



VISUALISATION DES DONNEES DU DIAGNOSTIC DE CANCER WISCOSIN WDBC

LICENCE 3 INFORMATIQUE
2019-2020

Sylia RAHMANI
11707524

PLAN



INTRODUCTION



Visualisation des
données



Analyse et
interprétation des
données

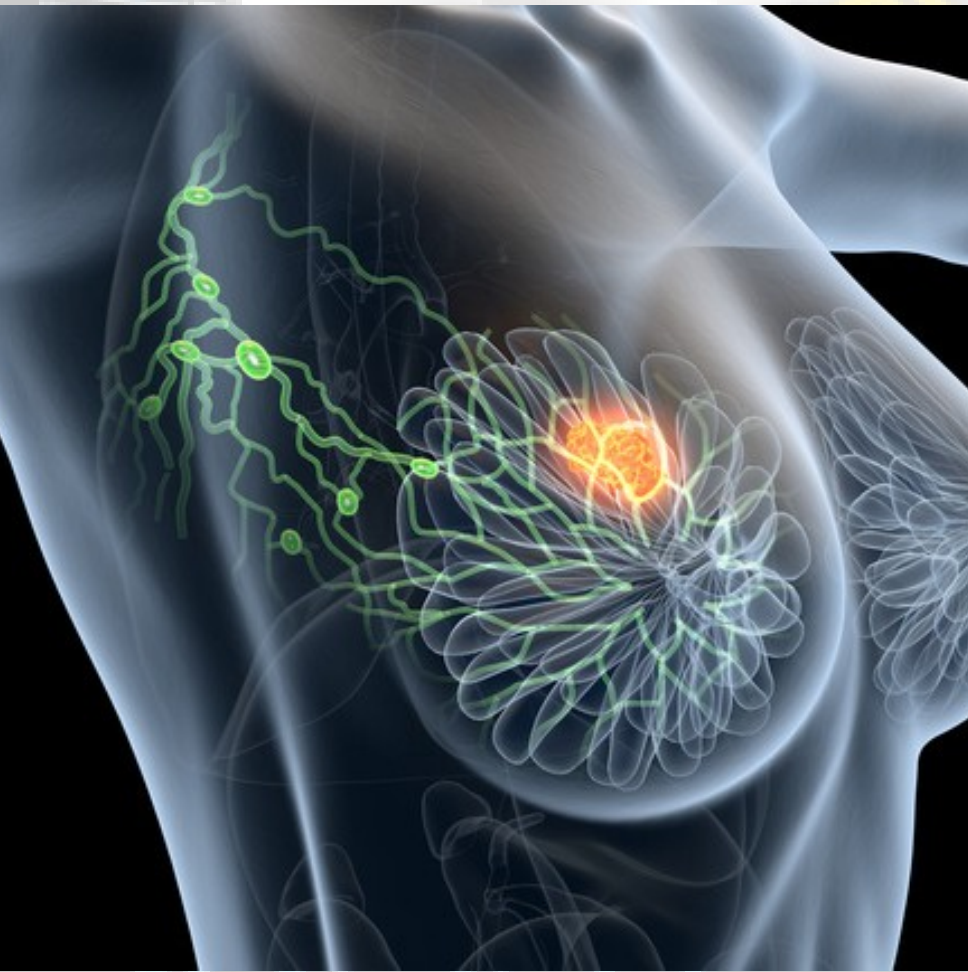


conclusion

INTRODUCTION

Open for Innovation

Le cancer du sein est une tumeur maligne se développant à partir des cellules constituant la glande mammaire. Une cellule, initialement normale, dont les gènes ont subi des modifications pour diverses raisons, se développe alors de façon anarchique, conduisant à l'apparition d'une tumeur. Afin de mieux comprendre cette maladie, nous allons explorer les données WDBC avec Knime qui est un outil de data préparation et de data science, ce qui nous permettrait d'en tirer rapidement des informations grâce aux représentations graphiques, les analyser et conclure avec des prédictions.



Row ID	S Col1	D Col2	D Col3	D Col4	D Col5	D Col6	D Col7	D Col8	D Col9	D Col10	D Col11	D Col12	D Col13	D Col14
Row0	M	17.99	10.38	122.8	1,001	0.118	0.278	0.3	0.147	0.242	0.079	1.095	0.905	8.589
Row1	M	20.57	17.77	132.9	1,326	0.085	0.079	0.087	0.07	0.181	0.057	0.543	0.734	3.398
Row2	M	19.69	21.25	130	1,203	0.11	0.16	0.197	0.128	0.207	0.06	0.746	0.787	4.585
Row3	M	11.42	20.38	77.58	386.1	0.142	0.284	0.241	0.105	0.26	0.097	0.496	1.156	3.445
Row4	M	20.29	14.34	135.1	1,297	0.1	0.133	0.198	0.104	0.181	0.059	0.757	0.781	5.438
Row5	M	12.45	15.7	82.57	477.1	0.128	0.17	0.158	0.081	0.209	0.076	0.335	0.89	2.217
Row6	M	18.25	19.98	119.6	1,040	0.095	0.109	0.113	0.074	0.179	0.057	0.447	0.773	3.18
Row7	M	13.71	20.83	90.2	577.9	0.119	0.165	0.094	0.06	0.22	0.075	0.584	1.377	3.856
Row8	M	13	21.82	87.5	519.8	0.127	0.193	0.186	0.094	0.235	0.074	0.306	1.002	2.406
Row9	M	12.46	24.04	83.97	475.9	0.119	0.24	0.227	0.085	0.203	0.082	0.298	1.599	2.039
Row10	M	16.02	23.24	102.7	797.8	0.082	0.067	0.033	0.033	0.153	0.057	0.38	1.187	2.466
Row11	M	15.78	17.89	103.6	781	0.097	0.129	0.1	0.066	0.184	0.061	0.506	0.985	3.564
Row12	M	19.17	24.8	132.4	1,123	0.097	0.246	0.206	0.112	0.24	0.078	0.956	3.568	11.07
Row13	M	15.85	23.95	103.7	782.7	0.084	0.1	0.099	0.054	0.185	0.053	0.403	1.078	2.903
Row14	M	13.73	22.61	93.6	578.3	0.113	0.229	0.213	0.08	0.207	0.077	0.212	1.169	2.061
Row15	M	14.54	27.54	96.73	658.8	0.114	0.16	0.164	0.074	0.23	0.071	0.37	1.033	2.879
Row16	M	14.68	20.13	94.74	684.5	0.099	0.072	0.074	0.053	0.159	0.059	0.473	1.24	3.195
Row17	M	16.13	20.68	108.1	798.8	0.117	0.202	0.172	0.103	0.216	0.074	0.569	1.073	3.854
Row18	M	19.81	22.15	130	1,260	0.098	0.103	0.148	0.095	0.158	0.054	0.758	1.017	5.865
Row19	B	13.54	14.36	87.46	566.3	0.098	0.081	0.067	0.048	0.189	0.058	0.27	0.789	2.058
Row20	B	13.08	15.71	85.63	520	0.107	0.127	0.046	0.031	0.197	0.068	0.185	0.748	1.383
Row21	B	9.504	12.44	60.34	273.9	0.102	0.065	0.03	0.021	0.181	0.069	0.277	0.977	1.909
Row22	M	15.34	14.26	102.5	704.4	0.107	0.213	0.208	0.098	0.252	0.07	0.439	0.71	3.384
Row23	M	21.16	23.04	137.2	1,404	0.094	0.102	0.11	0.086	0.177	0.053	0.692	1.127	4.303
Row24	M	16.65	21.38	110	904.6	0.112	0.146	0.152	0.092	0.2	0.063	0.807	0.902	5.455
Row25	M	17.14	16.4	116	912.7	0.119	0.228	0.223	0.14	0.304	0.074	1.046	0.976	7.276
Row26	M	14.58	21.53	97.41	644.8	0.105	0.187	0.142	0.088	0.225	0.069	0.255	0.983	2.11
Row27	M	18.61	20.25	122.1	1,094	0.094	0.107	0.149	0.077	0.17	0.057	0.853	1.849	5.632
Row28	M	15.3	25.27	102.4	732.4	0.108	0.17	0.168	0.088	0.193	0.065	0.439	1.012	3.498
Row29	M	17.57	15.05	115	955.1	0.098	0.116	0.099	0.08	0.174	0.061	0.6	0.823	4.655
Row30	M	18.63	25.11	124.8	1,088	0.106	0.189	0.232	0.124	0.218	0.062	0.831	1.466	5.574
Row31	M	11.84	18.7	77.93	440.6	0.111	0.152	0.122	0.052	0.23	0.078	0.482	1.03	3.475
Row32	M	17.02	23.98	112.8	899.3	0.12	0.15	0.242	0.12	0.225	0.064	0.601	1.398	3.999
Row33	M	19.27	26.47	127.9	1,162	0.094	0.172	0.166	0.076	0.185	0.063	0.556	0.606	3.528
Row34	M	16.13	17.88	107	807.2	0.104	0.156	0.135	0.078	0.2	0.065	0.334	0.686	2.183

PRÉSENTATION DES DONNÉES

569 cas.

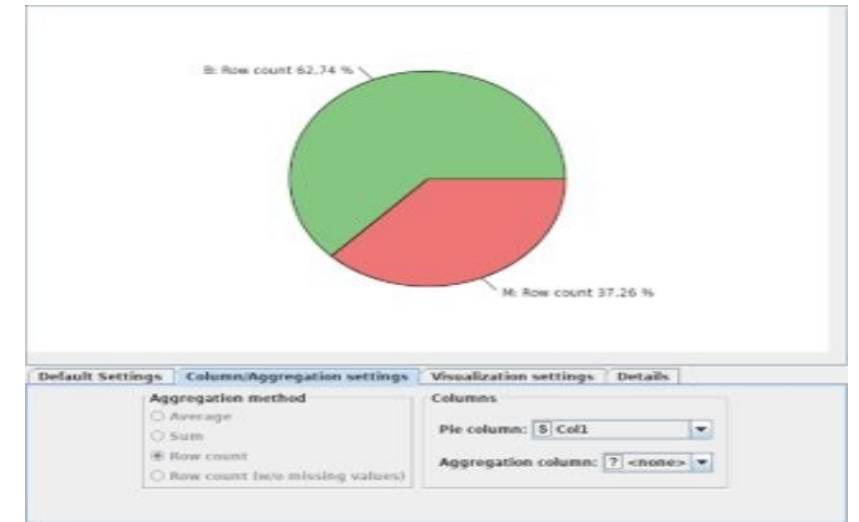
32 attributs:

Id, diagnostic et 30 fonctions à valeur réelles

Deux diagnostics possibles:

Tumeur Maligne: 212

Tumeur Bénigne: 357



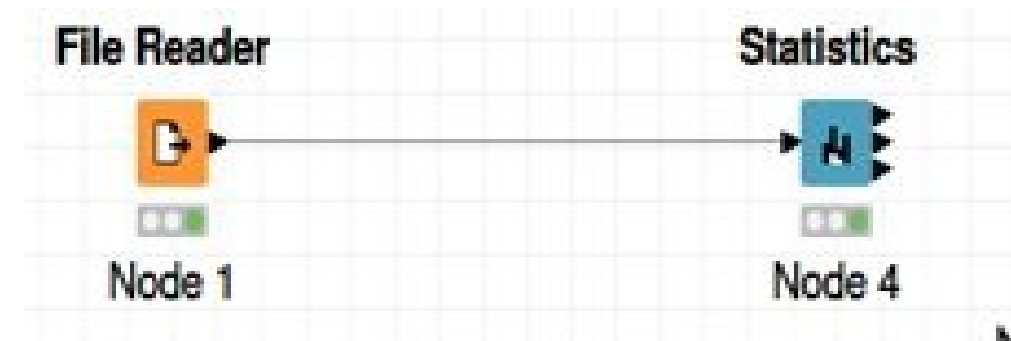
Numéro d'identification		0	
Diagnostic (M= maligne ,B= bénigne)		1	
Rayon	2	12	22
Texture	3	13	23
Périmètre	4	14	24
Zone	5	15	25
Douceur	6	16	26
Compacité	7	17	27
Concavité	8	18	28
Concaves	9	19	29
Symétrie	10	20	30
Dimension fractale	11	21	31


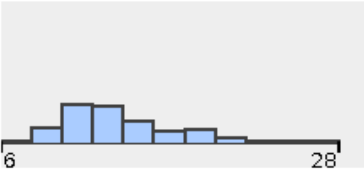
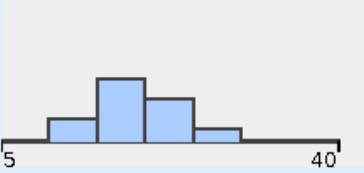
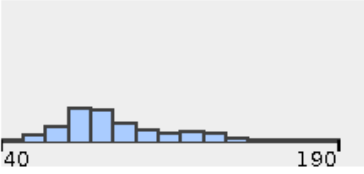
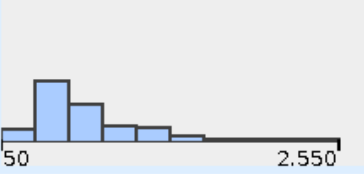
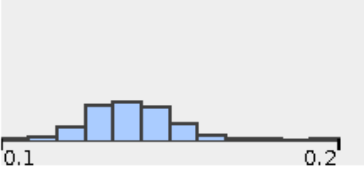
A dark blue, irregular ink splatter shape centered on a white background. The splatter has a rough, textured edge with some lighter blue and white speckles around it.

TRAITEMENT DES DONNÉES AVEC L'OUTIL "KNIME"

VUE STATISTIQUE

Cette technique nous permet de d'avoir un peu plus d'informations sur les données dont on dispose comme l'aplatissement des données et leurs position par rapport à la médiane.

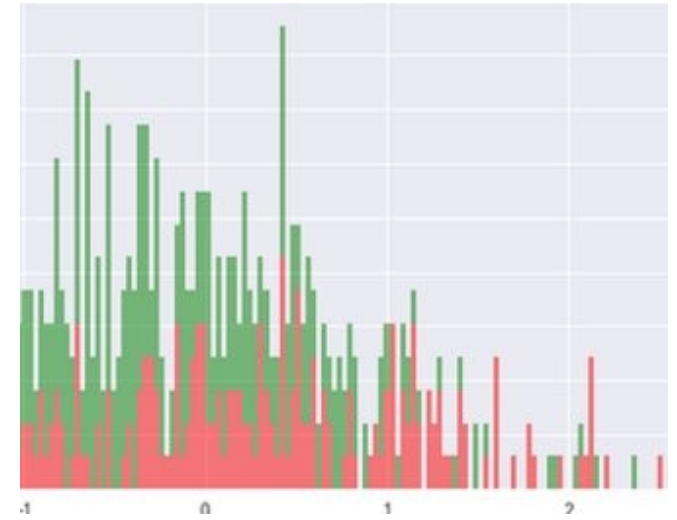


Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Col0	8 670	30 371 831,4323	906 024	9,11E8	1,25E8	6,4738	42,1932	0	0	0	
Col2	6,981	14,1273	13,37	28,11	3,524	0,9424	0,8455	0	0	0	
Col3	9,71	19,2896	18,84	39,28	4,301	0,6504	0,7583	0	0	0	
Col4	43,79	91,969	86,24	188,5	24,299	0,9907	0,9722	0	0	0	
Col5	143,5	654,8891	551,1	2 501	351,9141	1,6457	3,6523	0	0	0	
Col6	0,0526	0,0964	0,0959	0,1634	0,0141	0,4563	0,856	0	0	0	
Col7	0,0194	0,1043	0,0926	0,3454	0,0528	1,1901	1,6501	0	0	0	

Quelques Hypothèses

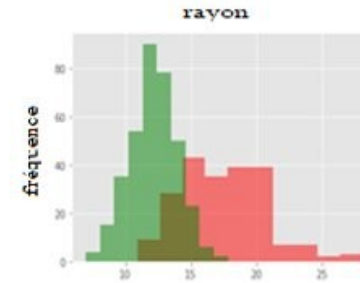
Pas de relation entre
la symétrie et le
critère bénigne ou
maligne de la tumeur

Fréquence

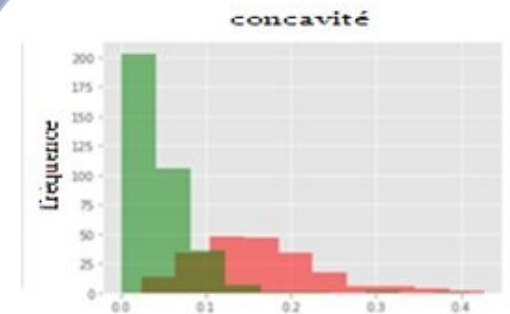


symétrie

Plus le rayon
augmente plus
la tumeur
risque d'être
maligne

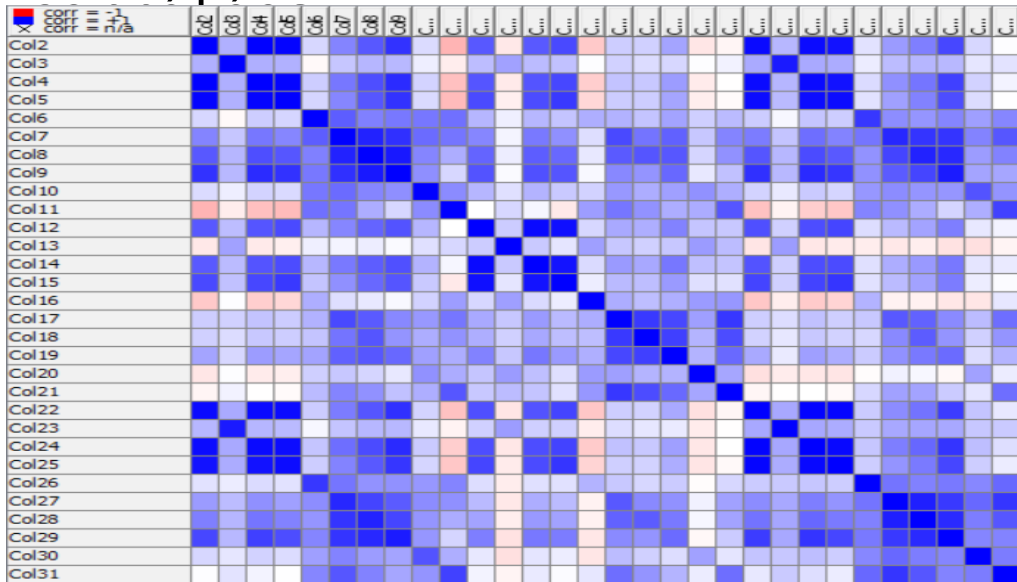


Plus la concavité
augmente plus la
tumeur risque
d'être maligne

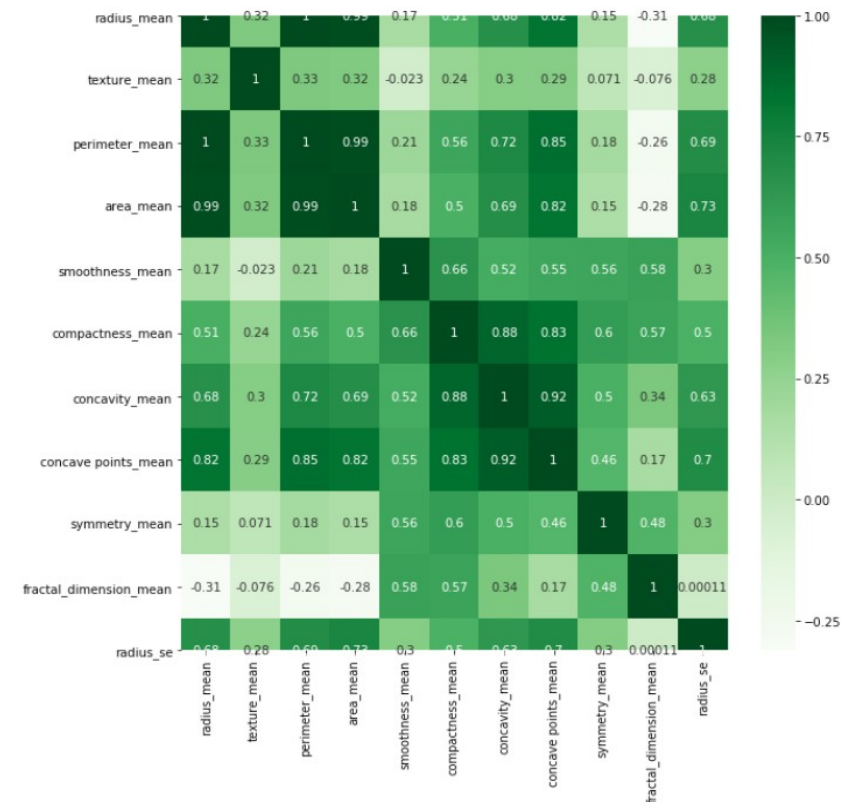


LINEAR CORRELATION


la matrice de corrélation nous sert d'un modèle de Filtrage pour l'outil Correlation Filter, afin d'éliminer les colonnes qui sont très




Matrice de corrélation entre les différentes caractéristiques pour un diagnostic du cancer de sein



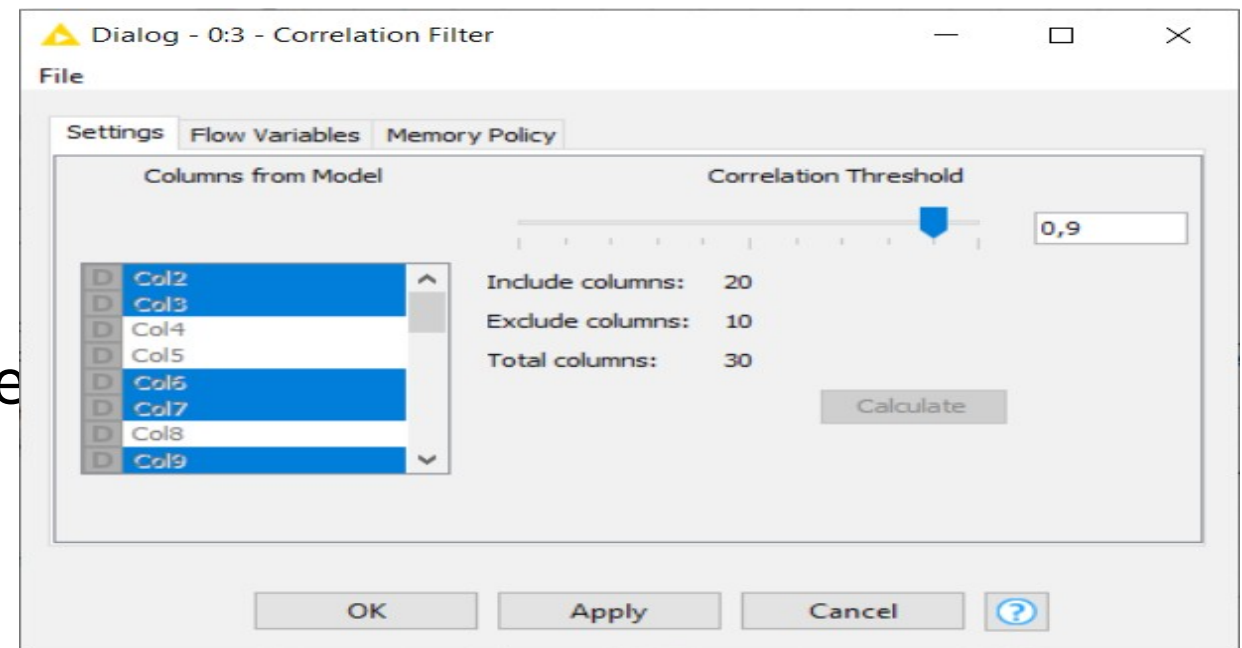
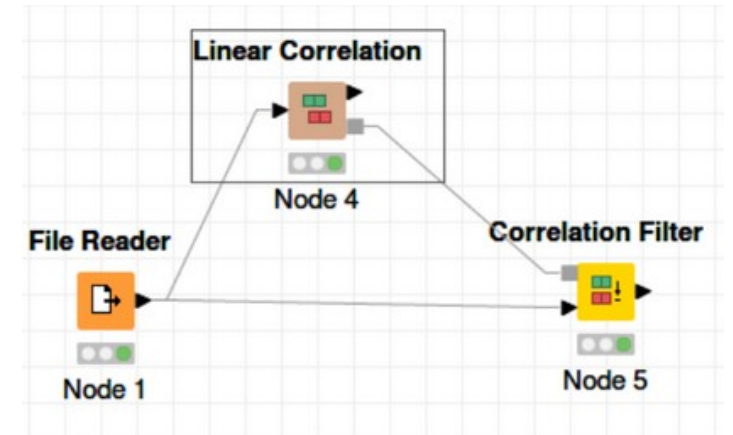
Row ID	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Col13	Col14	Col15	Col16	Col17	Col18	Col19
Col1	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Col2	?	1	0.324	0.998	0.987	0.171	0.506	0.677	0.823	0.148	-0.312	0.679	-0.097	0.674	0.736	-0.223	0.206	0.194	0.376
Col3	?	0.324	1	0.33	0.321	-0.023	0.237	0.302	0.293	0.071	-0.076	0.276	0.386	0.282	0.26	0.007	0.192	0.143	0.164
Col4	?	0.998	0.33	1	0.987	0.207	0.557	0.716	0.851	0.183	-0.261	0.692	-0.087	0.693	0.745	-0.203	0.251	0.228	0.407
Col5	?	0.987	0.321	0.987	1	0.177	0.499	0.686	0.823	0.151	-0.283	0.733	-0.066	0.727	0.8	-0.167	0.213	0.208	0.372
Col6	?	0.171	-0.023	0.207	0.177	1	0.659	0.522	0.554	0.558	0.585	0.301	0.068	0.296	0.247	0.332	0.319	0.248	0.381
Col7	?	0.506	0.237	0.557	0.499	0.659	1	0.883	0.831	0.603	0.565	0.497	0.046	0.549	0.456	0.135	0.739	0.571	0.642
Col8	?	0.677	0.302	0.716	0.686	0.522	0.883	1	0.921	0.501	0.337	0.632	0.076	0.66	0.617	0.099	0.67	0.691	0.683
Col9	?	0.823	0.293	0.851	0.823	0.554	0.831	0.921	1	0.462	0.167	0.698	0.021	0.711	0.69	0.028	0.49	0.439	0.616
Col10	?	0.148	0.071	0.183	0.151	0.558	0.603	0.501	0.462	1	0.48	0.303	0.128	0.314	0.224	0.187	0.422	0.343	0.393
Col11	?	-0.312	-0.076	-0.261	-0.283	0.585	0.565	0.337	0.167	0.48	1	0	0.164	0.04	-0.09	0.402	0.56	0.447	0.341
Col12	?	0.679	0.276	0.692	0.733	0.301	0.497	0.632	0.698	0.303	0	1	0.213	0.973	0.952	0.165	0.356	0.332	0.513
Col13	?	-0.097	0.386	-0.087	-0.066	0.068	0.046	0.076	0.021	0.128	0.164	0.213	1	0.223	0.112	0.397	0.232	0.195	0.23
Col14	?	0.674	0.282	0.693	0.727	0.296	0.549	0.66	0.711	0.314	0.04	0.973	0.223	1	0.938	0.151	0.416	0.362	0.556
Col15	?	0.736	0.26	0.745	0.8	0.247	0.456	0.617	0.69	0.224	-0.09	0.952	0.112	0.938	1	0.075	0.285	0.271	0.416
Col16	?	-0.223	0.007	-0.203	-0.167	0.332	0.135	0.099	0.028	0.187	0.402	0.165	0.397	0.151	0.075	1	0.337	0.269	0.328
Col17	?	0.206	0.192	0.251	0.213	0.319	0.739	0.67	0.49	0.422	0.56	0.356	0.232	0.416	0.285	0.337	1	0.801	0.744
Col18	?	0.194	0.143	0.228	0.208	0.248	0.571	0.691	0.439	0.343	0.447	0.332	0.195	0.362	0.271	0.269	0.801	1	0.772
Col19	?	0.376	0.164	0.407	0.372	0.381	0.642	0.683	0.616	0.393	0.341	0.513	0.23	0.556	0.416	0.328	0.744	0.772	1
Col20	?	-0.104	0.009	-0.082	-0.072	0.201	0.23	0.178	0.095	0.449	0.345	0.241	0.412	0.266	0.134	0.414	0.395	0.309	0.313
Col21	?	-0.043	0.054	-0.006	-0.02	0.284	0.507	0.449	0.258	0.332	0.688	0.228	0.28	0.244	0.127	0.427	0.803	0.727	0.611
Col22	?	0.97	0.353	0.969	0.963	0.213	0.535	0.688	0.83	0.186	-0.254	0.715	-0.112	0.697	0.757	-0.231	0.205	0.187	0.358
Col23	?	0.297	0.912	0.303	0.287	0.036	0.248	0.3	0.293	0.091	-0.051	0.195	0.409	0.2	0.196	-0.075	0.143	0.1	0.087
Col24	?	0.965	0.358	0.97	0.959	0.239	0.59	0.73	0.856	0.219	-0.205	0.72	-0.102	0.721	0.761	-0.217	0.261	0.227	0.395
Col25	?	0.941	0.344	0.942	0.959	0.207	0.51	0.676	0.81	0.177	-0.232	0.752	-0.083	0.731	0.811	-0.182	0.199	0.188	0.342
Col26	?	0.12	0.078	0.151	0.124	0.805	0.566	0.449	0.453	0.427	0.505	0.142	-0.074	0.13	0.125	0.314	0.227	0.168	0.215
Col27	?	0.413	0.278	0.456	0.39	0.472	0.866	0.755	0.667	0.473	0.459	0.287	-0.092	0.342	0.283	-0.056	0.679	0.485	0.453
Col28	?	0.527	0.301	0.564	0.513	0.435	0.816	0.884	0.752	0.434	0.346	0.381	-0.069	0.419	0.385	-0.058	0.639	0.663	0.55
Col29	?	0.744	0.295	0.771	0.722	0.503	0.816	0.861	0.91	0.43	0.175	0.531	-0.12	0.555	0.538	-0.102	0.483	0.44	0.602
Col30	?	0.164	0.105	0.189	0.144	0.394	0.51	0.409	0.376	0.7	0.334	0.095	-0.128	0.11	0.074	-0.107	0.278	0.198	0.143
Col31	?	0.007	0.119	0.051	0.004	0.499	0.687	0.515	0.369	0.438	0.767	0.05	-0.046	0.085	0.018	0.101	0.591	0.439	0.311


 Paire de colonne
moins corrélées


 Paire de colonne très
corrélées

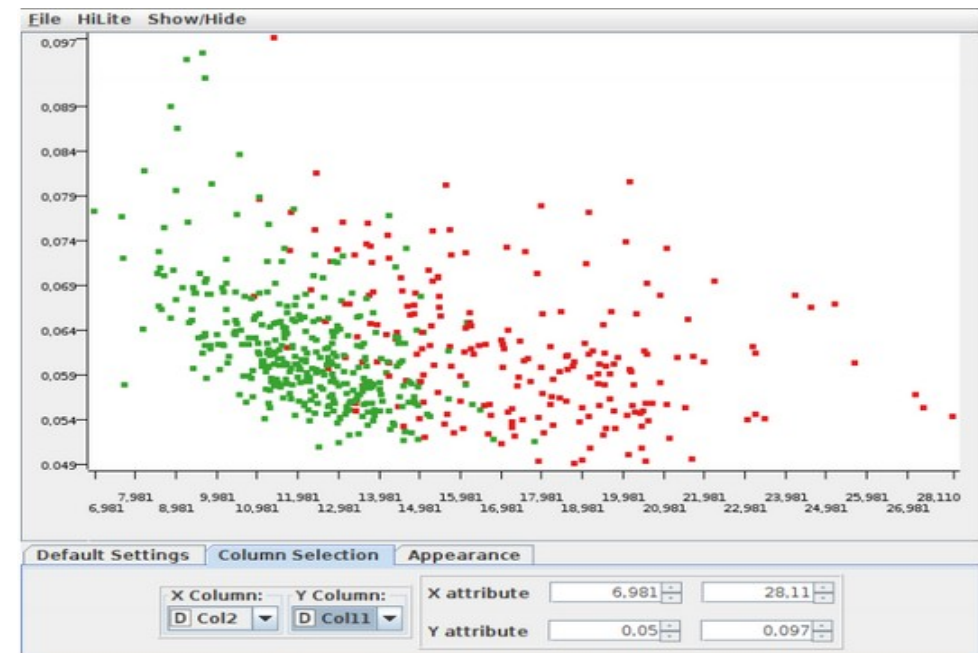
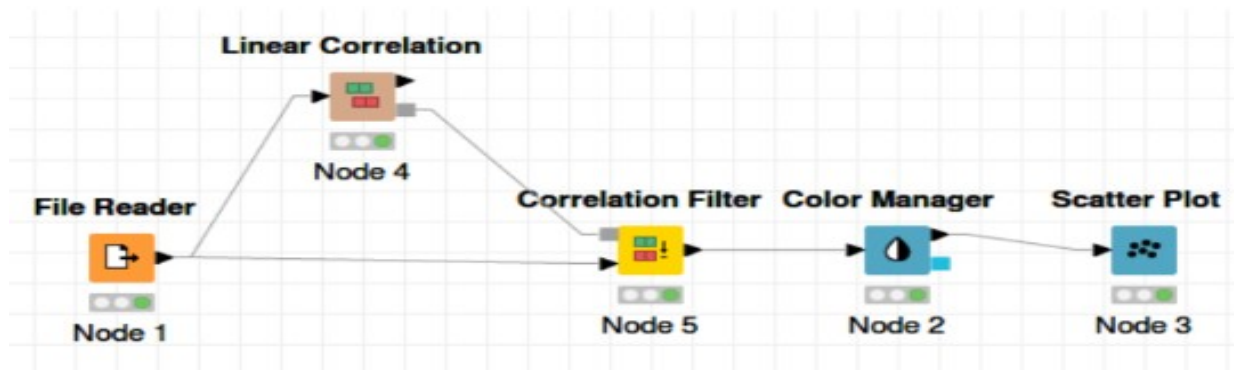
CORRELATION FILTER

Pour notre Correlation filter on a choisi de prendre l'indice de corrélation à 0.9 un indice assez élevé pour garder le maximum d'information.
(20 au lieu de 30 sur notre modèle).



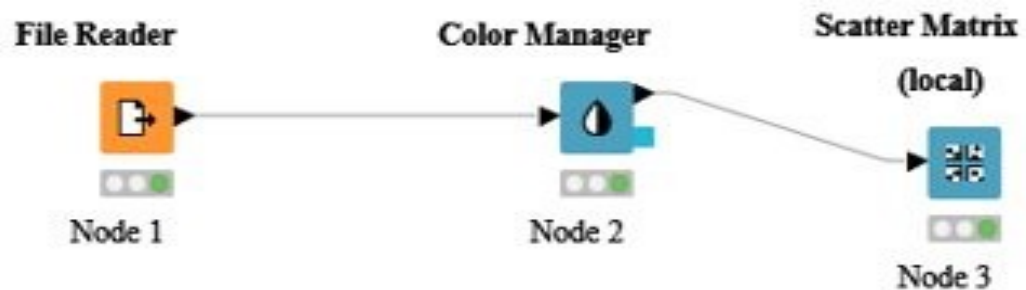
SCATTER PLOT

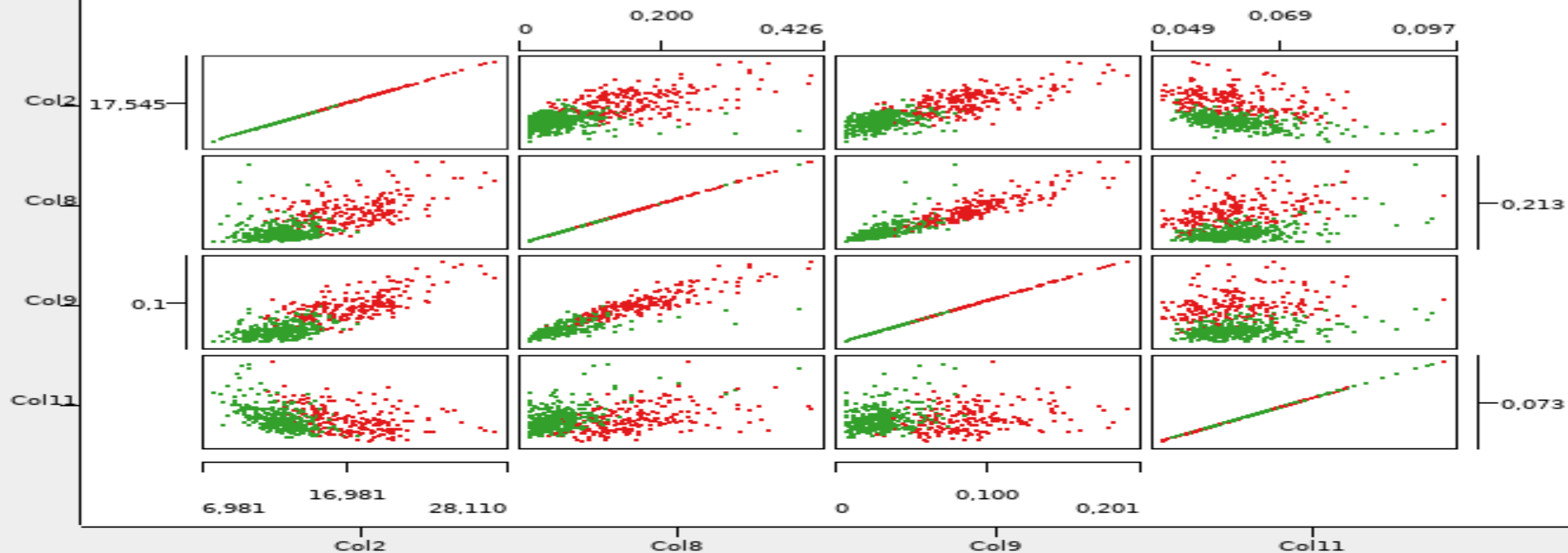
Ce type de graphique s'avère utile pour déceler les relation entre des valeurs et de dégager les valeurs hors norme dans un ensemble de données.



SCATTER MATRIX

Cet outil nous permet de visualiser toutes les colonnes choisies au même temps





Default Settings

Column Selection

Appearance

Exclude

Filter

I Col0
S Col1
D Col3
D Col4
D Col5
D Col6
D Col7
D Col10
D Col12
D Col13

>

>>

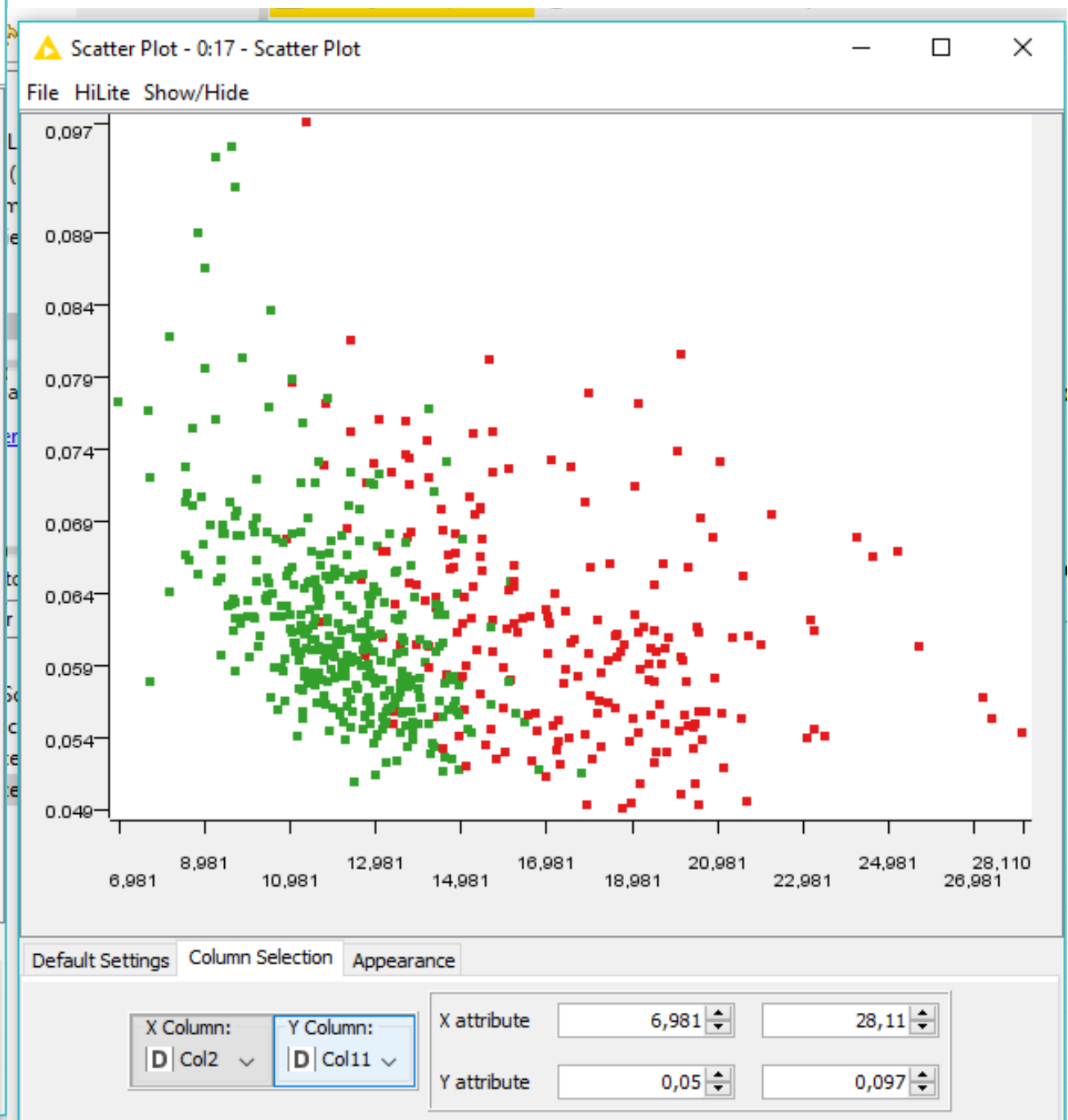
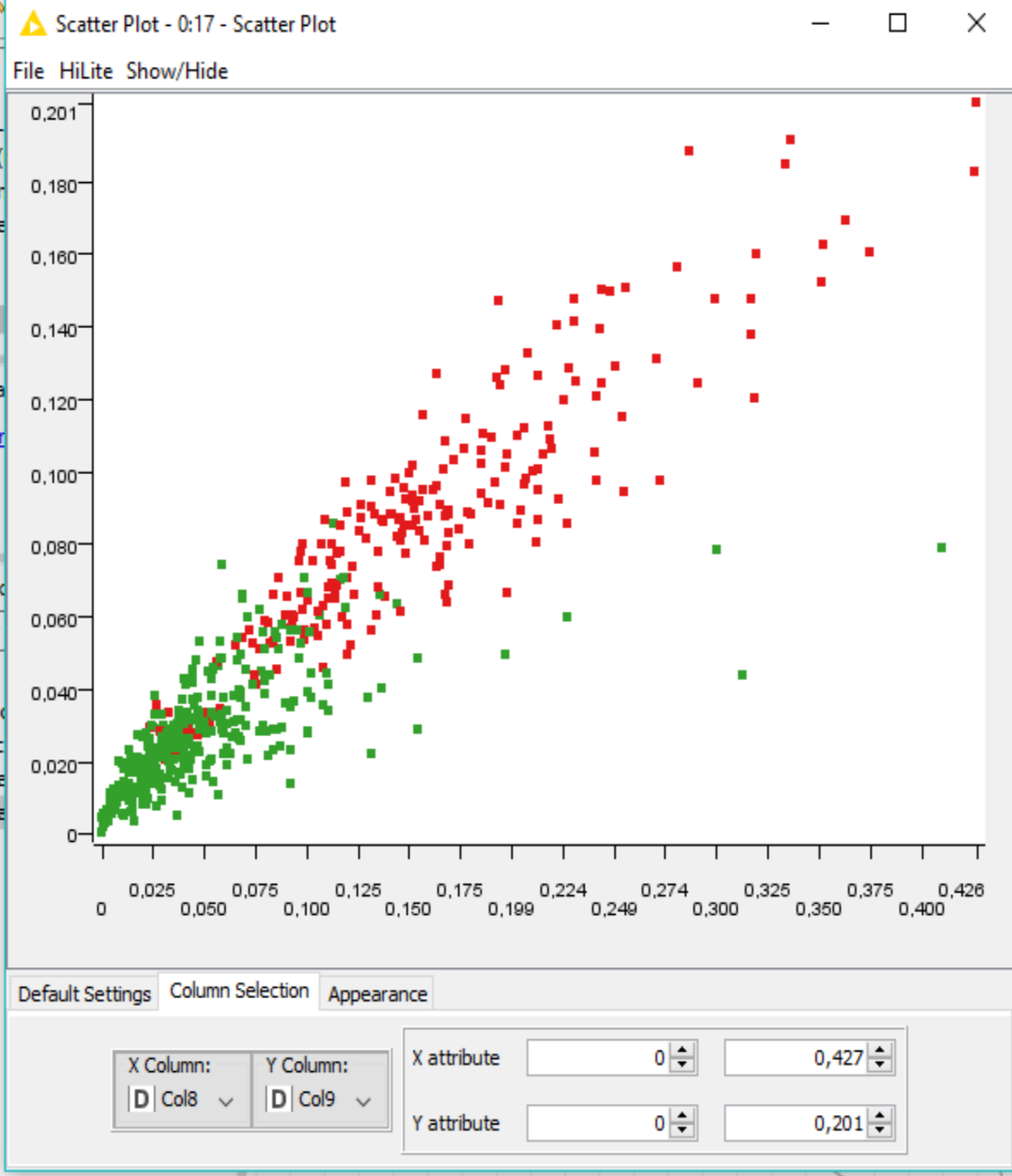
<

<<

Include

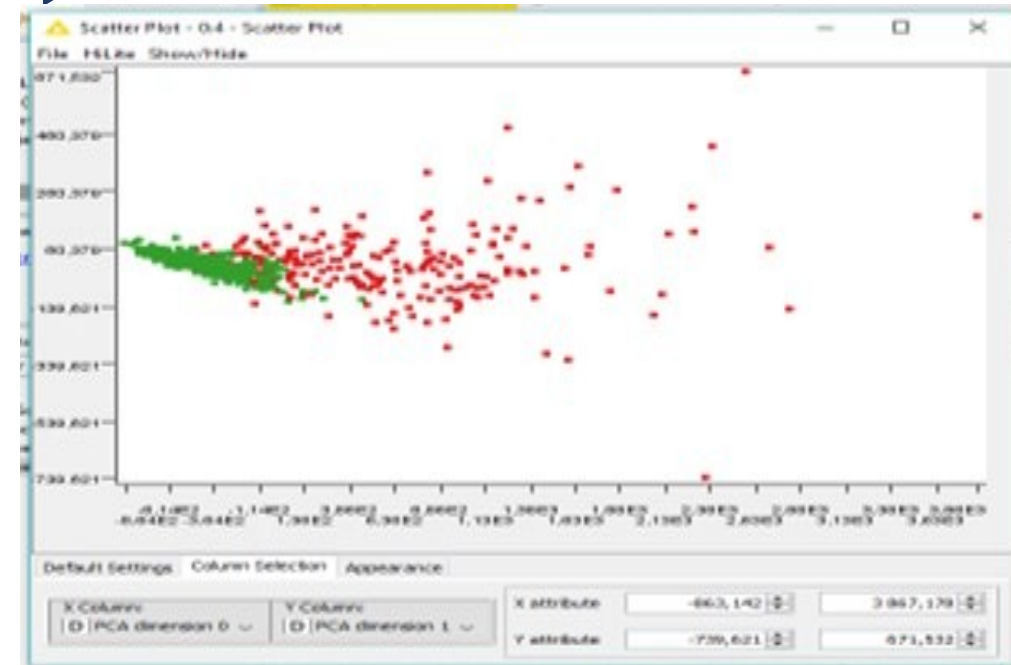
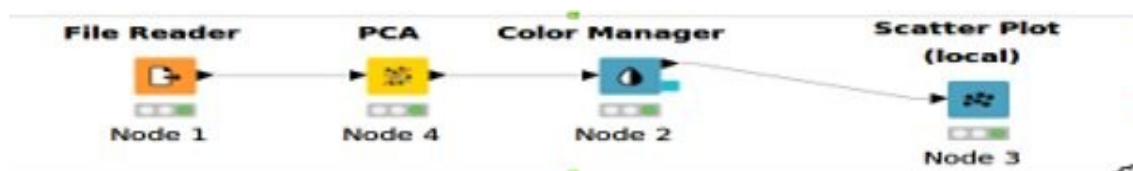
Filter

D Col2
D Col8
D Col9
D Col11



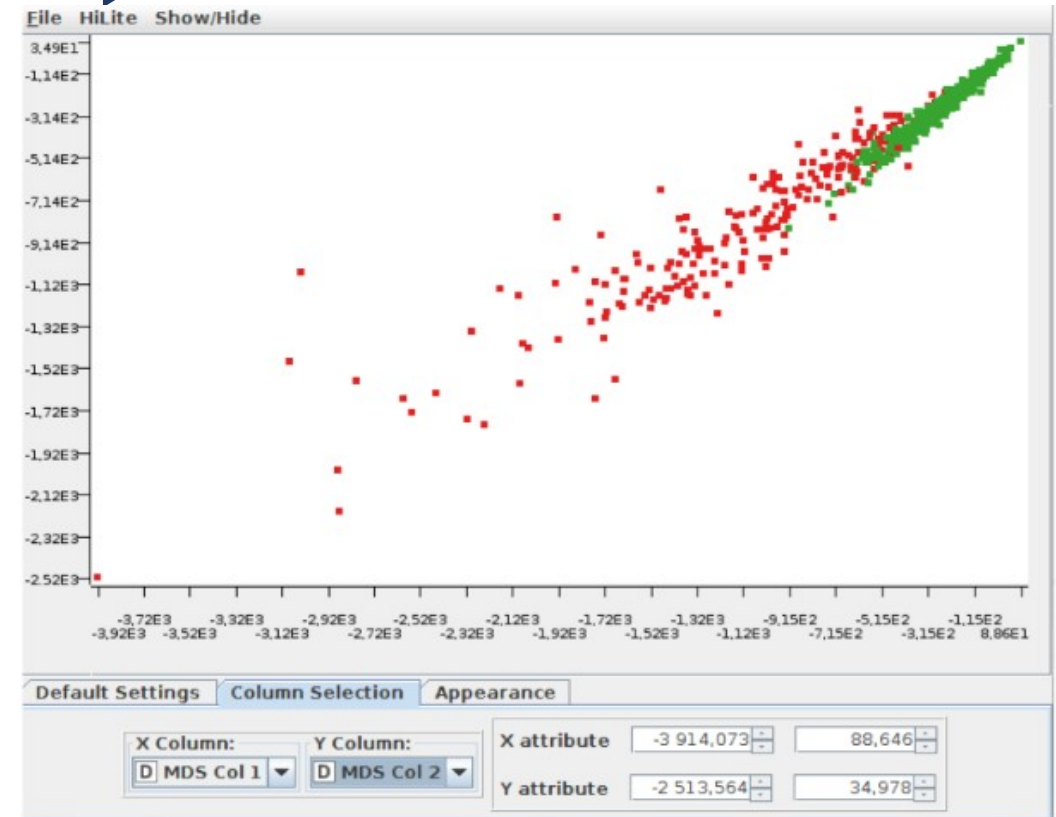
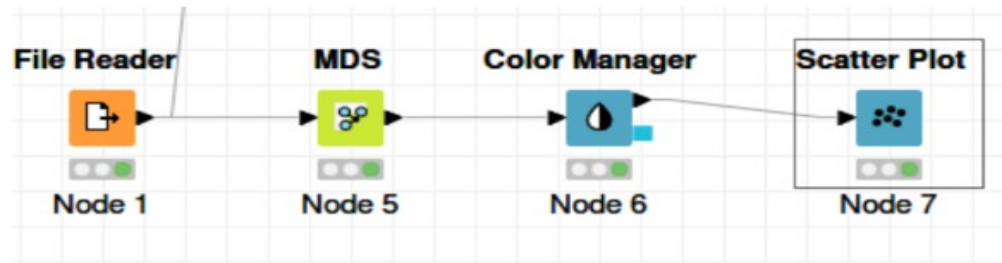
PCA
(Analyse en composantes principales)

PCA nous donne la possibilité
de réduire nos vecteurs en
moins de dimensions grâce à
l'indice de corrélation



MDS

- On remarque que MDS à tout simplement renversé le diagramme obtenu avec PCA



CLUSTERING

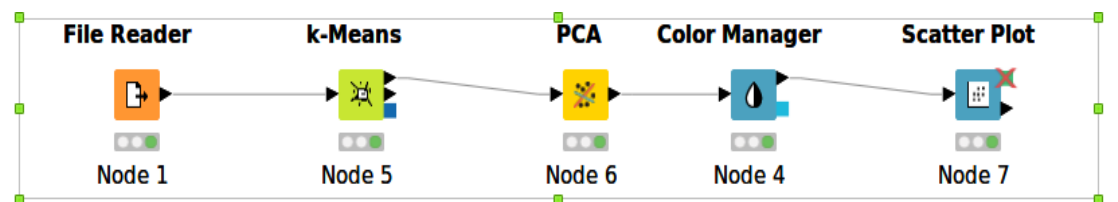
- Ce modèle consiste à utiliser les arbres de décision qui apprend sur une partie des données et prédit la classe du reste des données*

Beaucoup d'erreurs pour les données se trouvant au centre du nuage de points.

Deux clusters:

Cluster_0: tumeur maligne

Cluster_1: tumeur bénigne



Input data and view ...

File Hilite Navigation View

Flow Variables

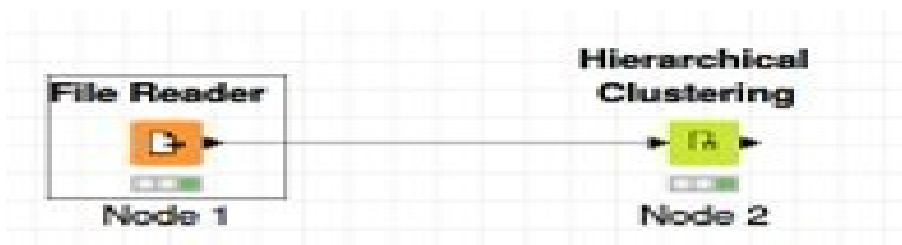
Spec - Columns: 36 Properties

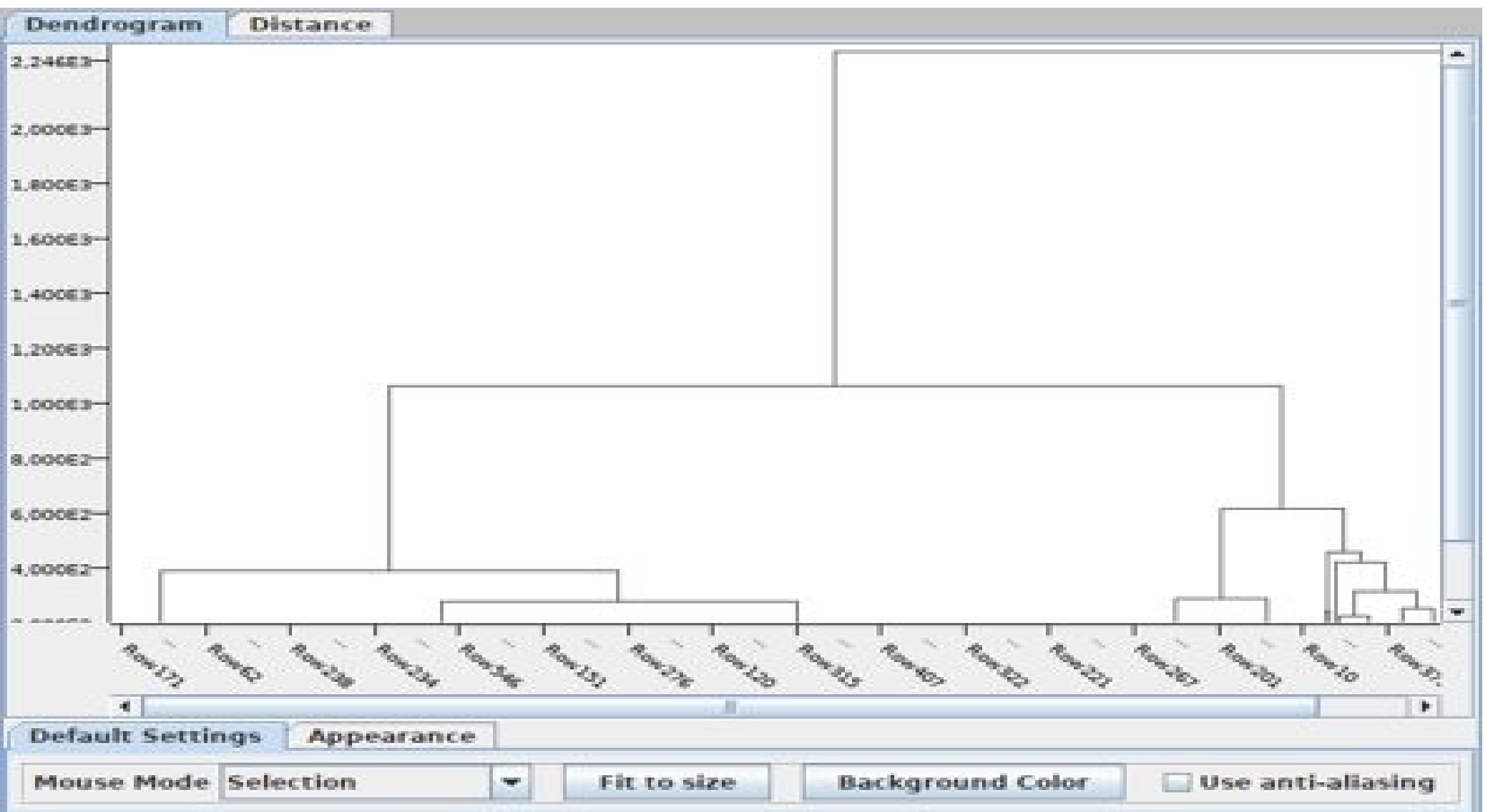
Table "default" - Rows: 569

Row ID	Cluster	Col1
Row131	cluster_0	M
Row287	cluster_0	B
Row291	cluster_0	B
Row403	cluster_0	B
Row47	cluster_0	M
Row95	cluster_1	M
Row96	cluster_0	B
Row108	cluster_1	M
Row111	cluster_0	B
Row112	cluster_0	B
Row121	cluster_1	M
Row125	cluster_0	B
Row153	cluster_0	B
Row166	cluster_0	B
Row171	cluster_0	M
Row172	cluster_0	M
Row203	cluster_0	M
Row204	cluster_0	B
Row282	cluster_1	M
Row304	cluster_0	B
Row306	cluster_0	B
Row307	cluster_0	B
Row326	cluster_0	B
Row339	cluster_1	M
Row340	cluster_0	B
Row342	cluster_0	B
Row347	cluster_0	B
Row382	cluster_0	B
Row383	cluster_0	B
Row385	cluster_0	M
Row389	cluster_1	M

HIERARCHICAL CLUSTERING

En fonction de la distance
considérer entre les éléments
on obtient un nombre de
cluster différent



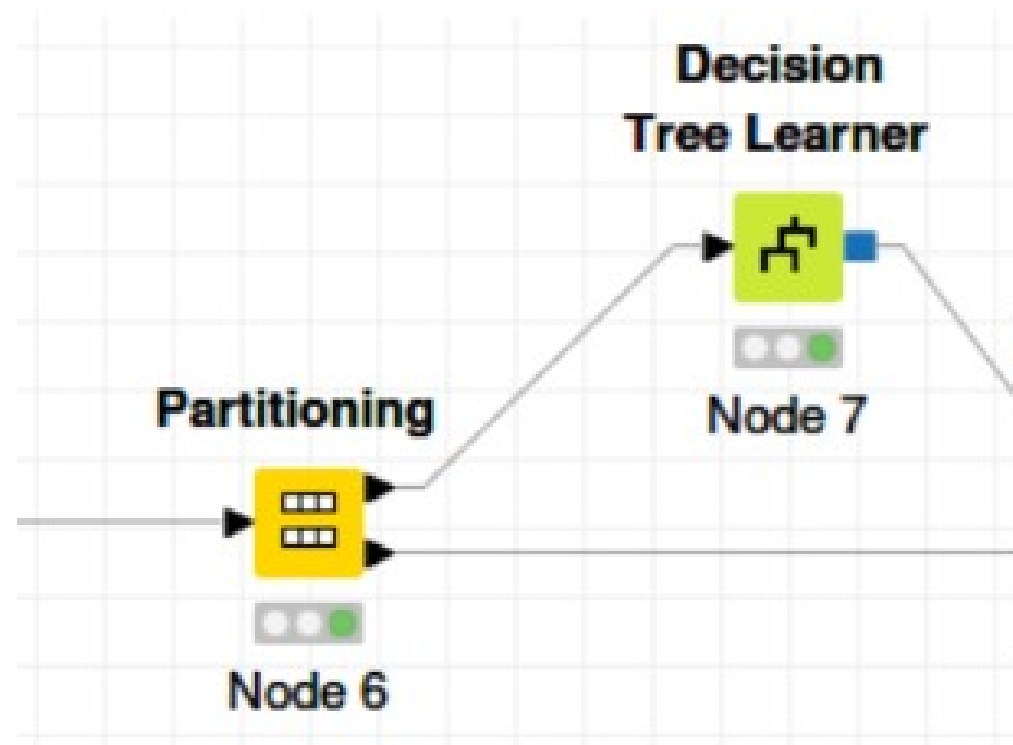


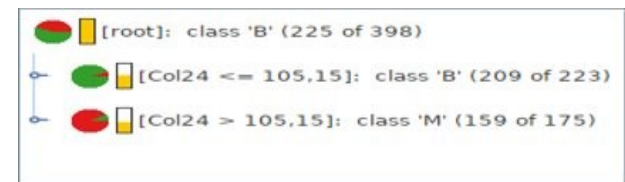
A dark blue, irregular ink splatter shape centered on a white background. The splatter has a rough, textured edge with some smaller droplets and splatters extending outwards. The text is centered within the main body of the splatter.

PARTIE PRÉDICTION

DECISION TREE PREDICTOR

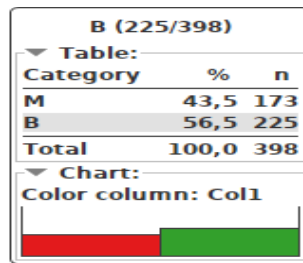
Après avoir générer l'arbre de décision avec Decision tree learner, Decision tree predictor quant à lui permet de nous prédire si les données appartiennent à la classe maligne ou bénigne en utilisant toutes nos colonnes





Exemple avec la donnée row5

Row ID	D Col14	D Col15	D Col16	D Col17	D Col18	D Col19	D Col20	D Col21	D Col22	D Col23	D Col24	D Col25	D Col26	D Col27
Row5	2.217	27.19	0.008	0.033	0.037	0.011	0.022	0.005	15.47	23.75	103.4	741.6	0.179	0.525

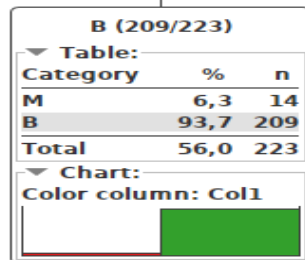


Col24

⊖

$\leq 105,15$

$> 105,15$

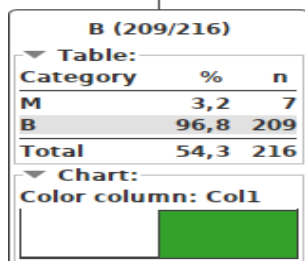


Col26

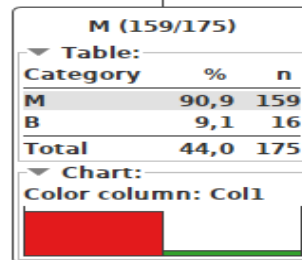
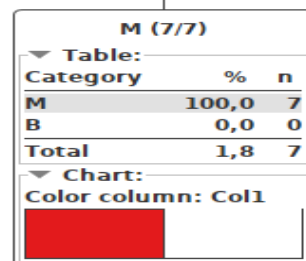
⊖

$\leq 0,1759$

$> 0,1759$



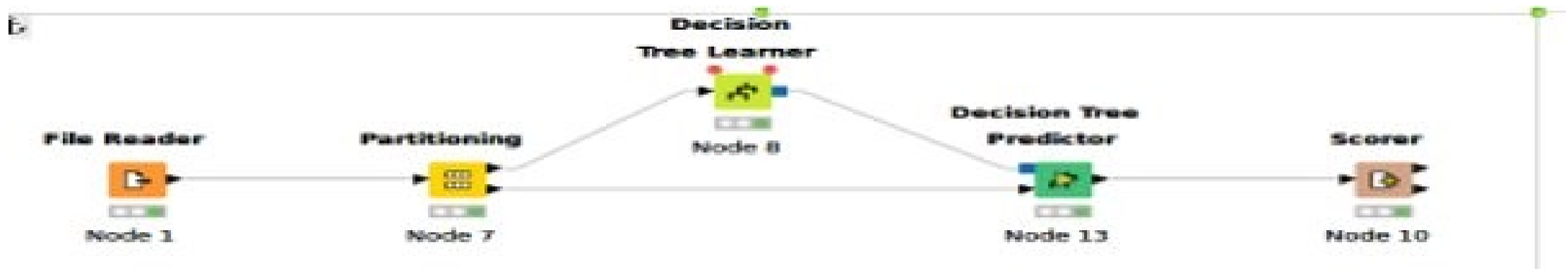
⊕



⊕

SCORER

Ce modèle consiste à utiliser les arbres de décision qui apprend sur une partie des données et prédit la classe du reste des données.



LES PRÉDICTIONS

Ce modèle consiste à utiliser les arbres de décision qui apprend sur une partie des données et prédit la classe du reste des données.

Col1 \ Pre...	M	B
M	37	2
B	14	118

Correct classified: 155
Accuracy: 90,643 %
Cohen's kappa (κ) 0,76

Wrong classified: 16
Error: 9,357 %

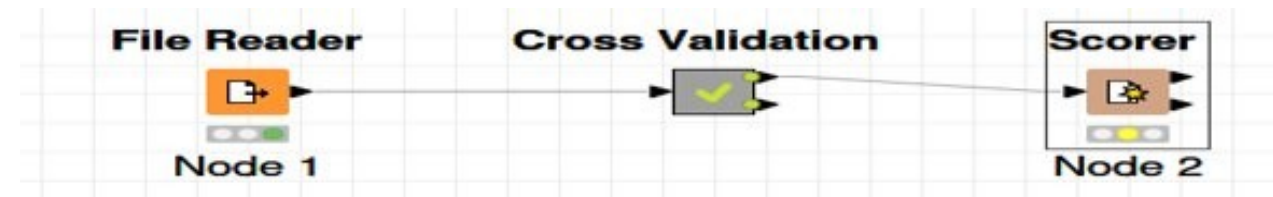
File Hilite Navigation View

Table "wdbc.data" - Rows: 569 Sp

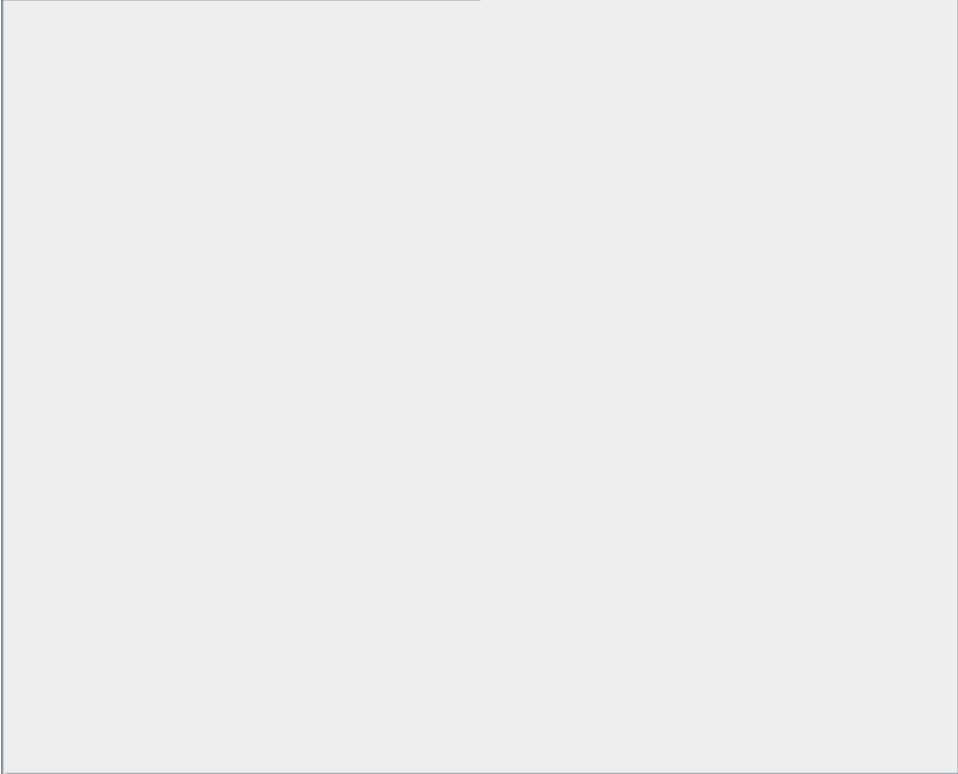
Row ID	Col1	Predic...
Row3	M	M
Row37	B	B
Row48	B	B
Row72	M	M
Row80	B	B
Row81	B	M
Row89	B	B
Row117	M	M
Row133	B	B
Row134	M	M
Row167	M	M
Row170	B	B
Row172	M	M
Row192	B	B
Row195	B	B
Row196	M	M
Row208	B	B
Row217	B	B
Row220	B	B
Row223	M	M
Row224	B	B
Row262	M	M
Row265	M	M
Row278	B	B
Row299	B	B
Row302	M	M
Row306	B	B
Row322	B	B

CROSS VALIDATION

Cross validation est un outil de prédiction qui sert à estimer la fiabilité du modèle d'apprentissage fondé sur la technique d'échantillonnage



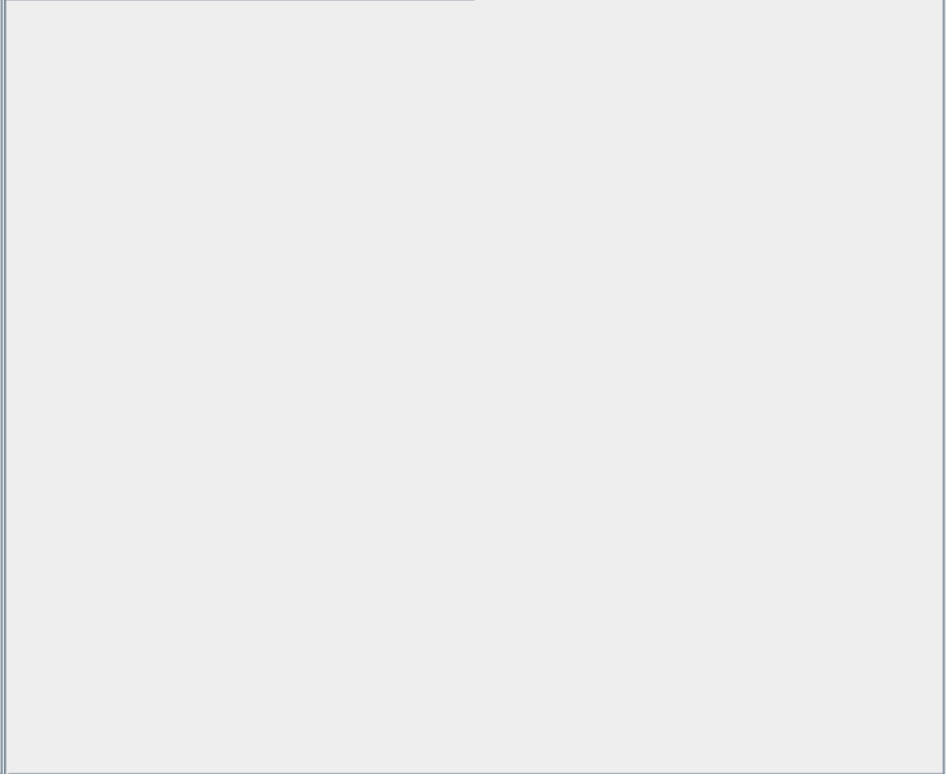
Col1 \ Pre...	M	B
M	193	19
B	20	337



Correct classified: 530
Accuracy: 93,146 %
Cohen's kappa (κ) 0,854

Echantillonnage aléatoire

Col1 \ Pre...	M	B
M	192	20
B	19	338

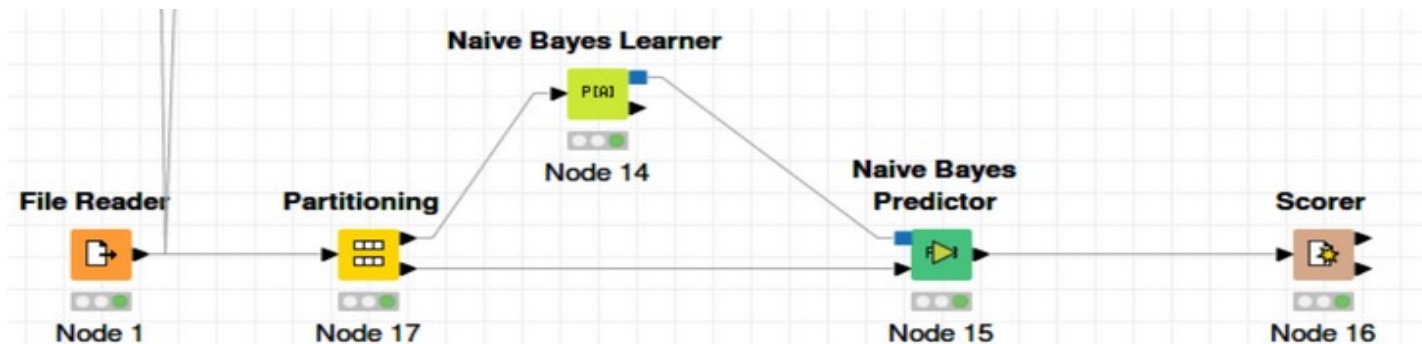


Correct classified: 530
Accuracy: 93,146 %
Cohen's kappa (κ) 0,853

Echantillonnage linéaire

NAÏVE BYES

La particularité de cet outil est de supposer l'existence d'une caractéristique spécifique pour identifier les classes

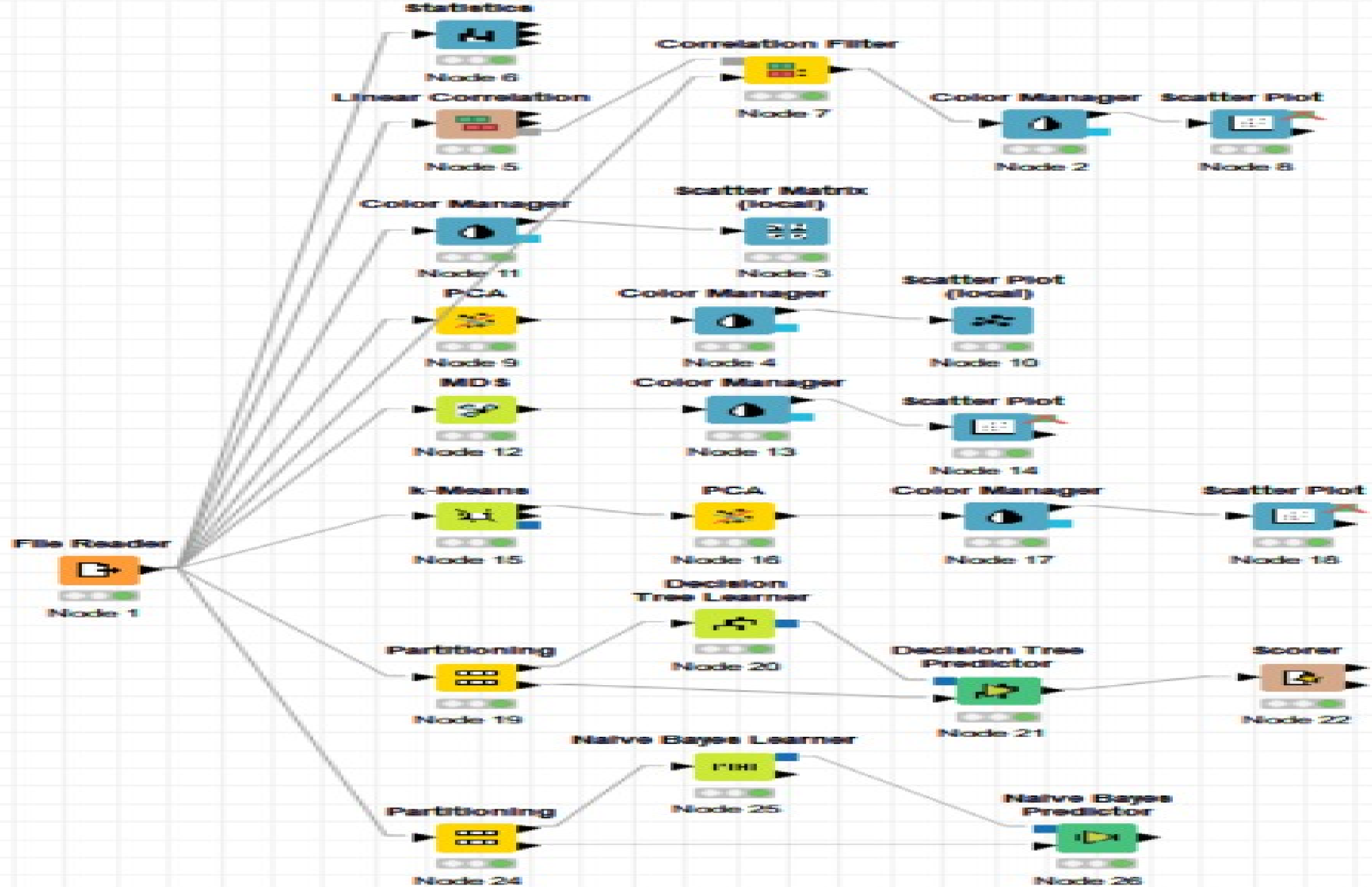


File Hilite		
Col1 \ Pre...	M	B
M	115	11
B	12	231

Correct classified: 346	Wrong classified: 23
Accuracy: 93,767 %	Error: 6,233 %
Cohen's kappa (κ) 0,862	



VUE COMPLETE DU WORKFLOW



CONCLUSION

- ✓ Plus le rayon est élevé plus on risque d'avoir une tumeur maligne, dont les cellules se multiplient et se propagent pour envahir les organes voisins.