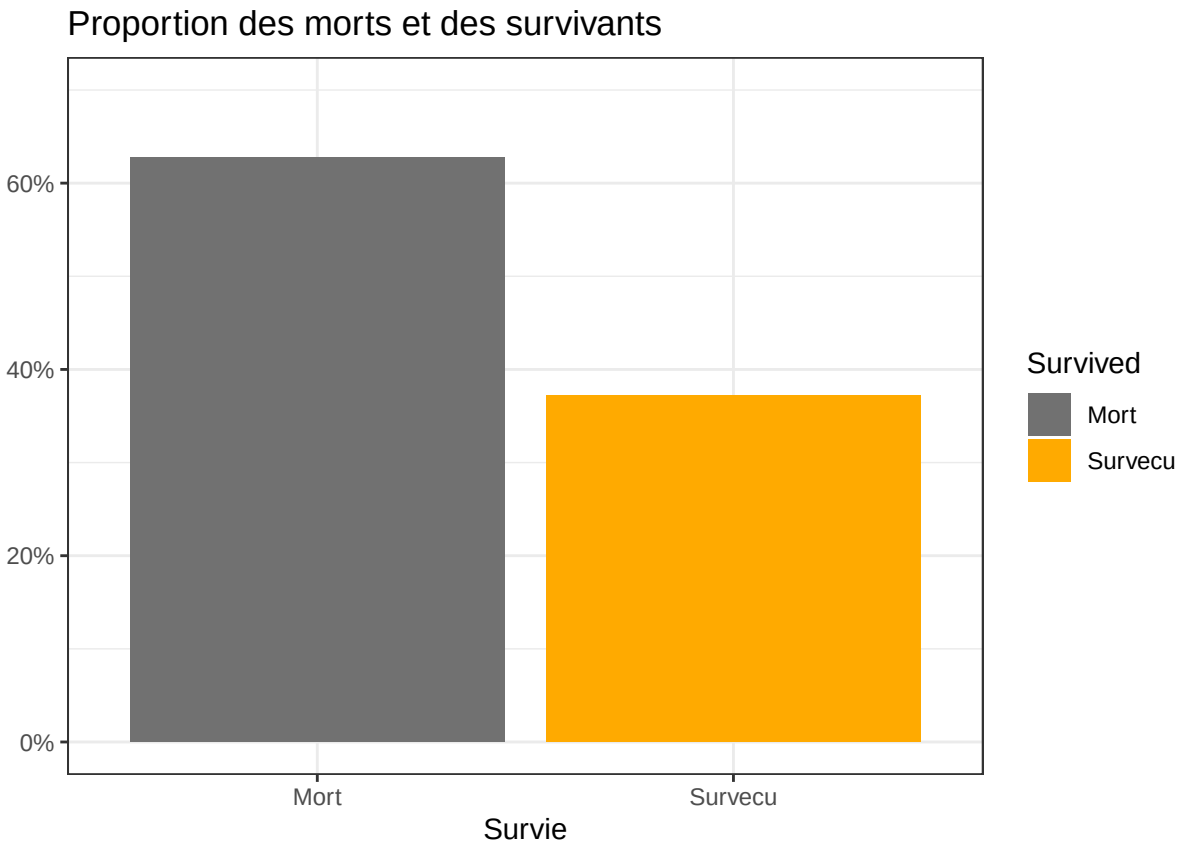

Figure 1: *Proportion des morts et des survivants*

Ce graphique nous montre que près des 2/3 des passagers (373) ont périt lors du naufrage du navire.

```
ggplot(data = train, aes(x = "", fill = Sex)) + geom_bar(aes(y = (..count..)/sum(..count..))) +
    ylab("") + scale_y_continuous(labels = percent) + xlab("") +
    scale_fill_manual(values = c("#ffef00", "#0080FF")) + ggtitle("Diagramme empilé de la proportion d'
    theme(legend.position = "none", plot.title = element_text(face = "bold",
        hjust = 0.5, size = 16)) + theme_bw()
```

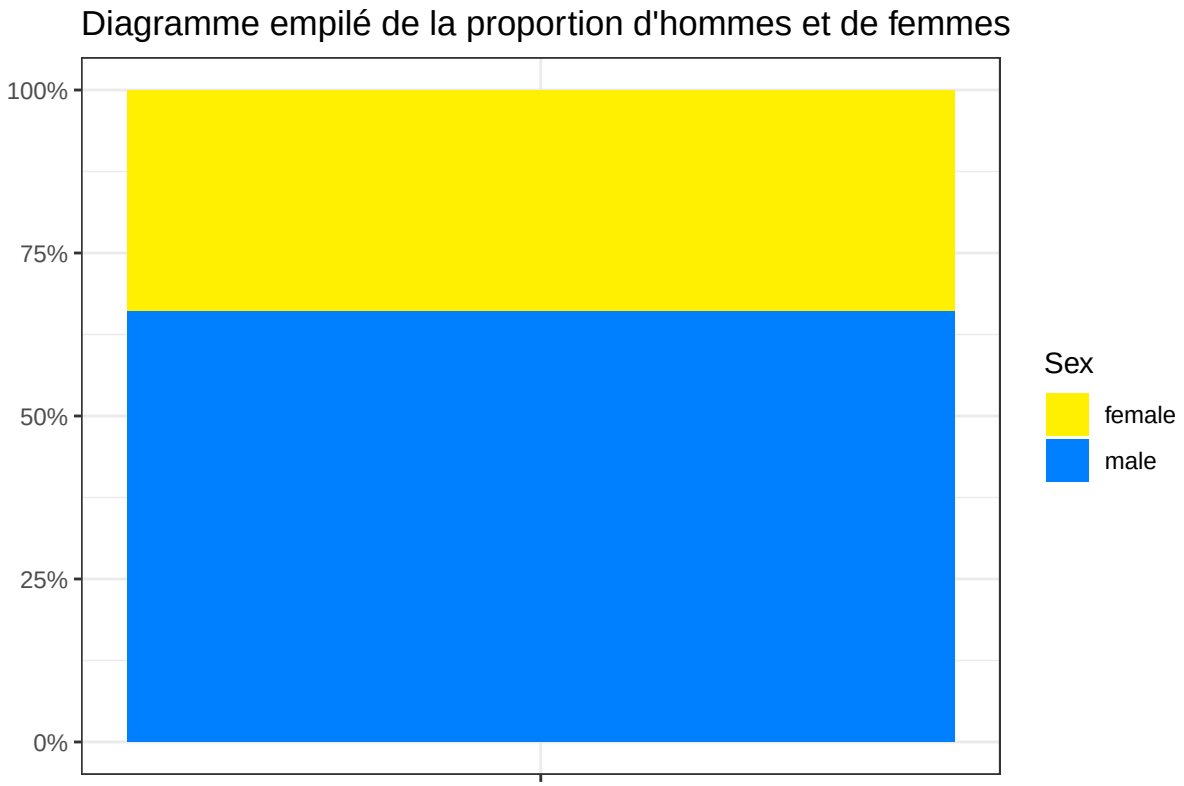

Figure 2: *Diagramme empilé de la variable Sex*

Sur les 594 passagers, la grande majorité sont des hommes puisqu'ils représentent 66% de l'ensemble des passagers (393 hommes) et les femmes ne représentent que 34% des passagers, soit 201 femmes.

```
ggplot(data = train, aes(fill = Pclass)) + geom_bar(mapping = aes(x = as.factor(Pclass))) +
    ylab("") + xlab("Classe de voyage") + scale_x_discrete(limits = c("3",
    "2", "1")) + coord_flip() + theme_bw() + ggtitle("Nombre de passagers en fonction de la classe de v
    theme(legend.position = "none", plot.title = element_text(face = "bold",
        hjust = 0.5, size = 16))
```

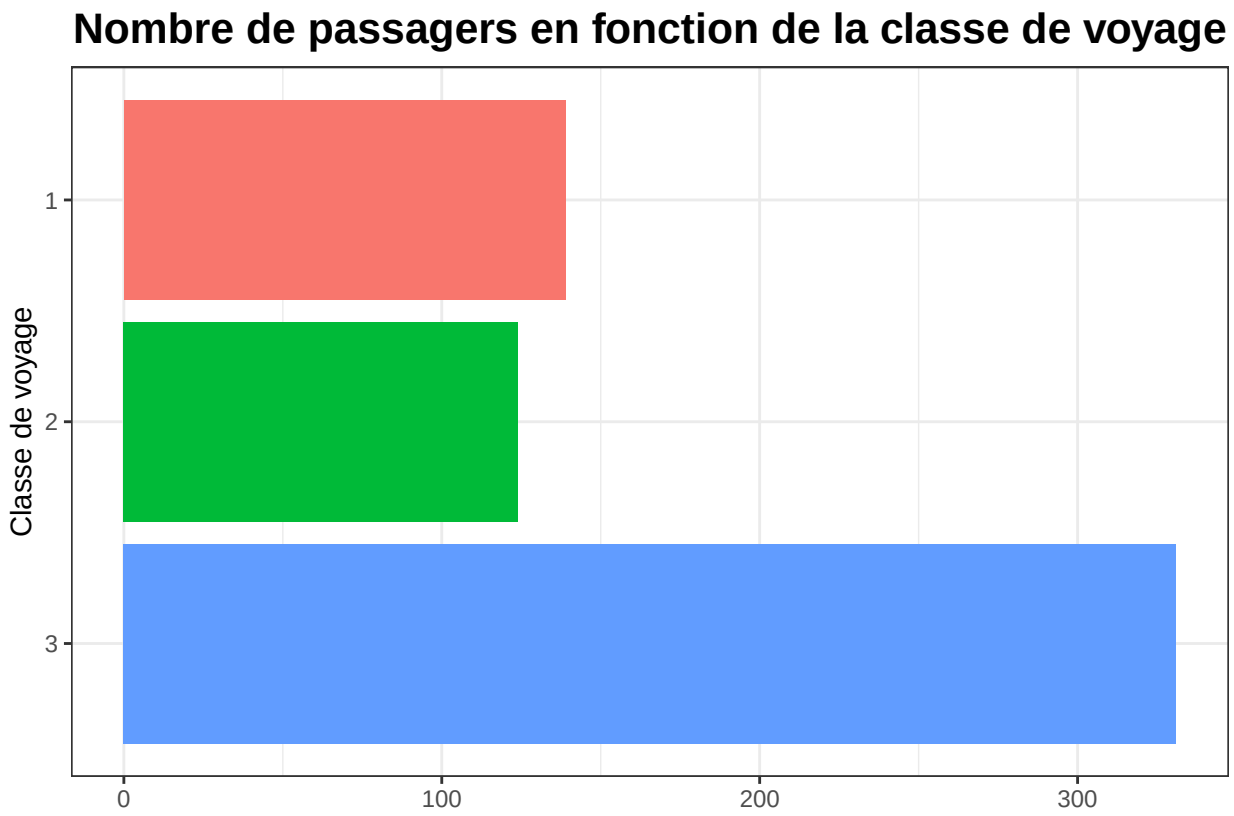

Figure 3: *Répartition du nombre de passagers en fonction de la classe de voyage*

Le bateau était divisé de la première à la troisième classe; la première classe acueillant les passagers les plus fortunés et la dernière était occupé par des passagers venant de l'immagration. Sur les 594 passagers, 139 voyageaient en première classe et 124 dans la deuxième. La majorité des passagers étaient installés en troisième classe puisqu'ils y étaient 331, soit 56% des passagers.

Concernant l'âge des passagers, on étudie sa distribution grâce au graphhique suivant:

```
ggplot(data = train, aes(x = "", y = Age)) + geom_boxplot(outlier.colour = "red",
    outlier.shape = 8, outlier.size = 4, fill = "#0083ff") +
    stat_summary(fun.y = mean, geom = "point", na.rm = T, colour = "red") +
    theme_bw() + ggtitle("Distribution de l'âge des passagers") +
    theme(legend.position = "none", plot.title = element_text(face = "bold",
        hjust = 0.5, size = 16))
```

```
## Warning: Removed 121 rows containing non-finite values (stat_boxplot).
```

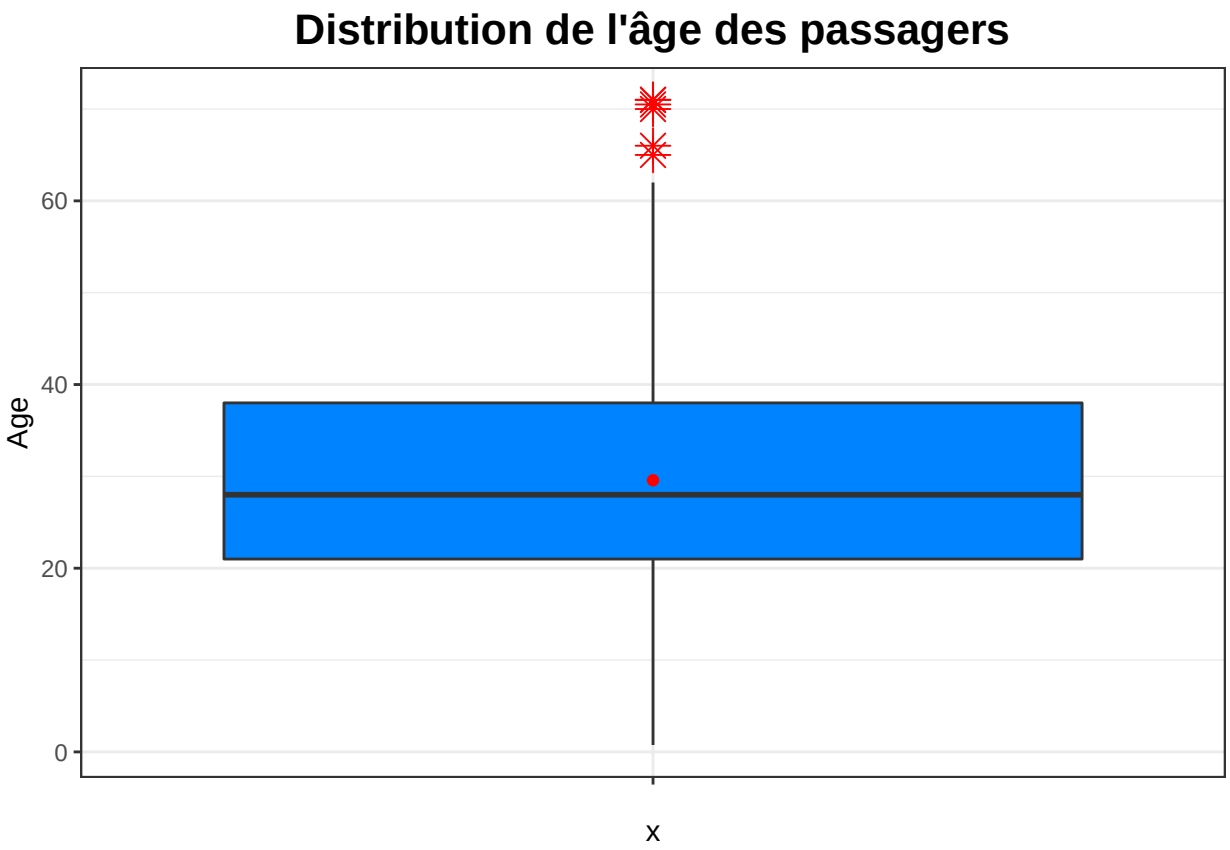

Figure 4: *Distribution de l'âge des passagers*

Age par niveau de quantiles:

0%	25%	50%	75%	100%
0.75	21	28	38	71

A bord du navire les passagers étaient plutôt jeunes puisque la moyenne d'âge est de 29 ans et la moitié d'entre eux avaient moins de 28 ans. Les 3/4 des passagers avaient moins de 38 ans. Egalement, 25% des passagers avaient entre 38 ans et 71 ans.

Afin de simplifier l'étude de cette variable nous allons la discrétiser en 4 classe d'âges. On considère les catégories d'âges par tranches de 20 ans, allant de 0 à 80 ans : (0, 20], (20, 40], (40, 60] et (60, 80] ans. La nouvelle variables créée est appelée *cA*.

```
train$cA = cut(Age, 4, c("(0;20]", "(20;40]", "(40;60]", "(60;80]"),
    right = T)
```

```
summary(train$cA)
```

```
##  (0;20] (20;40] (40;60] (60;80]    NA's
##      92     238     113      30     121
```

```
ggplot(data = train, aes(fill = cA)) + geom_bar(mapping = aes(x = cA,
    y = (..count..)/sum(..count..))) + scale_y_continuous(labels = percent) +
    ylab("") + xlab("Classe d'âge") + ggtitle("Répartition des passagers en fonction de la classe d'âge
    theme(legend.position = "none", plot.title = element_text(face = "bold",
        hjust = 0.5, size = 16)) + # scale_x_discrete(limits = c('(0;20]',
```

```
# '(20;40]','(40;60]','(60;80]'))+
theme_bw()
```

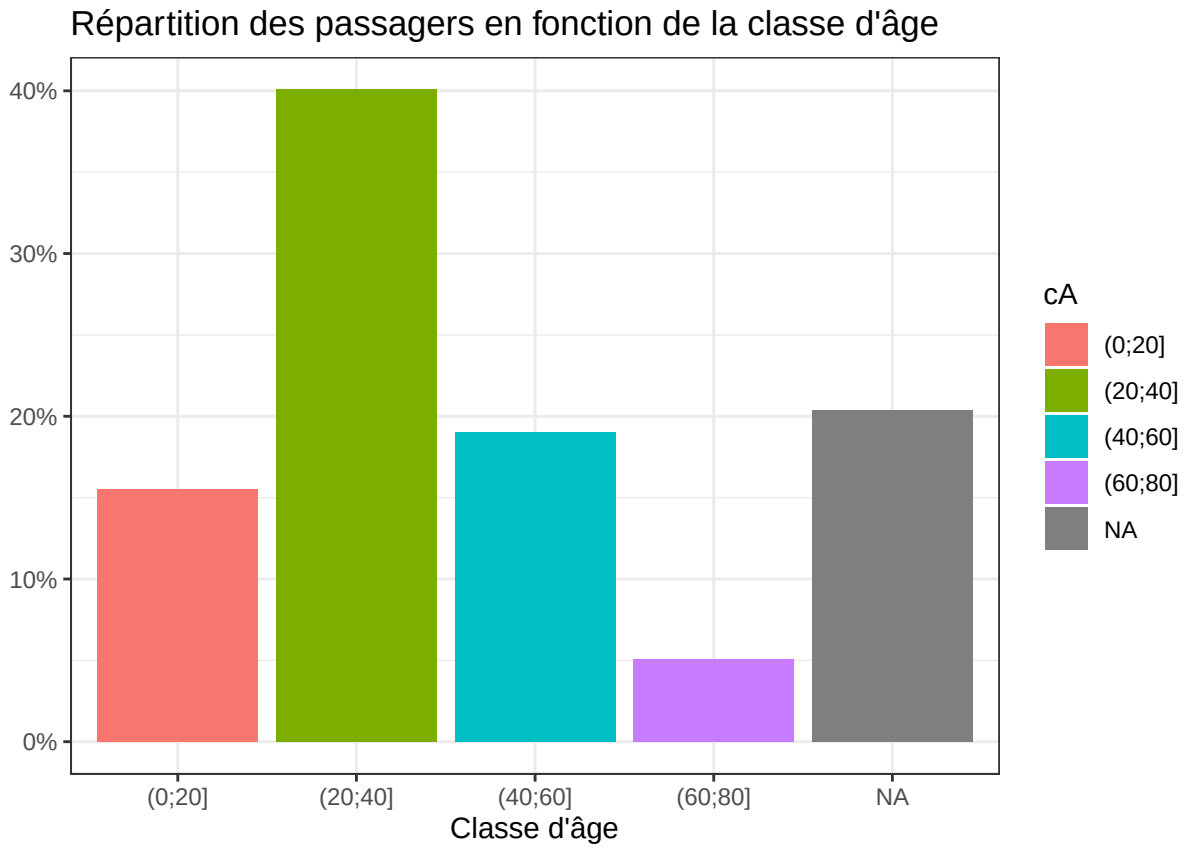

Figure 5: *Réparition des passagers en fonction de la classe d'âge*

En discrétisant la variable *Age* cela nous permet de simplifier l'information.

Avec le graphique ci-dessus, on constate qu'une grande partie des passagers (40%) ont entre 20 ans et 40 ans, et comme nous l'avions vu dans la figure que peu de personnes ont plus de 60 ans; seulement 6% des passagers ont entre 60 et 80 ans.

3. Lien entre les variables

Le but de notre étude est de prédire la chance de survie des passagers en fonciton des différents facteurs (l'âge, le sexe et la classe de voyage) que nous venons d'étudier. A présent, nous allons donc étudier le lien de la variable *Survived* en fonction de ces différentes variables.

```
ggplot(data = train, aes(x = factor(Sex), fill = Survived)) +
    geom_bar(aes(y = (..count..)/sum(..count..)), position = "fill") +
    scale_y_continuous(labels = percent) + ylab("") + xlab("Sexe de l'individu") +
    ggtitle("Lien entre la variable Survived et Sex") + scale_fill_manual(values = c("#717171",
    "#FFAA00")) + theme_bw()
```

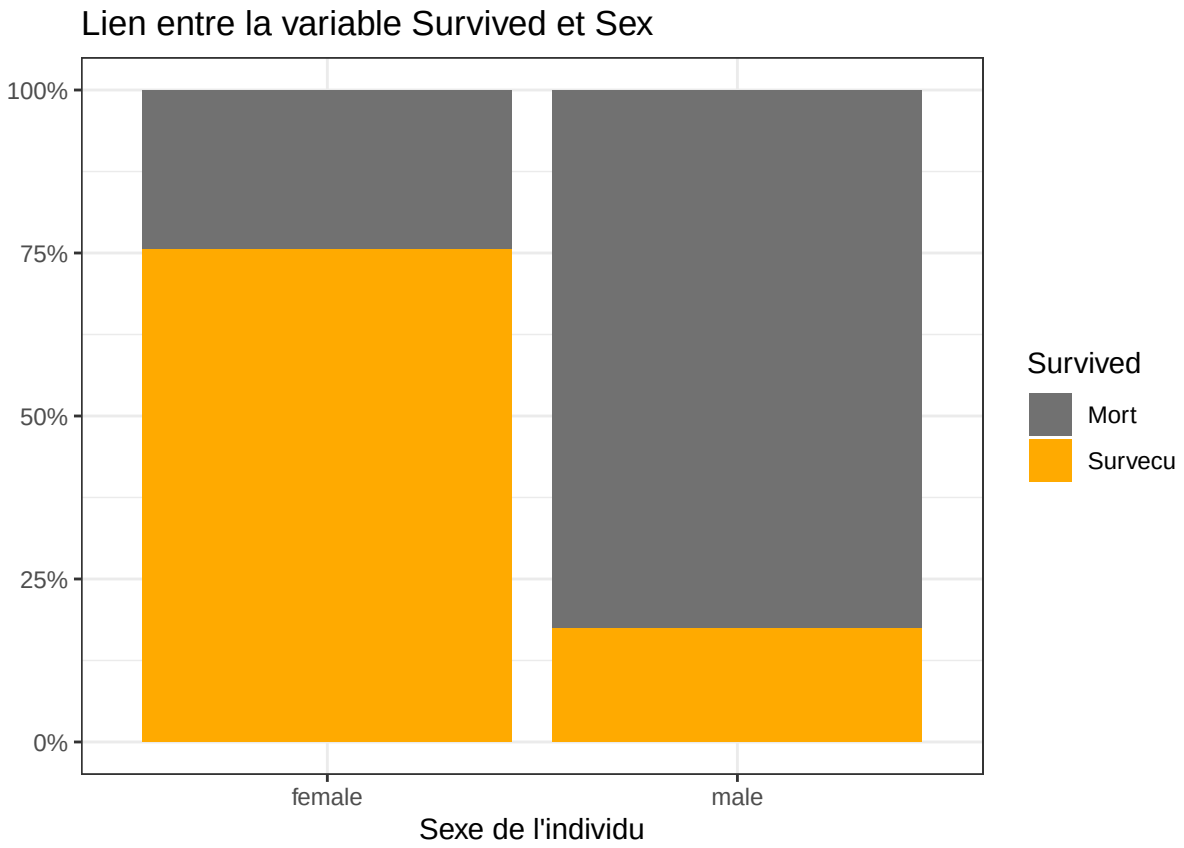

Figure 6: *Lien entre la variable Survived et Sex*

Si on s'intéresse à la proportion de passagers qui ont survécu en fonction de leur sexe, nous constatons que c'est parmi les femmes qu'on trouve le plus large taux de survie; en effet, sur 100 femmes les $\frac{3}{4}$ ont survécu au naufrage. Pour les hommes, le constat s'inverse puisque la plupart d'entre eux n'ont pas survécu; plus de 80% d'entre eux sont morts pendant le naufrage. Ainsi, il semblerait que le sexe du passager aient une influence sur sa chance de survie.

```
t_c = table(train$Survived, train$Pclass)
res = addmargins(prop.table(t_c, 2))
knitr::kable(round(res[, -4], 2))
```

	1	2	3
Mort	0.35	0.55	0.78
Survecu	0.65	0.45	0.22
Sum	1.00	1.00	1.00

Table 1: *Table de contingence de la survie en fonction de la classe de voyage*

Cette table de contingence nous donne la proportion de passagers morts ou vivants en fonction de la classe de voyage. Cette information nous est résumée de manière plus explicite dans le graphique suivant:

```
mosaicplot(Pclass ~ Survived, data = train, main = "Lien entre la variable Survived et Pclass",
    xlab = "Classe de voyage", color = c("#717171", "#FFAA00"))
```

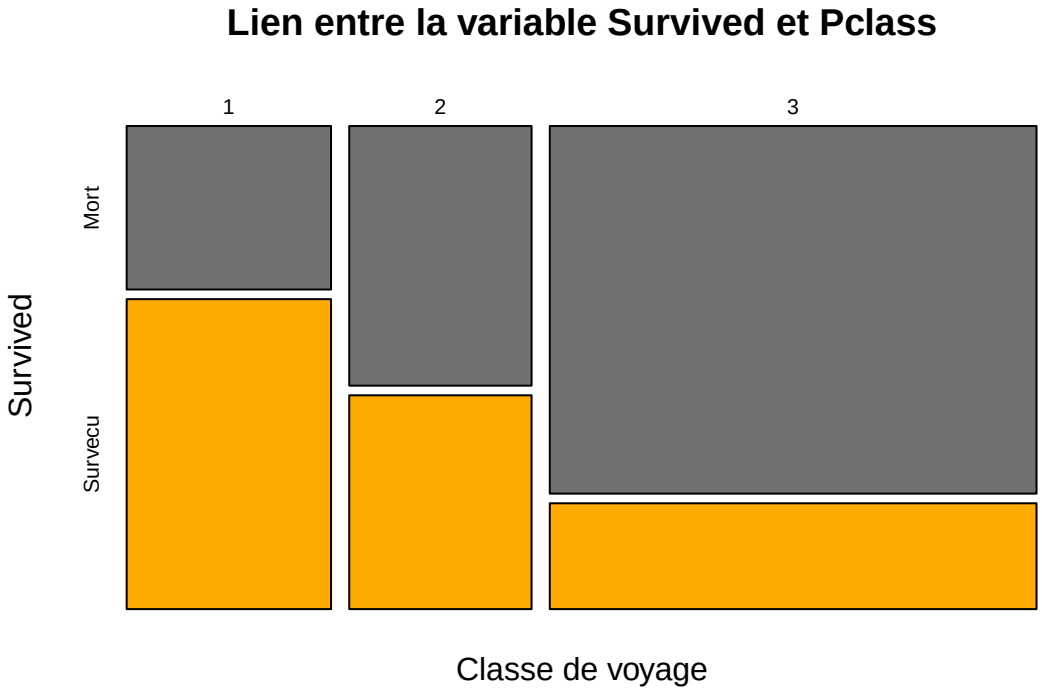

Figure 7: *Lien entre la variable Survived et Pclass*

Le taux de survivants varie également en fonction de la classe dans laquelle a voyagé le passager. En effet, plus la classe de voyage diminue et plus le taux de décès augmente. Dans la première classe, $\frac{1}{3}$ des passagers n'ont pas survécu (soit 48 sur les 139 passagers de la première classe); ce taux augmente de plus 20 points de pourcentage pour les passagers de la deuxième classe puisqu'ils sont 55% à avoir périt. Le taux de survie le plus faible est pour les passagers de la dernière classe puisque dans cette classe seulement 22% ont survécu. Ces chiffres nous laissent supposer qu'il existe un lien entre la classe de voyage et la probalité de survie d'un passager.

Le dernier facteur auquel nous nous intéressons est l'âge.

```
qplot(factor(Survived), Age, data = train, geom = "boxplot",
    xlab = "Survivant ou non", main = "Boites à moustaches de la variable Age en fonction de la variable
    theme_bw()
```

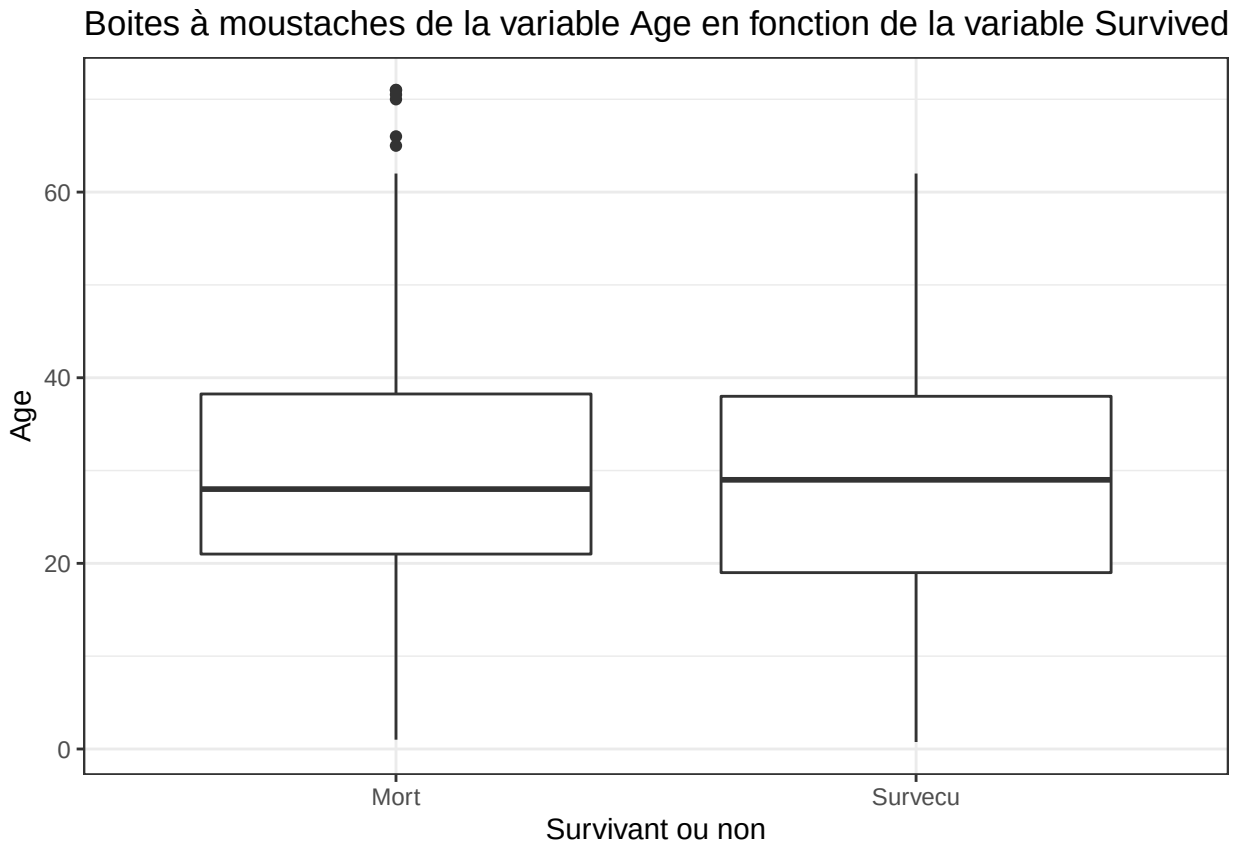

Figure 8: *Boites à moustaches de la variable Age en fonction de la variable Survived*

On peut voir que l'âge médian est assez proche chez les survivants ou non, de même que la répartition. Il ne semble pas y avoir de lien entre l'âge et le fait de survivre.

```
ggplot(data = train, aes(x = cA, fill = Survived)) + geom_bar(aes(y = (..count..)/sum(..count..)),
    position = "fill") + scale_y_continuous(labels = percent) +
    scale_fill_manual(values = c("#717171", "#FFAA00")) + ylab("") +
    ggtitle("Lien entre la variable Survived et la classe d'âge") +
    xlab("Classe d'âge") + theme_bw()
```

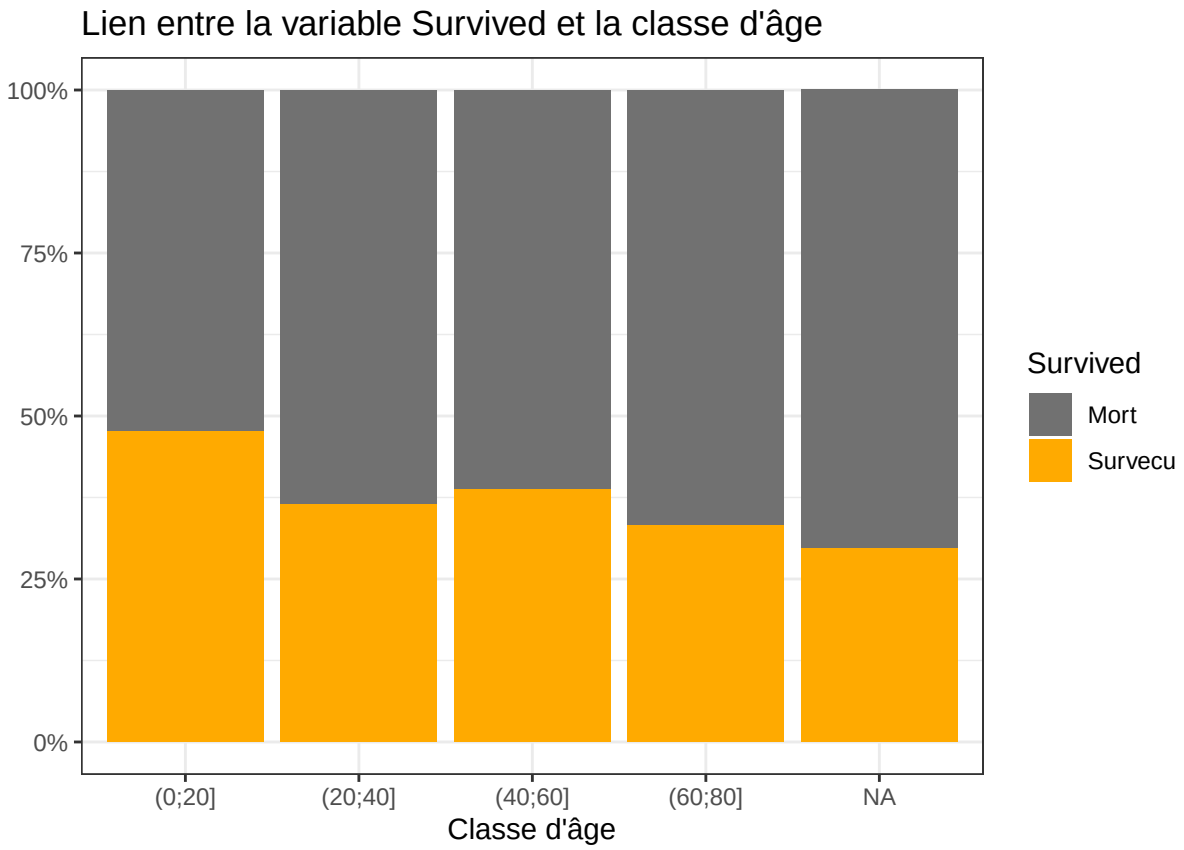

Figure 9: *Lien entre la variable Survived et la classe d'âge*

Nous avons précedemment décider de découper cette variable en plusieurs classes d'âge, nous allons donc regarder le lien entre les différentes classes et le taux de survie. Sur la figure 9, nous remarquons que la proportion de passagers mort est assez similaire entre les différentes classes d'âges. 52% des passagers ayant entre 0 et 20 ans sont morts, cette proportion s'élève 63% pour les passagers ayant entre 20 et 40 ans. La classe d'âge pour lequel le taux de survivants est le plus faible est pour ceux ayant entre 60 et 80 ans puisqu'ils ne sont qu'$\frac{1}{3}$ à avoir survécu. Concernant les passagers pour lequel nous n'avons aucune information concernant leur âge, le taux de décès est proche de 75%. Contrairement aux autres variables, le lien entre la chance de survie et l'âge semble moins significatif, bien que l'on remarque une faible augmentation du taux de décès lorsque l'âge augmente.

Ainsi ces analyses nous ont permis de mettre en avant plusieurs relations. Comme nous pouvions le supposer plus de femmes que d'hommes ont été sauvées du naufrage (75% contre 20%). Lors du naufrage c'est probablement le protocole *les femmes et les enfants d'abord* qui a été mis en place. Le sexe du passager semble donc avoir un impact déterminant sur sa chance de survie. Celle-ci semble également être influencée par la classe dans laquelle voyageait le passager. En effet nous avons vu que plus la classe de voyage était élevée et plus le taux de survie augmentait (70% des voyageurs de première classe ont survécu contre 20% pour ceux de la troisième classe). Concernant l'âge, il ne semble pas y avoir un impact très significatif de celui-ci sur la chance de survie du passager. Les taux de survie selon les différentes classes sont assez similaires.

Nous allons à présent estimer la probabilité de survie du passager en fonction de ces différents critères.

4. Prédiction de la survie

Dans le but de contruire notre modèle de prédiction dont nous avons besoin d'estimer la probabilité de survie conditionnelement à d'autres variables. Cette mesure nous l'avons grâce à la formule suivante:

$P_B(A) = \frac{P(A \cap B)}{P(B)}$ que l'on applique sur nos données. Nous pouvons donc estimer la probabilité de survie selon que le passager soit une femme:

$P_{Sx=female}(S = 1)$ est:

```
nrow(subset(train, subset = Survived == "Survecu" & Sex == "female"))/nrow(subset(train,
    subset = Sex == "female"))
```

```
## [1] 0.7562189
```

La probabilite de survie sachant que le passager soit un homme ($P_{Sx=male}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & Sex == "male"))/nrow(subset(train,
    subset = Sex == "male"))
```

```
## [1] 0.1755725
```

La probabilite de survie sachant que le passager voyage en première classe ($P_{P=1}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & Pclass ==
    1))/nrow(subset(train, subset = Pclass == 1))
```

```
## [1] 0.6546763
```

La probabilite de survie sachant que le passager voyage en deuxième classe ($P_{P=2}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & Pclass ==
    2))/nrow(subset(train, subset = Pclass == 2))
```

```
## [1] 0.4516129
```

La probabilite de survie sachant que le passager voyage en troisième classe ($P_{P=3}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & Pclass ==
    3))/nrow(subset(train, subset = Pclass == 3))
```

```
## [1] 0.223565
```

La probabilite de survie sachant que le passager a entre 0 et 20 ans ($P_{cA=(0;20]}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & cA == "(0;20]"))/nrow(subset(train,
    subset = cA == "(0;20]"))
```

```
## [1] 0.4782609
```

La probabilite de survie sachant que le passager a entre 20 et 40 ans ($P_{cA=(20;40]}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & cA == "(20;40]"))/nrow(subset(train,
    subset = cA == "(20;40]"))
```

```
## [1] 0.3655462
```

La probabilite de survie sachant que le passager a entre 40 et 60 ans ($P_{cA=(40;60]}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & cA == "(40;60]"))/nrow(subset(train,
    subset = cA == "(40;60]"))
```

```
## [1] 0.3893805
```

La probabilite de survie sachant que le passager a entre 60 et 80 ans ($P_{cA=(60;80]}(S = 1)$) est:

```
nrow(subset(train, subset = Survived == "Survecu" & cA == "(60;80]"))/nrow(subset(train,
    subset = cA == "(60;80]"))
```

```
## [1] 0.3333333
```

Afin de construire notre modèle de prédiction nous mettons en place le classificateur naïf de Bayes par lequel dépend notre modèle. Pour coder la fonction qui implémente cet outil, nous construisons les tables de probabilité conditionnelle suivantes:

S_P contient les probabilités $P_P(S)$

```
S_P <- prop.table(table(train$Pclass, train$Survived), margin = 2)

colnames(S_P) <- c("0", "1")
```

S_Sx contient les probabilités $P_Sx(S)$

```
S_Sx = prop.table(table(Sex, Survived), margin = 2)

colnames(S_Sx) <- c("0", "1")
```

S_Ca contient les probabilités $P_cA(S)$

```
S_Ca = prop.table(table(train$cA, Survived), margin = 2)

rownames(S_Ca) <- c("(0,20]", "(20,40]", "(40,60]", "(60,80]")
colnames(S_Ca) <- c("0", "1")
```

La table S contient les probailités suivantes: $P(S = 0)$ (= probabilté de ne pas survivre) et $P(S = 1)$ (=probabilité de survivre)

```
S <- prop.table(table(train$Survived))
names(S) <- c("0", "1")
```

Nous codons à présent la fonction **prob_prediction(Sex,Pclass,cAge)** qui implémente le classificateur naïf de Bayes et rend en sortie la probabilité $P_{Sx,P,cA}(S = 1)$ correspondante aux valeurs que l'on donne en entrée. Nous nous aiderons pour cela des tables que nous venons de construire.

```
prob_prediction = function(Sex, Pclass, cAge) {
    p = (S_Sx[Sex, "1"] * S_P[Pclass, "1"] * S_Ca[cAge, "1"] *
        S["1"])/((S_Sx[Sex, "1"] * S_P[Pclass, "1"] * S_Ca[cAge,
        "1"] * S["1"]) + (S_Sx[Sex, "0"] * S_P[Pclass, "0"] *
        S_Ca[cAge, "0"] * S["0"]))

    return(p)
}
```

5. Evaluation de la performance du classificateur

Le data frame test contient un deuxième échantillon de passagers qui permet d'évaluer la qualité du modèle de prédiction. La table test contient uniquement les variables *Survived*, *Sex*, *Pclass*, *cAge* et n'a pas de valeurs manquantes.

Nous chargeons cette base de données dans notre environnement:

```
load("/users/master/ih04752/Documents/PROJET/titanic_test.Rdata")
```

Notre fonction **prob_prediction** va nous permettre de prédire la probabilté de survie de chaque passager de *test*. Notre fonction prend en argument des vecteurs de types charactères. On convertit donc les arguments en type character grâce à la fonction **as.character()**.

```
test$Proba_Survie = prob_prediction(c(as.character(test$Sex)),
    c(as.character(test$Pclass)), c(as.character(test$cAge)))
```

La fonction rend en sortie un vecteur de probabilités que l'on rajoute à notre data frame en créant une nouvelle colonne que l'on nomme **Proba_Survie**.

```
knitr::kable(test[1:6, ])
```

	Survived	Pclass	Sex	cAge	Proba_Survie
7	0	1	male	(40,60]	0.4035067
24	1	1	male	(20,40]	0.3793449
28	0	1	male	(0,20]	0.4930081
55	0	1	male	(60,80]	0.3465803
63	0	1	male	(40,60]	0.4035067
67	1	2	female	(20,40]	0.7945467

Grâce à notre modèle nous avons la probabilté de survie de tous les passagers et nous pouvons donc prédire la survie de chacun; on sert pour cela de la règle de *Maximum a Posteriori Proabbility (MAP)*, c'est-à-dire qu'un passager sera classifié comme survivant si et seulement si la probabilité de survie est supérieure à 0.5.

A l'aide de la fonction **ifelse** qui permet de comparer un vecteur à un nombre, ici on compare les probabilités de survie à la valeur 0.5; si celle-ci est est supérieur à 0.5, la nouvelle colonne créée nommée **Prediction_survie** contiendra la valeur 1 (c'est-à-dire survivant), sinon elle vaudra 0.

```
test[, "Prediction_survie"] = ifelse(test$Proba_Survie > 0.5,
    1, 0)
```

Les prédictions de survie pour chaque passager est le suivant:

```
knitr::kable(test[1:6, ])
```

	Survived	Pclass	Sex	cAge	Proba_Survie	Prediction_survie
7	0	1	male	(40,60]	0.4035067	0
24	1	1	male	(20,40]	0.3793449	0
28	0	1	male	(0,20]	0.4930081	0
55	0	1	male	(60,80]	0.3465803	0
63	0	1	male	(40,60]	0.4035067	0
67	1	2	female	(20,40]	0.7945467	1

Notre data frame *test* contient le vrai status de survie de chaque passager. Nous pouvons donc comparer notre modèle avec les véritables données et ainsi déterminer si notre modèle est efficace pour prédire la survie d'un passager.

La table ci-dessus est la table de contingence entre nos valeurs prédites et les vraies valeurs de survie et de mort des passagers

```
table_modele = round(prop.table(table(test$Prediction_survie,
    test$Survived)), 2)

colnames(table_modele) = c("Mort", "Survecu")

rownames(table_modele) = c("Prediction_Mort", "Prediction_Survecu")


table(test$Prediction_survie, test$Survived)
```

```
##
```

```
##      0  1
##   0 22 15
##   1  2 27
```

```
table_modele
```

```
##
##                       Mort Survecu
##   Prediction_Mort     0.33    0.23
##   Prediction_Survecu 0.03    0.41
```

Cette table nous permet de calculer la proportion de passagers que le modèle à réussit à bien classer. On constate que 41% des passagers qui ont survécu ont été classés correctement par notre modèle.

Pour calculer la performance de notre modèle , on utilise le critère *d'accuracy* (ou justesse) qui désigne la proportion des prédiction correctes effectuées par le modèle.

Ainsi $accuracy = \frac{22+27}{22+15+2+27}$ 0.7424242. Nous obtenons une valeur de justesse égale à 74% (sur 100 prédiction, 74 sont correctes). Notre modèle de prédiction de survie des passagers affiche donc des résultats assez satisfaisant du point de vue de la prédiction de survie d'un passager du Titanic sur la base de variables qui sont l'âge, le sexe et sa classe de voyage.

6. Conclusion

Avec notre modèle, on obtient une proportion de bonnes prédictions de 0.74. Ce score est assez satisfaisant mais plusieurs axes sont possibles afin d'améliorer notre modèle. La qualité du modèle pourrait être améliorer si nous avions moins de valeurs manquantes.