الجـــــــــــــــــمهـوريـــــة الجـــــــــــــزائـريـة الديمـــــــقـراطيـة الـشـــــــــعبيـة

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التـــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـــــــــــــامعة أبي بكـــر بـلـقـايد – تـلمســـــــــــــان –

Université Aboubakr Belkaïd – Tlemcen –

Faculté de TECHNOLOGIE

THESE

Présentée pour l'obtention du **grade** de **DOCTORAT 3ème Cycle**

En : Génie Biomédical

Spécialité : Informatique Biomédicale

Par : Mme. GUILAL Rima ep. HADJ KADDOUR

Sujet

Sélection de variables et analyse d'association pour les données cliniques du Myélome Multiple

Soutenue publiquement, le 07/10/2023 , devant le jury composé de :

MESSADI Mohamed	Professeur	**Université de Tlemcen**	**Président**
CHIKH Mohammed Amine	Professeur	**Université de Tlemcen**	Directeur
BOUAMRANE Karim	Professeur	**Université d'Oran 1**	Examinateur 1
BENMOUNA Youcef	MCA	**Université de Tlemcen**	Examinateur 2
BOUKLI HACENE Ismail	MCA	**Université de Tlemcen**	Examinateur 3
SETTOUTI Nesma	MCA		Invité 1
BENDAHMANE Ahmed Fouad	Hématologue		Invité 2

Je dédie ce travail à :

Mes parents,
Mes sœurs,
Mon frère,
Mon Mari,
Mon neveu Mazen.

Qu'ils trouvent ici l'expression de toute ma reconnaissance.

Remerciements

Avant tous, Je remercie **ALLAH** Tout-Puissant de m'avoir donné le courage, la volonté, la force d'y croire et la patience d'aller jusqu'au Bout.

Au Prophète محمد صلى الله عليه و سلم Notre guide et notre exemple bien aimé. Qu'il nous oriente dans le droit chemin.

«من لا يشكر الناس لا يشكر الله» - حديث نبوي -

C'est avec un grand plaisir que je réserve cette page en signe de gratitude et de reconnaissance à tous ceux qui ont contribué à ce travail.

En premier lieu, je voudrais exprimer mes sincères remerciements à mes deux directeurs de thèse, le professeur **CHIKH Mohamed El Amine** et **Mme SETTOUTI Nesma**. Je tiens à vous adresser mes sincères remerciements, Mr. Professeur CHIKH, pour votre investissement scientifique et humain, ainsi que pour vos précieux conseils et encouragements.

Je ne vous remercierai jamais assez Mme Settouti pour toutes les discussions intéressantes et fructueuses que nous avons eues, mais aussi pour la patience dont vous m'avez fait part à mon égard.

J'exprime également ma gratitude aux membres du jury pour avoir accepté de juger ce travail : un grand honneur pour moi.

Je remercie Mme professeur **MESLI Naima**, chef service d'hématologie au CHU Tlemcen de m'avoir accordé d'effectuer un stage au sein de son service pour la mise en pratique de mon sujet de thèse.

Je remercie aussi **Mr. BENDAHMANE Ahmed Fouad**, Maitre-assistant

hospitalo-universitaire au CHU Tlemcen, de m'avoir aidé et donné de son temps pour que je passe faire la collecte des données.

Je souhaite également exprimer ma gratitude à **Mme Nafissa CHABNI**, Professeur d'épidémiologie, et **Mme Lamia BOUBLENZA**, Maitre-assistant classe A, Université de Tlemcen, Algérie pour leur aide et leurs précieuses suggestions qui ont grandement enrichi mon travail.

Je tiens également à remercier le professeur **Gonzalo MARTÍNEZ MUÑOZ**, Escuela Politéctica Superior, Universidad Autónoma de Madrid (Spain) pour avoir partagé avec moi ses connaissances dans le domaine de l'intélligence artificielle et de l'aide au diagnostic médical , lorsque j'ai bénéficié d'une bourse de doctorat "Erasmus+" à l'école polytechnique de l'Université Autónoma de Madrid en Espagne.

Mes salutations vont aux différents membres du Laboratoire de Génie Biomédical (GBM) et plus particulièrement à l'équipe du CREDOM, ainsi qu'à tous les chercheurs que j'ai rencontrés lors de conférences et stages de recherche, je ne peux pas tous les nommer.

Mes remerciements vont également au pharmacien **SEMAINE Mohamed**, et mes collègues de travail (Meriem, Amel, Youcef). Merci beaucoup pour votre support et vos encouragements.

Je tiens particulièrement à remercier ma véritable amie : Dr. **Ahlem BENAZZOUZ** pour avoir été à mes côtés dans les moments les plus difficiles, pour tous les moments que nous avons passé à travailler ensemble et pour tous les plaisirs que nous avons eus. Ce sont de merveilleux souvenirs que je garderai avec moi pour les années à venir.

Et bien sûr, Je ne terminerai pas mes remerciements sans exprimer ma gratitude envers **mes parents**, pour leur incroyable soutien tout au long de ma vie, leurs encouragements et leur motivation et pour avoir cru en moi. Un grand merci également à mes sœurs **Chaimaa, Asma, Douaa**, mon petit frère **Ahmed Yacine**, et mon mari **YOUCEF** pour m'avoir soutenu dans les moments difficiles, m'avoir encouragé et motivé pendant que je travaillais sur ma thèse.

Enfin, mes remerciements vont également à tous ceux qui ont participé plus ou moins indirectement au bon déroulement de ma thèse.

RIMA GUILAL

Résumé

Le myélome multiple est un cancer complexe du sang caractérisé par une prolifération incontrôlée de plasmocytes. En raison de l'absence de signes ou symptômes évidents dans les premiers stades, le diagnostic précoce de cette maladie est difficile et compliqué. Cela représente un défi considérable pour les patients, et le processus de diagnostic long peut être décourageant.

Cette thèse a pour objectif de prédire les tests les plus importants pour le diagnostic du myélome multiple et d'établir la relation entre les variables et les stades de ce cancer. L'objectif est d'améliorer l'efficacité du processus de diagnostic tout en réduisant les coûts. Les données ont été collectées auprès du Centre de lutte contre le cancer du CHU de Tlemcen. Après avoir traité les données déséquilibrées, nous avons utilisé des méthodes de sélection de variables basées sur une approche de filtre pour identifier les tests les plus pertinents. La méthode de sélection FCBF s'est révélée la plus pratique en raison de sa robustesse et de sa capacité à éliminer les caractéristiques non pertinentes.

Nous avons également exploré l'utilisation de méthodes d'ensemble basées sur les arbres de décision pour estimer l'importance des variables. Après avoir ajusté les hyperparamètres à l'aide de la technique GridSearchCV, les résultats ont montré que XGBoost a donné le meilleur classement pour les caractéristiques considérées comme les facteurs pronostiques les plus importants.

Dans cette étude, nous avons également abordé l'intégration d'une méthode basée sur les réseaux bayésiens. L'objectif était d'optimiser l'inférence clinique et de découvrir des relations causales intéressantes. Un modèle bayésien optimal a été obtenu, qui a exploré 46 relations causales pertinentes lors de la phase d'apprentissage des paramètres. Cela nous a permis de prédire de nouveaux scénarios.

En résumé, cette thèse se concentre sur l'amélioration du diagnostic du myélome multiple en prédisant les tests les plus importants et en explorant les relations entre les variables et les stades du cancer. Les résultats obtenus peuvent contribuer à améliorer l'efficacité du processus de diagnostic, tout en réduisant les coûts associés. De plus, l'utilisation de méthodes d'ensemble et de réseaux bayésiens a permis d'identifier les facteurs pronostiques les plus importants et d'explorer des relations causales intéressantes.

Mots clés : Myélome multiple, sélection de variables, approche par filtre, méthodes d'ensemble, importance des variables, réseau bayésien, association, données déséquilibrées.

Abstract

The multiple myeloma is a complex blood cancer characterized by uncontrolled proliferation of plasma cells. In the early stages, multiple myeloma may not cause any signs or symptoms, making early diagnosis difficult and challenging. The diagnostic process poses a considerable challenge for patients, and its prolonged duration can be discouraging.

This thesis aims to predict the most important tests in the diagnosis of multiple myeloma and determine the relationship between variables and stages of this cancer. The goal is to enhance the efficiency of the diagnostic process while reducing costs. We collected comprehensive data on multiple myeloma from the Cancer Center at Tlemcen University Hospital. After handling imbalanced data, we employed variable selection methods based on a filter approach to extract the most relevant tests. The FCBF selection technique proved to be more practical due to its robustness and ability to eliminate irrelevant features.

We also explored ensemble methods based on decision trees to estimate variable importance. The results obtained after hyperparameter tuning using the GridSearchCV technique showed that XGBoost yielded the best ranking for the considered features, which are considered the most significant prognostic factors.

In this study, we also addressed the integration of a Bayesian network-based method. The aim was to optimize clinical inference and discover interesting causal relationships. An optimal Bayesian model was obtained, exploring 46 relevant causal relationships during the parameter learning phase. This enabled us to predict new scenarios.

In summary, this thesis focuses on improving the diagnosis of multiple myeloma by predicting the most important tests and exploring the relationships between variables and cancer stages. The results obtained can

contribute to enhancing the efficiency of the diagnostic process while reducing associated costs. Furthermore, the use of ensemble methods and Bayesian networks has helped identify the most significant prognostic factors and explore interesting causal relationships.

Keywords : Multiple Myeloma, feature selection, filter approach, ensemble methods, feature importance, bayesian network, association, imbalanced data.

Table des matières

Table des figures

Liste des tableaux

Glossaire

AA	:	Apprentissage Artificiel.
AdaBoost	:	Adaptive Boosting.
ANN	:	Artificiel Neural Network.
ASR	:	Age-Standardized Rate.
BOM	:	Biopsie Ostéo-Médullaire.
CatBoost	:	Category Boosting.
CFS	:	Correlation-based Feature Selection.
CHU-Tlm	:	Centre Hospitalo-Universitaire du Tlemcen.
CLCC	:	Centre de Lutte Contre le Cancer.
CRAB	:	hyperCalcemia, Renal Failure, Anemia, Bone lesions.
DAG	:	Directed Acyclic Graph.
EMG	:	ÉlectroMyoGramme.
ExtraTree	:	Extremely Randomized Trees classifier.
FCBF	:	Fast correlation-Based Feature Selection.
FDG	:	FluoroDéoxyGlucose.
FLC	:	Free Light Chain.
GB	:	Globules Blancs.
GB	:	Gradient Boosting.
GR	:	Globules rouges.
IARC	:	International Agency of Research on Cancer.
IFx	:	Immunofixation des protéines.
IG-FS	:	Information Gain- Feature Selection.
IMWG	:	International Myeloma Working Group.
IRM	:	Imagerie par Résonnance Magnétique.
ISS	:	International Staging System.
KNN	:	K-Nearest Neighbors.
LightGBM	:	Lightweight Gradient Boosting machines.
MGUS	:	Momoclonal gammopathy of undetermind significance.
ML	:	Machine Learning.

MM	:	Myélome Multiple.
NB	:	Naïve Bayes.
NFS	:	Numération de la Formule Sanguine.
OMS	:	Organisation Mondiale de la Santé.
OOB	:	Out-Of-Bag.
PET	:	Positron Emission Tomography.
PGM	:	Probabilistic Graphical Modelling.
PltS	:	Plaquettes sanguines.
RB	:	Réseau Bayésien.
RF	:	Random Forest.
SADM	:	Systèmes d'Aide à la Décision Médicale.
SMOTE	:	Synthetic Minority Oversampling Technique.
SVM	:	Support Vector Machine.
TDM	:	TomoDensitoMétrie.
TEP	:	Tomographie à Emission de Positons.
VS	:	Vitesse de Sédimentation.
XGBoost	:	Extreme Gradient Boosting.

Introduction Générale

1 Contexte de la thèse

L'ingénierie biomédicale est un domaine pluridisciplinaire qui combine plusieurs domaines, tels que la biologie et la médecine. Elle occupe une place centrale dans les recherches scientifiques actuelles de l'industrie médicale et pharmaceutique, avec pour objectif le développement de systèmes d'aide au diagnostic médical pour de nombreuses maladies. Les recherches en bio-ingénierie portent notamment sur les bactéries modifiées pour produire des produits chimiques, les nouvelles technologies d'imagerie médicale, les dispositifs portables et rapides de diagnostic des maladies, les prothèses et les produits biopharmaceutiques, etc.

En médecine, la prise de décision clinique est une activité intellectuelle où les médecins utilisent leurs connaissances spécialisées, leur expérience et leur jugement clinique, combinés aux informations de test disponibles, pour déterminer l'investigation, le traitement ou la prise en charge cliniquement appropriés de l'état d'un patient, en tenant souvent compte du point de vue du patient et de sa compréhension de son état.

L'intelligence artificielle (IA) a pour principal défi de développer des outils et des technologies informatiques puissants pour simplifier et automatiser la tâche du médecin, voire de remplacer l'intervention humaine dans le raisonnement clinique.

Les systèmes d'aide à la décision médicale (SADM) sont des outils informatiques conçus pour soutenir la prise de décision clinique. Les professionnels de la santé utilisent ces outils pour traiter l'ensemble des caractéristiques d'un patient donné afin de générer les diagnostics probables

de son état clinique (aide au diagnostic) ou les traitements qui lui seraient adaptés (aide à la thérapeutique).

Ce domaine de recherche suscite un grand intérêt dans la communauté de l'apprentissage automatique, car il est porté par l'évolution des systèmes de collecte et de stockage de données, d'une part, et les exigences des systèmes pour soutenir les diagnostics, d'autre part.

Cependant, la représentation des données utilise souvent de nombreuses caractéristiques, dont seules quelques-unes peuvent être liées au concept cible. Les données inutiles peuvent ralentir le modèle d'apprentissage et le modèle peut mal apprendre de ces données, ce qui peut entraîner des erreurs. De plus, le volume et la spécificité de ces jeux de données, constitués d'un nombre de variables très largement supérieur au nombre d'expériences (échantillons), conduisent à des traitements faisant appel aux outils de sélection des caractéristiques et aux méthodes d'association. Ces techniques sont destinées à la fois à accélérer l'apprentissage et à améliorer la qualité du modèle.

2 Motivations et objectifs

Le sujet de cette thèse de doctorat vise à collecter des données biologiques auprès de différents centres hospitalo-universitaires. En collaboration avec les biologistes, nous réfléchissons à l'utilisation de ces données fondamentales dans la pratique clinique et à leur influence sur la prise en charge des patients. Ce domaine de recherche est crucial pour le dépistage, le traitement et la prédiction de l'évolution clinique des patients.

Grâce à une collaboration entre la faculté de Technologie de l'Université de Tlemcen et le centre hospitalo-universitaire de Tlemcen, nous avons effectué un stage pratique au sein du service d'hématologie du Centre de Lutte Contre le Cancer (CLCC) de Tlemcen. Nous avons pu contacter des biologistes et des médecins spécialistes en hématologie qui nous ont aidés à comprendre les différentes notions de base tout en nous orientant vers des pistes de recherche critiques d'intérêt local. L'objectif était de collecter des données sur le cancer du myélome multiple afin de les utiliser en pratique clinique et d'analyser leur impact sur le suivi des patients.

Le choix du cancer du myélome multiple a été fait en collaboration avec les hématologues qui nous ont orientés vers cette maladie en raison de son importance et de son activité de recherche dans le service d'hématologie.

Le diagnostic de cette maladie est complexe car certains symptômes et résultats de test peuvent également se produire dans d'autres maladies.

Le myélome multiple est un cancer du sang qui est associé à une croissance incontrôlée des plasmocytes dans la moelle osseuse [2]. Les cellules cancéreuses peuvent affecter plusieurs régions du corps, d'où le terme "multiple". Aux premiers stades de la maladie, un patient atteint de MM peut ne présenter aucun signe ni symptôme, ce qui en fait l'un des nombreux cancers compliqués. Une fois la tumeur développée, le diagnostic du MM nécessite l'observation des symptômes qui sont décrits avec l'acronyme "CRAB" : hypercalcémie, insuffisance rénale, anémie, douleurs osseuses et lésions ostéolytiques.

Le processus de diagnostic du myélome multiple (MM) est long et décourageant pour les patients, car il implique de nombreux examens et tests médicaux tels que des tests d'hématologie, des examens cytologiques, des tests de protéines, des tests d'imagerie médicale et des tests de biologie/chimie du sang [3]. Il est essentiel de déterminer le stade du MM pour aider les médecins à choisir le traitement approprié et à prédire le pronostic du patient, ce qui peut offrir une chance de guérison. Malheureusement, de nombreux patients sont diagnostiqués au stade III, ce qui crée un déséquilibre dans les exemples, certains stades étant plus fréquents que d'autres. Les facteurs causaux de cette maladie sont encore inconnus et les recherches sont en cours.

Dans ce contexte, notre objectif est de mieux comprendre l'évolution du cancer du MM à tous les stades en étudiant la relation entre les stades du MM et les examens médicaux effectués lors du diagnostic de cette maladie, afin d'identifier les examens les plus prédictifs dans le diagnostic et la stadification du MM. Nous cherchons notamment à répondre aux questions de recherche suivantes :

- Quelle famille de méthodes de sélection et d'association de caractéristiques est la plus appropriée ?
- Quelle méthode sera la plus adaptée à la classification des données biologiques réelles ?

3 Contributions

Le myélome multiple est une maladie très complexe dont les symptômes ne sont généralement pas détectés dans les premiers stades. Les

patients doivent subir une série de tests répétitifs et fréquents, ce qui rend le processus de diagnostic et de stadification du MM long et décourageant pour les malades.

Nos modèles proposés sont basés sur l'idée de prédire avec précision les tests les plus importants à effectuer dans le processus de stadification du MM. Cette approche permettrait de réduire le nombre de tests et de diminuer le coût total, un problème majeur pour tous les patients.

Les contributions de notre thèse de recherche sont les suivantes :

1. Nous avons collecté un nouveau jeu de données contenant les résultats de différents examens de diagnostic du MM (MM-dataset) [4]. Cette base de données peut être utilisée pour résoudre des problèmes tels que la détection des facteurs pronostiques du MM en utilisant des méthodes de sélection de caractéristiques, ainsi que pour résoudre des problèmes de classification de données déséquilibrées basés sur l'apprentissage automatique supervisé avec une sortie multi-classes.
2. Nous nous sommes intéressés à la corrélation entre les stades du MM et les résultats des différents examens diagnostiques de ce cancer, afin d'extraire les examens les plus pertinents pour le pronostic du myélome multiple. À cette fin, nous avons proposé des stratégies de sélection de caractéristiques basées sur une approche de filtre.
3. Nous avons proposé une analyse pratique de diverses approches de méthodes d'ensemble basées sur les arbres de décision, notamment les méthodes d'ensemble basées sur la randomisation et celles basées sur l'optimisation. L'utilisation de scores d'importance proposés par les méthodes d'ensemble nous a permis d'identifier les variables les plus pertinentes pour la construction d'un modèle prédictif plus performant. Ces scores peuvent être utilisés pour sélectionner les variables les plus importantes et éliminer celles qui ne sont pas significatives, tout en maintenant des performances similaires ou meilleures dans un temps d'apprentissage beaucoup plus court.
4. Afin d'optimiser l'inférence clinique, nous avons exploré et analysé des modèles graphiques probabilistes : les Réseaux Bayésiens (RB). Ces derniers ont pour objectif de comprendre les relations causales entre les variables du jeu de données MM et le stade du MM. Les réseaux bayésiens nous ont permis de produire une représentation graphique fiable et transparente, afin de mieux comprendre les relations entre les paramètres qui influencent le diagnostic du MM, avec une possibilité de prédire de nouveaux scénarios.

4 Organisation du manuscrit

Ce travail est structuré en trois chapitres :

Chapitre 1 présente les données et les connaissances sur le myélome multiple, y compris l'épidémiologie, la physiologie, les signes cliniques et les manifestations du MM, les critères de diagnostic et les analyses de sang/d'urine nécessaires, ainsi que les différentes formes cliniques du MM et les systèmes de stadification utilisés par les cliniciens.

Chapitre 2 traite de notre contribution à l'identification des caractéristiques pertinentes dans le diagnostic du MM. Nous détaillons l'approche proposée ainsi que les méthodologies utilisées pour mener à bien notre recherche. Avant cela, nous présentons les différentes approches de sélection de variables, leur importance et leurs avantages, ainsi que les travaux récents dans la littérature. Nous exposons également notre base de données collectée pour notre étude. Nous décrivons les méthodes d'ensemble basées sur les arbres de décision qui visent à interpréter l'importance des caractéristiques dans la stadification du MM. Enfin, nous discutons des résultats obtenus et expliquons la contribution que nos travaux peuvent apporter à la recherche scientifique.

Chapitre 3 vise à explorer et analyser des modèles graphiques probabilistes : un réseau bayésien (RB), dont le but est de comprendre les relations causales entre les variables de l'ensemble de données sur le MM. Nous commençons par introduire les concepts de base des réseaux bayésiens, leurs propriétés, les étapes de construction d'un modèle bayésien, ainsi que les différents algorithmes utilisés en apprentissage de structure. Nous citons également certains travaux de la littérature sur l'utilisation des réseaux bayésiens dans le domaine médical, notamment dans la prise de décision. Enfin, nous présentons notre méthodologie et les expérimentations réalisées avec une discussion des résultats obtenus.

Enfin, nous terminons par une conclusion générale qui fournit un bref résumé de nos contributions tout au long de ce travail, ainsi que des suggestions pour des perspectives de travaux futurs.

CHAPITRE 1

LE MYÉLOME MULTIPLE EN CLINIQUE

1 Définition

Le myélome multiple (MM) est aussi appelé « maladie de Kahler » du nom du médecin autrichien Otto Kahler qui l'a décrite pour la première fois en 1889 [5]. C'est une hémopathie maligne qui affecte les plasmocytes dans la moelle osseuse [6]. Elle est la plus fréquente des dysglobulinémies malignes qui rentre dans le cadre des syndromes lymphoprolifératifs chroniques. Le mot "multiple" est utilisé parce que les cellules cancéreuses touchent souvent plusieurs parties du corps. Le MM se développe à partir des cellules plasmatiques qui sont un type de globules blancs.

2 Épidémiologie

2.1 Dans le monde

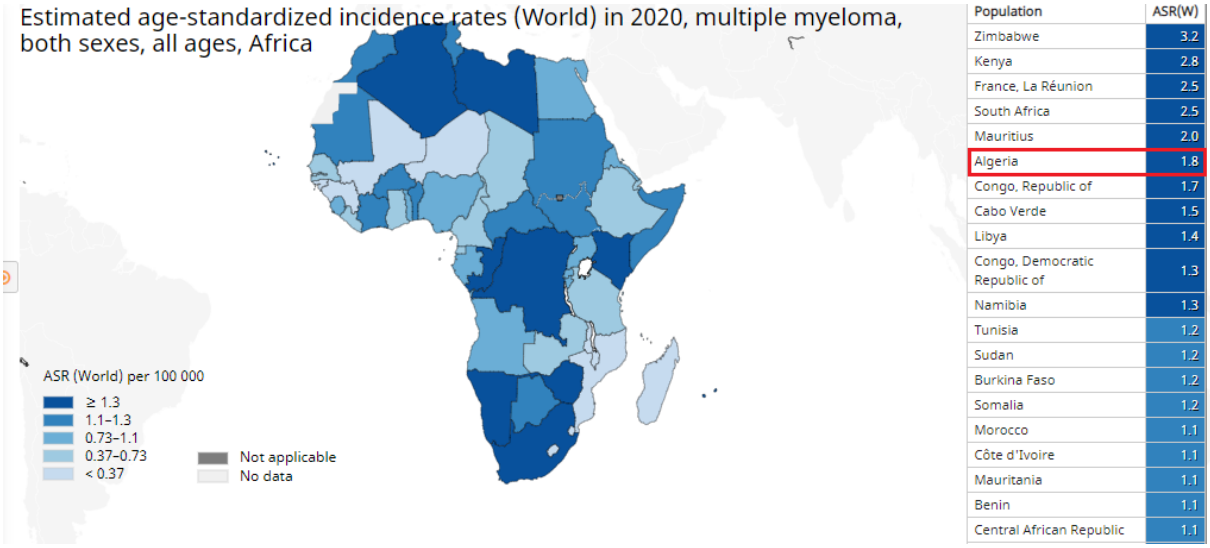
Bien qu'il s'agisse d'une maladie rare, le MM représente 13% des hémopathies malignes avec 159 985 nouveaux cas en 2018 [7], et est le 23ème cancer le plus fréquent dans le monde [8]. Il est plus fréquent chez les hommes que chez les femmes et chez les personnes d'origine afro-américaine. Ce cancer ne touche pas les enfants.

L'incidence du MM varie selon le statut socio-démographique, elle est de 4 à 6 cas pour 100 000 habitants dans les pays développés avec un âge médian de 70 ans [9], dont 37% des patients ont un âge inférieur à 65 ans, 26% ont entre 65 et 74 ans, et 37% des patients ont plus de 75 ans.

2.2 En Algérie

Selon les statistiques du Centre international de recherche sur le cancer (IARC) de l'Organisation mondiale de la santé (OMS) [1], l'Algérie se classe sixième en Afrique, en 2020, en termes de taux d'incidence estimés standardisés pour l'âge (ASR : Age-Standardized Rate) du MM, pour les deux sexes et tous les âges (Voir figure 1) [2].

Le MM est le 19ème cancer le plus fréquent, et la troisième maladie sanguine la plus fréquente en Algérie après les lymphomes et les leucémies aiguës [10]. Il y a environ 752 nouveaux cas pour 43 millions d'habitants en 2020, avec une incidence de 4,13 pour 100 000 au cours des cinq dernières années [11].



Population	ASR(W)
Zimbabwe	3.2
Kenya	2.8
France, La Réunion	2.5
South Africa	2.5
Mauritius	2.0
Algeria	1.8
Congo, Republic of	1.7
Cabo Verde	1.5
Libya	1.4
Congo, Democratic Republic of	1.3
Namibia	1.3
Tunisia	1.2
Sudan	1.2
Burkina Faso	1.2
Somalia	1.2
Morocco	1.1
Côte d'Ivoire	1.1
Mauritania	1.1
Benin	1.1
Central African Republic	1.1

Figure 1 – Taux d'incidence estimés standardisés sur l'âge (Monde) en 2020, Myélome Multiple, pour les deux sexes, tous âge, Région africaine de l'OMS.

3 Physiologie

La moelle osseuse est une matière spongieuse que l'on trouve principalement au centre des os longs. C'est là que sont produites les cellules sanguines. La moelle osseuse est composée d'une variété de cellules qui arrivent progressivement à maturité pour former (Voir figure 2) :

1. *Les globules rouges (GR)* : ou hématies ou érythrocytes, assurent le transport d'oxygène inhalé des poumons vers les autres organes et transportent le dioxyde de carbone des organes vers les poumons pour être expiré. C'est l'hémoglobine, un pigment situé dans les globules

1. https://gco.iarc.fr/today/home
2. https://gco.iarc.fr/today/online-analysis-map?v=2020

rouges, qui assure cette fonction. Un déficit en GR provoque une anémie.

2. *Les plaquettes (Plqts)* : également appelées thrombocytes, sont les cellules sanguines circulantes qui sont responsables de la coagulation du sang en collaboration avec certaines protéines et avec les cellules des vaisseaux. S'il y a un déficit en plqts, les plaies cicatrisent moins vite et les hématomes peuvent apparaître spontanément.
3. *Les globules blancs (GB)* : connus aussi sous le nom de leucocytes, défendent l'organisme contre les micro-organismes infectieux et les substances étrangères. Un déficit en GB peut se traduire par une augmentation dans la fréquence d'apparition des infections comme la pneumonie, la grippe, etc. Ces cellules se déclinent en plusieurs variétés : granulocytes, lymphocytes et monocytes, etc., chacun ayant un rôle différent dans le système immunitaire. Les lymphocytes sont désignés comme étant des lymphocytes B ou T qui sont présents dans les ganglions lymphatiques et autres tissus lymphatiques [6].

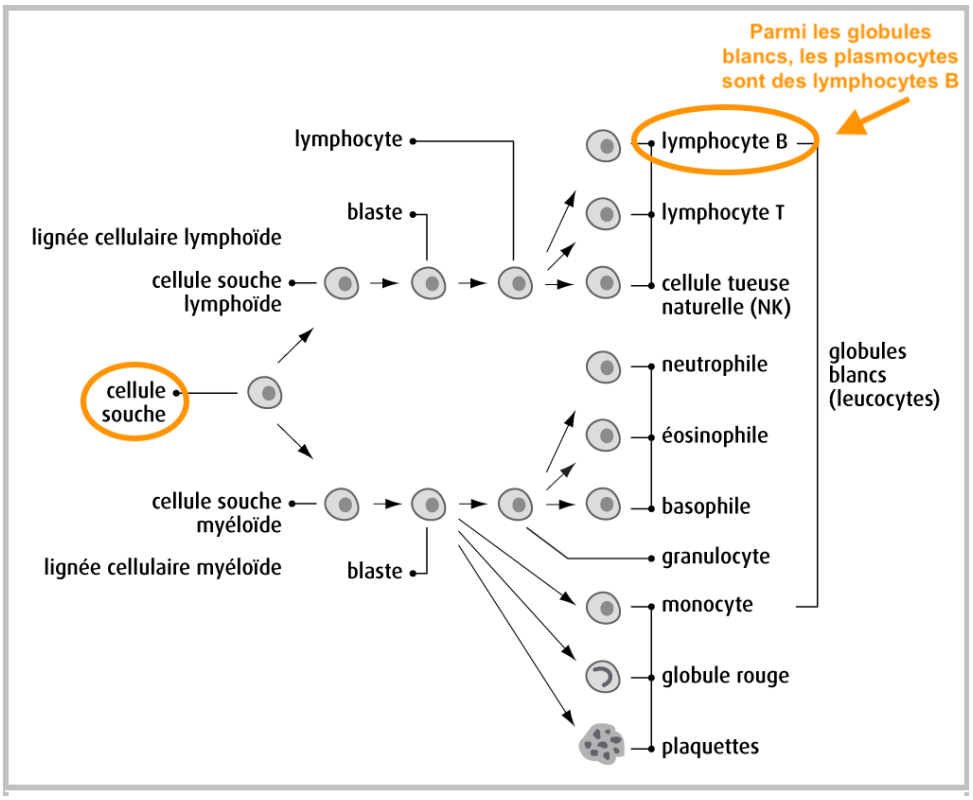


Figure 2 – Développement des cellules sanguines.

Les plasmocytes se développent eux-mêmes à partir des cellules lymphocytes B. Ils constituent moins de 5% des cellules de la moelle osseuse. Ils combattent les infections en fabriquant des anticorps (immunoglobu-

lines) qui reconnaissent et attaquent les agents responsables des maladies tels que les virus et les bactéries. Chaque plasmocyte ne peut produire qu'un seul type spécifique d'anticorps. Les plasmocytes individuels peuvent ensuite se diviser à plusieurs reprises pour former des copies d'eux-mêmes, appelées clones. Ce groupe de clones peut produire de grandes quantités d'un seul type d'anticorps pour combattre une infection spécifique.

4 Étiopathogénie

L'étiopathogénie (ou étiologie) est l'étude des causes et facteurs de risque qui modifient la chance d'une personne de contracter une maladie, telle que le cancer. Elle touche plusieurs aspects de la vie socioprofessionnelle du patient, notamment la profession, le mode de vie, les antécédents médicaux et la génétique [12].
Les facteurs de risque sont habituellement classés du plus important au moins important, mais dans la plupart des cas, il est impossible de les classer avec une certitude absolue. De plus, ces facteurs ne sont pas suffisants pour expliquer tous les cas de la maladie, car des personnes n'ayant aucun facteur de risque peuvent tout de même développer la maladie, tandis que d'autres ayant plusieurs facteurs de risque ne seront pas atteintes de la maladie.

L'étiopathogénie précise du MM est encore inconnue à ce jour, car il n'y a pas de cause bien définie pour cette maladie [13]. Son apparition est une conséquence de plusieurs événements oncogéniques chromosomiques et moléculaires liés à la série lymphocytaire B [14]. Le seul facteur de risque clairement identifié est l'exposition aux rayonnements ionisants et aux agents chimiques [15]. Cependant, comme pour d'autres types de cancer, il existe certains facteurs qui peuvent influencer les chances d'apparition du MM, tels que :

1. **Âge :** Le risque d'atteindre un MM augmente avec l'âge. L'âge moyen au moment du diagnostic est de 70 ans. Moins de 1% des cas sont diagnostiqués chez des personnes de moins de 35 ans.
2. **Genre & origine ethnique :** L'homme est légèrement plus touché que la femme. Aussi, le MM survient souvent deux fois plus chez la population noire que chez la population blanche, il est plus fréquent au Moyen-Orient, en Afrique et en Méditerranée. Les raisons ne sont pas connues.
3. **Antécédents médicaux familiaux :** Quelques études montrent

que le myélome semble être présent dans certaines familles. Le risque est élevé pour une personne dont la mère ou le père ou bien un frère ou une sœur en souffre du MM. Cependant, il n'y a aucune preuve d'un facteur héréditaire du myélome et pas d'excès dans la famille des patients atteints de ce cancer.

4. **Obésité & habitudes alimentaires :** L'obésité et les habitudes alimentaires ont été classées comme probablement moins impliquées dans le processus du développement du MM [10]. Les personnes ayant un indice de masse corporelle (IMC) élevé sont souvent plus susceptibles de développer un myélome que celles ayant un poids santé idéal.

5. **Autres maladies des plasmocytes :** La personne atteinte d'une gammapathie monoclonale de signification indéterminée ou d'un plasmocytome solitaire a un risque plus élevé (1 à 2 % de chances par an) de développer un myélome, un lymphome ou un autre cancer lié au sang appelé macroglobulinémie de Waldenstrom [16].

6. **Système immunitaire affaibli :** Les personnes dont le système immunitaire est affaibli ont un risque plus élevé de développer un myélome multiple. Il s'agit notamment des personnes atteintes du VIH ou du SIDA et des personnes ayant subi une transplantation d'organe et devant prendre des médicaments pour affaiblir leur système immunitaire. Certains chercheurs pensent aussi que les infections et les inflammations chroniques pourraient jouer un rôle dans la pathogenèse de cette maladie [17]. Pour combattre ces infections, les plasmocytes produisent de grandes quantités d'anticorps. En conséquence, certains plasmocytes peuvent dévier de leur schéma de croissance normal et se multiplier en un groupe de cellules anormales qui produisent toutes le même anticorps, appelé protéine M. Cependant, il n'y a aucune preuve de ce lien entre l'inflammation chronique et la maladie de Kahler.

5 Physiopathologie

La physiopathologie fait référence aux changements dans les processus corporels qui résultent d'une maladie. Dans le cas du myélome multiple

(MM), qui est un type de cancer de la moelle osseuse, la physiopathologie est complexe. Elle peut entraîner des problèmes osseux, sanguins, rénaux et parfois neurologiques.

La progression du MM commence par l'état pathologique précurseur asymptomatique de la MGUS. Des données récentes indiquent que la MGUS, précédemment caractérisée par la croissance des cellules myélomateuses sans destruction osseuse ni atteinte d'autres organes, est en fait associée à des altérations osseuses [18].

Dans le cas du myélome multiple, les plasmocytes anormaux prolifèrent et évincent les cellules sanguines saines. Au lieu de produire des anticorps utiles, ces cellules cancéreuses produisent une protéine anormale appelée paraprotéine ou protéine M, qui ne joue aucun rôle dans l'immunité et peut entraîner des complications. Cette prolifération plasmocytaire s'accompagne d'une inhibition de la lymphopoïèse B normale, qui est responsable d'une diminution du taux d'immunoglobulines polyclonales et, par conséquent, d'un risque accru d'infection. À son tour, cela génère également une suppression de l'hématopoïèse normale [14].

La physiopathologie du MM a un impact étendu sur l'organisme. C'est pourquoi une personne peut avoir besoin de plusieurs analyses de sang, de tests d'imagerie et de biopsies pour que les médecins puissent diagnostiquer le MM et déterminer un plan de traitement adapté.

6 Signes et manifestations cliniques du MM

Dans la majorité des cas, le MM ne cause aucun signe ou symptôme au stade précoce de la maladie. Les patients sont souvent diagnostiqués avec des maladies asymptomatiques d'étiologie inconnue. Une fois la maladie développée, l'augmentation du nombre de cellules myélomateuses, ainsi que les niveaux élevés de protéine monoclonale qu'elles produisent, peuvent provoquer plusieurs symptômes qui sont liés soit à une infiltration de la moelle osseuse par des plasmocytes, soit à des lésions organiques.

6.1 Manifestations osseuses

Elles sont dominées par la douleur, qui est le signe le plus fréquent chez 70 à 90% des cas au moment du diagnostic, parfois des fractures pathologiques et rarement des tassements vertébraux. Ces douleurs sont d'intensité et d'horaire variables, persistantes, non calmées par le repos ni par les antalgiques de palier I, II ou III. Elles peuvent être localisées ou diffuses, notamment au niveau du rachis, de la cage thoracique ou du bassin. Les fractures pathologiques surviennent lors de traumatismes minimes des

membres ou d'un certain type de tassements vertébraux, parfois responsables d'un syndrome compressif neurologique qui constitue une urgence diagnostique et thérapeutique.

6.2 Manifestations hématologiques

Le syndrome anémique est très fréquent (60%) en cas de suspicion de MM. Ses symptômes cliniques varient selon la gravité et comprennent :

- Fatigue et perte d'énergie.
- Peau pâle ou jaunâtre.
- Rythme cardiaque irrégulier et rapide, essoufflement.
- Vertiges ou maux de tête.
- Douleur à la poitrine.
- Mains et pieds froids.

Le dosage de l'hémoglobine dans le sang permet de diagnostiquer l'anémie. D'autres analyses sanguines sont utiles pour comprendre ses causes.

6.3 Syndrome d'hyperviscosité

Chez certains patients, de grandes quantités de protéines du myélome peuvent provoquer un "épaississement" du sang. Cet épaississement est appelé hyperviscosité. Les signes d'hyperviscosité sont extrêmement rares, mais s'ils surviennent, ils peuvent ralentir le flux sanguin vers le cerveau et provoquer une asthénie, des maux de tête, de la confusion, des étourdissements, des troubles de la conscience, des bourdonnements d'oreilles, des troubles de la vision, ainsi que des symptômes d'un accident vasculaire cérébral, tels qu'une faiblesse d'un côté du corps et des troubles de l'élocution.

6.4 Manifestations neurologiques

Elles sont un type de compression médullaire ou tronculaire (sciatique, cruralgie), au niveau de la gaine nerveuse par la prolifération plasmocytaire, par contiguïté liée à l'atteinte osseuse et/ou par une vertèbre fracturée [6]. La compression médullaire se manifeste cliniquement par des signes radiculaires, puis une paraplégie qui nécessite, après une IRM, un traitement chirurgical ou une radiothérapie en urgence.

6.5 Manifestations rénales

La complication rénale la plus fréquente du myélome est associée à une néphropathie tubulo-interstitielle par précipitation intratubulaire des

chaînes légères d'immunoglobulines monoclonales. Cela provoque une faiblesse, un essoufflement et un gonflement des membres inférieurs en raison de la rétention d'eau [19]. Cliniquement, la tubulopathie myélomateuse se présente comme une insuffisance rénale de constitution progressive qui est souvent aggravée par la déshydratation, l'infection, l'hypercalcémie, les médicaments néphrotoxiques, en particulier l'administration intraveineuse de produits iodés. La persistance d'une insuffisance rénale impacte fortement la survie des patients atteints du MM.

6.6 Manifestations infectieuses

Un système immunitaire affaibli est responsable de l'apparition de complications infectieuses fréquentes et sévères, qui sont souvent la cause de décès chez les patients atteints de myélome multiple. La prolifération de plasmocytes malades à l'intérieur de la moelle et les traitements (corticoïdes et chimiothérapie) sont des causes majeures de ce désordre des défenses immunitaires. Ces infections sont le plus souvent bactériennes, respiratoires dans 50% des cas liées au Streptococcus pneumoniae, au Staphylococcus aureus, à l'Haemophilus influenzae. Elles peuvent être rénales dans 30% des cas, liées à l'Escherichia coli, au Pseudomonas, au Proteus ou au Klebsiella, ou systémiques dans 8% des cas [6].
Ces infections doivent être traitées sans hésitation avec des antibiotiques car elles peuvent dégénérer très vite. La vaccination contre certaines infections bactériennes ou virales (pneumocoque et grippe) est également recommandée.

7 Diagnostic et bilans initiaux

- Le processus diagnostique du MM commence habituellement par une visite chez le médecin à cause d'un ensemble de symptômes et de douleurs gênantes (douleurs osseuses, asthénie, etc.).
- Après un examen physique et en se basant sur les informations extraites, un bilan NFS (Numération de la Formule Sanguine) est réalisé afin de voir la quantité des globules rouges, des globules blancs et des plaquettes.
- Par la suite, un examen des frottis sanguins périphériques est fait afin d'observer s'il y a des rouleaux érythrocytaires [3] et de rechercher des plasmocytes circulants [20]. Après cela, une analyse de la vitesse de sédimentation (VS) est faite pour détecter s'il y a une inflammation,

3. Il s'agit d'un phénomène dans lequel les érythrocytes (globules rouges) de taille variable sont regroupés et disposés en piles de plaques. Cela est dû à la présence de protéines de haut poids moléculaire dans le plasma.

une infection, etc.
- Un examen d'électrophorèse des protéines est obligatoire afin d'analyser le mélange des immunoglobulines (les anticorps) dans le sang, en cherchant un pic monoclonal.
- En cas de production excessive d'anticorps, certains patients vont avoir une sécrétion également dans les urines, donc des analyses urinaires sont très importantes. Et d'autant plus lorsqu'il y a ce qu'on appelle les chaînes légères libres dans les urines. Pour la caractérisation et l'identification des immunoglobulines monoclonales, une technique d'immunofixation est utilisée.
- Une ponction de la moelle osseuse ou un myélogramme est un autre examen essentiel pour le diagnostic du MM. Cet examen permet de quantifier les plasmocytes présents (<10%), d'identifier et de détecter ceux qui sont anormaux [21]. Parfois, une biopsie de la moelle osseuse est réalisée afin de voir si celle-ci contient des cellules cancéreuses.
- De plus, des radiographies ou un scanner permettent de visualiser les anomalies caractéristiques dans les os.
- La dernière série d'examens qui est importante dans le diagnostic de MM, ce sont toutes les examens biologiques de sang et d'urine qui vont permettre d'évaluer le fonctionnement de certains organes et aussi de détecter des anomalies (calcémie, bilan rénal, ionogramme sanguin, sérologie, bilan hépatique, etc.).

7.1 Bilans hématologiques

7.1.1 Numération de la Formule Sanguine (NFS)

L'hémogramme est le premier examen biologique utilisé pour dépister, explorer et suivre la plupart des hémopathies. Ses indications sont très nombreuses et dépassent largement le cadre des pathologies hématologiques.

Le NFS permet une étude qualitative et quantitative des trois lignées des cellules sanguines circulatoires (GB, GR et les plaquettes). Cet examen est essentiel afin d'évaluer le dysfonctionnement de la moelle osseuse ou des troubles périphériques (anémie, polyglobulie, leucocytose, problèmes de coagulation, etc.). Le NFS est influencé par l'âge, le sexe et les manifestations cliniques. Les résultats de cet examen contiennent :
- **Hémoglobine (Hb)** : l'hémoglobine (Hb) est une protéine transporteuse d'oxygène dans les globules rouges. Le taux s'exprime en grammes pour 100 ml de sang. Sa valeur normale est comprise entre 12 et 16 g/100 ml. Si le taux d'Hb est supérieur à la normale, on pense vers une polyglobulie. Si l'inverse, on pense directement vers

une anémie qu'il faut caractériser.

- **Hématocrite (Hct)** : l'hématocrite désigne le pourcentage relatif du volume des globules rouges par rapport au volume total du sang. Ce chiffre permet, entre autres, le calcul du volume globulaire moyen (VGM) et la concentration corpusculaire moyenne en hémoglobine (CCMH) [22]. Les valeurs normales sont :
  - Chez l'homme : **40 à 52%**
  - Chez la femme : **37 à 46%**
- **Volume Globulaire Moyen (VGM)** : il s'agit du volume moyen qu'occupent les globules rouges (hématies) au sein d'un échantillon sanguin donné. Cette mesure est exprimée en fL (femtolitre) ou en microns cube. Elle est calculée comme suit :

$$VGM = \frac{Hct}{nbrs_{GR}}$$

  Le VGM est normalement compris entre 80 et 100 μ^3. Sous le seuil de 80, on parle de **microcytose** et au-dessus de 100 de **macrocytose**.
- **Teneur Corpusculaire Moyenne en hémoglobine (T.C.M.H)** : c'est la masse moyenne d'hémoglobine contenue dans une seule globule rouge. Elle est calculée comme suit :

$$TCMH = Hb/nbrs_{GR}$$

  Cette masse est donc très faible et elle s'exprime en picogramme ; sa valeur normale est comprise entre 27 et 32 pg.
- **La Concentration Corpusculaire Moyenne en hémoglobine**
  CCMH est la quantité d'hémoglobine contenue dans 100 ml d'hématies qui seraient débarrassées du plasma, elle est calculée comme suit :

$$CCMH = Hb(g/100ml)/Hct$$

  Sa valeur normale est comprise entre 32 et 36 g / 100 ml.
  Les valeurs de **TCMH** et **CCMH** permettent de préciser l'origine de l'anémie.
- **Les plaquettes** Le taux de plaquette peut diminuer suite à une atteinte de la moelle osseuse, une maladie immunologique ou la prise de certains médicaments ; il peut au contraire augmenter en présence d'un état inflammatoire. Les valeurs normales sont comprises entre : 150 000 et 400 000 / mm^3.
- **Leucocytes** Il s'agit du nombre des globules blancs. Une augmentation (hyperleucocytose) ou une diminution (hypoleucocytose) de ce nombre peut par exemple signifier une infection bactérienne ou parasitaire, un syndrome inflammatoire, une réaction allergique médica-

menteuse. Une valeur normale est comprise entre : 4000 et 11 000 / mm^3.

– **Formule Leucocytaire** Cette formule étudie la proportion des différents globules blancs. (Voir Table 1)

Globules Blancs (GB)	Valeurs normales, exprimées :	
	en pourcentage (%)	en valeur absolue/mm^3
Polynucléaires neutrophiles	60 à 70	2000 à 8000
Polynucléaires éosinophiles	1 à 3	40 à 400
Polynucléaires basophiles	0.5 à 1	0 à 100
Lymphocytes	20 à 40	1000 à 3000
Monocytes	2 à 10	500 à 800

Table 1 – Globules Blancs

7.1.2 Frottis sanguin périphérique

Cet examen biologique est une étude morphologique du sang qui permet de savoir si l'aspect des éléments sanguins est normal. Il est demandé lorsque les résultats de NFS sont anormaux. Le frottis est obligatoire lors du diagnostic de myélome multiple, dont le but est d'observer s'il y a des rouleaux érythrocytaires.

7.1.3 Myélogramme

Le myélogramme est un examen hématologique qui consiste à faire une ponction de la moelle osseuse pour prélever des cellules afin de pouvoir les analyser. Cet examen est réalisé sous anesthésie locale en effectuant une ponction au niveau du sternum (os plat de la face antérieure du thorax) ou au niveau de la crête iliaque (la partie supérieure d'un des os du bassin). Le myélogramme est prescrit en cas d'anomalies détectées sur la numération de la formule sanguine. Il est systématiquement prescrit en cas de suspicion de cancer hématologique. Le diagnostic du myélome multiple est souvent évoqué sur des examens réalisés à partir d'une ponction de la moelle osseuse (un myélogramme) qui montrent un nombre trop élevé de plasmocytes ($>$ à 10%) [23].

7.1.4 Biopsie ostéomédullaire (BOM)

La biopsie ostéo-médullaire est un examen visant à prélever un fragment de tissu osseux et de la moelle osseuse à l'aide d'une aiguille creuse. Ce prélèvement est effectué le long de l'os iliaque (la partie haute de la fesse), donnant une idée sur l'architecture de la moelle osseuse, la

densité des travées osseuses, la densité cellulaire à l'intérieur des logettes médullaires, et de la présence ou non de fibrose (myélofibrose[4]). Il est généralement indiqué si le myélogramme est difficile à réaliser ou s'il n'est pas concluant. Cet examen est plus douloureux, donc il ne dispense pas d'une anesthésie locale en utilisant plutôt une prémédication par voie veineuse, ainsi qu'un traitement antalgique chez le sujet âgé. Il dure en moyenne un quart d'heure, et il n'est pas nécessaire d'être à jeun.

7.2 Imagerie

La prolifération de plasmocytes malins affecte l'ensemble du squelette osseux à des degrés divers, provoquant la destruction de certains types de cellules osseuses. Par conséquent, l'imagerie du MM est souvent nécessaire lorsque des complications surviennent, pouvant être révélatrices de la maladie. Les clichés radiographiques normaux sont généralement suffisants pour identifier les lésions lytiques ou les signes d'ostéopénie diffuse, ainsi que les complications fracturaires telles que les tassements vertébraux. Mais parfois, une IRM ou une TEP est demandée afin de rechercher une compression médullaire ou de planifier une intervention chirurgicale, et aussi pour évaluer l'étendue de la maladie et la réponse au traitement.

7.2.1 Radiologie conventionnelle

Les lésions lytiques sur les radiographies conventionnelles, basées sur l'absorption des rayons X, sont typiquement des lésions rondes à l'emporte-pièce, sans reconstruction, bien visibles sur la voûte crânienne, l'os iliaque ou sur les os longs, essentiellement les fémurs et les humérus. Un bilan complet peut être nécessaire dès le diagnostic de myélome symptomatique. Il comprend : cliché de crâne face plus profil, rachis cervical dorsal et lombaire face plus profil, gril costal, bassin de face et os longs, humérus et fémur seulement.

7.2.2 Tomodensitométrie (TDM-Scan)

La TDM-Scan permet la détection de petites lésions osseuses dans le myélome qui ne sont pas visibles sur les radiographies standards. Elle fournit une excellente reconstruction des images 3D. De plus, le scanner peut montrer l'étendue de lésions extra-osseuses de type plasmocytomes extra-médullaires [24].

4. La myélofibrose est une pathologie dans laquelle du tissu fibreux remplace progressivement les cellules souches sanguines dans la moelle osseuse.

7.2.3 Imagerie par Résonance Magnétique (IRM)

Cet examen est devenu très important dans l'évaluation des lésions du myélome. Il est plus sensible que la radiologie conventionnelle. L'IRM permet une discrimination entre une moelle normale et une moelle envahie, un diagnostic très précis en cas de suspicion de compression médullaire ou de compression neurologique, avec une très bonne visualisation des masses extra-médullaires, . . .etc. [24]

7.2.4 Imagerie par Transmission et Émission de Positons (TEP)

La TEP est une technique d'imagerie médicale performante qui permet une cartographie plus exhaustive des métastases osseuses.
L'intérêt du TEP scanner au 18-FDG (FDG-PET) dans la prise en charge des patients atteints de myélome multiple (MM) pour le bilan de diagnostic et pour l'évaluation thérapeutique a été récemment démontré. C'est un outil d'imagerie puissant pour la détection des lésions osseuses lors du diagnostic initial, avec des valeurs de sensibilité et de spécificité élevées.
Une étude comparative faite sur le TEP scanner et l'IRM dans le myélome multiple (MM), en ce qui concerne le nombre de lésions osseuses au moment du diagnostic [25], a montré que les deux techniques sont tout aussi efficaces pour identifier le nombre de lésions osseuses, mais TEP a eu un meilleur impact pronostique en termes de PFS et d'OS.

7.2.5 Scintigraphie osseuse

La scintigraphie osseuse est un examen isotopique basé sur le même principe que le PET Scan. Il consiste à injecter dans l'organisme un produit légèrement radioactif (un isotope) qui se fixe sur les zones de forte activité métabolique osseuse. Cela inclut les tumeurs et les métastases osseuses, qui ne sont pas simplement des masses inertes mais des groupes de cellules qui se divisent rapidement et de manière incontrôlable, consommant beaucoup d'énergie.

Cependant, la scintigraphie au technétium 99 est très peu utilisée en raison de sa sensibilité médiocre et de la présence de nombreux faux négatifs [6]. Elle est réservée aux diagnostics difficiles pour éliminer d'autres pathologies néoplasiques ou infectieuses.

7.3 Bilans protidiques

Le bilan protidique est un ensemble d'examens sanguins qui permettent d'évaluer les protéines présentes dans le sang et de détecter des anomalies dans leur quantité ou leur qualité.

7.3.1 Vitesse de Sédimentation (VS)

La vitesse de sédimentation est un examen qui permet de mesurer la distance parcourue par les hématies quand elles sédimentent dans un tube vertical pendant un temps donné. Ce test est utilisé pour déceler et surveiller les maladies s'accompagnant d'un syndrome inflammatoire ou infectieux. Elle est mesurée à deux moments : une heure et deux heures après un prélèvement sanguin à jeun. Ses valeurs normales sont comme suit :

$1^{ere} heure < 7mm$ $2^{eme} heure < 20mm$

La valeur du VS peut varier en fonction de certains facteurs, comme : l'âge (>45 ans), la grossesse, les médicaments (les anti-inflammatoires diminuent la VS, les œstrogènes l'augmentent). Chez les patients atteints de maladie de Kahler (myélome multiple), la vitesse de sédimentation est généralement accrue.

7.3.2 Protéine C-réactive (CRP)

La protéine C-réactive est une protéine sécrétée par le foie sous contrôle de l'IL-6 en cas d'infection ou d'inflammation aiguë ou chronique dans l'organisme. Son test sanguin est l'une des analyses les plus régulièrement demandées lors d'une prise de sang. La valeur normale de CRP doit être inférieure à 6 mg/L. Un taux supérieur à cette valeur peut être le signe d'une infection banale, ou d'autres pathologies telles que la pyélonéphrite, les maladies néoplasiques, les cancers, les maladies auto-immunes, etc. Les résultats peuvent varier selon les laboratoires et les techniques qu'ils utilisent. Il est important de savoir que ces résultats seuls ne constituent pas un diagnostic et que le médecin doit prescrire des examens complémentaires ou un traitement éventuel.

La CRP est un biomarqueur qui reflète l'activité du myélome multiple et indique un mauvais pronostic [26]. Sa production par l'interleukine IL-6 est stimulée en grande quantité par les cellules stromales du microenvironnement tumoral et par les plasmocytes malins eux-mêmes, constituant le principal facteur de croissance de la prolifération des cellules myélomateuses. Les myélomes multiples en rémission ont une CRP plus faible que les myélomes qui rechutent. Cependant, la protéine C-réactive n'est pas un facteur spécifique de l'activité de la maladie, car elle peut être augmentée et modifiée par de nombreux autres facteurs, tels que le tabagisme, le stress, l'obésité, l'hypertension, etc. Par conséquent, l'utilisation de la CRP comme marqueur diagnostique pour le myélome multiple doit être interprétée avec prudence et doit être combinée avec d'autres tests diagnostiques pour une évaluation clinique précise. [6].

7.3.3 Électrophorèse des protéines sériques (EPP)

L'électrophorèse des protéines sériques (EPP) est une technique d'analyse d'un mélange de protéines qui permet d'identifier et de séparer les protéines en soumettant le mélange à l'action d'un champ électrique en fonction de leurs charges, de leurs tailles, ou des deux à la fois.

Cet examen est utilisé en immunologie, notamment pour confirmer le diagnostic de certaines maladies du système immunitaire, d'infections et de certains types de cancers, en particulier le myélome multiple. Deux techniques d'électrophorèse des protéines sériques sont disponibles :

- L'électrophorèse en gel d'agarose.
- L'électrophorèse capillaire.

Le but de cet examen est de rechercher des immunoglobulines monoclonales dans le sérum. L'EPP permet de confirmer la monoclonalité sans ambiguïté [27] en mettant en évidence six fractions protéiques :

a/- L'albumine (AL) : est la protéine la plus abondante dans le sang (60%) et est fabriquée par les hépatocytes. Cette protéine joue un rôle majeur dans le maintien de la pression oncotique du sang.

b/- α-1 globulines et α-2 globulines : reflètent la production d'organismes des protéines. Ce sont des globulines plasmatiques ayant la plus grande mobilité en EPP à pH=8,6. Le taux des α-globulines augmente au cours des maladies inflammatoires et néoplasiques.

c/- β-1 globulines et β-2 globulines : constituent entre 9 et 15% du plasma sanguin et affichent une mobilité électrophorétique intermédiaire entre celle des α-globulines et celle des γ-globulines.

d/- γ-globuline : est une protéine du plasma sanguin qui migre après les alpha et bêta globulines. Elle est diminuée en cas de déficit de l'immunité hormonale et augmentée en cas d'état inflammatoire, infectieux et de cirrhose. Le taux élevé des gammaglobulines dans le sang peut également être un signe de cancer, comme dans le cas du myélome multiple. L'EPP permet de révéler un pic monoclonal (présence d'une bande étroite) dans la zone de migration des gammaglobulines, le plus souvent (voir figure 3). Ce pic est souvent détecté, au contraire, dans la région des bêta-globulines ou alpha-globulines, plus rarement [28].

La confirmation de la présence d'une gammapathie monoclonale doit être faite par immunofixation ou immuno- soustraction sérique ou urinaire. Les valeurs de référence pour les six fractions protéiques qui sont mise en évidence par l'EPPs sont dans le tableau suivant (voir Table2) :

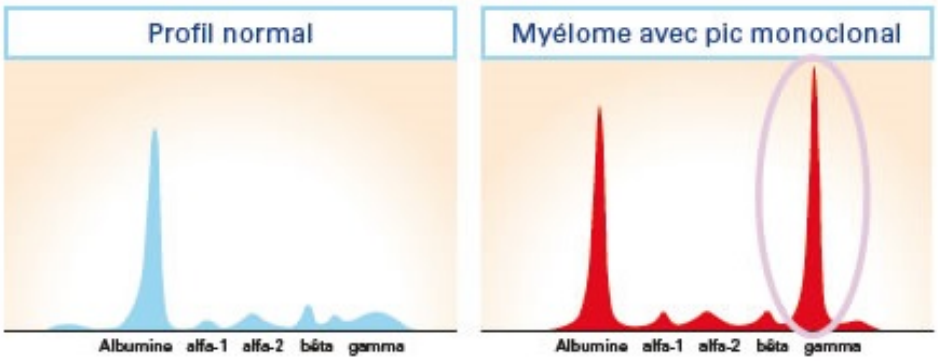


Figure 3 – EPPs (profil normal et en cas de MM)

Fractions	valeur en %	valeur absolue en g/l
Albumine	55.8 - 66.1	40.2 - 47.6
α-1	2.9 - 4.9	2.1 - 3.5
α-2	7.1 - 11.8	5.1 - 8.5
β-1	4.7 - 7.2	3.4 - 5.2
β-2	3.2 - 6.5	2.3 - 4.7
γ	11.1 - 18.8	8.0 - 13.5

Table 2 – Les valeurs de référence pour les six fractions protéiques

7.3.4 Immunofixation des protéines (IFx)

L'immunofixation des protéines est une technique immunochimique connue depuis 1969, qui repose sur le principe de l'électrophorèse des protéines, découvert dès les années 1930 par Tiselius. Elle permet de révéler et d'identifier de manière qualitative les immunoglobulines monoclonales dans le sérum, les urines et éventuellement le liquide céphalorachidien [29]. Cette technique est largement utilisée dans les laboratoires d'analyses médicales et l'interprétation des résultats est généralement facile, bien que certaines situations puissent poser des problèmes d'interprétation.
L'IFx est souvent utilisée pour diagnostiquer le myélome multiple ou la macroglobulinémie de Waldenström lorsque les symptômes de ces troubles sont présents. Elle est effectuée sur un échantillon de sang en utilisant une technique de précipitation : des anticorps spécifiques à chaque type d'immunoglobuline sont déposés sur un gel après application d'un courant électrique qui permet de les séparer selon leur taille. Ce phénomène de précipitation est visible à l'œil nu ou avec un appareil.
Les techniques d'immunofixation sont plus sensibles que les méthodes d'électrophorèse des protéines, elles peuvent détecter de faibles bandes monoclonales qui ne sont pas visibles à l'électrophorèse [27]. Un résultat négatif indique qu'il n'y a pas des immunoglobulines (Ig) anormales.

7.3.5 Analyse des chaînes légères libres sériques (CLL ou Free light chains FLC)

Ce qui est plus un obstacle pour les cliniciens que pour les chercheurs, c'est le manque de biomarqueurs, car leur présence est essentielle pour diagnostiquer les maladies, ainsi que pour surveiller l'efficacité et la réponse au traitement. Pour les gammapathies monoclonales, notamment le MM, des biomarqueurs sont présents, ce qui est l'une des caractéristiques de ces maladies. Ces biomarqueurs sont des immunoglobulines produites par les plasmocytes, représentées par la lettre **"Y"** (voir figure 4), et constituées de deux types de chaînes : les chaînes lourdes et les chaînes légères.

Il existe cinq types de chaînes lourdes, chacun étant nommé par une lettre spécifique (G, A, D, E et M). En revanche, il n'existe que deux types de chaînes légères (kappa (κ) et lambda (λ)).

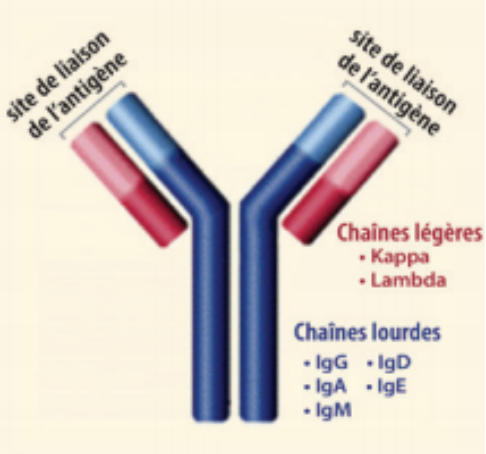


Figure 4 – Structure d'une Immunoglobuline (Anticorp)

Chaque immunoglobuline est constituée d'un type de chaîne lourde et d'un type de chaîne légère. Ainsi, au total, il y a seulement 5 sous-types d'immunoglobulines normales pour les chaînes lourdes (G, A, D, E et M) et 2 sous-types pour les chaînes légères (Kappa (κ) et Lambda (λ)) (voir Table3).

Lorsque les chaînes légères sont associées aux chaînes lourdes, elles sont appelées chaînes légères liées. En revanche, lorsqu'elles ne sont pas liées à des chaînes lourdes, elles sont appelées chaînes légères libres (Free light chains FLC).

chaine lourde	chaine légère	sous-type d'Ig
G	κ	**IgG** κ
	λ	**IgG** λ
A	κ	**IgA** κ
	λ	**IgA** λ
D	κ	**IgD** κ
	λ	**IgD** λ
E	κ	**IgE** κ
	λ	**IgE** λ
M	κ	**IgM** κ
	λ	**IgM** λ

Table 3 – Liste des sous-types d'immunoglobulines

En cas de prolifération anormale des plasmocytes, ces derniers produisent pour des raisons inconnues un nombre très élevé de chaînes légères, dépassant le nombre nécessaire à la fabrication des immunoglobulines. Ces chaînes légères excédentaires entrent dans la circulation sanguine sous forme de chaînes légères libres.

Le dosage sérique des chaînes légères libres (également appelé "analyse FreeLite") est un test sanguin qui détecte ces biomarqueurs afin de diagnostiquer et de surveiller les pathologies plasmocytaires telles que le myélome multiple, ainsi que pour évaluer l'efficacité du traitement [30]. Les résultats doivent toujours être interprétés en conjonction avec ceux de l'EPP.

Les chaînes légères libres sont normalement présentes en faible quantité dans le sang. Les valeurs de référence sont les suivantes :

Le taux de chaînes légères κ : 3,3-19,4 mg/L.

Le taux de chaînes légères λ : 5,7-26,3 mg/L.

Le rapport κ/λ devrait être d'environ 0,26 à 1,65.

7.3.6 Protéinurie des 24 heures

Chaque jour, 10 à 15 g des protéines sériques traversent les reins, mais seulement 100 à 150 mg sont excrétés dans les urines des 24 heures [31]. La protéinurie est définie par la présence de quantités anormales de protéines dans les urines. La recherche et le dosage de ces protéines renseignent sur le bon fonctionnement des reins. Les résultats de ce test sont exprimés en grammes/24 heures et ne doivent pas dépasser 0,15 g/24 heures. Une augmentation du taux de ces protéines dans l'urine peut être due à l'effort, au myélome multiple ou à une atteinte rénale.

7.3.7 Protéinurie de Bence Jones

Dans la majorité des cas de myélome multiple, des chaînes légères d'immunoglobulines sont anormalement produites et excrétées dans les urines en raison de leur faible poids moléculaire. Celles-ci sont appelées protéines de Bence Jones. Elles sont constituées de chaînes légères libres (Free light chains) monoclonales d'immunoglobulines (Ig) d'isotype kappa (κ) ou lambda (λ) et ont un poids moléculaire de 22-24 kDa (22000-24000 Dalton) [32].
La présence de cette protéine constitue un argument très fort pour le diagnostic du myélome multiple, et son taux permet de suivre l'évolution de la maladie et l'efficacité du traitement. Les résultats obtenus sont dit anormaux en cas de présence de protéines de Bence-Jones dans les urines.

7.4 Bilans biochimiques complémentaires

7.4.1 Bilan phosphocalcique

Le calcium est le minéral le plus abondant dans le corps humain. 99% du calcium de l'organisme contribue à la formation et à la solidité des os et des dents. Le 1% restant est important pour la coagulation sanguine, la contraction musculaire, la conduction nerveuse et la libération d'hormones. Sa valeur normale est comprise entre 2.2-2.6 mmol/l soit 85-105 mg/l.

L'hypercalcémie aiguë et chronique est un état caractérisé par l'augmentation anormale du taux de calcium dans le plasma. Quatre causes sont à l'origine de 90% des hypercalcémies : cancer avec ou sans métastase, myélome multiple, hyperparathyroïdie primaire, intoxication à la vitamine D.
L'hypercalcémie supérieure à 110 mg/l soit 2.75 mmol/l est la complication métabolique la plus fréquente chez 10 à 30% des patients atteints de myélome [33]. Elle peut avoir un impact sur l'évolution de la maladie et le taux de survie global. La principale cause de l'hypercalcémie est l'ostéolyse induite par les plasmocytes malins.

7.4.2 Ionogramme sanguin

L'ionogramme sanguin est l'un des examens de laboratoire les plus demandés. Il correspond au dosage des principaux constituants ioniques du sang :

- **Des ions positifs (des cations) :** tels que le sodium (Na^+), le potassium (K^+), le magnésium (Mg^{2+}), etc.

- **Des ions négatifs (des anions) :** tels que le chlorure (Cl^-), les bicarbonates HCO_3^-, les phosphates HPO_4^{2-}, etc.

Ce bilan sanguin sert à surveiller l'équilibre hydro-électrolytique de l'organisme, qui est assuré par les reins, la peau, la respiration et le système digestif.
Un déséquilibre hydro-électrolytique peut avoir des conséquences sur le métabolisme ou sur le contrôle des apports hydriques lors de perfusion. Le tableau 4 présente les valeurs de référence des principaux composants ioniques du sang qui sont les plus demandés dans le cas du myélome multiple.

Ions	valeur normale
Sodium (Na^+)	135 - 145 Meq/L
Potassium (K^+)	3.5 - 5 mmol/L
Calcium (Ca^+)	90 - 100 mg/L

Table 4 – Les valeurs de référence des principaux composants ioniques du sang

7.4.3 Bilan rénal

Les reins ne sont pas simplement des filtres placés sur la circulation sanguine : ils ont également de nombreux rôles essentiels au fonctionnement du corps humain, notamment l'épuration du sang, la régulation de l'équilibre en eau et en sels minéraux, ainsi que la production d'hormones, d'enzymes et de vitamines.

L'insuffisance rénale est l'un des problèmes les plus courants et les plus graves qu'un patient atteint de myélome multiple peut rencontrer. Cela est dû aux fragments d'immunoglobulines excrétés dans l'urine qui endommagent les reins et les empêchent de remplir correctement leur fonction de filtration. Des examens médicaux peuvent être réalisés pour évaluer la fonction rénale, qui peuvent être effectués ensemble ou séparément dans le sang ou dans les urines.

A/- Urée : est une molécule résultant de la dégradation des protéines formées dans le foie à partir de l'ammoniac et excrétées par les reins. Leurs valeurs normales vont de :
- Pour l'urée urinaire : 15-35 g/24 heures.
- Pour l'urée sanguine : 0.18-0.50 g/L chez l'homme et 0.15-0.50 g/L chez la femme.

B/- Créatinine (créat) : n'est pas un composant très connu bien qu'elle soit présente dans le sang, elle est produite lors d'un effort par la dégradation de la créatine qui est une protéine fabriquée par le foie et stockée dans les muscles. Lorsque tout va bien, la créatinine est

régulièrement éliminée par les reins et seule une petite quantité reste dans le sang.

Le dosage de la créatinine fournit des informations sur la fonction rénale et est souvent prescrit dans le cadre d'un bilan de santé systématique ou pour les personnes atteintes de certaines maladies chroniques pouvant avoir des répercussions sur les reins, comme le diabète et l'hypertension artérielle. Il est le principal test effectué pour diagnostiquer et surveiller la progression de la maladie rénale associée au myélome et pour établir la gravité de la maladie [34].

Un taux élevé de créatinine est souvent le signe d'une insuffisance rénale, et un faible taux peut être le signe d'une myopathie (atrophie musculaire sévère). Les valeurs normales sont les suivantes [22] :

a/ Dans le sang :

6 à 11 mg/L chez la femme (soit 50 à 100 µmol/L).

7 à 14 mg/L chez l'homme (soit 65 à 120 µmol/L).

b/ Dans les urines de 24 heures :

8 à 16 mmol chez la femme.

9 à 18 mmol chez l'homme.

C/- Clairance de créatinine : est un test supplémentaire qui aide à déterminer la cause de l'insuffisance rénale. Il s'agit de calculer le rapport entre la créatinine présente dans le sang et celle retrouvée dans les urines après filtration rénale. Selon la formule utilisée, il est nécessaire de recueillir les urines de 24 heures.

Un taux élevé de clairance de créatinine peut être le signe d'une insuffisance rénale, de leucémie, de myélome multiple, etc. Au contraire, un faible taux peut s'observer chez des personnes souffrant d'une myopathie.

L'estimation de sa valeur peut s'effectuer selon différentes formules qui prennent en compte l'âge, le poids et la couleur de la peau. Parmi les formules les plus utilisées, nous avons :

La formule de Cockcroft : permet le calcul de la clairance uniquement à partir d'un prélèvement sanguin de la créatinine selon la formule suivante :

$$clairance = \frac{k * poids * (140 - age)}{creatinine}$$

Avec : **k=1.23** chez l'homme et **k=1.04** chez la femme.

Cette formule n'est pas fiable chez l'enfant, la femme enceinte, les sujets obèses ou âgés (>65 ans).

La formule de MDRD (Modification of Diet in Renal Disease) : dans cette formule, le poids n'est pas nécessaire. Les formules pour les hommes et les femmes sont les suivantes :

– Chez l'homme :

$$result = 186.3 * (creatinine/88.4)^{-1.154} * age^{-0.203}$$

– Chez la femme :

$$result = 186.3 * (creatinine/88.4)^{-1.154} * age^{-0.203} * 0.742$$

– Si le sujet est noir de peau : $result * 1.212$

7.4.4 Bilan hépatique

Un bilan hépatique est un ensemble de tests sanguins qui permettent d'évaluer le fonctionnement du foie et d'identifier certaines pathologies. L'atteinte hépatique au cours d'un MM est généralement due à une amylose hépatique ou à une obstruction des voies biliaires extra-hépatiques. Bien que cela soit exceptionnel et très rare, c'est extrêmement dangereux et rapidement mortel. C'est pourquoi il est important d'effectuer un bilan hépatique lors du diagnostic du MM [35].

1/- **Les Transaminases :** sont des enzymes ayant une activité intracellulaire, présentes dans plusieurs tissus tels que le foie, le cœur, les reins ou les muscles. Il existe deux types :

– **ALAT** (Alanine Aminotransférase) appelée aussi **SGPT** (Sérum Glutamate Pyruvate Transférase) : elle se trouve essentiellement dans le foie, les reins mais également en faible quantité dans les muscles striés et dans les globules rouges. Un taux élevé d'ALAT indique une lésion hépatique. Les valeurs normales sont :

8 à 35 UI/L chez l'homme **6 à 25 UI/L** chez la femme.

– **ASAT** (Aspartate Aminotransférase) appelé aussi **SGOT** (Sérum Glutamo Oxaloacétate Transférase) : se trouve plus spécifiquement dans les muscles striés, les globules rouges et le foie. Une ASAT élevé indique une forme de lésion hépatique, également une forme de lésion musculaire. De plus, si le patient a eu un infarctus du myocarde, le niveau d'ASAT peut être élevé. Ses Valeurs normales sont :

8 à 30 UI/l chez l'homme **6 à 25 UI/l** chez la femme.

2/- **GGT** (Gamma Glutamyl transférase) : est une enzyme présente dans les cellules tapissant les voies biliaires, dans de nombreux organes tels

que les reins et les intestins, mais plus particulièrement dans le foie. Une augmentation du taux de GGT peut être un indice d'anomalie du foie, telle qu'une cirrhose hépatique, une nécrose hépatique, des tumeurs ou cancers hépatiques, une hépatite (virale ou microbienne), une exposition à des substances toxiques ou à des médicaments hépatotoxiques, ou encore une consommation d'alcool.
Chez les hommes, le taux normal de GGT est d'environ **inférieur à 45 UI/L**, tandis que chez les femmes, il devrait être **inférieur à 35 UI/L**. Cependant, il convient de noter que le taux de GGT peut diminuer pendant la grossesse et augmenter avec l'âge à partir de 60 ans.

3/- **PAL** (Les phosphatases alcalines) : sont des enzymes présentes dans tout l'organisme, mais surtout à 90% dans le foie et les os. Leur dosage sanguin est souvent effectué pour diagnostiquer diverses pathologies, en particulier des pathologies hépatiques ou osseuses.

Les valeurs de référence pour un adulte selon les normes de laboratoire sont comprises entre environ 38 UI/L et 125 UI/L. Une diminution des PAL peut survenir en raison d'une insuffisance hépatocellulaire, d'une cirrhose, d'une hépatite, d'une inflammation du foie, d'une anémie pernicieuse, d'une anémie aplasique, d'une hypophosphatasie, etc.

Il est normal que le taux des PAL soit élevé pendant la grossesse et qu'il soit plus élevé chez les enfants jusqu'à l'adolescence. En dehors de cela, lorsque le taux de GGT et de PAL est élevé, cela indique fortement que le patient a une forme de cholestase. D'autre part, si le taux de PAL est élevé seul et que les valeurs de GGT sont normales, cela peut signifier que le patient présente une forme de dégradation osseuse accrue dans le corps.

7.4.5 Lactate-déshydrogénase (LDH)

LDH est une enzyme très importante pour transformer les sucres en énergie en catalysant la conversion du lactate en acide pyruvique. Elle est 100 fois plus élevée dans les globules rouges que dans le plasma et est largement exprimée dans les tissus de l'organisme tels que les cellules sanguines, le muscle cardiaque, etc. Le dosage de LDH est souvent prescrit lors du diagnostic d'un myélome multiple pour comprendre jusqu'où le cancer s'est propagé dans le corps. Sa valeur normale est comprise entre 190 UI/L et 400 UI/L, mais peut varier en fonction de l'âge et de la méthode de dosage.

7.4.6 Bilan d'hémostase

L'hémostase est le processus qui consiste à maintenir le sang à l'intérieur d'un vaisseau sanguin endommagé en empêchant les saignements de se produire. Les mécanismes d'hémostase sont très efficaces pour traiter les blessures dans les petits vaisseaux sanguins, y compris les artérioles, les veinules et les capillaires, qui sont souvent rompus lors de traumatismes mineurs de la vie quotidienne et constituent donc la source la plus courante de saignement.
L'exploration de l'hémostase repose sur deux examens principaux qui permettent d'apprécier la coagulation du sang :

1/- **TP** (Taux de prothrombine) : est appelé aussi "temps de Quick" (TQ), il s'agit d'un test exprimé en secondes ou en pourcentage avec un taux normalement compris entre 70% et 100%. La mesure de TP explore la voie extrinsèque de la coagulation. Si la valeur de TP baisse, cela signifie que la coagulation est plus lente et le sang est plus fluide.

2/- **TCA** (Temps de céphaline activée) : est le temps de coagulation, mesuré à 37 °C, d'un plasma pauvre en plaquettes citraté après addition de céphaline et d'un activateur. Cet examen permet de mesurer la fonctionnalité de la voie intrinsèque. Lorsque l'échantillon met plus de temps que la normale pour coaguler, le TCA est dit "allongé". Sa valeur normale est comprise entre 24 secondes et 41 secondes.

7.4.7 Dosage de la glycémie

La glycémie correspond à la concentration de glucose dans le sang. La mesure de la glycémie permet de savoir s'il y a une régulation adéquate du taux de sucre dans le sang. Sa valeur peut varier en fonction des apports et des besoins énergétiques de chaque personne. Cet examen est prescrit lorsqu'une hyperglycémie est suspectée, ce qui peut permettre de détecter un diabète, mais est également prescrit pour détecter une hypoglycémie.

La glycémie normale à jeun doit être d'environ 0,7 à 1,1 g/L. Deux heures après un repas, elle doit être d'environ 1 à 1,4 g/L.

7.5 Autres bilans

7.5.1 β_2 microglobuline (B2M) :

B2M est une protéine non glycosylée présente à la surface de nombreuses cellules, en particulier les lymphocytes et toutes les cellules cancéreuses. Elle a un faible poids moléculaire (11 800 Daltons) [36]. Le test de B2M

aide à déterminer le stade du cancer et l'efficacité du traitement (marqueur pronostique initial et de suivi thérapeutique). Il est effectué lors de l'évaluation de certains types de cancer affectant les globules blancs, notamment la leucémie lymphoïde chronique, le lymphome non hodgkinien et le myélome multiple [37]. Tout traitement en cours doit être signalé avant de faire le prélèvement, car de nombreux médicaments peuvent modifier le taux de B2M dans le sang, notamment les traitements ayant des effets secondaires indésirables sur les reins, comme certains antibiotiques. L'échantillon à analyser est obtenu soit par le sang veineux via une ponction au pli du coude, parfois les urines des 24 heures, ou plus rarement un échantillon de liquide céphalo-rachidien (LCR).

Les valeurs de référence de B2M peuvent varier selon les différentes techniques et méthodes de dosages utilisées. Chez un adulte et par immunonéphélémétrie :

- *Dans le sang* : <2,5 mg/L
- *Dans les urines* : <0,37 mg/24 heures ou <0,28 mg/g de créatinine.
- *Dans le LCR* : <2,3 mg/L.

7.5.2 Dosage de la ferritine

La ferritine est une protéine qui stocke le fer et régule son absorption intestinale. Il est souvent préférable de doser la ferritine plutôt que le taux de fer dans le sang car elle est plus représentative. Le dosage de ferritine aide au dépistage précoce d'une carence en fer ou d'une surcharge dans l'organisme, et permet également de contrôler l'efficacité du traitement prescrit. Sa valeur est augmentée en cas de réaction inflammatoire.

7.5.3 Fibrinogène

Le fibrinogène est une protéine présente dans le plasma sanguin, qui joue un rôle important dans la formation des caillots. Sous l'action de la thrombine, le fibrinogène se transforme en fibrine, une protéine insoluble essentielle à la coagulation du sang. Le test de fibrinogène permet de détecter des syndromes inflammatoires (infections, cancer, lymphomes, rhumatismes, etc.), des syndromes hémorragiques ou une dysfibrinogénémie au cours de thromboses veineuses. Les valeurs normales de fibrinogène dans le sang sont comprises entre environ 2 et 4 g/L.

7.5.4 Sérologie virale

La sérologie virale est un ensemble de tests sanguins qui permettent de rechercher et étudier les anticorps dans le sang correspondant à des

maladies virales données, reflétant l'immunité individuelle. Elle est demandée lors du diagnostic du MM en raison de sa sensibilité. Parmi les tests sérologiques les plus prescrits, nous avons :

a/* **La sérologie HCV** (Hepatitis C Virus Anticorps) ou VHC (virus d'hépatite C) qui consiste à rechercher la présence d'anticorps anti-HCV, ce qui signifie la présence d'une infection au virus d'hépatite C. En l'absence de contact récent ou ancien avec le VHC, il n'y a normalement pas d'anticorps anti-VHC dans le sang.

b/* **La sérologie Ag HBs :** L'antigène du virus de l'hépatite B (Ag HBs) est utilisé pour détecter une infection par le virus de l'hépatite B (VHB). Un résultat négatif peut indiquer que la personne n'a pas été infectée par ce virus, qu'elle en a guéri ou que son système immunitaire a éradiqué la souche virale.

c/* **La sérologie VIH** (virus de l'immunodéficience humaine) est un type de virus qui peut causer une maladie appelée SIDA [5]. Un résultat négatif signifie l'absence d'une infection par ce virus.

7.5.5 Examens cardiaques

L'évaluation de la fonction cardiaque est essentielle dans la prise en charge des patients atteints du myélome multiple. Dans de rares cas, le myélome multiple peut causer des troubles cardiaques, notamment en raison de l'hypercalcémie qui peut entraîner des complications mortelles. Les techniques d'imagerie cardiaque permettent d'estimer la fraction d'éjection (FE) du ventricule gauche et d'évaluer la fonction cardiaque.

7.5.6 Examen électromyographique (EMG)

La neuropathie périphérique fait référence à l'ensemble des maladies des nerfs appartenant au système nerveux périphérique. Dans le cas du myélome multiple, la neuropathie périphérique (NP) est souvent détectée en raison des dépôts endo-neuraux d'immunoglobulines. Un examen électromyographique (EMG) peut être prescrit pour étudier la fonction des nerfs et des muscles du système nerveux périphérique.

5. Le syndrome d'immunodéficience acquise (SIDA) est le dernier stade de l'infection par le VIH

8 Critères diagnostiques

En 2003 [38], le groupe international du travail sur le myélome (IMWG [6]) a publié les critères de diagnostic du myélome multiple. Il est important de savoir que ces critères évoluent constamment et sont mises à jour toutes les quelques années.
Les critères du MM les plus récents ont été publiés par l'IMWG en novembre 2014 [1]. Selon ces critères (Voir le tableau5), pour poser le diagnostic d'un MM symptomatique, il faut obligatoirement retrouver la plasmocytose soit sur un myélogramme >10%, ou sur une biopsie osseuse, ou encore sur un plasmocytome extra médullaire.
A côté de cela, il faut avoir au moins un de ces évènements :

- Un critère du **CRAB** qui traduisent les manifestations cliniques ou biologiques directement liés au MM. L'ensemble de ces éléments sont regroupés dans le tableau 6.
- L'existence d'une protéine monoclonale sérique/urinaire et/ou des plasmocytes clonaux dans la moelle osseuse.

Formes cliniques	Critères de définition
MGUS	- Protéine monoclonale sérique <30g/dl Ou - Plasmocytes médullaires<10% - Absence des critères difinissant le MM
MM indolent	- Protéine monoclonale sérique >=30g/dl Ou urinaire >=500mg/24heures Et/Ou - Plasmocytes médullaires entre 10% et 60% - Absence d'évènements définissant le MM actif ou d'amyloïdose
MM actif	- Au moins un signe CRAB - Plasmocytose médullaire >=60% - Un ratio des chaines légères sériques >=100 - Plus d'une lésion focale à l'IRM

Table 5 – Les critères diagnostiques IMWG 2014 [1]

6. International Myeloma Working Group : est une émanation de la fondation internationale du myélome. Il s'agit d'un consortium international regroupant plus de 200 chercheurs. Ce groupe a élaboré les principales bases de diagnostic de MM, de prise en charge, de critères de réponse, . . .etc.

hyperCalcemia	Ca >2.75 mmol/l (>11 mg/dl) Ou Une augmentation plus de 0.25 mmol/l (1mg/dl) au-dessus de la limite supérieur normale.
Renal failure	Clairance de créatinine <40ml/min Ou créat >20 mg/dl (117 mol/l).
Anemia	Hg <10g/dl Ou Une diminution au moins de 2g/dl sous la limite inférieure normale.
Bone lesion	Au moins une lésion ostéolytique à la radiographie du squelette, scanner ou PET-scan.

Table 6 – Les critères du CRAB

9 Formes cliniques

Selon les critères diagnostique(Voir le tableau5), le myélome multiple passe par de nombreux états précancéreux lors de sa progression (voir figure 5).

9.1 MGUS

Les gammapathies monoclonales de signification indéterminée (MGUS : Monoclonal Gammopathy of Undetermined Significance) sont des troubles plasmocytaires précancéreux qui précèdent souvent l'apparition du myélome chez de nombreux patients. Les personnes atteintes de MGUS ont un petit nombre de cellules myélomateuses dans leur moelle osseuse, mais ces cellules ne forment pas de tumeur et les symptômes du myélome ne sont pas présents. En général, cette condition est découverte lors d'un examen sanguin de routine qui montre des niveaux inhabituels de protéines dans le sang.
Un bilan de contrôle doit être effectué tous les six mois afin de surveiller la maladie et s'assurer qu'elle ne se transforme pas en myélome multiple, même si cela ne se produit que chez un petit nombre de patients.

9.2 Myélome multiple asymptomatique

Le myélome multiple asymptomatique est une forme de myélome caractérisée par des manifestations cliniques et biologiques telles que décrites dans les critères CRAB (Voir tableau6). Il est important de souligner que la majorité des patients diagnostiqués avec un myélome multiple asymptomatique développeront un myélome multiple actif au cours des cinq

prochaines années. Le traitement est généralement recommandé pour les patients présentant un myélome multiple asymptomatique à haut risque de progression. Cependant, pour les patients à faible risque de progression, la surveillance étroite peut être recommandée. Les options de traitement pour le SMM peuvent inclure une chimiothérapie, une transplantation de cellules souches et une thérapie ciblée.

9.3 Myélome multiple actif

Ce type de myélome (appelé aussi MM symptomatique) cause des dommages au corps et doit être traité rapidement. Une personne atteinte de myélome symptomatique présente plus de cellules myélomateuses qu'une personne atteinte de myélome asymptomatique ou de MGUS, et elle répond également à au moins un des critères CRAB (Voir tableau6). Le traitement peut inclure une combinaison de chimiothérapie, de radiothérapie, de greffe de cellules souches et d'autres médicaments. Le choix du traitement dépendra de l'état de santé général du patient et de la gravité de la maladie.

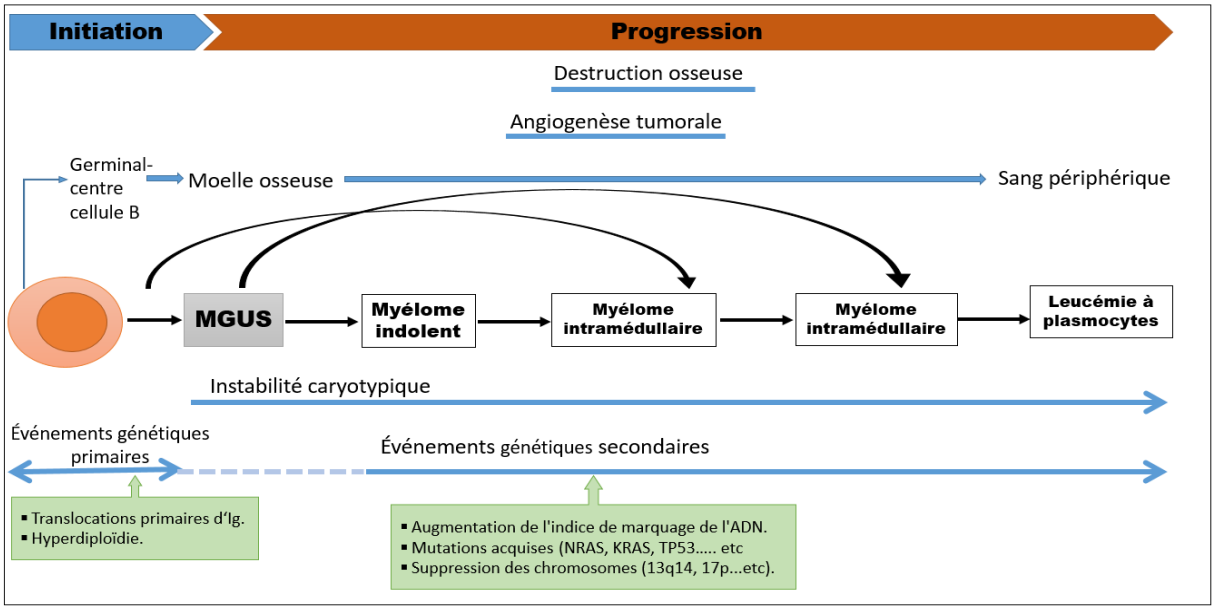


Figure 5 – Initiation et progression du myélome multiple

10 Classifications et facteurs pronostiques

Après le diagnostic d'un myélome multiple, le médecin doit le classer en stades, en fonction de son degré d'évolution, en se basant sur l'un des systèmes de stadification disponibles. Connaître le stade du myélome aide

les médecins à comprendre la gravité de la maladie et à établir un plan de traitement.

10.1 Classification Durie-Salmon

Le système de stadification Durie-Salmon a été développé en 1975 [39]. Il démontre la corrélation entre la masse cellulaire myélomateuse mesurée et les dommages qu'elle a causés. Ce système est basé sur la quantité de myélome dans le corps (le nombre total de cellules myélomateuses) et un ensemble de paramètres cliniques, biologiques et radiologiques, incluant le taux d'hémoglobine, la calcémie, la créatinine, le taux de protéine monoclonale sérique et urinaire, ainsi que le nombre de lésions osseuses.

Selon la gravité de la maladie, les patients sont classés en stade I, II ou III. Ces stades du myélome multiple sont ensuite divisés en sous-classes (classes A et B) en fonction du taux de créatinine dans le sang, qui révèle la fonction rénale (voir Tableau 7).

Stade	Critère
Stade I : Faible masse cellulaire	Tous les éléments suivants : • Valeur d'hémoglobine > **10 g/dL**. • Calcium sérique normal ou <**10,5 mg/dL**. • Structure osseuse normale (échelle 0) ou plasmocytome osseux solitaire uniquement. • Faibles taux de production de composant M (**IgG** < **5g/dL ; IgA** < **3 g/dL**). • Taux des chaînes légères urinaire < **4 g/24h**.
Stade II : Masse cellulaire intermédiaire	Ne convient ni au **stade I** ni au **stade III**.
Stade III : Masse cellulaire élevée	Un ou plusieurs des éléments suivants : • Valeur d'hémoglobine <**8.5 g/dL**. • Calcémie <**12 mg/dL**. • Lésions osseuses lytiques avancées (échelle 3). • Taux de production élevés de composant M (**IgG** > **7 g/dL ; IgA** > **5 g/dL**). • Taux des chaînes légères urinaire > **12 g/24h**.
Sous-classification (soit A ou B)	Sous-classe A : • Fonction rénale normale : Taux de créat < **20 mg/L**.
	Sous-classe B : • Problème rénal : Taux de créat > **20 mg/L**.

Table 7 – Critères de classification Durie-Selmon

10.2 Système international de Stadification (ISS)

ISS (International Staging System) est un nouveau système de stadification qui a été développé en 2005 [40] par le groupe IMWG. Il est simple, basé sur des variables biologiques faciles à utiliser (B2-microglobuline et le taux d'albumine) Les données cliniques et biologiques utilisées dans cette recherche ont été recueillies sur 10750 patients atteints d'un myélome n'ayant jamais été traités, dans 17 établissements (en Amérique du Nord, en Europe et en Asie). Les résultats de cette recherche ont ensuite été validés en démontrant son efficacité :

- Sur des patients d'Amérique du Nord, d'Europe et d'Asie.
- Sur des patients âgés de moins et de plus de 65 ans.
- Avec un traitement standard ou une autogreffe ; et en comparaison avec le système de stadification de Durie-Salmon [39].

L'ISS permet de séparer les patients en 3 groupes à risque, avec des médianes de survie significativement différentes (Voir Tableau8).

Stade	Critère
stade I	• β2M $<$ **3.5 mg/L** • Albumine $>=$ **35g/L**
stade II	• β2M $<$ **3.5 mg/L** et Albumine $<$ **35 g/L** Ou • **3.5 mg/L** $<$ β2M $<$ **5,5 mg/L**, et quelle que soit le taux d'albumine.
stade III	• β2M $>$**5.5 mg/L**

Table 8 – Critères de ISS

10.3 Système international révisé de Stadification (R-ISS)

R-ISS (Revised International Staging System) est un nouveau systeme de stratification avec une puissance pronostique améliorée par rapport aux systeme ISS, il a été publié en Août 2015 par le groupe IMWG [41], ils ont intégrè deux facteurs pronostiques supplémentaires simples, fiables et largement utilisés : marqueurs génétiques (voir Table 9), LDH (voir Table 10). Ce systeme permet d'établir un pronostic du MM nouvellement diagnostiqué, en identifiant trois stades différents (voir Table 11).

Risque	Critère
Risque standard	Pas d'anomalies chromosomiques à haut risque.
Risque élevé	Présence de del(17p), et/ou translocation t(4 ;14), et/ou translocation t(14,16).

Table 9 – Risque selon les anomalies chromosomiques par FISH

Risque	Critère
Normal	$<$ à la normale définie par le laboratoire.
Élevé	$>$ à la normale définie par le laboratoire.

Table 10 – Risque selon le niveau de LDH

Stade	Critère
stade I	• Stade I du **ISS** • Cytogénétique de risque standard en FISH • LDH normale
stade II	Ne convient ni au **stade I** ni au **stade III**.
stade III	• Stade III du **ISS** • Cytogénétique de haut risque en FISH • LDH élevée

Table 11 – Critères du R-ISS

11 Conclusion

Le myélome multiple est une hémopathie maligne très courante, commençant généralement par des états précurseurs asymptomatiques. Ce cancer est un néoplasme des cellules B de la moelle osseuse qui s'accompagne d'une série complexe de manifestations cliniques, y compris l'anémie, les lésions osseuses, le dysfonctionnement rénal, etc. Dans ce chapitre, nous avons présenté des données et des connaissances très importantes sur le myélome multiple, notamment son épidémiologie, sa physiopathologie, sa pathogenèse, tous les signes cliniques et toutes les formes de MM ainsi que le processus de diagnostic et les tests/examens effectués. L'objectif principal est de comprendre l'aspect médical de cette maladie avant de passer à nos expérimentations et la présentation des approches que nous proposons pour aider au diagnostic et stadification du MM, dans les chapitres suivants.

CHAPITRE 2

ÉTUDE DES FACTEURS DE DIAGNOSTIC/CAUSES DU MYÉLOME MULTIPLE

1 Introduction

Le myélome multiple est une maladie complexe qui est souvent diagnostiquée tardivement. Les patients doivent effectuer une série de tests fréquents pour diagnostiquer et stadifier la maladie, ce qui peut être décourageant. Les facteurs causaux du myélome multiple restent inconnus et les recherches sont en cours pour mieux comprendre cette maladie.

Le but principal de ce chapitre est de rechercher une méthode de prédiction précise pour les tests les plus importants dans le processus de stadification du myélome multiple. Cela permettrait de réduire le nombre de tests nécessaires et de réduire le coût total pour les patients, qui est un problème majeur pour de nombreux patients atteints de cette maladie.

Dans la littérature récente, les travaux proposés pour aider au diagnostic du MM se sont principalement concentrés sur l'utilisation de bases de données génétiques [42], [43], [44], [45]. Cependant, la collecte de données peut parfois inclure des données inutiles qui peuvent ralentir l'apprentissage du modèle et réduire sa précision. Pour résoudre ce problème, la sélection de caractéristiques est utilisée pour ne garder que les données les plus pertinentes et réduire le bruit et les données redondantes.

Dans ce chapitre, nous proposons de tester des méthodes supervisées de sélection de variables pour améliorer les performances de classification. Nous présentons les différentes approches de sélection de variables, leur importance et leurs avantages, ainsi que des travaux récents dans la littérature. Nous présentons également notre base de données collectée pour notre étude, ainsi que l'approche proposée et les méthodologies utilisées

pour mener à bien notre recherche. Enfin, nous discutons des résultats obtenus et de la contribution que notre travail peut apporter à la recherche scientifique.

2 Outils statistiques de sélection des variables

En choisissant les mesures statistiques adéquates en fonction du type de données d'entrée et de sortie (catégorique ou numérique), les méthodes de sélection peuvent être rapides et efficaces. Ces outils statistiques permettent d'évaluer la relation entre chaque variable d'entrée et la variable cible, et de sélectionner les variables d'entrée qui ont la relation la plus forte avec la variable cible.

La Figure 6 présente un aperçu des outils statistiques qui peuvent être utilisés en fonction des différents types de données étudiées. Ainsi, lorsque la variable d'entrée et de sortie sont toutes deux catégoriques, le test du χ^2 ou l'information mutuelle peuvent être utilisés. Les coefficients de corrélation de Pearson ou de Spearman sont utilisés lorsque la variable d'entrée et de sortie sont toutes deux numériques. En revanche, lorsque la variable d'entrée est numérique et la sortie est catégorique, on peut utiliser le test Anova ou le coefficient de rang de Kendall. Ces mêmes mesures statistiques peuvent également être appliquées pour des entrées catégorielles et des sorties numériques.

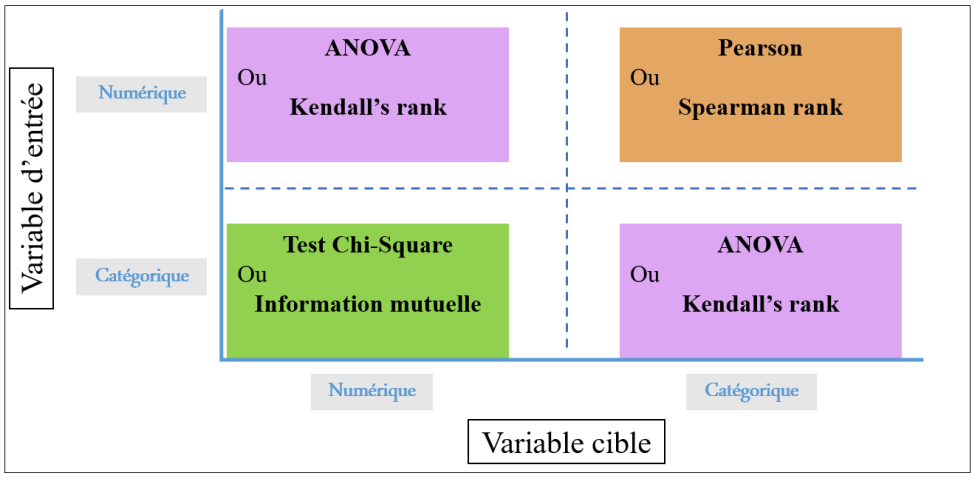


Figure 6 – Outils statistiques de sélection des variables

3 Approches de sélection des variables

Les différentes méthodologies et techniques qui peuvent être utilisées pour sélectionner le meilleur sous-ensemble de caractéristiques et aider les modèles à être plus performants et efficaces, sont regroupées en trois approches principales : l'approche filtre, l'approche enveloppe et l'approche intégrée.

3.1 Approche Filtre

L'approche filtre (Filter) est une étape de pré-traitement dans laquelle les variables sont éliminées en fonction de leur corrélation avec la sortie. Les méthodes "Filter" permettent de sélectionner les variables indépendamment du modèle d'apprentissage (voir figure 7), ce qui les rend assez rapides. C'est l'un des avantages de cette approche [46]. Il existe deux types de méthodes de sélection basées sur l'approche filtre : univariées et multivariées [47]. Dans les méthodes univariées, différents types de critères de classement peuvent être utilisés, par exemple le score de Fisher, l'information mutuelle et la variance de la variable. Dans les méthodes multivariées, la relation mutuelle entre les variables est prise en compte pour éliminer celles qui sont redondantes.

3.2 Approche enveloppe

L'approche enveloppe (Wrapper) mesure l'utilité des variables en optimisant les performances du classifieur (voir figure 7). Elle considère la sélection d'un ensemble de caractéristiques comme un problème de recherche, où différentes combinaisons sont préparées, évaluées et comparées à d'autres combinaisons. Un modèle prédictif est utilisé pour évaluer une combinaison de caractéristiques et attribuer un score basé sur la précision du modèle.

Les méthodes wrapper sont généralement très coûteuses en termes de temps de calcul par rapport aux méthodes de filtrage en raison des étapes d'apprentissage répétées.

3.3 Approche intégrée

L'approche intégrée (Embedded) est assez similaire à l'approche Wrapper, car elle est également utilisée pour optimiser la fonction objectif ou les performances d'un modèle d'apprentissage (voir figure 7). La différence avec l'approche Wrapper est qu'une métrique intrinsèque de construction de modèle est utilisée pendant l'apprentissage.

Les méthodes intégrées combinent les qualités des méthodes Filter et Wrapper pour créer un meilleur sous-ensemble. Elles sont implémentées par des algorithmes qui ont leurs propres méthodes de sélection de variables intégrées.

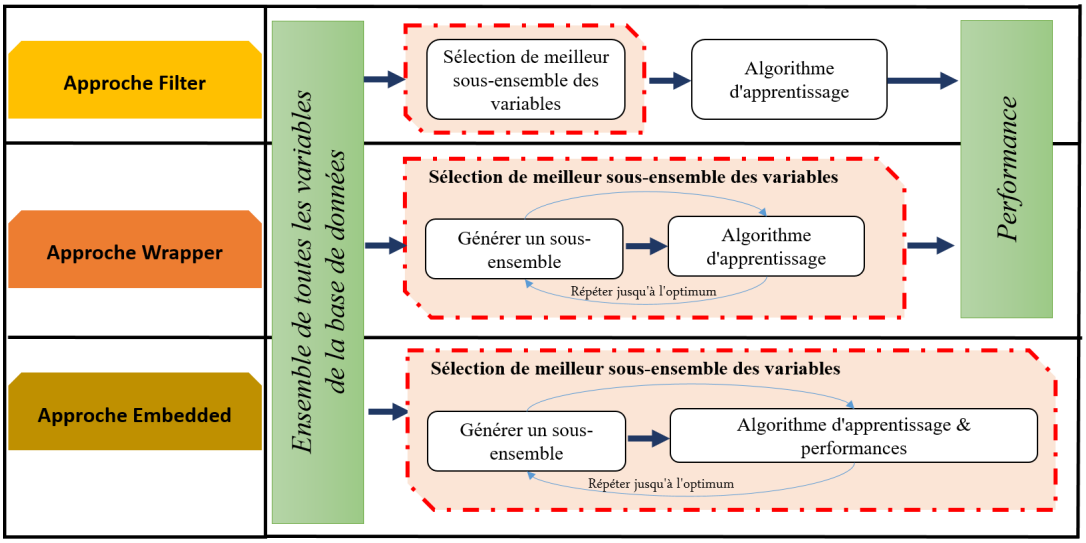


Figure 7 – Approches de sélection des variables

4 État de l'art

La sélection des variables est une étape cruciale dans la construction de modèles d'apprentissage automatique, car elle permet de choisir les caractéristiques les plus pertinentes pour obtenir une meilleure précision de prédiction tout en utilisant moins de données [48]. Dans la littérature, de nombreux articles scientifiques sont disponibles pour donner un aperçu des méthodes de sélection existantes. Dans leur article, Saeys et al. [49] présentent un aperçu des différentes techniques de sélection de caractéristiques et de leurs utilisations en bio-informatique. Leur objectif principal est de sensibiliser les praticiens aux avantages et à la nécessité d'appliquer des techniques de sélection. Ils décrivent également les domaines d'application les plus importants en bio-informatique et mettent en évidence les efforts de la communauté bio-informatique dans le développement de nouvelles procédures adaptées. Enfin, ils orientent le lecteur intéressé vers des packages utiles pour l'exploration de données et la bio-informatique qui peuvent être utilisés pour la sélection de variables. D'autre part, Heinze et al. [50] expliquent les concepts de base et les conséquences importantes des méthodes de sélection de variables qui peuvent encore être flous pour de nombreux praticiens et développeurs de logiciels. Ils suggèrent que certaines quantités soient calculées de manière routinière chaque fois qu'un

utilisateur demande la mise en œuvre d'un algorithme de sélection de variables, et expliquent comment ces quantités peuvent aider l'analyste.

La sélection de variables dans la construction des modèles d'apprentissage automatique est une étape cruciale qui vise à obtenir un modèle prédictif précis en choisissant les caractéristiques les plus pertinentes tout en réduisant la quantité de données nécessaires [48]. Dans la littérature, plusieurs techniques de sélection de caractéristiques sont disponibles pour les chercheurs, notamment les méthodes filtres qui analysent la structure des données pour déterminer le sous-ensemble optimal et les méthodes wrapper qui effectuent une recherche parmi les sous-ensembles possibles. Les algorithmes de pondération sont également utilisés pour classer chaque caractéristique selon son niveau d'importance, mais cette approche n'aide pas à réduire la dimensionnalité [50].

Les méthodes d'ensemble basées sur les arbres de décision ont permis de mettre en pratique l'importance des caractéristiques, qui mesure la fréquence et le degré d'utilisation de chaque caractéristique dans le modèle. Les travaux de Jeremy et al. [51] ont proposé une technique pour déterminer la pertinence des caractéristiques en utilisant le gain d'information moyen obtenu lors de la construction d'ensembles d'arbres de décision. Cette technique prend en compte la complexité des nœuds et utilise une méthode statistique pour mettre à jour la distribution d'échantillonnage des caractéristiques en fonction d'intervalles de confiance pour contrôler le taux de convergence. Les résultats de leurs expériences ont montré que la pondération et la sélection des caractéristiques sont essentielles pour optimiser la généralisation du modèle. Ils ont également comparé leurs résultats à ceux obtenus par l'algorithme de sélection de caractéristiques basé sur la corrélation CFS [52].

De nombreuses applications ont été développées dans le domaine médical pour aider le personnel de santé dans la prise de décision. La classification de données déséquilibrées représente un défi dans bon nombre de ces applications [53]. En effet, c'est un problème très important d'un point de vue algorithmique et de performance [54, 55]. Dans la littérature, les solutions proposées pour traiter le problème d'apprentissage à partir de données déséquilibrées ont été classées en trois catégories [56] : les méthodes opérant au niveau des données [57] (les techniques de suréchantillonnage et de sous-échantillonnage), les méthodes opérant au niveau algorithmique et les méthodes d'ensemble. Cette dernière catégorie est sensible à l'asymétrie, notamment via l'approche de boosting ou de

bagging qui est devenue un axe de recherche majeur [58–60].

Tanha et al. [61] ont examiné les performances de 14 algorithmes de boosting sur 19 ensembles de données multi-classes non équilibrées. Les résultats expérimentaux ont montré que les algorithmes CatBoost et LogitBoost sont meilleurs que les autres algorithmes de boosting sur les jeux de données multi-classes, déséquilibrés et volumineux, respectivement.

D'un autre côté, Chen et al. [62] ont proposé deux méthodes pour utiliser la forêt aléatoire (RF) sur des ensembles de données non équilibrés. La première, "Balanced random forest (BRF)", est basée sur une technique d'échantillonnage. La seconde approche, appelée "Weighted random forest (WRF)", est basée sur un apprentissage sensible aux coûts. Les résultats obtenus ont montré que les deux méthodes, WRF et BRF, ont donné de meilleures performances par rapport à la plupart des techniques existantes qu'ils ont étudiées.

Dans cette thèse de recherche, nous proposons de combiner les méthodes d'apprentissage d'ensemble basées sur des arbres de décision avec des méthodes d'échantillonnage (SMOTE) pour traiter le problème de déséquilibre élevé dans notre jeu de données "MM_dataset" [4]. Cette combinaison est basée sur plusieurs propositions récentes dans la littérature qui ont obtenu des résultats positifs [55, 63].

Les problèmes liés à chaque type de sélection de variables sont très différents et la littérature sur ce sujet est très vaste [64]. De nombreux chercheurs, dans leurs publications scientifiques, tentent d'améliorer et de démontrer les performances prédictives des méthodes de sélection et leurs utilisations dans diverses sciences de la vie [65], [66], [67].

Cependant, une revue de la littérature montre qu'il existe relativement peu de publications présentant des résultats de travaux concernant les méthodes d'aide au diagnostic et de détection automatique du myélome multiple. Par exemple, David et al. [68] présentent dans leur article une étude comparative d'une variété d'algorithmes d'apprentissage supervisés (SVM, réseaux bayésiens, arbres de décision) sur un ensemble de données contenant plus de 100 échantillons de puces à ADN d'expression génétique. Ce travail leur permet de tirer des leçons importantes pour l'exploration de données de puces à ADN, notamment : les réseaux et ensembles de Bayes fonctionnent au moins aussi bien que d'autres approches, mais fournissent sans doute un aperçu plus direct ; la recherche de différences cohérentes dans l'expression peut être plus importante que les grandes différences.

Ces résultats fournissent des preuves et des références pour des travaux futurs qui pourraient être utiles dans d'autres applications d'exploration de données supervisées pour l'étude des maladies basées sur les données de puces à ADN.

En outre, Hwang et al. [69] ont mené une étude rétrospective portant sur 467 patients, basée sur les résultats d'imagerie par résonance magnétique (IRM) de la colonne lombaire. L'objectif de cette étude était de construire un modèle d'apprentissage automatique à l'aide d'un classifieur de texture SVM, capable d'isoler les modèles d'infiltration suspects de maladies hématologiques sur les IRM de la colonne lombaire. La comparaison des résultats obtenus avec ceux de radiologues expérimentés a démontré avec succès la faisabilité des SVM pour différencier la moelle osseuse atteinte de maladies hématologiques de celle qui ne l'est pas.
D'autre part, Chen et al. [70] ont proposé une combinaison de la spectroscopie de dégradation induite par laser à base de sérum (serum-based LIBS) avec des méthodes d'apprentissage automatique pour le diagnostic et la détermination du stade du myélome multiple (MM). Ils ont appliqué et tenté d'optimiser des statistiques multivariées et des méthodes d'apprentissage automatique, y compris les classifieurs PCA, kNN, SVM et ANN, via une validation croisée (10-fold), et les ont évaluées en termes de précision, de sensibilité, de spécificité et de courbes ROC. Les résultats obtenus ont montré que les classifieurs kNN, SVM et ANN atteignaient des performances de discrimination similaires, avec des précisions supérieures à 90% pour le diagnostic et la stadification du MM.

Dans le même domaine, d'autres travaux ont été proposés pour améliorer le diagnostic du myélome multiple en utilisant des méthodes de sélection de caractéristiques. Liu et al. [71] ont proposé une méthode appelée Recursive Feature Addition (RFA) qui combine l'apprentissage supervisé et les mesures de similarité statistique pour sélectionner les gènes pertinents à partir des données d'expression génétique des biopuces MAQ-II pour le cancer du sein et le myélome multiple. Ils ont comparé cette méthode avec d'autres méthodes de sélection de gènes telles que SVM Recursive Feature Elimination (SVMRFE), Leave-One-Out Calculation Sequential Forward Selection (LOOCSFS) et Gradient based Leave-one-out Gene Selection (GLGS) en utilisant plusieurs classifieurs d'apprentissage populaires. Les résultats montrent que l'approche proposée est plus performante que les autres méthodes comparées.

De même, Zhang et al. [72] ont proposé une approche de sélection

conjointe bayésienne appelée méthode Overlap-HSVS pour identifier plusieurs combinaisons de gènes et de voies qui sont significativement associées aux résultats cliniques de la maladie du myélome multiple. Les résultats montrent que la méthode Overlap-HSVS permet d'identifier la plupart des groupes ainsi que les variables individuelles au sein d'un groupe par rapport à la méthode Lasso, qui est l'une des méthodes de sélection de variables les plus populaires. En outre, certains des gènes et des voies sélectionnés ont été identifiés dans les recherches biologiques comme des biomarqueurs importants du MM.

5 Présentation de la base de données collectée

Nous avons effectué un stage pratique au sein du Centre de Lutte Contre le Cancer (CLCC) en collaboration avec la faculté de Technologie de l'université de Tlemcen et le Centre Hospitalo-Universitaire de Tlemcen (CHU-Tlm) pendant une période de 8 mois. L'objectif de ce stage était de collecter des données sur le cancer du myélome multiple, afin de les utiliser en pratique clinique et d'analyser leur impact sur le suivi des patients.

Nous avons pu collecter une base de données de **203 patients** diagnostiqués entre 2008 et 2019, qui contient **57 paramètres**. Cette base de données couvre toutes les informations démographiques des patients, leurs antécédents personnels et familiaux, ainsi que les résultats de divers examens médicaux et tests de diagnostic du myélome multiple. Nous avons rendu cette base de données publique le 24/12/2019 [4]. Cette base de données peut être utilisée pour résoudre les problèmes auxquels les médecins cliniciens sont confrontés dans le diagnostic du myélome multiple, tels que la détection des facteurs pronostiques du MM en utilisant des méthodes de sélection de variables.

Cependant, cette base de données contient des valeurs manquantes en raison de la situation actuelle du service d'Hématologie du CHU-Tlm. En effet, les archives médicales sont uniquement sous forme papier, les dossiers médicaux sont mal organisés et plusieurs fiches cliniques sont rédigées par différents médecins pour un même patient. De plus, certains patients refusent ou ignorent certains tests pour des raisons financières ou parce que le processus de diagnostic leur semble trop long. Tous ces facteurs peuvent entraîner une abondance de documents dans le service et ralentir le diagnostic, empêchant ainsi les chercheurs de mener des études et des recherches scientifiques afin de trouver des réponses sur les causes et les facteurs affectant cette maladie en Algérie et en particulier dans la région de l'Ouest algérien.

La classe de sortie de notre base de données contient les stades du cancer du MM, étiquetés par les spécialistes en hématologie du CHU-Tlm en utilisant la stadification de Durie-Selmon [39] et le système international de stadification (ISS) [40]. Cependant, la distribution des classes dans notre jeu de données est très déséquilibrée (voir figure 8). En effet, comme mentionné précédemment, pour ce cancer, aucun symptôme n'est détecté à un stade précoce et que plusieurs patients, lorsqu'ils se rendent à l'hôpital pour la première fois, sont directement diagnostiqués au stade III. La description des attributs du jeu de données MM est présentée dans le tableau 12.

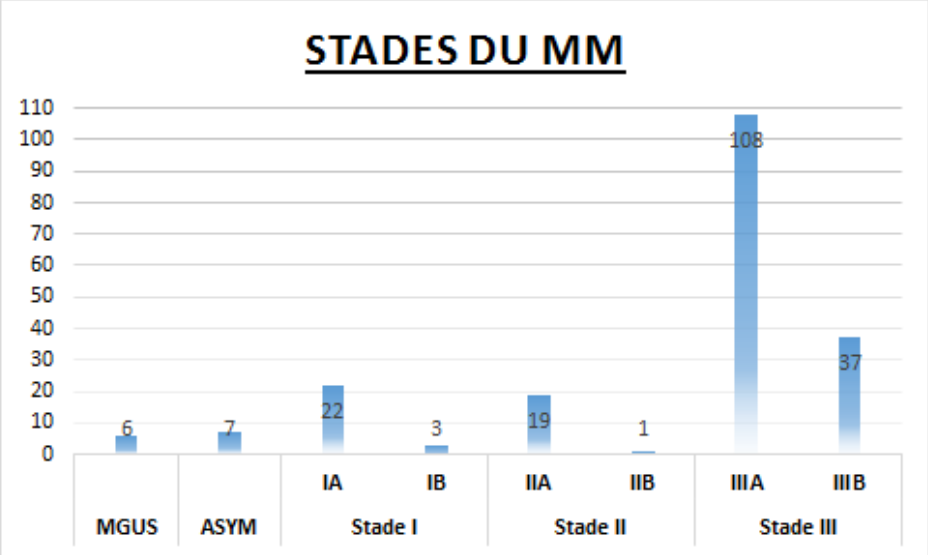


Figure 8 – Distribution des stades du MM

Les distributions d'âge pour les deux groupes Homme/Femme sont présentées dans la figure 9. L'âge moyen pour les deux sexes est d'environ 65 ans, avec une légère différence entre l'âge maximum et minimum. En effet, l'âge maximal est de 98 ans pour les femmes et de 89 ans pour les hommes, tandis que l'âge minimal est respectivement de 38 ans et 39 ans pour les hommes et les femmes.

Par ailleurs, il convient de noter que la majorité des patients résident dans la wilaya de Tlemcen, comme indiqué dans la figure 10.

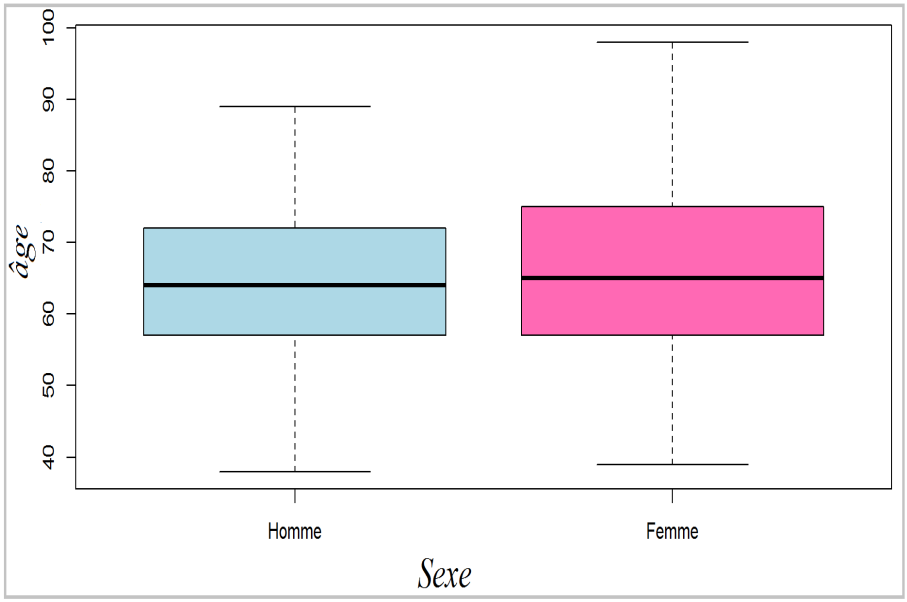


Figure 9 – Répartition d'âge par sexe

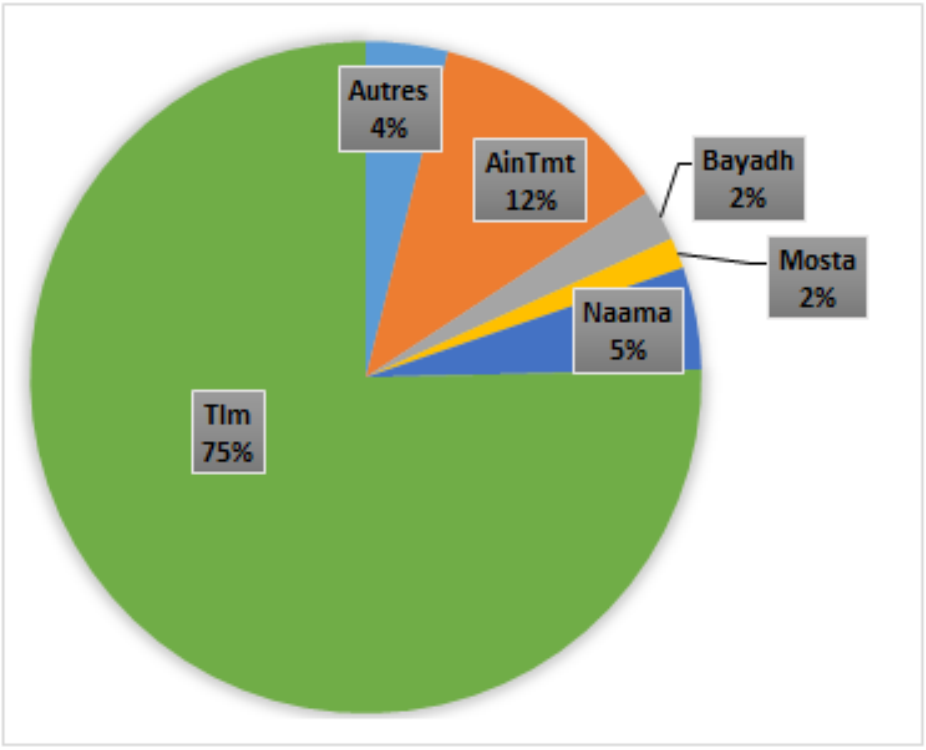


Figure 10 – Répartition par wilaya de résidence

Information	Code	Attribute
Données démographiques	gender	Sexe
	age	Age
	city	Wilaya de résidence
	married	Marié(e)
	nbr_child	nombre d'enfants
Examens cliniques	weight	Poids
	body_surf	Surface corporelle
	blood	Groupe sanguin
	asth&bone	Asthénie et douleurs osseuses
	anemia	Signes cliniques d'anémie
Antécédents personnels et familiaux	HBP	Hypertension artérielle
	diabete	Diabète
	tobbacco	Tabac
	chron_disea	Maladies familiales chroniques
	hrd_blo_disea	Maladies sanguines héréditaires
Tests d'hématologie	CBC_WBC	Taux de globules blancs
	CBC_RBC	Taux de globules rouges
	CBC_plats	Taux de plaquettes
	CBC_Hgb	Taux d'hémoglobine
	CBC_Hct	Hématocrite
	CBC_MCV	Volume Globulaire Moyen
	CBC_MCHC	Concentration corpusculaire moyenne en hémoglobine
Examens de cytologie	roll_RBC	Présence ou absence des rouleaux érythrocytaires
	plasma_cells	Taux de plasmocytes
Proteins Tests	B2M	Test de β-2 Microglobuline
	prot_rate	Taux de protéines
	alb	Albumine
	α_glob	α_globuline
	β_glob	β_globuline
	γ_glob	γ_globuline
	BJp	Protéines de Bence Jonce
	24h_prot	Protéinurie de 24 heures
	Ig	Type d'immunoglobuline anormale
	chain	Type de chaîne légère libre

Suite sur la page suivante

Medical imaging	ost_les	Lésions ostéolytiques
Bilans biologique et chimique (urinaire et sanguine)	VS	Vitesse de sédimentation
	Ca	Taux de calcium
	K	Taux de potassium
	Na	Taux de sodium
	P	Taux de phosphore
	CRP	Protéine C-réactive
	creat	Taux de créatinine
	urea	Urée
	clair_creat	Taux de clairance de la créatinine
	SGOT	Glutamate-oxaloacetate-transaminase
	SGPT	Glutamate-pyruvate-transaminase
	GGT	Gamma-glutamyl transférase
	PAL	Phosphatase alcaline
	Ac_Anti_HCV	Test d'anticorps anti-hépatite C
	HIV	Test du VIH
	Ag_HBS	Test d'antigène de surface de l'hépatite B
	gly	Test de glycémie
	TCA	Temps de céphaline activée
	TP	Taux de prothrombine
	Fib	Taux de fibrinogène
	Ferr	Taux de ferritine
	LDH	Lactate déshydrogénase
	cardio_EF	Fraction d'éjection ventriculaire gauche

Table 12: Description des variables de la base de donnée

6 Méthodologie proposée

La sélection de variables n'a pas de méthode universelle ou de meilleur choix. Au lieu de cela, une approche systématique et expérimentale est nécessaire pour déterminer ce qui fonctionne le mieux pour notre problème spécifique en utilisant différentes mesures statistiques et en essayant différents modèles.

Dans cette section, nous présentons notre méthodologie pour étudier la relation entre les stades du MM et les examens médicaux effectués lors du diagnostic de MM. Nous avons utilisé une base de données très déséquilibrée pour identifier les examens les plus pertinents pour la tâche de stadification de ce type de cancer. Notre méthodologie est illustrée dans la figure 11. Nous avons exploré deux approches principales : les méthodes d'ensemble basées sur des arbres de décision et les méthodes de sélection de variables supervisées basées sur l'approche Filter.

Les méthodes d'ensemble basées sur des arbres de décision sont couramment utilisées en apprentissage automatique pour leur capacité de généralisation et leur interprétabilité. Nous avons utilisé deux types de méthodes d'ensemble : celles basées sur la randomisation et celles basées sur l'optimisation.

D'autre part, nous avons également utilisé des méthodes de sélection de variables supervisées basées sur l'approche Filter. Ces méthodes peuvent être divisées en deux groupes : celles qui évaluent la qualité (pertinence) de chaque variable individuellement sans considérer l'interaction avec les autres variables, et celles qui sélectionnent un sous-ensemble de variables pour la classification en prenant en compte l'interaction entre les variables dans chaque sous-ensemble candidat évalué.
Dans ce chapitre, nous avons apporté plusieurs contributions importantes, qui peuvent être synthétisées en trois points clés.

- ☞ Premièrement, nous avons effectué un prétraitement de données déséquilibrées en utilisant l'algorithme SMOTE, qui est une méthode populaire pour la génération de données synthétiques. Contrairement à d'autres méthodes, SMOTE ne crée pas de doublons de données, mais plutôt des points de données synthétiques légèrement différents des données originales.
- ☞ Deuxièmement, nous avons abordé le problème des valeurs manquantes dans notre base de données en utilisant des méthodes de traitement appropriées.

☞ Troisièmement, nous avons analysé et comparé les variables sélectionnées comme pertinentes par les différentes méthodes utilisées, et nous avons discuté ces résultats avec des hématologues pour valider nos conclusions.

☞ Enfin, nous avons étudié le diagnostic et la stadification du MM en tant que problème d'apprentissage supervisé multi-classes. Nous avons évalué les performances des algorithmes d'apprentissage individuels après la tâche de sélection de variables basée sur l'approche Filter, ainsi que les performances des méthodes d'ensemble basées sur les arbres de décision, qui ont été utilisées en raison de leur capacité à calculer l'importance des variables.

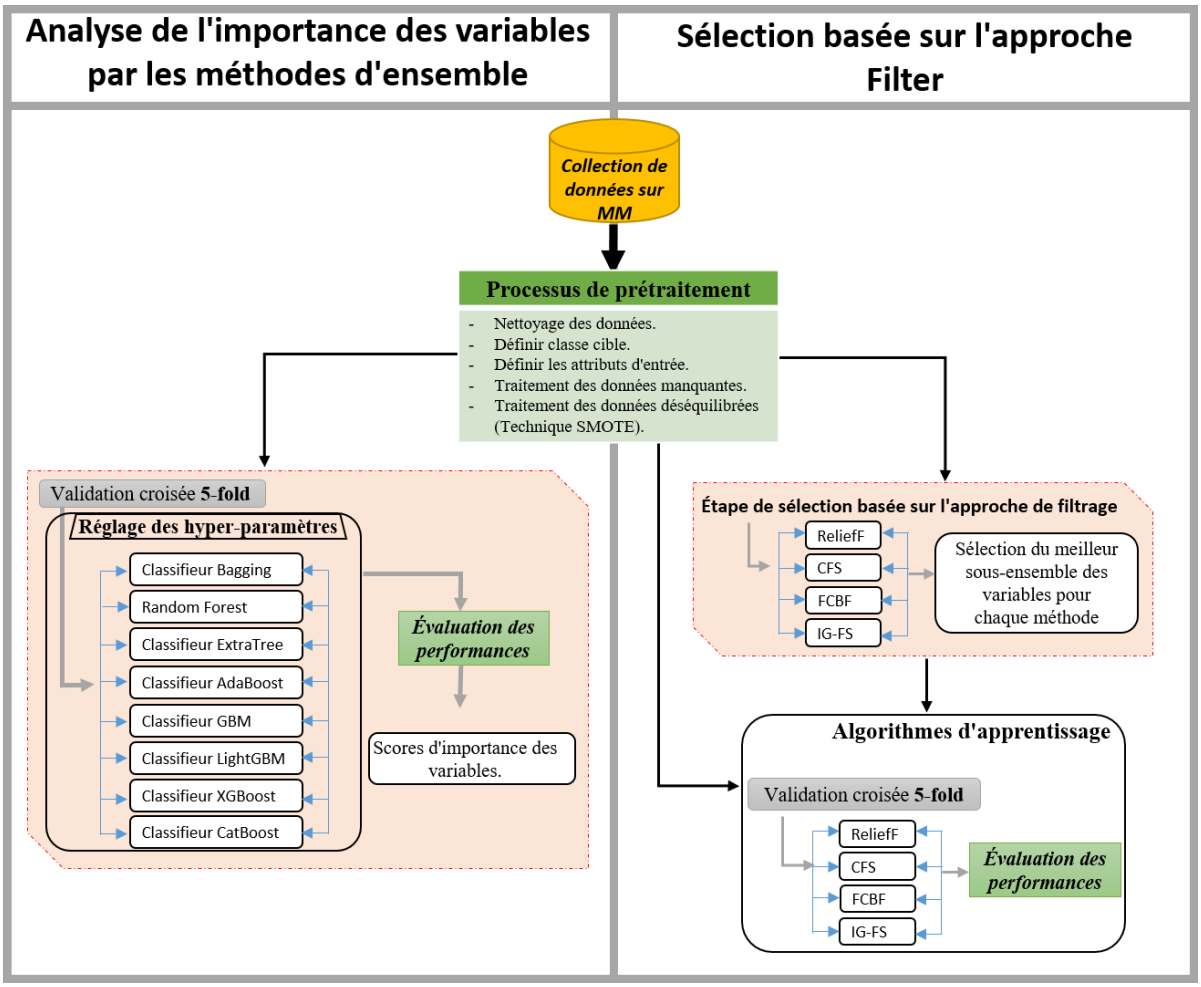


Figure 11 – Diagramme du l'approche proposée

7 Étape de pré-traitement

La stadification du myélome multiple est essentielle pour aider le médecin à prédire l'état du patient et à choisir le type de traitement. Cependant,

comme nous l'avons mentionné précédemment, ce cancer est difficile à détecter à un stade précoce, ce qui entraîne une distribution très déséquilibrée des exemples dans notre base de données, avec une fréquence plus élevée du stade III par rapport aux autres stades (voir figure 6).

Afin de résoudre ce problème de déséquilibre des données, nous utilisons l'algorithme SMOTE [73]. Mais avant cela, nous avons traité les valeurs manquantes présentes dans notre base de données collectée en raison de dossiers médicaux incomplets.

7.1 Traitement de données manquantes

La présence de données manquantes est fréquente dans les ensembles de données du monde réel. Cela peut être dû à plusieurs raisons, comme des données non enregistrées ou la corruption des données. Pour traiter ce problème dans notre ensemble de données, nous avons utilisé deux stratégies simples :

⇒ Remplacer les valeurs manquantes pour les variables catégorielles par la valeur de l'attribut le plus fréquent.

⇒ Pour les variables numériques, nous les remplaçons par la valeur moyenne de l'attribut.

7.2 Traitement de données déséquilibrées

La classification de données déséquilibrées est un nouveau défi dans de nombreuses applications, notamment dans le domaine médical. Les données déséquilibrées se produisent lorsqu'une ou plusieurs classes ont des proportions très faibles dans les données d'apprentissage par rapport aux autres classes [74]. En machine learning, les méthodes classiques ne sont pas toujours adaptées aux données déséquilibrées et peuvent conduire à des résultats trompeurs et optimistes, car les points de la classe minoritaire sont souvent considérés comme des "outliers" ne contenant pas d'informations. Nous avons utilisé l'algorithme SMOTE pour résoudre ce problème en générant des données synthétiques pour les classes minoritaires plutôt que de dupliquer les points de données existants.

Il existe de nombreuses approches pour pallier le problème de classification déséquilibrée, qui sont généralement regroupées en deux catégories principales :

1. **Méthodes au niveau d'algorithme** (Algorithm-level methods) : ces méthodes consistent à adapter les modèles de machine learning classiques pour qu'ils soient en mesure de mieux gérer le déséquilibre.

L'apprentissage sensible aux coûts est une technique qui appartient à cette famille, elle consiste à affecter un poids important à la classe minoritaire. En pratique, cela signifie que nous spécifions à notre modèle que le fait de bien classer un point de la classe minoritaire est plus important que de bien classer un point de la classe majoritaire. De cette façon, nous parvenons à considérer le mauvais classement d'un point de la classe minoritaire par le modèle comme plus grave que le mauvais classement d'un point de la classe majoritaire.

2. **Méthodes au niveau des données** (Data-level methods) : l'idée derrière ce type de méthodes est de transformer les données pour atténuer le déséquilibre en utilisant des techniques d'échantillonnage pour équilibrer le rapport de classe [57], telles que la suppression de représentants de la classe majoritaire (under-sampling) ou l'ajout de représentants à la classe minoritaire (over-sampling).

Pour résoudre notre problème de données déséquilibrées, nous avons étudié une méthode de sur-échantillonnage qui consiste à compléter le jeu de données d'origine par des observations synthétiques des classes minoritaires. Nous nous sommes particulièrement intéressés à la technique SMOTE (Synthetic Minority Over-Sampling TEchnique) [73], qui a été prouvée efficace par plusieurs études récentes dans la littérature [75, 76].

La méthode SMOTE s'inspire d'une technique bien établie dans la reconnaissance de caractères manuscrits [77]. Elle consiste à générer des exemples synthétiques dans «l'espace de la variable cible» plutôt que dans «l'espace des données». Pour sur-échantillonner la classe minoritaire, on prend chaque échantillon de la classe minoritaire et on introduit des exemples synthétiques le long des segments de ligne reliant certains (ou tous) des k plus proches voisins de la classe minoritaire. Les voisins parmi les k plus proches sont choisis au hasard en fonction de la quantité de sur-échantillonnage requise (voir Algorithme 1).

Algorithm 1 SMOTE Algorithm

INPUT :

P number of minority class sample ; S amount of synthetic to be generated ; k number of nearest neighbors

OUTPUT :

N_s=(S/100)*P synthetic samples

BEGIN

1. Create function ComputKNN ($i \leftarrow 1toP$, P_i, P_j)

{**For** $i \leftarrow 1toP$

- Compute k nearest neighbors of each minority instance P_i and other minority instances P_j.
- Save the indices in the $nnarray$.
- Populate (N_s, i, $nnarray$) to generate new instance.

End for}

2. N_s=**(**S**/100)***P

While $N_s \neq 0$

3. Create function GenerateS (P_i, P_j)

{Choose a random number between 1 and k, call it nn.

For $attr \leftarrow 1$ to $numattrs$

$dif = P_i[nnarray[nn]][attr] - P_j[i][attr]$

$gap =$ random number between 0 and 1

Synthetic $[newindex][attr] = P_i[i][attr] + gap * dif$

End for

$newindex = newindex + 1$}

$N_s = N_s - 1$

End while

4. Return (*End of Populate.*)

END

De nombreuses modifications et extensions ont été apportées à la méthode SMOTE depuis son développement. Parmi celles-ci, on peut citer SVMSMOTE [78], Kmeans-SMOTE [79], BorderlineSMOTE [80], ADASYN [81], qui ont permis de gérer des variables nominales (catégorielles) ainsi que des points à la frontière.

8 Étape de sélection basée sur l'approche de filtrage

La sélection des variables par l'approche Filter est largement utilisée à ce jour pour l'analyse des données biologiques. Cette méthode permet de filtrer les variables sur la base de certaines métriques, telles que le calcul de la corrélation. Elle évalue la pertinence des caractéristiques en dehors des modèles prédictifs et ne modélise ensuite que les caractéristiques qui répondent à un certain critère. Elle est plus rapide et constitue généralement la meilleure approche lorsque le nombre de variables est élevé. Contrairement aux méthodes wrapper et embedded qui demandent

un temps de calcul long par rapport à la méthode Filter, bien qu'elles soient aussi caractérisées par la pertinence des attributs sélectionnés.

8.1 Méthode ReliefF

Relief est un algorithme développé par Kira et Rendell [82], inspiré des algorithmes d'apprentissage basés sur des instances [83]. Il est construit sur la base d'une approche Filter, qui est particulièrement sensible aux interactions entre les caractéristiques. À l'origine, il a été conçu pour être appliqué à des problèmes de classification binaire avec des caractéristiques discrètes ou numériques. Cet algorithme s'appuie notamment sur la mesure des similarités et des dissimilarités entre les valeurs d'entrée et la variable cible, et permet ainsi d'estimer la pertinence des différentes caractéristiques à l'aide d'un score global.

La méthode ReliefF [84] a été étendue à partir de la famille Relief pour traiter des problèmes multi-classes. Elle sélectionne une instance aléatoire puis recherche les k plus proches voisins de la même classe que l'instance sélectionnée (Near Hit) ainsi que les k plus proches voisins de chacune des autres classes (Near Misses) [85] (voir Algorithme 2).

Algorithm 2 ReliefF Algorithm :

INPUT :

S dataset of N features and m instances ; α predefined adjustable relevance threshold ; C Output class

BEGIN

Initialize all weights $w = 0$

For i = 1 : m

- Randomly select an instance x_i
- Find k nearest neighbors of x_i having the same class of x_i (hits)
- Find k nearest neighbors of x_i having a different class of x_i (misses).

For A = 1 : N

$$w(A) = w(A) - \sum_{j=1}^{k} \frac{diff(A, x_i, hits_j)}{m*k} + \sum_{c \neq class(x_i)} \frac{P(c)}{1-P(class(x_i))} \sum_{j=1}^{k} \frac{diff(A, x_i, misses_j)}{m*k}$$

Select w greater than α

End For **End For** **END**

OUTPUT :

S_{best} : Optimal subset of attributs that have w greater than α

8.2 Méthode de sélection basée sur la corrélation

Les caractéristiques sont pertinentes si leurs valeurs varient systématiquement en fonction de l'appartenance à une catégorie. En d'autres termes, une caractéristique est utile si elle est corrélée ou prédictive de la classe ;

sinon elle n'a aucun intérêt. En probabilité et statistique, l'analyse de la corrélation entre deux ou plusieurs attributs permet d'étudier le degré d'association pouvant exister entre ces attributs. Cette corrélation peut être positive ou négative, linéaire ou non linéaire, monotone ou non monotone [86]. CFS (Correlation-based feature selection) [52] est un algorithme de sélection qui permet d'évaluer la valeur d'un sous-ensemble d'attributs en examinant la capacité prédictive individuelle de chaque attribut ainsi que le degré de redondance entre eux (voir Algorithme 3).

Algorithm 3 CFS Algorithm :

INPUT : $S(F_1, .., F_N)$: a dataset of N features.
α : predefined threshold value
Y : Output class
OUTPUT : S_{best} : Optimal subset.
BEGIN
For i = 1 : N
Calculate $r(F_i, Y)$: the correlation between each attribute F_i and the class Y.
IF $r(F_i, Y) > \alpha$
F_i added to S_best
End if
End for
END

Cet algorithme est généralement combiné à des stratégies de recherche telles que la sélection en avant, l'élimination en arrière, la recherche bidirectionnelle, la recherche du meilleur en premier et la recherche génétique, entre autres.

Le coefficient de corrélation est utilisé pour estimer la corrélation entre le sous-ensemble d'attributs et la classe [87]. Il nous indique non seulement si deux variables évoluent dans la même direction ou dans la direction opposée comme la covariance, mais il indique également la force de la relation. Sa valeur varie de -1 à 1, où 1 indique la plus forte corrélation possible, considérée comme une proportionnalité directe parfaite. D'autre part, -1 représente la plus forte corrélation inverse possible, ou on dit proportionnalité inverse parfaite. Si le coefficient est égal à 0, cela indique qu'il n'y a aucune corrélation.

La stratégie CFS proposée dans notre travail de recherche est basée sur le coefficient de Pearson [88], qui est le plus populaire. Il est calculé en divisant la covariance entre les deux variables par le produit de leurs écarts-types (voir l'équation 2.1).

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y} \tag{2.1}$$

avec : **Cov**(**X**, **Y**) désigne la covariance entre les variables X et Y.

$\sigma_{\mathbf{X}}$ et $\sigma_{\mathbf{Y}}$ sont les écarts-types des variables X et Y respectivement.

8.3 Méthode de selection basée sur la corrélation rapide

FCBF (Fast Correlation-based Filter method) [89] est une très simple méthode de sélection multivariées basée sur la théorie de l'information, où la pertinence de la variable cible et la dépendance entre chaque paire d'attributs sont prises en considération [90]. Cet algorithme utilise un nouveau concept de corrélation prédominante basé principalement sur l'incertitude symétrique pour calculer les dépendances des caractéristiques, filtrer les caractéristiques non pertinentes (ou redondantes) et trouver le meilleur sous-ensemble, en utilisant la technique de sélection en arrière avec une stratégie de recherche séquentielle, dont le but est d'améliorer la qualité de la classification (voir Algorithme 3).

Algorithm 4 Fast Correlation-based Filter Algorithm :

INPUT :
S (F_1 , F_2 ,. . .,F_N) : a dataset of N features
α : predefined threshold value
C : Output label
OUTPUT :
S_{best} : Optimal subset of the selected attributes
BEGIN
For i = 1 : N
Calculate SU(F_i,C) : Symmetrical Uncertainty
for F_i
IF $SU(F_i, C) >= \alpha$
F_i added to S'_{list}
end
end
Order S'_{list} in descending SU(F_i,C) values
F_x= First Element in S'_{list}
F_y= Next Element of F_x in S'_{list}
do begin
F'_y=F_y
IF $SU(F_x, F_y) >= SU(F_y, C)$
remove F_y from S'_{list}
F_y= Next Element of F'_y in S'_{list}
else F_y= Next Element of F_y in S'_{list}
end until (F_y==NULL)
F_x= Next Element of F_x in S'_{list}
end until (F_x==NULL)
S_{best}=S'_{list}
END

L'incertitude symétrique est une mesure normalisée de l'information mutuelle, qui est utilisée pour évaluer la dépendance entre deux variables aléatoires en se basant sur leurs entropies et leur entropie conditionnelle [91]. Elle permet de mesurer la pertinence d'une caractéristique par rapport à la variable cible. Une caractéristique présentant une valeur élevée d'incertitude mutuelle est considérée comme importante.

Soient X et Y deux variables aléatoires, où P(x) et P(y) représentent leurs probabilités respectives et P(x, y) est leur probabilité conjointe. L'incertitude symétrique est définie comme suit :

$$SU(X, Y) = 2[\frac{H(X) - H(X|Y)}{H(X) + H(Y)}] \tag{2.2}$$

Où :

$H(X)$ et $H(Y)$ est l'entropie de X et Y respectivement (équation 2.3).

$H(X|Y)$ l'entropie conditionnelle moyenne de X sur Y (équation 2.4).

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i)) \tag{2.3}$$

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) log_2(P(x_i y_j)) \tag{2.4}$$

8.4 Méthode IG-FS

La méthode IG-FS (Information Gain-Feature Selection) [92] est une méthode d'évaluation des caractéristiques basée sur l'entropie. Elle permet de calculer l'impact d'une modification de l'ensemble de données sur sa pureté. Une entropie plus petite suggère une plus grande pureté ou une moindre incertitude.

Le gain d'information est utilisé pour la sélection de variables. Il évalue le gain de chaque caractéristique dans le contexte de la variable cible afin de détecter la caractéristique présentant le plus d'informations. Le gain d'information est défini comme suit (voir équation 2.5).

$$Gain(S_j) = E(P_j) - E(S_j) \tag{2.5}$$

avec :

$$E(P) = \sum_{i=1}^{n} P_i log_2 P_i \tag{2.6}$$

$$E(S_j) = \sum_{i=1}^{S_j} I_j * E(Y_j) \tag{2.7}$$

Où :
P_i est le ratio de l'attribut conditionnel.
La valeur de l'information $E(S_j)$ est définie par l'équation 2.7, lorsque S_j possède $|S_j|$ types de valeurs d'attributs, et que P_i divise l'ensemble en utilisant l'attribut S_j.

8.5 Comparaison des méthodes de sélection des variables

Le but de cette partie de la recherche est de déterminer les facteurs pertinents qui influencent le diagnostic et la stadification du MM en utilisant l'approche de sélection de variables. Nous avons mené une étude comparative de quatre algorithmes de sélection basés sur l'approche Filtre : CFS, ReliefF, FCBF et FS-IG décrits précédemment. Nous avons appliqué ces quatre algorithmes de sélection avec les paramètres présentés dans le tableau 13 sur notre base de données réelle (MM_dataset [4]), qui contient

les stades du MM et les résultats des différents examens diagnostiques de ce cancer.

Algorithmes	**Réglage des paramètres**
FCBF	Threshold $\alpha = 0$
ReliefF	K-nearest neighbors $K = 10$
CFS	Threshold $\alpha = 0$
FS-IG	Threshold $\alpha = 0$

Table 13 – Réglage des paramètres pour chaque algorithme utilisé

Une comparaison des résultats obtenus a montré que la méthode FCBF a nécessité moins de temps d'exécution que les autres algorithmes utilisés (voir tableau 14). Contrairement aux autres méthodes, FCBF commence avec l'ensemble complet des variables et utilise le calcul des dépendances de caractéristiques (incertitude symétrique) pour trouver le meilleur sous-ensemble à l'aide de la technique de sélection en arrière avec une stratégie de recherche séquentielle. De plus, FCBF dispose d'un critère d'arrêt interne qui permet à l'algorithme de s'arrêter lorsqu'il n'y a plus de caractéristiques à éliminer. Ce processus de fonctionnement a permis à FCBF d'être plus rapide que les autres méthodes de sélection utilisées dans cette étude.

En revanche, la méthode ReliefF s'est avérée la plus lente (voir tableau 14). Ceci peut être dû au fait que la recherche des plus proches voisins d'une variable peut prendre plus de temps par rapport au calcul d'incertitude symétrique utilisé par FCBF.

Algorithme	FCBF	ReliefF	CFS	FS-IG
Temps d'exécution (en ms)	65.41	529.43	206.4	166.01
# de variables sélectionnées	18	31	36	25

Table 14 – Temps d'exécution et nombre de variables sélectionnées pour chaque algorithme

La sélection de variables est une technique très utile pour réduire le nombre d'entrées dans un classifieur, ce qui conduit à des modèles prédictifs plus performants et moins complexes en termes de calcul. Dans le domaine de la recherche médicale, cette approche est particulièrement intéressante car elle permet de réduire les tests, les coûts et d'accélérer le diagnostic. En examinant la figure 12 et le tableau 14, on constate que FCBF a sélectionné le plus petit nombre d'attributs, soit **18**, par rapport aux autres algorithmes utilisés. Ces résultats confirment que l'algorithme

FCBF atteint un niveau de réduction de la dimensionnalité supérieur. En ce qui concerne les méthodes CFS et ReliefF, le nombre de variables sélectionnées représente respectivement **57%** et **66%** du nombre total de variables de la base de données originale.

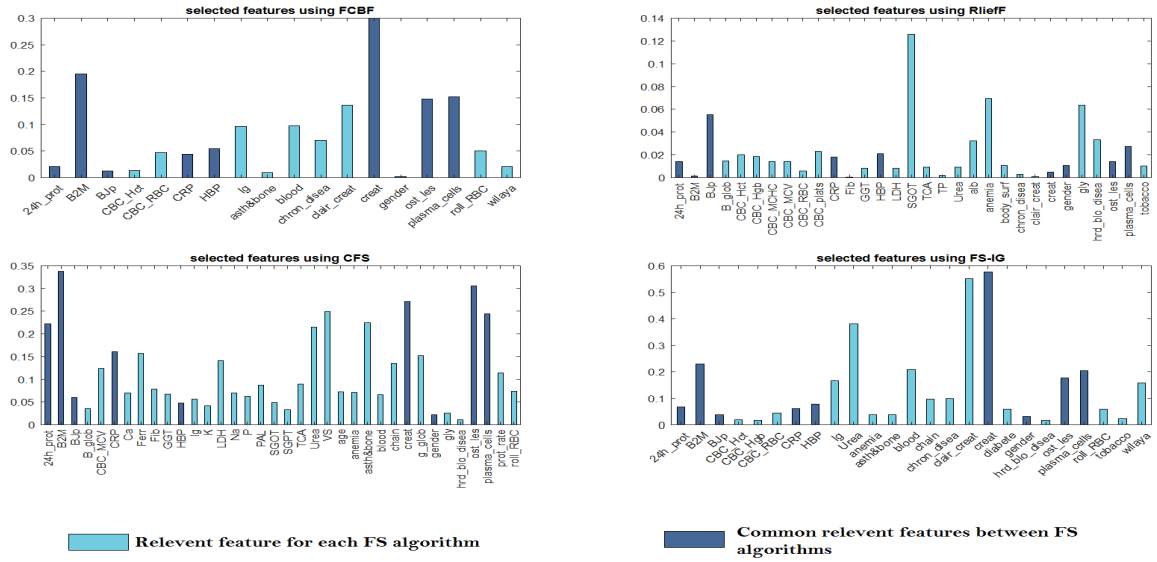


Figure 12 – Variables sélectionnées à l'aide des quatre algorithmes de sélection utilisés.

La figure 12 montre que les méthodes de sélection de variables utilisées dans cette étude produisent des sous-ensembles de variables différents, non seulement en termes de variables sélectionnées mais aussi en termes de l'ordre de leur importance (Ranking). Il est important de noter que lors de la discussion de nos résultats avec des spécialistes en hématologie, ces derniers ont remis en question la pertinence de certaines variables sélectionnées par la méthode ReliefF. En revanche, les variables sélectionnées par la méthode FCBF ont été considérées comme ayant une importance médicale significative dans le diagnostic du MM, de même que les variables sélectionnées par la méthode CFS. Ces variables sont les suivantes : douleurs osseuses et asthénie (asth&bone), tests hématologiques et cytologiques (CBC_RBC, CBC Hct, roll RBC, plasma cells), lésions ostéolytiques (ost_les), tests de protéines (B2M, BJP, 24h_prot, Ig, chain), tests biologiques et chimiques du sang et urinaire (CRP, creat, clair_creat).

9 Analyse de l'importance des variables par les méthodes d'ensemble

Lorsqu'on conçoit des systèmes de diagnostic assisté par ordinateur de haute performance, il est important d'améliorer la précision des algorithmes d'apprentissage automatique. Pour cela, les méthodes d'ensemble [93] sont considérées comme l'une des plus efficaces, car elles combinent plusieurs modèles individuels pour produire un modèle plus précis. Cette méthode repose sur le principe de la "Sagesse de la foule" (Wisdom of the crowd), qui postule que la décision prise par un groupe de personnes est souvent plus précise qu'une décision prise par un individu seul. Ce concept est ancien et a été mentionné par le scientifique Aristote dans son ouvrage "Politique" [94]. Même si seulement un ou deux spécialistes sont concernés par le diagnostic d'un patient, une décision précise et robuste est généralement prise par un ensemble de médecins, qui décident par consensus. C'est une idée ancienne dont le scientifique Aristote a parlé dans son ouvrage "Politique" [94], il disait :

«Il est possible que de nombreux individus, dont aucun homme n'est vertueux, quand ils s'assemblent soient meilleurs que ceux qui sont meilleurs mais peu nombreux, non pas individuellement, mais collectivement, comme les repas collectifs sont meilleurs que ceux qui sont organisés aux frais d'une seule personne . . .C'est aussi pourquoi la multitude est meilleur juge en ce qui concerne les arts et les artistes : en effet, les uns jugent une partie, les autres une autre, et tous jugent le tout.»

ARISTOTE, Politiques

Les méthodes d'ensemble basées sur les arbres de décision sont des techniques de méta-apprentissage qui combinent plusieurs algorithmes de machine learning en un seul modèle prédictif pour produire de meilleures prédictions qu'un modèle unique [95]. Elles sont largement utilisées en machine learning pour leur robustesse et leur capacité à améliorer les performances des modèles prédictifs [96]. Ces méthodes ont également remporté de nombreuses compétitions de machine learning, telles que Netflix Competition, KDD 2009 et Kaggle [97].

En outre, l'utilisation de scores d'importance proposés par les méthodes d'ensemble permet d'identifier les variables les plus pertinentes pour la construction d'un modèle prédictif plus performant. Ces scores peuvent être utilisés pour sélectionner les variables les plus importantes et éliminer celles qui ne sont pas significatives et qui ont des performances simi-

laires ou meilleures dans un temps d'apprentissage beaucoup plus court. La compréhension de la logique sous-jacente du modèle améliore également sa vérification et son amélioration.

Dans la suite, nous examinons diverses approches pour les méthodes d'ensemble basées sur les arbres de décision qui visent à interpréter l'importance des caractéristiques dans la stadification du myélome multiple en utilisant un ensemble de données très déséquilibré. Nous présentons une analyse pratique de ces différentes approches, notamment les méthodes d'ensemble basées sur la randomisation et celles basées sur l'optimisation. Ces approches sont largement utilisées dans les tâches d'apprentissage automatique en raison de leur capacité à généraliser les résultats et de leur interprétabilité.

9.1 Méthodes d'ensemble basées sur la randomisation

La diversité des unités d'apprentissage individuelles est la clé fondamentale de l'apprentissage en ensemble. Les algorithmes basés sur la randomisation sont un choix idéal pour l'apprentissage en ensemble car ils sont naturellement diversifiés [98]. Dans ces types d'ensembles, la diversité des classifieurs de base est générée pendant le processus d'apprentissage par une forme de randomisation, telle que le sous-échantillonnage d'instances, le sous-échantillonnage d'attributs, la randomisation d'hyperparamètres, etc.

9.1.1 Arbres de décision de bagging

Dans le bagging [99], également connu sous le nom d'agrégation bootstrap, un ensemble d'apprentissage est créé en formant une réplique bootstrap de l'ensemble d'apprentissage d'origine. Cela signifie qu'à partir d'un ensemble d'apprentissage S contenant m exemples, un nouvel ensemble d'apprentissage S_0 est construit en sélectionnant aléatoirement m exemples à partir de S. Le bagging est adapté aux algorithmes présentant une forte variance [100], tels que les réseaux de neurones et les arbres de décision pour la classification ou la régression. Un avantage supplémentaire de cette méthode est que la génération des modèles de base dans l'ensemble peut être naturellement parallélisée.

9.1.2 Forêt aléatoire (Random forest)

Les algorithmes d'arbres de décision [101] sont faciles à entraîner, faciles à utiliser et très interprétables, ce qui permet de connaître rapidement les

règles qui permettent la prédiction ou la classification d'un exemple. Cependant, derrière ces qualités apparentes, ces algorithmes peuvent conduire à des prédictions sous-optimales. Pour pallier ce problème, les forêts aléatoires (RF) [102] ont été créées. C'est l'un des algorithmes d'apprentissage supervisé les plus populaires et les plus puissants, utilisé pour effectuer à la fois des tâches de régression et de classification. Comme son nom l'indique, cet algorithme fonctionne comme une grande collection d'arbres de décision non corrélés. Chaque arbre classe un objet en fonction de ses attributs et la forêt choisit la classification ayant le plus de votes. Dans le cas de la régression, on prend la moyenne des sorties des différents arbres. L'un des principaux avantages de RF est qu'elle produit des modèles très précis avec peu ou pas de réglage des hyper-paramètres. En outre, elle utilise des échantillons bootstrap pour entraîner chaque classifieur de base, ce qui la rend facile à utiliser. Un autre avantage de RF est qu'il est facile de mesurer l'importance relative de chaque caractéristique qui a contribué à la prédiction. Deux techniques peuvent être utilisées pour calculer les importances des caractéristiques à partir de RF :

a/- ***Importance du Gini :*** Cette mesure est calculée à partir de la structure de la RF. Elle quantifie la façon dont chaque variable contribue à l'homogénéité des nœuds et des feuilles dans la forêt aléatoire construite.

b/- ***La diminution moyenne de la précision :*** Cette méthode calcule l'importance de la caractéristique sur les échantillons out-of-bag permutés (OOB) en fonction de la diminution moyenne de la précision. Elle est similaire à la méthode d'importance basée sur la permutation.

9.1.3 Classifieur Extra-Trees

Le classifieur Extra-Trees (EXTR) [103] (appelé aussi Extremely Randomized Trees Classifier) est une méthode d'ensemble basée sur la randomisation, dont le concept est similaire à celui de RF et ne diffère que par la manière dont les arbres de décision sont construits. EXTR génère un ensemble d'arbres de décision ou de régression non ajustés selon la procédure classique descendante. La principale différence avec un classifieur arborescent traditionnel est qu'il divise les nœuds en choisissant des points de coupure entièrement au hasard et utilise l'ensemble de l'échantillon d'apprentissage (plutôt qu'une réplique bootstrap) pour développer les arbres. Cet algorithme est plus rapide que RF. L'importance d'une caractéristique est calculée comme la réduction totale du critère fourni par cette caractéristique (l'importance de Gini).

9.2 Méthodes d'ensemble basées sur l'optimization

Le principe de ce type de méthodes d'ensemble est différent de celui des méthodes d'ensemble basées sur la randomisation. L'idée est de former un ensemble de modèles de manière séquentielle, de sorte que chaque nouveau modèle se concentre sur les erreurs de ses prédécesseurs pour les corriger. Deux algorithmes majeurs suivent cette idée : AdaBoost et Gradient Boosting.

9.2.1 Classifieur AdaBoost

La méthode d'ensemble adaptatif "AdaBoost" (Adaptive Boosting) [104] est mise en œuvre pour améliorer la performance de prédiction en convertissant un certain nombre d'apprenants faibles en apprenants forts. Cet algorithme commence par donner des poids égaux à tous les points de données. Ensuite, un modèle faible est entraîné à l'aide de ces poids en utilisant l'ensemble d'apprentissage complet. Les résultats obtenus sont analysés en attribuant des poids plus élevés aux points mal classés (et des poids faibles aux points correctement classés) [105]. L'algorithme le plus couramment utilisé avec AdaBoost est celui des arbres de décision à un niveau, c'est-à-dire avec des arbres de décision à une seule division. Ces arbres sont également appelés "Decision Stumps".

L'importance d'une caractéristique pour AdaBoost est dérivée de l'importance de la caractéristique fournie par son classifieur de base. En supposant que l'on utilise un arbre de décision comme classifieur de base, l'importance de la caractéristique AdaBoost est déterminée par l'importance moyenne de la caractéristique fournie par chaque arbre de décision.

9.2.2 Gradient boosting

Le gradient boosting (GB) [106] construit de manière séquentielle un modèle additif visant à minimiser une fonction de perte donnée ("Loss function"). Le modèle global s'améliore à chaque itération, en ajoutant un nouveau modèle adaptatif qui aide les apprenants faibles. L'idée principale est de surmonter les erreurs de prédiction de l'apprenant précédent. Contrairement à AdaBoost, les poids des instances mal classées ne sont pas directement incrémentés dans GB. Ce type de méthode de boosting comporte trois éléments principaux :

(a)- La fonction de perte qui doit être améliorée ;

(b)- Un arbre de décision pour calculer les prédictions et former un apprenant fort ;

(c)- Un modèle additif qui régularise la fonction de perte.

Tout comme AdaBoost, l'algorithme de gradient boosting peut être utilisé pour les problèmes de classification et de régression. L'un des avantages de GB est que, après avoir construit les arbres de décision améliorés, on peut simplement récupérer les scores d'importance qui indiquent l'utilité de chaque caractéristique dans la construction des arbres de décision améliorés dans le modèle [107]. Le calcul de l'importance permet de classer les caractéristiques, de les comparer entre elles et de sélectionner les plus pertinentes.

9.2.3 Classifieur XGBoost

XGBoost (signifiant "Extreme Gradient Boosting") [97] est une version avancée des méthodes de gradient boosting conçue pour améliorer la vitesse de calcul et l'efficacité du modèle. La raison de l'introduction de ce modèle est que l'algorithme GB calculait la sortie à un rythme très lent car il y avait une analyse séquentielle de l'ensemble de données, ce qui prenait plus de temps. XGBoost est une variante de GB dans laquelle les arbres sont construits séquentiellement. Il implémente ce que l'on appelle des méthodes de calcul distribué pour évaluer tous les modèles volumineux et complexes. Cet algorithme utilise également le calcul Out-of-Core pour analyser des ensembles de données énormes et variés. Il met également en œuvre l'optimisation du cache pour faire le meilleur usage du matériel et des ressources utilisés.

Dans XGBoost, l'importance d'une variable peut être mesurée par plusieurs métriques, telles que le poids, le gain moyen, etc. Le poids correspond au nombre de fois qu'une variable est utilisée pour diviser les données sur tous les arbres boostés. Les caractéristiques les plus importantes sont utilisées plus fréquemment dans la construction des arbres boostés, et les restantes sont utilisées pour améliorer les résidus. Pour le gain, il mesure la réduction réelle des impuretés des nœuds plutôt que de calculer les fractionnements. Il correspond au gain moyen sur tous les fractionnements dans lesquels la caractéristique est utilisée [108].

9.2.4 Classifieur LightGBM

LightGBM [109] est l'acronyme de "Lightweight Gradient Boosting Machines", une méthode de gradient boosting qui utilise l'apprentissage basé sur les arbres. Sa mise en œuvre s'est concentrée sur la création d'un algorithme efficace qui repose sur le pré-calcul de l'histogramme des caractéristiques. Cela accélère l'apprentissage et réduit l'utilisation de la mémoire. La vitesse d'apprentissage rapide est l'un des avantages clés de

l'utilisation de LightGBM. En outre, cette méthode a la capacité d'atteindre un bon équilibre entre la réduction du nombre d'instances de données et le maintien de la précision des arbres de décision appris. Un autre avantage clé de l'utilisation de LightGBM est qu'il prend en charge les caractéristiques catégorielles qui ne sont pas prises en charge par d'autres méthodes de gradient boosting. On peut éviter l'overfitting dans LightGBM en ajustant certains paramètres tels que $n\_estimators$, $learning\_rate$, $feature\_fraction$, num_leaves, etc. et en évitant la croissance d'un arbre très profond.

9.2.5 Classifieur CatBoost

Le nom "CatBoost" vient de deux mots : "Category" et "Boosting". Il s'agit d'une nouvelle génération d'algorithmes de boosting de gradient à haute performance, qui a été récemment développée [110]. Contrairement à la plupart des outils d'apprentissage automatique qui ne fonctionnent qu'avec des données numériques, CatBoost peut travailler avec différents types de données pour aider à résoudre un large éventail de problèmes dans de multiples domaines. Il convertit les valeurs catégorielles en nombres en utilisant diverses statistiques sur des combinaisons de caractéristiques catégorielles et numériques, ce qui permet de préserver la structure de ces données.

9.3 Réglage des hyperparamètres

Avoir un modèle fonctionnel est une bonne chose, mais avoir un modèle optimisé est encore mieux. En machine learning, l'optimisation ou le réglage des hyper-paramètres est une méthode très efficace pour sélectionner les meilleurs paramètres ajustables qui permettent de contrôler le processus d'apprentissage d'un modèle. Trois méthodes sont couramment utilisées pour régler les hyper-paramètres : l'optimisation bayésienne (Bayesian optimization), la recherche par grille (Grid search) et la recherche aléatoire (Random search).

Dans cette étude, nous avons utilisé la technique GridSearchCV pour trouver le modèle avec les meilleurs hyper-paramètres. Cette technique consiste à effectuer une recherche par grille sur l'ensemble d'apprentissage pour affiner les hyper-paramètres. Pour chaque paramètre, nous avons formulé une hypothèse contenant les valeurs optimales pour le modèle. Une fois l'entraînement terminé, les meilleurs paramètres sont retenus pour le modèle ayant obtenu le meilleur score. Ce modèle peut être sauvegardé et testé sur les données de test pour évaluer ses performances.

Le tableau 15 présente les valeurs de recherche par grille sur lesquelles nous avons concentré notre étude pour régler les modèles utilisés sur le jeu de données du myélome multiple.

Pour cette étude, nous avons appliqué une validation croisée à 5-fold plutôt qu'à 10-fold afin de nous assurer que les ensembles d'entraînement et de test contenaient suffisamment d'exemples de chaque élément de la classe cible. Nous avons d'abord effectué une recherche par GridSearchCV du nombre d'arbres pour chaque algorithme en évaluant une série de 600 arbres avec un pas de 25. La figure 13 présente la précision de chaque modèle en fonction du nombre d'arbres. Nous avons constaté que le nombre optimal d'arbres variait selon l'algorithme. Pour le RandomForest, nous avons observé que sa performance atteignait la valeur maximale à partir de 50 arbres. En général, il n'est pas nécessaire de régler le nombre d'arbres pour RandomForest. Il suffit de le fixer à un grand nombre réalisable en termes de calcul et de laisser le comportement asymptotique des nombres faibles ou grands faire le reste [111]. Il n'y a pas de risque de surapprentissage dans RandomForest avec un nombre croissant d'arbres car ils sont formés indépendamment les uns des autres. Pour les classifieurs Bagging et ExtraTree, nous n'avons pas observé une grande différence dans le nombre d'arbres optimal entre 200 et 400, comme le montrent les graphiques de la figure 13. Le nombre optimal d'arbres pour les classifieurs GB, LightGBM, XGBoost et AdaBoost est respectivement de 225, 275, 400 et 325 arbres.

Dans une deuxième stratégie, nous avons examiné de plus près les hyper-paramètres importants pour chaque algorithme d'apprentissage d'ensemble utilisé. Le tableau 15 montre l'ensemble d'hyper-paramètres de la recherche de grille qui a donné la meilleure précision moyenne et qui sera utilisé lors du réglage des modèles sur notre jeu de données. De plus, les ensembles d'hyper-paramètres par défaut pour chaque méthode ont également été utilisés.

Hyper-paramètre	Valeur par défaut	Valeurs de GridSearchCV	Meilleure valeur
Bagging Decision Tree (BDT) :			
max_ features	1	3, 5, 10	3
max_ samples	1	5, 10, 20	20
Random Forest (RF) :			
max_ depth	Unlimited	3, 5, 7, 10, 15	3
max_ features	auto	log2, sqrt, 0.25, 2, 3	0.25
min_ samples_ leaf	1	3, 25, 50, 70	1
min_ samples_ split	2	5, 7, 10, 15	5
ExtraTree (EXTR)			
max_ depth	Unlimited	3, 5, 7, 10, 15	7
max_ features	auto	log2, sqrt, 0.25, 2, 3	0.25
min_ samples_ leaf	1	3, 25, 50, 70	1
min_ samples_ split	2	5, 7, 10, 15	5
Gradient Boosting (GB) :			
learning rate	0.1	0.0001, 0.025, 0.5, 1	0.025
max_ depth	3	5, 7, 10, 15	10
subsample	1	0.15, 0.25, 0.5, 0.75	0.75
LightGBM (lightgbm) :			
learning rate	0.1	0.0001, 0.025, 0.1, 0.5, 1	0.0001
max_ depth	-1	-1, 3,5,7,10,15	-1
num_ leaves	31	4, 6, 10, 31, 100	4
XGBOOST (xgb) :			
learning rate	0.1	0.0001, 0.025, 0.1, 0.5, 1	0.1
max_ depth	3	3, 5, 7, 10, 15	2
gamma	0	0, 0.1, 0.3, 0.5, 1, 2	2
subsample	1	0.15, 0.25, 0.5, 0.75, 1	0.25
AdaBoost (adaboost) :			
learning_ rate	1	0.0001, 0.025, 0.1, 0.5, 1	0.5
CatBoost (catboost) :			
learning_ rate	Var	0.0001, 0.025, 0.1, 0.5, 1	0.05
max_ depth	6	3, 5, 7, 10, 15	3
l2_leaf_reg	3	1, 3, 5, 7, 10	1

Table 15 – Réglage des hyperparamètres pour les méthodes d'ensemble utilisées.

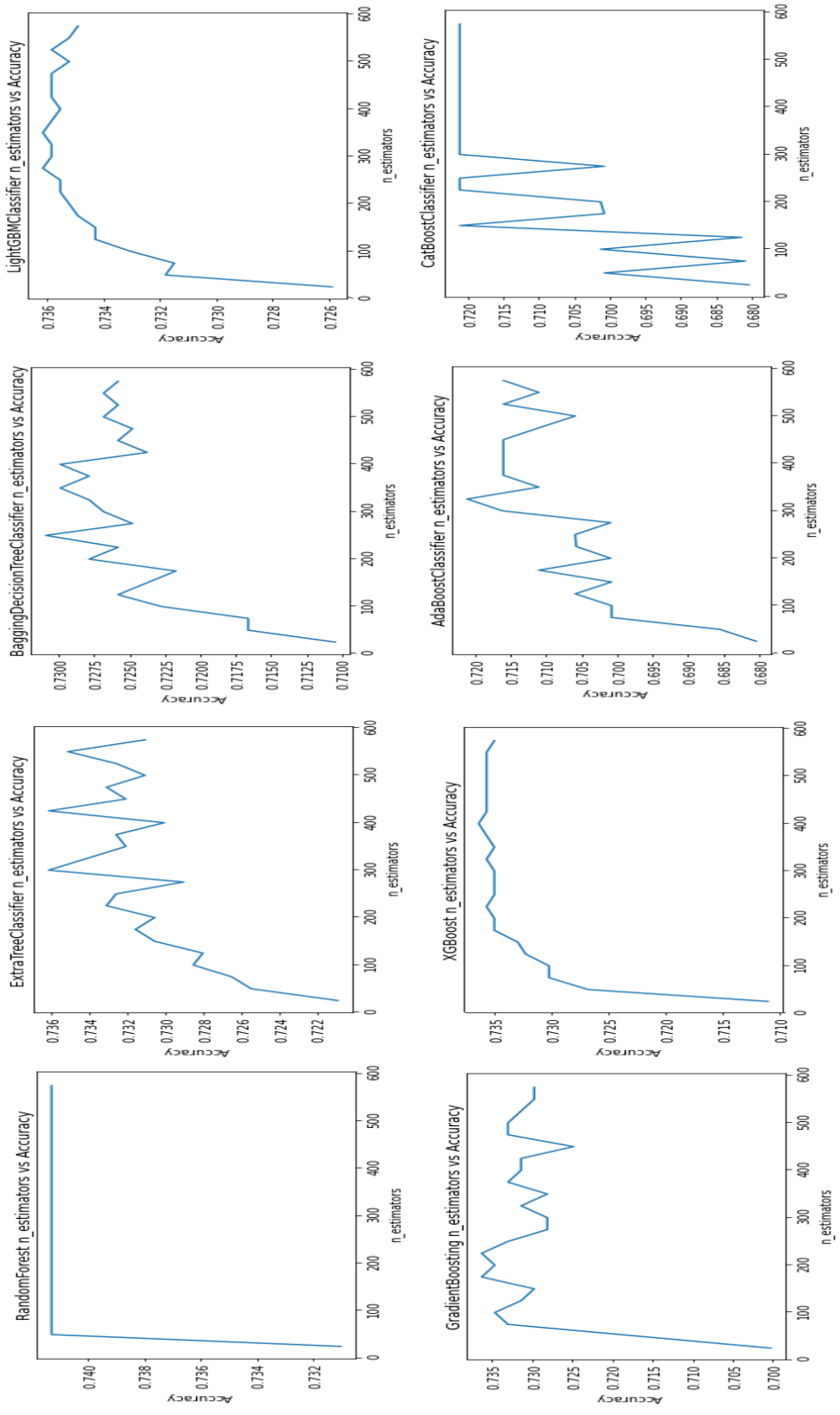


Figure 13 – Scores de précision pour chaque modèle en fonction de n _estimators

9.4 Analyse de l'importance des variables

Une fois notre modèle formé, il est intéressant d'analyser quelles caractéristiques sont les plus prédictives et devraient avoir la plus grande influence sur les valeurs de résultat. À la suite de nos expérimentations, nous avons comparé les scores d'importance des caractéristiques calculés par les différentes méthodes d'apprentissage d'ensemble que nous avons utilisées afin de déterminer celles qui sont basées sur les caractéristiques les plus pertinentes pour le diagnostic et la stadification du MM. Les résultats sont présentés dans la figure 14 pour les méthodes d'ensemble basées sur la randomisation et sur l'optimisation.

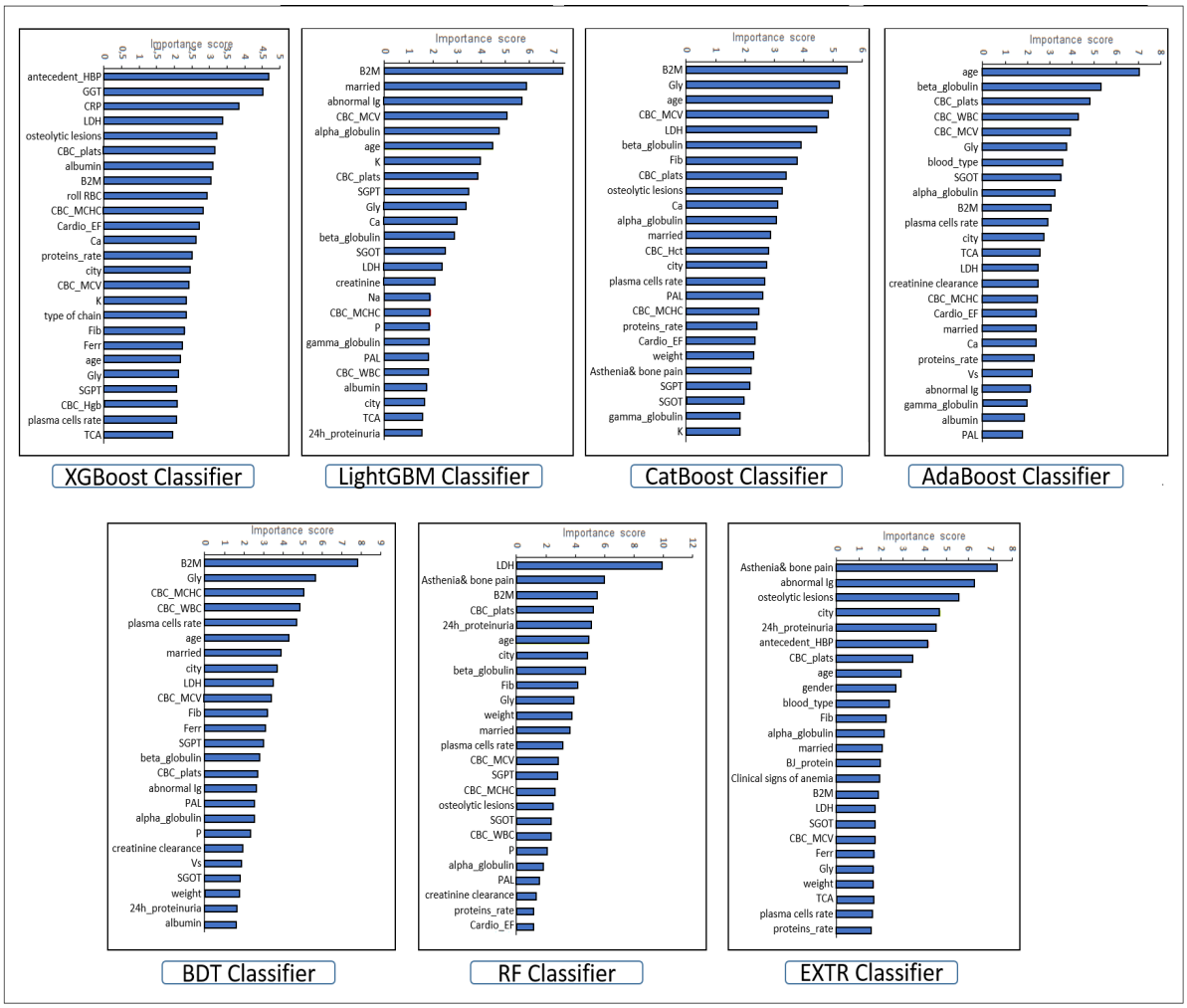


Figure 14 – Scores d'importance des caractéristiques par des méthodes d'ensembles basés sur l'optimisation et sur la randomisation

D'après la figure présentée, il est possible de constater que la plupart des caractéristiques ont une importance relative supérieure à 2% dans tous les modèles d'ensemble pour la prédiction des stades de MM. Les caracté-

ristiques les plus importantes selon tous les modèles sont : la formule sanguine complète (CBC_plats, CBC_MCV), la B2-Microglobuline (B2M), l'albumine, les tests de protéines (alpha_globuline, beta_globuline, gamma _globuline, prot_rate) et le taux de LDH.

Cependant, certains modèles d'ensemble attribuent une importance élevée à certaines variables qui sont importantes dans le diagnostic et la stadification du MM, tandis que d'autres méthodes les ont éliminées. Ces variables incluent : les tests hématologiques (CBC_WBC, CBC_Hgb, CBC_Hct), le taux de plasmocytes (plasma_cells_rate), le taux de calcium (Ca), la clairance de la créatinine (creatinine_clearance), les immunoglobulines anormales (abnormal_Ig) et les lésions ostéolytiques (osteolytic_lesions).

En revanche, les caractéristiques considérées comme les moins importantes sont : les informations démographiques (gender, age), les antécédents personnels et familiaux (antecedent_HBP, antecedent_diabete, hereditary_blood_diseases) et certains tests biologiques (ionogramme sanguin (K, P, Na), test CRP, mécanismes hémostatiques (TP)).

Ces résultats ont été validés par des spécialistes en hématologie. Ces derniers ont souligné que les symptômes détectés dans les analyses sanguines et les scanners sont la première piste de diagnostic pour le MM (CRAB), qui comprend un taux élevé de calcium, des problèmes rénaux, une anémie et des lésions osseuses. Par la suite, le MM est classé comme symptomatique (avec symptômes) ou asymptomatique (sans aucun symptôme). Si un patient est diagnostiqué avec un MM, les médecins tentent de déterminer l'étendue de sa propagation en procédant à une stadification.

En conséquence, les experts en hématologie ont confirmé que les modèles qui classent correctement les caractéristiques considérées comme des facteurs pronostiques pour la stadification du MM sont : le classifieur XGBoost, GB, LightGBM, RF et CatBoost. En revanche, pour les autres méthodes (BDT, ExtraTree et AdaBoost), les experts ont souligné que les caractéristiques choisies comme importantes n'ont aucune signification médicale car elles ne participent pas à la stadification du MM.

10 Évaluation des performances de classification

Dans cette section, nous présentons les résultats obtenus par différents algorithmes d'apprentissage automatique appliqués à notre jeu de

données. Nous avons suivi une méthodologie consistant à appliquer trois classifieurs d'apprentissage automatique supervisés - K-plus proche voisin (KNN) [112], support vector machine (SVM) [113], arbre de décision C4.5 [114] - sur notre jeu de données complet, ainsi que sept algorithmes d'apprentissage d'ensemble basés sur la randomisation (BDT [115], Random forest RF [100], Extra-Tree EXTR [103]) et sur l'optimisation (Adaboost [104], gradient boosting GB [106], Xgboost [97], lightGBM [109], Catboost [110]). Pour cette étude comparative, nous avons mis en place une validation croisée à 5-fold plutôt qu'à 10-fold afin d'assurer une répartition suffisante des instances de toutes les classes dans les sous-ensembles d'apprentissage et de test, compte tenu du déséquilibre des données (voir figure 8).

Dans le même contexte d'évaluation des performances, nous avons appliqué tous les modèles d'ensemble (Boosting et Bagging) choisis avec les meilleurs hyper-paramètres et ceux avec la configuration par défaut sur le jeu de données MM, après une étape de sur-échantillonnage utilisant la technique SMOTE. Les modèles individuels ont également été appliqués à des sous-ensembles de données contenant uniquement les facteurs pertinents pour chaque algorithme de sélection de variables basé sur l'approche Filtre utilisée dans l'étape de sélection précédente.

Tous les résultats obtenus sont résumés dans les tableaux 16 et 17.

Classifieur	Sans la technique SMOTE	Avec la technique SMOTE			
	Jeu de données complet	FCBF	CFS	ReliefF	IG-FS
KNN	56.6%	75.1%	75.9%	75.6%	75.3%
SVM	57.8%	72.0%	72.9%	70.3%	71.7%
C4.5	62.2%	77.4%	79.4%	74.0%	75.1%

Table 16 – Scores de précision pour les algorithmes ML avec et sans étape de sélection par Filtre

En examinant les résultats présentés dans le tableau 16, nous constatons que les performances les plus élevées ont été obtenues en utilisant des sous-ensembles de caractéristiques pertinentes sélectionnées plutôt que l'ensemble complet de caractéristiques.

De plus, nous remarquons que l'algorithme d'arbre de décision C4.5 a obtenu de meilleurs résultats par rapport aux autres classifieurs utilisés. Ce modèle d'apprentissage automatique est réputé pour sa robustesse et sa facilité d'interprétation dans les tâches de classification et de régression, car il peut identifier les variables contribuant à la classification ou à la régression et leur importance relative en fonction de leur position dans

l'arbre. C4.5 utilise le concept d'entropie d'information pour construire un arbre de décision à partir des données d'apprentissage [116]. Il sélectionne l'attribut avec le gain d'information le plus élevé parmi les échantillons actuels comme attribut de test pour diviser l'ensemble d'échantillons.

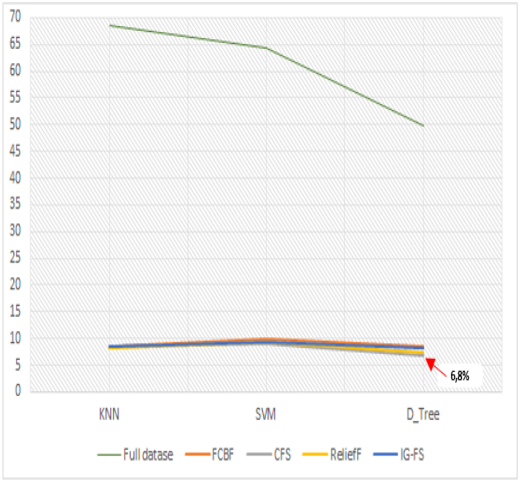


Figure 15 – Taux de faux positifs des classifieurs

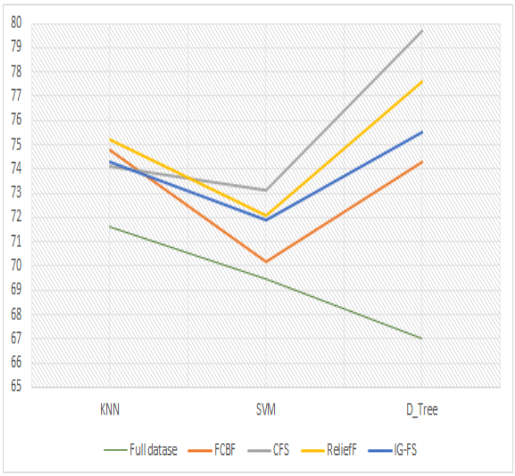


Figure 16 – Taux de vraix positifs des classifieurs

D'après les résultats présentés dans la figure 12, nous avons remarqué que la méthode de sélection des variables "CFS" associée à l'algorithme C4.5 offre la meilleure précision (79,4%) et le taux de faux positifs le plus bas. Quant à la technique FCBF, elle a permis d'atteindre le niveau de réduction de dimensionnalité le plus élevé, avec une différence de précision peu significative par rapport à la méthode CFS (voir tableaux 14 et 16 et figure 12).

Il est important de noter que les méthodes de sélection de variables basées sur l'approche Filtre ne nécessitent pas de modèle d'apprentissage automatique. Elles choisissent des sous-ensembles de caractéristiques en fonction de leur corrélation avec la variable cible. Les mesures statistiques utilisées pour la sélection doivent être sélectionnées avec soin en fonction du type de données et de la variable cible.

En revanche, les méthodes de sélection Wrapper et Embedded créent plusieurs modèles avec différents sous-ensembles de caractéristiques d'entrée et les entraînent pour sélectionner les caractéristiques qui produisent le modèle le plus performant selon une mesure de performance.

En fouille de données, les arbres de décision sont l'une des méthodes de prise de décision les plus efficaces, car ils permettent de présenter et

de hiérarchiser les informations. Le Boosting et le Bagging sont des techniques d'apprentissage en ensemble où plusieurs modèles sont formés pour résoudre un même problème, et combinés pour obtenir de meilleurs résultats.

Dans la suite de cette section, nous effectuons une analyse comparative approfondie de l'efficacité des modèles d'apprentissage en ensemble basés sur les arbres de décision, en utilisant notre jeu de données original et le jeu de données MM équilibré avec la technique d'Oversampling "SMOTE". Les classifieurs que nous avons évalués sont : Bagging Decision Trees (BDT), Random Forest (RF), ExtraTree (EXTR), AdaBoost, Gradient Boosting (GB), XGBoost, LightGBM et CatBoost.

Le tableau 17 montre la précision moyenne et le temps d'exécution pour les différents algorithmes appliqués, avec les meilleurs réglages de paramètres obtenus grâce à la technique de réglage des hyperparamètres "GridSearchCV", ainsi que les réglages par défaut des packages correspondants. Les modèles réglés sont identifiés par la lettre "T" précédant leur nom, tandis que les modèles par défaut sont identifiés par la lettre "D".

En observant les scores de précision des modèles sans SMOTE dans le tableau 17, nous remarquons que l'ordre des performances est similaire à celui des modèles avec SMOTE, mais les performances sont nettement plus faibles. Nous constatons également que le modèle LightGBM est le plus rapide pour les deux scénarios, avec ou sans SMOTE, tandis que GB et CatBoost sont les méthodes les plus lentes, prenant plus de 50 fois plus de temps que LightGBM dans notre cas d'étude. LightGBM est considéré comme étant léger car il est rapide et puissant tout en conservant une précision supérieure. Cette technique est une version améliorée de XGBoost qui est plus ancienne et plus populaire chez les Data Scientists. Elle repose sur deux concepts clés pour traiter respectivement un grand nombre d'instances de données et de caractéristiques : GOSS (Gradient-based One-Side Sampling) et EFB (Exclusive Feature Bundling).

L'utilisation de la technique SMOTE en combinaison avec Random Forest nous a permis d'obtenir les meilleurs résultats de précision, avec une moyenne de 97,41%. Les autres méthodes qui ont donné une bonne précision moyenne supérieure à 90% sont, par ordre décroissant : D.CatBoost, T.GB, T.LightGBM, D.GB, T.EXTR, T.CatBoost, T.XGBoost, D.XGBoost et T.BDT.

Classifieur	Sans la technique SMOTE		Avec la technique SMOTE	
	Précision (%)	Temps d'éxécution (s)	Précision (%)	Temps d'éxécution (s)
D.BDT	68.00	27.00	87.07	14.11
T.BDT	74.00	18.35	92.00	5.88
D.RF	73.62	3.23	97.07	1.95
T.RF	74.62	8.60	97.41	1.80
D.EXTR	73.60	0.81	90.01	3.82
T.EXTR	74.12	9.11	94.31	5.06
D.GB	70.62	5.13	94.66	13.54
T.GB	73.12	37.58	95.00	52.77
D.LightGBM	72.09	0.63	84.66	0.93
T.LightGBM	73.62	0.62	95.00	1.98
D.XGBoost	70.59	1.62	92.93	3.80
T.XGBoost	71.59	2.33	94.00	6.14
D.AdaBoost	64.99	0.87	67.24	1.34
T.AdaBoost	72.60	1.24	88.62	25.42
D.CatBoost	72.13	46.10	96.72	49.56
T.CatBoost	70.60	15.59	94.14	21.07

Table 17 – Temps d'exécution et précision moyenne pour chaque modèle d'apprentissage d'ensemble avec et sans SMOTE

Figure 17 – Matrices de confusion pour Random Forest avant (à gauche) et après (à droite) l'étape de rééchantillonnage

Enfin, la figure 17 illustre les matrices de confusion résumant les performances de classification du classifieur Random Forest avant et après l'étape de ré-échantillonnage. Dans la matrice de droite de cette figure, la première ligne correspond aux 145 patients au stade ASYM (myélome asymptomatique) correctement classés. Les deuxième et troisième lignes indiquent que deux patients du stade I et trois patients du stade II ont été incorrectement classés comme appartenant au stade III. Cependant, 143 et 142 patients ont été correctement classés comme appartenant respectivement aux stades I et II. En revanche, dans la matrice de gauche de la figure 17, nous constatons qu'un seul patient du stade II et 144 patients du stade III ont été correctement classés, tandis que tous les autres patients ont été incorrectement classés.

11 Conclusion

Dans ce chapitre, nous avons étudié le classement et la sélection de caractéristiques pour le cancer MM en utilisant un ensemble de données collecté au CLCC-CHU Tlemcen en Algérie. Cet ensemble de données est déséquilibré car la plupart des patients sont au stade III. Pour compenser la distribution déséquilibrée des classes, nous avons utilisé SMOTE comme méthode de ré-échantillonnage.

Nous avons d'abord développé un modèle pour mesurer la corrélation entre les variables d'entrée et la classe cible en utilisant des méthodes de sélection de variables basées sur l'approche Filtre. Les résultats ont montré que les performances de classification étaient améliorées en utilisant les sous-ensembles de caractéristiques pertinentes retenues par les méthodes de sélection, plutôt que l'ensemble complet de caractéristiques. La technique de sélection CFS a donné de bons résultats lorsqu'elle a été utilisée avant le classifieur d'arbre de décision (C4.5), tandis que FCBF a été plus pratique pour sa robustesse et sa capacité à éliminer les caractéristiques non pertinentes.

Dans la deuxième partie de notre travail, nous avons abordé les méthodes d'ensemble basées sur les arbres de décision pour estimer l'importance des variables. Nous avons effectué un réglage approfondi des hyper-paramètres pour les algorithmes d'ensemble proposés en utilisant la technique GridSearchCV, obtenant des résultats très prometteurs et encourageants. LightGBM a été la méthode la plus rapide testée, tandis que Random Forest a donné une précision moyenne de plus de 97%, et XGBoost a donné le meilleur classement pour les caractéristiques considérées comme les facteurs les plus pronostiques.

Comme travail de recherche potentiel dans le chapitre suivant, nous pourrions proposer une intégration de méthodes d'apprentissage automatique basées sur les réseaux bayésiens pour découvrir des relations intéressantes entre les variables de notre base de données. Ces méthodes ont récemment été largement utilisées pour la découverte de connaissances dans les bases de données en utilisant certaines mesures d'intérêt.

CHAPITRE 3

ASSOCIATION DES VARIABLES POUR LE DIAGNOSTIC DU MM A L'AIDE D'UN MODÈLE BAYÉSIEN

1 Introduction

L'un des principaux objectifs de l'intelligence artificielle (IA) est de concevoir des systèmes capables de prendre des décisions similaires à celles du raisonnement humain. Ces dernières années, les avancées technologiques ont facilité l'acquisition et la collecte d'un grand nombre de données, notamment dans le domaine médical. Cette collecte de données de santé permet d'améliorer le suivi des patients, les relations entre les patients et les professionnels de santé, ainsi que la recherche clinique.

Cependant, il arrive parfois que les connaissances acquises ne soient pas suffisantes pour permettre au système de prendre la décision la plus appropriée. Dans le domaine médical, où l'incertitude est inhérente, il est nécessaire d'adopter une approche probabiliste, ce qui nous amène naturellement aux **"Réseau bayésiens"**.

Un réseau bayésien permet de représenter graphiquement les relations entre les attributs de manière fiable et transparente, en offrant la possibilité de prédire de nouveaux scénarios. Il est devenu une méthode largement utilisée pour modéliser des connaissances incertaines, en raison de sa capacité à modéliser les relations complexes entre les données et à estimer la probabilité de différents événements en fonction des données observées [117].

Dans ce chapitre, notre objectif est d'explorer et d'analyser des modèles graphiques probabilistes, tels que les réseaux bayésiens (BN), afin de mieux comprendre les relations entre les paramètres qui influencent le diagnostic du MM et son stade.

2 Réseaux Bayésiens

2.1 Définition

Les réseaux bayésiens (RB), qui tirent leur nom des travaux de **Thomas Bayes** sur la "probabilité des causes" au XVIIIe siècle [118], sont le fruit de recherches menées dans les années 80. Ils font partie des techniques de modélisation graphique probabiliste (Probabilistic Graphical Modelling (PGM)), qui reposent sur un formalisme basé sur les théories des probabilités et des graphes [119].

Un réseau bayésien permet de modéliser des connaissances incertaines et complexes, en représentant les relations d'influence (dépendances et indépendances) sous forme de graphes acycliques dirigés [120]. Il est largement utilisé dans le domaine de l'intelligence artificielle et de l'apprentissage automatique, notamment pour le diagnostic (médical et industriel) [121], la bioinformatique [122], l'analyse des risques [123, 124], la détection de spams [125], etc.

2.2 Graphe Acyclique Orienté (DAG)

Comme tout autre graphe statistique, un graphe acyclique orienté (DAG) est composé d'un ensemble de nœuds et de liens qui représentent les relations entre les nœuds. Par exemple, en observant la figure 18, nous pouvons voir des liens ou des flèches dirigés vers le nœud *Var3*, ce qui indique que *Var3* dépend à la fois de *Var1* et de *Var2*. Cela signifie que *Var1* et *Var2* sont les nœuds parents de *Var3*. De la même manière, *Var4* dépend de *Var3* et est donc l'enfant du nœud *Var3*. Il s'agit d'une relation simple que nous pouvons comprendre en observant la figure 18.

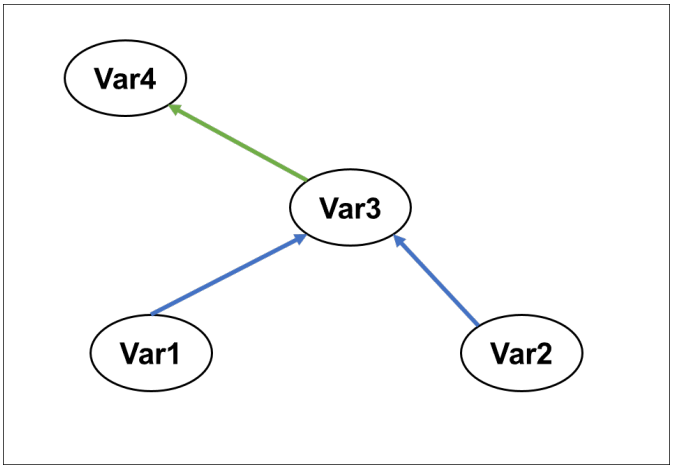


Figure 18 – Exemple d'un graphe acyclique dirigé

2.3 Théorème de Bayes

Le théorème de Bayes [118] est l'un des principaux théorèmes de la théorie des probabilités. Il permet de calculer la probabilité d'un événement en fonction de connaissances préalables sur les conditions qui pourraient être liées à cet événement. Le théorème de Bayes peut être formulé à partir des axiomes de base de la théorie des probabilités, en particulier la probabilité conditionnelle.

Supposons que A et B soient deux événements. La probabilité conditionnelle de l'événement A est la probabilité que l'événement se produise sachant que l'événement B s'est déjà produit. Mathématiquement, le théorème de Bayes peut être dérivé à partir de la définition de la probabilité conditionnelle :

$$P(A|B) = \frac{P(B \cap A)}{P(B)} \quad \text{si } P(B) \neq 0 \tag{3.1}$$

La formule du théorème de Bayes est donnée par (voir équation 3.2) :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{si } P(B) \neq 0 \tag{3.2}$$

Avec :

- P(A) et P(B) sont les probabilités de A et B respectivement, sans condition préalable. Elles sont connues sous le nom de probabilités a priori et probabilités marginales.
- $P(B \bigcap A)$ est la probabilité que A et B se réalisent simultanément.
- $P(A|B)$ est une probabilité conditionnelle : la probabilité que l'événement A sachant que B est vrai. Elle est également appelée probabilité postérieure de A étant donné B.
- $P(B|A)$ est également une probabilité conditionnelle : la probabilité que l'événement B se produise si A est vrai. Elle peut également être interprétée comme la probabilité de A étant donné un B fixe, car $P(B|A) = L(A|B)$

2.4 Indépendance conditionnelle

Un graphe acyclique orienté (DAG) permet de modéliser l'incertitude d'un événement en se basant sur la distribution de probabilité conditionnelle des variables dans un domaine donné.

La notion de probabilité conditionnelle conduit naturellement à celle d'indépendance. On dit que les événements A et B sont indépendants si

la probabilité conjointe de A et B est égale au produit des probabilités marginales de A et B, et on l'écrit de la manière suivante :

$$\forall A, B : \quad P(A, B) = p(A)P(B) \iff A \perp B$$

L'indépendance implicite dans un réseau bayésien peut être classée en deux types :

1. **Indépendances locales** : toute variable du réseau est indépendante de ses non-descendants étant donné ses parents. Cela peut s'écrire : $A \perp NonDesc(A) | Par(A)$, où *NonDesc(A)* est l'ensemble des variables qui ne sont pas des descendants de A et *Par(A)* est l'ensemble des variables qui sont les parents de A.
2. **Indépendances globales** : Pour discuter de ce type d'indépendances dans un RB, il faut examiner les différentes structures possibles du réseau. Par exemple, dans le cas d'un RB avec 2 nœuds, il n'y a que 2 façons possibles de les connecter (voir figure 19). Par conséquent, tout changement dans l'une des variables entraîne un changement dans l'autre variable.

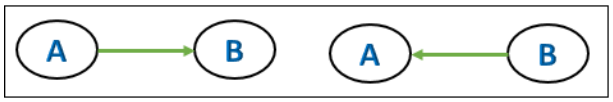


Figure 19 – Exemple 1 : Connexions possibles entre deux nœuds

2.4.1 Indépendance de *n* événements

Cette notion est plus complexe à comprendre et présente de nombreux pièges. Afin de faciliter la compréhension de cette idée, nous proposons un exemple d'un réseau bayésien à 3 nœuds.

Pour observer le flux d'influence de A vers C dans différents scénarios, il existe quatre configurations de connexion possibles (voir figure 20) :

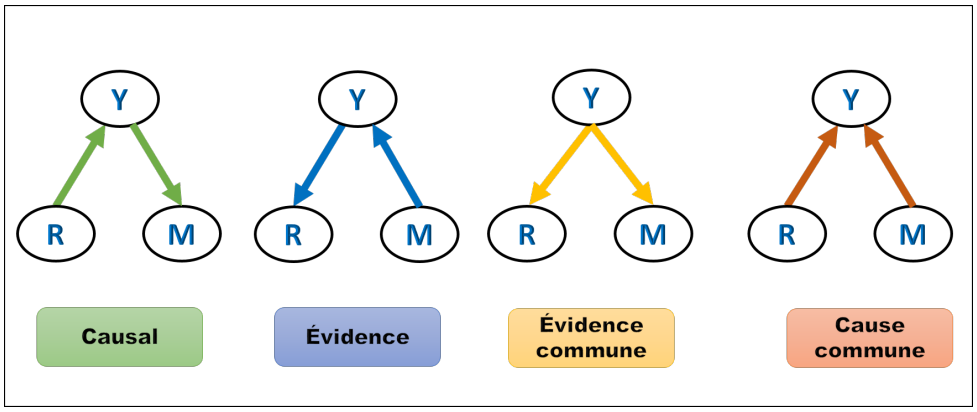


Figure 20 – Exemple 2 : Connexions possibles entre 3 nœuds

a/- Connexion série (causale) : Comme nous l'avons mentionné précédemment (dans le cas de deux nœuds), lorsque nous modifions la variable A, cela a un effet sur la variable B, et ce changement de B modifie ensuite les valeurs de C. Un autre cas possible est lorsque B est observé, c'est-à-dire que nous connaissons sa valeur. Dans ce cas, tout changement dans A n'affectera pas B car nous connaissons déjà sa valeur. Par conséquent, il n'y aura pas de changement dans C, car il dépend uniquement de B.

b/- Connexion divergente (évidence commune) : La même logique s'applique dans ce cas. Lorsque B est observé, cela rend C indépendant de A. Sinon, lorsque B n'est pas observé, l'influence passe de A à C.

Dans ces deux types de connexions (série et divergente) :

– A et C sont indépendants conditionnellement à B.
– $(A \perp C|B)$.

c/- V-structure (évidence commune) : Ce cas est un peu différent des autres. Lorsque B n'est pas observé, tout changement dans A se reflète dans un changement de B mais pas de C. Dans ce cas :
- A et C sont indépendants ;
- A et C sont dépendants conditionnellement à B ;
- $(A \not\perp C|B)$.

d/- Cause commune : L'influence passe de A à C lorsque B n'est pas observé, mais lorsque B est observé, le changement dans A n'affecte pas C car ce dernier dépend uniquement de B. On peut l'écrire comme suit : $(A \perp C|B)$.

3 Application des réseaux bayésiens dans le domaine médical

Les réseaux bayésiens sont de plus en plus utilisés en biomédecine, dans les soins de santé et dans les systèmes d'aide au diagnostic pour résoudre divers problèmes [126]. De nombreux travaux de recherche ont été proposés dans la littérature, basés sur des modèles bayésiens, afin d'analyser les données médicales, d'améliorer le processus de diagnostic et de suivre l'évolution des patients.

Dans le domaine médical réel, les ensembles de données sont généralement de petite taille, ne contenant que quelques centaines d'instances, et ils présentent souvent des données manquantes. Cette caractéristique est courante dans les domaines médicaux. Seuls les troubles très répandus, tels que le cancer du poumon ou l'infarctus du myocarde, sont généralement associés à des ensembles de données plus volumineux [127]. Cela soulève la question de la faisabilité d'apprendre les structures de réseaux bayésiens à partir de données dans le contexte des domaines médicaux réels.

Pour répondre à cette question, Wu et al. [128] ont réalisé une étude sur l'accident vasculaire cérébral. Ils ont construit un réseau bayésien basé sur un modèle causal conçu en collaboration avec un médecin spécialiste. Deux algorithmes d'apprentissage de structure différents ont été comparés, et les avantages et les limites de ces algorithmes ont été discutés en fonction des résultats expérimentaux obtenus. Leurs résultats ont montré que lorsque le nombre de cas était inférieur à 1000, la précision était très faible, même en l'absence de données manquantes. L'ensemble de données sur les accidents vasculaires cérébraux utilisé était trop petit pour les algorithmes disponibles, et la présence de nombreuses données manquantes a également affecté les résultats. Malgré cela, ils ont réussi à identifier des relations causales importantes entre les variables. Des améliorations sont attendues lorsque les ensembles de données sont plus importants.

En 2004, Cauchemez et al. [129] ont développé un modèle bayésien pour estimer les caractéristiques clés de la transmission de la grippe au sein des ménages. L'objectif de cette étude était d'estimer simultanément la durée de la période de transmission et les risques d'infection instantanés. L'estimation des paramètres du modèle de transmission était complexe car une grande partie du processus infectieux n'était pas directement observée : seules les dates de détection des nouveaux cas étaient disponibles. La durée moyenne de l'infection grippale a été estimée à 3,8 jours, avec un écart-type de 2 jours. Les chercheurs ont également observé que le risque

immédiat de transmission de la grippe entre une personne infectée et une personne saine au sein d'une famille diminuait avec la taille de la famille. De plus, ils ont constaté que les enfants (moins de 15 ans) étaient plus susceptibles de transmettre l'infection que les adultes (probabilité postérieure supérieure à 99%), bien que la durée moyenne de la période infectieuse était similaire chez les enfants et les adultes. La probabilité postérieure que les enfants présentent un risque communautaire plus élevé était de 76%, et la probabilité postérieure qu'ils soient plus sensibles que les adultes était de 79%.

Une autre étude réalisée en 2009 a introduit une approche bayésienne pour aider les médecins à prendre des décisions diagnostiques [130]. Les auteurs ont proposé un modèle bayésien naïf flou (Fuzzy Naive Bayesian - FNB) pour l'aide au diagnostic, basé sur l'extension de l'approche bayésienne floue proposée par Okuda [131]. Pour appliquer ce modèle, un système d'information flou orthogonal basé sur les symptômes a été défini à partir d'entretiens avec des médecins. Pour développer et évaluer les caractéristiques, l'algorithme a été appliqué à un ensemble de données simulées simple et comparé à une approche naïve bayésienne traditionnelle (NB). Pour évaluer les performances du FNB dans un scénario réel, la comparaison a été répétée sur un ensemble de données floues réelles comprenant 81 patients diagnostiqués avec des maladies infectieuses. Les résultats ont montré que le FNB pouvait être optimal par rapport au NB pour le diagnostic des patients lorsque des informations floues et imprécises étaient disponibles.

L'apprentissage de la structure des réseaux bayésiens (RB) à partir de données d'observation suscite un intérêt croissant dans divers domaines scientifiques et industriels, notamment dans le domaine médical [132], afin d'assurer la sécurité des patients. Dans leur travail, Zoullouti et al. [133] ont proposé des approches intégrées pour la gestion des risques dans un système hospitalier. Ils ont développé des méthodes qui prennent en compte différents aspects du risque et du type d'information disponible. La première approche est conçue pour un contexte où des données sur les événements à risque sont disponibles. Elle utilise les réseaux bayésiens pour analyser le risque de sécurité des patients dans la salle d'opération, qui est une zone à haut risque d'événements indésirables. Les réseaux bayésiens fournissent un cadre permettant de représenter les relations causales et de réaliser une inférence probabiliste entre un ensemble de variables. Dans la deuxième approche proposée, les chercheurs ont utilisé des réseaux bayésiens flous pour modéliser et analyser le risque. La logique floue permet d'utiliser les

opinions des experts lorsque les données quantitatives sont insuffisantes et que seules des déclarations qualitatives ou vagues peuvent être formulées. Les résultats ont montré que cette deuxième approche fournit un modèle exploitable qui soutient avec précision la cognition humaine en utilisant des variables linguistiques. Pour illustrer l'application de la méthode proposée, une étude de cas portant sur le risque de sécurité des patients en salle d'opération est utilisée.

L'objectif principal de la plupart des chercheurs dans la littérature est de fournir aux lecteurs une introduction aux réseaux bayésiens, qui sont des outils de représentation des connaissances et d'apprentissage automatique permettant d'estimer les risques en science médicale. Dans leur article, Arora et al. ont examiné comment les réseaux bayésiens sont des représentations graphiques compactes et intuitives des distributions de probabilités conjointes (JPD) qui peuvent être utilisées pour effectuer un raisonnement causal et une analyse d'estimation des risques. Ils ont souligné les avantages offerts par les réseaux bayésiens par rapport aux méthodes basées sur la régression, tout en abordant les défis associés aux méthodes traditionnelles de prédiction des risques. Ils ont ensuite décrit la construction, l'application et les avantages des réseaux bayésiens dans la prédiction des risques, en se basant sur des exemples liés au cancer et aux maladies cardiaques [134].

Malgré la disponibilité de nombreux logiciels open source pour les réseaux bayésiens, aucun d'entre eux n'est capable de traiter efficacement les données à la fois petites et grandes de l'espace des caractéristiques, tout en récupérant des structures de réseau avec une précision acceptable. C'est dans ce contexte que le logiciel bAIcis a été développé par le BERG. Il vise à apprendre les réseaux bayésiens à partir de "Big Data" dans le domaine de la santé, qui souvent dépassent des centaines de milliers de caractéristiques lors de la recherche en génomique ou en multi-omique. Dans leur article, Lixia et al. [135] ont présenté une évaluation complète des performances de bAIcis et l'ont comparée à d'autres algorithmes RB open source. L'étude a été réalisée sur des ensembles de données synthétiques discrètes, continues et mixtes, dans des espaces de caractéristiques petits et grands respectivement. Les résultats obtenus ont démontré que bAIcis surpassait les algorithmes accessibles au public en termes de précision de récupération de structure, atteignant des taux de vrais positifs de 90% et une précision de 80%.

4 Construction d'un réseau bayésien

La construction d'un réseau bayésien peut être réalisée de deux manières : **une construction manuelle** ou **une construction automatique** (appelée "apprentissage") à partir de bases de données [136]. La construction manuelle suppose une connaissance préalable du domaine par un expert. La première étape consiste à construire un graphe acyclique dirigé, suivie de l'évaluation de la distribution de probabilité conditionnelle dans chaque nœud.

Contrairement à la construction manuelle, les réseaux bayésiens basés sur l'apprentissage automatique ne nécessitent pas de connaissance préalable du domaine. Ils peuvent être appris automatiquement à partir de bases de données à l'aide d'algorithmes souvent intégrés à des logiciels spécialisés. Cependant, l'inconvénient est que la construction automatique impose des exigences plus élevées sur les données. La plupart des algorithmes d'apprentissage automatique supposent que les données sont complètes, ce qui est souvent une hypothèse très forte dans la pratique. Si des données sont manquantes, elles doivent être importées, imputées ou estimées à partir d'autres sources.

La construction d'un réseau bayésien implique trois étapes principales (voir figure 21) :

1. Identification des variables et de leurs espaces d'états : Cette étape qualitative consiste à déterminer l'ensemble des variables X_i, qu'elles soient catégorielles ou numériques, qui caractérisent le système. Ensuite, il est nécessaire de spécifier l'espace d'états de chaque variable X_i, c'est-à-dire l'ensemble de ses valeurs possibles. C'est la seule étape de la construction du réseau qui nécessite l'intervention humaine.
2. Définition de la structure du réseau : La détermination de la structure du réseau revient à répondre à la question suivante pour chaque variable X du réseau : quelles sont les variables que nous considérons comme les causes directes de X ?
3. Définition de la loi de probabilité conjointe des variables : La difficulté et l'objectivité de cette étape varient considérablement d'un problème à l'autre. Les tableaux de probabilités conditionnelles peuvent parfois être : déterminés entièrement à partir de l'énoncé du problème par des considérations objectives, le reflet de croyances subjectives, ou estimés à partir des données.

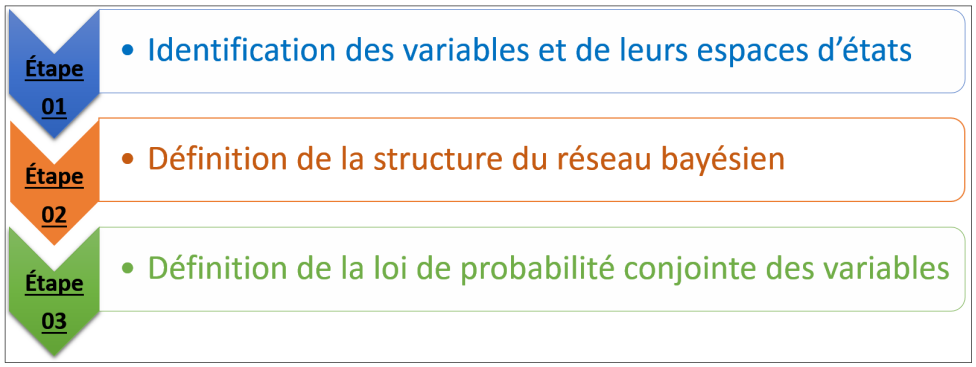


Figure 21 – Étapes de construction d'un réseau bayésien

4.1 Discrétisation

En mathématiques, les données continues ont un nombre infini de degrés de liberté (DDL), ce qui peut entraîner des problèmes de calcul potentiellement infinis [137]. C'est pourquoi les scientifiques des données utilisent la discrétisation.

La discrétisation est le processus de conversion de variables continues en une forme discrète [138]. Cette technique est largement utilisée dans la littérature pour créer des fonctions de densité de probabilité.

Les méthodes de discrétisation des variables continues dans les modèles bayésiens sont généralement classées en trois catégories principales [139] : manuelle, supervisée et non supervisée.

1. **Méthode de discrétisation supervisée** : Il s'agit d'une méthode informative qui prend en compte la variable de sortie lors de la définition des limites de discrétisation, ce qui est souvent une approche bénéfique. La discrétisation supervisée nécessite une variable de sortie discrète pour effectuer la discrétisation des variables d'entrée continues. Cela signifie que si la variable de sortie est continue, des connaissances a priori, des hypothèses ou une méthode de discrétisation non supervisée sont nécessaires pour la discrétiser avant de pouvoir utiliser une méthode supervisée pour les variables d'entrée [139]. Les algorithmes de discrétisation supervisée sont couramment utilisés en informatique et ont démontré leur efficacité dans l'utilisation de réseaux bayésiens sur divers ensembles de données [140]. Cependant, ils sont peu présents dans la littérature sur les réseaux bayésiens en raison des limitations des packages existants pour les réseaux bayésiens et des disparités persistantes entre la modélisation des réseaux bayésiens en informatique et l'apprentissage automatique.
2. **Méthodes de discrétisation non supervisée** : Cette méthode est une étape cruciale et populaire dans de nombreuses tâches de découverte

de connaissances, car elle est simple à calculer et objective. Elle est basée sur la distribution intrinsèque des données de chaque variable individuelle et utilise une stratégie de découpage descendant et une stratégie de fusion ascendante [140]. Les algorithmes de discrétisation en intervalles de même taille (equal width intervals) et en intervalles de même fréquence (equal frequency intervals) sont les plus couramment utilisés dans de nombreuses applications [141]. Le premier type divise les données continues en un nombre prédéfini d'intervalles de largeur égale, tandis que le second type divise les données en un nombre prédéfini d'intervalles de fréquence égale.

3. **Méthodes de discrétisation manuelle** : Également appelée discrétisation experte, cette méthode consiste à sélectionner manuellement les seuils de discrétisation en fonction de leur signification physique, de connaissances théoriques ou d'une interprétation experte du domaine du problème. Une revue des articles sur l'application et la modélisation des réseaux bayésiens [142] a noté que la discrétisation manuelle est souvent privilégiée dans les études nécessitant la discrétisation de données continues, en particulier dans le domaine médical. Elle permet de discrétiser les variables continues en intervalles interprétables et pertinents pour les objectifs du modèle d'étude, sans nécessiter d'algorithme de discrétisation supplémentaire.

4.2 Apprentissage de la structure

L'apprentissage de la structure des réseaux bayésiens (ASRB) est le processus qui consiste à apprendre les liens d'un réseau bayésien et la structure du graphe acyclique dirigé (DAG) à partir d'un ensemble de données [143].

L'ASRB peut être formulé pour répondre à de nouvelles requêtes. Dans ce cas, même si l'on dispose d'une certaine expertise dans le domaine, elle peut ne pas être suffisante pour produire un modèle utilisable. Il est alors possible d'obtenir de meilleurs résultats en apprenant à partir des données et en identifiant réellement les dépendances les plus significatives que les données indiquent.

Le deuxième scénario est celui où l'on n'a pas nécessairement l'intention d'utiliser le réseau, mais où l'on veut simplement le découvrir. Ce type d'utilisation du modèle se produit, par exemple, dans les ensembles de données scientifiques ou biologiques, où l'objectif est de découvrir les interrelations entre les variables afin de mieux comprendre le domaine.

Les algorithmes d'ASRB sont regroupés en trois approches principales : les algorithmes basés sur les scores (score-based learning), les algorithmes basés sur les contraintes (constraint-based learning) et les algorithmes hybrides.

4.2.1 Algorithmes basés sur les scores

Les algorithmes basés sur les scores (Score-based algorithms) sont des applications de techniques d'optimisation générales. Chaque DAG candidat se voit attribuer un score de réseau maximisé en tant que fonction objective. Cette approche consiste à définir d'abord un critère d'évaluation de l'adéquation du réseau bayésien aux données, puis à rechercher dans l'espace des DAG une structure qui atteint le score maximum. Le problème se compose essentiellement de deux parties :

- L'algorithme de recherche qui optimise l'espace de recherche de tous les DAG possibles, tels que ExhaustiveSearch, Hillclimbsearch et Chow-Liu.

- La fonction de score qui indique dans quelle mesure le réseau bayésien s'adapte aux données. Les fonctions de score couramment utilisées sont les scores de Dirichlet bayésiens tels que BDeu ou K2, ainsi que le critère d'information bayésien (BIC, également appelé MDL).

Les métriques de score pour une structure G et des données D peuvent être généralement définies comme suit :

$$score(G : D) = LL(G : D) - \Phi(|D|)||G|| \tag{3.3}$$

Où :

$LL(G : D)$ est la mesure de log-vraisemblance des données sous la structure du graphe G. $|D|$ est le nombre d'échantillons de données. $||G||$ est le nombre de paramètres dans le graphe G. $\Phi(|D|)||G||$ est un terme de régularisation qui favorise les modèles simples.

➥ Lorsque $\Phi(t) = 1$, la fonction de score est connue sous le nom de critère d'information d'Akaike (AIC).

➥ Lorsque $\Phi(t) = \frac{\log(t)}{2}$, la fonction de score est connue sous le nom de critère d'information bayésien (BIC).

Il existe une autre famille de fonctions de score bayésiennes appelée score bayésien de Dirichlet (BD). Pour ce score, on définit d'abord la probabilité des données D conditionnellement à la structure du graphe G comme suit : $P(D|G) = \int P(D|G, \theta_G)P(\theta_G|G)d\theta_G$.

Où $P(D|G, \theta_G)$ est la probabilité des données compte tenu de la structure et des paramètres du réseau, et $P(\theta_G|G)$ est la probabilité a priori des paramètres.

Les algorithmes de recherche les plus courants sont la recherche locale et la recherche gloutonne. Dans le premier type, on commence avec un graphe vide ou un graphe complet. À chaque étape, on tente de modifier la structure du graphe en ajoutant, supprimant ou inversant une arête. Si le score augmente, la tentative est adoptée et le changement est effectué ; sinon, une autre tentative est faite. Dans le cas de la recherche gloutonne (par exemple, l'algorithme K3), on suppose un ordre topologique du graphe.

4.2.2 Algorithmes basés sur les contraintes

Les algorithmes basés sur les contraintes (Constraint-based algorithms) identifient un ensemble de contraintes d'arêtes pour le graphe en utilisant des tests d'indépendance, puis trouvent le meilleur DAG qui satisfait ces contraintes. Cette approche fonctionne bien en présence de connaissances préalables sur la structure, mais nécessite un grand nombre d'échantillons de données pour garantir la puissance du test. Par conséquent, elle est moins fiable lorsque le nombre d'échantillons est faible.

4.2.3 Algorithmes hybrides

Les algorithmes hybrides (Hybrid algorithms) comprennent une phase de restriction qui met en œuvre une stratégie basée sur les contraintes pour réduire l'espace des DAG candidats, ainsi qu'une phase de maximisation qui met en œuvre une stratégie basée sur les scores pour trouver le DAG optimal dans l'espace restreint.

4.3 Apprentissage des paramètres

Les réseaux bayésiens fournissent des résultats de décision pour les systèmes experts, qui sont basés sur la structure du réseau et les tables de probabilités conditionnelles (CPT). Lorsque la table de probabilités conditionnelles est inconnue, il est nécessaire d'apprendre à partir des

données observées afin d'obtenir les paramètres de probabilités conditionnelles. L'apprentissage des paramètres des réseaux bayésiens est une étape importante dans l'apprentissage de ces réseaux, car elle permet d'estimer les paramètres de probabilité conditionnelle pour toutes les relations causales du réseau [144]. Cette procédure s'effectue à partir des données relatives au problème à modéliser, qui peuvent être complètes ou incomplètes. Dans cet exposé, nous nous intéressons aux problèmes impliquant des données complètes.

5 Méthodologie proposée

Dans le but d'optimiser l'inférence clinique, nous avons exploré et analysé les modèles graphiques probabilistes, en particulier les Réseaux Bayésiens (RB). Les RB visent à comprendre les relations causales entre les variables du jeu de données MM et le stade de MM, afin de produire une représentation graphique fiable et transparente. Cela nous permet de mieux comprendre les relations entre les paramètres influençant le diagnostic du MM et offre la possibilité de prédire de nouveaux scénarios.

Pour rappel, notre jeu de données (MM_dataset [4]), collecté dans le CLCC (CHU-Tlm), contient toutes les informations trouvées dans les rapports de diagnostic des patients atteints du MM (voir Tableau 12). De plus, le stade du cancer a été déterminé par les médecins spécialistes pour chaque patient lors du premier diagnostic, en utilisant les systèmes de stadification internationale.

Dans le cadre de l'apprentissage structurel, notre objectif est de déterminer la structure du graphe qui capture au mieux les dépendances causales entre les variables de l'ensemble de données. En d'autres termes, nous cherchons à savoir quel DAG correspond le mieux à nos données sur le MM et quelles variables ont un effet causal direct sur la classe cible.

Pour effectuer des inférences, deux éléments sont nécessaires : le DAG et les tables de probabilités conditionnelles (CPT) des données. Comme mentionné précédemment, le DAG est déterminé lors de l'étape précédente. Les CPT peuvent être calculées en utilisant l'apprentissage des paramètres. Ainsi, nous commencerons par l'apprentissage des paramètres avant de passer à l'inférence. Les CPT sont essentielles pour décrire quantitativement la relation statistique entre chaque nœud et ses parents.

Dans notre travail, la conception du RB repose principalement sur des attributs discrets. Par conséquent, nous avons effectué une étape de discrétisation des variables avant de procéder à l'apprentissage et à la construction du graphe.

La Figure 22 présente un diagramme expliquant les différentes étapes impliquées dans la conception du réseau.

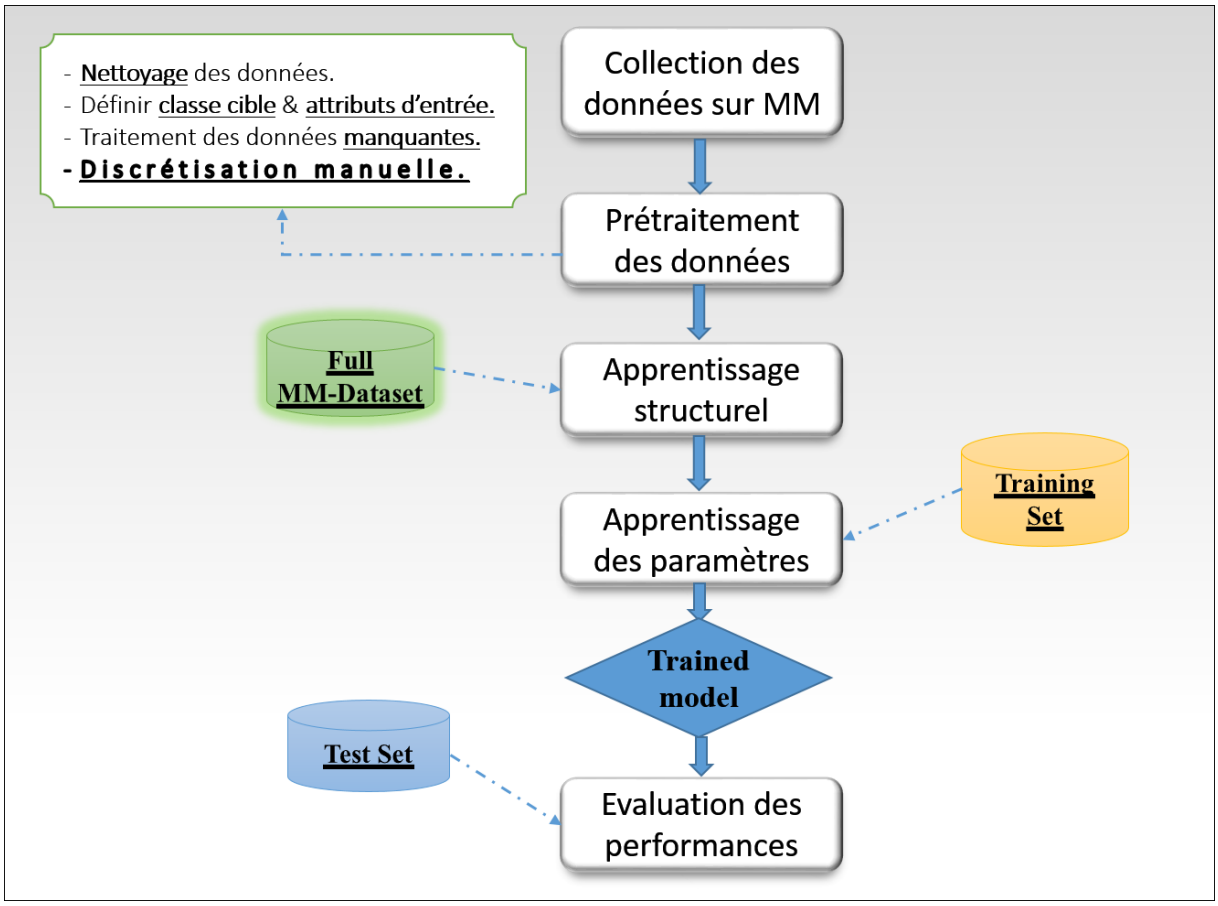


Figure 22 – Diagramme de la conception du réseau proposé.

Nous procéderons à un apprentissage structurel sur l'ensemble de données, suivi d'un apprentissage des paramètres en utilisant un ensemble d'apprentissage qui représente 70 % de l'ensemble de données total. Ensuite, nous évaluerons la précision du modèle en utilisant un ensemble de test qui comprend les 30 % restants de l'ensemble de données.

Cette section présente les différents matériels et méthodes utilisés dans le cadre de cette étude. Elle aborde en détail l'objectif de la recherche, le prétraitement des données et les outils utilisés pour l'étude.

5.1 Discrétisation des données du jeu de données MM-dataset

Bien que les données réelles contiennent souvent un mélange de variables discrètes et continues, de nombreux algorithmes bayésiens ne fonctionnent qu'avec des variables discrètes. Par conséquent, il est souvent nécessaire de discrétiser les données continues afin de les utiliser dans la modélisation et l'apprentissage d'un RB.

Il existe trois principales méthodes de discrétisation des données continues pour les utiliser dans un RB : (i) la discrétisation manuelle, effectuée par un expert du domaine ; (ii) la discrétisation supervisée, dans laquelle la variable de sortie est utilisée pour optimiser automatiquement la discrétisation des autres variables ; et (iii) la discrétisation non supervisée, basée sur la distribution de chaque variable individuelle, sans utiliser la variable de sortie ou si celle-ci n'est pas disponible.

Dans le domaine médical, la méthode de discrétisation manuelle est souvent préférée car elle permet de définir des intervalles de discrétisation en se basant sur des seuils naturels et interprétables.

Dans notre ensemble de données, 34 des 53 attributs sont continus. Étant donné que les variables des modèles de RB sont par nature discrètes, nous devons donc convertir ces données continues en données catégoriques. Nous nous appuyons sur l'expertise médicale pour effectuer la discrétisation de nos données. Le tableau ci-dessous (voir Table 18) présente la description des variables continues dans notre jeu de données ainsi que leurs valeurs discrétisées.

Variable	Règles de discrétisation	Variable	Règles de discrétisation	Variable	Règles de discrétisation
age	age <65 ans	palsma_cells	Normal = <10 %	alb (g/L)	alb >= 35
	age >= 65 ans		High >10 %		alb >35
body_surf	body_surf <1.71	Ca (mg/L)	hypoCa <85	alpha (g/L)	Low
	body_surf >= 1.71		85 <Normal <110		Normal
weight	underweight		hyperCa >110		High
	healthy	K (mmol/L)	Low <3.5	beta (g/L)	Low
	obese		3.5 <Normal <5		Normal
	overweight		High >5		High
CBC_WBC (*10^3/mm^3)	Low <3.8	P (mg/dL)	hypoPhos <2.5	gamma (g/L)	Low
	3.8 <Normal <11		2.5 <Normal <5		Normal
	High >11		hyperPhos >5		High
CBC_RBC (*10^6/mm^3)	Low <3.8	Na (Meq/L)	hypoNa <135	prot_rate (g/L)	Low
	3.8 <Normal <5.9		135 <Normal <145		Normal
	High >5.9		hyperNa >145		High
CBC_plats (*10^3/mm^3)	Low <150	B2M (mg/L)	Normal <3.5	TGO (UI/L)	Low
	150 <Normal <450		[3.5 , 5.5]		High
	High >450		High >5.5	TGP (UI/L)	Low
CBC_Hgb (g/dL)	anemia <11	creat (mg/L)	creat =<20		High
	no_anemia >11		creat >20	GGT (UI/L)	Low
CBC_Hct	Low <38%	urea (g/L)	Low <0.15		High
	38% <Normal <49%		0.15 <Normal <0.50	PAL (UI/L)	Low
	High >49%		High >0.50		High
CBC_VGM (u^3)	microcytose <80	clair_creat (ml/min)	prob <40	TP	Low <70 %
	80 <normocytose <100		no_prob >40		Normal >= 70 %
	macrocytose >100	gly (g/L)	hypoGly <0.70	LDH (U/L)	LDH <400
CBC_CCMH (g/dL)	Low <32		0.70 <Normal <1.10		LDH >400
	32 <Normal <36		HyperGly >1.10	FE	FE <55 %
	High >36	Fib (g/L)	Low <2		FE >= 55 %
Ferr (ng/mL)	Low <13		2 <Normal <4		
	13 <Normal <225		High >4		

Table 18 – Intervalles de discrétisation des variables continues MM

5.2 Phase d'apprentissage structurel

Après avoir effectué le prétraitement des données pour les adapter au modèle à construire, nous passons à la phase d'apprentissage structurel dans les réseaux bayésiens en utilisant la bibliothèque *pgmpy*[1] en Python. Cette étape permet de déterminer les dépendances entre les variables du réseau.

pgmpy est une bibliothèque puissante et flexible qui permet de construire et de manipuler des modèles graphiques probabilistes (PGM) en Python, notamment des réseaux bayésiens [145]. Elle peut être utilisée dans diverses applications telles que le raisonnement probabiliste, la prise de décision et la prédiction. Actuellement, pgmpy propose l'implémentation de trois algorithmes principaux : (1) l'algorithme PC (Constraint-Based Estimator) avec des variantes stables et parallèles, (2) l'algorithme Hill-ClimbSearch et (3) l'algorithme Exhaustive Search.

Dans notre travail, nous utiliserons un algorithme d'apprentissage de structure spécifique, que nous présenterons dans la section suivante.

Algorithme Hill-ClimbSearch

Hill-ClimbSearch (Hc) [146] est un algorithme de recherche utilisé pour construire des réseaux bayésiens à partir de données. Son objectif principal est de découvrir un DAG optimal qui correspond le mieux aux données étudiées. L'algorithme Hc construit progressivement un modèle bayésien en ajoutant ou supprimant des liens entre les différentes variables, jusqu'à ce que la configuration qui maximise la vraisemblance des données soit atteinte. Cet algorithme de recherche utilise la descente de gradient itérative pour optimiser la probabilité jointe du réseau bayésien et améliorer ainsi la prédiction ou l'inférence.

Les fonctions de score pouvant être utilisées avec l'algorithme Hc sont : K2-Score, BDeu et Bic.

L'architecture de l'algorithme Hc est simple et facile à mettre en œuvre, ce qui en fait un choix populaire pour résoudre un large éventail de problèmes d'optimisation. Cependant, comme il s'agit d'un algorithme de recherche locale, il peut être piégé dans des optima locaux ou atteindre des solutions sous-optimales.

Il est important de noter que l'implémentation spécifique de l'algorithme Hill-ClimbSearch pour les réseaux bayésiens peut varier en fonction du logiciel ou du langage de programmation utilisé. Les étapes suivantes décrivent la procédure générale pour trouver et optimiser la structure op-

1. `https://github.com/pgmpy/pgmpy`

timale d'un réseau bayésien à l'aide de cet algorithme :

1. **Initialisation** : Commencer avec une structure de réseau bayésien vide ou pré-définie. Dans la plupart des cas, l'algorithme Hc pour les RB commence par une structure vide et déconnectée. L'étape d'initialisation permet donc de créer un RB vide, où chaque nœud représente une variable et n'a aucune relation avec les autres nœuds. Au début de l'algorithme, chaque nœud est considéré comme étant indépendant des autres. Il convient de noter que l'utilisation d'une structure vide et déconnectée comme point de départ n'est pas une obligation stricte. Dans certains cas, une structure initiale pré-définie ou une structure générée aléatoirement peut être utilisée.
2. **Définition d'une méthode de score** : Sélectionner une fonction de score, telle que "K2Score" ou "BDeuScore", pour évaluer la qualité de la structure du RB.
3. **Évaluation de la structure initiale** : Calculer le score de la structure initiale du réseau en utilisant la métrique de score choisie.
4. **Recherche locale** : Effectuer de manière itérative les étapes suivantes jusqu'à ce qu'aucune amélioration ne soit possible ou qu'un critère d'arrêt soit atteint :
   - Sélectionner aléatoirement un nœud ou une arête de la structure RB actuelle.
   - Évaluer les modifications possibles, telles que l'ajout ou la suppression d'arêtes ou le changement de direction d'une arête.
   - Évaluer la structure modifiée en utilisant la métrique de score choisie.
   - Mettre à jour la structure du réseau avec la modification ou conserver la structure actuelle.
5. **Renvoi de la structure de réseau optimale** : Lorsque le critère d'arrêt est atteint, la meilleure structure de réseau avec le score le plus élevé obtenu pendant la recherche est renvoyée.

5.3 Phase d'apprentissage des paramètres

L'apprentissage des paramètres dans les réseaux bayésiens est un sujet très important qui implique l'estimation des tables de probabilités conditionnelles (CPT) pour chaque nœud du réseau [144]. Il existe principalement deux catégories de méthodes pour l'estimation des paramètres dans les réseaux bayésiens : l'une convient pour traiter les données complètes, l'autre pour les données incomplètes. Nous nous sommes concentrés sur

deux méthodes couramment utilisées dans la première catégorie : l'estimation du maximum de vraisemblance (Maximum Likelihood Estimate) et la méthode d'estimation bayésienne (Bayesian Estimation method).

5.3.1 Estimation bayésienne

L'estimation bayésienne (BE) est une méthode statistique populaire utilisée pour l'apprentissage des paramètres dans les réseaux bayésiens. Elle se base sur le théorème de Bayes pour estimer la distribution postérieure des paramètres en fonction des données [147]. Elle nécessite la définition d'une distribution préalable des paramètres qui est combinée à la vraisemblance des données pour obtenir une distribution postérieure.

L'estimation bayésienne présente plusieurs avantages pour les réseaux bayésiens. Elle permet d'intégrer des connaissances préalables ou des hypothèses sur les valeurs des paramètres, ce qui peut influencer les estimations finales. De plus, elle permet de traiter les données manquantes et les observations bruyantes, ainsi que d'effectuer une sélection et une comparaison de modèles [148].

Cependant, l'estimation bayésienne peut être très gourmande en ressources informatiques et nécessiter une puissance de calcul importante. De plus, elle nécessite la spécification de distributions préalables, ce qui peut être difficile, notamment si les connaissances préalables sont limitées.

5.3.2 Estimation du Maximum de Vraisemblance

L'estimation du maximum de vraisemblance (MLE) est une méthode qui consiste à trouver les valeurs des paramètres qui maximisent la vraisemblance des données observées [149]. Dans un réseau bayésien, les paramètres des nœuds du réseau sont des probabilités. La méthode du MLE suppose que les données sont générées indépendamment à partir de la table de probabilité conditionnelle (CPT) de chaque nœud. Cela implique de trouver la valeur des paramètres qui maximise la fonction de vraisemblance, qui est calculée en multipliant les probabilités de chaque observation dans les données.

Une fois l'estimation du maximum de vraisemblance calculée, elle peut être utilisée pour mettre à jour les croyances préalables concernant les probabilités dans le réseau bayésien. Ces connaissances actualisées peuvent être utilisées pour effectuer de meilleures prédictions et améliorer la précision du réseau bayésien.

6 Résultats et discussions

Cette section aborde les résultats de la recherche et compare les modèles bayésiens estimés à l'aide de l'algorithme HillClimbSearch, en utilisant les méthodes de scoring "K2Score" et "BDeuScore" respectivement. Nous commençons par présenter les DAG obtenus lors de l'apprentissage de la structure du modèle à partir de l'ensemble de données complet. Ensuite, nous effectuons l'apprentissage des paramètres en utilisant un ensemble d'entraînement. Enfin, nous évaluons la performance des différents modèles bayésiens obtenus en utilisant un ensemble de test, en prenant en compte divers critères tels que la précision.

6.1 Apprentissage de la structure avec l'algorithme HillClimbSearch

En utilisant l'algorithme HillClimbSearch et en fonction de la méthode de scoring utilisée, nous effectuons une recherche locale par escalade pour estimer la structure DAG optimale. Le point de départ de la recherche locale est généralement déterminé de manière aléatoire, mais il peut également être défini manuellement. L'idée est de commencer la recherche à un point aléatoire ou spécifié par l'utilisateur dans l'espace de recherche, puis de se déplacer itérativement vers le meilleur voisin jusqu'à atteindre un optimum local. Dans la bibliothèque pgmpy pour les RB en Python, un réseau complètement déconnecté est utilisé par défaut comme point de départ.

Les figures 23 et 24 présentent les graphes acycliques dirigés (DAG) optimaux générés par l'algorithme HillClimbSearch en utilisant les méthodes de scoring K2Score et BDeuScore respectivement. Nous pouvons observer que les nœuds des deux modèles RB estimés par K2Score et BDeuScore représentent les variables de notre ensemble de données, tandis que les arêtes (arcs) reflètent les relations parent/enfant entre ces variables (voir Tables 19 et 20). Les deux graphes sont acycliques, ce qui signifie qu'il n'y a pas de boucles.

Lors de la comparaison des deux DAG, plusieurs critères peuvent être pris en compte pour évaluer leur similarité ou leur différence. Les critères à utiliser dépendent de l'objectif de la comparaison et des spécifications du problème.

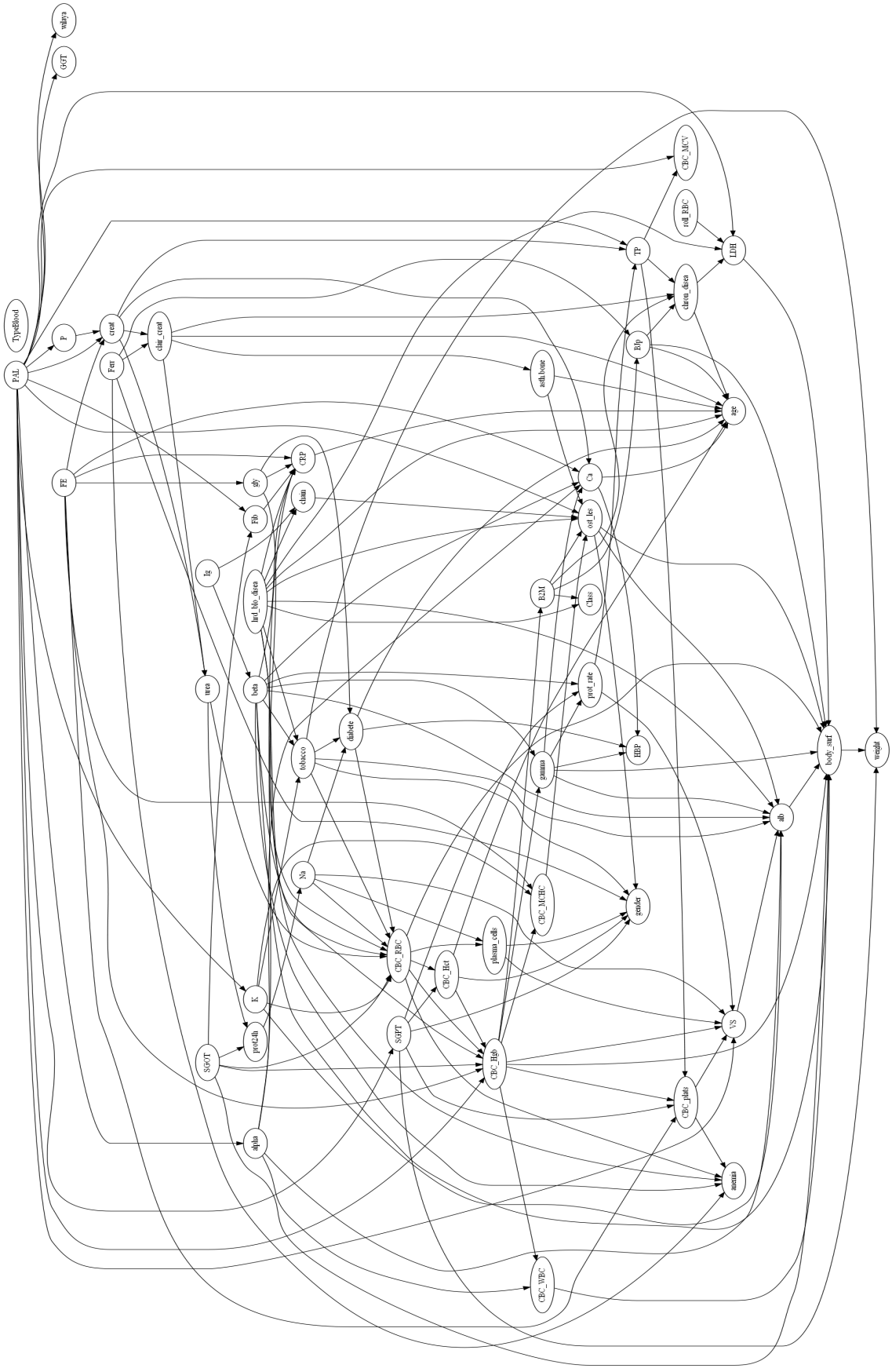


Figure 23 – DAG1 estimé par l'algorithme HillClimbsearch ; Fonction de score "K2Score".

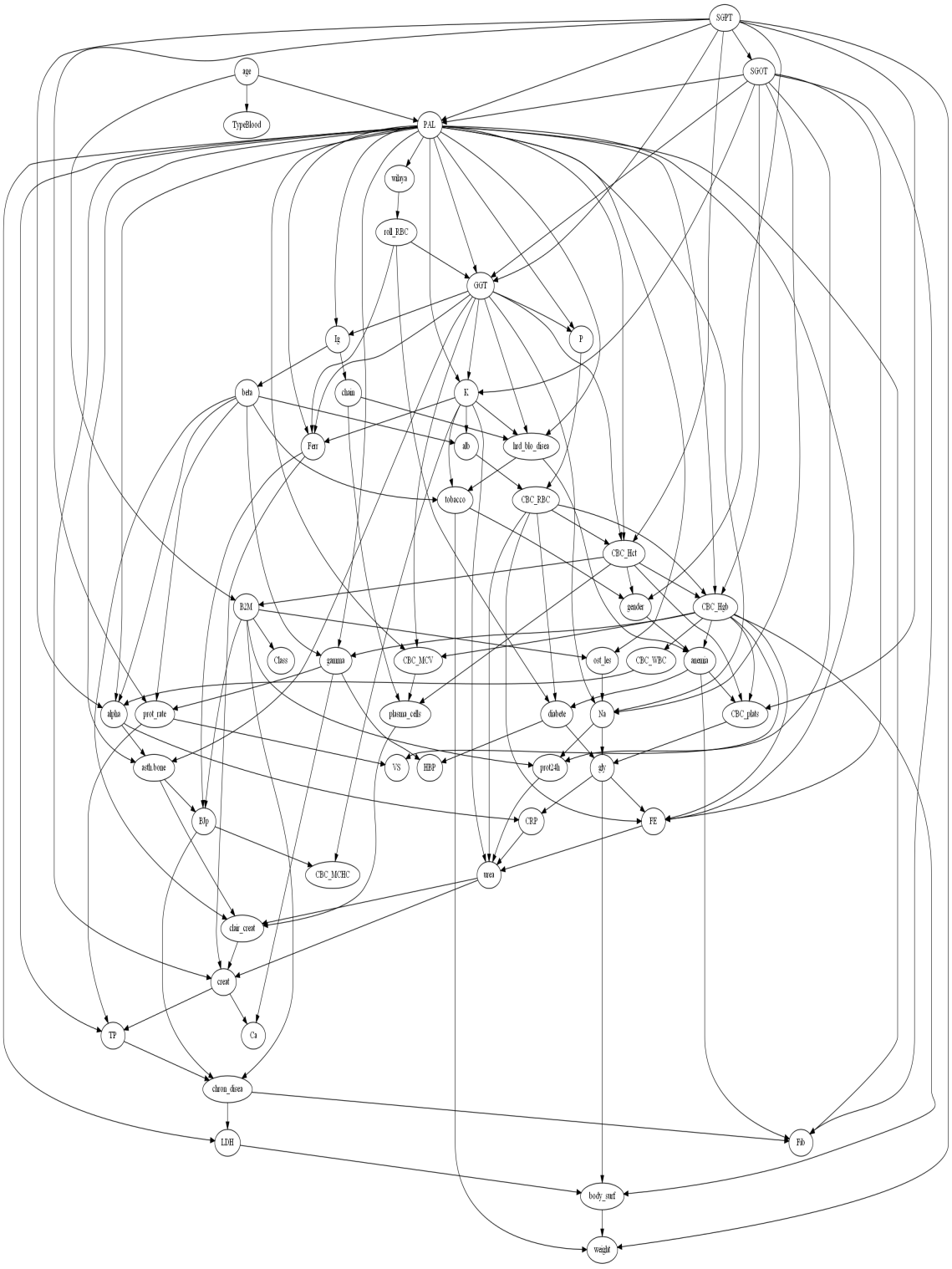


Figure 24 – DAG2 estimé par l'algorithme HillClimbsearch ; Fonction de score "BDeuScore".

Nœuds Parent	Nœuds Enfants
body_surf	weight
diabete	HBP, CBC_RBC, age
tobacco	gender, diabete, weight, alb, CBC_RBC
chron_disea	age, LDH
hrd_blo_disea	**Class**, anemia, ost_les, CRP, alb, tobacco, CBC_Hgb, age, chain, LDH
asth.bone	age, ost_les
CBC_WBC	body_surf
CB_RBC	CBC_Hgb, anemia, plasma_cells, CBC_Hct, body_surf
CBC_plats	VS, anemia
CBC_Hgb	VS, gamma, CBC_plats, CBC_WBC, CBC_MCHC, B2M, body_surf
CBC_Hct	gender, CBC_Hgb, age
CBC_MCHC	ost_les
VS	alb
roll_RBC	LDH
plasma_cells	gender, VS
Ca	age, HBP
K	CBC_MCHC, alb, tobacco, CBC_RBC
P	creat
Na	diabete, VS, plasma_cells, CBC_RBC
B2M	chron_disea, **Class**, ost_les, BJp
CRP	age
creat	clair_creat, Ca, TP, urea
urea	prot24h, CBC_RBC
clair_creat	urea, age, asth.bone, chron_disea
prot24h	Na
BJp	chron_disea, age, body_surf
alb	body_surf
alpha	CRP, alb, CBC_WBC, Ca
beta	prot_rate, alb, gamma, anemia, CRP, body_surf, Ca, tobacco, CBC_RBC
gamma	prot_rate, alb, Ca, HBP, body_surf
prot_rate	TP, VS
Ig	chain, beta
chain	ost_les
ost_les	alb, body_surf, gender
SGOT	CBC_Hgb, prot24h, body_surf, Fib, CBC_RBC
SGPT	weight, gender, CBC_Hct, CBC_plats, prot_rate
GGT	PAL, SGPT, TP, ost_les, CBC_Hgb, P, LDH, VS, alpha, CBC_MCV, Fib
PAL	wilaya, K, creat
gly	diabete, CRP, CBC_RBC
TP	chron_disea, CBC_plats, CBC_MCV
Fib	CRP
Ferr	BJp, anemia, clair_creat, gender
LDH	body_surf
FE	gly, creat, CRP, CBC_MCHC, Ca, CBC_plats, CBC_Hgb

Table 19 – Les arcs Parent/Enfants pour DAG1 (avec K2Score)

Nœuds Parent	Nœuds Enfants
wilaya	roll_RBC
gender	anemia
age	B2M
body_surf	weight
diabete	HBP, gly
tobacco	gender, weight
chron_disea	LDH, Fib
hrd_blo_disea	anemia, PAL, tobacco
asth.bone	clair_creat, urea, BJp, PAL
anemia	CBC_plats, diabete, Fib
CBC_RBC	CBC_Hgb, CBC_Hct, TP, gamma, diabete
CBC_plats	TP, gly
CBC_Hgb	VS, anemia, CBC_WBC, CBC_MCV, body_surf, CBC_plats, B2M
CBC_Hct	gender, CBC_Hgb, CBC_plats, plasma_cells, P
CBC_MCV	plasma_cells
CBC_MCHC	FE
VS	prot_rate
roll_RBC	diabete
Ca	age
K	GGT, hrd_blo_disea, CBC_MCHC, tobacco
Na	prot_rate
B2M	chron_disea, **Class**, ost_les, BJp
creat	Ca, urea
urea	CBC_RBC, prot24h
clair_creat	creat, urea, age
prot24h	Na, SGOT, CBC_RBC
BJp	chron_disea
alpha	asth.bone, CRP, CBC_WBC
beta	gamma, alb, tobacco, alpha, CBC_RBC
gamma	Ca, alb, HBP
prot_rate	gamma
Ig	beta, Ferr, VS
chain	Ig, plasma_cells, hrd_blo_disea
SGOT	PAL, Fib, CBC_Hgb, Na, SGPT, hrd_blo_disea
SGPT	, CBC_Hct, CBC_plats, gender, weight, gamma
GGT	PAL, SGPT, SGOT, P, CBC_MCV, alpha, CBC_Hct, LDH, CBC_Hgb, FE, ost_les, hrd_blo_disea, creat, Fib, TP, clair_creat, asth.bone, gamma
PAL	wilaya, SGPT, P, CBC_Hct, CBC_MCV
gly	CRP, body_surf
TP	prot_rate, chron_disea
Ferr	K, creat, BJp, SGOT, GGT, prot24h
LDH	body_surf
FE	clair_creat, alpha, gly, PAL, SGOT

Table 20 – Les arcs Parent/Enfants pour DAG2 (avec BdeuScore)

D'après le tableau 21, il est important de souligner que le nombre de nœuds dans le DAG obtenu grâce à BDeuScore représente le nombre total de variables considérées dans l'ensemble de données (53 nœuds). Certaines variables sont directement connectées entre elles, tandis que d'autres sont reliées de manière indirecte via des variables intermédiaires. De plus, il est à noter qu'il existe des variables qui ne sont pas du tout connectées dans le DAG.

	Hc avec K2Score	Hc avec Bdeu	DAG_Common
Nombre des noeuds	52	53	48
Nombre des arcs	146	135	67
Profondeur du DAG	18	30	10

Table 21 – Quelques critères des pathway DAGs menant à la classe cible

Les mêmes observations peuvent être faites pour le graphe obtenu en utilisant la méthode de scoring "K2Score", à l'exception d'une seule variable éliminée (il y a 52 nœuds). Cette variable correspond au groupe sanguin de chaque patient (*"TypeBlood"*).

D'après les connaissances médicales actuelles, il n'y a aucune interprétation médicale directe reliant le groupe sanguin et le MM. Par conséquent, le groupe sanguin n'est pas considéré comme un critère de diagnostic ou de stadification spécifique pour le MM. Cela signifie que la suppression de cette variable n'affecte pas les autres aspects de nos expériences.

Le DAG1 obtenu par la méthode de scoring K2Score comporte 146 arcs (arêtes), tandis que le deuxième DAG obtenu avec la méthode de scoring BDeuScore ne compte que 135 arcs (voir Table 21).

Les arêtes diffèrent entre ces deux modèles, ce qui indique qu'ils sont structurellement différents et représentent des hypothèses d'indépendance conditionnelle différentes. Cela peut avoir un impact significatif sur le comportement de chaque modèle et sur les résultats de toute tâche d'inférence ou d'apprentissage effectuée à l'aide de ces modèles. Il est donc important d'examiner attentivement les différences dans les arêtes entre les deux modèles et de comprendre leurs implications pour la tâche de modélisation en cours.

Cependant, il est important de souligner qu'il y a **67 arcs communs** entre les deux DAG obtenus (voir Table 21 et figure 25).

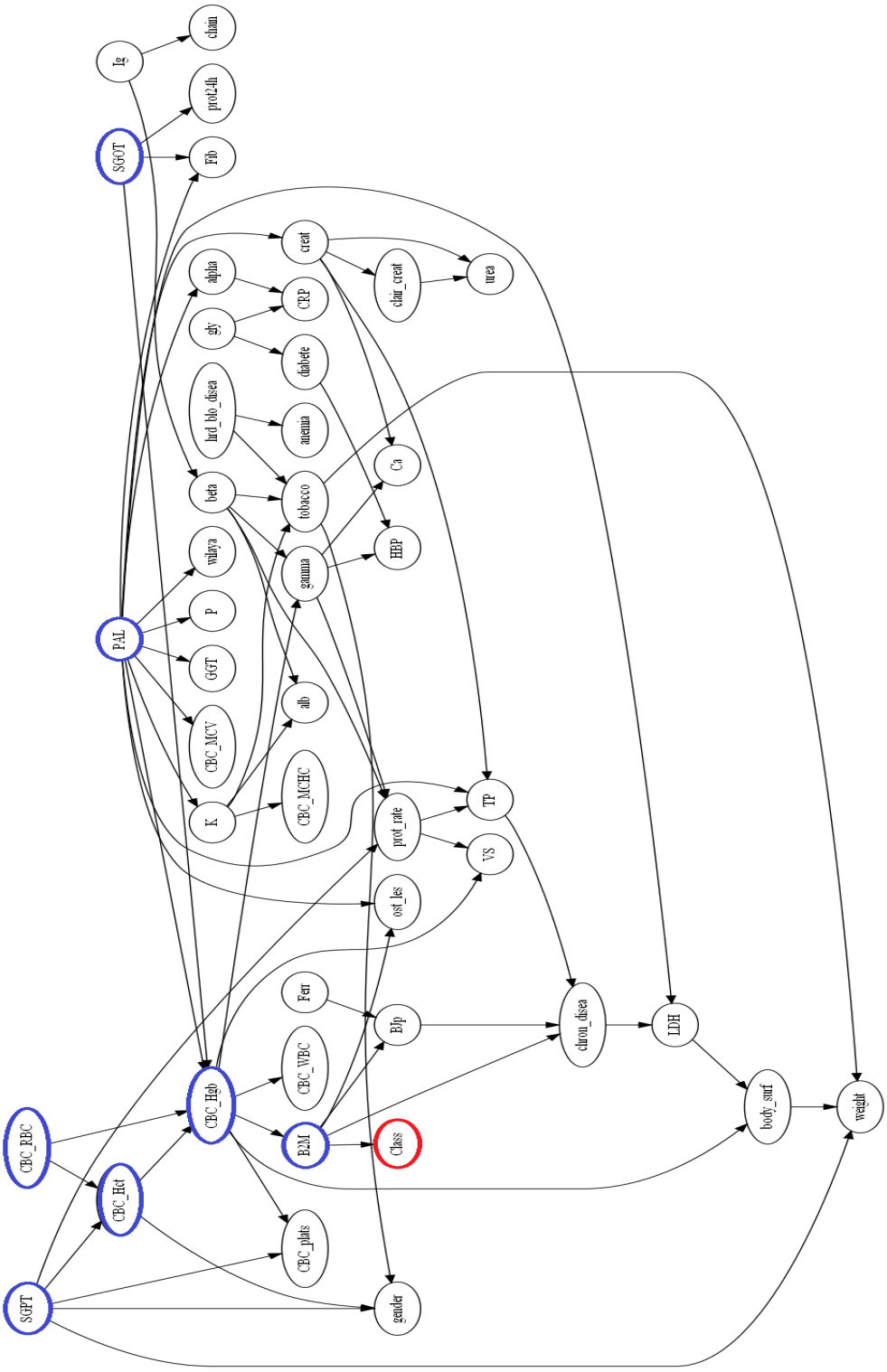


Figure 25 – Pathway DAG commun entre "K2score" et "BDeuScore".

Pour évaluer le degré de similarité entre les deux DAGs obtenus, nous avons utilisé le coefficient de Jaccard, qui est une mesure de similarité structurelle permettant de comparer les ensembles de paires de nœuds adjacents dans les deux DAGs. Le coefficient de Jaccard est défini comme le rapport entre la taille de l'intersection des ensembles et la taille de leur union, et sa valeur varie entre 0 et 1.

Dans notre étude, la valeur du coefficient de Jaccard entre les deux DAGs modélisés par l'algorithme HillClimbSearch avec les méthodes de scoring K2Score et BDeuScore (Figure 23 et 24 respectivement) est de 0.31. Cela indique que 31% des relations entre les variables présentes dans l'un des DAGs sont également présentes dans l'autre DAG (voir Figure 25).

Le DAG commun (voir Figure 25) représente la structure et les dépendances partagées entre les deux DAGs obtenus par l'algorithme HillClimbSearch en utilisant les méthodes de scoring K2Score et BDeuScore. Ces dépendances communes peuvent être exploitées pour estimer les paramètres du réseau bayésien à modéliser. Il est important d'examiner attentivement ces relations et de comprendre leur impact sur les résultats avant de décider de les utiliser.

Dans le contexte du diagnostic du myélome multiple à l'aide d'un graphe acyclique dirigé (DAG), les nœuds racines représentent généralement les variables considérées comme indépendantes et n'ayant pas de nœuds parents. Ils peuvent être considérés comme les points de départ ou les variables initiales du DAG. Cependant, les nœuds racines spécifiques et leurs relations dans un DAG pour le diagnostic du myélome multiple peuvent varier en fonction du contexte, des données disponibles et des méthodes de modélisation de structure utilisées.

En examinant la Figure 27, nous constatons que la variable "age" est considérée comme un nœud racine dans le DAG construit. Cela implique que cette variable est considérée comme indépendante et sert de point de départ pour le pronostic du MM. Dans le contexte du MM, l'âge peut avoir une relation causale avec le pronostic et la stadification de la maladie. Il est bien connu que le MM est plus fréquent chez les personnes âgées, généralement de plus de 65 ans. Cette corrélation peut être due à l'accumulation de lésions génétiques et à d'autres changements cellulaires liés au vieillissement.

Cependant, il convient de souligner que l'âge seul n'est pas suffisant pour diagnostiquer ou stadifier le MM. Dans notre analyse bayésienne des

données cliniques collectées, nous constatons l'existence d'une relation d'influence indirecte entre l'âge et d'autres variables liées au diagnostic. Bien qu'ils ne soient pas directement connectés par une arête dans le DAG, d'autres éléments doivent également être pris en compte dans le processus de diagnostic.

Dans le même sens, la Figure 26 montre que le DAG repose sur six variables en tant que nœuds racine ('hrd_blo_disea', 'Ig', 'SGOT', 'PAL', 'Ferr', 'FE').

Il est important de noter que le MM est généralement considéré comme une maladie sporadique, ce qui signifie qu'elle survient de manière aléatoire et n'est pas directement héritée. Cependant, selon les spécialistes en hématologie ou en génétique, la présence de membres de la famille atteints de maladies du sang peut augmenter le risque de développer un MM. Cette maladie est considérée comme ayant une prédisposition génétique, ce qui signifie qu'elle peut être plus fréquente chez les individus ayant des antécédents familiaux de la maladie.

En ce qui concerne la clairance de la créatinine (*"clair_creat"*), la créatinine (*"creat"*) et l'urée (*"urea"*), en revenant à la Figure 25, nous constatons que la relation parent/enfant peut être interprétée comme suit :

Clairance de la créatinine comme parent : la clairance de la créatinine peut être considérée comme le parent de l'urée. Cela signifie que la clairance de la créatinine peut influencer les valeurs d'urée dans le contexte du myélome multiple. Sur le plan médical, une diminution de la clairance de la créatinine peut indiquer une altération de la fonction rénale, ce qui peut entraîner une accumulation d'urée dans le sang.

Urée comme enfant : l'urée est considérée comme l'enfant de la créatinine et de la clairance de la créatinine. Cela suggère que les niveaux d'urée peuvent être influencés par les niveaux de créatinine et la clairance de la créatinine dans le contexte du myélome multiple.

Dans le diagnostic et la stadification du myélome multiple, la mesure de la créatinine, de la clairance de la créatinine et de l'urée peut fournir des informations sur la fonction rénale et l'accumulation de déchets dans le corps. Des niveaux élevés de créatinine, une clairance de la créatinine réduite et des niveaux élevés d'urée peuvent indiquer une détérioration de la fonction rénale et peuvent être utilisés pour évaluer la gravité de la maladie et guider les décisions de traitement.

Dans le même contexte de la relation parent/enfant, l'algorithme Hil-

lClimbSearch avec le score K2Score a considéré la créatinine comme un prédicteur de la clairance de la créatinine dans le pathway DAG vers la classe cible (voir Figure 26).

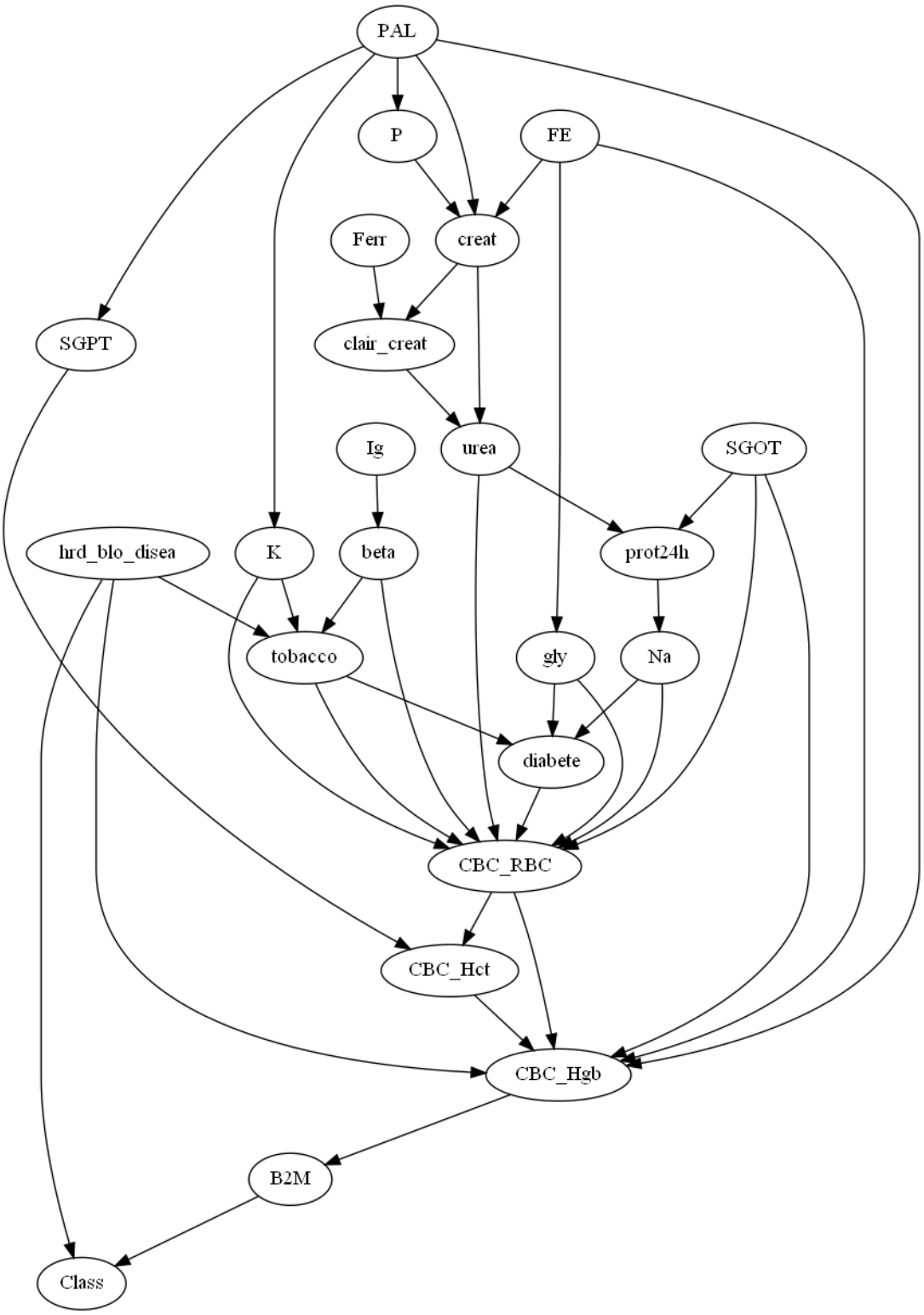


Figure 26 – DAG3 : Pathway DAG vers la classe cible basé sur l'algorithme Hc avec "K2Score".

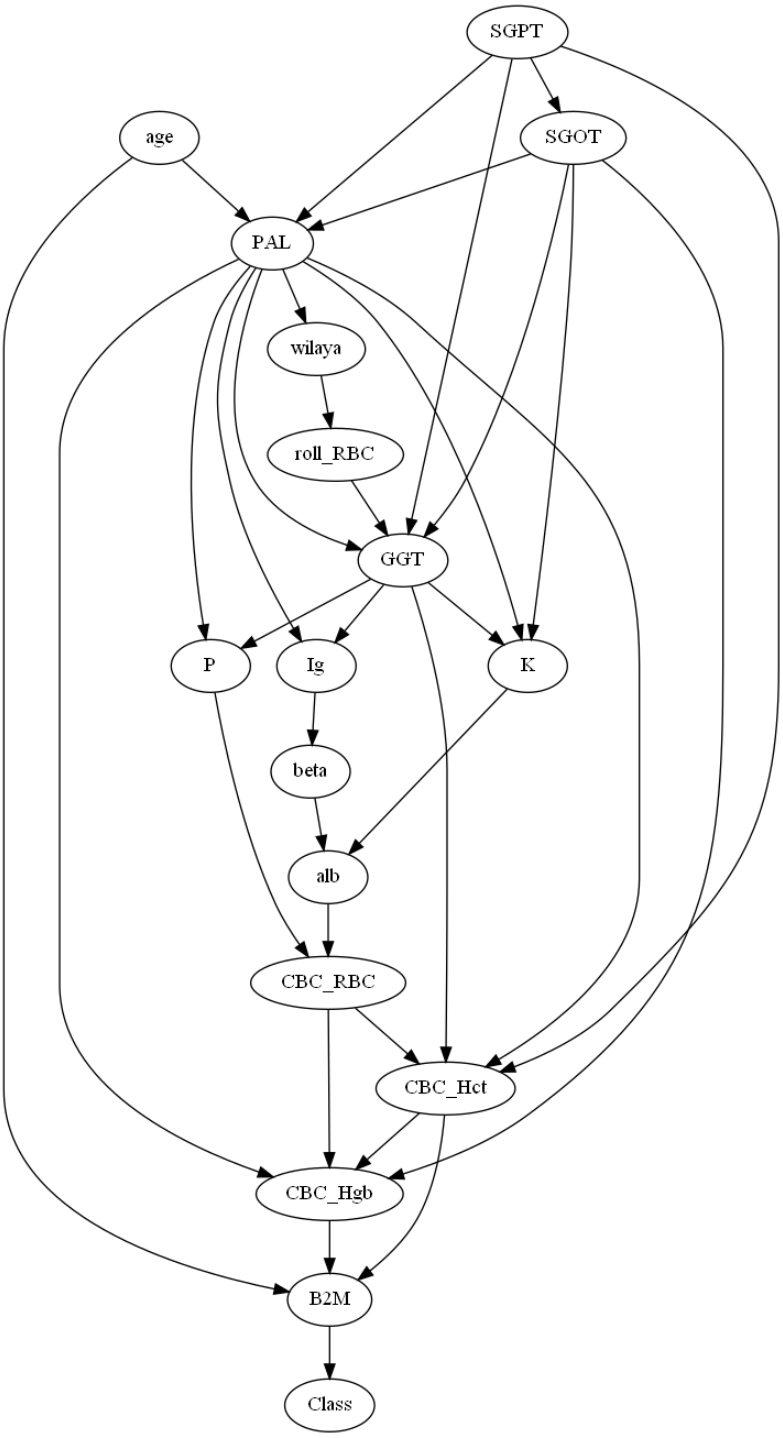


Figure 27 – DAG4 : Pathway DAG vers la classe cible basé sur l'algorithme Hc avec "BDeuScore".

Ensuite, nous procédons à l'apprentissage des paramètres pour estimer de manière empirique les probabilités conditionnelles, afin de représenter les relations de dépendance entre les variables dans le modèle obtenu par l'algorithme Hc.

6.2 Apprentissage des paramètres avec l'algorithme HillClimbSearch

L'apprentissage des paramètres est généralement réalisé à partir d'un ensemble de données d'entraînement dans lequel les valeurs des variables et leurs relations sont connues. L'algorithme d'apprentissage utilise ces données pour estimer les probabilités conditionnelles, en optimisant différents critères tels que la vraisemblance des données observées ou la minimisation de l'erreur de prédiction.

En utilisant l'estimateur du maximum de vraisemblance (MaximumLikelihoodEstimator), nous avons estimé les paramètres de notre modèle construit à partir de nos données d'entraînement. Ensuite, nous avons affiché les distributions de probabilité conditionnelle (CPTs). Les CPTs décrivent la relation de chaque nœud en termes de probabilités conditionnelles ou a priori (Voir Annexe 1).

Nous avons présenté clairement, dans le tableau 22, les règles d'association les plus pertinentes. Notre modèle construit retourne **43 règles conditionnelles** qui expriment les relations de dépendances probabilistes entre les 46 variables.

À partir de ces règles, et en utilisant les probabilités conditionnelles du réseau, il est possible d'extraire des informations sur les relations causales. Le tableau 23 montre que la relation causale entre un taux élevé de B2M et le stade 3 du MM (**P (Class : 4|B2M :3)**) est considérée comme forte. Cependant, la probabilité conditionnelle de cette relation causale est légèrement supérieure à celle de la relation causale entre un taux de B2M compris entre [3.5, 5.5] et le stade 3.

La probabilité conditionnelle d'une relation causale dépend de plusieurs facteurs, tels que la nature de la relation, la présence de variables de confusion et la validité des données utilisées pour évaluer la relation. Lorsqu'il s'agit d'évaluer une relation causale, il est important de prendre en compte les concepts de corrélation et de causalité. Il était également important pour nous de consulter un professionnel de la santé qui nous a fourni des informations précises et a interprété nos résultats spécifiques dans le contexte de l'état général. Le stade 3 du MM correspond à un stade avancé de la maladie où le cancer s'est largement propagé et peut avoir affecté plusieurs organes ou tissus. Le stade du MM est généralement déterminé en fonction de plusieurs facteurs. La bêta-2 microglobuline (B2M) est une protéine souvent utilisée comme marqueur pronostique dans le myélome multiple. Ainsi, des niveaux élevés de B2M dans le sang sont associés à une maladie plus agressive, indiquant une charge tumorale plus importante

et une activité accrue de la maladie.

01	P(Class :4 | B2M :3)
02	P(B2M :3 | CBC_Hgb :2)
03	P(CBC_Hgb :2 | CBC_Hct :3, CBC_RBC :2, PAL :2, SGOT :2)
04	P(SGOT :2)
05	P(PAL :2)
06	P(CBC_RBC :2)
07	P(CBC_Hct :3 | CBC_RBC :2, SGPT :2)
08	P(SGPT :2)
09	P(BJp :2 | B2M :3, Ferr :3)
10	P(ost_les :2 | B2M :3, PAL :2)
11	P(chron_disea :2 | B2M :3, BJp :2, TP :2)
12	P(wilaya :10 | PAL :2)
13	P(alpha :3 | PAL :2)
14	P(CRP :2 | alpha :3, gly :3)
15	P(LDH :2 | PAL :2, chron_disea :2)
16	P(body_surf :2 | CBC_Hgb :2, LDH :2)
17	P(hrd_blo_disea :2)
18	P(anemia :2 | hrd_blo_disea :2)
19	P(K :3 | PAL :2)
20	P(alb :2 | K :3, beta :3)
21	P(CBC_plats :2 | CBC_Hgb :2, SGPT :2)
22	P(HBP :2 | diabete :2, gamma :3)
23	P(TP :2 | PAL :2, creat :2, prot_rate :3)
24	P(CBC_WBC :3 | CBC_Hgb :2)
25	P(CBC_MCV :3 | PAL :2)
26	P(Fib :3 | PAL :2, SGOT :2)
27	P(tobacco :2 | K :3, beta :3, hrd_blo_disea :2)
28	P(weight :4 | SGPT :2, body_surf :2, tobacco :2)
29	P(diabete :2 | gly :3)
30	P(GGT :2 | PAL :2)
31	P(gender :2 | CBC_Hct :3, SGPT :2, tobacco :2)
32	P(CBC_MCHC :3 | K :3)
33	P(creat :2 | PAL :2)
34	P(VS :2 | CBC_Hgb :2, prot_rate :3)
35	P(P :3 | PAL :2)
36	P(Ig :3)
37	P(chain :2 | Ig :3)
38	P(Ca :3 | creat :2, gamma :3)
39	P(clair_creat :2 | creat :2)
40	P(Ferr :3)
41	P(urea :3 | clair_creat :2, creat :2)
42	P(prot24h :2 | SGOT :2)
43	P(gly :3)

Table 22 – Règles de probabilité conditionnelle associées à notre modèle

P(Class :4|B2M :3)			
	B2M(**High**)	B2M(**Normal**)	B2M(**[3.5 ;5.5]**)
Class(ASYM)	0.036	0.071	0.083
Class(stade I)	0.108	0.285	0.083
Class(stade II)	0.072	0.285	0.250
Class(stade III)	0.783	0.357	0.583

Table 23 – Table de probabilité conditionnelle de la variable "Class".

P(B2M :3|CBC_Hgb :2)		
	CBC_Hgb(**anemia**)	CBC_Hgb(**no_anemia**)
B2M(High)	0.847	0.750
B2M(Normal)	0.058	0.173
B2M([3.5 ;5.5])	0.094	0.076

Table 24 – Table de probabilité conditionnelle de la variable "B2M".

P(CBC_Hgb :2 | CBC_Hct :3, CBC_RBC :2, PAL :2, SGOT :2)		
CBC_Hct	...	CBC_Hct(**Normal**)
CBC_RBC	...	CBC_RBC(**Normal**)
PAL	...	PAL(**Normal**)
SGOT	...	SGOT(**Normal**)
CBC_Hgb(anemia)	...	0.0044
CBC_Hgb(no_anemia)	...	0.9955

Table 25 – Table de probabilité conditionnelle de la variable "CBC_Hgb".

P(CBC_Hct :3 | CBC_RBC :2, SGPT :2)			
CBC_RBC	CBC_RBC(**Low**)	...	CBC_RBC(**Normal**)
SGPT	SGPT(**High**)	...	SGPT(**Normal**)
CBC_Hct(High)	0.0	...	0.023
CBC_Hct(Low)	1.0	...	0.441
CBC_Hct(Normal)	0.0	...	0.534

Table 26 – Table de probabilité conditionnelle de la variable "CBC_Hct".

P(CBC_RBC :2)	
CBC_RBC(Low)	0.678
CBC_RBC(Normal)	0.321

Table 27 – Table de probabilité conditionnelle de la variable "CBC_RBC".

P(PAL :2)	
PAL(High)	0.0072
PAL(Low)	0.9927

Table 28 – Table de probabilité conditionnelle de la variable "PAL".

P(SGOT :2)	
SGOT(High)	0.024
SGOT(Normal)	0.975

Table 29 – Table de probabilité conditionnelle de la variable "SGOT".

P(SGPT :2)	
SGPT(High)	0.038
SGPT(Normal)	0.961

Table 30 – Table de probabilité conditionnelle de la variable "SGPT".

6.3 Analyse du réseau bayésien construit

L'analyse d'un réseau bayésien permet d'explorer différents aspects du modèle afin de comprendre ses propriétés et son comportement. Par exemple, il est possible de vérifier les indépendances conditionnelles et les pistes actives (*active trails*) par rapport à une preuve (*evidence*) donnée, ou de déterminer la couverture de Markov (*Markov blanket*) d'un nœud.

La couverture de Markov, également appelée bouclier de Markov, d'un nœud est l'ensemble de ses parents, de ses enfants et de ses co-parents directs. Une fois que l'on connaît les valeurs de ces nœuds, toutes les autres variables du réseau deviennent statistiquement indépendantes du nœud en question [150].

L'analyse de la couverture de Markov permet de comprendre le flux d'informations au sein du réseau bayésien. Les variables parentes fournissent des informations sur la variable cible, tandis que les variables enfants peuvent être influencées par elle. En incluant les co-parents des variables

enfants, nous prenons en compte les influences indirectes sur la variable cible. Dans l'ensemble, l'interprétation de la couverture de Markov nous aide à comprendre les relations causales et les dépendances entre les variables dans un réseau bayésien, à contrôler les effets de confusion et à simplifier les calculs d'inférence pour une prise de décision plus efficace.

D'après le tableau 31, la couverture de Markov du nœud "Class" ne comprend que la variable "B2M", qui est un nœud parent de cette variable cible. Cela signifie que si nous connaissons la valeur de B2M, alors les autres variables du réseau deviennent statistiquement indépendantes de la variable "Class".

Nous constatons également que la couverture de Markov de la variable "B2M" est basée sur 8 variables (voir Table 31), à savoir :

- *CBC_Hgb* : il s'agit de son nœud parent, ce qui signifie que sa valeur peut influencer la probabilité de B2M dans notre modèle construit. Dans le domaine médical, l'hémoglobine et le B2M ne sont pas directement liés en tant que "nœud parent". L'hémoglobine est une protéine qui transporte l'O_2 dans le sang, tandis que le bêta-2-microglobuline (B2M) est une autre protéine qui joue plusieurs rôles, notamment dans le système immunitaire. Cependant, dans le cas du MM, ces deux mesures peuvent être utilisées en combinaison pour fournir des informations plus précises sur la santé d'une personne atteinte du MM et pour aider au diagnostic de ce type de cancer. C'est ce qui peut expliquer la relation causale que nous obtenons dans notre modèle.
- *"BJp", "chron_disea", "Class", "ost_les"* : ce sont des nœuds enfants de B2M. Cela signifie que, en connaissant les valeurs du nœud parent B2M, nous pouvons tirer des conclusions sur les nœuds enfants.
- *"PAL", "TP", "Ferr"* : ce sont des nœuds co-parents qui partagent des relations parentales avec B2M (pour les nœuds "ost_les", "chron_disea" et "BJp" respectivement), mais qui ne sont pas directement connectés à celui-ci.

D'autre part, dans le même contexte lié à notre objectif de recherche abordé dans ce chapitre, et en utilisant les tables 31, 22 et la figure 25, nous pouvons identifier plusieurs relations causales de dépendance et d'indépendance entre les variables (nœuds) du RB obtenu. Afin de mieux comprendre le concept de couverture de Markov de ces nœuds du RB, nous les utiliserons par la suite comme hypothèses dans des problèmes d'inférence statistique.

Nœud cible	Couverture de Markov
'beta'	'prot_rate', 'gamma', 'hrd_blo_disea', 'CBC_Hgb', 'alb', 'tobacco', 'K', 'Ig', 'SGPT'
'gamma'	'Ca', 'prot_rate', 'HBP', 'diabete', 'beta', 'CBC_Hgb', 'creat', 'SGPT'
'prot_rate'	'VS', 'gamma', 'TP', 'PAL', 'beta', 'CBC_Hgb', 'creat', 'SGPT'
'PAL'	'Fib', 'wilaya', 'prot_rate', 'chron_disea', 'P', 'CBC_RBC', 'ost_les', 'CBC_Hct', 'CBC_MCV', 'GGT', 'TP', 'SGOT', 'B2M', 'alpha', 'K', 'LDH', 'CBC_Hgb', 'creat'
'wilaya'	'PAL'
'alpha'	'CRP', 'PAL', 'gly'
'CRP'	'alpha', 'gly'
'CBC_Hgb'	'CBC_WBC', 'CBC_RBC', 'prot_rate', 'body_surf', 'VS', 'CBC_Hct', 'gamma', 'SGOT', 'B2M', 'PAL', 'beta', 'LDH', 'CBC_plats', 'SGPT'
'B2M'	'BJp', 'chron_disea', 'Class', 'ost_les', 'CBC_Hgb', 'PAL', 'TP', 'Ferr'
'LDH'	'body_surf', 'CBC_Hgb', 'chron_disea', 'PAL'
'body_surf'	'weight', 'tobacco', 'LDH', 'CBC_Hgb', 'SGPT'
'hrd_blo_disea'	'tobacco', 'anemia', 'K', 'beta'
'anemia'	'hrd_blo_disea'
'K'	'hrd_blo_disea', 'alb', 'tobacco', 'CBC_MCHC', 'PAL', 'beta'
'alb'	'K', 'beta'
'CBC_plats'	'CBC_Hgb', 'SGPT'
'HBP'	'diabete', 'gamma'
'TP'	'BJp', 'chron_disea', 'prot_rate', 'B2M', 'PAL', 'creat'
'CBC_WBC'	'CBC_Hgb'
'CBC_MCV'	'PAL'
'SGOT'	'Fib', 'CBC_RBC', 'prot24h', 'CBC_Hct', 'PAL', 'CBC_Hgb'
'Fib'	'SGOT', 'PAL'
'tobacco'	'weight', 'body_surf', 'CBC_Hct', 'hrd_blo_disea', 'gender', 'beta', 'K', 'SGPT'
'weight'	'body_surf', 'tobacco', 'SGPT'
'diabete'	'HBP', 'gly', 'gamma'

Table 31 – la couverture de Markov de tous les nœuds du réseau bayésien.

suite de Table 31

Nœud cible	Couverture de Markov
'BJp'	'Ferr', 'TP', 'B2M', 'chron_disea'
'GGT'	'PAL'
'CBC_RBC'	'CBC_Hct', 'SGOT', 'PAL', 'CBC_Hgb', 'SGPT'
'chron_disea'	'BJp', 'B2M', 'PAL', 'LDH', 'TP'
'SGPT'	'weight', 'prot_rate', 'CBC_RBC', 'body_surf', 'CBC_Hct', 'gamma', 'tobacco', 'gender', 'beta', 'CBC_Hgb', 'CBC_plats'
'gender'	'tobacco', 'CBC_Hct', 'SGPT'
'CBC_Hct'	'CBC_RBC', 'SGOT', 'tobacco', 'PAL', 'gender', 'CBC_Hgb', 'SGPT'
'CBC_MCHC'	'K'
'creat'	'Ca', 'prot_rate', 'urea', 'gamma', 'clair_creat', 'PAL', 'TP'
'VS'	'CBC_Hgb', 'prot_rate'
'P'	'PAL'
'Ig'	'beta', 'chain'
'chain'	'Ig'
'Ca'	'creat', 'gamma'
'clair_creat'	'urea', 'creat'
'Class'	'B2M'
'gly'	'diabete', 'alpha', 'CRP'
'Ferr'	'BJp', 'B2M'
'urea'	'clair_creat', 'creat'
'prot24h'	'SGOT'
'ost_les'	'B2M', 'PAL'

6.4 Inférences

Grâce à la propriété de d-séparation induite par la couverture de Markov, il est possible d'effectuer des inférences sur un nœud de manière efficace en ne considérant que les valeurs de sa couverture de Markov. Cela réduit la complexité de l'inférence et permet des calculs plus rapides.

Dans un réseau bayésien, les inférences sont réalisées en utilisant les paramètres appris et les informations disponibles dans le réseau. Pour effectuer ces inférences et faire des prédictions, nous nous appuyons sur un ensemble de nouvelles évidences ou requêtes extraites à partir de notre modèle bayésien construit.

Les résultats présentés dans le tableau 32 fournissent des informations cruciales sur les meilleures inférences réalisées dans notre réseau bayésien construit en utilisant la base d'apprentissage avec des évidences spécifiques. Nous avons analysé les variables d'intérêt, leurs meilleures valeurs et les probabilités conditionnelles correspondantes en fonction des évidences fournies.

Ces résultats obtenus mettent en évidence plusieurs variables médicales d'intérêt. Tout d'abord, nous remarquons que la variable "B2M" est indiquée comme étant élevée avec une probabilité de 0,8108. Cette variable a une relation parentale directe avec la variable "Class" qui représente les stades du MM. Plus précisément, notre modèle bayésien suggère une relation causale entre un taux élevé de B2M et le stade III du MM (P(Class|B2M)) avec une probabilité conditionnelle d'environ 0,78 (voir Table 23). Cette inférence relationnelle est cohérente avec la littérature existante, notamment le travail de Salih et al. [151], où les auteurs ont indiqué qu'un niveau élevé de B2M est souvent observé chez les patients atteints de MM. Cette étude a probablement utilisé le système ISS [40] pour classer les patients en fonction des niveaux de B2M et d'albumine sérique (voir Table 8). Un stade avancé correspond généralement à une charge tumorale plus importante et à une progression de la maladie.

Le taux de bêta 2-microglobuline (B2M) et le taux d'hémoglobine (CBC_Hgb) sont deux facteurs biologiques utilisés pour évaluer le profil de santé d'une personne atteinte de MM. Il n'y a pas de relation directe entre le taux de B2M et CBC_Hgb. Cependant, dans le contexte du diagnostic du MM, l'évaluation conjointe du taux de B2M et du CBC_Hgb peut fournir des informations supplémentaires pour étayer le diagnostic. Un niveau élevé de B2M combiné à une diminution du CBC_Hgb peut être

un indicateur de la présence et de la progression du myélome multiple. Ces connaissances médicales renforcent la confiance envers notre modèle bayésien qui a identifié une forte relation entre un taux élevé de B2M et un faible taux de CBC_Hgb avec une probabilité conditionnelle de 0,98 (voir Table 24).

De plus, il existe une corrélation bien établie entre un niveau élevé de B2M et la présence de lésions osseuses (ost_les) dans le contexte du myélome multiple (MM), avec une probabilité conditionnelle d'environ 0,91. Plusieurs études ont montré que des niveaux élevés de B2M peuvent indiquer une charge tumorale plus importante, ce qui peut conduire à une destruction osseuse accrue et à des lésions osseuses plus importantes visibles sur l'imagerie médicale. Ces connaissances renforcent l'évaluation des inférences tirées de notre modèle bayésien construit.

Selon nos résultats, nous pouvons également noter la coexistence d'une leucocytose (taux élevé de CBC_WBC) et d'une anémie (taux bas de CBC_Hgb), avec une probabilité conditionnelle de 0,435. Il est important de noter que la leucocytose et l'anémie ne sont pas spécifiques au MM et peuvent être présentes dans d'autres conditions médicales. Nous savons que le MM se caractérise par une prolifération anormale des plasmocytes dans la moelle osseuse, ce qui peut entraîner des altérations dans la production normale de différentes cellules sanguines.

En ce qui concerne la variable PAL (Phosphatase alcaline), nous avons observé une probabilité conditionnelle de "0,99" pour un taux bas de PAL. Cette variable est une enzyme présente dans différentes parties du corps, notamment les os, le foie, les reins, les intestins et le placenta chez les femmes enceintes. Cela pourrait être une réponse médicale aux relations de causalité inférées entre les variables "Fib" (Fibrinogène) et "ost_les" (lésions osseuses) d'une part, et la variable "PAL" d'autre part.

En général, le niveau de PAL dans le sang n'est pas directement lié au myélome multiple. Cependant, il convient de noter qu'il peut y avoir une relation indirecte entre eux. Dans certains cas de myélome multiple, les cellules plasmatiques cancéreuses peuvent perturber le fonctionnement normal des cellules osseuses, ce qui peut entraîner une augmentation de PAL due à une résorption osseuse accrue. Cela n'exclut cependant pas la possibilité d'un myélome multiple en cas de faible taux de PAL dans le sang.

Variable d'intérêt	**Valeur**	**Evidence**	φ
alb	>=35	K= 'Low' ; beta= 'Low'	0,710
alpha	Low	PAL= 'Low'	0,750
anemia	No	hrd_blo_disea= 'No'	0,813
B2M	High	CBC_Hgb= 'anemia'	0,847
BJp	Positive	CBC_Hgb= 'anemia' ; Ferr= 'Normal'	0,970
body_surf	>=1,71	CBC_Hgb= 'anemia ; LDH= '=<400'	0,657
Ca	Normal	creat= '<=20' ; gamma= 'Low'	0,750
CBC_Hct	Low	CBC_RBC= 'Low' ; SGPT= 'Normal'	0,812
CBC_Hgb	anemia	CBC_Hct= 'Low' ; CBC_RBC= 'Low' ; PAL= 'Low' ; SGOT= 'Normal'	0,938
CBC_MCHC	Normal	K= 'Low'	0,734
CBC_MCV	normo	PAL= 'Low'	0,867
CBC_plats	Normal	CBC_Hgb= 'anemia' ; 'SGPT= 'Normal'	0,761
CBC_WBC	High	CBC_Hgb= 'anemia'	0,435
chain	Kappa	Ig= 'IgG'	0,901
chron_disea	No	B2M= 'High' ; BJp= 'Positive' ; TP= 'Normal'	0,797
clair_creat	no_prob	creat= '<=20'	0,845
Class	Stade III	B2M= 'High'	0,783
creat	<=20	PAL= 'Low'	0,808
CRP	Negative	alpha= 'Low' ; gly= 'Normal'	0,728
diabete	No	gly= 'Normal'	0,886
Fib	High	PAL= 'Low' ; SGOT= 'Normal'	0,711
gender	Female	CBC_Hct= 'Low' ; SGPT= 'Normal' ; tobacco= 'No'	0,589
GGT	Normal	PAL= 'Low'	0,992
HBP	No	diabete= 'No' ; gamma= 'Low'	0,777
K	Low	PAL= 'Low'	0,941
LDH	=<400	PAL= 'Low' ; chron_disea= 'No'	0,886
ost_les	Yes	B2M= 'High' ; PAL= 'Low'	0,909
P	hypo	PAL= 'Low'	0,904
prot24h	Positive	SGOT= 'Normal'	0,860
tobacco	No	K= 'Low' ; beta= 'Low' ; hrd_blo_disea= 'No'	0,909
TP	Normal	PAL= 'Low' ; creat= '<=20' ; prot_rate= 'Low'	0,952
urea	Normal	clair_creat= 'no_prob' ; creat= '<=20'	0,784
VS	High	CBC_Hgb= 'anemia' ; prot_rate= 'Low'	0,561
weight	healthy	SGPT= 'Normal' ; body_surf= '>=1.71' ; tobacco= 'No'	0,487
wilaya	Tlemcen	PAL= 'Low'	0,720

Table 32 – Meilleures inférences avec variables d'intérêt et évidences

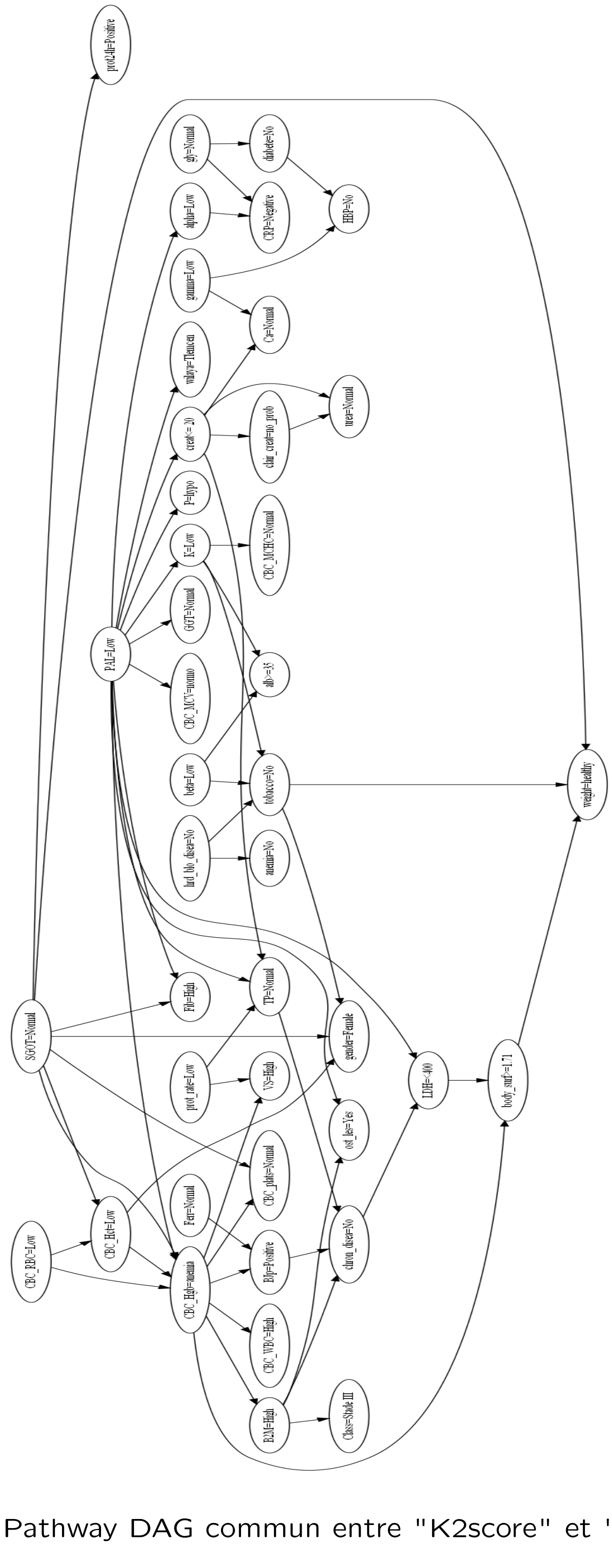


Figure 28 – Pathway DAG commun entre "K2score" et "BDeuScore".

Il convient également de noter que, bien que les taux élevés de fibrinogène ne soient pas spécifiques au MM et puissent également être présents dans d'autres conditions inflammatoires ou cancéreuses, ils peuvent être associés à plusieurs mécanismes pathologiques dans le cas du MM, tels que la réaction inflammatoire, le risque de thrombose et l'interaction avec les cellules tumorales, entre autres.

En ce qui concerne la variable d'intérêt "clair_creat" (clairance de la créatinine), nous avons observé une probabilité conditionnelle de 0.845 lorsque l'évidence "creat" (taux de créatinine) était fixée à "<=20 mg/L". Cela suggère qu'une concentration basse de créatinine est généralement associée à une clairance normale de la créatinine. Il est important de mentionner que le MM peut avoir des effets sur la fonction rénale en raison de l'accumulation de protéines anormales. Cependant, une faible concentration de créatinine peut être associée à une clairance normale chez un patient atteint du MM, et dans ce cas, elle ne peut pas être considérée comme le seul facteur d'évaluation pour le diagnostic de cette maladie.

Les résultats de l'analyse du réseau bayésien fournissent également des informations précieuses sur la relation entre les niveaux d'albumine (alb) et de protéine bêta-globuline (beta) chez les patients atteints de myélome multiple (MM). En examinant les probabilités inférées et les dépendances conditionnelles, nous observons une forte relation entre ces deux variables. Lorsque les taux de protéines bêta-globulines diminuent, les taux d'albumine ont tendance à augmenter.

La diminution des taux de protéine bêta-globuline peut être un signe de progression de la maladie ou d'un défaut de synthèse des protéines, ce qui entraîne une augmentation des taux d'albumine en tant que mécanisme compensatoire. Cela suggère qu'une relation inverse entre ces biomarqueurs peut être envisagée, c'est-à-dire que des niveaux plus faibles de protéine bêta-globuline pourraient être associés à des niveaux plus élevés d'albumine chez les patients atteints de MM.

La relation entre l'anémie et l'existence de personnes ayant des antécédents de maladies du sang dans la famille est un aspect crucial à explorer dans le domaine médical. Notre étude fournit des informations importantes sur cette relation. Notre modèle bayésien met en évidence une association significative entre l'anémie (anemia) et l'existence de personnes ayant des antécédents familiaux de maladies du sang (hrd_blo_disea) chez les patients atteints de MM, avec une probabilité conditionnelle d'environ 0,66.

Cette constatation suggère un lien potentiel entre les deux conditions et met en évidence l'importance de l'hérédité dans le développement des maladies du sang. L'identification de cette relation dans le cas du myélome multiple revêt une importance clinique. Ces informations peuvent aider les professionnels de la santé à évaluer le risque de développement de la maladie chez les personnes présentant une anémie et des antécédents familiaux de maladies du sang. De plus, cela peut influencer les décisions de dépistage et de suivi chez les membres de la famille présentant un risque.

Les inférences effectuées dans notre réseau bayésien en utilisant les évidences spécifiées ont fourni des informations intéressantes sur les variables d'intérêt. Les probabilités conditionnelles obtenues nous permettent de comprendre les relations entre les variables d'intérêt et les évidences fournies, contribuant ainsi à une meilleure compréhension du domaine étudié. Il est important de noter que ces inférences sont basées sur les paramètres et les hypothèses du réseau bayésien utilisé.

Ces résultats illustrent l'importance de prendre en compte les évidences lors de l'inférence dans un réseau bayésien. Il est nécessaire de considérer ces résultats dans le contexte plus large de la recherche et de les interpréter avec prudence.

6.5 Évaluation des performances

L'évaluation des performances de notre modèle bayésien a été réalisée en utilisant une base de test distincte, qui contient 30

La précision est une mesure d'évaluation qui quantifie la proportion de prédictions positives correctes parmi toutes les prédictions positives faites par un modèle. Elle est utile lorsque les faux positifs sont coûteux ou indésirables. Elle est calculée à l'aide de la formule :

$$Prcision = V_{positifs}/(V_{positifs} + F_{positifs})$$

Le F1-score est une autre métrique d'évaluation qui combine à la fois la précision et le rappel pour évaluer les performances d'un modèle de classification. Cette mesure est particulièrement utile lorsque les classes sont déséquilibrées ou lorsque les faux positifs et les faux négatifs ont des conséquences différentes. Le F-score est calculé à l'aide de la formule :

$$F - score = 2 * (Prcision * Rappel)/(Prcision + Rappel)$$

D'après le tableau 6.5, qui présente les scores de précision et de F1-score pour différents modèles de réseau bayésien, nous pouvons tirer plusieurs

observations importantes.

Modèle de RB	Précision	F1-score
Hc avec K2Score	72.26%	83.89%
Hc avec BdeuScore	73%	84%
modèle commun optimisé	72.26%	83.89%

Table 33 – Scores de précision et F1-score

Tout d'abord, nous avons utilisé deux méthodes de scoring différentes, à savoir "BdeuScore" et "K2Score", pour construire les modèles bayésiens avec l'algorithme Hc. Les performances des modèles construits avec ces deux méthodes sont comparables en termes de précision et de F1-score. Le modèle construit avec la méthode "BdeuScore" a une précision de 73% et un F1-score de 84%, tandis que le modèle construit avec la méthode "K2Score" a une précision légèrement inférieure de 72,26% et un F1-score de 83,89%. Ces résultats suggèrent que les deux méthodes de scoring sont efficaces pour la construction de modèles de réseau bayésien dans le contexte de notre étude sur le myélome multiple.

De plus, nous avons introduit un modèle commun optimisé qui présente des performances similaires aux deux modèles précédents. Ce modèle a une précision de 72,26% et un F1-score de 83,89%. Bien que les performances de ce modèle soient légèrement inférieures à celles du modèle construit avec la méthode "BdeuScore", la différence est minime. Par conséquent, le modèle commun optimisé peut être considéré comme une alternative valide pour prédire les stades du myélome multiple.

En résumé, les résultats de notre évaluation des performances indiquent que les modèles de réseau bayésien construits avec les méthodes "BdeuScore" et "K2Score", ainsi que le modèle commun optimisé, ont des performances comparables en termes de précision et de F1-score. Ces résultats suggèrent que ces modèles sont capables de prédire avec précision les stades du myélome multiple, ce qui peut être utile dans le cadre du diagnostic et de la prise en charge de cette maladie complexe.

7 Conclusion

Dans ce chapitre, nous avons entrepris d'explorer et d'analyser des modèles graphiques probabilistes en utilisant l'algorithme du réseau bayésien "Hc". Notre objectif principal était d'optimiser l'inférence clinique et de mieux comprendre les relations causales entre les variables de l'ensemble de données du myélome multiple (MM). Pour ce faire, nous avons déve-

loppé différents modèles en utilisant l'algorithme Hc avec deux méthodes de scoring différentes, à savoir le K2score et le BdeuScore.

L'évaluation des performances de ces modèles a révélé des résultats prometteurs. En comparant les modèles bayésiens construits avec les deux méthodes de scoring, nous avons pu produire une représentation graphique fiable et transparente du MM. Notre modèle optimal final, qui contient 67 arcs et 46 relations causales pertinentes, nous a permis de mieux comprendre les liens entre les paramètres influençant le diagnostic du MM.

Grâce à cette analyse approfondie, nous avons pu identifier les variables les plus significatives et les relations causales clés liées au MM. Ces informations sont cruciales pour améliorer la compréhension de la maladie et pour développer de nouveaux scénarios prédictifs. En comprenant les facteurs qui contribuent au diagnostic du MM, nous pouvons mieux anticiper les résultats et optimiser les décisions cliniques.

Ces résultats ouvrent de nouvelles perspectives de recherche et offrent des opportunités pour une prise en charge améliorée des patients atteints de MM. En utilisant notre modèle optimal, les cliniciens pourront prendre des décisions plus éclairées et personnalisées, ce qui peut conduire à un diagnostic plus précoce, à des interventions plus ciblées et à de meilleurs résultats pour les patients.

En conclusion, l'utilisation de modèles graphiques probabilistes basés sur le réseau bayésien "Hc" a permis d'obtenir une meilleure compréhension des relations causales dans le contexte du MM. Ces résultats fournissent une base solide pour des recherches futures visant à développer des outils de diagnostic plus précis et des approches thérapeutiques personnalisées pour les patients atteints de cette maladie complexe.

Conclusion Générale

L'objectif principal de notre sujet de recherche pour notre thèse de doctorat était de sélectionner des variables à partir des données cliniques du myélome multiple (MM) et d'analyser leurs associations. Le diagnostic de cette maladie est complexe car certains symptômes et résultats de tests peuvent également se produire dans d'autres maladies.

Le premier défi de notre travail de recherche a été de collecter des données biologiques auprès de différents centres hospitalo-universitaires. Grâce à une collaboration entre la Faculté de Technologie de l'Université de Tlemcen et le CHU de Tlemcen, nous avons bénéficié d'un stage pratique dans le service d'hématologie du Centre de Lutte Contre le Cancer (CLCC) de Tlemcen. L'objectif était de collecter des données sur le cancer du myélome multiple afin de les utiliser en pratique clinique et d'analyser leur impact sur le suivi des patients. Nos contacts avec des biologistes et des médecins spécialistes en hématologie nous ont permis de comprendre les différentes notions de base sur le diagnostic du MM qui représente, selon eux, un terrain de recherche crucial.

Le myélome multiple est un cancer du sang malin très fréquent. Il s'agit d'un néoplasme des plasmocytes dans la moelle osseuse qui s'accompagne d'une série complexe de manifestations cliniques. Ces plasmocytes font partie du système immunitaire et jouent un rôle dans la production d'anticorps qui aident à combattre les infections. Nous avons donc commencé, dans le premier chapitre, par une présentation précieuse des connaissances très importantes sur le myélome multiple, en particulier son épidémiologie, sa physiopathologie, sa pathogénie, tous les signes cliniques et toutes les formes de MM, ainsi que le processus diagnostique et les tests/examens effectués. L'objectif principal était de comprendre l'aspect médical de cette

maladie avant de passer à nos expérimentations.

Le diagnostic du MM est parfois très complexe, les symptômes n'étant généralement pas détectés aux premiers stades. Les patients doivent subir une série d'examens/tests fréquents et répétés, ce qui rend le processus de diagnostic et de stadification long et stressant pour les patients. Par conséquent, l'objectif des modèles que nous avons proposés était de s'appuyer sur l'idée d'une prédiction rétrospective des tests les plus importants à effectuer. Cette initiative permettrait de réduire le nombre de tests et le coût global, ce qui constitue un problème majeur pour tous les patients.

Dans le deuxième chapitre de notre travail, nous avons essayé d'étudier le classement et la sélection des caractéristiques de diagnostic et de stadification du MM en utilisant notre ensemble de données collectées [4] au CLCC-CHU de Tlemcen en Algérie. Cet ensemble de données est déséquilibré car la plupart des patients sont au stade III. Pour compenser cette distribution déséquilibrée des classes, nous avons utilisé SMOTE comme méthode de rééchantillonnage.

Pour mesurer la corrélation entre les variables d'entrée et la classe cible, nous avons développé un modèle basé sur des méthodes de sélection de variables fondées sur l'approche du filtre. Les résultats ont montré que la performance de la classification était améliorée en utilisant les sous-ensembles de caractéristiques pertinents retenus par les méthodes de sélection, plutôt que l'ensemble complet de caractéristiques. La technique de sélection CFS a donné de bons résultats lorsqu'elle a été utilisée avant le classifieur "Arbre de décision" (C4.5), tandis que FCBF s'est avéré plus pratique en raison de sa robustesse et de sa capacité à éliminer les caractéristiques non pertinentes.

Dans la suite de notre travail, nous avons abordé les méthodes d'ensemble basées sur les arbres de décision pour estimer l'importance des variables. Nous avons procédé à un réglage approfondi des hyperparamètres pour les algorithmes d'ensemble proposés en utilisant la technique GridSearchCV, et nous avons obtenu des résultats très prometteurs et encourageants. LightGBM a fait preuve de la plus grande rapidité, tandis que Random Forest a obtenu une précision moyenne de plus de 97% et que XGBoost a obtenu le meilleur classement pour les caractéristiques considérées comme les facteurs de pronostic les plus importants.

Dans le dernier chapitre, nous avons proposé une intégration de méthodes d'apprentissage automatique basées sur les réseaux bayésiens pour

découvrir des relations causales intéressantes entre les variables de notre base de données. Nous avons utilisé l'algorithme "HillClimbSearch" qui a récemment été largement utilisé pour la découverte de connaissances dans les bases de données à l'aide de certaines mesures d'intérêt (méthode de scoring). Dans le but d'optimiser l'inférence clinique et de mieux comprendre les relations causales entre les variables du jeu de données MM, nous avons tenté, dans la phase d'apprentissage structurel, d'explorer et d'analyser des modèles graphiques probabilistes en utilisant cet algorithme de réseau bayésien, mais en changeant la méthode de scoring (K2score et BdeuScore). Afin de produire une représentation graphique fiable et transparente, une comparaison des modèles bayésiens obtenus a permis de mettre en évidence un modèle optimal contenant 67 arcs. Ce modèle explore 46 relations causales pertinentes dans la phase d'apprentissage des paramètres. Cela nous a permis de mieux comprendre les relations entre les paramètres influençant le diagnostic du MM et offre la possibilité de prédire de nouveaux scénarios.

Perspectives

En ce qui concerne les perspectives à court terme, il serait bénéfique de développer des collaborations entre différents centres hospitalo-universitaires en Algérie pour collecter des données cliniques sur le myélome multiple. L'utilisation de l'intelligence artificielle et de l'analyse de données massives (big data) permettrait d'exploiter ces données et de découvrir de nouvelles connaissances, des associations et des schémas cliniques pertinents. Cette approche favoriserait également les échanges d'expertise et la mise en place de recommandations pour améliorer la pratique clinique.

Une autre perspective à court terme serait d'inclure des sujets sains dans les études sur le myélome multiple. Cela permettrait de comparer les caractéristiques et les résultats des patients atteints de myélome multiple avec ceux d'un groupe témoin, ce qui pourrait aider à identifier les facteurs de risque et à mettre en évidence des schémas cliniques spécifiques associés à la maladie.

À plus long terme, l'intégration des données omiques serait une perspective très prometteuse. L'analyse du génome, du transcriptome, du protéome et de l'épigénome des patients atteints de myélome multiple permettrait de mieux comprendre les mécanismes sous-jacents de la maladie. Cela pourrait conduire à l'identification de signatures moléculaires spécifiques, de nouvelles cibles thérapeutiques et à une personnalisation accrue

des schémas thérapeutiques.

En résumé, l'avenir de la recherche sur la sélection des variables et l'analyse des données cliniques du myélome multiple en Algérie est prometteur. L'intégration de nouvelles sources de données, l'utilisation de l'intelligence artificielle et l'exploration des données omiques ouvrent des perspectives passionnantes pour une meilleure compréhension de la maladie et une amélioration des soins aux patients. Ces avancées contribueront à renforcer les capacités de diagnostic en Algérie, à améliorer la prise en charge des patients atteints de myélome multiple et à réduire les retards de diagnostic.

Contributions scientifiques

Cette thèse a généré de nombreuses contributions méthodologiques qui ont été intégrées de manière transversale dans différents chapitres de la thèse, reflétant l'interconnexion des différentes parties du travail de recherche. Plusieurs travaux en collaboration avec des collègues ont influencé également les réflexions directes et indirectes liées à la thèse, apportant ainsi une valeur ajoutée à la recherche.

Journaux indexés

- **R. GUILAL**, N. SETTOUTI, G. MARTINEZ MUNOZ and MA. CHIKH. *Feature importance analysis for highly imbalanced multiple myeloma staging*. In press the International Journal of Medical Engineering and Informatics (IJMEI). January 2022. DOI : 10.1504/IJMEI.2022.10046878.

Base de données

- **R. GUILAL**, A. BENDAHMEN, N. SETTOUTI, N. MESLI and MA. CHIKH. *Multiple Myeloma Dataset (MM-dataset)*. Mendeley Data, v1. 2019. `http://dx.doi.org/10.17632/7wpcv7kp6f.1`

Communications Internationales

- **R. GUILAL**, N. SETTOUTI, AF. BENDAHMENE, A. BENAZZOUZ, MA. CHIKH and N. MESLI. *Selection of prognostic features for multiple myeloma diagnosis in the region of Tlemcen, Algeria.*. In Proceedings of the 1st International Conference on Intelligent Systems and

Pattern Recognition (ISPR '20). Association for Computing Machinery, New York, NY, USA, 17–21. DOI : 10.1145/3432867.3432899.

- **R. GUILAL**, AF. BENDAHMANE, N. SETTOUTI, A. BENAZZOUZ and MA. CHIKH, "Clinical and paraclinical factors selection for multiple myeloma diagnosis," 2019 International Conference on Advanced Electrical Engineering (ICAEE), Algiers, Algeria, 2019, pp. 1-6, doi : 10.1109/ICAEE47123.2019.9014837.
- K. BOUKHOBZA, **R. GUILAL**, M. SAIDI, N. SETTOUTI. *A Variable Importance Measure for Multiple Myeloma Staging Disease Prediction under a Budget*. Conférence Internationale sur les Mathématiques Financières, Outils et applications (MFOA'2019). 28-29 Octobre 2019. Bejaia. Algérie.
- **R. GUILAL**, N. SETTOUTI, A.F. BENDAHMANE, A. BENAZZOUZ, MA. CHIKH and N. MESLI. *Prognostic factors selection for multiple myeloma diagnosis in wilaya of Tlemcen*. In The First International Conference on Pattern Analysis and Recognition (ICPAR 2019) October 22- 24, 2019 in Tebessa, Algeria. (**Best Paper Award**).
- **R. GUILAL**, N. SETTOUTI and MA. CHIKH. *Analyse de corrélation des facteurs pronostic pour le diagnostic du Myélome Multiple dans la Wilaya de Tlemcen*. A la 9ème édition du colloque Tendances dans les Applications Mathématiques en Tunisie Algérie Maroc, 23-27 Février 2019, Tlemcen, Algérie.

Communications Nationales

- **R. GUILAL**, N. SETTOUTI and MA. CHIKH. *Extraction and recognition of clinical and para-clinical factors for multiple myeloma diagnosis : Fast Correlation-based Filter Solution*. Journée Doctorale GBM2019, Tlemcen, Algérie.
- **R. GUILAL**, N. SETTOUTI and MA. CHIKH. La détection des facteurs pronostique du Myélome Multiple. Journée d'étude : Vers des méthodes innovantes pour le diagnostic précoce dans la recherche du cancer. 10 Décembre 2018, Tlemcen, Algérie.

Rapport de recherche

- **R. GUILAL**, N. SETTOUTI and MA. CHIKH. *Myélome Multiple : étude descriptive des données en pratique clinique*. Rapport de Recherche de Doctorat en Génie Biomédical 2020.
  `https://hal.archives-ouvertes.fr/hal-02435378`.

Co-encadrement de projet de fin d'études de Master

– Sujet de fin d'études de l'étudiante "Khadidja Boukhobza" Master II Génie Biomédical option IBM (Informatique Biomédicale) intitulé : *A Variable Importance Measure for a Cost Sensitive Random Forest Predition on a Budget.*. 2019

Projet de recherche

– Système d'aide à la décision médicale pour la gestion du schéma thérapeutique adapté dans le traitement du myélome multiple. PRFU N° C00L07UN130120200001. 2020-2024.

Bibliographie

[1] S Vincent Rajkumar, Meletios A Dimopoulos, Antonio Palumbo, Joan Blade, Giampaolo Merlini, María-Victoria Mateos, Shaji Kumar, Jens Hillengass, Efstathios Kastritis, Paul Richardson, et al., "International myeloma working group updated criteria for the diagnosis of multiple myeloma," *The lancet oncology*, vol. 15, no. 12, pp. e538–e548, 2014.

[2] Leon Furchtgott, Arnold Bolomsky, Fred Gruber, Mehmet Kemal Samur, Jonathan J Keats, Jennifer Yesil, Kathrin Stangelberger, Michel Attal, Philippe Moreau, Hervé Avet-Loiseau, et al., "Multiple myeloma drivers of high risk and response to stem cell transplantation identified by causal machine learning : Out-of-cohort and experimental validation," *Blood*, vol. 130, pp. 3029, 2017.

[3] Khadidja Madini, Amal Nasri, Sidahmed Bentrari, Mohamed El-Bachir Benkhaldia, and Oussama Youcefi, *Myélome Multiple : Réponses thérapeutique au service d'hématologie CHU Tlemcen entre 2014 et 2016*, Ph.D. thesis, Université de Tlemcen-Abou Bekr Belkaid, 2016.

[4] Rima Guilal, Ahmed Fouad Bendahmane, Nesma settouti, Mohammed Amine Chikh, and Naima Mesli, "Multiple myeloma dataset (mm-dataset)," Mendeley Data, v1. http ://dx.doi.org/10.17632/7wpcv7kp6f.1, December 2019.

[5] Robert A Kyle and David P Steensma, "History of multiple myeloma," in *Multiple myeloma*, pp. 3–23. Springer, 2011.

[6] Ahmed Fouad Bendahmane, *Bortezomib "bihebdomadaire" versus "hebdomadaire" dans le traitement du myélome multiple en premiére ligne*, Ph.D. thesis, Faculté de médecine, Universtité de Tlemcen, 2019.

[7] Jacques Ferlay, M Colombet, I Soerjomataram, C Mathers, DM Parkin, M Piñeros, A Znaor, and F Bray, "Estimating the global cancer incidence and mortality in 2018 : Globocan sources and methods," *International journal of cancer*, vol. 144, no. 8, pp. 1941–1953, 2019.

[8] Dickran Kazandjian, "Multiple myeloma epidemiology and survival : A unique malignancy," in *Seminars in oncology*. Elsevier, 2016, vol. 43, pp. 676–681.

[9] Kari Hemminki, Asta Försti, Richard Houlston, and Amit Sud, "Epidemiology, genetics and treatment of multiple myeloma and precursor diseases," *International Journal of Cancer*, vol. 149, no. 12, pp. 1980–1996, 2021.

[10] Hadjira Ahmidatou and Nadia Belarbi Boudjerra, *Impact pronostique de l'insuffisance rénale chez les patients atteints de myélome multiple de novo á l'ére des nouvelles molécules*, Ph.D. thesis, 2021.

[11] J Ferlay, M Ervik, F Lam, M Colombet, L Mery, M Piñeros, et al., "Observatoire mondial du cancer, centre international de recherche sur le cancer [cancer today]," https ://gco.iarc.fr/today, 2020, [consulté en ligne le 25 Mars 2022].

[12] Blade Joan, Bruno Benedetto, and Mohty Mohamad, *Multiple Myeloma*, p. 603, Springer, 2019.

[13] Chérifa Guezlane, *Evaluation des facteurs pronostiques dans le myélome multiple*, Ph.D. thesis, Université Saad Dahleb de Blida 1, 2018.

[14] Marta Gonzalez Martinez, "Advances in the current treatment of multiple myeloma," 2020.

[15] Elizabeth Cardis, M Vrijheid, M Blettner, E Gilbert, M Hakama, C Hill, G Howe, J Kaldor, CR Muirhead, M Schubauer-Berigan, et al., "The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry : estimates of radiation-related cancer risks," *Radiation research*, vol. 167, no. 4, pp. 396–416, 2007.

[16] Ola Landgren, Kyle Robert A., Pfeiffer Ruth M., Katzmann Jerry A., Caporaso Neil E., Hayes Richard B., Dispenzieri Angela, Kumar Shaji, Clark Raynell J., Baris Dalsu, Hoover Robert, and Rajkumar S. Vincent, "Monoclonal gammopathy of undetermined significance (mgus) consistently precedes multiple myeloma : a prospective study," *Blood*, vol. 113, no. 22, pp. 5412–5417, 2009.

[17] Adrien Bosseboeuf, Sophie Allain-Maillet, Nicolas Mennesson, Anne Tallet, Cédric Rossi, Laurent Garderet, Denis Caillot, Philippe Moreau, Eric Piver, Francois Girodon, Héléne Perreault, Sophie Brouard,

Arnaud Nicot, Edith Bigot-Corbel, Sylvie Hermouet, and Jean Harb, ''Pro-inflammatory state in monoclonal gammopathy of undetermined significance and in multiple myeloma is characterized by low sialylation of pathogen-specific and other monoclonal immunoglobulins,'' *Frontiers in Immunology*, vol. 8, 2017.

[18] Heather Fairfield, Carolyne Falank, Lindsey Avery, and Michaela R Reagan, ''Multiple myeloma in the marrow : pathogenesis and treatments,'' *Annals of the New York Academy of Sciences*, vol. 1364, no. 1, pp. 32–51, 2016.

[19] Benmoussa Mehdi, *Les atteintes rénales au cours du myélome multiple*, Ph.D. thesis, Faculté de médecine et de pharmacie, Universtité Mohammed V de Rabat, Royaume du Maroc, 2021.

[20] ''Tumeur maligne, affection maligne du tissu lymphatique ou hématopoiétique : Myélome multiple,'' Tech. Rep., Institut National du Cancer, decembre 2010.

[21] Jean-Luc Harousseau, *Idées Vraies ou Fausses sur le Myélome Multiple*, Association Francaise des Malades du Myélome Multiple, 2012.

[22] Association Francaise des Polyarthritiques & des Rhumatismes Inflammatoires Chroniques, *Boite a outils : L'analyse de sang*.

[23] Anne Cairoli and Michel André Duchosal, ''Myélome multiple : diagnostic et perspectives thérapeutiques,'' p. 6 pages, 2013.

[24] C Touzeau and P Moreau, ''Imagerie du myélome multiple,'' *Journal de Radiologie diagnostique et interventionnelle*, vol. 94, no. 2, pp. 196–198, 2013.

[25] Christos Sachpekidis, Jens Hillengass, Hartmut Goldschmidt, Jennifer Mosebach, Leyun Pan, Heinz-Peter Schlemmer, Uwe Haberkorn, and Antonia Dimitrakopoulou-Strauss, ''Comparison of 18f-fdg pet/ct and pet/mri in patients with multiple myeloma,'' *American journal of nuclear medicine and molecular imaging*, vol. 5, no. 5, pp. 469, 2015.

[26] Zohra Ouzzif, Mounya Bouabdillah, Nouzha Jaouhar, Fatiha Aoufir, Farida Aoufi, Layachi Chabraoui, et al., ''Myélome multiple à immunoglobuline d,'' in *Annales de Biologie Clinique*, 2011, vol. 69, pp. 581–587.

[27] InternationaL Myeloma Foundation, North Hollywood, California, USA, *Comprendre L'électrophorése des protéines*, 2011.

[28] Frédérique Retornaz, Isabelle Potard, Caroline Franqui, Luc Benezech, Philippe Halfon, Frédérique Rousseau, Michelle Merlin, and

Catherine Molines, ''Conduite à tenir devant la découverte d'un pic monoclonal à l'électrophorése des protéines,'' *Ann Gerontol*, p. 7 pages, 2010.

[29] Raoul Karfo, Elie Kabré, Nadia Safir, Mounya Bouabdellah, Laila Benchekroun, Jean Sakandé, and Layachi Chabraoui, ''Interprétation délicate de l'immunofixation des protéines sériques,'' *Pan African Medical Journal*, vol. 30, no. 130, 2018.

[30] Jennifer LJ Heaney, John P Campbell, Punit Yadav, Ann E Griffin, Meena Shemar, Jennifer H Pinney, and Mark T Drayson, ''Multiple myeloma can be accurately diagnosed in acute kidney injury patients using a rapid serum free light chain test,'' *BMC nephrology*, vol. 18, no. 1, pp. 247, 2017.

[31] Moulin Bruno and Peraldi Marie-Noelle, *Néphrologie : Collége Universitaire des Enseignants de Néphrologie.*, chapter 8 : protéinurie et syndromes néphrotiques, pp. 111 – 125, 2016.

[32] Srinivasan Siva, ''A case report of false positive bence jones proteinuria,'' *University Journal of Pre and Para Clinical Sciences*, p. 5 pages, 2017.

[33] Babatunde O Oyajobi, ''Multiple myeloma/hypercalcemia,'' *Arthritis research & therapy*, vol. 9, no. 1, pp. 1–6, 2007.

[34] Ihssane BEN-TEBBA, *Rein et myélome multiple : Prévalence, facteurs de risque et pronostic*, Ph.D. thesis, Faculté de médecine et de pharmacie, Universtité CADI AYYAD, Marrakech, Maroc, 2013.

[35] Fadi E Rahhal, Robert R Schade, Asha Nayak, and Teresa A Coleman, ''Hepatic failure caused by plasma cell infiltration in multiple myeloma,'' *World journal of gastroenterology : WJG*, vol. 15, no. 16, pp. 2038, 2009.

[36] Abdollahzadeh Estakhri Mohammad reza, Kavakeb Parviz, Fathi Delaram, Karimi Gholamreza, Mohammadpour Amirhosein, and Rahbar Maryam, ''An investigation of the relationship between beta-2 microglobulin (b2m) and inflammatory factors (serum levels of crp and albumin) and high density lipoproteins (hdl) in hemodialysis patients,'' *Modern Medical Laboratory Journal*, p. 6 pages, 2017.

[37] Bailee Sliker, Cassie Liu, Brittany Poelaert, Benjamin Goetz, and Joyce C Solheim, ''Beta 2-microglobulin promotes human pancreatic cancer cell migration,'' *AACR*, 2018.

[38] International Myeloma Working Group, ''Criteria for the classification of monoclonal gammopathies, multiple myeloma and related

disorders : a report of the international myeloma working group,'' *British journal of haematology*, vol. 121, no. 5, pp. 749–757, 2003.

[39] Brian GM Durie and Sydney E Salmon, ''A clinical staging system for multiple myeloma correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival,'' *Cancer*, vol. 36, no. 3, pp. 842–854, 1975.

[40] Philip R Greipp, Jesus San Miguel, Brian GM Durie, John J Crowley, Bart Barlogie, Joan Bladé, Mario Boccadoro, J Anthony Child, Hervé Avet-Loiseau, Robert A Kyle, et al., ''International staging system for multiple myeloma,'' *Journal of clinical oncology*, vol. 23, no. 15, pp. 3412–3420, 2005.

[41] Antonio Palumbo, Hervé Avet-Loiseau, Stefania Oliva, Henk M Lokhorst, Hartmut Goldschmidt, Laura Rosinol, Paul Richardson, Simona Caltagirone, Juan José Lahuerta, Thierry Facon, et al., ''Revised international staging system for multiple myeloma : a report from international myeloma working group,'' *Journal of clinical oncology*, vol. 33, no. 26, pp. 2863, 2015.

[42] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis, ''Machine learning applications in cancer prognosis and prediction,'' *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[43] M Rossi, Maria Teresa Di Martino, Pietro Hiram Guzzi, Pierosandro Tagliaferri, and Pierfrancesco Tassone, ''New approaches to predict outcome and personalize therapy in multiple myeloma : from micrornas to integrated genomics,'' 2015.

[44] Aurore Palmaro, Martin Gauthier, Cécile Conte, Pascale Grosclaude, Fabien Despas, and Maryse Lapeyre-Mestre, ''Identifying multiple myeloma patients using data from the french health insurance databases : Validation using a cancer registry,'' *Medicine*, vol. 96, no. 12, 2017.

[45] Joske Ubels, Pieter Sonneveld, Erik H van Beers, Annemiek Broijl, Martin H van Vliet, and Jeroen de Ridder, ''Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects,'' *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[46] Jason Brownlee, *Data preparation for machine learning : data cleaning, feature selection, and data transforms in Python*, Machine Learning Mastery, 2020.

[47] Jason Brownlee, ''An introduction to feature selection,'' *Machine learning process*, vol. 6, 2014.

[48] Cen Wan, Wan, and Wheeler, *Hierarchical feature selection for Knowledge discovery*, Springer, 2019.

[49] Yvan Saeys, Inaki Inza, and Pedro Larranaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[50] Georg Heinze, Christine Wallisch, and Daniela Dunkler, "Variable selection–a review and recommendations for the practicing statistician," *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.

[51] Jeremy D Rogers and Steve R Gunn, "Ensemble algorithms for feature selection," in *International Workshop on Deterministic and Statistical Methods in Machine Learning*. Springer, 2004, pp. 180–198.

[52] Mark A Hall, *Correlation-based feature selection for machine learning*, Ph.D. thesis, The University of Waikato, 1999.

[53] T Elhassan and M Aljurf, "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method," *Global J Technol Optim S*, vol. 1, 2016.

[54] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[55] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera, *Learning from imbalanced data sets*, vol. 10, Springer, 2018.

[56] Bartosz Krawczyk, "Learning from imbalanced data : open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[57] Nathalie Japkowicz and Shaju Stephen, "The class imbalance problem : A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[58] Juan J Rodríguez, José F Díez-Pastor, and César García-Osorio, "Ensembles of decision trees for imbalanced data," in *International workshop on multiple classifier systems*. Springer, 2011, pp. 76–85.

[59] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera, "A review on ensembles for the class imbalance problem : bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2011.

[60] Nuno Moniz, Paula Branco, and Luís Torgo, ''Evaluation of ensemble methods in imbalanced regression tasks,'' in *First International Workshop on Learning with Imbalanced Domains : Theory and Applications*. PMLR, 2017, pp. 129–140.

[61] Jafar Tanha, Yousef Abdi, Negin Samadi, Nazila Razzaghi, and Mohammad Asadpour, ''Boosting methods for multi-class imbalanced data classification : an experimental review,'' *Journal of Big Data*, vol. 7, no. 1, pp. 1–47, 2020.

[62] Chao Chen, Andy Liaw, Leo Breiman, et al., ''Using random forest to learn imbalanced data,'' *University of California, Berkeley*, vol. 110, no. 1-12, pp. 24, 2004.

[63] Wei Feng, Wenjiang Huang, and Jinchang Ren, ''Class imbalance ensemble learning based on the margin theory,'' *Applied Sciences*, vol. 8, no. 5, pp. 815, 2018.

[64] Max Kuhn and Kjell Johnson, ''An introduction to feature selection,'' in *Applied predictive modeling*, pp. 487–519. Springer, 2013.

[65] Shelley Derksen and Harvey J Keselman, ''Backward, forward and stepwise automated subset selection algorithms : Frequency of obtaining authentic and noise variables,'' *British Journal of Mathematical and Statistical Psychology*, vol. 45, no. 2, pp. 265–282, 1992.

[66] Farideh Bagherzadeh-Khiabani, Azra Ramezankhani, Fereidoun Azizi, Farzad Hadaegh, Ewout W Steyerberg, and Davood Khalili, ''A tutorial on variable selection for clinical prediction models : feature selection methods in data mining could improve the results,'' *Journal of clinical epidemiology*, vol. 71, pp. 76–85, 2016.

[67] Loann David Denis Desboulets, ''A review on variable selection in regression analysis,'' *Econometrics*, vol. 6, no. 4, pp. 45, 2018.

[68] David, Fenghuang Zhan, James Cussens, Michael Waddell, Johanna Hardin, Bart Barlogic, and John Shaughnessy JR, ''Comparative data mining for microarrays : A case study based on multiple myeloma,'' Tech. Rep., University of Wisconsin-Madison Department of Computer Sciences, 2002.

[69] Eo-Jin Hwang, Joon-Yong Jung, Seul Ki Lee, Sung-Eun Lee, and Won-Hee Jee, ''Machine learning for diagnosis of hematologic diseases in magnetic resonance imaging of lumbar spines,'' *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[70] Xue Chen, Yao Zhang, Xiaohui Li, Ziheng Yang, Aichun Liu, and Xin Yu, ''Diagnosis and staging of multiple myeloma using serum-based laser-induced breakdown spectroscopy combined with machine

learning methods,'' *Biomedical Optics Express*, vol. 12, no. 6, pp. 3584–3596, 2021.

[71] Qingzhong Liu, Andrew H Sung, Zhongxue Chen, Jianzhong Liu, Xudong Huang, and Youping Deng, ''Feature selection and classification of maqc-ii breast cancer and multiple myeloma microarray gene expression data,'' *PloS one*, vol. 4, no. 12, pp. e8250, 2009.

[72] Lin Zhang, Jeffrey S Morris, Jiexin Zhang, Robert Z Orlowski, and Veerabhadran Baladandayuthapani, ''Bayesian joint selection of genes and pathways : applications in multiple myeloma genomics,'' *Cancer informatics*, vol. 13, pp. CIN–S13787, 2014.

[73] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, ''Smote : synthetic minority over-sampling technique,'' *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[74] Jason Brownlee, *Imbalanced classification with Python : better metrics, balance skewed classes, cost-sensitive learning*, Machine Learning Mastery, 2020.

[75] Zhang Xing Zhong, Akotonou J. Michael, Zhao Jie Lun, and Dong Hong Yue, ''Ecg classification using machine learning techniques and smote oversampling technique,'' in *2020 2nd International Conference on Image Processing and Machine Vision*, 2020, pp. 10–13.

[76] Chetna Kumari, Muhammad Abulaish, and Naidu Subbarao, ''Using smote to deal with class-imbalance problem in bioactivity data to predict mtor inhibitors,'' *SN Computer Science*, vol. 1, no. 3, pp. 1–7, 2020.

[77] Thien M Ha and Horst Bunke, ''Off-line, handwritten numeral recognition by perturbation method,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 535–539, 1997.

[78] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz, ''Applying support vector machines to imbalanced datasets,'' in *European conference on machine learning*. Springer, 2004, pp. 39–50.

[79] Georgios Douzas, Fernando Bacao, and Felix Last, ''Improving imbalanced learning through a heuristic oversampling method based on k-means and smote,'' *Information Sciences*, vol. 465, pp. 1–20, 2018.

[80] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, ''Borderline-smote : a new over-sampling method in imbalanced data sets learning,'' in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[81] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li, "Adasyn : Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.

[82] Kenji Kira and Larry A Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*, pp. 249–256. Elsevier, 1992.

[83] David W Aha, Dennis Kibler, and Marc K Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[84] Igor Kononenko, "Estimating attributes : Analysis and extensions of relief," in *European conference on machine learning*. Springer, 1994, pp. 171–182.

[85] Salim Chikhi and Sadek Benhammada, "Reliefmss : a variation on a feature ranking relieff algorithm.," *Int. J. Bus. Intell. Data Min.*, vol. 4, no. 3/4, pp. 375–390, 2009.

[86] Ricco Rakotomalala, "Analyse de corrélation. étude des dépendances–variables quantitatives (version 1.1.)," *Tiré de http ://www. eric. univ-lyon2. fr/ ricco/cours/Analyse de*, 2015.

[87] Krzysztof Michalak and Halina Kwasnicka, "Correlation based feature selection method," *International Journal of Bio-Inspired Computation*, vol. 2, no. 5, pp. 319–332, 2010.

[88] Philip Sedgwick, "Pearson's correlation coefficient," *British Medical Journal*, vol. 345, 2012.

[89] Lei Yu and Huan Liu, "Feature selection for high-dimensional data : A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[90] Youness Khourdifi and Mohamed Bahaj, "Feature selection with fast correlation-based filter for breast cancer prediction and classification using machine learning algorithms," in *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*. IEEE, 2018, pp. 1–6.

[91] Baris Senliol, Gokhan Gulgezen, Lei Yu, and Zehra Cataltepe, "Fast correlation based filter (fcbf) with a different search strategy," in *2008 23rd international symposium on computer and information sciences*. IEEE, 2008, pp. 1–4.

[92] B Azhagusundari, Antony Selvadoss Thanamani, et al., "Feature selection based on information gain," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 2, pp. 18–21, 2013.

[93] Thomas G Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[94] Jules Barthélemy Saint-Hilaire et al., *Politique d'Aristote*, Ladrange, 1874.

[95] Zhi-Hua Zhou, *Ensemble methods : foundations and algorithms*, A Chapman & Hall Book, Machine Learning & Pattern Recognition Series, Taylor & Francis Group, 2012.

[96] Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.

[97] Tianqi Chen and Carlos Guestrin, "Xgboost : A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[98] Chao Ren, Yan Xu, Yuchen Zhang, and Chunchao Hu, "A multiple randomized learning based ensemble model for power system dynamic security assessment," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2018, pp. 1–5.

[99] Leo Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[100] Leo Breiman, "Arcing classifiers," Tech. Rep., Citeseer, 1996.

[101] J. Ross Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[102] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[103] Pierre Geurts, Damien Ernst, and Louis Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[104] Yoav Freund and Robert E Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[105] Robert E Schapire, "Explaining adaboost," in *Empirical inference*, pp. 37–52. Springer, 2013.

[106] Jerome H Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[107] Afek Ilay Adler and Amichai Painsky, "Feature importance in gradient boosting trees with cross-validation feature selection," *Entropy*, vol. 24, no. 5, pp. 687, 2022.

[108] Xiupeng Shi, Yiik Diew Wong, Michael Zhi-Feng Li, Chandrasekar Palanisamy, and Chen Chai, "A feature learning approach based on xgboost for driving assessment and risk prediction," *Accident Analysis & Prevention*, vol. 129, pp. 170–179, 2019.

[109] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "Lightgbm : A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[110] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "Catboost : unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[111] Philipp Probst and Anne-Laure Boulesteix, "To tune or not to tune the number of trees in random forest," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6673–6690, 2017.

[112] James M Keller, Michael R Gray, and James A Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, , no. 4, pp. 580–585, 1985.

[113] Johan AK Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[114] S Rasoul Safavian and David Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[115] J Sunil Rao and William JE Potts, "Visualizing bagged decision trees.," in *KDD*, 1997, pp. 243–246.

[116] Shin-Jye Lee, Zhaozhao Xu, Tong Li, and Yun Yang, "A novel bagging c4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making," *Journal of biomedical informatics*, vol. 78, pp. 144–155, 2018.

[117] Rónán Daly, Qiang Shen, and Stuart Aitken, "Learning bayesian networks : approaches and issues," *The knowledge engineering review*, vol. 26, no. 2, pp. 99–157, 2011.

[118] Thomas Bayes, ''Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s,'' *Philosophical transactions of the Royal Society of London*, , no. 53, pp. 370–418, 1763.

[119] Nir Friedman, Dan Geiger, and Moises Goldszmidt, ''Bayesian network classifiers,'' *Machine learning*, vol. 29, pp. 131–163, 1997.

[120] Youcef Benmouna, Mourtada Benazzouz, Mohammed Amine Chikh, and Saïd Mahmoudi, ''New method for bayesian network learning,'' *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 02, pp. 1959005, 2019.

[121] Philippe Leray and Olivier François, ''Réseaux bayésiens pour la classification méthodologie et illustration dans le cadre du diagnostic médical.,'' *Rev. d'Intelligence Artif.*, vol. 18, no. 2, pp. 169–193, 2004.

[122] Darren J Wilkinson, ''Bayesian methods in bioinformatics and computational systems biology,'' *Briefings in bioinformatics*, vol. 8, no. 2, pp. 109–116, 2007.

[123] Ilias Maglogiannis, Elias Zafiropoulos, A Platis, and Costas Lambrinoudakis, ''Risk analysis of a patient monitoring system using bayesian network modeling,'' *Journal of Biomedical informatics*, vol. 39, no. 6, pp. 637–647, 2006.

[124] Zhi Yuan, Nima Khakzad, Faisal Khan, and Paul Amyotte, ''Risk analysis of dust explosion scenarios using bayesian networks,'' *Risk analysis*, vol. 35, no. 2, pp. 278–291, 2015.

[125] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz, ''A bayesian approach to filtering junk e-mail,'' in *Learning for Text Categorization : Papers from the 1998 workshop*. Citeseer, 1998, vol. 62, pp. 98–105.

[126] Youcef Benmouna, *Amélioration des performances des réseaux bayésiens dans le domaine medical*, Ph.D. thesis, Université de Tlemcen-Abou Bekr Belkaid, 2019.

[127] Evangelia Kyrimi, Scott McLachlan, Kudakwashe Dube, Mariana R Neves, Ali Fahmi, and Norman Fenton, ''A comprehensive scoping review of bayesian networks in healthcare : Past, present and future,'' *Artificial Intelligence in Medicine*, vol. 117, pp. 102108, 2021.

[128] Xiaofeng Wu, Peter Lucas, Susan Kerr, and Roelf Dijkhuizen, ''Learning bayesian-network topologies in realistic medical domains,'' in *Medical Data Analysis : Second International Symposium, ISMDA*

2001 Madrid, Spain, October 8–9, 2001 Proceedings 2. Springer, 2001, pp. 302–307.

[129] Simon Cauchemez, Fabrice Carrat, Cécile Viboud, Alain Jacques Valleron, and PierreYves Boëlle, "A bayesian mcmc approach to study transmission of influenza : application to household longitudinal data," *Statistics in medicine*, vol. 23, no. 22, pp. 3469–3487, 2004.

[130] Kavishwar B Wagholikar, Sundararajan Vijayraghavan, and Ashok W Deshpande, "Fuzzy naive bayesian model for medical diagnostic decision support," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 3409–3412.

[131] Tetsuji Okuda, Hideo Tanaka, and Kiyoji Asai, "A formulation of fuzzy decision problems with fuzzy information using probability measures of fuzzy events," *Information and Control*, vol. 38, no. 2, pp. 135–147, 1978.

[132] Dania Abed Aljawad, Ebtesam Alqahtani, AL-Kuhaili Ghaidaa, Nada Qamhan, Noof Alghamdi, Saleh Alrashed, Jamal Alhiyafi, and Sunday O Olatunji, "Breast cancer surgery survivability prediction using bayesian network and support vector machines," in *2017 International Conference on Informatics, Health & Technology (ICIHT)*. IEEE, 2017, pp. 1–6.

[133] Bouchra Zoullouti, Mustapha Amghar, and Sbiti Nawal, "Using bayesian networks for risk assessment in healthcare system," *Bayesian networks-Advances and novel applications*, 2019.

[134] Paul Arora, Devon Boyne, Justin J Slater, Alind Gupta, Darren R Brenner, and Marek J Druzdzel, "Bayesian networks for risk prediction using real-world data : a tool for precision medicine," *Value in Health*, vol. 22, no. 4, pp. 439–445, 2019.

[135] Lixia Zhang, Leonardo O Rodrigues, Niven R Narain, and Viatcheslav R Akmaev, "baicis : A novel bayesian network structural learning algorithm and its comprehensive performance evaluation against open-source software," *Journal of Computational Biology*, vol. 27, no. 5, pp. 698–708, 2020.

[136] Michal Hornỳ, "Bayesian networks," *Boston University School of Public Health*, vol. 17, 2014.

[137] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Discretization techniques : A recent survey," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.

[138] Farnaz Nojavan, Song S Qian, and Craig A Stow, "Comparative analysis of discretization methods in bayesian networks," *Environmental Modelling & Software*, vol. 87, pp. 64–71, 2017.

[139] Tomas Beuzen, Lucy Marshall, and Kristen D Splinter, "A comparison of methods for discretizing continuous variables in bayesian networks," *Environmental modelling & software*, vol. 108, pp. 61–66, 2018.

[140] James Dougherty, Ron Kohavi, and Mehran Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine learning proceedings 1995*, pp. 194–202. Elsevier, 1995.

[141] Serena H Chen and Carmel A Pollino, "Good practice in bayesian network modelling," *Environmental Modelling & Software*, vol. 37, pp. 134–145, 2012.

[142] Pedro Aguilera Aguilera, Antonio Fernández, Rosa Fernández, Rafael Rumí, and Antonio Salmerón, "Bayesian networks in environmental modelling," *Environmental Modelling & Software*, vol. 26, no. 12, pp. 1376–1388, 2011.

[143] Daphne Koller and Nir Friedman, *Probabilistic graphical models : principles and techniques*, MIT press, 2009.

[144] Zhiwei Ji, Qibiao Xia, and Guanmin Meng, "A review of parameter learning methods in bayesian network," in *Advanced Intelligent Computing Theories and Applications : 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part III 11*. Springer, 2015, pp. 3–12.

[145] Ankur Ankan and Abinash Panda, "pgmpy : Probabilistic graphical models using python," in *Proceedings of the 14th python in science conference (scipy 2015)*. Citeseer, 2015, vol. 10.

[146] Bart Selman and Carla P Gomes, "Hill-climbing search," *Encyclopedia of cognitive science*, vol. 81, pp. 82, 2006.

[147] Michael J Zyphur and Frederick L Oswald, "Bayesian estimation and inference : A user's guide," *Journal of Management*, vol. 41, no. 2, pp. 390–420, 2015.

[148] Anton J Haug, *Bayesian estimation and tracking : a practical guide*, John Wiley & Sons, 2012.

[149] In Jae Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[150] Shunkai Fu and Michel C Desmarais, "Markov blanket based feature selection : a review of past decade," in *Proceedings of the world*

congress on engineering. Newswood Ltd. Hong Kong, China, 2010, vol. 1, pp. 321–328.

[151] Zahraa Mudher M Salih and Haithem Ahmed Al-Rubaie, ‘‘Evaluation of angiopoietin-2 level in patients with multiple myeloma at presentation and in remission state,’’ 2023.