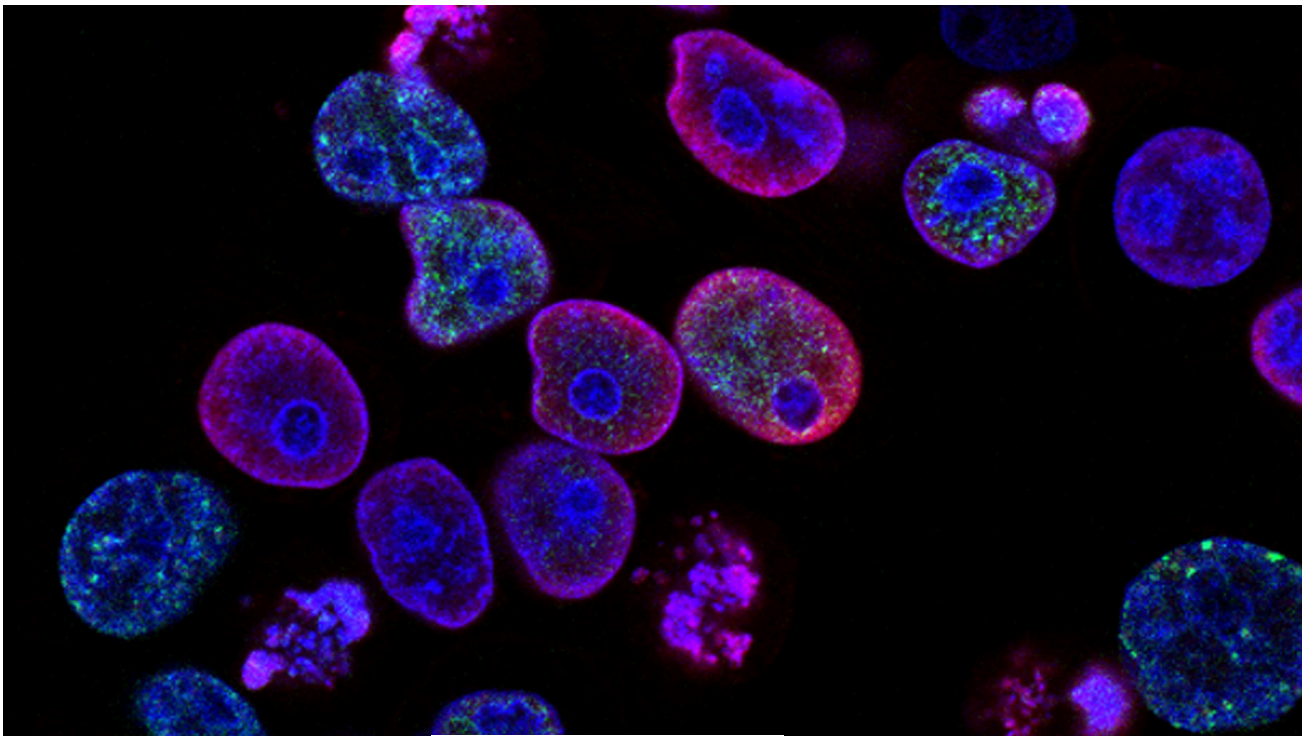
Master Bio-informatique et
Génomique Environnementale

Rapport de fouille :

La fouille de donnée au service de la médecine



Yves Pommier, Rozenn Josse, 2014

Antonin MÉNARD
Florent LE QUELLEC

Introduction **3**
Matériel et Méthode **3**
Résultats **5**
Analyse exploratoire 5
Adenopathy 6
Focality 7
Gender 7
Hx Radiotherapy 7
Métastase (M) 7
Ganglion touché (N) 7
Taille (T) 7
Pathology 8
Response 8
Risk 8
Stage 8
Clustering 11
Classification 13
Conclusion **14**
Bibliographie **15**
Annexe **17**
Annexe 1 : Tableau description des variables utilisées 17
Annexe 2 : Tableau classification TNM [6] [7] 17
Annexe 3 : Tableau classification des cancers par stade [8] 18
Annexe 4 : Caractéristique des Clusters 19
Annexe 5 : Diagrammes en barre des distributions des modalités de nos variables. 20
Annexe 6 : Graphique silhouette 21

Introduction

Une tumeur se caractérise par la prolifération anormale de cellules dont le système de division est déréglé, conduisant à la formation d'une masse. Cette tumeur peut être bénigne (localisée, sans caractère invasif) ou maligne (capable d'envahir les tissus voisins ou de se propager à distance). Lorsqu'une tumeur est maligne, on parle alors de tumeur cancéreuse, ou plus simplement de cancer.[1][2]

Dans notre cas, nous nous intéressons au cancer de la thyroïde, qui représente environ 5 % des tumeurs thyroïdiennes, et qui est essentiellement diagnostiqué chez des femmes [3]. Ce cancer reste relativement rare en France, mais son incidence est en augmentation constante depuis plusieurs années, notamment grâce à l'amélioration des techniques de détection. D'après la Société canadienne du cancer, sur 6 600 nouveaux cas de cancer de la thyroïde diagnostiqués en une année, environ 280 décès sont enregistrés, soit un taux de mortalité de 4,2 %. Il s'agit en réalité d'un cancer au pronostic globalement favorable, mais qui nécessite un suivi attentif en raison des risques de récidive. [4]

Grâce à la fouille de données, il devient possible d'explorer de grands ensembles de données cliniques afin de détecter des profils à risque, des relations cachées entre variables, ou encore de prédire des issues médicales comme dans notre cas : la récidive. Cela permet de transformer des données statiques en connaissances utilisables pour la recherche médicale.

Ce projet a pour objectif d'explorer un jeu de données clinique regroupant 383 patients atteints de cancer de la thyroïde, dans le but de :

- Identifier les variables les plus fortement liées à la récidive
- Regrouper les patients selon des profils communs (clustering)
- Tester des modèles de classification pour prédire la récidive

Pour répondre à ces objectifs, une série d'analyses ont été menées : analyse exploratoire des données, application d'une Analyse des Correspondances Multiples (ACM) pour la réduction de dimension, clustering non supervisé par k-means pour détecter des groupes de patients similaires, et enfin des modèles supervisés (arbre de décision, forêt aléatoire, modèle bayésien) ont été évalués pour prédire la récidive.

Matériel et Méthode

Notre jeu de données provient du site Kaggle. Ce jeu de données porte sur la récidive du cancer de la thyroïde après un traitement par iode radioactif (RAI). Il regroupe les informations cliniques de 383 patients, décrites à travers 13 variables clés, telles que l'âge, le sexe, le type de pathologie, le stade du cancer, la classification du risque, la réponse au traitement et la survenue éventuelle d'une récidive.

Ces données sont particulièrement utiles pour prédire le risque de récidive, mieux comprendre les facteurs de risque associés et évaluer l'efficacité des traitements administrés.

Le jeu de données comprend 13 variables décrivant les caractéristiques cliniques, pathologiques et thérapeutiques de 383 patients atteints d'un cancer de la thyroïde. Parmi ces variables, une seule est quantitative : l'âge du patient. Les douze autres sont qualitatives et couvrent des aspects variés tels que le sexe, les antécédents de radiothérapie, la présence de ganglions lymphatiques atteints (Adenopathy), le type histologique de la tumeur (Pathology), la focalité tumorale, ou encore la classification TNM (T, N, M). On y trouve également des variables résumant le stade global du cancer (Stage), la réponse au traitement (Response), et enfin la présence ou non d'une récidive (Recurred), qui constitue la variable cible dans les analyses de classification. L'ensemble de ces variables permet d'analyser en profondeur les profils de patients et d'explorer les facteurs potentiels de récidive. Pour plus d'information, consulter les tableaux annexes 1, 2 et 3.

Les données ont été traitées à l'aide de R (version 4.4.2) et de RStudio pour la phase d'analyse exploratoire et de clustering. Les packages utilisés incluent :

- tidyverse (Wickham et al., 2019), ggplot2 (Wickham, 2016) et patchwork (Pedersen, 2024) pour la manipulation des données et la création de visualisations graphiques,
- Cluster (Maechler et al., 2025), FactoMineR (Lê, Josse & Husson, 2008) et factoextra (Kassambara & Mundt, 2020) pour la réalisation de l'analyse des correspondances multiples (ACM),
- plotly (Sievert, 2020) pour les visualisations interactives en trois dimensions,

L'ensemble des données a été soumis à plusieurs étapes de prétraitement avant l'analyse :

1. Conversion des variables catégorielles au format factor afin de permettre une analyse statistique adaptée,
2. Analyse univariée par la création de statistiques descriptives, d'histogrammes pour les variables numériques et de diagrammes en barres pour les variables catégorielles,
3. Réalisation de tests de normalité (Shapiro-Wilk) et de variance (F-test), ainsi qu'un test de Wilcoxon sur la variable Âge afin de comparer les groupes avec et sans récidive,
4. Application du test du Chi² pour évaluer l'indépendance entre les variables qualitatives et la variable cible Recurred,
5. Transformation des données au format long pour les représentations graphiques avec ggplot2,
6. Réalisation d'une ACM afin de réduire la dimensionnalité des données catégorielles en amont du clustering et réalisation d'un diagramme de silhouette pour représenter le nombre idéal de groupements de patients pour nos analyses, sur base d'un calcul de score de silhouette qui détermine la qualité de ces groupements,
7. Clustering non supervisé par la méthode des k-means sur les premières dimensions issues de l'ACM,
8. Interprétation des clusters par croisement avec les variables cliniques et la variable cible.

La phase de classification supervisée a été effectuée à l'aide de KNIME (version 5.4.3). À partir du fichier de données complet, plusieurs modèles ont été testés. Une

validation croisée a été mise en place à l'aide des nœuds Partitioning (pour séparer les jeux d'apprentissage et de test) et X-Aggregator (pour compiler les résultats). Cela a permis d'évaluer la robustesse des modèles selon des métriques telles que le taux d'erreur moyen, le nombre moyen d'erreurs et la taille moyenne du jeu de test.

Résultats

Analyse exploratoire

Répartition de nos variables catégorielles (annexe 5)

Bien que notre nombre d'individus soit d'une taille respectable compte tenu des analyses qui vont être réalisées dessus, il demeure certains déséquilibres dans la répartition des modalités de nos données.

Les modalités présentes en majorité sont celles indiquant que les patients n'ont pas subi de radiographie Hx, n'ont pas d'adénopathie, ont un cancer à un niveau de développement intermédiaire, n'ont pas de métastase ou de tumeur qui aurait atteint au moins un ganglion, ont un cancer unilocal, une excellente réponse au traitement et un risque de propagation du cancer faible (un nombre inférieur, mais non négligeable d'individus possèdent un risque intermédiaire).

Tous ces éléments nous indiquent que la majorité de nos patients sont atteints de cancers relativement peu développés et qui réagissent bien au traitement, en lien avec le taux relativement faible de récidive.

En outre, la majorité de nos individus sont atteints d'un cancer de type papillaire et sont des femmes (en conformité avec le sex ratio des patients atteints d'un cancer de la thyroïde), bien que notre échantillon d'individus hommes soit d'une taille acceptable (312 contre 71).

Répartition de nos variables selon la récidive

***Figure 1** : Diagramme en barres empilées de répartition des modalités des variables catégorielles selon la présence ou non d'une récidive*

La figure 1 est une représentation de la distribution des différentes modalités pour chacune de nos variables catégorielles en fonction d'une récidive ou non du cancer (Recurred). Ce sont des barres empilées exprimées en pourcentage, afin d'observer la proportion relative de récidive pour chaque modalité.

Adenopathy

Les patients avec récidive sont présents chez 100% des patients des groupes "Extensive", "Posterior" et sont très représentés dans les groupes "Left", "Bilateral" et "Right, tandis que l'absence d'adénopathie ("No") est majoritairement associée à l'absence de récidive.

Cela suggère un lien fort entre le fait d'avoir une adénopathie et le risque de récidive, certaines adénopathies semblant être associées à des risques plus élevés de récidive.

Focality

Les cas de tumeurs multi-focaux présentent une proportion nettement plus élevée de récidive que les cas uni-focaux. Cela renforce l'idée qu'un cancer multifocal est plus agressif et par conséquent, plus difficile à éradiquer complètement.

Gender

Les hommes semblent avoir une fréquence de récidive plus élevée que les femmes. Ceci pourrait indiquer qu'il existe un facteur de risque lié au sexe.

Hx Radiotherapy

Les patients ayant déjà subi une radiothérapie ont plus de récidives que ceux qui n'en ont pas eu. Cela pourrait refléter un cancer plus avancé au départ ou une résistance accrue. Il est cependant important de prendre en compte qu'une forte majorité des patients n'ont pas eu de radiothérapie.

Métastase (M)

La présence d'une métastase est associée fortement à la récidive. Ce résultat est en cohérence avec le fait que ce cancer soit à un niveau de développement plus avancé

Ganglion touché (N)

Le pourcentage de récidive est faible pour les patients N0, moyen pour les patients N1a et élevé pour les patients N1b. Pour rappel, le taux de récidive sur l'ensemble des patients est de 28%, posséder un cancer N1b est donc un facteur plus fréquemment associé à la récidive.

Ce résultat met en évidence l'importance du statut ganglionnaire dans le risque de rechute.

Taille (T)

On observe une augmentation progressive du risque de récidive avec le développement du cancer. Les tailles tumorales les plus petites (T1 et T2) sont associées à peu de récidives, contrairement aux tumeurs de grande taille (T3-T4) qui montrent des fréquences élevées de récidive (T3a possède un risque de récidive intermédiaire).

Pathology

Les cancers papillaires, folliculaires et les cancers à cellules de Hürthle sont davantage associés à la récidive (aux alentour de 35% des cas pour les trois cancers). Aucune récidive de cancer n'a été observée chez les patients atteints d'un cancer micro-papillaire.

Response

Une réponse structurale incomplète au traitement est fortement associée à la récidive. Une réponse biochimique incomplète est associée à la récidive dans près de la moitié des cas. Pour rappel, le taux de récidive sur l'ensemble des patients est de 28%, une réponse biochimique incomplète est donc un facteur plus fréquemment associé à la récidive. Une

réponse excellente aux traitements est associée à un taux de récidive très faible, conformément aux attentes cliniques.

Risk

Tous les patients à haut risque ont récidivé, tandis que ceux à faible risque récidivent rarement (4.8%). Les patients ayant un risque intermédiaire sont associés à un risque de récidive situé entre ces deux valeurs, mais l'association demeure assez forte. Ces résultats montrent l'existence d'une association positive entre le risque de propagation et le risque de récidive, conformément à ce que l'on pourrait attendre d'un cancer plus agressif.

Stage

On constate une augmentation du risque de récidive avec l'évolution du cancer au fil des stades, en cohérence avec nos résultats sur les classifications TNM. Le stade I est le seul associé à un risque de récidive assez faible (19.5%), le stade II est associé majoritairement à des patients ayant connu une rechute, et les stades III, IVB, IVA sont associés à de la récidive dans 100% des cas.

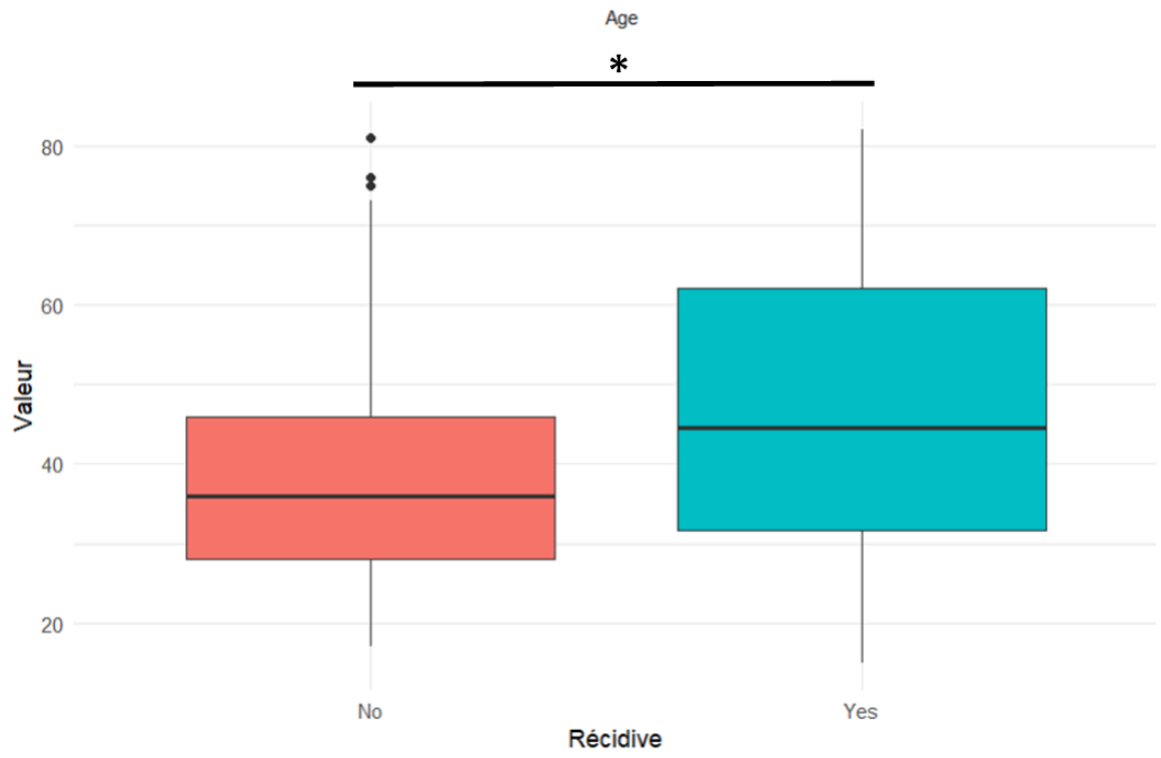


Figure 3 : *Distribution de l'âge des patients selon la survenue d'une récidive*

Les âges du groupe ayant connu une récidive et ceux ayant connu une récidive ne suivant pas une loi normale et n'étant pas homoscédastique (même après une transformation log et carré), nous avons donc réalisé test non paramétrique de Wilcoxon-Mann-Whitney unilatéral pour comparer les moyennes des âges de ces deux groupes.

Ce dernier nous a indiqué que les individus ayant connu une récidive sont en moyenne plus âgés que ceux n'en ayant pas connu (p-value = 1.777e-05, voire la figure 3).

Ces résultats pouvaient être attendus dans la mesure où des individus plus âges pourraient avoir des organismes plus fragiles qui auraient plus de difficultés à lutter contre un cancer.

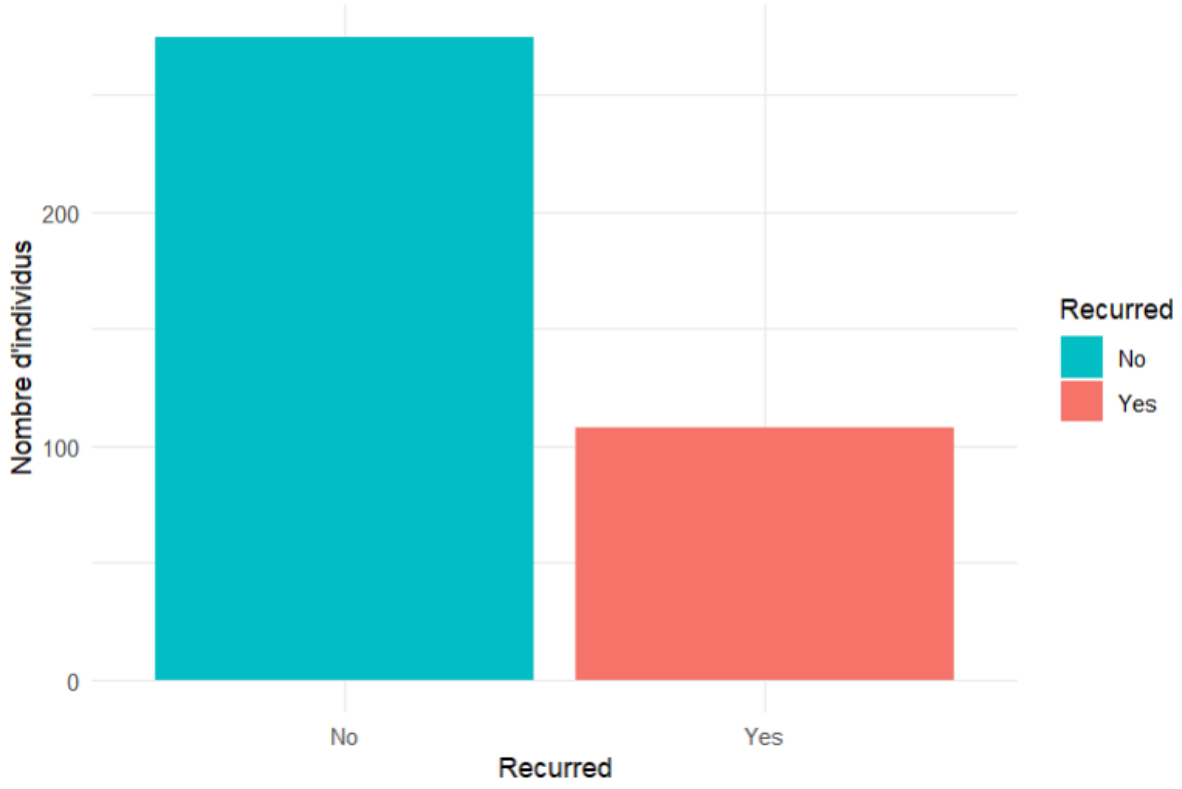


Figure 3 : *diagramme en barres du nombre d'individus ayant connu une récidive et n'en ayant pas connu.*

Le taux de récidive sur l'ensemble des patients est de 28%. 72% des patients n'ont donc pas de récidive, comme visible sur la figure 3.

Tableau 1 : Résultats du test du Chi² entre les variables catégorielles et la récidive

	variable	p_value	statistic
1	Response	8.863124e-67	309.472321
2	Risk	4.507816e-46	208.826203
3	N	5.443985e-34	153.186764
4	Adenopathy	4.220721e-32	157.044293
5	T	5.353537e-28	141.290246
6	Stage	3.161073e-20	97.617971
7	Focality	1.446380e-13	54.641643
8	M	2.616837e-11	44.444406
9	Gender	3.458852e-10	39.396676
10	Pathology	3.546645e-05	23.270435
11	Hx Radiothreapy	2.795966e-03	8.936051

Ce tableau présente les résultats du test du Chi² d'indépendance appliqué à chaque variable catégorielle du jeu de données en relation avec la variable cible Recurred (récidive). Pour chaque variable, sont indiquées la statistique de test (statistic) et la valeur p associée (p_value), permettant d'évaluer l'existence d'une dépendance significative entre la variable en question et la survenue d'une récidive.

Ces résultats confirment statistiquement les observations issues de l'analyse exploratoire : certaines variables comme la réponse au traitement, la classification du risque, ou le statut ganglionnaire (N) sont fortement liées à la récidive du cancer de la thyroïde. Ces variables constituent ainsi des candidats solides pour la construction de modèles de classification ou la définition de sous-groupes à risque.

Clustering

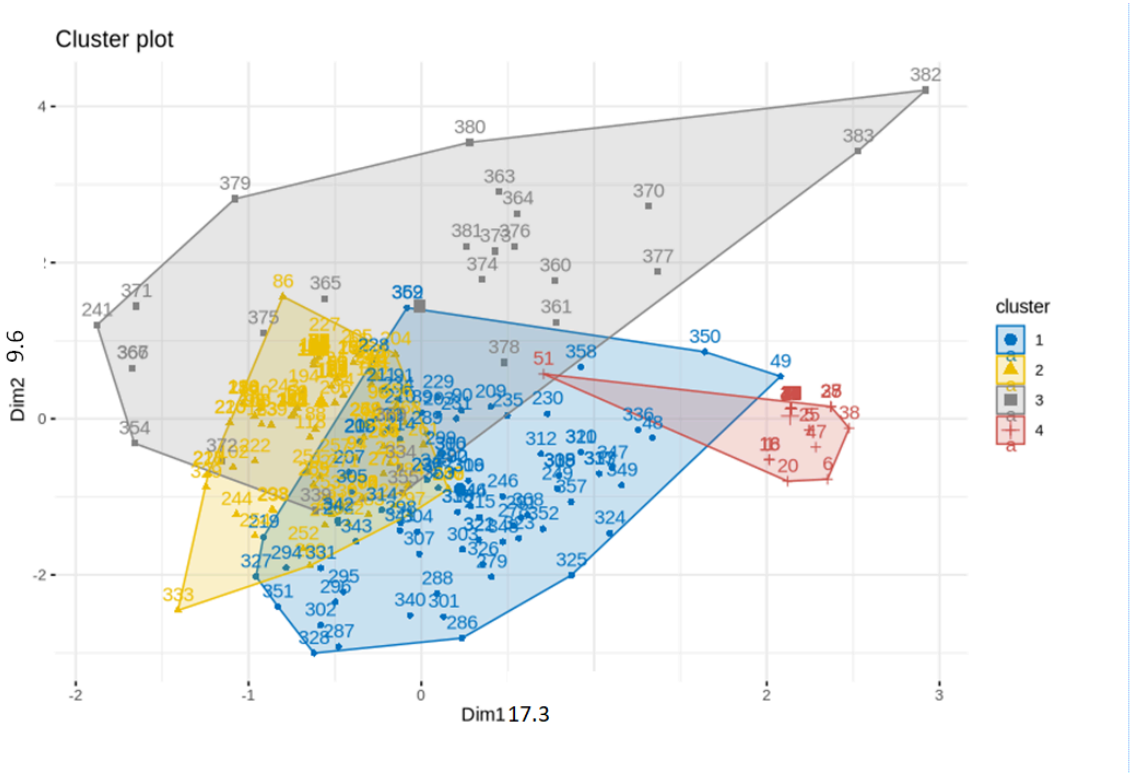


Figure 4 : *Visualisation des clusters de patients projetés sur les deux premiers axes de l'ACM*

La figure 4 présente la visualisation en deux dimensions des clusters obtenus via k-means appliqué aux coordonnées issues de l'Analyse des Correspondances multiples (ACM). Chaque point représente un individu (patient), positionné selon les deux premiers axes de l'ACM, qui expliquent ensemble 26,8% de la variance totale, ce qui peut expliquer le chevauchement des clusters. Une version en 3D permettant de prendre en compte tous les axes est disponible sur le gitlab Les patients sont colorés selon leur cluster d'appartenance, les zones convexes délimitant les regroupements. Le nombre de clusters a été déterminé à la suite d'un test de silhouette qui indiquait un nombre de clusters optimal entre 4 et 8. Dans un souci de simplicité d'interprétation, nous avons choisi de ne garder que 4 clusters (voire l'annexe 6).

Analyse des clusters (annexe 4) :

- Cluster 1 (bleu - points) : Ce cluster illustre un profil ambigu : malgré un stade I et une absence de métastases, la combinaison d'un ganglion atteint (N1b), d'une tumeur multifocale et d'une réponse incomplète au traitement est associée à une probabilité accrue de récidive. Ce groupe représente des patients à risque clinique non négligeable, même dans des stades précoces.

- Cluster 2 (jaune - triangles) : Ce groupe représente des patients à faible risque, typiques d'une prise en charge efficace du cancer thyroïdien. Ils ont des caractéristiques favorables à tous les niveaux (absence d'adénopathie, réponse complète, stade I), et ne montrent aucune récidive. Ce cluster peut être considéré comme un groupe de référence ou un profil de guérison attendue.

- Cluster 3 (gris - carrés) : Ce cluster regroupe un profil alarmant, avec des caractéristiques cliniques sévères à tous les niveaux : grosse tumeur, adénopathies multiples, métastases, mauvaise réponse au traitement. Tous les patients de ce groupe présentent une récidive, ce qui en fait un profil critique dans la prise en charge thérapeutique.

- Cluster 4 (rouge - croix) : Ce cluster concentre les caractéristiques les plus favorables observables dans l'ensemble du jeu de données. Le type histologique (Micropapillary) est reconnu pour son faible potentiel de récidive, et toutes les autres variables vont dans le même sens. Il s'agit du profil idéal, typique des formes très localisées et traitées efficacement.

Ces clusters mettent en évidence la capacité du clustering à restituer des profils patients cohérents cliniquement, permettant d'envisager une stratification des risques fondée sur des combinaisons de variables, et non sur un facteur isolé. Cette approche peut constituer un outil d'aide à la décision complémentaire aux classifications traditionnelles.

Classification

L’objectif de cette étape est d’évaluer la capacité de différents modèles de classification supervisée à prédire la variable cible Recurred (récidive du cancer de la thyroïde), à partir de données cliniques et pathologiques. Quatre modèles ont été testés et comparés sur la base des performances observées lors de validations croisées :

- Arbre de décision
- Forêt aléatoire
- Régression logistique
- Classifieur naïf bayésien

Tableau 2 : tableau de l'efficacité de différents modèles de classification

Modèle	Taux Erreur (%)	Nombre moyen Erreur	Nombre moyen patient test
Arbre de décision	3,394	1.3	38.3
Forêt aléatoire (500 arbres)	3,914	1,5	38,3
Régression linéaire (l'âge non inclus)	4,69	1,8	38,3
Naïf bayésien	7,834	3	38,3

- L’arbre de décision s’impose comme le meilleur compromis entre performance et interprétabilité. Son taux d’erreur très faible (3,4 %) en fait un outil potentiellement exploitable dans un cadre clinique.
- La forêt aléatoire, plus complexe mais robuste, obtient des performances similaires, avec une meilleure capacité de généralisation sur des données plus bruitées.
- La régression logistique, bien que légèrement moins performante, permet d’identifier les contributions individuelles des variables à la prédiction, ce qui peut être utile dans un objectif explicatif.
- Le classifieur naïf bayésien, basé sur l’indépendance des variables, est nettement moins performant ici, ce qui confirme que les relations entre variables sont interdépendantes et que ce modèle n’est pas le plus adapté à ce jeu de données.

La comparaison des modèles montre que les méthodes d’arbres sont particulièrement efficaces dans ce contexte. Leur capacité à modéliser des interactions complexes entre variables catégorielles, tout en maintenant une bonne lisibilité du modèle, les rend particulièrement pertinentes pour une application médicale.

Conclusion

Ce travail de fouille de données avait pour objectif d'explorer les facteurs cliniques associés à la récidive du cancer de la thyroïde, à partir d'un jeu de données réel comportant 383 patients. À travers une démarche, combinant analyses exploratoires, réduction de dimension, clustering non supervisé et classification supervisée, plusieurs enseignements ont pu être tirés.

L'analyse exploratoire a permis de mettre en évidence des associations fortes entre certaines variables et la récidive, en particulier la présence d'adénopathie, le type de réponse au traitement, le niveau de risque clinique, ou encore les classifications TNM (T, N, M). Le test du Chi² a confirmé statistiquement ces liens, identifiant ainsi des variables potentiellement discriminantes.

Le clustering par k-means, appliqué aux dimensions issues de l'Analyse des Correspondances Multiples (ACM), a révélé quatre profils distincts de patients. Ces profils sont globalement cohérents : un cluster à très haut risque, un cluster intermédiaire avec facteurs isolés de gravité, et deux clusters correspondant à des cas peu avancés et bien pris en charge. Ce regroupement non supervisé permet ainsi d'envisager une stratification automatisée des patients selon leur niveau de risque.

Enfin, plusieurs modèles de classification ont été comparés pour prédire la récidive. Les résultats montrent que les arbres de décision et les forêts aléatoires offrent les meilleures performances, avec des taux d'erreur moyens inférieurs à 4 %. Ces méthodes, tout en étant performantes, présentent l'avantage d'être explicables, ce qui est essentiel dans un contexte médical.

En conclusion, ce projet démontre l'apport concret de la fouille de données dans le domaine de la santé, et en particulier dans le suivi post-thérapeutique des cancers. Il ouvre la voie à des outils de soutien à la décision permettant d'anticiper les récidives et d'adapter les parcours de soins. Des pistes d'amélioration pourraient consister à enrichir le jeu de données (imagerie, durée de suivi, informations biologiques, notre jeu de données contenant en outre une majorité de patient femme et avec un cancer papillaire) ou à tester d'autres approches de modélisation (réseaux de neurones, gradient boosting, modèles hybrides).

Bibliographie

Thyroid cancer gender disparity

[1] Contributeurs aux projets Wikimedia. (2024, 11 novembre). Cancer de la thyroïde. https://fr.wikipedia.org/wiki/Cancer_de_la_thyro%C3%AFde

[2] Contributeurs aux projets Wikimedia. (2024, octobre 7). Tumeur. https://fr.wikipedia.org/wiki/Tumeur

[3] Rahbari, R., Zhang, L., & Kebebew, E. (2010). Thyroid cancer gender disparity. Future Oncology, 6(11), 1771-1779. https://doi.org/10.2217/fon.10.127

[4] Du Cancer, C. C. S. /. S. C. (2024, 1 janvier). Statistiques sur le cancer de la thyroïde. Société canadienne du Cancer. https://cancer.ca/fr/cancer-information/cancer-types/thyroid/statistics

[5] Fréquence et risque du cancer de la thyroïde - VIDAL. (s. d.). VIDAL. https://www.vidal.fr/maladies/cancers/cancer-thyroide/frequence-risque.html

[6] Classification du cancer, fondation québécoise du cancer https://cancerquebec.ca/information-sur-le-cancer/le-cancer/classification-cancer/

[7] Classification TNM 8ème édition https://referentiels-aristot.com/wp-content/uploads/2_CPC_5_Classification.pdf https://www.arcagy.org/infocancer/uploads/pdf/la-stadification-classification-tnm-8%C3%A8me-%C3%A9dition-2017-7249.pdf

[8] Classification TNM Mc Millan cancer support https://www.macmillan.org.uk/cancer-information-and-support/thyroid-cancer/stages#:~:text=T3a%20means%20the%20tumour%20is,thyroid%20gland%20into%20nearby%20muscles.

[9] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

[10] Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

[11] Pedersen, T. L. (2024). patchwork: The Composer of Plots (R package version 1.3.0). https://CRAN.R-project.org/package=patchwork

[12] Sievert, C. (2020). Interactive web-based data visualization with R, plotly, and shiny. Chapman and Hall/CRC. https://plotly-r.com

[13] Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. Journal of Statistical Software, 25(1), 1–18. https://doi.org/10.18637/jss.v025.i01

[14] Kassambara, A., & Mundt, F. (2020). factoextra: Extract and visualize the results of multivariate data analyses (R package version 1.0.7). https://CRAN.R-project.org/package=factoextra

[15] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2025). cluster: Cluster analysis basics and extensions (R package version 2.1.8.1). https://CRAN.R-project.org/package=cluster

Annexe

Annexe 1 : Tableau description des variables utilisées

Nom de la variable	Type	Description	Modalités
Gender	qualitative	Sexe	"F" "M"
Hx Radiotherapy	qualitative	Antécédents de radiothérapie	"No" "Yes"
Adenopathy	qualitative	Atteint ganglionnaire	"No" "Right" "Extensive" "Left" "Bilateral" "Posterior"
Pathology	qualitative	Type de cancer thyroïdien	"Micropapillary" "Papillary" "Follicular" "Hurthel cell"
Focality	qualitative	Focalité tumorale	"Uni-Focal" "Multi-Focal"
Risk	qualitative	Niveau de risque de propagation du cancer	"Low" "Intermediate" "High"
T	qualitative	Tumeur primitive (Classifications TNM)	"T1a" "T1b" "T2" "T3a" "T3b" "T4a" "T4b"
N	qualitative	Tumeur dans les ganglions lymphatiques (node)	"N0" "N1b" "N1a"
M	qualitative	Métastase	"M0" "M1"
Stage	qualitative	Stade du cancer	"I" "II" "IVB" "III" "IVA"
Reponse	qualitative	Réponse au traitement	"Indeterminate" "Excellent" "Structural Incomplete" "Biochemical Incomplete"
Reccured	qualitative	Récidive	"No" "Yes"
Age	quantitative	Âge du patient	/

Annexe 2 : Tableau classification TNM [6] [7]

T1	< 2 cm. Elle ne s'est pas développée à l'extérieur de la glande thyroïde.

T1a	≤ 1cm
T1b	> 1 cm et ≤ 2 cm
T2	> 2 cm et ≤ 4cm. Elle ne s'est pas développée hors de la glande thyroïde.
T3	> 4 cm ou s'est légèrement développée à l'extérieur de la glande thyroïde.
T3a	> 4 cm et ne s'est pas développée en dehors de la glande thyroïde.
T3b	La tumeur est de n'importe quelle taille et s'est légèrement développée à l'extérieur de la glande thyroïde dans les muscles voisins.
T4	T4 signifie que la tumeur s'est développée à l'extérieur de la glande thyroïde et dans les structures voisines.
Nx Envahissement d'un/de plusieurs ganglions lymphatiques	
N1	Le cancer s'est propagé aux ganglions lymphatiques proches de la glande thyroïde ou dans la région du cou ou de la poitrine.
N1a	Le cancer s'est propagé aux ganglions lymphatiques situés au milieu du cou, près de la glande thyroïde.
N1b	Le cancer s'est propagé aux ganglions lymphatiques d'un ou des deux côtés du cou, ou de la partie supérieure de la poitrine.
Mx Métastases	
M0	Pas de métastases
M1	Métastase(s)

Annexe 3 : Tableau classification des cancers par stade [8]

Il est possible de classer les cancers par stade. Chaque cancer classé via la classification TNM possède une correspondance dans la classification par stade.

Stade 1	Tumeur unique et de petite taille (ex: T1N0M0)

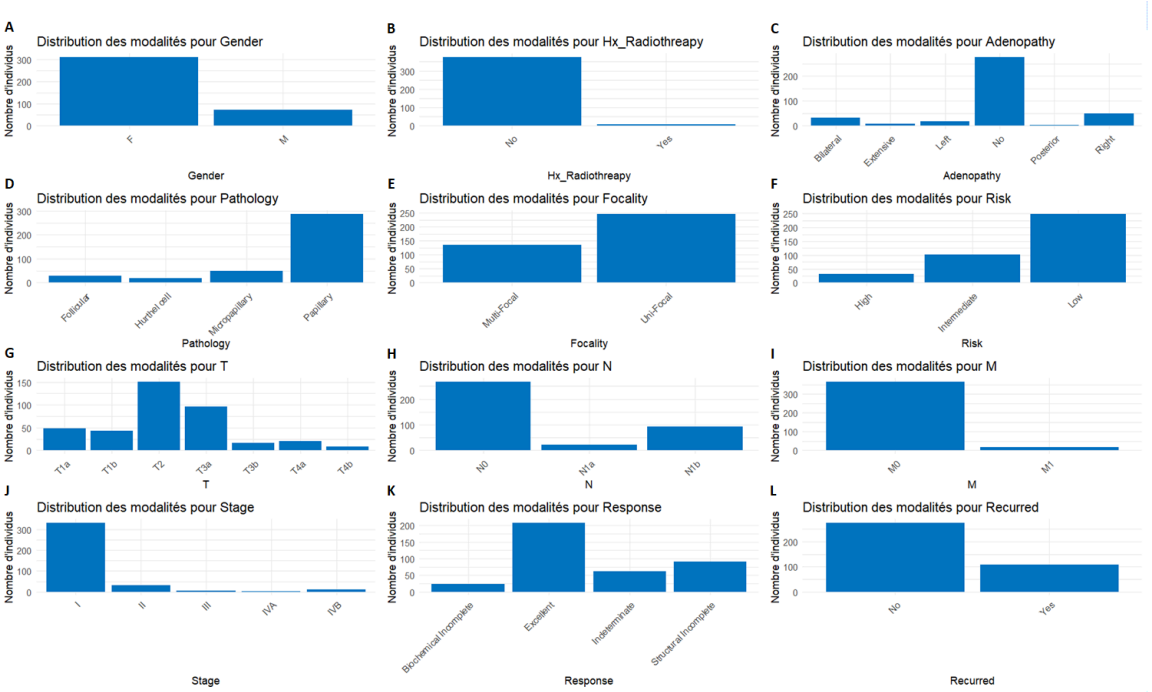
Stade 2	Correspond à un volume local plus important que le stade 1 (ex: T2N0M0)	
Stade 3	Envahissement des ganglions lymphatiques et/ou des tissus avoisinants (ex: T1N1M0 ou T3N0M0)	
Stade 4	Extension aux organes distants.	
	Stade IV-A - Nodules tumoraux séparés dans un lobe controlatéral, ou nodules pleuraux ou pleurésie maligne ou péricardite maligne ou - 1 seule métastase dans un seul site métastatique	**Stade IV-B** - Extension plus large et/ou une dissémination dans l'organisme sous forme de métastases

Annexe 4 : Caractéristique des Clusters

Cluster	gender	Hx Radiotherapy	Adenopathy	Pathology	Focality	Risk
1	F	No	Right	Papillary	Multi-Focal	Intermediate
2	F	No	No	Papillary	Uni-Focal	Low
3	M	No	Bilateral	Papillary	Multi-Focal	hight
4	F	No	No	Micropapillary	Uni-Focal	Low

Cluster	T	N	M	Stage	Response	Recurred
1	T3a	N1b	M0	I	Structural Incomplete	Yes
2	T2	N0	M0	I	Excellent	No
3	T4a	N1b	M1	IVB	Structural Incomplete	Yes
4	T1a	N0	M0	I	Excellent	No

Annexe 5 : Diagrammes en barre des distributions des modalités de nos variables



Annexe 6 : Graphique silhouette