Université Iba Der THIAM de Thiès

UFR Sciences et Technologies

Département Mathématique

Master Sciences de Données et Applications option Statistiques et Econométrie

Option : Statistique Econométrie et Modélisation

Projet d'Apprentissage statistiques

Présenté par :

Ismaël YODA & Amsatou DIOP

Nom du Professeur :

Dr Lucien GNING

Année Scolaire 2020 & 2021

Table des matières

I- Description du projet

Le cancer du sein est le cancer le plus fréquent chez les femmes dans le monde. Il représente 25 % de tous les cas de cancer et a touché plus de 2,1 millions de personnes en 2015 seulement.

Nous disposons des données sur le cancer des femmes. Notre étude a pour objectifs de comprendre l'ensemble de données, le nettoyer (si nécessaire) et ensuite de créer des modèles de classification pour prédire si le les tumeurs sont des tumeurs malignes (cancéreuses) ou bénignes (non cancéreuses). Nous allons donc construire pour cela un modèle d'apprentissage automatique qui permettra de classer les tumeurs selon qu'elles sont cancéreuses ou pas.

II- Description des données

1- Attributs et types des données et leurs propriétés statistiques

Toutes nos données sont de types décimales (float64) a l'exception de la variable diagnostic qui est de types object. Etant donné que la variable diagnosis est censé être une variable binaire, nous allons la recoder de telle sorte que la variable prenne la valeur 1 si la tumeur est cancéreuse et 0 si elle ne l'est pas. Nous allons ensuite changer le type de la variable en une variable catégorielle.

Par ailleurs, on constate qu'il y'a un grand écart entre les variances des features. Par exemple les variances des variables perimeter_worst(33,602542) , area_mean(351,914129) , area_worst (569,356993) sont beaucoup plus élevé que les variances des variables comme concave points_worst(0,065732), symmetry_worst(0,061867) et fractal_dimension_worst(0,055040). Le même constat est fait pour le cas de la moyenne. En plus, On constate également à travers les boxplots de nos features que certaines variables possèdent des outliers.

La matrice de corrélation des features, nous montre une très forte corrélation entre mean radius et mean perimeter, mean radius et mean area, mean perimeter et mean area.

2- Les difficultés qui se présentes

a- La normalisation et la standardisation et des données

La différence d'échelle et la présence d'outliers dans certaines features peut conduire à des performances moindres de notre algorithme de classification. Pour pallier à cela, nous allons normaliser et standardiser nos données afin de réduire l'espace de variation des features et par conséquent de réduire les outliers. Pour cela, nous allons faire le **Feature Scaling** qui comprend la **Standardisation** et la **Normalisation**.

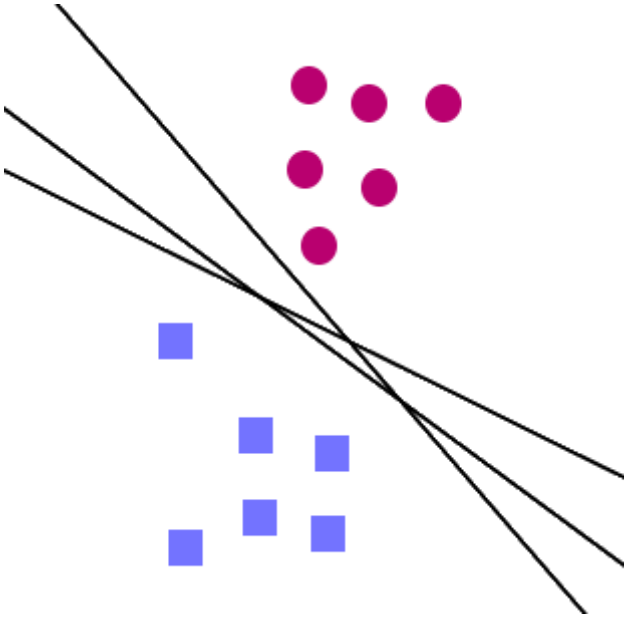
b- Fléau de dimensionnalité

Pour faire une représentation graphique, il faudra prendre en compte trente dimensions car nous avons trente variables explicatives. La visualisation de nos données sera donc très difficile car au-delà de deux dimensions, il est difficile de bien visualiser nos données. Nous nous proposons alors d'effectuer une réduction de la dimension de nos données afin d'avoir une meilleure visualisation. Cela nous permettra d'avoir un aperçue sur le type de modèle approprié pour classifier nos données (linéaire ou non linéaire). Pour pallier à ce problème de dimension, nous allons utiliser l'algorithme **PCA(Principal Component Analysis).**

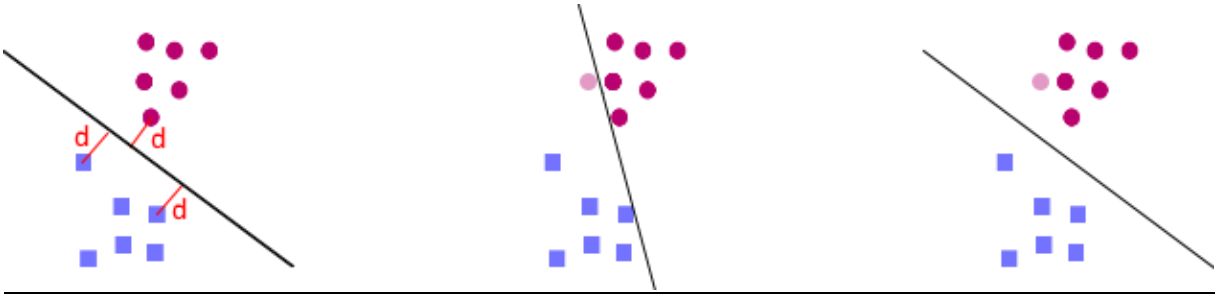
III- Traitement des données

Nous avons choisi d'implanter l'algorithme du Support Vector Machine (SVM).

Le but, pour un SVM, est d'apprendre à bien placer la frontière entre deux catégories. Cependant, la principale difficulté est de trouver la frontière optimale car quand on a un ensemble de points d'entraînement, il existe plusieurs lignes droites qui peuvent séparer nos catégories. La plupart du temps, il y en a une infinité comme l'illustre la figure ci-dessous

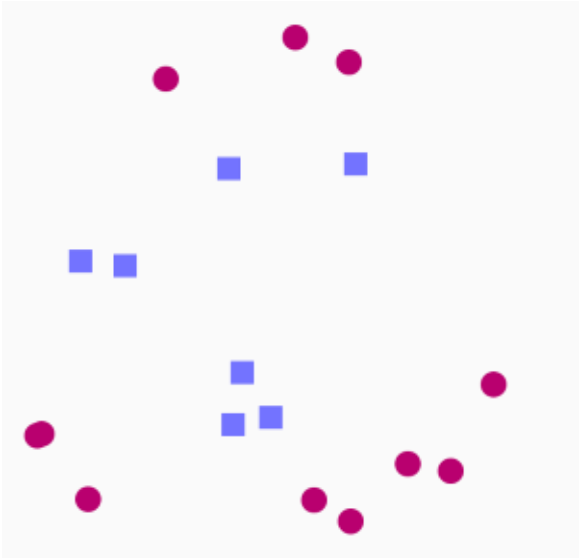


De façon Intuitive, on se dit que si nous avons un nouveau point, très proche des ronds rouges, alors il est très probable que ce point soit un rond rouge lui aussi. Inversement, plus un point est près des carrés bleus, plus il est probable qu'il soit lui-même un carré bleu. Pour cette raison, un SVM va placer la frontière aussi loin que possible des carrés bleus, mais également aussi loin que possible des ronds rouges. La frontière optimale est donc celle la plus éloignées des points des deux groupes comme le montre l'image de droite représenté dans la figure ci-dessous. On dit qu'elle a la meilleure **capacité de généralisation**



Les exemples ci-dessus représente les cas où les données peuvent être séparées par des lignes droites pourtant, dans la plupart des cas les données ne sont pas séparables avec des lignes droites.

Considérons l'exemple suivant :



Puisque les carrés sont entourés de ronds de toute part, il est impossible de trouver de ligne droite qui soit une frontière : on dit que les données d'entraînement ne sont pas **linéairement séparables**. Cependant, imaginons qu'on arrive à trouver une transformation qui fasse en sorte que notre problème ressemble à ça :

A partir de là, il est facile de trouver une séparation linéaire. Il suffit donc de trouver une transformation qui va bien pour pouvoir classer les objets. Cette méthode est appelée **kernel trick**, ou astuce du noyau en français. Vous vous en doutez, toute la difficulté est de trouver la bonne transformation.

Les composantes clefs du modèles SVM sont :

Les vecteurs support : Ce sont les points d'entraînement les plus proches de la frontière sont**.**

L'**hyperplan séparateur :** c'est la frontière de séparation entre les classes

L'**hyperplan optimal :** c'est l'hyperplan à marge maximale

La séparabilité linéaire : On dit qu'un jeu de données est **linéairement séparable** s'il existe au moins un hyperplan tels que tous les points positifs soient d'un coté de cet hyperplan et que tous les points de l'autre.

Marge d'un hyperplan séparateur : La marge d'un hyperplan séparateur est la distance de cet hyperplan l'observation du jeu d'entrainement la plus proche.

Le choix du modèle SVM est motivé par le fait en raison de sa capacité à travailler avec des données de grandes dimensions, ses garanties théoriques et les bons résultats réalisés en pratique. De plus, ce modèle requiert un faible nombre de paramètres et il est simple d'usage.

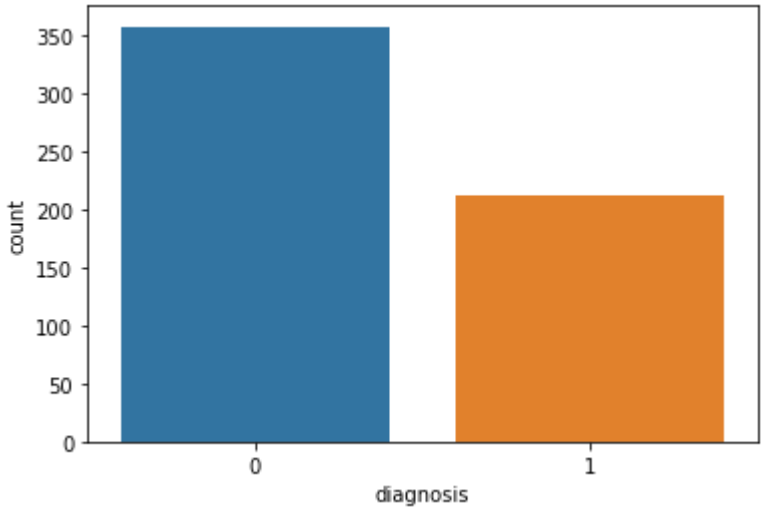
Nous avons considéré l'alternative de la régression logistique pour faire notre classification mais ce modèle s'est avéré moins performante que le modèle SVM car ce dernier affiche un score de 0,99 alors que le modèle logistique a un score de 0,96.

IV- Résultats final

Tableau 1 : Statistiques descriptives

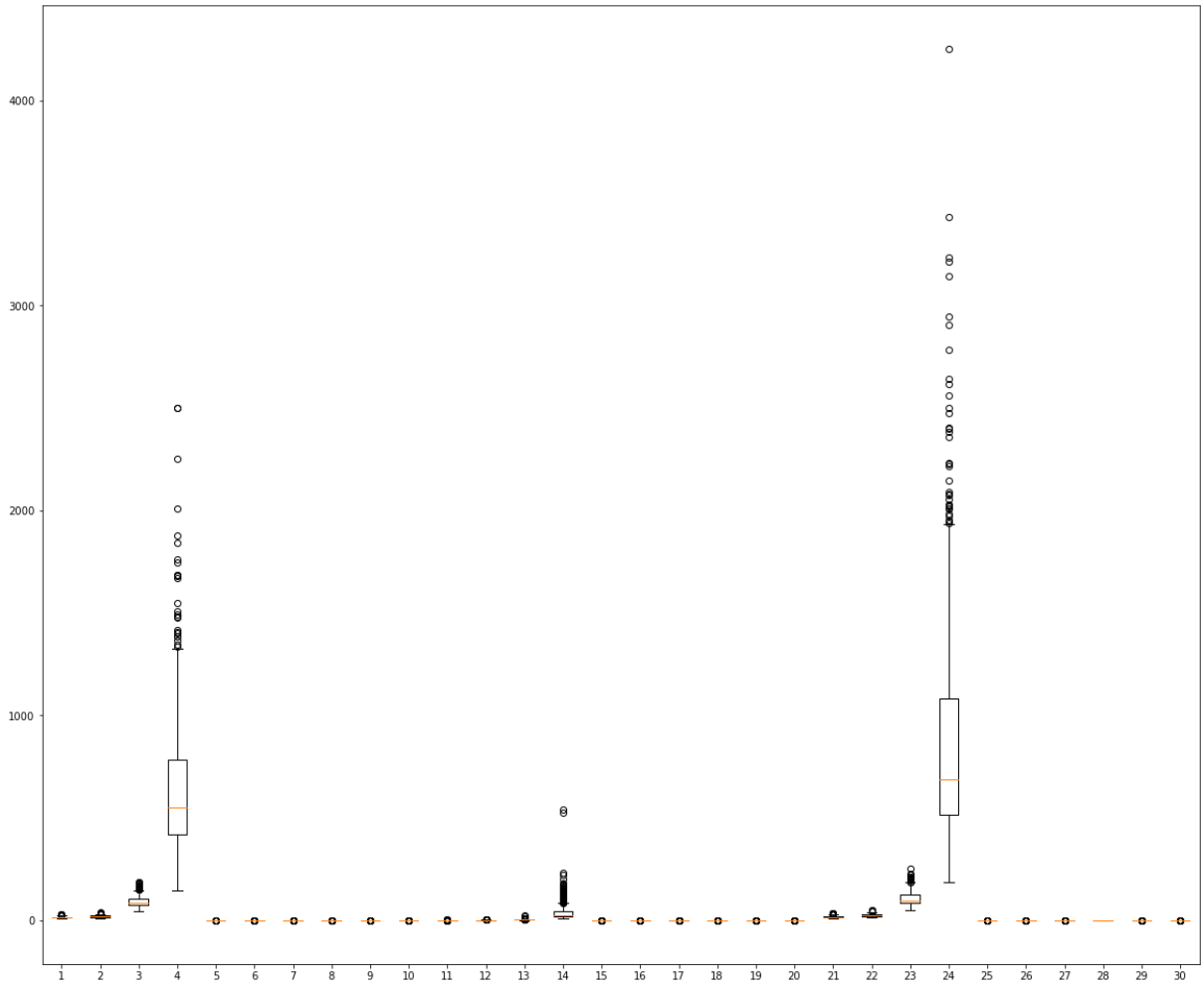
	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	conc poin
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.0
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201

Figure 1 : Répartition des céllules cancéreuses et non cancéreuses



L'effectif des cellules cancéreuses est de 212 soit 37,25% et celui des cellules non cancéreuses est de 357 soit 62,74%.

Figure 2 : Boxplot des variables quantitatives



Nous constatons à travers la figure 2 qu'il existe des données aberrantes dans certaines variables. La normalisation et la standardisation des données que nous effectuerons contribuera résorber ce problème.

Figure 3 : Matrice de corrélation des variables explicatives

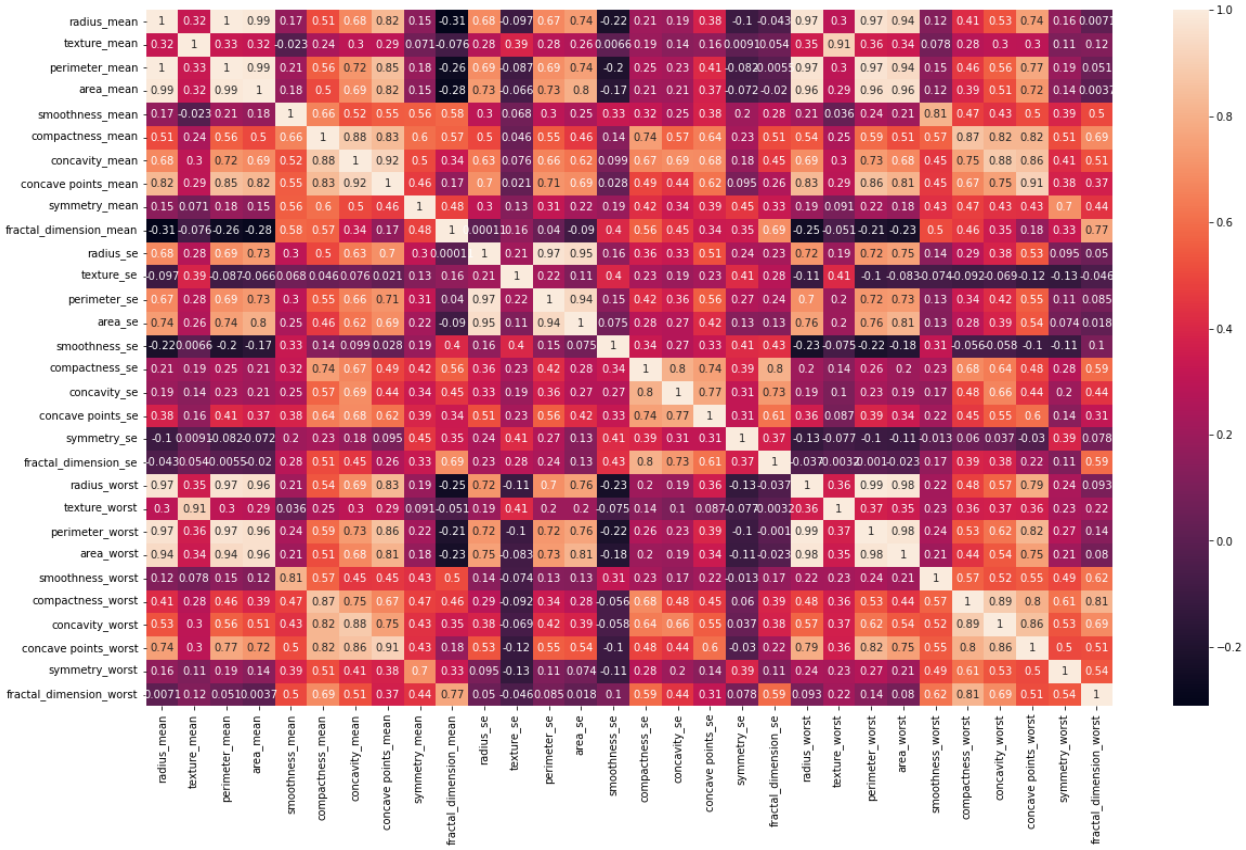
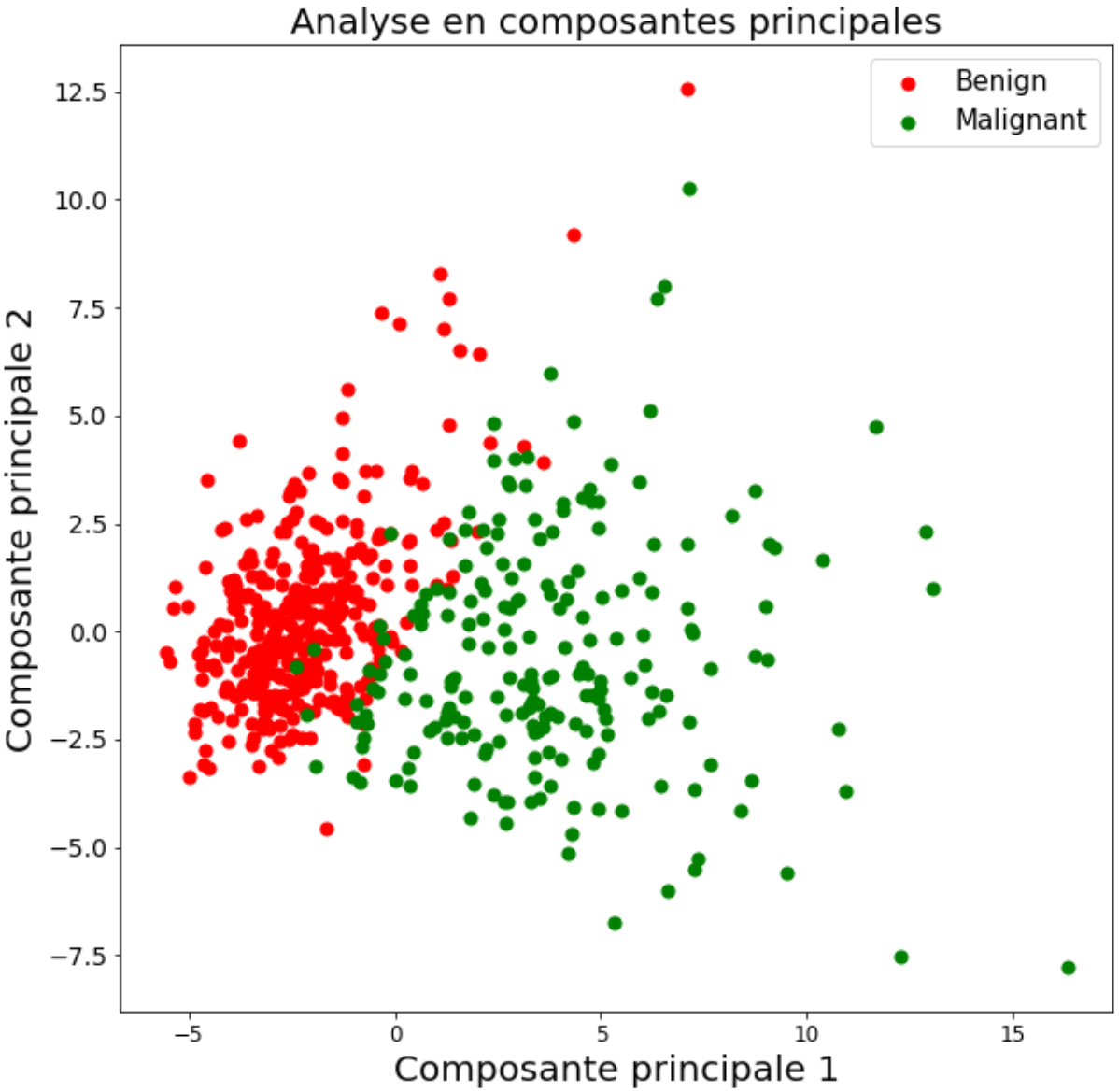

radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave points_worst
symmetry_worst
fractal_dimension_worst
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Tableau 2 : Aperçue des deux dimensions réduites

	principal component 1	principal component 2
564	6.439315	-3.576817
565	3.793382	-3.584048
566	1.256179	-1.902297
567	10.374794	1.672010
568	-5.475243	-0.670637

La composante principal 1 explique 44,27% de l'inertie total alors que composante principal 2 explique 18,97% l'inertie total. Les deux composantes principales expliquent donc 63,24% de l'inertie totale. Bien que le pourcentage d'inertie des deux axes principales n'atteigne pas 90%, leur pourcentage d'inertie est supérieur à 60% et la représentation des données sur ces deux dimensions nous permettra d'avoir une idée sur la répartition des données dans l'espace de dimension 2.

Figure 4 : Représentation graphique des données sur deux dimensions.



La figure nous montre que la séparabilité linéaire de nos données n'est pas si évident que cela.

Certains points Malignants se retrouve encerclés par les points de l'autre classe (Benign).

Une séparation linéaire de nos données pourrait donc entrainer beaucoup d'erreur de classification.

Nous allons donc privilégier un SVM avec un noyau non linéaire notamment un Rbf (Radial basis function).

Classification obtenue avec un svm non linéaire avec un noyau rbf

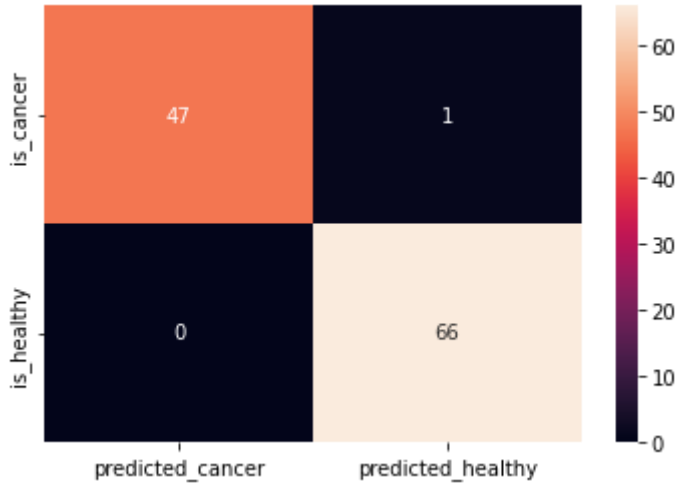
Tableau 3 : Statistiques pertinentes de notre modèle

	précision	rappel	f1-score	Nombre d'individus
0	0,99	1	0,99	66
1	1	0,98	0,99	48

Etant donné que notre classe de référence est la classe des tumeurs cancéreuses (recodé 1), les statistiques qui nous intéresse le plus sont celles coloriées en rouge.

Au vu des résultats, nous pouvons conclure que les prévisions positives sont précises à 100% c'est-à-dire que toute les tumeurs prédites par notre modèle comme cancéreuses sont réellement cancéreuse. Aussi, le taux de vrai positif est de 98%.

Tableau 4 : Matrice de confusion



Les résultats consignés dans la matrice de confusion nous montrent que les performances de notre modèle sont bonnes. En effet, le nombre de faux positifs est de 0 c'est-à-dire qu'il y'a 0 tumeurs prédites cancéreuse mais qui sont non cancéreuses en vrai. Cela signifie que notre modèle prédit sans aucune erreurs les tumeurs non cancéreuses.

En plus, le nombre de faux négatif est de 1 c'est-à-dire qu'il y'a une seule tumeur prédite non cancéreuses mais qui est cancéreuses en vrai. Notre modèle prédit donc avec une seule erreur les tumeurs cancéreuses.

Conclusion : Notre modèle est plus performant dans la prévision des tumeurs non cancéreuses que dans la prévision des tumeurs cancéreuses. Cela pourrais s'expliquer par le fait qu'il y'ai plus de tumeurs non cancéreuses que de tumeurs cancéreuses. Il est plus facile pour le modèle de se tromper sur la classe moins nombreuse(cancéreuse) que la classe la plus nombreuse (non cancéreuse).