Utilisation de Caret avec un model predictif

```
#Installation de l'ensemble de ces packages
#install.packages("tidyverse")
#install.packages("caret")
#install.packages("mlbench")
#install.packages("lattice")
#install.packages("rpart")
#install.packages("xlsx")
#tinytex::install_tinytex()
library(here)
```

```
## here() starts at C:/Users/manu_/Desktop
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.0.3
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.3
```

```r
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.0.3
```

```r
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.0.3
```

```r
library(xlsx)
```

```
## Warning: package 'xlsx' was built under R version 4.0.3
```

```r
set.seed(12345)

# Import des donnees
df <- readxl::read_xlsx(path ="C:/Users/manu_/Desktop/caret.xlsx")

# Changement du format de la date
date <- as.Date(df$dateRep)
df3 <- cbind(df,date)
head(df3)
```

```
##   countriesAndTerritories    dateRep day month year cases deaths       date
## 1                  France 2020-12-14  14    12 2020 11533    150 2020-12-14
## 2                  France 2020-12-13  13    12 2020 13947    194 2020-12-13
## 3                  France 2020-12-12  12    12 2020 13406    627 2020-12-12
## 4                  France 2020-12-11  11    12 2020 13750    292 2020-12-11
## 5                  France 2020-12-10  10    12 2020 14595    296 2020-12-10
## 6                  France 2020-12-09   9    12 2020 13713    831 2020-12-09
```

```r
# D'abord, on cree notre echantillon d'entrainement en prenant 75%
# du dataset et un echantillon test.
train_data <- df3[df3$date < '2020-09-19',]
test_data  <- df3[df3$date > '2020-09-19',]

death <- test_data$deaths
attach(df3)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     date
```

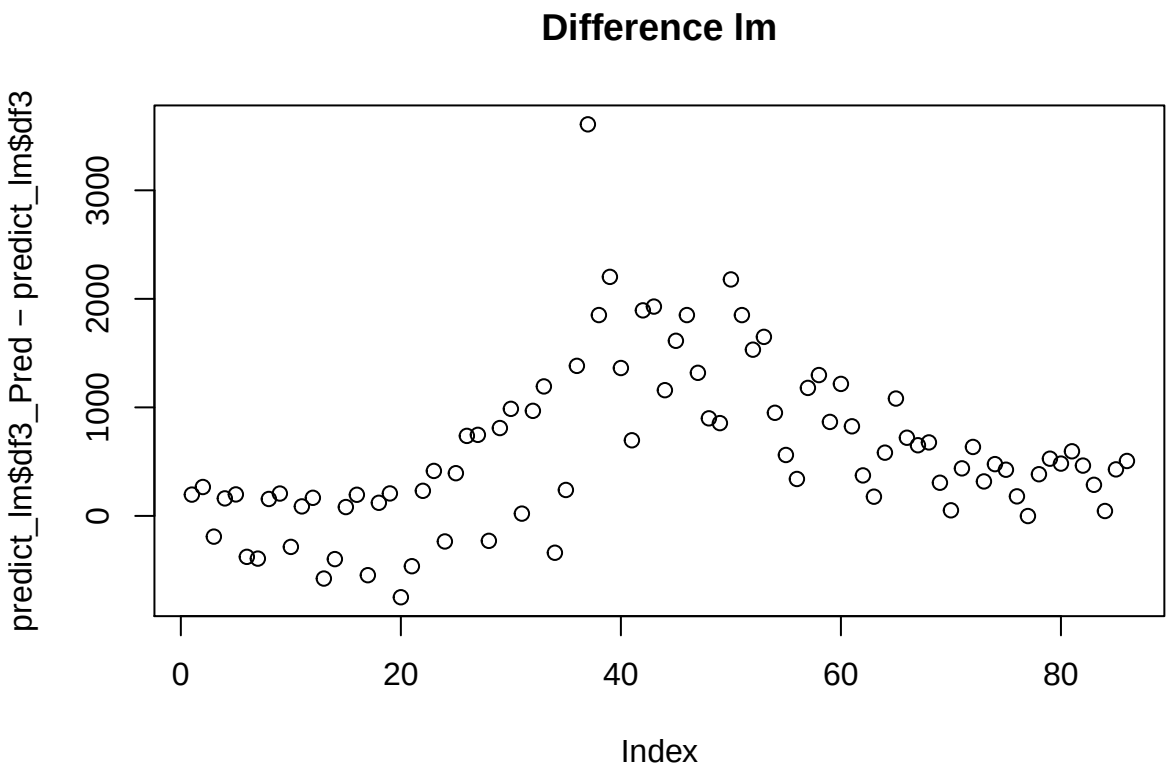
```r
# Model 1: Creattion d'un modele de regression lineaire
# entrainer avec l'echantillon d'entrainement
model_lm <- lm(deaths ~  cases + date, data = train_data)
#Model 2: Utilisation de rpart pour cree un
# arbre de decison avec l'echantillon d'entrainement
model_rpart  <- rpart(deaths ~  cases + date, data = train_data)

# On utlise les deux models pour faire
# une prediction en utilisant l'echantillon test
predict_lm <- predict(model_lm, test_data)
predict_lm <- data.frame(df3_Pred = predict_lm, df3 = test_data$deaths, date = test_data$date)
predict_rpart  <- predict(model_rpart,  test_data)
predict_rpart <- data.frame(df3_Pred = predict_rpart, df3 = test_data$deaths, date = test_data$date)

# On verifie les resultats en affichant les
# premieres lignes des dataset de predictions
head(predict_lm)
```

```
##   df3_Pred df3       date
## 1 346.1586 150 2020-12-14
## 2 460.2733 194 2020-12-13
## 3 436.0360 627 2020-12-12
## 4 453.2341 292 2020-12-11
## 5 493.8888 296 2020-12-10
## 6 453.6860 831 2020-12-09
```

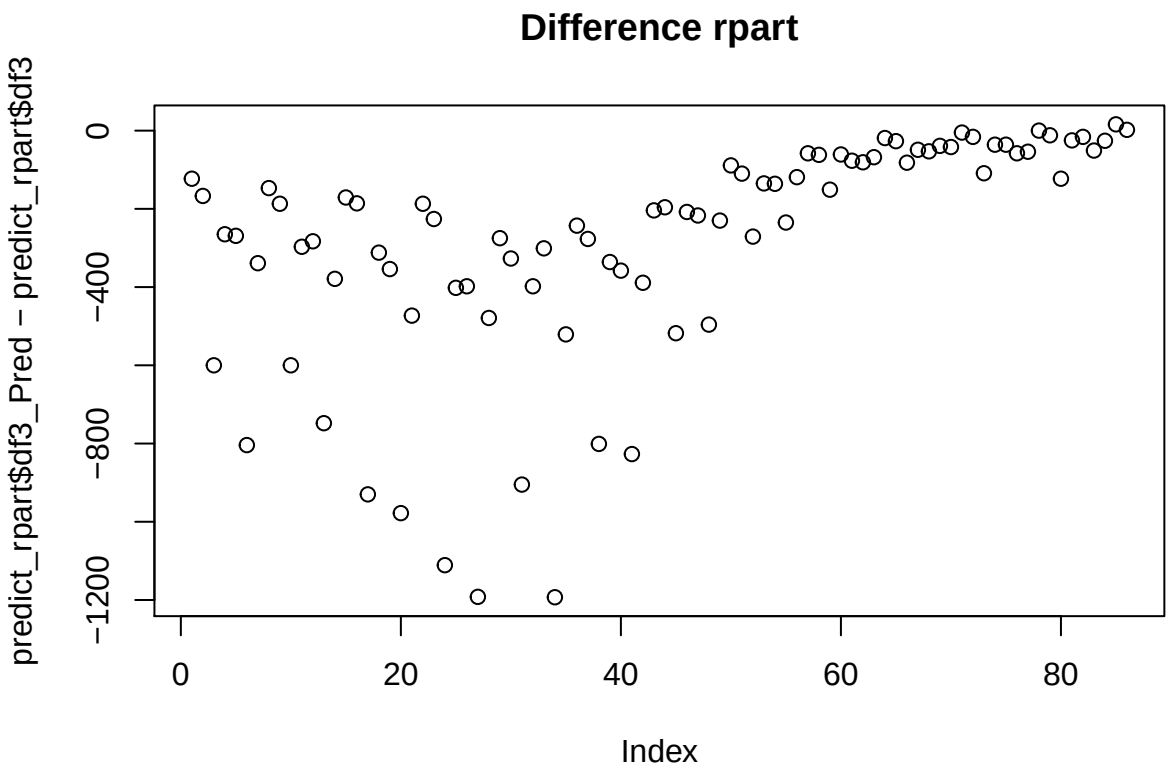
```r
head(predict_rpart)
```

```
##   df3_Pred df3       date
## 1 27.04348 150 2020-12-14
## 2 27.04348 194 2020-12-13
## 3 27.04348 627 2020-12-12
## 4 27.04348 292 2020-12-11
## 5 27.04348 296 2020-12-10
## 6 27.04348 831 2020-12-09
```

```r
# On compare les resultats obtenue a l'aide d'un graph
par(mfrow = c(1, 1))
plot(predict_lm$df3_Pred - predict_lm$df3, main = "Difference lm")
```




```
plot(predict_rpart$df3_Pred  - predict_rpart$df3,  main = "Difference rpart")
```




```
# D'apres les resultats des 2 graphiques, on peut retenir le modele de regression lineaire
# car le nuage de points comme une reaction normal d'un virus, c'est a dire sur une premiere
# phase croissante qui correspond a la propagation du virus puis une deuxieme phase de pic
# des morts et pour finir une phase decroissante de la lethalite.

# Nous pouvons voir qu'il y aura eventuelle troisieme vague.
```

#Bibliographie

https://docs.microsoft.com/fr-fr/sql/machine-learning/tutorials/r-predictive-model-train?view=sql-server-ver15

http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_package_caret.pdf

https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

https://github.com/dsobo/caret-beginner-tutorial

https://rstudio-pubs-static.s3.amazonaws.com/253860_05f11cddd938407a9cb3b06d9dc38c9a.html