RAPPORT DE PROJET VISUALISATION DES DONNEES «DIAGNOSTIC DE CANCER WISCOSIN WDBC»

Réaliser par :

Sylia RAHMANI 11707524

Souade HADJ ALI 11708934

Contenu

Introduction

[slide 1, 2, 3]

Dans ce rapport nous allons vous exposer une étude analytique faite sur la base de données WDBC pour comprendre les facteurs influents et les causes de la maladie du cancer de sien afin d'essayer de prévenir cette dernière.
Notre réflexion se déploie sur trois étapes majeures ; il convient de faire dans un premier temps une présentation générale des données afin de comprendre leurs provenances leurs caractéristiques, suivie des chaines de traitements faites avec KINME & PYTHON, deux outils qui s'avèrent utiles pour une fouille efficace des données. Enfin nous allons analyser chaque résultat ce qui nous permettront par la suite de faire des préventions, de mettre des hypothèses qui faciliteront le traitement de cette maladie.
Ce document vient pour expliquer les résultats décrits dans le Powerpoint qui se trouve dans le même répertoire que lui.

Présentation des données et du DATAFRAME :

[Slide 4]

L'image ci-dessous nous présente caractéristiques de notre DATAFRAME; le nombre de lignes, de colonnes, le nom de chaque colonne ainsi que le type de donnée dans chaque colonne.

```
wdbc.info
<bound method DataFrame.info of            id diag  rayon1  texture1  perimetre1   zone1  douceur1  \
0      842302    M   17.99     10.38      122.80  1001.0   0.11840
1      842517    M   20.57     17.77      132.90  1326.0   0.08474
2    84300903    M   19.69     21.25      130.00  1203.0   0.10960
3    84348301    M   11.42     20.38       77.58   386.1   0.14250
4    84358402    M   20.29     14.34      135.10  1297.0   0.10030
..        ...  ...     ...       ...         ...     ...       ...
562    926424    M   21.56     22.39      142.00  1479.0   0.11100
563    926682    M   20.13     28.25      131.20  1261.0   0.09780
564    926954    M   16.60     28.08      108.30   858.1   0.08455
565    927241    M   20.60     29.33      140.10  1265.0   0.11780
566     92751    B    7.76     24.54       47.92   181.0   0.05263

     compacite1  concavite1  concave1  ...  rayon3  texture3  perimetre3  \
0       0.27760     0.30010   0.14710  ...  25.380     17.33      184.60
1       0.07864     0.08690   0.07017  ...  24.990     23.41      158.80
2       0.15990     0.19740   0.12790  ...  23.570     25.53      152.50
3       0.28390     0.24140   0.10520  ...  14.910     26.50       98.87
4       0.13280     0.19800   0.10430  ...  22.540     16.67      152.20
..          ...         ...       ...  ...     ...       ...         ...
562     0.11590     0.24390   0.13890  ...  25.450     26.40      166.10
563     0.10340     0.14400   0.09791  ...  23.690     38.25      155.00
564     0.10230     0.09251   0.05302  ...  18.980     34.12      126.70
565     0.27700     0.35140   0.15200  ...  25.740     39.42      184.60
566     0.04362     0.00000   0.00000  ...   9.456     30.37       59.16

      zone3  douceur3  compacite3  concavite3  concave3  symetrie3  \
0    2019.0   0.16220     0.66560      0.7119    0.2654     0.4601
1    1956.0   0.12380     0.18660      0.2416    0.1860     0.2750
2    1709.0   0.14440     0.42450      0.4504    0.2430     0.3613
3     567.7   0.20980     0.86630      0.6869    0.2575     0.6638
4    1575.0   0.13740     0.20500      0.4000    0.1625     0.2364
..      ...       ...         ...         ...       ...        ...
562  2027.0   0.14100     0.21130      0.4107    0.2216     0.2060
563  1731.0   0.11660     0.19220      0.3215    0.1628     0.2572
564  1124.0   0.11390     0.30940      0.3403    0.1418     0.2218
565  1821.0   0.16500     0.86810      0.9387    0.2650     0.4087
566   268.6   0.08996     0.06444      0.0000    0.0000     0.2871

     dimfractale3
0         0.11890
1         0.08902
2         0.08758
3         0.17300
4         0.07678
..            ...
562       0.07115
563       0.06637
564       0.07820
565       0.12400
566       0.07039

[567 rows x 32 columns]>
```

Les données que nous allons traiter proviennent de vrais patients, les caractéristiques sont calculées à partir d'une image numérisée d'une aiguille fine d'aspiration d'une masse au sein, elles décrivent les caractéristiques des noyaux de cellules présentes dans l'image.
Les données sont extraites de 569 individus possédant des tumeurs et sont répertoriées dans un tableau selon 32 attributs tels qu'un attribut permet d'identifier la personne, un autre permet de savoir si la tumeur est bénigne (inoffensif) ou maligne (offensif). Les 30 autres colonnes sont des fonctions à valeurs réelles correspondent à dix caractéristiques de trois noyaux différents de la tumeur.

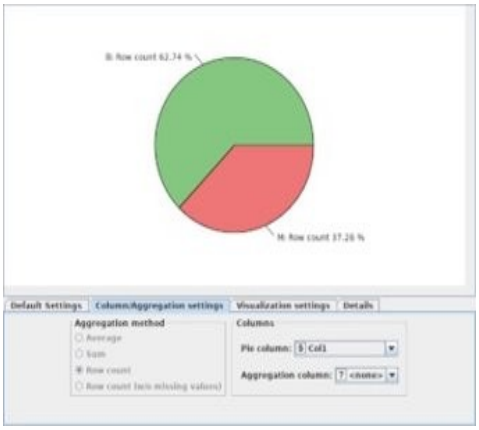
visualisation

Paie chart

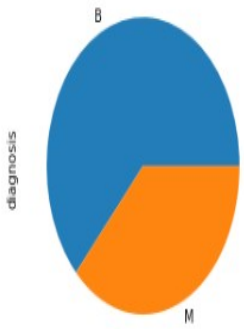
[Slide 5]

Nous avons utilisé pie chart (Camembert) afin d'observer la répartition de nos données.

On remarque que deux classes se dégagent assez rapidement, sur les 569 individus on trouve un pourcentage de 62,74% de personnes possédants une tumeur bénigne (partie verte) et 37.26% maligne (partie rouge).



Pie chart avec KNIME - SLIDE 4



Pie chart avec python

La fonction ci-dessous montre que la base de données contient : 357 personnes possèdent une tumeur bénigne et 212 malignes.

```
total = wdbc['diagnosis'].count()
maligne = wdbc[wdbc['diagnosis'] == "M"]['diagnosis'].count()
print("tumeur Maligne", maligne)
print("Tumeur Benigne", total-maligne)
```

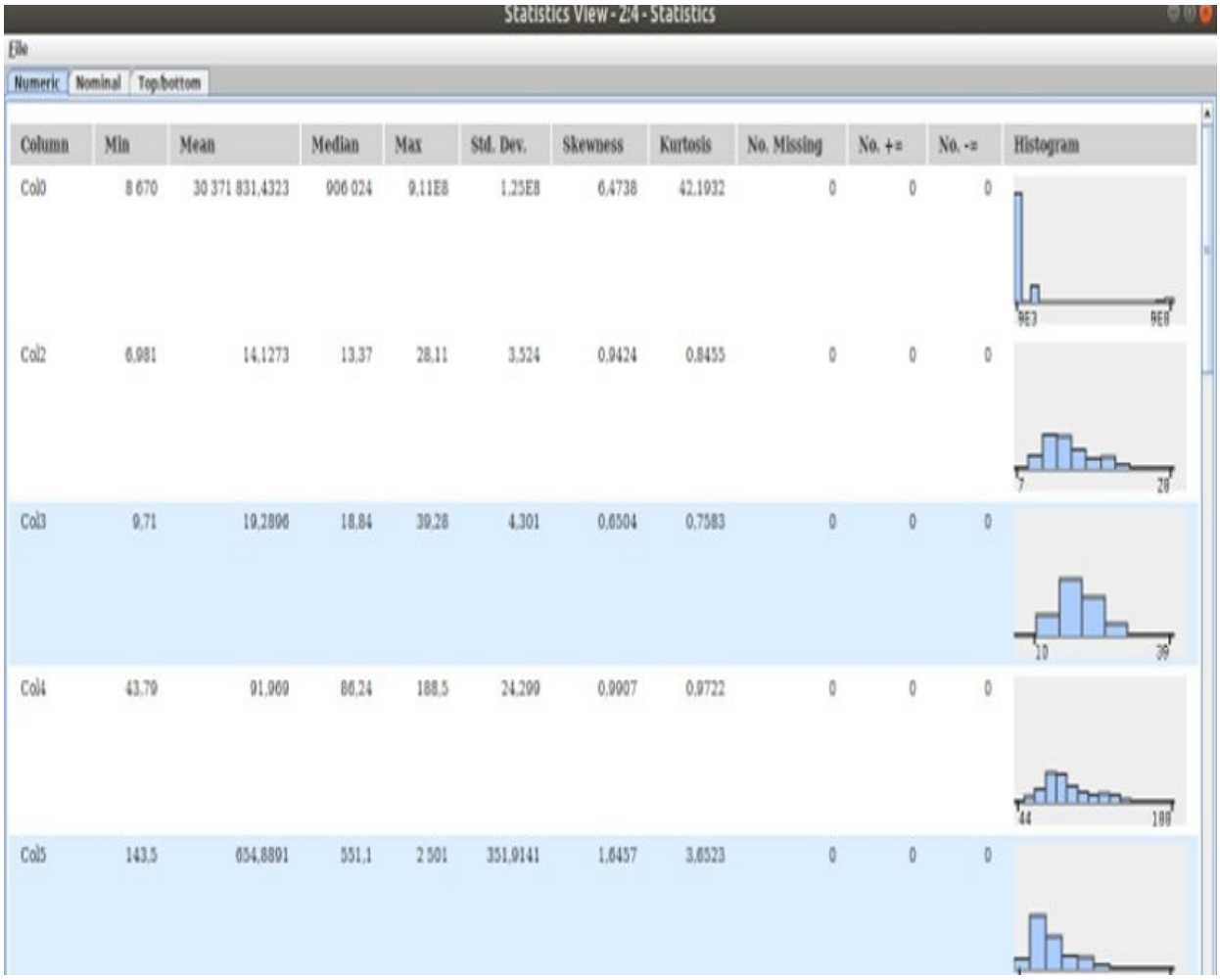
```
tumeur Maligne 212
Tumeur Benigne 357
```

Traitement des données

[Slide 7,8]

Vue statistique

Avec l'outil « Statistics » on peut calculer pour chaque colonne de nos données les différends constatent statistique telles que le minimum, le maximum, la médiane etc. Ainsi leurs représentations sous forme d'histogramme.



Pour une statique rapide avec python on obtient le tableau suivant :

```
wdbc = wdbc.drop(['id'] , axis =1)
wdbc.describe()
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fract
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	

8 rows × 31 columns

Pour voir la répartition des données par rapport à la médiane, l'écart-type, et voir aussi l'aplatissement des données. Par exemple, un écart-type assez faible fait prédire que les données sont assez recentrées et on n'a pas de valeur aberrante. De plus, on peut voir que le minimum et le maximum sont assez proches de la médiane.
En revanche, quand on regarde la moyenne de la colonne 5 (654.9) et la moyenne de la colonne 2 (14.1273) par exemple, on constate que les moyennes sont assez étendues.
L’étude statistique nous donne un premier aperçu et nous aide à décider si on va normaliser ou non les données afin de les rendre plus analysables et d'avoir des valeurs les plus significatives possible.

Quelques hypothèses

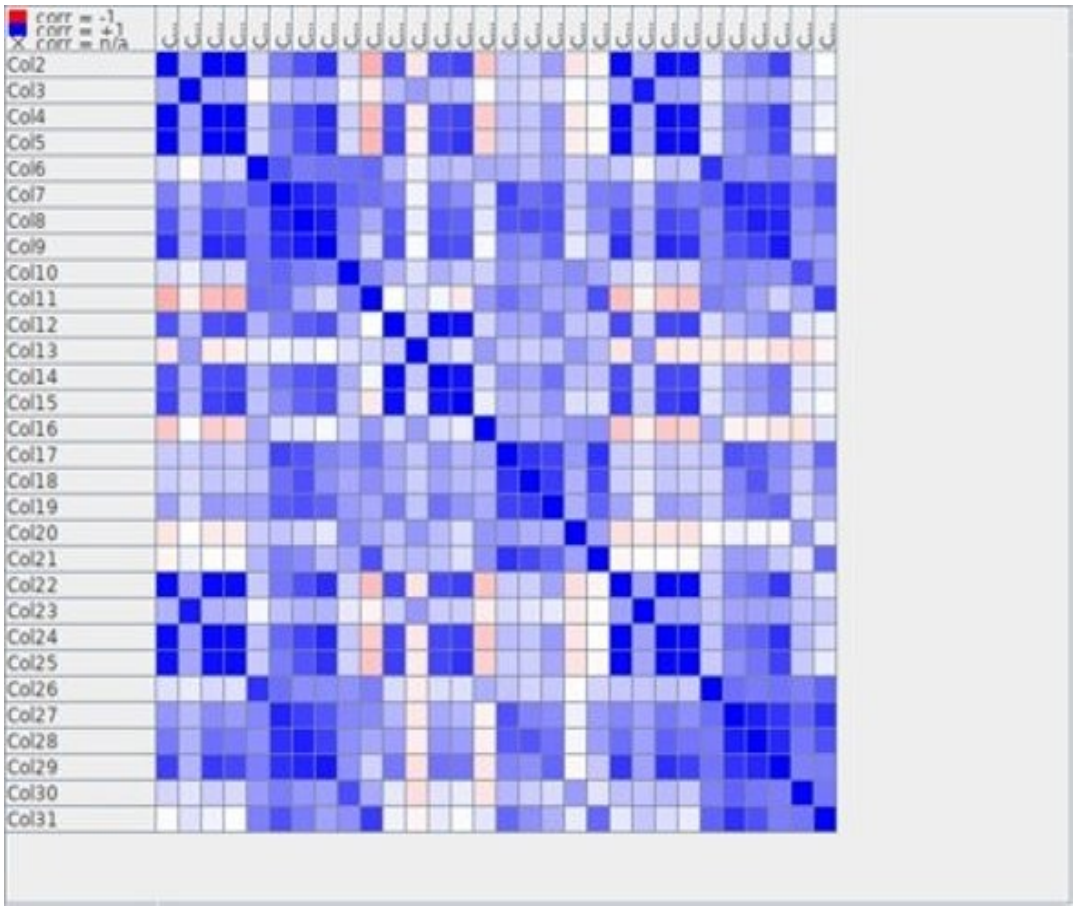
[Slide 9]

L’observation des histogrammes obtenus à l’aide de python nous donne une première impression que plus la concavité ou le rayon de la tumeur augmente, plus la tumeur a des risques d'être maligne.
En revanche, avec l'histogramme 3, on voie clairement que la symétrie de la tumeur n'a aucune relation avec le critère bénin ou malin d'une tumeur car y a un chevauchement important entre les barres de l’histogramme.
On peut dire alors que cette colonne apporte peu d’informations à notre étude, il est envisageable de la supprimer dans la suite de notre analyse.

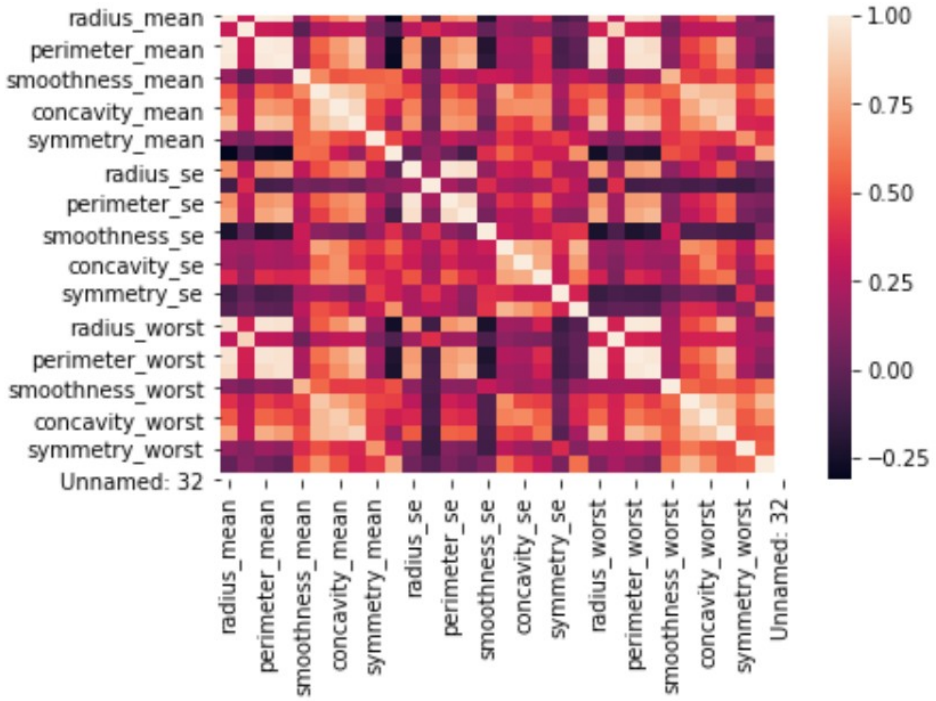
Linear Correlation et Linear Filter

[Slide 10]

En plus du paie chart vu précédemment, on dispose aussi de la matrice de corrélation qui est une autre façon de voir la présence de liens aux non entre les données, ainsi que la force de ces liens.
Ici on voit que les variables sont corrélées deux à deux, donc on procède par élimination des données les plus corrélées entre elles car elles donnent la même information puis on s’en serre des données les moins possibles corrélées entre elles afin d’extraire plus facilement des différences.



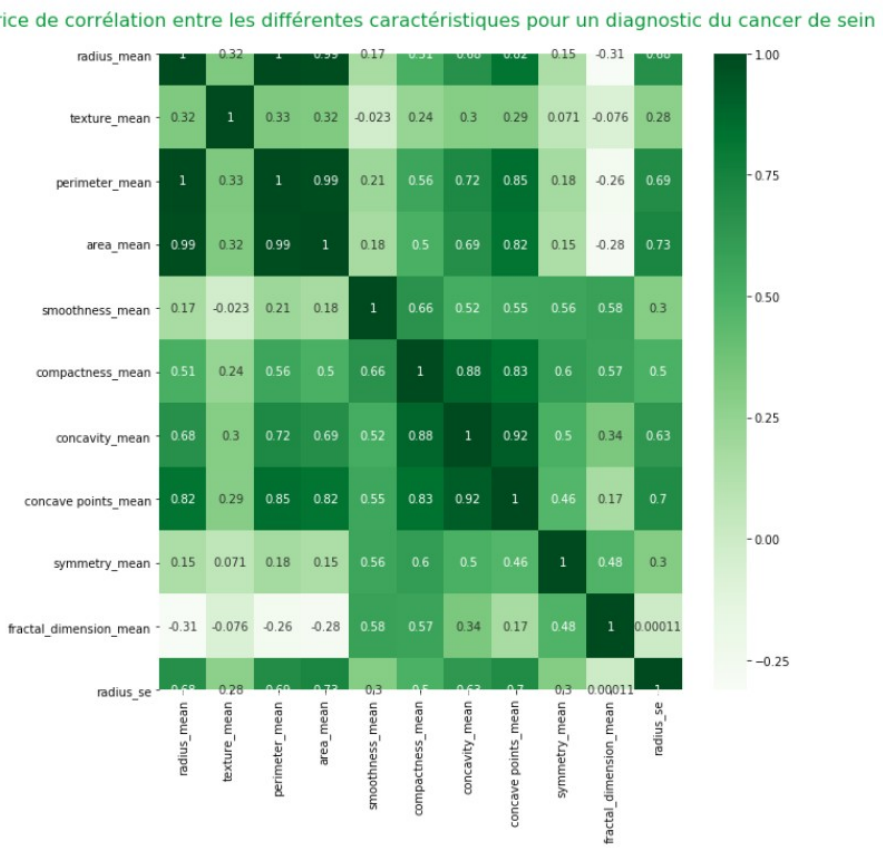
On a ici l'équivalence en python, tel que nous avons utilisé Seaborn on lui précisant comme argument .corr() et nous avons obtenu la matrice de corrélation suivante :



À première vue, il est plutôt compliqué de comprendre ce graphique c'est pourquoi nous avons apporté quelques modifications afin que ce soit plus compréhensible :

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

plt.figure(figsize=(10,10))
sns.heatmap(wdbc.iloc[:,1:12].corr(), annot=True, cmap='Greens')
plt.title("Matrice de corrélation entre les différentes caractéristiques pour un diagnostic du cancer de sein \n", fontsize=16,
```



Analyse des résultats :
en prenant les colonnes 8 et 4 qui représentent la concavité et le taux de concavité de la tumeur on voit qu'elles sont corrélées à 0.92 (très proche de 1) ce qui nous causerait donc des difficultés à les étudier 5(elle contient des informations similaires). Par ailleurs, les colonnes 2 et 3 représentants le rayon et la dimension fractale de la tumeur sont peu corrélés à -0.31, elles vont donc nous servir à choisir des axes lors d'une éventuelle projection dans l'espace des données pour trouver leur classe d'équivalence.
Un autre exemple des colonnes corrélées et non corrélées disponible sur le Slide 11

Correlation filter

[Slide12]

Cependant, en s'appuyant sur les résultats précèdent nous avons pu filtrer au mieux nos données grâce à l'outil « Correlation Filter » de KNIME avec comme indice de corrélation 0.9 ce qui nous a permis de garder que 20 colonnes dont l'indice de corrélation est inférieur à 90.

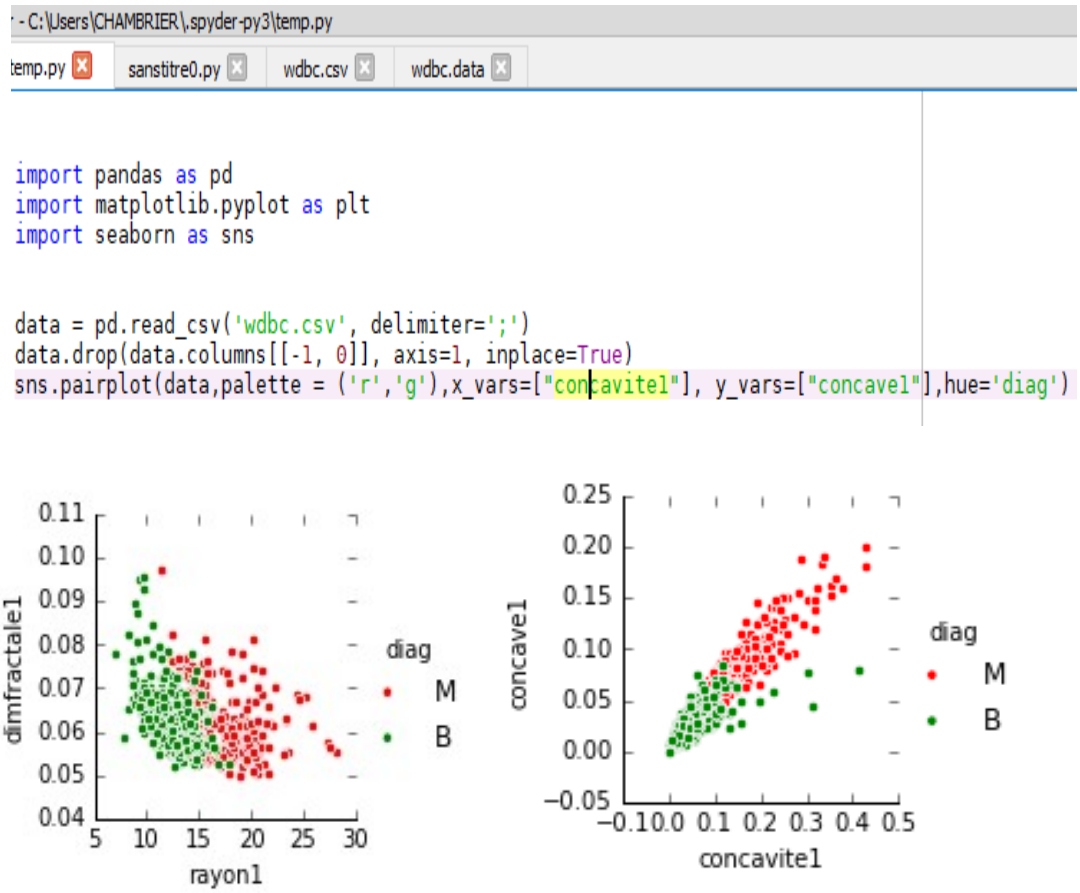
Scatter Plot

[Slide 13]

Nous avons passé pour scatter plot les colonnes 2 et 11 qui étaient corrélées a -0.312 on remarque que plus le rayon de la tumeur est élevé, plus le risque d'avoir une tumeur maligne augmente. Par ailleurs la dimension fractale importe peu sur le caractère malin ou bénin.

Scatter plot matrix

[Slide 14, 15, 16]

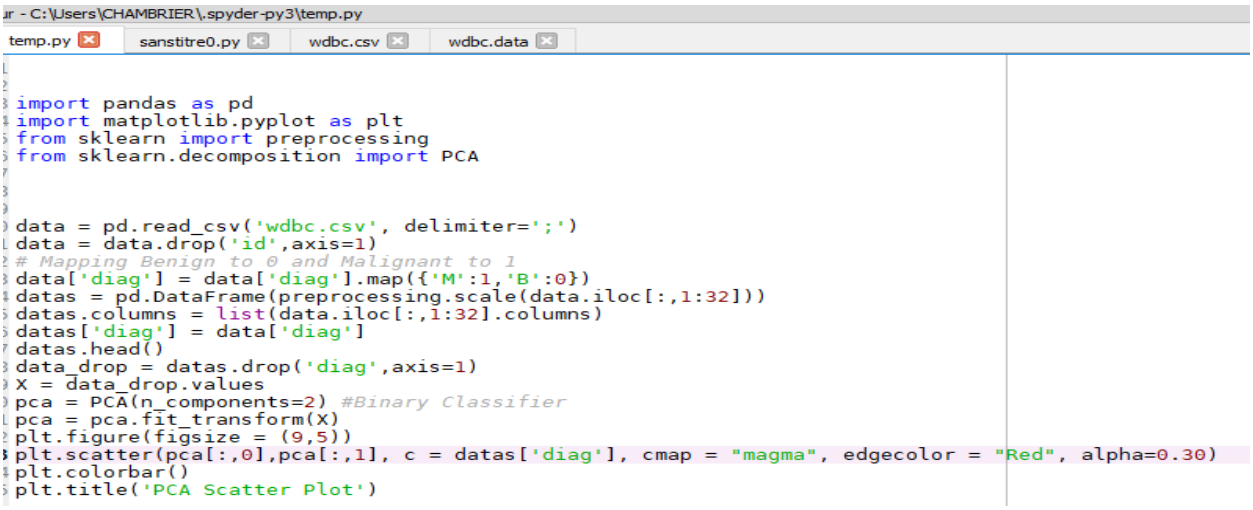


La visualisation des colonnes 8 et 9 et 2 et 11, le nuage de points montre que plus la concavité de la tumeur augmente, plus la personne à un risque d'avoir une tumeur maligne. En effet si le taux de concavité est entre 0 et 0.75, aucune tumeur n'est maligne, or si la concavité est entre 0.125 et 0.5, plus de 98% des tumeurs sont malignes.

Avec l'exemple à gauche nous observons que plus le rayon de la tumeur est élevé, plus elle a de risque d'être maligne. En effet entre 5 et 12 millimètres de rayon aucune tumeur n'est maligne et entre 17 et 30 elles sont tout malignes. En revanche la dimension fractale de la tumeur (colonne 11) impacte peu sur le caractère malin ou bénin de la tumeur. En effet, nous voyons que si la dimension fractale est entre 0.06 et 0.08, le risque que la tumeur soit maligne est similaire au risque qu'elle soit bénigne.

On arrive donc à distinguer deux classes grâce à différent choix de variables mais les données restes assez mélangées dans l'ensemble. Et il est assez difficile de donner une tendance.

PCA & MDS



[**Slide 17, 18]**

Dans le but de réduire le choix des dimensions pour une meilleure visualisation, nous avons décidé de faire une analyse en composante principale.
Ainsi, on obtient après la projection deux nouvelles colonnes PCA_dimension_0 et PCA_dimension_1 (On aurait pu aussi faire une deuxième option, c'est-à-dire garder un certain pourcentage d'information. Le problème est que nous entrons un problème de machine Learning pour trouver le meilleur pourcentage) une visualisation différente, plus précise dans l'étude des données. On voit ainsi une nette séparation des verts et des rouges. De plus, les verts sont assez concentrés car les caractéristiques des tumeurs bénignes sont assez similaires et on observe un net une séparation vers la droite et de manière assez dispatchée des points rouges, car les caractéristiques des tumeurs malignes sont propres à un individu.

CLUSTERING

[Slide 19]

Dans le but de trouver des groupes d'objets de manière automatique par un modèle d'apprentissage nous avons fait clustering en utilisant K-means, un outil qui permet de rassembler les données avec plus de similitude possible. D'après le diapositif 18 on voit que les données sont regroupées au milieu ce qui nous cause des difficultés à notre modèle d'apprentissage de savoir dans quel cluster la donnée se trouve. On constate alors que ce modèle n'est pas fiable et se peut tromper car en effet k-means se base sur la densité, la moyenne et les centres.

Hiérchicale clustering

[Slide 20, 21]

La méthode suppose qu'on dispose d'une mesure de dissimilarité entre les individus.

La classification ascendante hiérarchique est dite ascendante car elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes suivant leur proximité les unes des autres.

En remarque que pour une distance de 8000 on obtient deux clusters.

Partie prédiction

Decision Tree Predictor

[Slide 22, 23, 24]

Afin de prédire à partir d'une donnée sa classe, nous devons construire un modèle d'apprentissage. Pour cela, nous avons commencé par la répartition de nos classes en deux (70% pour chaque classe) grâce à l'outil "Partitionnig "

On obtient alors l'arbre de décision de slide 24

[Slide 25]

Prenons un exemple avec la donnée de la ligne 5 suivantes : On regarde la valeur à la colonne 24 correspondants au périmètre du noyau de la tumeur, on a 103.4 qui est inférieur à 105.14, on va donc à gauche puis on regarde la colonne 26 correspondants à la zone, elle est de 0.179 supérieurs à 0.1759 donc on va à droite

et il n'y a plus de colonne à regarder donc d'après notre modèle la tumeur est maligne. Ce que confirme la colonne 1.

SCORER & prédictions

[Slide 26, 27]

Cette étape a pour objectif de voir si un modèle d'apprentissage est fiable ou pas, plus il a un scorer élevé plus il est considéré fiable.

Le score obtenu est de 90.641 un indice élevé, on peut dire que notre modèle est valide.

CROSS validation

[Slide 28]

Avec cet outil de prédiction notamment à estimer la fiabilité du modèle d'apprentissage fondé sur la méthode d'échenillage, nous avons pour cela partitionné nos données en 10 afin d'extraire le meilleur apprentissage parmi ces dernières.

Puis à l'aide d'un scorer nous avons trouvé un taux de réussite de 92,091%, un résultat proche de la valeur de prédiction (90.643 %) faite avant ce qui montre que notre modèle est valide.

[Slide 29]

On peut encore améliorer ce modèle en utilisant l'option "Leave-One-Out". Ce modèle sera le meilleur modèle de validation car il apprend sur toutes les données, il en prend une, puis la met de côté, il apprend sur le reste, On a donc 600 prédictions. On trouve un modèle pas très loin de ce qu'on avait déjà trouvé avec 92,267 %.

[Slide 30]

L'échantillonnage aléatoire et linéaire nous donne un résultat de 93.146 un peu mieux que ceux d'avant.

Naive Byes

[Slide 31]

Nous avons utilisé cet outil de prédiction dont la particularité est de supposer l'existence d'une caractéristique spécifique pour identifier les classes.

Nous avons obtenu un pourcentage de prédiction à 93,767%, on remarque qu'il élevé par rapport à ce que nous avons obtenue avec les autres outils.

On déduit que Naive Byes est plus fiable.

Déduction

[Slide 34]

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

Conclusion

L'application de l'analyse de données à la médecine offre une perspective essentielle à la prévention et le traitement de nombreuses maladies.

Cette innovation tend vers un seul et unique objectif « combattre la mort et la maladie »