

DATA HÔPITAL

1/ Compréhension de métier: (1 SEUL MÉTIER)

- Objectifs Métiers (Détecter les caractéristiques):
 - Améliorer la qualité des soins médicaux fournis par l'hôpital.**
 - Optimiser les processus internes
 - Personnaliser les traitements des patients
 - Réduire les risques de maladies cardiovasculaires
- Objectif Data Mining: (chaque obj métier doit avoir les objectifs DM)

L'objectif du data mining est d'extraire à partir des données des patients des relations entre les données et créer des clusters et classes qui nous aideront dans la classification des patients.

2/ Compréhension des données:

DIM DATASET + boxplot

```
> summary(hospital)
  male      age      education  currentSmoker  cigsPerDay  BPMeds
Min.   :0.0000  Min.   :32.00  Min.   :1.000  Min.   :0.0000  Min.   : 0.000  Min.   :0.00000
1st Qu.:0.0000  1st Qu.:42.00  1st Qu.:1.000  1st Qu.:0.0000  1st Qu.: 0.000  1st Qu.:0.00000
Median :0.0000  Median :49.00  Median :2.000  Median :0.0000  Median : 0.000  Median :0.00000
Mean   :0.4292  Mean   :49.58  Mean   :1.979  Mean   :0.4941  Mean   : 9.006  Mean   :0.02962
3rd Qu.:1.0000  3rd Qu.:56.00  3rd Qu.:3.000  3rd Qu.:1.0000  3rd Qu.:20.000  3rd Qu.:0.00000
Max.   :1.0000  Max.   :70.00  Max.   :4.000  Max.   :1.0000  Max.   :70.000  Max.   :1.00000
NA's   :105

prevalentStroke  prevalentHyp  diabetes  totChol  sysBP  diaBP
Min.   :0.000000  Min.   :0.0000  Min.   :0.00000  Min.   :107.0  Min.   : 83.5  Min.   : 48.0
1st Qu.:0.000000  1st Qu.:0.0000  1st Qu.:0.00000  1st Qu.:206.0  1st Qu.:117.0  1st Qu.: 75.0
Median :0.000000  Median :0.0000  Median :0.00000  Median :234.0  Median :128.0  Median : 82.0
Mean   :0.005896  Mean   :0.3106  Mean   :0.02571  Mean   :236.7  Mean   :132.4  Mean   : 82.9
3rd Qu.:0.000000  3rd Qu.:1.0000  3rd Qu.:0.00000  3rd Qu.:263.0  3rd Qu.:144.0  3rd Qu.: 90.0
Max.   :1.000000  Max.   :1.0000  Max.   :1.00000  Max.   :696.0  Max.   :295.0  Max.   :142.5
NA's   :50

BMI  heartRate  glucose  TenYearCHD
Min. :15.54  Min.   : 44.00  Min.   : 40.00  Min.   :0.0000
1st Qu.:23.07  1st Qu.: 68.00  1st Qu.: 71.00  1st Qu.:0.0000
Median :25.40  Median : 75.00  Median : 78.00  Median :0.0000
Mean   :25.80  Mean   : 75.88  Mean   : 81.96  Mean   :0.1519
3rd Qu.:28.04  3rd Qu.: 83.00  3rd Qu.: 87.00  3rd Qu.:0.0000
Max.   :56.80  Max.   :143.00  Max.   :394.00  Max.   :1.0000
NA's   :19    NA's   :1    NA's   :388
```

INTERPRETATIONS : valeurs manquantes + aberrantes + valeurs qualit et quantit + cible
sup + non sup

diabetes

diaBP

Histogrammes pour les distributions (Age, sysbp, diabb, bmi)

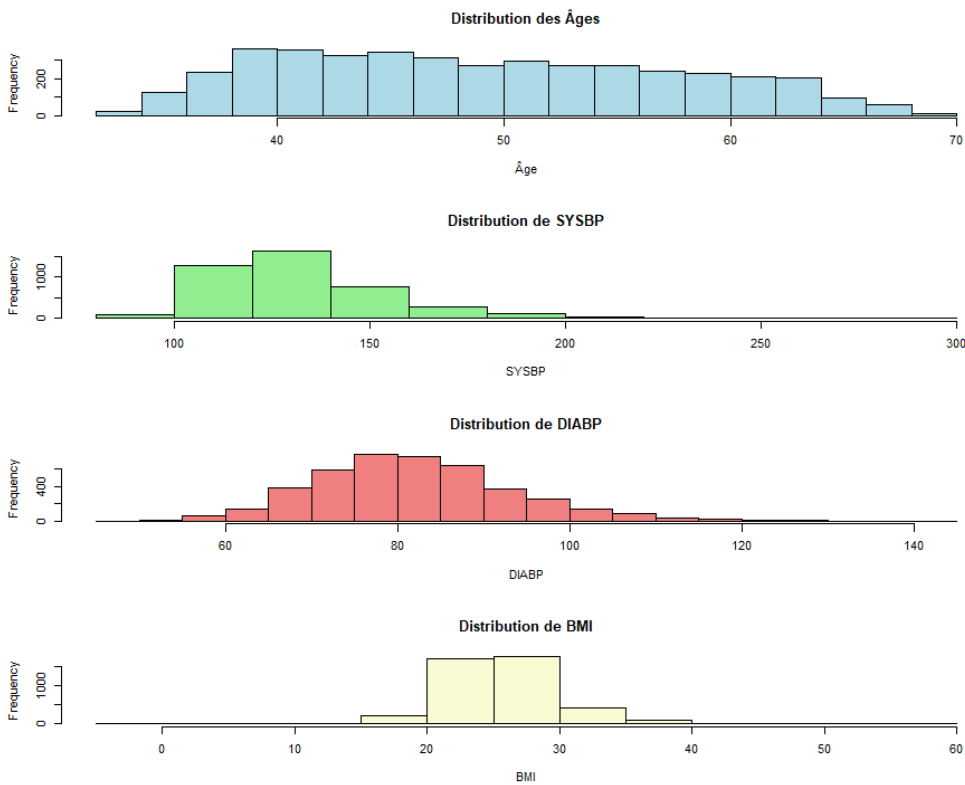
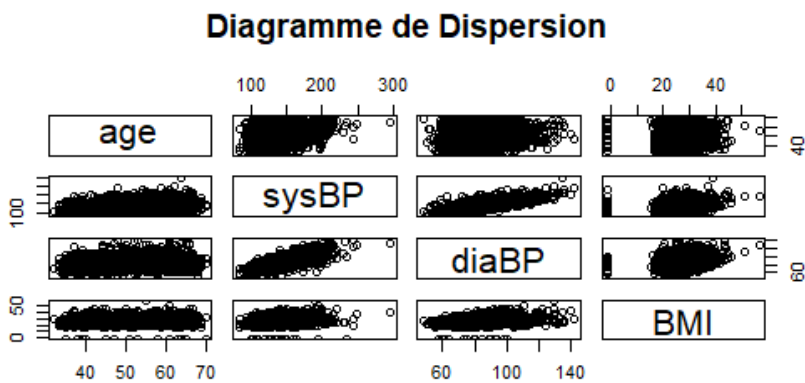
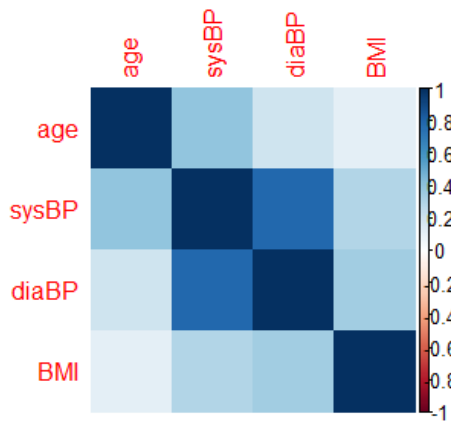


Diagramme de dispersion



Matrice de corrélation



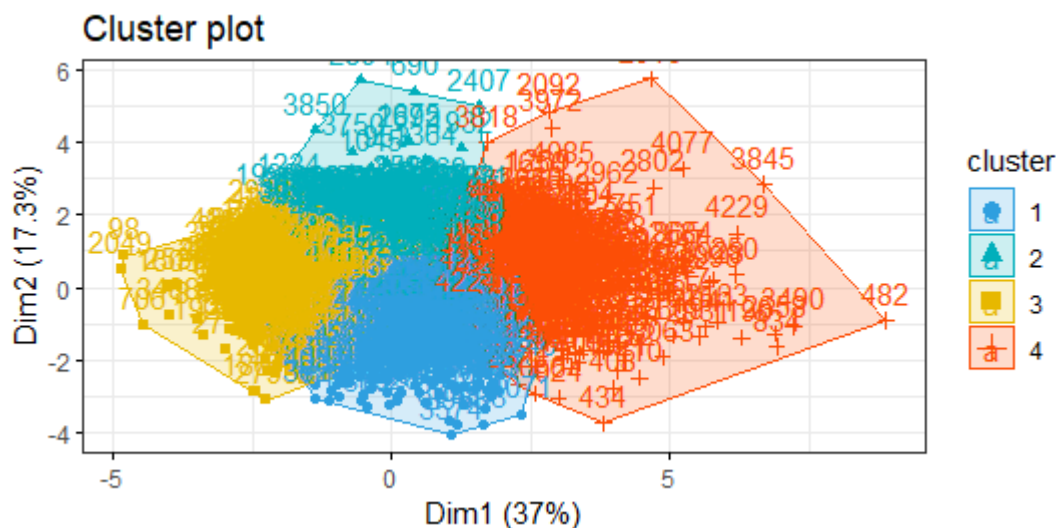
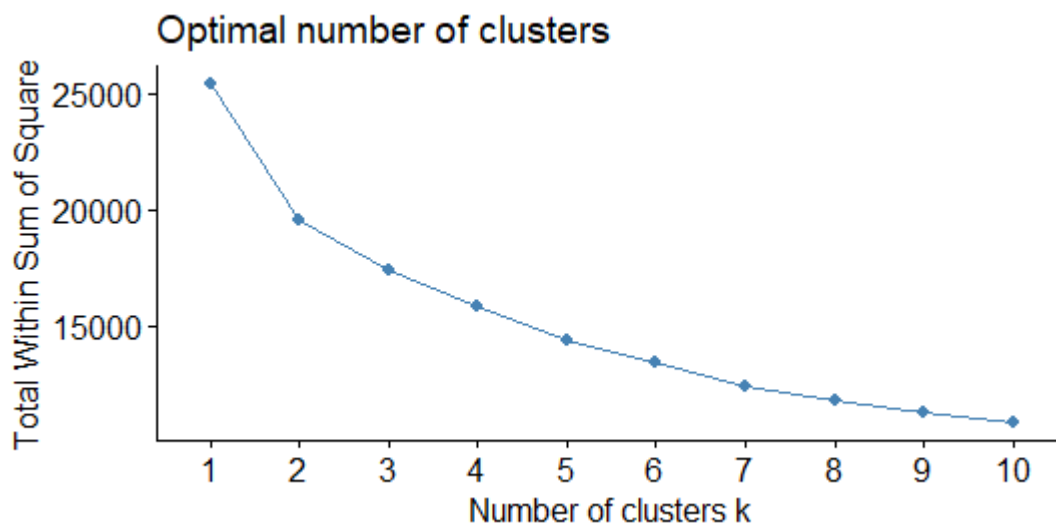
3/Préparation des données:

- **Nettoyage des données** : Identifier et traiter les valeurs manquantes, les valeurs aberrantes ...
- **Préparation des données** : Normalisation, encodage des variables ...

- Remplacer les valeurs manquantes MÉDIANE sinon **REEMPLACER PAR KNN**
- Convertir les variables binaires en facteurs

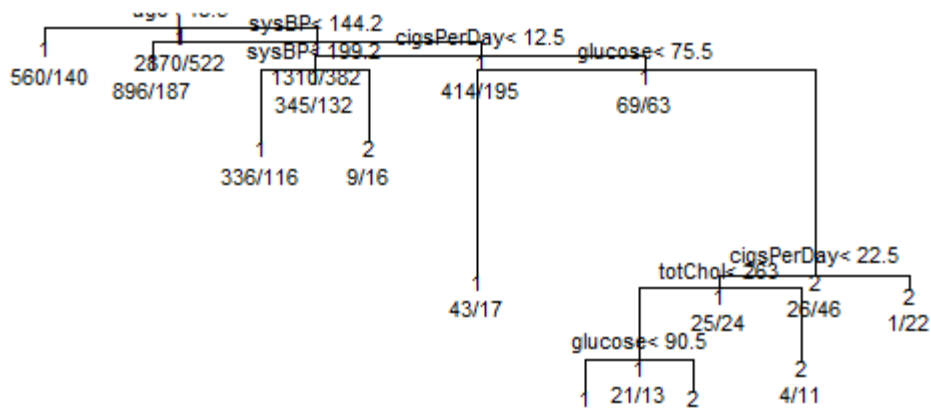
ACP : si je veux comprendre le comportement de chaque variable : il joue le role de modelisation

ACP juste comprehension : visualiser les données



4/Modélisation:

Arbre de décision:



```
> conf_matrix_tree <- confusionMatrix(pred_tree,
> conf_matrix_tree
Confusion Matrix and Statistics
```

```

      Reference
Prediction  0   1
      0  691 123
      1   25   9

      Accuracy : 0.8255
      95% CI   : (0.7982, 0.8504)
      No Information Rate : 0.8443
      P-Value [Acc > NIR] : 0.9392

      Kappa : 0.0477

      Mcnemar's Test P-Value : 1.544e-15

      Sensitivity : 0.96508
      Specificity : 0.06818
      Pos Pred Value : 0.84889
      Neg Pred Value : 0.26471
      Prevalence : 0.84434
      Detection Rate : 0.81486
      Detection Prevalence : 0.95991
      Balanced Accuracy : 0.51663

      'Positive' Class : 0

```

```
> |
```

5/Evaluation: tester l'algorithme

```
> conf_matrix_tree
```

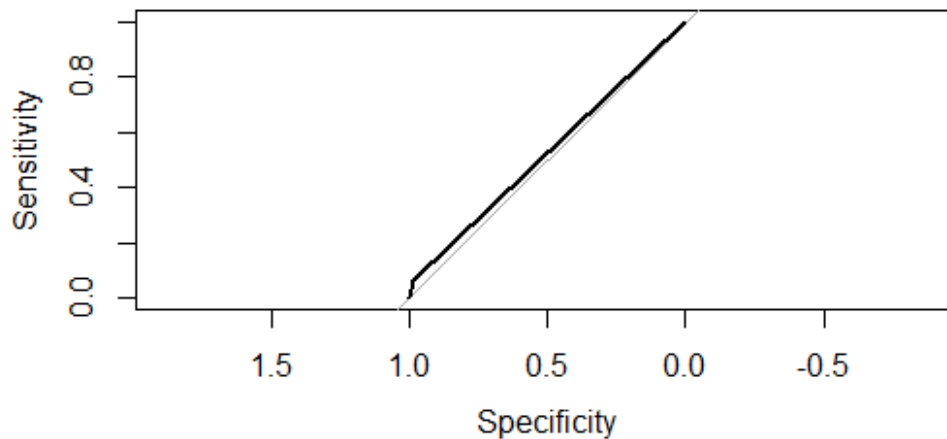
```

pred_tree  1   2
      1  714 114
      2   12   8

```

```
> |
```

Courbe ROC



```
> auc(roc_curve)
Area under the curve: 0.5245
> |
```

```
> summary(modele_lineaire)
```

Call:

```
lm(formula = TenYearCHD ~ age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol +
    sysBP + diaBP + BMI + heartRate + glucose, data = train_data)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-0.63169 -0.18429 -0.10226 -0.02172  1.05740
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.845e-01	8.757e-02	5.533	3.43e-08	***
age	7.425e-03	8.617e-04	8.617	< 2e-16	***
education	-9.402e-04	6.372e-03	-0.148	0.88270	
currentSmoker	6.880e-03	2.026e-02	0.340	0.73425	
cigsPerDay	2.637e-03	8.524e-04	3.094	0.00199	**
BPMeds1	5.352e-02	3.940e-02	1.358	0.17444	
prevalentStroke1	1.839e-01	8.670e-02	2.121	0.03400	*
prevalentHyp1	3.742e-02	1.949e-02	1.920	0.05497	.
diabetes1	8.034e-03	4.946e-02	0.162	0.87097	
totChol	9.012e-05	1.507e-04	0.598	0.54981	
sysBP	2.198e-03	5.497e-04	3.999	6.51e-05	***
diaBP	-1.098e-03	9.021e-04	-1.217	0.22358	
BMI	1.742e-03	1.733e-03	1.006	0.31466	
heartRate	-1.021e-03	5.461e-04	-1.870	0.06158	.
glucose	8.529e-04	3.444e-04	2.477	0.01331	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3412 on 2910 degrees of freedom

Multiple R-squared: 0.09197, Adjusted R-squared: 0.0876

F-statistic: 21.05 on 14 and 2910 DF, p-value: < 2.2e-16