Modelisation des risques d'AVC avec la régression logistique

Adama Sall

2025-04-22

Contexte:

Dans le cadre de mon mémoire d'intutilé ***« Modélisation des risques d'AVC à l'aide de la statistique des extrêmes et des technique de machine learning »*** qui vise à identifier et quantifier le risque d'AVC grâce aux différentes méthodes listée dans le plan. Dans cette section nous faisons la modelisation avec l'une de ces méthodes (**régression logistique)** pour identifier l'impact des différents facteurs sur la survenue des AVC grace aux variables disponible (***Genre,Âge ,Hypertension ,Maladie_Cardiaque ,Situation_Matrimoniale ,Type_travail ,Résidence Taux_glucose_moyen IMC ,Statut_Fumer ,AVC***). La description et l'analyse descriptive a été détaillé dans le rapport d'Analyse exploratoire des données

Question de recherche :

- Quels sont les facteurs de risque majeurs d'AVC dans la population étudiée ?
- Comment varie le risque individuel d'AVC à partir de ces facteurs et par interaction ?
- Comment ces facteurs influencent-ils la probabilité d'AVC **?**

Description des données

- **Source :** (non renseigné pour le moment)
- **Nombre d'observation:** 5 110
- **Variables :** 11 (3 numériques et 8 catégorielles)

Variable	Description	Type
Genre	Sexe du patient (Male, Female, Other)	Catégorique
Age	Âge du patient	Numérique
Hypertension	1 si le patient souffre d'hypertension, 0 sinon	Catégorique
Maladie_Cardiaque	1 si le patient a une maladie cardiaque, 0 sinon	Catégorique
Situation_Matrimoniale	Statut matrimonial (Yes/No)	Catégorique
Type_travail	Type d'emploi (Private, Self-employed, etc.)	Catégorique

Residence	Lieu de résidence (Urban/Rural)	Catégorique
Taux_glucose_moyen	Moyenne du taux de glucose	Numérique
IMC	Indice de masse corporelle	Numérique
Statut_Fumer	Statut tabagisme (never smoked, formerly smoked, smokes, unknown)	Catégorique
AVC	1 si AVC, 0 sinon	Catégorique

METHODOLOGIE ADAPTÉE

1. **Modélisation initiale**
2. **Sélection de variables (Backward)**
3. **Ajusttement , Calcul de l'odds ratio (OR) et intervalle de confiance**
4. **calcul de probabilite prédite d'AVC selon le modèle réduit**

- Probabilté predite de 'AVC l'Age uniquement
- Probabilté predite de 'AVC selon l'age et la presence de l'hypertension et maladie cardiaque
- Boxplot hypertension (Probabilté predite de 'AVC selon la presence de l'hypertension uniquement)
- Boxplot hypertension (Probabilté predite de 'AVC selon la presence de maladie cardiaque uniquement)

5. **Analyse par classes d'âge**

Courbes de probabilité d'AVC selon l'âge, l'hypertension et la maladie cardiaque

Probabilté predite de 'AVC selon la classe d'age et taux de gulcose

6. **Prédiction avec le modèle retenu**

Réechantillonnage avec smote

Normalisation puis developement du modele

```r
library(evd)

library(readxl)
library(dplyr)

library(tidyverse)

library(extRemes)

library(ggplot2)
data <- read.csv("C:/Users/Hp/Desktop/formation R/Données/base_avc.csv")
head(data)
```

```
##   Genre Age Hypertension Maladie_Cardiaque Situation_Matrimoniale Type_tra
vail
## 1     1  67            0                 1                       1
2
## 2     0  61            0                 0                       1
3
## 3     1  80            0                 1                       1
2
## 4     0  49            0                 0                       1
2
## 5     0  79            1                 0                       1
3
## 6     1  81            0                 0                       1
2
##   Residence Taux_glucose_moyen      IMC Statut_Fumer AVC
## 1         1           169.3575 36.60000            0   1
## 2         0           169.3575 28.89324            1   1
## 3         0           105.9200 32.50000            1   1
## 4         1           169.3575 34.40000            2   1
## 5         0           169.3575 24.00000            1   1
## 6         1           169.3575 29.00000            0   1
```

```r
attach(data)
```

Ajustement avec toutes les variables

```r
model_full <- glm(
  AVC ~ Genre + Age + Hypertension + Maladie_Cardiaque + Situation_Matrimonia
le +
       Type_travail + Residence + Taux_glucose_moyen + IMC + Statut_Fumer,
  data = data,
  family = binomial
)
summary(model_full)
```

```
##
## Call:
```

```
## glm(formula = AVC ~ Genre + Age + Hypertension + Maladie_Cardiaque +
##     Situation_Matrimoniale + Type_travail + Residence + Taux_glucose_moyen
+
##     IMC + Statut_Fumer, family = binomial, data = data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -7.735999   0.582587 -13.279  < 2e-16 ***
## Genre                   0.054689   0.140206   0.390 0.696489
## Age                     0.070876   0.005356  13.232  < 2e-16 ***
## Hypertension            0.382506   0.163357   2.342 0.019205 *
## Maladie_Cardiaque       0.325918   0.189546   1.719 0.085529 .
## Situation_Matrimoniale -0.189512   0.219065  -0.865 0.386987
## Type_travail           -0.051919   0.072299  -0.718 0.472685
## Residence               0.095975   0.137687   0.697 0.485770
## Taux_glucose_moyen      0.005970   0.001784   3.346 0.000819 ***
## IMC                     0.004460   0.011923   0.374 0.708335
## Statut_Fumer            0.025758   0.109921   0.234 0.814729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1990.4  on 5109  degrees of freedom
## Residual deviance: 1589.5  on 5099  degrees of freedom
## AIC: 1611.5
##
## Number of Fisher Scoring iterations: 7
```

) Ajustement avec variable reduite (Backward)

```
model_reduit <- glm(AVC ~ Age + Hypertension + Maladie_Cardiaque +
         Taux_glucose_moyen  ,
  data = data,
  family = binomial

)
summary(model_reduit)
```

```
##
## Call:
## glm(formula = AVC ~ Age + Hypertension + Maladie_Cardiaque +
##     Taux_glucose_moyen, family = binomial, data = data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -7.686410   0.381126 -20.168  < 2e-16 ***
## Age                 0.069263   0.005133  13.493  < 2e-16 ***
## Hypertension        0.380244   0.162647   2.338 0.019395 *
## Maladie_Cardiaque   0.335933   0.187442   1.792 0.073100 .
```

```
## Taux_glucose_moyen  0.006096   0.001743   3.497 0.000471 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1990.4  on 5109  degrees of freedom
## Residual deviance: 1591.6  on 5105  degrees of freedom
## AIC: 1601.6
##
## Number of Fisher Scoring iterations: 7
```

Calcul de l'odds ratio (OR) et intervalle de confiance

```
OR1 <- exp(coef(model_reduit))["Age"]
OR2 <- exp(coef(model_reduit))["Hypertension"]
OR3 <- exp(coef(model_reduit))["Maladie_Cardiaque"]
OR4 <- exp(coef(model_reduit))["Taux_glucose_moyen"]

confint_OR1 <- exp(confint(model_reduit))["Age", ]

## Attente de la réalisation du profilage...

confint_OR2 <- exp(confint(model_reduit))["Hypertension", ]

## Attente de la réalisation du profilage...

confint_OR3 <- exp(confint(model_reduit))["Maladie_Cardiaque", ]

## Attente de la réalisation du profilage...

confint_OR4 <- exp(confint(model_reduit))["Taux_glucose_moyen", ]

## Attente de la réalisation du profilage...

# Affichage
cat("Odds Ratio (Age) :", round(OR1, 2), "\n")

## Odds Ratio (Age) : 1.07

cat("Intervalle de confiance 95% : [", round(confint_OR1[1], 2), ";", round(c
onfint_OR1[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 1.06 ; 1.08 ]

cat("Odds Ratio (Hypertension) :", round(OR2, 2), "\n")

## Odds Ratio (Hypertension) : 1.46

cat("Intervalle de confiance 95% : [", round(confint_OR2[1], 2), ";", round(c
onfint_OR2[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 1.06 ; 2 ]
```

```r
cat("Odds Ratio (Maladie_Cardiaque) :", round(OR3, 2), "\n")
```

```
## Odds Ratio (Maladie_Cardiaque) : 1.4
```

```r
cat("Intervalle de confiance 95% : [", round(confint_OR3[1], 2), ";", round(c
onfint_OR3[2], 2), "]","\n")
```

```
## Intervalle de confiance 95% : [ 0.96 ; 2.01 ]
```

```r
cat("Odds Ratio (Taux_glucose_moyen) :", round(OR4, 2), "\n")
```

```
## Odds Ratio (Taux_glucose_moyen) : 1.01
```

```r
cat("Intervalle de confiance 95% : [", round(confint_OR4[1], 2), ";", round(c
onfint_OR4[2], 2), "]","\n")
```

```
## Intervalle de confiance 95% : [ 1 ; 1.01 ]
```

Interpretation

Aprés avoir appliquer la méthode de backward nous avons le résultat ci dessus

Le modèle identifie l'âge comme facteur dominant, suivi de l'hypertension et de l'hyperglycémie

Intercept : La probabilité de base d'AVC est très faible lorsque toutes les autres variables sont à 0

Age : Chaque année supplémentaire augmente le log-odds d'AVC de 0.069 (OR = exp(0.069) ≈ 1.071). Effet hautement significatif ($p < 0.001$). Chaque année supplémentaire augmente le risque d'AVC de 7.2%.

Chaque unité supplémentaire augmente le log-odds d'AVC de 0.006 (OR ≈ 1.006). Effet très significatif ($p = 0.0005$).

Les patients hypertendus ont un risque d'AVC 46% plus élevé (OR = 1.46) que les non-hypertendus.

Maladie_Cardiaque : Augmentation non significative au seuil 5% ($p = 0.073$), mais effet marginalement significatif (OR ≈ 1.40). Les patients avec une maladie cardiaque ont un risque d'AVC 40% plus élevé (OR = 1.40)

Une augmentation d'une unité du taux de glucose augmente légèrement le risque 0.6%.

calcul de probabilite

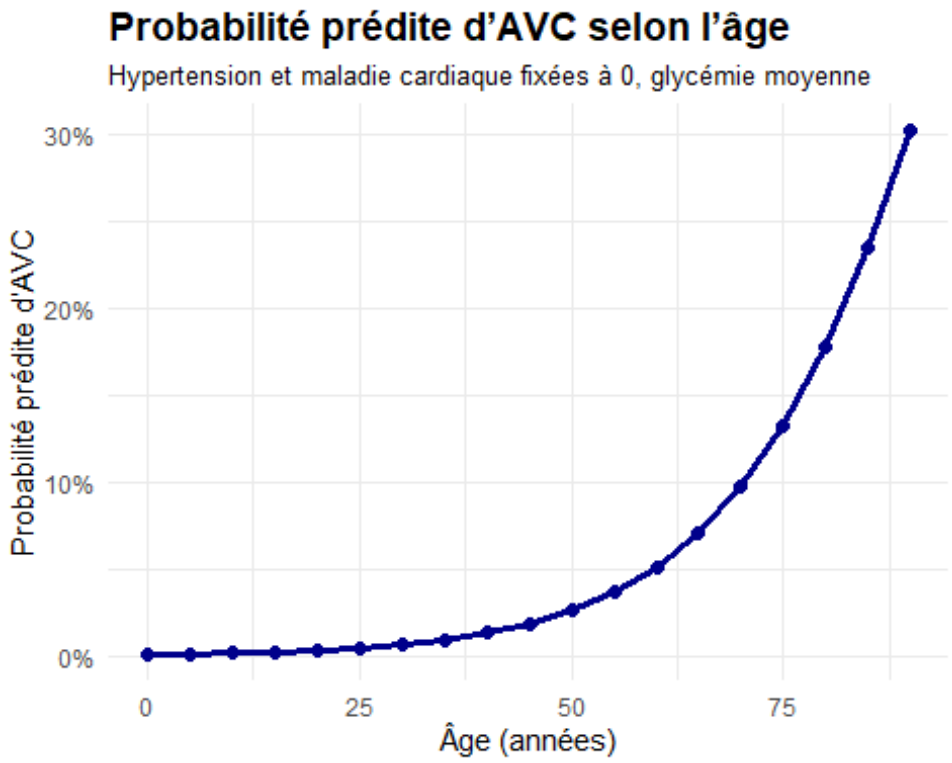
```r
# Ajouter les probabilités prédites au dataframe
data$Prob_AVC <- predict(model_reduit, type = "response")

# Aperçu des 5 premières lignes
head(data[, c("Age", "Hypertension", "Maladie_Cardiaque", "Taux_glucose_moyen
", "Prob_AVC")], 5)
```

```
##   Age Hypertension Maladie_Cardiaque Taux_glucose_moyen   Prob_AVC
## 1  67            0                 1           169.3575 0.15742566
## 2  61            0                 0           169.3575 0.08098633
## 3  80            0                 1           105.9200 0.23798111
## 4  49            0                 0           169.3575 0.03696289
## 5  79            1                 0           169.3575 0.30958909
```

Probabilté predite de ’AVC l’Age uniquement

```
# Création  jeu de données  avec uniquement l'âge variable
new_data_age <- data.frame(
  Age = seq(0, 90, by = 5),
  Hypertension = 0,
  Maladie_Cardiaque = 0,
  Taux_glucose_moyen = mean(data$Taux_glucose_moyen, na.rm = TRUE)
)

# Calculer les probabilités prédites d'AVC selon l'âge uniquement
new_data_age$Prob_AVC <- predict(model_reduit, newdata = new_data_age, type =
"response")

# Visualisation

ggplot(new_data_age, aes(x = Age, y = Prob_AVC)) +
  geom_line(color = "darkblue", size = 1.2) +
  geom_point(color = "darkblue", size = 2) +
  labs(
    title = "Probabilité prédite d’AVC selon l’âge",
    subtitle = "Hypertension et maladie cardiaque fixées à 0, glycémie moyenn
e",
    x = "Âge (années)",
    y = "Probabilité prédite d’AVC"
  ) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 10)
  )
```

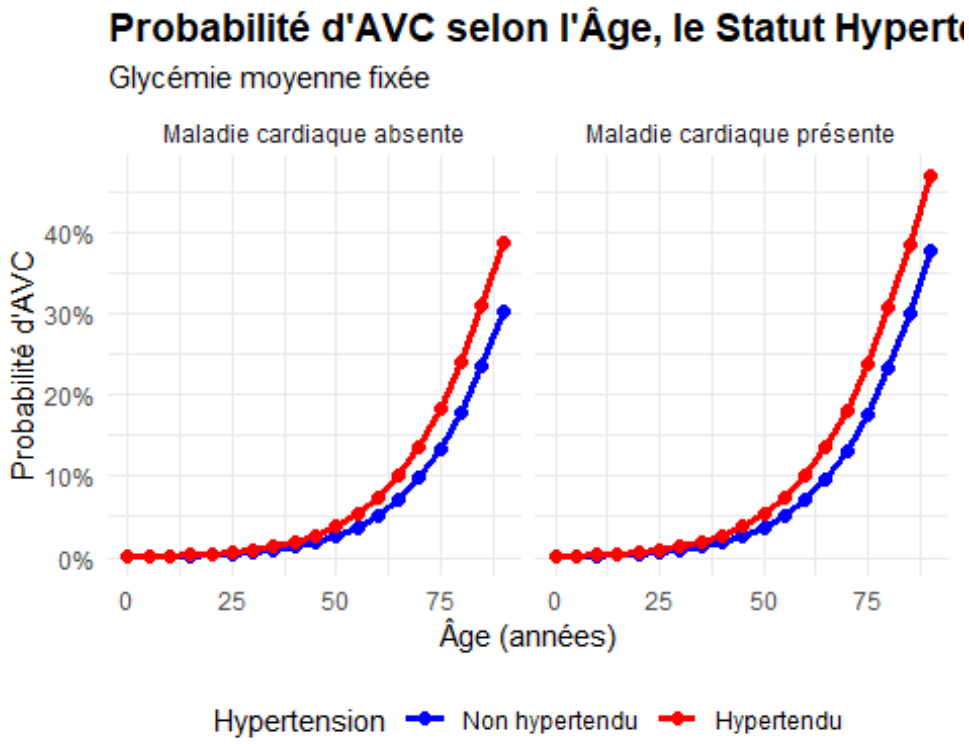
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Probabilté predite de ’AVC selon l’age hypertension et maladie cardiaque

```
# Créer un jeu de données avec Maladie_Cardiaque à 0 et 1
new_data <- expand.grid(
  Age = seq(0, 90, by = 5),
  Hypertension = c(0, 1),
  Maladie_Cardiaque = c(0, 1),
  Taux_glucose_moyen = mean(data$Taux_glucose_moyen)
)

# Calculer les probabilités prédites
new_data$Prob_AVC <- predict(model_reduit, newdata = new_data, type = "respon
se")

# Graphique avec facet_wrap pour Maladie_Cardiaque
ggplot(new_data, aes(x = Age, y = Prob_AVC, color = factor(Hypertension))) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  scale_color_manual(
    values = c("0" = "blue", "1" = "red"),
    labels = c("Non hypertendu", "Hypertendu")
  ) +
  labs(
    title = "Probabilité d'AVC selon l'Âge, le Statut Hypertensif et la Malad
ie Cardiaque",
```

```
    subtitle = "Glycémie moyenne fixée",
    x = "Âge (années)",
    y = "Probabilité d'AVC",
    color = "Hypertension"
  ) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  facet_wrap(~ Maladie_Cardiaque, labeller = labeller(Maladie_Cardiaque = c(`
0` = "Maladie cardiaque absente", `1` = "Maladie cardiaque présente"))) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    plot.title = element_text(face = "bold", size = 14)
  )
```



Interpretation

La probabilité d'AVC augmente de façon exponentielle avec l'âge, quel que soit le statut hypertensif.

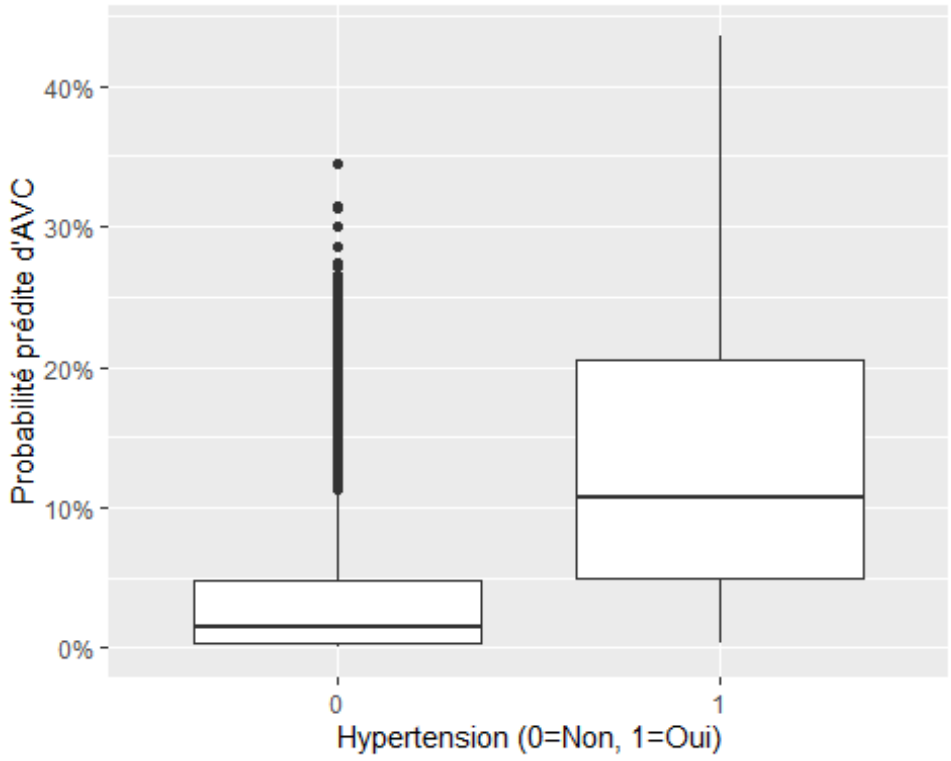
De 0 Jusqu'à 50 ans, la probabilité reste faible (<5%), puis elle croît rapidement après 60 ans. Àpres l'âge 30, la courbe rouge (hypertension) est au-dessus de la courbe bleue (non hypertension), ce qui montre que l'hypertension augmente significativement le risque d'AVC. À 60 ans, la différence est visible mais modérée. A80 ans, la probabilité d'AVC depasse 45% chez les hypertendus contre 37% environ chez les non hypertendus. Cet effet est observé aussi bien en l'absence qu'en présence de maladie cardiaque

La presence de la maladie cardiaque augemente la probabilité mais cette derniere est plus élevé chez les hypertendu L'effet de l'hypertension et de la maladie cardiaque est cumulatif : le risque maximal est observé chez les patients âgés, hypertendus et avec maladie cardiaque.

Conclusion : L'hypertension augmente de façon remarquable le risque d'AVC à tous age , et cet effet est plus marquant chez les personnes âgées (aprés 60 ans) La maladie cardiaque ajoute un risque supplémentaire.

Boxplot hypertension

```r
data$Predicted_Prob <- predict(model_reduit, type = "response")

ggplot(data, aes(x = factor(Hypertension), y = Predicted_Prob)) +
  geom_boxplot() +
  labs(x = "Hypertension (0=Non, 1=Oui)", y = "Probabilité prédite d'AVC") +
  scale_y_continuous(labels = scales::percent)
```



Interpretation

Chez les hypertendu la médian est plus haute , ce qui confirme que ce groupe a un risque prédictif d'AVC beaucoup plus élevé. La distribution est plus étalée, avec une médiane autour de 10-12% et 75% des individus de ce groupe qui dépasse 20%. Certains cas atteignent ou dépassent 40%.

Chez les non hypertendus, la majorité a une probabilité prédite faible (<10%), avec quelques valeurs extrêmes au-dessus de 10%.Donc il y'a des individus non hypertendus qui présentent un risque élevé, probablement à cause d'autres facteurs (âge, maladie cardiaque, taux de glucose...)

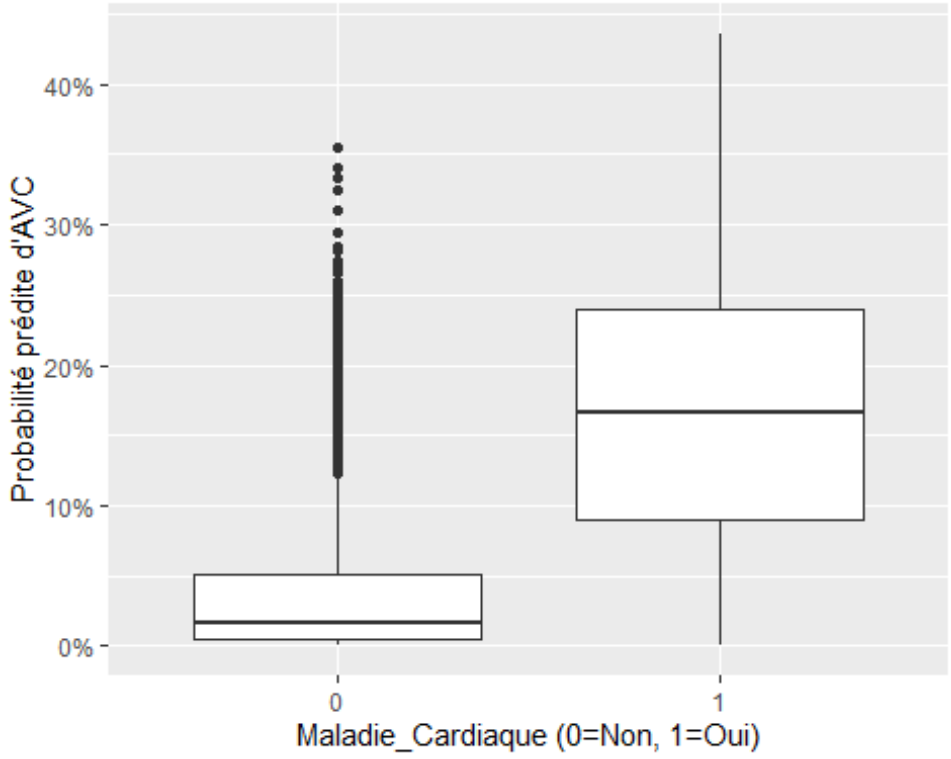
Conclusion:

L'hypertension est un facteur de risque majeur et indépendant d'AVC.

Même chez les non hypertendus, le risque peut augmenter avec l'âge ou d'autres facteurs.

Boxplot maladie cardiaque

```
data$Predicted_Prob <- predict(model_reduit, type = "response")

ggplot(data, aes(x = factor(Maladie_Cardiaque), y = Predicted_Prob)) +
  geom_boxplot() +
  labs(x = "Maladie_Cardiaque (0=Non, 1=Oui)", y = "Probabilité prédite d'AVC
") +
  scale_y_continuous(labels = scales::percent)
```



Decomposition de l'age en 3

```
Age_classe  = cut(Age, breaks=c(25,45,65,82))
levels(Age_classe ) = c("classe 1","classe 2","classe 3")
eff_age = table(Age_classe )
eff_age
```

```
## Age_classe
## classe 1 classe 2 classe 3
##     1325     1527      965
```

```
model_age_cat <- glm(
  AVC ~ Age_classe + Hypertension + Maladie_Cardiaque + Taux_glucose_moyen ,
  data = data,
  family = binomial
)
summary(model_age_cat)
```

```
##
## Call:
## glm(formula = AVC ~ Age_classe + Hypertension + Maladie_Cardiaque +
##     Taux_glucose_moyen, family = binomial, data = data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.399313   0.349533 -15.447  < 2e-16 ***
## Age_classeclasse 2  1.744817   0.325509   5.360 8.31e-08 ***
## Age_classeclasse 3  2.859938   0.320936   8.911  < 2e-16 ***
## Hypertension        0.415723   0.161155   2.580 0.009890 **
## Maladie_Cardiaque   0.416435   0.185866   2.241 0.025058 *
## Taux_glucose_moyen  0.006024   0.001744   3.454 0.000552 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1830.1  on 3816  degrees of freedom
## Residual deviance: 1580.8  on 3811  degrees of freedom
##   (1293 observations effacées parce que manquantes)
## AIC: 1592.8
##
## Number of Fisher Scoring iterations: 7
```

```
OO1 <- exp(coef(model_age_cat))["Age_classeclasse 2"]
OO11 <- exp(coef(model_age_cat))["Age_classeclasse 3"]
OO2 <- exp(coef(model_age_cat))["Hypertension"]
OO3 <- exp(coef(model_age_cat))["Maladie_Cardiaque"]
OO4 <- exp(coef(model_age_cat))["Taux_glucose_moyen"]

confint_OO1 <- exp(confint(model_age_cat))["Age_classeclasse 2", ]

## Attente de la réalisation du profilage...

confint_OO11 <- exp(confint(model_age_cat))["Age_classeclasse 3", ]

## Attente de la réalisation du profilage...

confint_OO2 <- exp(confint(model_age_cat))["Hypertension", ]
```

```
## Attente de la réalisation du profilage...

confint_OO3 <- exp(confint(model_age_cat))["Maladie_Cardiaque", ]

## Attente de la réalisation du profilage...

confint_OO4 <- exp(confint(model_age_cat))["Taux_glucose_moyen", ]

## Attente de la réalisation du profilage...

# Affichage
cat("Odds Ratio (classe 2) :", round(OO1, 2), "\n")

## Odds Ratio (classe 2) : 5.72

cat("Intervalle de confiance 95% : [", round(confint_OO1[1], 2), ";", round(c
onfint_OO1[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 3.16 ; 11.46 ]

cat("Odds Ratio (classe 3) :", round(OO11, 2), "\n")

## Odds Ratio (classe 3) : 17.46

cat("Intervalle de confiance 95% : [", round(confint_OO11[1], 2), ";", round(
confint_OO11[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 9.74 ; 34.69 ]

cat("Odds Ratio (Hypertension) :", round(OO2, 2), "\n")

## Odds Ratio (Hypertension) : 1.52

cat("Intervalle de confiance 95% : [", round(confint_OO2[1], 2), ";", round(c
onfint_OO2[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 1.1 ; 2.07 ]

cat("Odds Ratio (Maladie_Cardiaque) :", round(OO3, 2), "\n")

## Odds Ratio (Maladie_Cardiaque) : 1.52

cat("Intervalle de confiance 95% : [", round(confint_OO3[1], 2), ";", round(c
onfint_OO3[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 1.04 ; 2.17 ]

cat("Odds Ratio (Taux_glucose_moyen) :", round(OO4, 2), "\n")

## Odds Ratio (Taux_glucose_moyen) : 1.01

cat("Intervalle de confiance 95% : [", round(confint_OO4[1], 2), ";", round(c
onfint_OO4[2], 2), "]","\n")

## Intervalle de confiance 95% : [ 1 ; 1.01 ]
```

Le risque d'AVC augemente avec l'age , hypertension , maladie cardiaque taux de glucose.

L'âge est un facteur de risque majeur , les classes 2 et 3 ont un risque d'AVC multiplié par 5.7 et 17.5 respectivement par rapport à la classe 1

L'hypertension et la maladie cardiaque augmentent significativement le risque d'AVC, avec un effet similaire (OR ~1.5)

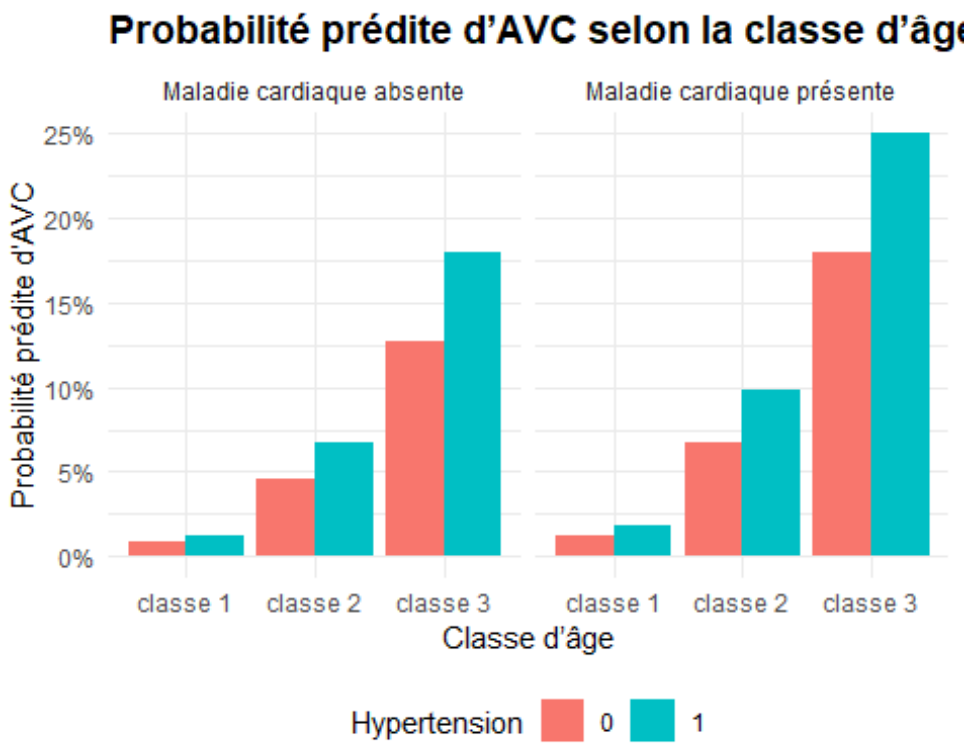
Glycémie a un impact faible (OR = 1,01) mais très significatif (+0,6% par unité, p = 0,00055).

```
data$Age_classe <- cut(
  data$Age,
  breaks = c(25, 45, 65, 82),
  labels = c("classe 1", "classe 2", "classe 3"),
  right = FALSE
)
```

Probabilté predite de 'AVC selon la classe d'age hypertension et maladie cardiaque

```
# Moyenne du taux de glucose dans vos données
mean_glucose <- mean(data$Taux_glucose_moyen, na.rm = TRUE)

# Jeu de données synthétique
new_data <- expand.grid(
  Age_classe = factor(c("classe 1", "classe 2", "classe 3"), levels = levels(
data$Age_classe)),
  Hypertension = c(0, 1),
  Maladie_Cardiaque = c(0, 1),
  Taux_glucose_moyen = mean_glucose
)
new_data$Prob_AVC <- predict(model_age_cat, newdata = new_data, type = "respo
nse")

ggplot(new_data, aes(x = Age_classe, y = Prob_AVC, fill = factor(Hypertension
))) +
  geom_bar(stat = "identity", position = position_dodge()) +
  facet_wrap(~ Maladie_Cardiaque,labeller = labeller(Maladie_Cardiaque = c(`0
` = "Maladie cardiaque absente", `1` = "Maladie cardiaque présente"))) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(
    title = "Probabilité prédite d'AVC selon la classe d'âge, l'hypertension
et la maladie cardiaque",
    x = "Classe d'âge",
    y = "Probabilité prédite d'AVC",
    fill = "Hypertension"
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
```

```
    plot.title = element_text(face = "bold", size = 14)
  )
```



Interpretation

La probabilité d'AVC augmente fortement avec l'âge, quel que soit le statut hypertensif ou cardiaque :

Classe 1 : Risque très faible (<2%)

Classe 2 : Risque modéré (5% à 10%)

Classe 3 : Risque élevé (jusqu'à 25% ou plus)

À chaque classe d'âge, la probabilité predite sue les hypertendus est plus haute que celle prredite sur les non hypertendu. et que l'écart entre hypertendus et non hypertendus s'agrandit avec l'âge :

En classe 1, l'effet est modeste, en classe 2 l'effet augmente un peu et en classe 3, la différence devient très marquée (plus de 10% d'écart).

La présence de maladie cardiaque augmente la probabilité d'AVC dans toutes les classes d'âge :

À classe d'âge et statut hypertensif identiques, le risque est toujours plus élevé chez les hypertendus en presence de la maladie cardiaque présente.

L’effet est particulièrement visible en classe 3, où la probabilité d’AVC peut dépasser 25% chez les patients hypertendus et cardiaques.

Conclusion :

Le risque maximal d’AVC est observé chez les patients de classe 3, hypertendus et avec une maladie cardiaque.

Le risque minimal concerne les sujets jeunes (classe 1), non hypertendus et sans maladie cardiaque.

L’âge est le principal facteur de risque

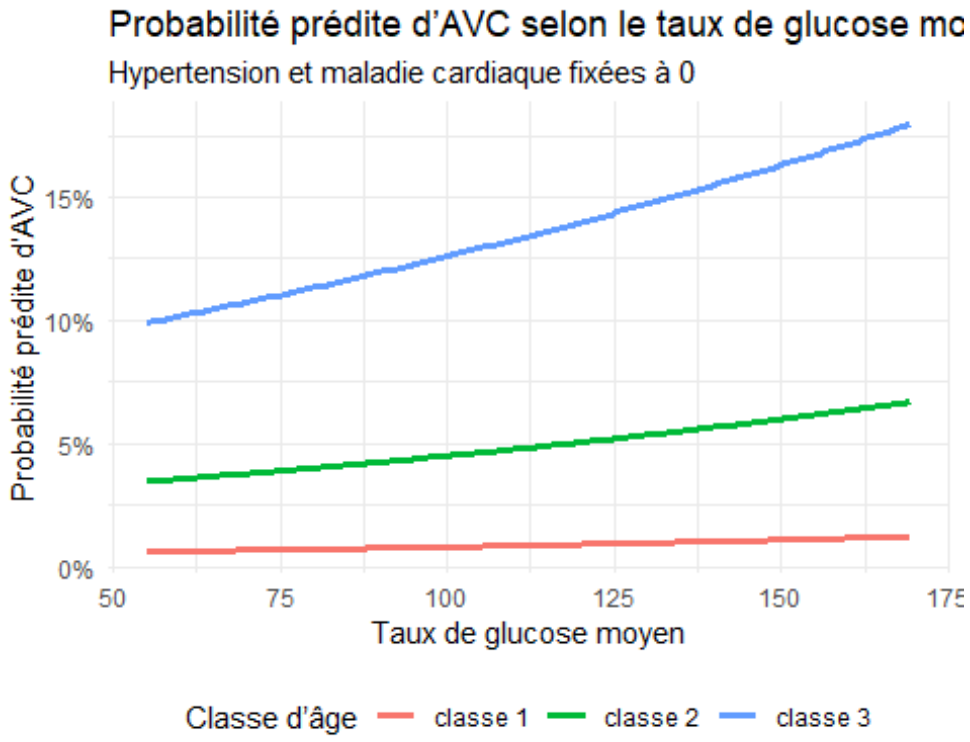
L’hypertension multiplie le risque dans toutes les tranches d’âge, mais surtout chez les personnes ayant plus 60 ans.

La maladie cardiaque ajoute un risque supplémentaire important, de façon cumulative avec l’âge et l’hypertension.

Probabilté predite de ’AVC selon la classe d’age et taux de gulcose

```r
# Séquence de taux de glucose couvrant la plage observée
glucose_seq <- seq(min(data$Taux_glucose_moyen, na.rm = TRUE),
                   max(data$Taux_glucose_moyen, na.rm = TRUE),
                   length.out = 100)

# Niveaux de la variable Age_classe (facteur)
age_classes <- levels(data$Age_classe)

hypertension_ref <- 0
maladie_cardiaque_ref <- 0

# Créer le data.frame combiné
new_data <- expand.grid(
  Taux_glucose_moyen = glucose_seq,
  Age_classe = age_classes,
  Hypertension = hypertension_ref,
  Maladie_Cardiaque = maladie_cardiaque_ref
)
new_data$Prob_AVC <- predict(model_age_cat, newdata = new_data, type = "respo
nse")



ggplot(new_data, aes(x = Taux_glucose_moyen, y = Prob_AVC, color = Age_classe
)) +
  geom_line(size = 1.2) +
  labs(
    title = "Probabilité prédite d’AVC selon le taux de glucose moyen et la c
lasse d’âge",
    subtitle = "Hypertension et maladie cardiaque fixées à 0",
```

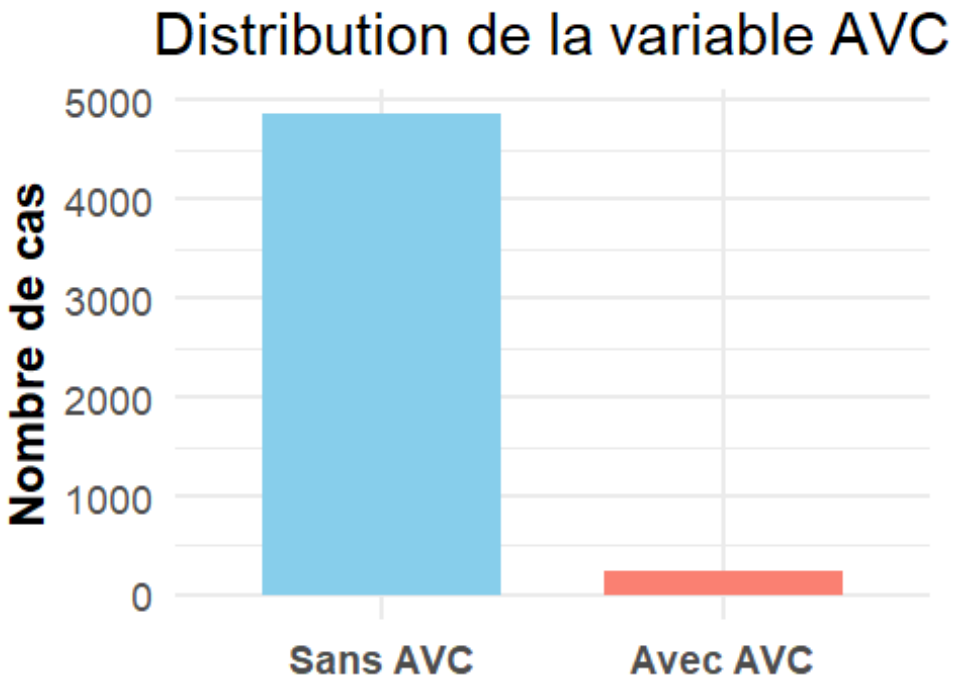
```
    x = "Taux de glucose moyen",
    y = "Probabilité prédite d'AVC",
    color = "Classe d'âge"
  ) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Prédiction

```
library(ggplot2)

# Préparation des données pour l'affichage
data$AVC_libelle <- factor(data$AVC, levels = c(0, 1), labels = c("Sans AVC",
"Avec AVC"))

# Bar plot
ggplot(data, aes(x = AVC_libelle, fill = AVC_libelle)) +
  geom_bar(show.legend = FALSE, width = 0.7) +
  scale_fill_manual(values = c("Sans AVC" = "skyblue", "Avec AVC" = "salmon")
) +
  labs(
    x = "",
    y = "Nombre de cas",
    title = "Distribution de la variable AVC"
  ) +
```

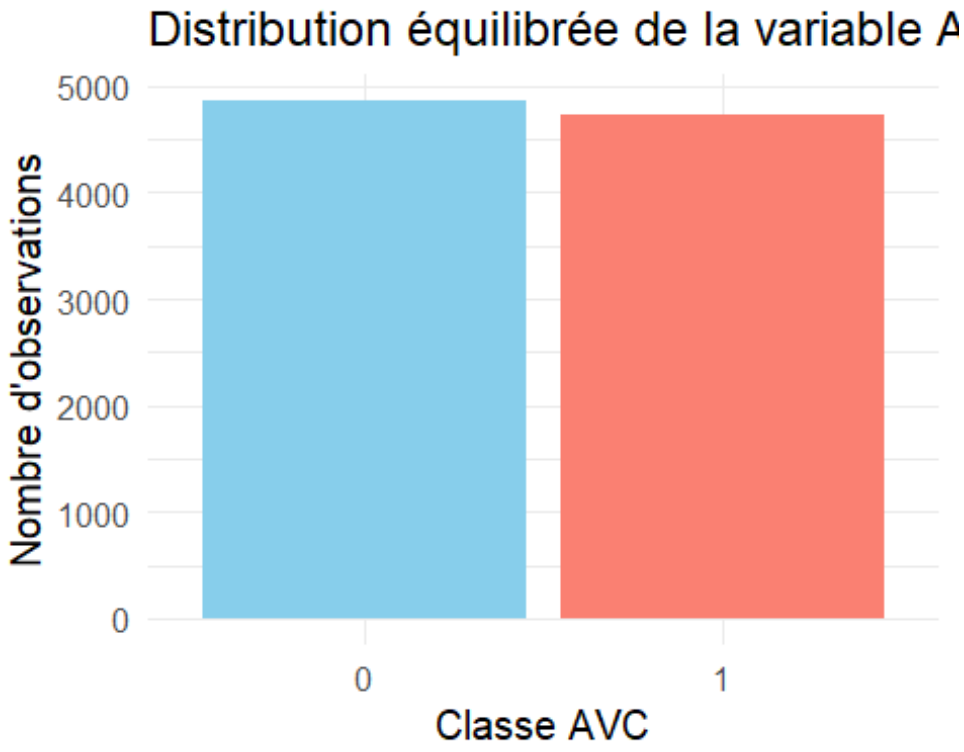
```r
  theme_minimal(base_size = 18) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  )
```



Oversampling

```r
library(smotefamily)

## Warning: le package 'smotefamily' a été compilé avec la version R 4.3.3

data$AVC <- as.factor(data$AVC)

# Séparer les variables explicatives et la cible
X <- data[, setdiff(names(data), c("AVC", "Prob_AVC", "Predicted_Prob", "Age_
classe", "AVC_libelle"))]

y <- data$AVC

# Appliquer SMOTE

smote_output <- SMOTE(X, y, K = 5, dup_size = 0)

# smote_output$data contient les données augmentées avec la variable cible en
dernière colonne
```

```r
data_smote <- smote_output$data
names(data_smote)[ncol(data_smote)] <- "AVC"
data_smote$AVC <- as.factor(data_smote$AVC)
ggplot(data_smote, aes(x = AVC, fill = AVC)) +
  geom_bar(show.legend = FALSE) +
  scale_fill_manual(values = c("skyblue", "salmon")) +
  labs(title = "Distribution équilibrée de la variable AVC",
       x = "Classe AVC",
       y = "Nombre d'observations") +
  theme_minimal(base_size = 15)
```



Normalisation des données

```r
var_quat <- c("Genre", "Hypertension", "Maladie_Cardiaque", "Situation_Matrim
oniale",
              "Type_travail", "Residence", "Statut_Fumer", "AVC")

# Sélectionner les colonnes numériques à normaliser
num_vars <- setdiff(names(data_smote), var_quat)

# Normaliser uniquement les variables numériques
data_smote[num_vars] <- scale(data_smote[num_vars])
```

Séparation en train et test puis entrainement des données

```r
library(caret)
```

```r
set.seed(123)
trainIndex <- createDataPartition(data_smote$AVC, p = 0.7, list = FALSE)

train <- data_smote[trainIndex, ]
test <- data_smote[-trainIndex, ]

# Entraînement du modèle
model_final <- glm(AVC ~ Age + Hypertension + Maladie_Cardiaque + Taux_glucos
e_moyen,
                   data = train, family = binomial)

# Résumé du modèle
summary(model_final)
```

```
##
## Call:
## glm(formula = AVC ~ Age + Hypertension + Maladie_Cardiaque +
##     Taux_glucose_moyen, family = binomial, data = train)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.37965    0.03775 -10.058  < 2e-16 ***
## Age                 1.70395    0.04823  35.328  < 2e-16 ***
## Hypertension        0.48263    0.09296   5.192 2.08e-07 ***
## Maladie_Cardiaque   0.50334    0.11624   4.330 1.49e-05 ***
## Taux_glucose_moyen  0.20944    0.03156   6.637 3.20e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9307.7  on 6714  degrees of freedom
## Residual deviance: 6342.3  on 6710  degrees of freedom
## AIC: 6352.3
##
## Number of Fisher Scoring iterations: 5
```

Prédiction sur la base de test

```r
predictions_prob <- predict(model_final, newdata = test, type = "response")
predictions_class <- ifelse(predictions_prob > 0.5, 1, 0)

# matrice de confusion
conf_matrix <- confusionMatrix(data = as.factor(predictions_class),
                               reference = as.factor(test$AVC),
                               positive = "1")
```

```r
#  Afficher les résultats
conf_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1094  291
##          1  364 1128
##
##                Accuracy : 0.7723
##                  95% CI : (0.7566, 0.7875)
##     No Information Rate : 0.5068
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5449
##
##  Mcnemar's Test P-Value : 0.004904
##
##             Sensitivity : 0.7949
##             Specificity : 0.7503
##          Pos Pred Value : 0.7560
##          Neg Pred Value : 0.7899
##              Prevalence : 0.4932
##          Detection Rate : 0.3921
##    Detection Prevalence : 0.5186
##       Balanced Accuracy : 0.7726
##
##        'Positive' Class : 1
##
```

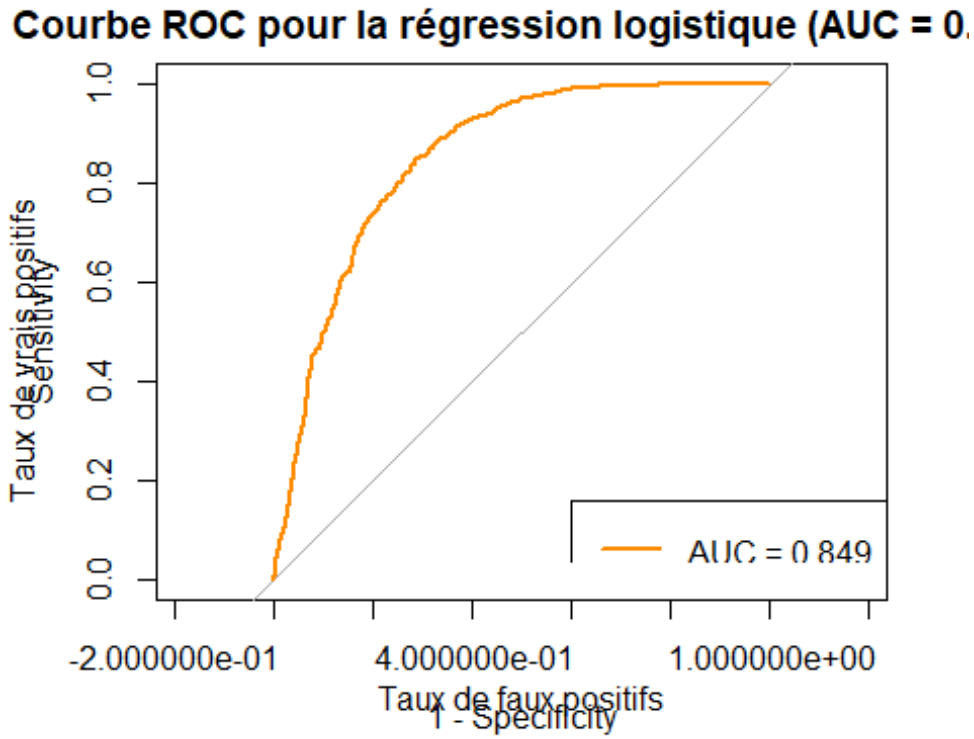
```r
library(pROC)

# Calcul de  ROC
roc <- roc(test$AVC, predictions_prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
# Calcul de l'AUC
auc_value <- auc(roc)

# Tracé de la courbe ROC
plot(roc, col = "darkorange", lwd = 2, legacy.axes = TRUE,
     main = paste0("Courbe ROC pour la régression logistique (AUC = ", round(
```

```
auc_value, 2), ")"))

# Axes et légende
xlabel <- "Taux de faux positifs"
ylabel <- "Taux de vrais positifs"
title(xlab = xlabel, ylab = ylabel)
legend("bottomright", legend = paste("AUC =", round(auc_value, 3)), col = "da
rkorange", lwd = 2)
```



Calcul les taux des taux de bon et de mauvais classement

```
total_cases <- nrow(test)
correct_classifications <- sum(predictions_class == test$AVC)
incorrect_classifications <- total_cases - correct_classifications

taux_bon_classement <- (correct_classifications / total_cases) * 100
taux_mauvais_classement <- (incorrect_classifications / total_cases) * 100

# 3. Afficher les résultats
cat("Taux de bon classement:", round(taux_bon_classement, 2), "%\n")

## Taux de bon classement: 77.23 %

cat("Taux de mauvais classement:", round(taux_mauvais_classement, 2), "%\n")

## Taux de mauvais classement: 22.77 %
```

Interprétation

Globalement le modèle a un taux de bon classement de 77.27% (accuracy) et un taux de mauvais classement de 22,73%. IL peut détecter correctement 79.41% des cas d'AVC réels, mais il rate environ 20.59% des cas d'AVC (faux négatifs).

Dans la classe 0 (sans AVC), sur 1388 cas le modèle réussit à prédire correctement 1096 cas et fais 292 erreurs. Dans la classe 1 (avec AVC), sur 1489 cas le modèle réussit à prédire correctement 1127 cas et fais 362 erreurs.