Objectif: Exploration de données et apprentissage de modèles supervisés avec RShiny.

Ce Document présente les résultats de nos analyses réalisé sur la base de données « breast-cancer-wisconsin ».

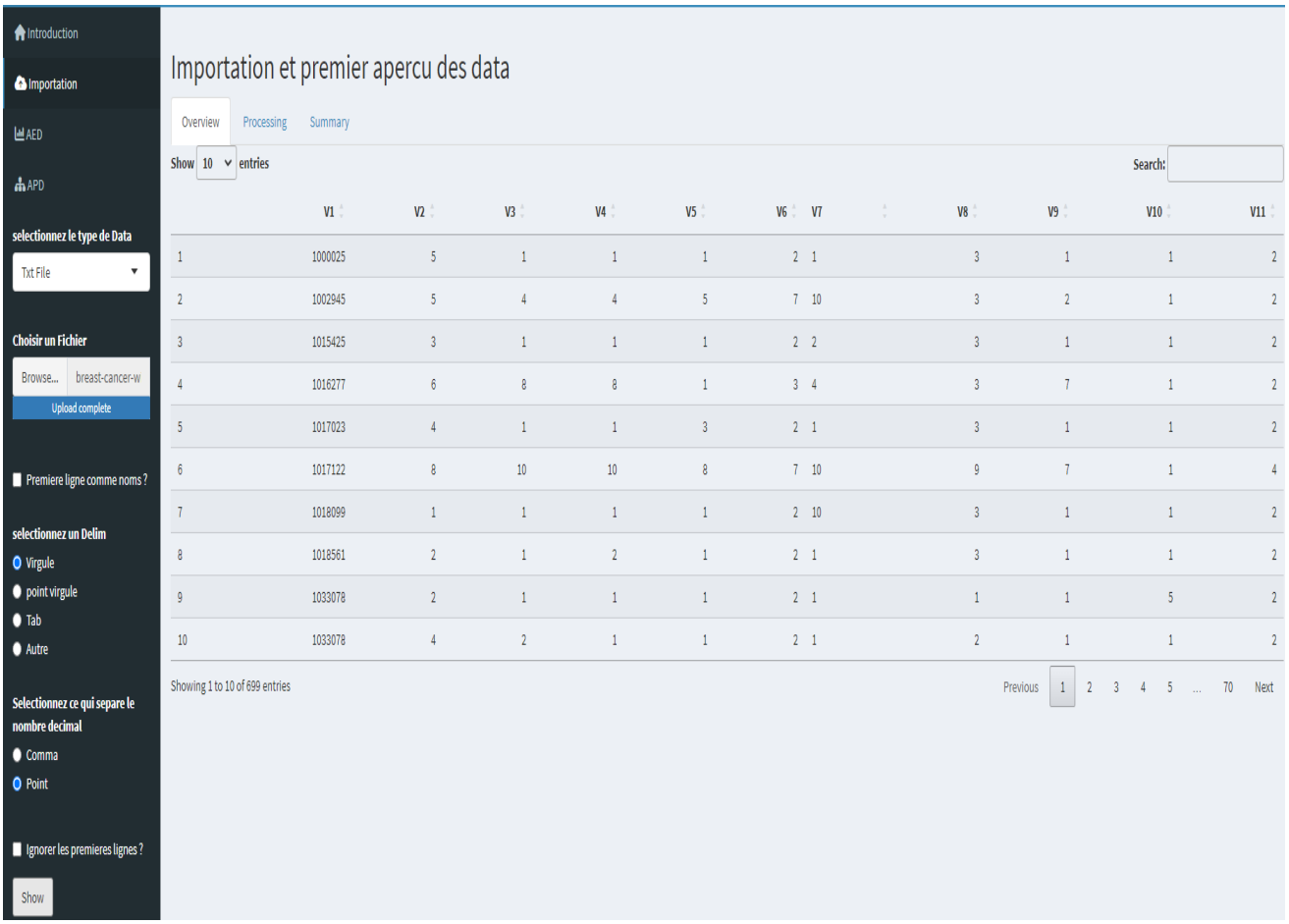
Cette base de données sur le cancer du sein a été obtenues auprès des hôpitaux de l'Université du Wisconsin, à Madison, par le Dr William H. Wolberg, du Wisconsin, Madison, auprès du Dr. William H. Wolberg.

Ce Projet est réalisé par :

- **Mohamed El Ayeb**
- **Ayoub Kabli**
- **Wassim Ouni.**

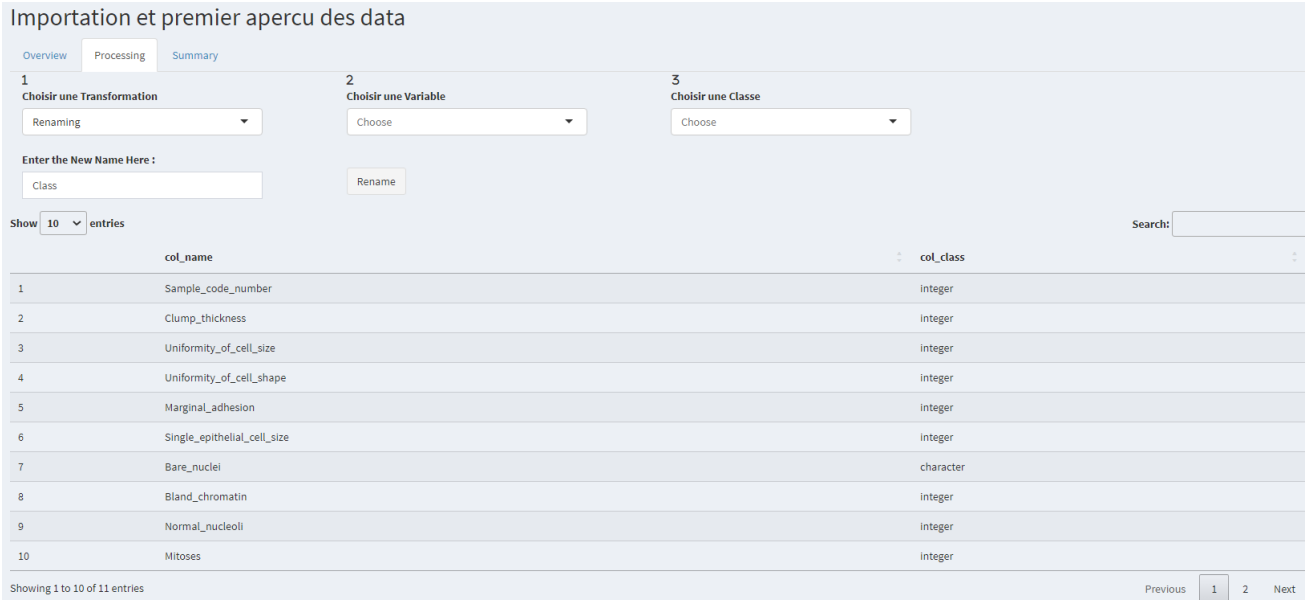
1. Importation des données

En Premier temps on a importé la base de donnée en utilisant notre Application Web :

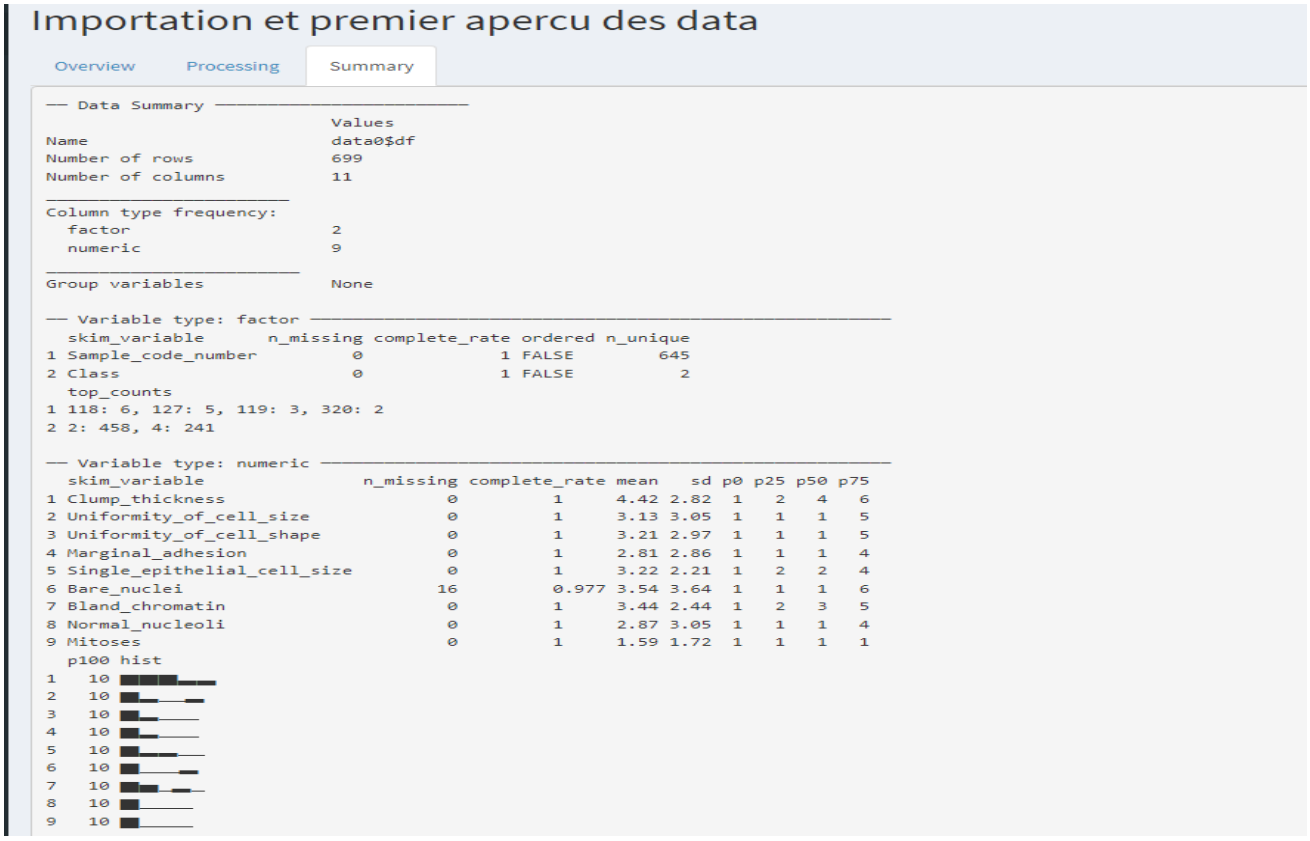


2. Préprocessing des données :

On a commencé par des opérations de Processing simples pour rendre la base plus lisible comme la conversions des variables aux types approprié et renommer les variables par les noms appropriés trouvé sur le site https://archive-beta.ics.uci.edu :



Après on fait un premier aperçu sur les statistiques descriptives :



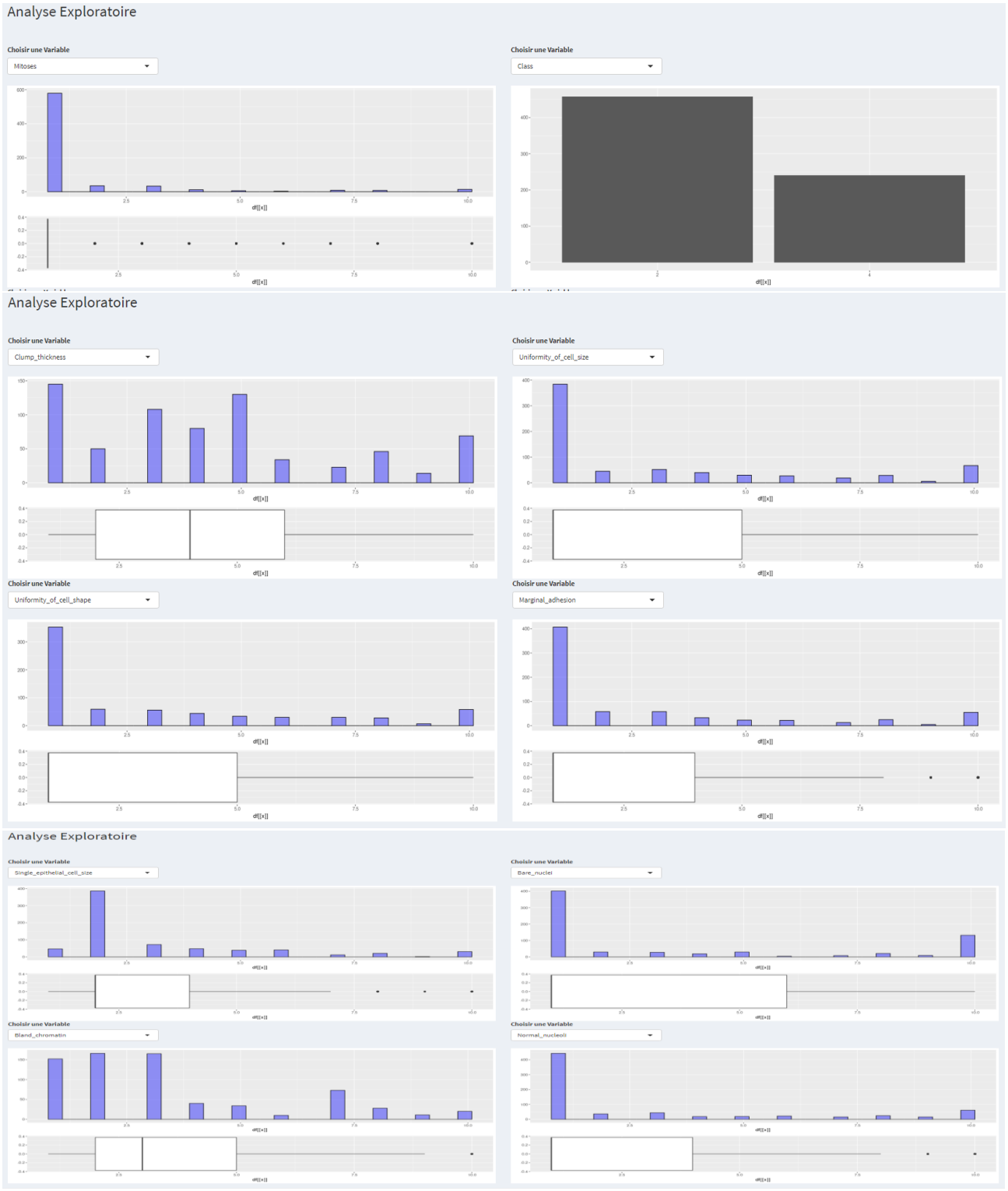
Remarques :

- D'après l'output, On a 699 observation et 11 variables (2 de types facteur et 9 numérique)
- La variable qu'on va prédire à deux modalités (2 :458, 4 :241) et on peut remarquer qu'on a un problème de déséquilibre des classes, on doit le résoudre pour obtenir un meilleur modèle de prédiction par une méthode de ré-échantillonnage (upsamplpin, downsampling, SMOTE, …)
- On peut aussi remarquer qu'on a des valeurs manquantes pour la variable « Bar_nuclei » on peut résoudre ce problème soit par supprimer les observations avec les valeurs manquantes, mais voyant que le nombre d'observations est limité ça sera mieux d'imputer les valeurs manquantes par le moyen ou la médiane de la variable « Bar_nuclei »

3. Analyse exploratoire :

3.1 Analyse uni-variée

Maintenant on passe au Data Visualisation et on commence par faire une analyse Uni-variée des variables :



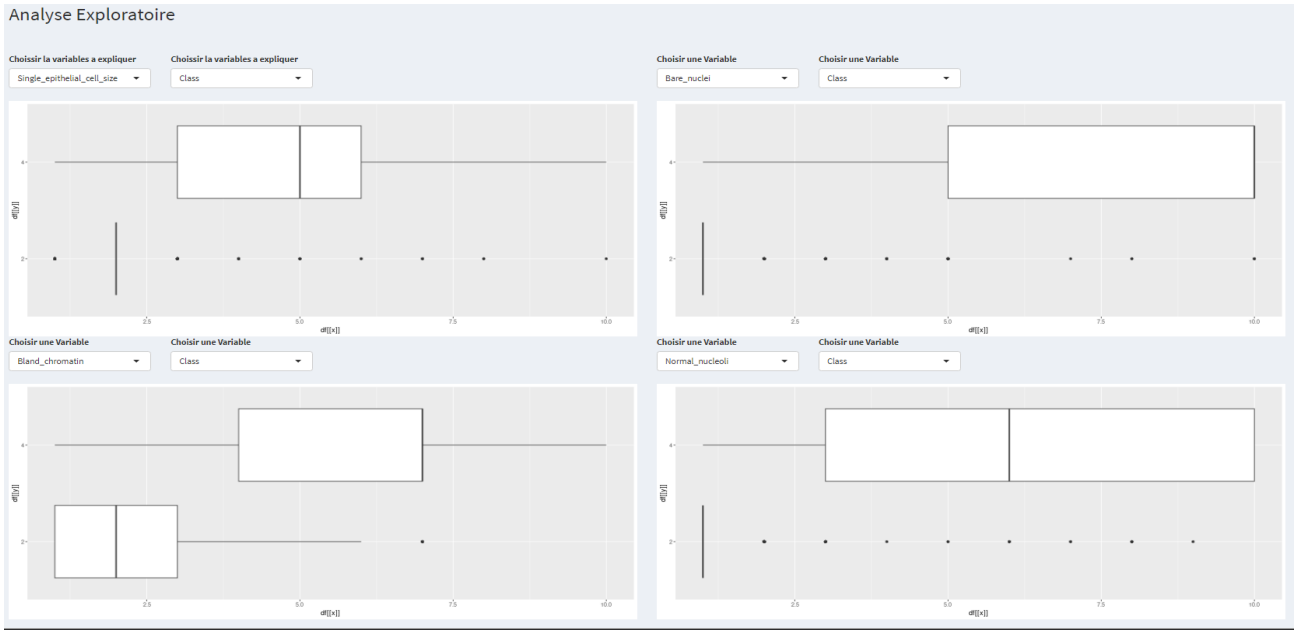
D'après les graphes, on peut remarquer l'existence des outliers pour certaines variables comme « Normal_nucleoli », « Single_epithelial_cell_size », « Marginal_adhesion » et « Mitoses »

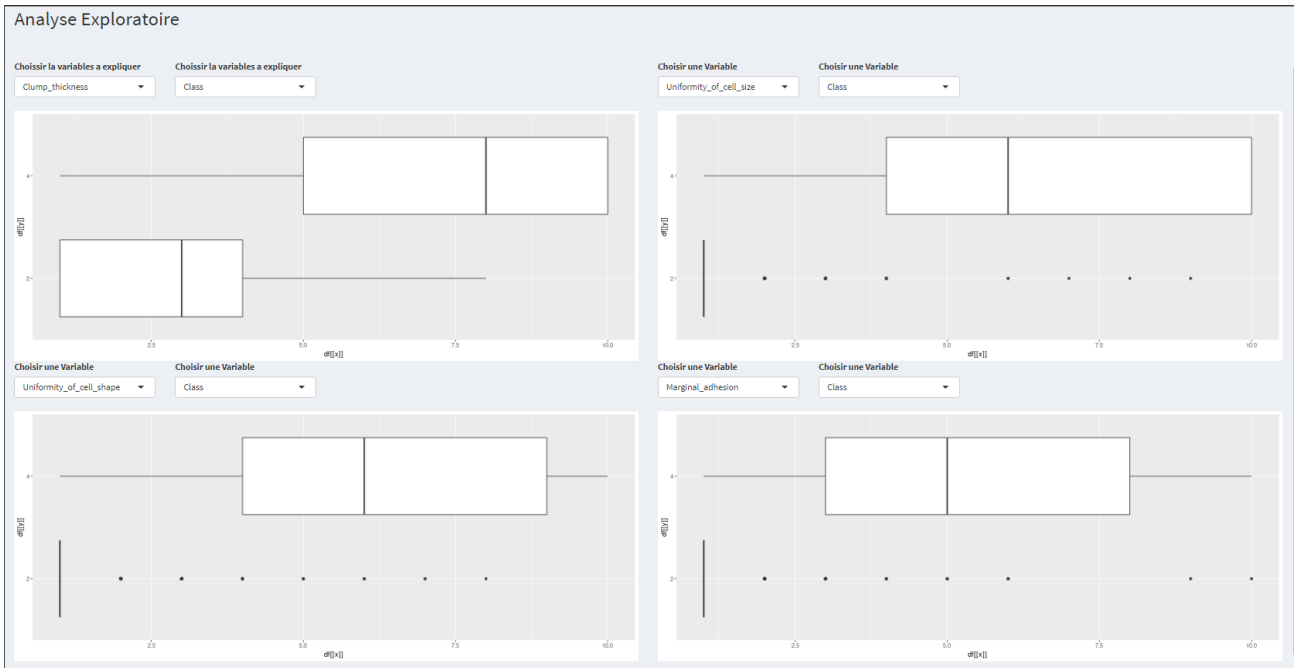
On peut résoudre ce problème par soit éliminer ces observations soit appliquer une transformation comme Log ou BoxCox pour ne pas perdre de l'information.

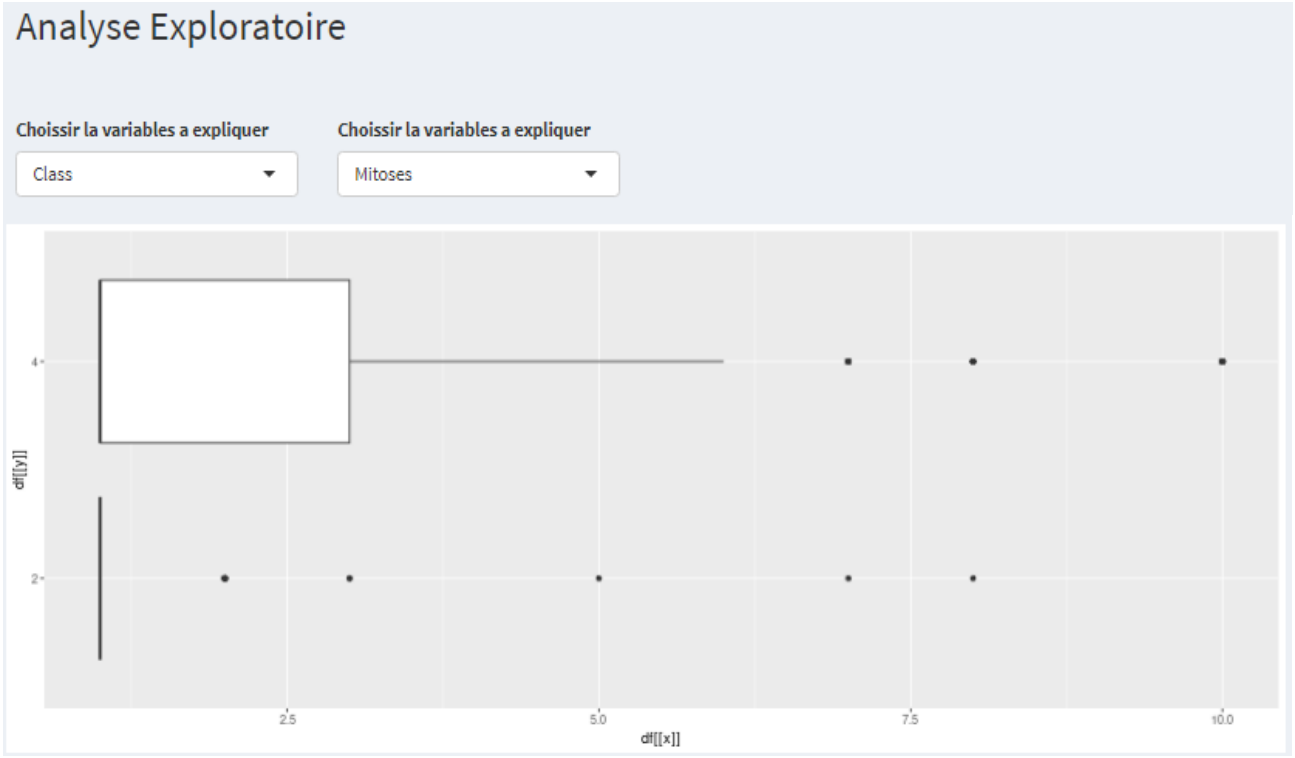
On peut aussi confirmer l'existence d'un déséquilibre des classes pour la variable « Class »

3.2 Analyse Bi-variée :

On passe maintenant à l'analyse Bi-varié :







Remarque :

- D'après les Graph on peut remarquer qu'il y a une différence significative entre la médiane des deux classes de la variable « Class » pour tous les variables explicatives sauf la variable « Mitoses », on peut remarquer que la médiane des deux classes est à peu près égaux et ça sera intéressant de voir son rang dans la courbe de l'importance des variables.

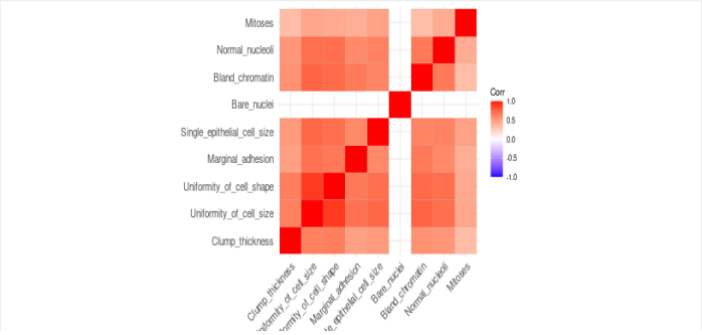
3.3 L'analyse de corrélation :

Analyse Exploratoire

Show 10 entries Search:

	Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses
Clump_thickness	1	0.6449125043512	0.654589080001924	0.486356243676702	0.521816219959653		0.558426162285396	0.535834549212977	0.350033856485964
Uniformity_of_cell_size	0.6449125043512	1	0.906881913052595	0.705581811557112	0.751799129877131		0.755720981100574	0.722864821906358	0.45869314741651
Uniformity_of_cell_shape	0.654589080001924	0.906881913052595	1	0.683079200230476	0.719668437170359		0.735948454023298	0.719446316953281	0.438910928928209
Marginal_adhesion	0.486356243676702	0.705581811557112	0.683079200230476	1	0.599599068425499		0.666715326264053	0.603352412216761	0.417632780056886
Single_epithelial_cell_size	0.521816219959653	0.751799129877131	0.719668437170359	0.599599068425499	1		0.616101840871849	0.628880685589092	0.479101477034748
Bare_nuclei						1			
Bland_chromatin	0.558426162285396	0.755720981100574	0.735948454023298	0.666715326264053	0.616101840871849		1	0.665877809425439	0.344169496265853
Normal_nucleoli	0.535834549212977	0.722864821906358	0.719446316953281	0.603352412216761	0.628880685589092		0.665877809425439	1	0.428335749353575
Mitoses	0.350033856485964	0.45869314741651	0.438910928928209	0.417632780056886	0.479101477034748		0.344169496265853	0.428335749353575	1

Showing 1 to 9 of 9 entries Previous 1 Next



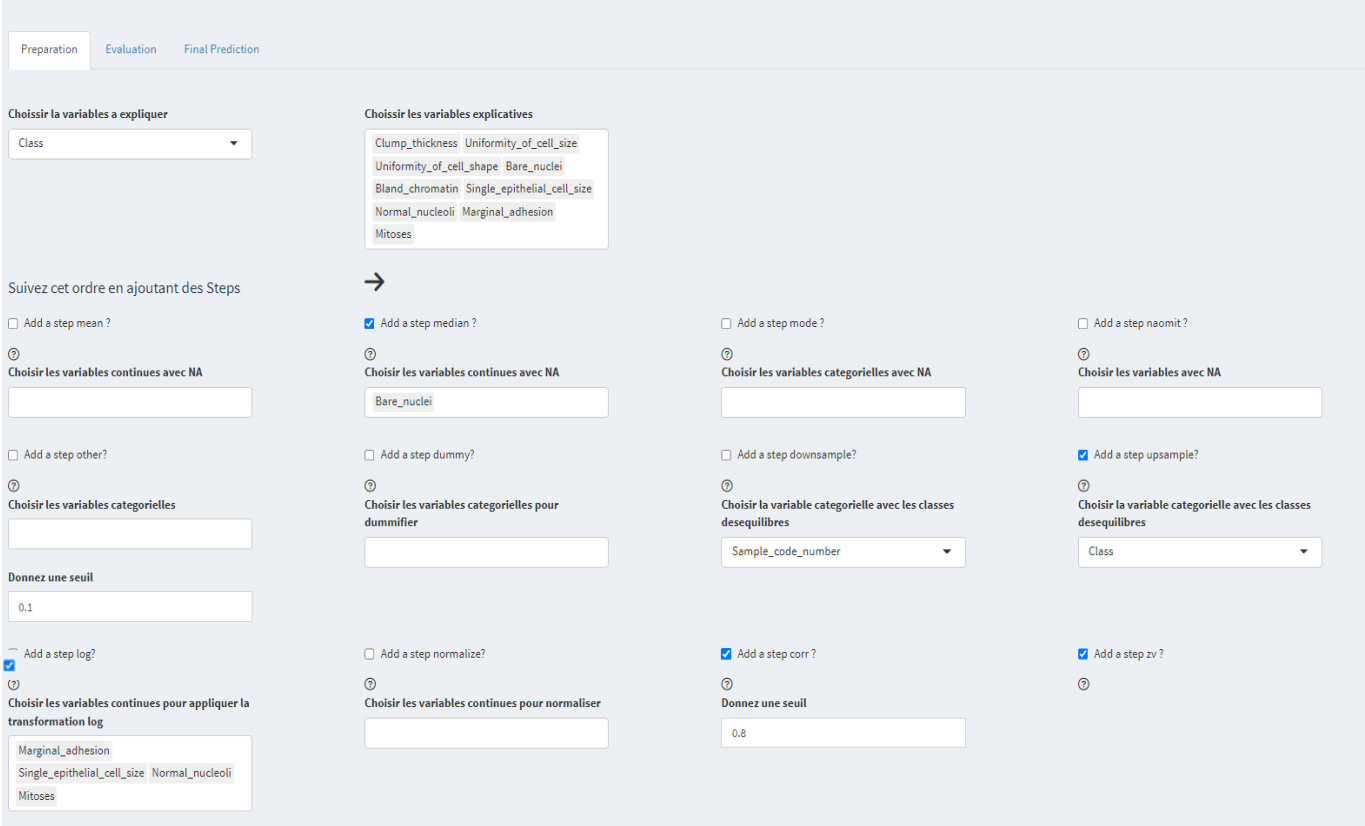
Remarques :

- D'après le tableau de corrélation et le Heatmap, On remarque qu'il y a une forte corrélation entre les variables « Uniformity_of_cell_shape » et « Uniformity_of_cell_size » qui est égale à 0.90688,
- Ça peut présenter un problème d'inclure deux variables corrélé entre eux dans notre modèle comme ils présentent la même information, ça sera mieux d'éliminer une entre eux.

Sinon, les autres variables sont moyennement corrélées entre eux et on peut passer à l'analyse prédictive.

4. Analyse prédictive :

Pour cette Partie on va commencer par choisir les variables à expliquer et les variables explicatives et on va choisir les Transformations (Steps) à appliquer :



Etapes :

1. On a utilisé Step_median pour imputer la variable « Bare_nuclei ».
2. On a utilisé Step_upsample pour résoudre le problème de déséquilibre des classes pour la variable « Class ».
3. On a utilisé Step_log pour appliquer la transformation Logarithme aux variables « Normal_nucleoli », « Single_epithelial_cell_size », « Marginal_adhesion » et « Mitoses » qui ont présentent des outliers.
4. On a utilisé Step_corr pour eliminer une des variables qui présentes une forte corrélation entre eux .
5. On a utilisé Step_zv pour éliminer les variables qui ont 0 variances.

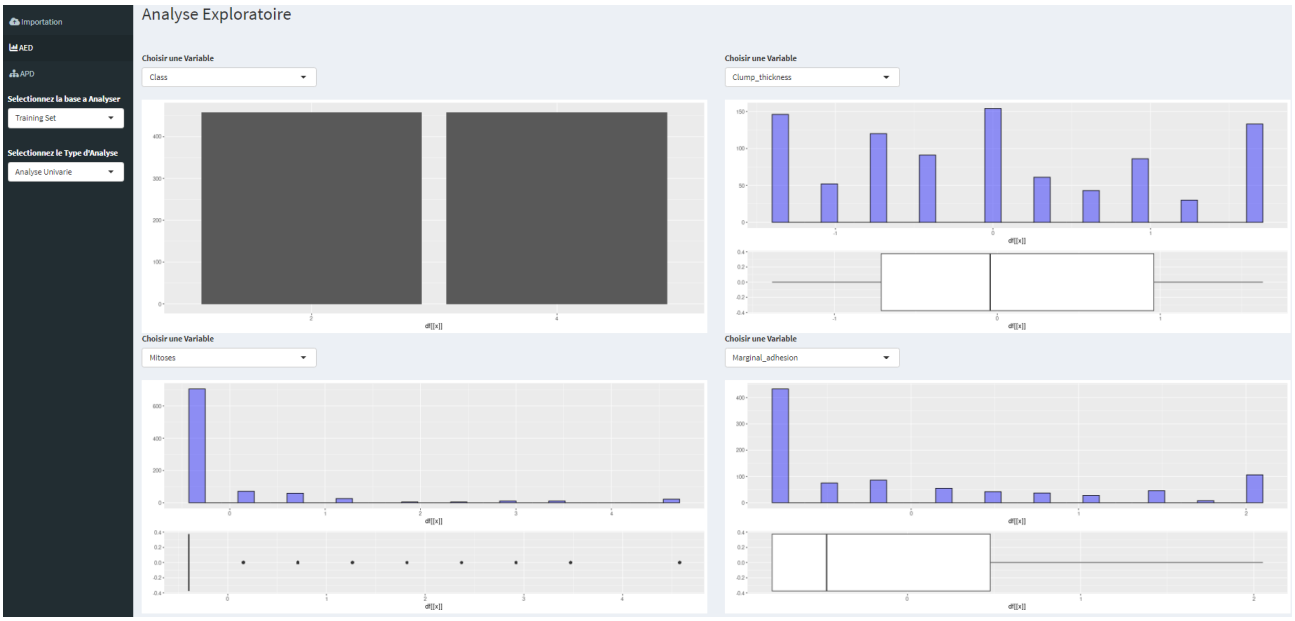
```
Recipe

Inputs:

      role #variables
   outcome         1
 predictor         9

Training data contained 699 data points and 16 incomplete rows.

Operations:

Median imputation for Bare_nuclei [trained]
$terms
<list_of<quosure>>

[[1]]
<quosure>
expr: ^Class
env:  0x558d9ab26b20


$over_ratio
[1] 1

$ratio
[1] NA

$role
[1] NA

$trained
[1] TRUE

$column
[1] "Class"

$target
[1] 458

$skip
[1] TRUE

$id
[1] "upsample_hDytO"

$seed
[1] 53311

attr(,"class")
[1] "step_upsample" "step"
Log transformation on Marginal_adhesion, Single_epithelial_cell_size... [trained]
Correlation filter on Uniformity_of_cell_size [trained]
Zero variance filter removed <none> [trained]
```

Voici la base de donnée après la transformation :

Show 10 entries Search:

	Clump_thickness	Uniformity_of_cell_shape	Bare_nuclei	Bland_chromatin	Single_epithelial_cell_size	Normal_nucleoli	Marginal_adhesion	Mitoses	Class
1	5	1	1	3	0.693147180559945	0	0	0	2
2	5	4	10	3	1.94591014905531	0.693147180559945	1.6094379124341	0	2
3	3	1	2	3	0.693147180559945	0	0	0	2
4	6	8	4	3	1.09861228866811	1.94591014905531	0	0	2
5	4	1	1	3	0.693147180559945	0	1.09861228866811	0	2
6	1	1	10	3	0.693147180559945	0	0	0	2
7	2	2	1	3	0.693147180559945	0	0	0	2
8	2	1	1	1	0.693147180559945	0	0	1.6094379124341	2
9	4	1	1	2	0.693147180559945	0	0	0	2
10	1	1	1	3	0	0	0	0	2

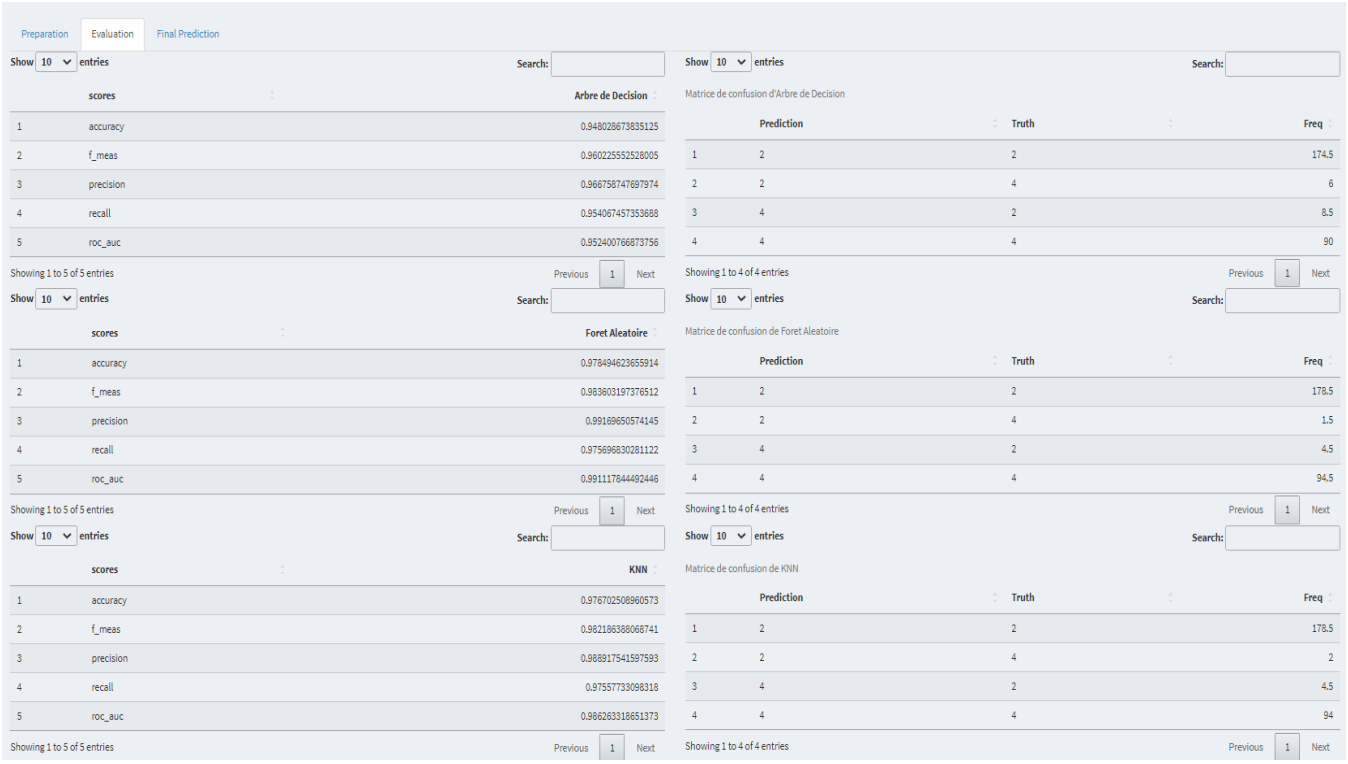
Showing 1 to 10 of 916 entries Previous 1 2 3 4 5 ... 92 Next



Remarque :

- Les classes sont équilibrées maintenant et la transformation Log a résolu le problème des outliers sauf pour la variable « Mitoses », ça sera intéressant de voir la performance du modèle sans cette variable.

Après On a appliqué 3 modèle pour faire la classification (le foret aléatoire, le KNN et l'arbre de décision) on a aussi entrainé nos modèles avec 5-fold CV pour éviter l'Overfitting et obtenir des résultats proches des résultats qu'on va trouver sur l'ensemble de Test :



Preparation | Evaluation | Final Prediction

Show 10 entries — Search:

	scores	Arbre de Decision
1	accuracy	0.948028673835125
2	f_meas	0.960225552528005
3	precision	0.966758747697974
4	recall	0.954067457353688
5	roc_auc	0.952400766873756

Showing 1 to 5 of 5 entries — Previous 1 Next

Show 10 entries — Search:

	scores	Foret Aleatoire
1	accuracy	0.978494623655914
2	f_meas	0.983603197376512
3	precision	0.99169650574145
4	recall	0.975696830281122
5	roc_auc	0.991117844492446

Showing 1 to 5 of 5 entries — Previous 1 Next

Show 10 entries — Search:

	scores	KNN
1	accuracy	0.976702508960573
2	f_meas	0.982186388068741
3	precision	0.988917541597593
4	recall	0.97557733098318
5	roc_auc	0.986263318651373

Showing 1 to 5 of 5 entries — Previous 1 Next

Show 10 entries — Search:

Matrice de confusion d'Arbre de Decision

	Prediction	Truth	Freq
1	2	2	174.5
2	2	4	6
3	4	2	8.5
4	4	4	90

Showing 1 to 4 of 4 entries — Previous 1 Next

Show 10 entries — Search:

Matrice de confusion de Foret Aleatoire

	Prediction	Truth	Freq
1	2	2	178.5
2	2	4	1.5
3	4	2	4.5
4	4	4	94.5

Showing 1 to 4 of 4 entries — Previous 1 Next

Show 10 entries — Search:

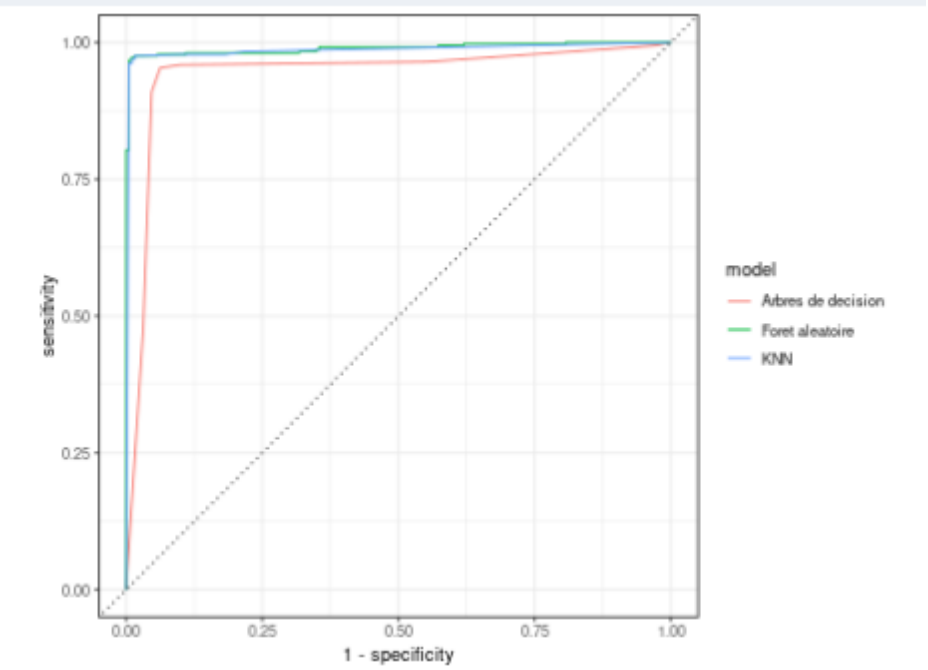
Matrice de confusion de KNN

	Prediction	Truth	Freq
1	2	2	178.5
2	2	4	2
3	4	2	4.5
4	4	4	94

Showing 1 to 4 of 4 entries — Previous 1 Next

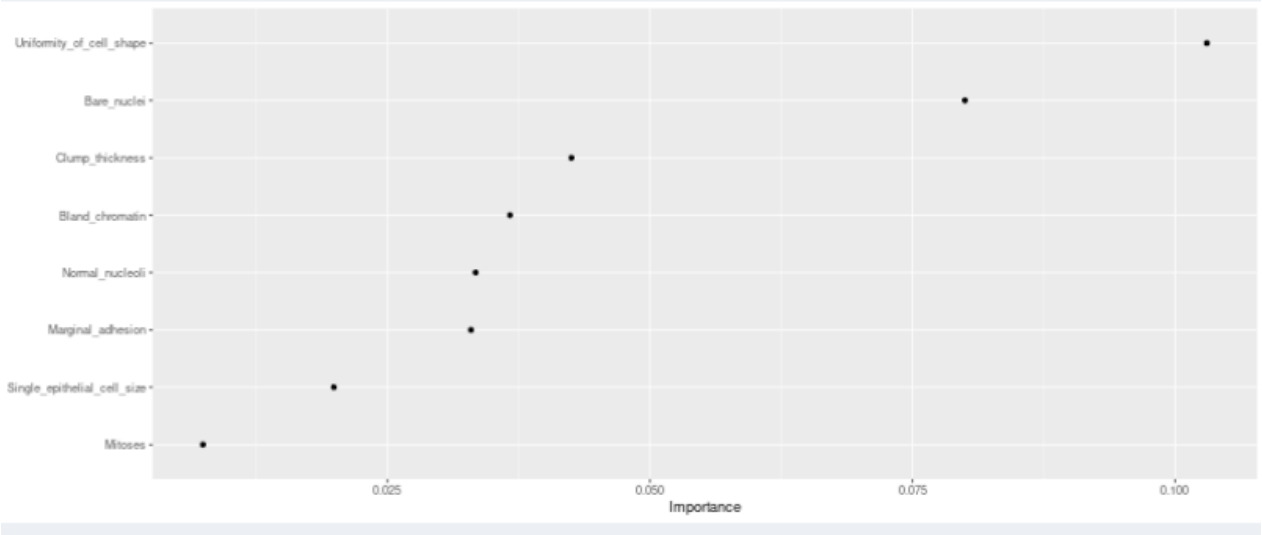
Les 3 modèles ont donné des bons résultats supérieurs à 0.9 pour tous les métriques mais le foret aléatoire est le meilleur avec un accuracy = 0.978, un recall = 0.975 et roc_auc = 0.991.

D'après les matrices de confusion, On peut remarquer que le foret aléatoire et le KNN ont une meilleure tendance à prédire la classe « 2 » par rapport au arbre de décision, par contre le foret aléatoire est le meilleur modèle pour prédire la classe « 4 ».



Le Graph des courbe ROC confirme les résultats, l'arbre de décision est le modèle le moins performant et la performance de KNN et la foret aléatoire est très proche.

On passe maintenant aux analyse de l'importance des variables :



On remarque que la variable « Uniformity_of_cell_size » est le plus important suivi par la variable « Bare_nuclei ».

La courbe confirme notre observation que la variable qui apport le moins d'information à ce modèle prédictive est la variable « Mitoses ».