UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODE D'EXTRACTION DE SIGNATURES DE BIOMARQUEURS MÉTABOLIQUES DANS LE CADRE DE PRÉDICTION DES MALADIES CHEZ LES BOVINS LAITIERS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

ABDOURAHMANE BALDE

NOVEMBRE 2024

RÉSUMÉ

Les maladies métaboliques chez les bovins laitiers sont une préoccupation majeure pour les producteurs laitiers, car elles affectent la production, la reproduction, le bien-être et la longévité des animaux. Les approches de classification informatique actuelles des maladies métaboliques manquent de précision et de simplicité pour permettre une classification effective en milieu de ferme. Cela est dû principalement à la variabilité naturelle des profils métaboliques entre les vaches et à la distribution non équilibrée des classes (malade, non malade) dans les jeux de données sur les maladies surveillées par l'industrie laitière. Cette disparité entre les classes rend la prédiction des maladies encore plus complexe. Pour permettre aux vétérinaires de mieux suivre les paramètres métaboliques en lien avec les principales maladies suivies en médecine vétérinaire, nous proposons une approche d'extraction des signatures métaboliques qui permettrait de caractériser les maladies, d'identifier une meilleure prédiction des risques associés et d'informer les classifieurs pour le diagnostique des maladies métaboliques. Notre approche est décomposée en trois étapes comme suit : la catégorisation des indicateurs métaboliques clés, l'extraction des signatures métaboliques présentes dans les maladies, la sélection des signatures discriminantes à l'aide de deux mesures statistiques (le test de Fisher et l'intervalle de confiance), et enfin l'entraînement des modèles de classification afin de prédire une maladie métabolique à partir d'un profil métabolique. Cette approche a été appliquée à un jeu de données contenant cinq indicateurs métaboliques pour 623 vaches. Ces indicateurs sont associés aux sept maladies principales à déclaration obligatoire au Québec. Les résultats préliminaires ont permis d'identifier 208 signatures métaboliques et 96 signatures discriminantes. Ces signatures ont ensuite permis de construire un classifieur pour identifier si une vache est diagnostiquée comme malade seulement à partir de son échantillon de données métaboliques. Ces signatures discriminantes constitueront un potentiel de biomarqueurs utiles pour le diagnostic des maladies. Les résultats de ces classifieurs sont intéressants avec des F-mesure avoisinant 0,91.

Mots-clés : Maladies métaboliques, Bovins laitiers, Producteurs laitiers du Canada, Profil métabolique, Signatures métabolique, Tests Statistiques, Apprentissage machine, Prédiction.

ABSTRACT

Metabolic diseases in dairy cattle are a major concern for dairy producers, as they affect animal production,reproduction, welfare and longevity. Current computer classification approaches for metabolic diseases lack the accuracy and simplicity to enable effective on-farm classification. This is mainly due to the natural variability of metabolic profiles between cows and the unbalanced distribution of classes (sick, not sick) in the disease datasets monitored by the dairy industry. This disparity between classes makes disease prediction even more complex. To enable veterinarians to better monitor the metabolic parameters associated with the main diseases monitored in veterinary medicine, we propose a metabolic signature extraction approach that would characterize diseases, identify better prediction of associated risks and inform classifiers for metabolic disease diagnosis. Our approach is broken down into three steps as follows : categorization of key metabolic indicators, extraction of metabolic signatures present in diseases, selection of discriminating signatures using two statistical measures (Fisher's test and confidence interval), and finally training of classification models to predict a metabolic disease from a metabolic profile. This approach was applied to a dataset containing five metabolic indicators for 623 cows. These indicators are associated with the seven major notifiable diseases in Quebec. Preliminary results identified 208 metabolic signatures and 96 discriminating signatures. These signatures were then used to build a classifier to identify whether a cow is diagnosed as sick based solely on her metabolic data sample. The results of these classifiers are interesting, with F-measurements approaching 0.91.

Keywords : Metabolic diseases, Dairy cattle, Canadian dairy producers, Metabolic profile, Metabolic signatures, Statistical tests, Machine learning, Prediction.

REMERCIEMENTS

Tout d'abord, je tiens à exprimer ma reconnaissance envers mon père pour son soutien. Il demeure un modèle pour moi, et je suis fier d'être son fils. Ensuite, je tiens à remercier ma mère, une femme d'une force inébranlable. Les mots me manquent pour exprimer mes sentiments à son égard. Je suis également reconnaissant envers mon grand frère et mentor, Cheikhna Balde, sans qui je ne serais pas l'homme que je suis aujourd'hui. Je voudrais également adresser mes remerciements à toute ma famille pour leur soutien inconditionnel, sans oublier mes deux nièces, Ndoura et Nene Bu Daw.

À Cheikhal Islam Cheikh Ibrahima Niasse (RTA), je vous exprime ma profonde gratitude pour m'avoir introduit à la voie de la spiritualité et m'avoir permis de découvrir la proximité avec Dieu, par le biais de Malawlana Sangue Aboubakrine Ndiaye. Vous êtes la preuve vivante que les Sénégalais peuvent s'élever et s'adapter dans ce monde globalisé.

Je souhaite exprimer ma reconnaissance envers mon directeur de recherche, le Professeur Abdoulaye Baniré Diallo, pour son leadership et ses précieux conseils qui ont guidé ma recherche. Je remercie également les membres de mon laboratoire, notamment Hayda, pour son professionnalisme et le partage de ses connaissances précieuses. Golrokh Kiani et Amanda, pour leur disponibilité et leur générosité à mon égard. Armand, pour son dynamisme et sa rigueur dans le travail. Thomas, pour son professionnalisme et sa disponibilité, Mohamed Amine Remita, pour ses conseils avisés et sa disponibilité. Je suis reconnaissant envers Mamadou, Khaly, Issa, Julie pour nos échanges enrichissants.

Je tiens à exprimer ma gratitude envers mes amis pour leur soutien indéfectible tout au long de ces dernières années. Enfin, je remercie sincèrement toutes les personnes qui ont contribué de près ou de loin au succès de mon projet de maîtrise.

"Sa andi a anda a andat Sa anda a anda a andata."

"Si tu sais que tu ne sais pas, alors tu sauras. Mais si tu ne sais pas que tu ne sais pas, alors tu ne sauras point"

Amadou Hampaté Bah

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

BHB: Beta-hydroxybutyrate.

DA: déplacement d'abomasum.

DCD: Le déplacement de caillette à droite.

DCG: Le déplacement de caillette à gauche.

DSA: Dossier de Santé Animale.

FN: nombre de faux négatifs.

FP: nombre de faux positifs.

IC: intervalle de confiance.

KNN: K-Nearest Neighbors.

MAP: Mycobacterium avium sous-espèce paratuberculosis.

MUN: Milk Urea Nitrogen.

NIH: National Institutes of Health.

NTIC: Nouvelles Technologies de l'Information et de la Communication.

OR: Rapports de Cotes.

PCA: Analyse en Composantes Principales.

RMN: résonance magnétique nucléaire.

SCC: nombre de cellules somatiques.

SVM: Support Vector Machines.

TFN: Taux de faux négatifs.

TFP: Taux de faux Positifs.

TVN: Taux de vrais négatifs.

VN: nombre de vrais négatifs.

VP: nombre de vrais positifs.

INTRODUCTION

L'élevage joue un rôle crucial dans de nombreux pays à travers le monde. Les projections des Nations Unies indiquent une augmentation de la population mondiale de 7,5 milliards à 9,7 milliards d'ici 2050, ce qui se traduirait par une augmentation des besoins en produits agricoles (Roy *et al.*, 2022). La prise de décision dans l'élevage traditionnel était généralement basée sur l'expérience du producteur. Afin de gérer efficacement les animaux, il est impératif de promouvoir le progrès technologique dans le domaine de l'élevage qui représente environ un tiers de la production agricole mondiale, comme indiqué dans (Häring, 2003). Ce qui donne naissance à de nouvelles notions telles que l'agriculture de précision (Bewley, 2010).

L'apprentissage automatique représente une composante essentielle de l'intelligence artificielle (IA), centrée sur l'utilisation d'algorithmes avancés pour simuler le processus d'apprentissage, ce qui améliore progressivement leur précision. Au cœur du domaine en pleine expansion de la science des données, les algorithmes sont conçus pour effectuer des tâches telles que la catégorisation ou la prédiction en utilisant des méthodes statistiques, ce qui permet de révéler des informations cruciales lors de l'analyse des données (Roy *et al.*, 2022). L'intégration de nouvelles technologies dans l'élevage de précision devient primordiale pour optimiser la rentabilité des exploitations tout en améliorant le suivi de la production et le bien-être des animaux (Caja *et al.*, 2016). Par exemple (Naghashi *et al.*, 2023) ont également mis en évidence l'importance des algorithmes d'apprentissage dans la prévision du revenu laitier, en utilisant des méthodes de séries chronologiques univariées et multivariées basées sur les attributs laitiers. (Karoui *et al.*, 2021) ont développé un modèle d'apprentissage profond visant à détecter la boiterie en se basant sur l'analyse des mouvements des différentes articulations des jambes des vaches. Cette approche pourrait être généralisée a l'inclusion d'autres types de mesures tels que les scores de boiterie évalués par des experts, et la température a la surface de l'onglon des vaches laitières (Nejati *et al.*, 2024).

Les maladies métaboliques chez les vaches laitières constituent un défi majeur pour les éleveurs, non seulement en termes de santé animale, mais également en matière de rentabilité et de bien-être animal. La détection précoce et la gestion efficace de ces affections revêtent une importance cruciale pour maintenir la productivité et la qualité du lait, tout en réduisant les coûts de traitement et les pertes économiques associées (Pyorala, 2003). Par exemple, dans les conditions normales, les vaches peuvent produire entre 12 à 15 litres de lait par jour, tandis que cette production peut chuter à 5 à 10 litres en cas de maladie tel que noté par (Wang *et al.*, 2022) pour l'élevage de bovin laitier en Inde. Au Québec, les fermes laitières

comptent en moyenne 73 vaches, chacune produisant plus de 9 300 litres de lait par année (Gouvernement du Québec, 2019).

L'utilisation des signatures de biomarqueurs métaboliques dans l'agriculture implique l'identification et l'analyse de métabolites spécifiques présents dans les échantillons biologiques, tels que le lait, le sang ou l'urine des animaux (Pyorala, 2003; Dervishi *et al.*, 2021). Ces biomarqueurs peuvent fournir des informations précieuses sur l'état de santé des animaux, la qualité de leur alimentation et leur environnement, ainsi que sur leurs performances de production. Divers biomarqueurs sont utilisés pour évaluer la qualité du lait et surveiller la santé des animaux ainsi que leur alimentation, comme les acides gras, les protéines, le lactose (BROLIS HerdLine, 2024) et les composés phénoliques (Dervishi *et al.*, 2021). Les biomarqueurs de la santé animale jouent un rôle crucial dans la détection précoce des maladies et des déséquilibres métaboliques chez les animaux d'élevage. Ces biomarqueurs comprennent des métabolites spécifiques associés à des affections telles que la cétose ou la mammite (Dervishi *et al.*, 2017). La compréhension de ces maladies, de leurs symptômes, et de leurs facteurs de risque est essentielle pour maintenir la santé et la productivité des vaches laitières. Une détection précoce et des mesures de prévention adéquates sont cruciales pour réduire l'impact de ces maladies sur les troupeaux laitiers (Kares, 2022). Dans ce contexte, l'utilisation de l'apprentissage supervisé pour prédire ces maladies revêt une importance significative (Tanyildizl et Yildirim, 2019).

La problématique de recherche soulevée est la suivante : Comment pouvons-nous, à partir d'un ensemble de dosages métaboliques associés à des maladies, identifier des signatures métaboliques et les intégrer dans un modèle de classification ? Ce mémoire se concentre sur la conception d'une approche pour la classification des maladies métaboliques chez les vaches laitières basée sur des signatures métaboliques. L'objectif de cette étude est de concevoir une méthodologie pour identifier des signatures de métabolites uniques associées aux maladies identifiables dans un échantillon contenant cinq métabolites essentiels et de les intégrer dans un modèle de prédiction des maladies métaboliques.

Le présent mémoire est structuré en cinq (05) chapitres :

— Dans le chapitre 1, nous posons les fondements en définissant les maladies les plus courantes chez les vaches laitières, concepts fondamentaux en lien avec les données métaboliques et les impacts de la santé animale sur la production laitière.

— Le chapitre 2 introduit les concepts d'analyses statistiques et d'apprentissage automatique pour comprendre et traiter les données.
— Le chapitre 3 est dédié à la revue de la littérature et à la définition du problème. Il établit le cadre de cohérence de notre recherche de solutions.
— Le chapitre 4 expose notre méthodologie avec une description détaillée des données utilisées. Nous exposons l'approche biostatistique proposée pour extraire les signatures métaboliques, et l'utilisation de l'apprentissage automatique pour prédire les maladies métaboliques chez les vaches laitières.
— Dans le chapitre 5, nous présentons les résultats obtenus à travers une analyse statistique approfondie des données métaboliques recueillies. Nous mettons en lumière les signatures discriminantes identifiées et évaluons les performances de classification des signatures à travers plusieurs algorithmes.

CHAPITRE 1
NOTIONS PRÉLIMINAIRES : SANTÉ DES VACHES AXÉE SUR LES DONNÉES

Dans ce chapitre, nous présenterons les maladies présentes dans notre jeu de données, définirons le concept de signatures biomarqueurs, et enfin, nous aborderons les impacts de la santé animale dans la production laitière.

1.1 Introduction aux maladies suivies par les vétérinaires et leurs incidences métaboliques

Dans l'agriculture, l'élevage des vaches laitières est essentiel pour fournir du lait aux consommateurs. La santé des vaches laitières est importante car elle influence directement leur capacité à produire du lait et donc la rentabilité des fermes. Un défi majeur pour les éleveurs et les vétérinaires est de gérer les maladies qui affectent les troupeaux laitiers. Parmi les maladies, on retrouve la mammite, la boiterie, l'acétonémie, le déplacement de la caillette et la métrite. Ces maladies peuvent perturber le métabolisme normal des vaches, ce qui peut modifier la production et la qualité du lait. Dans ce qui suit, nous examinerons ces différentes maladies.

Mammite : La mammite, une inflammation de la mamelle généralement provoquée par une bactérie, se manifeste par divers signes, dont des symptômes généraux tels que l'état de choc, la perte d'appétit, et des changements de température corporelle. Les conséquences économiques sont considérables, avec des pertes de productivité laitière, une diminution des performances reproductives et une augmentation de la mortalité lorsque la maladie n'est pas détectée dans les premières 24 heures (Tanyildizl et Yildirim, 2019; Fadul-Pacheco *et al.*, 2021). Une recherche portant sur 2 087 vaches en Floride a montré qu'une vache souffrant de mammite clinique au cours des 45 premiers jours de gestation a 2,7 fois plus de risques d'avorter dans les 90 jours suivants (Chantal, 2022). Les conséquences ne se limitent pas à la période d'affection, car ces vaches ne retrouvent pas leur potentiel de production laitière habituel (Fadul-Pacheco *et al.*, 2021).

Boiterie : La boiterie se caractérise par la présentation clinique de troubles douloureux, principalement associés au système locomoteur, provoquant une altération du mouvement ou une déviation par rapport à la démarche ou à la posture normale (Van Nuffel *et al.*, 2015). Cet enjeu crucial pour le bien-être des vaches laitières se traduit par d'importants coûts pour les producteurs. En outre, elle génère des douleurs, entraîne une diminution de la production laitière et réduit la longévité des animaux (Wells *et al.*, 1998),

en faisant ainsi l'un des problèmes de santé et de bien-être préoccupants dans les exploitations laitières modernes.

Acétonémie : L'acétonémie, est une condition qui peut survenir pendant la période post-partum (deuxieme velage), souvent dans les premières semaines après la mise bas. C'est une affection métabolique liée à un déséquilibre énergétique, où la vache ne parvient pas à couvrir ses besoins énergétiques en raison d'une ingestion alimentaire insuffisante par rapport à ses besoins énergétiques (Lopez, 2021).

Déplacement de caillette : Le déplacement de caillette, également appelé déplacement de réseau de caillettes, est une condition chez les ruminants, tels que les vaches, où l'un des compartiments de leur système digestif, appelé caillette, se déplace de sa position normale. Le déplacement d'abomasum (DA) est une situation qui a un impact significatif sur le bien-être des vaches laitières et occasionne d'importantes pertes économiques (Abdullah Basoglu et Gulersoy, 2020a). Le déplacement de caillette à gauche (DCG) constitue la majorité des cas de déplacements de caillette, représentant 85% du total (les 15% restants correspondant à des déplacements à droite avec ou sans torsion). Le DCG se caractérise par une déviation partielle de la caillette, normalement située à droite, entre le rumen et la paroi thoracoabdominale gauche. Cette condition s'accompagne d'une dilatation plus ou moins prononcée due à l'accumulation de gaz (Éleveurs de demain - Le blog, 2020).

Fièvre vitulaire : La fièvre de lait est un trouble métabolique causé par un manque de calcium (Skelly, 2023). Elle affecte principalement les vaches laitières qui produisent beaucoup de lait, généralement après leur deuxième mise bas. Cette condition résulte d'une augmentation soudaine des besoins en calcium lors du début de la lactation. La manifestation classique de la fièvre de lait se produit généralement dans les 48 heures suivant le vêlage. Initialement, la vache présente une diminution de l'appétit et de la consommation d'eau, entraînant un arrêt de la rumination. Elle éprouve des difficultés à se lever ou à rester debout. En l'absence d'un traitement rapide, la vache peut tomber dans le coma et succomber en moins de 24 heures (Clinique Vétérinaire de l'Aérodrome Saint Romain de Colbosc, 2024).

Métrite : La métrite est une maladie post-partum courante chez les vaches laitières qui se caractérise par des taux élevés de leucotriène B4, ce qui entraîne une inflammation et des lésions tissulaires de l'utérus (Wang *et al.*, 2023). La métrite puerpérale est une affection infectieuse courante chez les vaches laitières, constituant la deuxième raison principale de l'utilisation d'antimicrobiens. Elle impacte négativement la

rentabilité des troupeaux en réduisant la production laitière, en compromettant l'efficacité reproductive et en augmentant le risque d'élimination précoce. De plus, la métrite puerpérale est associée à des problèmes de bien-être animal. Les coûts estimés de la métrite englobent l'utilisation d'antimicrobiens, la perte de lait, la baisse de la production laitière et la diminution de l'efficacité reproductive (Garzon *et al.*, 2022).

1.2 Définition des biomarqueurs

Un biomarqueur, tel que défini par le groupe de travail sur les définitions des biomarqueurs du National Institutes of Health en 1998, se réfère à une caractéristique mesurée de façon objective qui sert d'indicateur pour les processus biologiques normaux, les processus pathogènes, ou les réponses pharmacologiques à une intervention thérapeutique (Marchand *et al.*, 2018). (Foroutan *et al.*, 2020) a révélé par la présence de 142 métabolites dans le sérum, 232 dans le liquide ruminal, 52 dans l'urine et 972 dans le lait.

Parallèlement, le terme "biomarqueur" regroupe différentes définitions qui varient selon le domaine d'application et les applications spécifiques, généralement désignant des mesures objectives destinées à décrire l'état actuel d'un système biologique (Gao *et al.*, 2017). Le biomarqueur reflète de manière précise et sensible un état pathologique. Il peut être utilisé pour diagnostiquer, prédire la réponse aux médicaments, ainsi que pour surveiller l'évolution de la maladie pendant et après le traitement (Jain, 2008). L'étude (La Santé Des Ruminants, 2023) a démontré que le temps de rumination ainsi que certains éléments sanguins tels que les lactates, les protéines totales, l'albumine et la créatinine peuvent être utilisés comme biomarqueurs pour détecter les signes d'acidose et de cétose subcliniques. L'analyse métabolomique peut souligner des signatures métaboliques propres à chaque maladie et spécifiques à chaque individu.

Enfin, les travaux de (Dervishi *et al.*, 2021) ont révélé des caractéristiques métaboliques communes entre différentes maladies chez les vaches laitières, mettant en évidence la complexité des interactions métaboliques dans le contexte de leur santé.

Le concept de signature englobe un ensemble distinctif de caractéristiques, de mesures ou de motifs étroitement associés à une condition médicale spécifique, à une réponse thérapeutique, ou à d'autres aspects médicaux. En résumé, une signature se présente comme un modèle ou un ensemble de marqueurs fournissant des informations précieuses pour la compréhension, le diagnostic, le traitement, ou la prédiction d'une condition médicale particulière.

1.3 Types de données et organisations des biomarqueurs métaboliques

Les biomarqueurs métaboliques peuvent être classés en différentes catégories en fonction de leur nature et de leur origine :

- **Métabolites :** sont des biomolécules essentielles dans les voies physiologiques, reflétant l'état physiochimique d'un individu. Ils constituent des biomarqueurs potentiels pour le diagnostic, le pronostic et les effets de l'exposition environnementale sur la santé (Aggarwal *et al.*, 2022).
- **Profil métabolique :** désigne l'analyse quantitative et qualitative des métabolites présents dans un échantillon biologique, tel que le sang, l'urine, le tissu, ou d'autres fluides biologiques. Les métabolites sont les molécules produites lors des processus métaboliques de l'organisme, et leur profil peut fournir des informations précieuses sur l'état physiologique et les réponses aux stimuli externes (Zhang *et al.*, 2022).
- **Profils spécifiques :** Certains biomarqueurs métaboliques peuvent être spécifiquement associés à une condition ou une maladie particulière. Par exemple, des métabolites tels que les acides biliaires, les acides aminés et les acides gras ont été identifiés comme biomarqueurs pour la détection précoce de la stéatose hépatique chez les vaches laitières, facilitant ainsi le diagnostic et la prise en charge (Zhang *et al.*, 2022) .
- **Modèles de signature :** Les modèles de signature métabolique consistent à combiner plusieurs biomarqueurs métaboliques pour former un modèle ou une signature diagnostique. Ces modèles peuvent être utilisés pour prédire ou classifier différents états physiologiques ou pathologiques (Zhang *et al.*, 2022).
- **Données multiomiques :** Les données multiomiques intègrent plusieurs types de données omiques, tels que les données génomiques, transcriptomiques, protéomiques et métabolomiques, pour obtenir une image plus complète des processus biologiques (Bersanelli *et al.*, 2016; Vilanova et Porcar, 2016).

1.4 Méthodes de collecte des biomarqueurs

Le profilage métabolique implique souvent l'utilisation de techniques analytiques avancées telles que la chromatographie liquide couplée à la spectrométrie de masse (LC-MS) (He *et al.*, 2022) ou la spectroscopie de résonance magnétique nucléaire (RMN) (Dervishi *et al.*, 2018) pour identifier et quantifier les métabolites dans les échantillons biologiques. Le prélèvement sanguin est une méthode courante pour collecter des biomarqueurs chez les vaches laitières. Dans une étude menée par (Favole *et al.*, 2023), les biomar-

queurs du bien-être des vaches laitières ont été collectés en analysant les métabolites plasmatiques, le facteur neurotrophique dérivé du cerveau et l'indoleamine 2,3-dioxygénase (IDO1) dans des échantillons de sang provenant de bovins abattus et en utilisant la spectrométrie de masse. Les méthodes de collecte des biomarqueurs du stress chronique chez les vaches laitières comprennent la surveillance de la perte de lait, de la fréquence cardiaque, de la rumination, du cortisol capillaire, de la fructosamine sanguine et d'autres paramètres physiologiques tels que le cortisol salivaire et la glycémie (Grelet *et al.*, 2022). Des échantillons de sang et de lait ont été prélevés chez des vaches Holstein en début de lactation aux 1er, 2e et 3e semaines après l'accouchement afin d'évaluer des biomarqueurs tels que les corps cétoniques, les acides gras, le bilan énergétique négatif et la fertilité (Mansour *et al.*, 2021). Leur article, (Ojo *et al.*, 2023) ont utilisé le séquençage de nouvelle génération pour détecter des animaux présentant des déséquilibres métaboliques tels que l'acidose, souvent observée chez les vaches laitières à haut rendement.

1.5 Impact de la santé animale et de la production laitière

L'impact de la santé animale et de la production laitière peut être considéré sous différents aspects.

impact sur l'environnement et la biodiversité : La production laitière intensive a un impact direct sur la biodiversité et la santé de l'écosystème, notamment par le biais des changements d'utilisation des terres à la ferme et des processus de production d'aliments pour animaux (Etienne, 2020). Les systèmes laitiers intensifs, qui dépendent souvent de pâturages monoculturels et utilisent des quantités élevées d'engrais, peuvent présenter des risques pour la biodiversité et la stabilité des écosystèmes (Etienne, 2020). Ces dernier facteurs impactent en retour la qualité du lait, et de même que la productivité des animaux. Les coûts des soins vétérinaires représentent une considération majeure, et la durabilité de l'industrie est également mise à l'épreuve par les défis posés par le changement climatique. Les réponses physiologiques des animaux face à des températures élevées incluent une augmentation du taux de respiration, de la température rectale et du rythme cardiaque, entraînant des conséquences directes sur la prise alimentaire, la croissance, la production laitière, et les performances reproductives, voire la mortalité dans des cas extrêmes (Das *et al.*, 2016).

impact sur la santé humaine : L'intensification de la production laitière peut également avoir des impacts indirects sur la santé humaine (Etienne, 2020). Par exemple, les fermes laitières peuvent contribuer à la pollution de l'environnement et à la production de gaz à effet de serre, ce qui peut avoir des conséquences

sur la qualité de l'air et de l'eau, ainsi que sur le changement climatique (Etienne, 2020). C'est pourquoi (Ezanno *et al.*, 2020) a suggéré de renforcer les équipes existantes autour du concept OneHealth. Les animaux d'élevage hébergent une multitude de micro-organismes, dont certains sont pathogènes. Cette diversité comprend sept genres bactériens (Campylobacter sp, Coxiella sp, Escherichia sp, Leptospira sp, Listeria sp, Salmonella sp et Yersinia sp), deux genres de parasites (Cryptosporidium sp et Giardia sp) ainsi qu'un virus (Influenza). Des études scientifiques ont confirmé la possibilité de transmission de ces agents pathogènes de l'environnement animal à l'humain (Institut National de Santé Publique Du Québec, 2023). Près de la moitié de l'ensemble des antibiotiques sont utilisés dans l'agriculture en Amérique du Nord. Cette pratique excessive contribue à accroître la résistance des populations bactériennes, qui peuvent ensuite être transmises aux humains (Institut National de Santé Publique Du Québec, 2023).

Bien-être animal : Le bien-être des animaux d'élevage, y compris les vaches laitières, est une préoccupation importante. Les fermes laitières doivent respecter des réglementations strictes en matière de bien-être animal, et de nombreux éleveurs s'engagent dans des chartes de bonnes pratiques pour veiller au bien-être de leurs animaux (Ministère de l'Agriculture et de la Souveraineté Alimentaire, 2019). Cette tache est complique maintenant par les changements climatiques. En effet le stress thermique supprime les systèmes immunitaire et endocrinien, ce qui rend les animaux plus sensibles à diverses maladies (Das *et al.*, 2016), y compris chez les animaux d'élevage, compromettant leur bien-être et ayant des répercussions négatives sur la production laitière.

Le bien-être animal constitue également un aspect crucial de la production laitière. Les éleveurs accordent une grande importance au traitement attentif de leurs vaches. Des études ont démontré que des soins appropriés, tels que des brossages et une attention particulière, peuvent augmenter la production laitière de jusqu'à 1 kg par jour et réduire jusqu'à 30% les risques de développement de mammites, une inflammation de la glande mammaire (Producteurs de Lait du Québec, 2014). Ces considérations soulignent l'importance d'une approche holistique pour assurer la santé et la durabilité de la production laitière. Cependant, il est important de noter que certaines pratiques d'élevage peuvent soulever des préoccupations, telles que la séparation précoce des veaux de leurs mères. La santé animale influe sur la production de lait, car des soins vétérinaires réguliers et un traitement rapide des maladies contribuent à réduire le risque d'infections telles que la mammite clinique, à garantir le bien-être des vaches laitières et à maintenir les niveaux de production de lait (Kares, 2022). (Carvalho *et al.*, 2019) montre que lorsque les vaches développent plusieurs maladies métaboliques, leur production annuelle de lait diminue de 703 kg, de gras de

27kg et de protéines de 19 kg sur une période de 305 jours. De plus, ces maladies ont été associées à une augmentation des problèmes de reproduction chez les vaches. Plusieurs études ont mis en lumière l'impact négatif des maladies de transition, comme la métrite ou la cétose, sur le bien-être et la rentabilité des vaches laitières (LEBLANC, 2010). Par exemple, (Macmillan *et al.*, 2020) ont observé qu'après le vêlage, près de 40% des vaches malades développent plus d'une maladie, soulignant ainsi la complexité des problèmes de santé chez ces animaux.

impact économique : La santé des animaux peut également avoir un impact économique sur les producteurs laitiers. Les maladies animales peuvent entraîner une mortalité et une réduction de la productivité des troupeaux laitiers, ce qui peut entraîner des pertes économiques importantes pour les producteurs, en particulier pour les petites exploitations laitières. Les affections animales entraînent une hausse du taux de mortalité et une réduction de la productivité au sein des troupeaux laitiers à l'échelle mondiale, engendrant ainsi d'importantes pertes économiques (Fao, 2024).

(Dillon et Hennessy, 2012) affirme que l'amélioration des pratiques de santé animale peut entraîner des gains en termes de coûts et d'efficacité de la production dans les exploitations laitières irlandaises. il ajoute que la réduction du nombre de cellules somatiques (SCC) dans le lait peut augmenter la marge brute de 6%, soit 87€ par vache. Par exemple, une équipe de chercheurs basée aux États-Unis a fourni des estimations des coûts associés à chaque cas de dermatite digitale. Selon leurs recherches, chaque cas entraîne une perte de 49$ en production laitière, une réduction de 58$ en termes de fertilité, et des dépenses de traitement s'élevant à 79$. En agrégeant ces chiffres, le coût total par vache infectée s'élève à 186$ par année. Pour un troupeau laitier canadien moyen, supposant un effectif de 100 vaches en lactation, cela implique des coûts annuels compris entre 2 790$ et 4092$ (Les Producteurs laitiers du Canada,).

Dans ce chapitre, nous avons présenté les maladies présentes dans notre jeu de données, défini le concept de signatures biomarqueurs et mis en évidence les impacts de la santé animale sur la production laitière. Ces éléments fondamentaux établissent une base solide pour comprendre l'importance de l'analyse des données en santé animale. Dans le chapitre suivant, nous approfondirons les concepts d'analyses statistiques et d'apprentissage automatique, outils essentiels pour extraire et interpréter les signatures biomarqueurs identifiées dans ce chapitre. Cette transition marque une étape clé dans notre démarche d'analyse et de prédiction des maladies métabolique

CHAPITRE 2

CONCEPTS D'ANALYSES STATISTIQUES ET D'APPRENTISSAGE AUTOMATIQUE

Dans ce chapitre, nous mettrons en relief les concepts de distribution statistique et les méthodes d'analyse. Ensuite, nous aborderons l'ingénierie des caractéristiques. Enfin, nous explorerons les différents concepts de l'apprentissage automatique.

2.1 Méthodes d'analyse statistique

2.1.1 Types de distribution statistiques

L'analyse statistique est une composante essentielle dans l'étude et l'interprétation des données. Elle offre des outils puissants pour comprendre les tendances, détecter les modèles et prendre des décisions éclairées. Dans cette section, nous explorerons brièvement différentes méthodes d'analyse statistique qui jouent un rôle crucial dans divers domaines. En statistique, les distributions jouent un rôle fondamental pour modéliser le comportement des données. Parmi les distributions les plus couramment utilisées, nous examinerons notamment les distributions normale, Bernoulli et binomiale. Chacune de ces distributions présente des caractéristiques uniques et trouve des applications spécifiques dans l'analyse statistique.

Normale : La distribution normale, communément appelée distribution gaussienne, est l'une des distributions de probabilité les plus utilisées dans le domaine des statistiques(Nadarajah, 2005). Il existe un grand nombre de façons de trouver des fonctions de densité théoriques qui peuvent correspondre de manière efficace aux données recueillies. La distribution gaussienne (normale) est le plus souvent supposée pour décrire la variation aléatoire qui se produit dans les données provenant de nombreuses disciplines scientifiques. La courbe en forme de cloche bien connue peut facilement être caractérisée et décrite par deux valeurs : la moyenne arithmétique μ et l'écart-type σ, de sorte que les ensembles de données sont couramment décrits par l'expression ($\mu \pm \sigma$) (Limpert *et al.*, 2001).

Comme illustré à la Figure 2.1, la distribution normale avec $\mu = 0$ et $\sigma = 1$ est présentée.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.1}$$

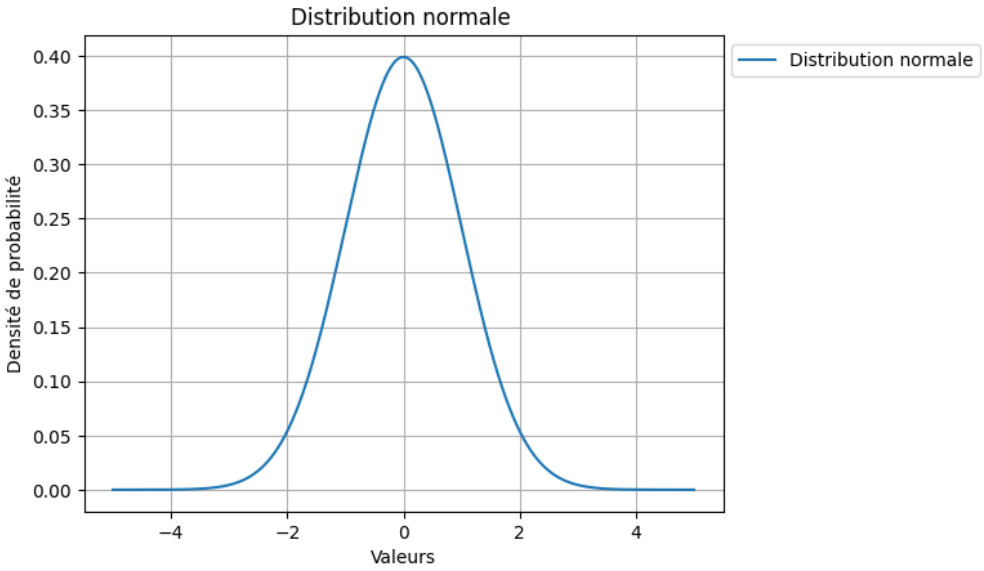

Figure 2.1 – La distribution normale pour $\mu = 0$ et $\sigma = 1$

Discrètes : Les distributions discrètes sont des modèles mathématiques qui décrivent la probabilité d'occurrence de différentes valeurs discrètes pour une variable aléatoire (Sinharay, 2023). Contrairement aux distributions continues, qui peuvent prendre n'importe quelle valeur dans un intervalle, les distributions discrètes sont définies uniquement pour des valeurs distinctes. Parmi les exemples courants de distribution discrète, on trouve la distribution de Bernoulli et la distribution binomiale. Ces deux distributions sont particulièrement pertinentes et largement utilisées dans divers domaines pour modéliser des phénomènes aléatoires discrets.

Bernoulli : La distribution de Bernoulli est une distribution discrète caractérisée par deux résultats possibles, notés $n = 0$ et $n = 1$ (Johnson, 1969). Dans ce contexte, le résultat $n = 1$, représentant le "succès", survient avec une probabilité p, tandis que le résultat $n = 0$, indiquant "l'échec", se produit avec une probabilité $q = 1 - p$, où $0 < p < 1$. La fonction de densité de probabilité associée s'exprime comme suit :

$$P(n) = \begin{cases} 1 - p & \text{si } n = 0, \\ p & \text{si } n = 1. \end{cases} \tag{2.2}$$

Cette formulation peut également être réécrite sous la forme

$$P(n) = p^n(1-p)^{1-n} \tag{2.3}$$

Binomiale : La distribution binomiale est une distribution de probabilité discrète qui modélise le nombre de succès dans un nombre fixe d'essais indépendants, chacun ayant deux résultats possibles (généralement étiquetés comme succès ou échec), et avec la même probabilité de succès à chaque essai (Mathworld, 2024). Elle est souvent utilisée dans des situations où des événements indépendants avec une probabilité constante de succès se produisent de manière répétée.

La fonction de masse de probabilité (pmf) de la distribution binomiale est donnée par :

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.4}$$

où n est le nombre total d'essais, k est le nombre de succès, p est la probabilité de succès dans un essai, $(1-p)$ est la probabilité d'échec dans un essai, et $\binom{n}{k}$ est le coefficient binomial, représentant le nombre de façons de choisir k succès parmi n essais.

La moyenne (espérance) de la distribution binomiale est $\mu = np$ et la variance est $\sigma^2 = np(1-p)$.

2.1.2 Interprétation des analyses statistiques

Dans cette section, nous examinerons de près certaines des mesures et des concepts clés utilisés dans l'interprétation des résultats statistiques. Chacun de ces éléments fournit des informations précieuses pour comprendre la signification des données analysées.

Ecart-type : l'écart type est une mesure de variabilité qui fournit au chercheur et au lecteur des informations sur la dispersion des scores, permettant une interprétation plus approfondie de la moyenne des accords entre observateurs (Carr *et al.*, 1996). Il nous permet de quantifier la variabilité des données et de comprendre à quel point les observations s'éloignent de la valeur moyenne. Une faible déviation standard suggère une concentration étroite des données autour de la moyenne, tandis qu'une déviation standard

élevée indique une dispersion plus importante.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{2.5}$$

— $\bar{x}$ est la moyenne,

— x_i sont les observations individuelles,

— n est le nombre d'observations.

P-value : La valeur de p, ou p-value, est une mesure cruciale lors de l'évaluation de l'importance statistique d'une observation. Sous l'hypothèse nulle, la valeur de p, qui est calculée à partir d'un test statistique continu, est uniformément répartie dans l'intervalle [0, 1], quelle que soit la taille de l'échantillon ; En revanche, lorsque l'on considère l'hypothèse alternative, la distribution de la valeur de p dépend à la fois de la taille de l'échantillon et des valeurs réelles ou de la plage de valeurs réelles du paramètre testé (Hung *et al.*, 1997). Elle indique la probabilité d'observer les résultats actuels ou plus extrêmes si l'hypothèse nulle est vraie. Une valeur de p faible (généralement < 0,05) suggère une forte évidence contre l'hypothèse nulle, ce qui peut conduire au rejet de cette dernière.

$$p\text{-value} = P(\text{Observation aussi extrême que celle observée} \mid \text{Hypothèse nulle est vraie}) \tag{2.6}$$

Odds-ratio : L'odds ratio (rapport des cote) quantifie la relation entre deux occurrences, généralement dans le domaine de la recherche épidémiologique. Il définit la probabilité proportionnelle qu'un événement se produise par rapport à un autre. Une valeur de rapport de cotes supérieure à 1 signifie une connexion positive, tandis qu'une valeur inférieure à 1 signifie une négative défavorable (Szumilas, 2010).

Les rapports de cotes servent à évaluer les cotes relatives de l'apparition d'un résultat particulier (tel qu'une maladie ou un trouble) en fonction de l'exposition à une variable spécifique (comme une caractéristique de santé ou un aspect de l'historique médical) (Szumilas, 2010). Ils permettent également d'établir si une exposition donnée constitue un facteur de risque pour un résultat particulier et de comparer l'importance de différents facteurs de risque associés à ce résultat (Szumilas, 2010).

L'association entre l'exposition et l'issue est évaluée à l'aide du rapport des cotes. Le tableau 2.1 ci-dessous montre le statut des résultats et de l'exposition, et l'équation 2.7 présente comment le rapport des cotes est calculé.

$$\text{OR} = \frac{\text{a} \times \text{d}}{\text{b} \times \text{c}} \tag{2.7}$$

— a : Nombre de cas exposés

Statut d'exposition	Statut des résultats +	Statut des résultats -
+	a	b
-	c	d

Table 2.1 – Rapport des cotes

— b : Nombre de cas non exposés

— c : Nombre de témoins exposés

— d : Nombre de témoins non exposés

Intervalle de confiance : L'intervalle de confiance (IC) est une plage de valeurs qui fournit une estimation de la précision d'une mesure statistique. Généralement, un intervalle de confiance est construit autour d'une estimation ponctuelle, telle qu'une moyenne ou un rapport de cotes, et il indique le niveau de confiance dans lequel la vraie valeur de la mesure pourrait se situer. Un IC large indique un faible niveau de précision de l'OR, tandis qu'un IC étroit indique une plus grande précision de l'OR. Il est important de noter cependant que, contrairement à la valeur de p, l'IC à 95% ne rend pas compte de la signification statistique d'une mesure.

Dans la pratique, l'intervalle de confiance à 95% est fréquemment utilisé comme un niveau de signification statistique si ses limites n'incluent pas la valeur nulle (par exemple, OR=1). Il serait incorrect de conclure à l'absence d'association entre l'exposition et le résultat uniquement sur la base d'un intervalle de confiance à 95% qui englobe la valeur nulle (Szumilas, 2010). Et les intervalles sont calculés comme suit 2.8 pour les limites de l'intervalle de confiance.

$$\begin{aligned} \text{Upper 95\% CI} &= \exp\left[\ln(\text{OR}) + 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right] \\ \text{Lower 95\% CI} &= \exp\left[\ln(\text{OR}) - 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right] \end{aligned} \tag{2.8}$$

2.2 Ingénierie des caractéristiques

L'ingénierie des caractéristiques consiste à améliorer la qualité des prédictions d'un modèle en modifiant les caractéristiques des données, généralement en explorant différentes transformations ou en ajoutant de nouvelles fonctionnalités, puis en sélectionnant les plus pertinentes (Nargesian *et al.*, 2017). Cette approche permet de créer des caractéristiques additionnelles tant pour l'apprentissage supervisé et non supervisé,

dans le but de simplifier et accélérer les transformations de données tout en améliorant la précision du modèle. (Popov, 2023).

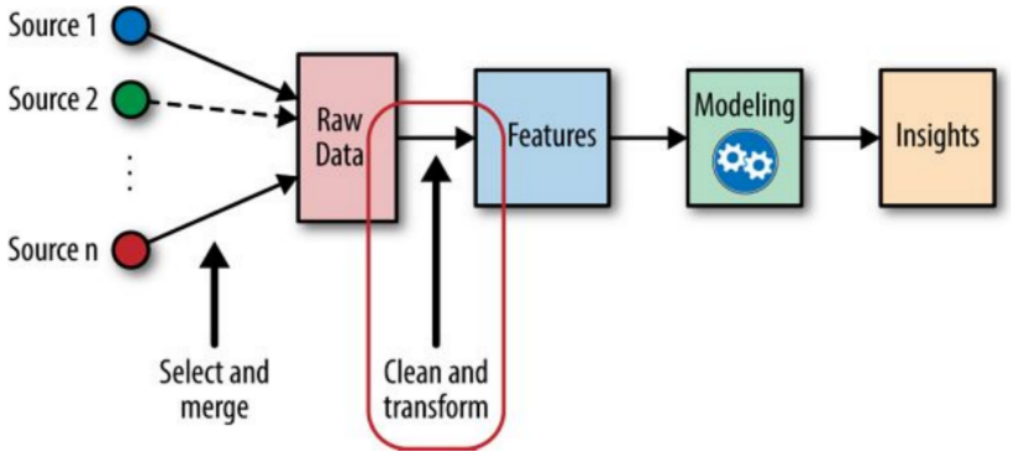

Figure 2.2 – La place de l'ingénierie des fonctionnalités de l'apprentissage automatique. Source :(Zheng et Casari, 2018)

Les méthodes d'ingénierie des caractéristiques pour les données relatives aux sciences de la vie et de l'agriculture font appel à diverses techniques pour préparer et extraire des caractéristiques significatives à partir des données brutes (Dineva et Atanasova, 2023).

Il existe différentes méthodes d'ingénierie des caractéristiques telles que : l'imputation, le traitement des valeurs aberrantes, la transformation logarithmique, l'encodage one-hot et la mise à l'échelle. Dans ce qui suit, nous présenterons ces méthodes et comment nous les avons utilisées dans notre méthodologies.

2.2.1 Imputation

Les valeurs manquantes sont fréquentes, résultant d'erreurs humaines, d'interruptions de flux de données, de conflits potentiels liés à la confidentialité et d'autres facteurs. Ces lacunes peuvent impacter les performances des modèles d'apprentissage automatique. L'imputation implique de remplacer les données absentes par des estimations statistiques. Son but est de générer un ensemble de données complet utilisable pour former des modèles d'apprentissage automatique. L'imputation moyenne/médiane remplace chaque valeur manquante (NA) dans une variable par la moyenne (pour une distribution gaussienne) ou la médiane (pour une distribution asymétrique) (Kunwar, 2020).

Index	Fat
1	low
2	medium
3	high

onehot →

Index	fat_low	fat_medium	fat_high
1	1	0	0
2	0	1	0
3	0	0	1

Figure 2.3 – Comparaison avant et après le codage one-hot.

2.2.2 One-hot encoding

L'encodage one-hot est une technique utilisée pour transformer les données avant de les utiliser dans un algorithme afin d'améliorer les prédictions.

Il consiste à créer une nouvelle colonne pour chaque catégorie, assignant une valeur binaire de 1 ou 0 à chaque colonne, où une seule colonne contient la valeur 1 correspondant à la catégorie de l'observation, tandis que les autres colonnes ont la valeur 0 (Educative, 2015).

Dans l'exemple ci-dessus 2.3 , nous appliquons un encodage One Hot à une caractéristique (Fat) qui possède trois catégories (faible, moyen et élevé).

2.2.3 La mise à l'échelle

Cette approche de l'ingénierie des caractéristiques est employée lorsqu'il y a une sensibilité excessive des algorithmes à certaines échelles de données. Le procédé min-max implique de redimensionner les valeurs pour les placer dans un intervalle allant de 0 à 1, ce qui revient à les normaliser (Filzinger, 2023).

— Normalisation : Les valeurs sont normalisées dans une plage spécifiée entre 0 et 1, via la normalisation (ou normalisation min-max), sans altérer la distribution de la caractéristique. Un impact disproportionné des valeurs aberrantes en raison des écarts-types réduits, d'où la recommandation de traiter les valeurs aberrantes préalablement à la normalisation(Ahmed *et al.*, 2022).

— standardisation :La standardisation (ou normalisation z-score) est le processus de mise à l'échelle des valeurs tout en tenant compte de la variabilité. Cela permet d'ajuster les plages des caractéristiques, même si leurs écarts types diffèrent. La standardisation réduit l'impact des valeurs extrêmes. Pour obtenir une distribution avec une moyenne de 0 et une variabilité de 1, on soustrait la moyenne de chaque point de données et on divise le résultat par l'écart type de la distribution (Alam, 2020).

2.2.4 Technique de Suréchantillonnage Synthétique pour les Minorités (SMOTE)

Le SMOTE, Technique de Suréchantillonnage Synthétique pour les Minorités (Chawla *et al.*, 2002), est une approche pour augmenter les observations de la classe minoritaire. Au lieu de simplement cloner les instances minoritaires existantes, le SMOTE adopte une stratégie différente : il crée de nouveaux exemples minoritaires qui partagent des caractéristiques similaires avec les instances existantes, mais qui ne sont pas exactement identiques. Cette méthode permet d'accroître la densité de la population des instances minoritaires de manière plus uniforme. Pour créer un individu synthétique, les étapes définies dans l'algorithme du SMOTE sont les suivantes (Tremblay, 2022) :

1. Sélectionner aléatoirement une observation minoritaire "initiale".
2. Identifier ses k plus proches voisins parmi les observations minoritaires.
3. Choisir aléatoirement l'un des k plus proches voisins.
4. Générer aléatoirement un coefficient λ.
5. Créer un nouvel individu entre l'observation initiale et le plus proche voisin choisi, selon la valeur du coefficient λ.

On répète ces étapes jusqu'à ce qu'un nombre spécifié par l'utilisateur d'individus générés soit atteint.

2.3 Apprentissage automatique

L'apprentissage automatique (machine learning) est un domaine de l'informatique qui se concentre sur le développement de techniques et d'algorithmes permettant aux systèmes informatiques d'apprendre à partir de données, de s'améliorer de manière autonome et d'effectuer des tâches sans être explicitement programmés (Géron, 2020). L'apprentissage automatique implique la création de modèles et d'algorithmes qui peuvent détecter des motifs, des relations ou des tendances dans les données, et utiliser ces informations pour prendre des décisions ou effectuer des prédictions (Géron, 2020).

Les domaines d'application de l'apprentissage automatique sont vastes, allant de la classification d'images (Bors et Pitas, 1999) à la prédiction de séries temporelles (Wagner *et al.*, 2021), la traduction automatique (Loffler-Laurian, 1996), la santé (Bohnsack *et al.*, 2023), la finance (Ghoddusi *et al.*, 2019), et bien d'autres (Ye *et al.*, 2019). Il offre un potentiel énorme pour automatiser des processus, améliorer la prise de décision, et développer des applications intelligentes.

L'apprentissage automatique peut être divisé en deux grands groupes qui sont : l'apprentissage supervisé et l'apprentissage non supervisé (nous ne traiterons pas l'apprentissage par renforcement dans ce mémoire). L'apprentissage supervisé se scinde en deux grands types de problèmes, appelés classification et régression. Dans le contexte de l'apprentissage non supervisé, on peut identifier des méthodes de transformation des données ainsi que des approches de regroupement (clustering).

2.3.1 Apprentissage supervisé

L'apprentissage supervisé désigne une méthodologie d'apprentissage automatique dans laquelle les algorithmes acquièrent des connaissances à partir de données d'entraînement annotées dans le but de fournir des prévisions ou des décisions basées sur des variables d'entrée. Dans le domaine de l'apprentissage supervisé, l'algorithme est doté d'un ensemble de paires d'entrées et de sorties. Les variables d'entrée représentent les attributs, et la variable de sortie représente le résultat ou le label souhaité que l'algorithme s'efforce de prédire (Bzdok *et al.*, 2018). L'objectif de l'apprentissage supervisé est d'extraire des principes ou des modèles généraux à partir de données d'entraînement annotées, qui peuvent ensuite être utilisés pour établir des pronostics ou catégoriser de nouvelles données. Les algorithmes d'apprentissage supervisé, tels que les machines à vecteurs de support linéaires (SVM) et les k-plus proches voisins (KNN), sont couramment utilisés dans divers domaines, notamment la biologie et la médecine, pour extraire des modèles et faire des prédictions sur la base de données d'entraînement étiquetée (Bzdok *et al.*, 2018; Osisanwo *et al.*, 2017).

Régression linéaire La régression linéaire est une méthode statistique qui vise à créer un modèle qui lie une variable dépendante (Y)avec ou plusieurs variables indépendantes (X) à l'aide d'une équation linéaire (Géron, 2020).

L'objectif de la régression linéaire est de trouver la meilleure ligne droite (ou hyperplan) qui représente au mieux la relation entre les variables [1]. Un modèle linéaire réalise des prédictions en effectuant une somme pondérée des variables d'entrée, à laquelle s'ajoute un terme constant (Géron, 2020). L'équation 2.9 montre comment la prédiction d'un modèle de regression linéaire est calculée

1.

Prédiction d'un modèle de régression linéaire.

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n \tag{2.9}$$

— Y est toujours la variable dépendante à prédire.
— n représente le nombre de variables.
— X_i est la valeur de la i$^{\text{ème}}$ variable.
— a_j est le j$^{\text{ème}}$ paramètre du modèle.

Régression logistique : La régression logistique est une méthodologie statistique utilisée pour construire un modèle trouvant l'association entre une variable dépendante binaire, caractérisée par deux catégories, et une série de variables indépendantes. Contrairement à la régression linéaire, qui vise à prévoir des résultats continus, la régression logistique est spécifiquement conçue pour estimer la probabilité qu'un événement particulier se produise (Géron, 2020).

La régression logistique est particulièrement utile pour la classification binaire, où on essaye de prédire si un événement sera observé (1) ou non (0) en fonction des caractéristiques d'un ensemble de données. Elle repose sur une fonction logistique (ou sigmoïde) pour modéliser la probabilité de l'événement en fonction des variables indépendantes.

De la même manière que la régression linéaire, un modèle de régression logistique effectue une somme pondérée des caractéristiques en entrée, mais au lieu de produire directement le résultat, il génère la logistique du résultat (Géron, 2020).

k-plus proches voisins (k-NN) : Le k-plus proches voisins, souvent abrégé K-NN, est un algorithme simple et largement utilisé en apprentissage automatique pour des tâches de classification et de régression. La classification ou l'évaluation d'un point de données non classé est établie en tenant compte du consensus dominant ou de la moyenne des k voisins les plus proches dans l'ensemble de données d'apprentissage (Bzdok *et al.*, 2018).

L'illustration 2.4 montre le fonctionnement de l'algorithme k-NN, où l'objectif est de classer le nouvel élément, représenté en rouge, en classe A ou en classe B. Lorsque nous considérons les trois plus proches voisins (k = 3), la classe majoritaire parmi ces trois éléments voisins est la classe B. Par conséquent, dans ce

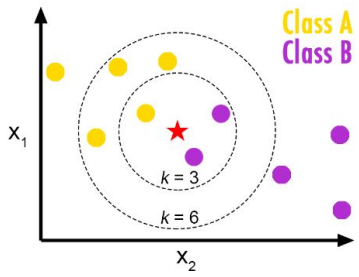

Figure 2.4 – Fonctionnement du k-NN. Source :(Genesis, 2018)

scénario, le nouvel élément sera classé dans la classe B. Sinon, si k = 6, le nouvel élément sera classé dans la classe A.

L'algorithme fait preuve de robustesse en présence de valeurs aberrantes, tout en conservant un cadre conceptuel simple (Genesis, 2018; Gandhi, 2018). La technique des k plus proches voisins repose de manière similaire sur les k échantillons d'apprentissage les plus proches de la nouvelle entrée x, en utilisant une mesure de distance telle que la distance euclidienne ou de Minkowski.

$$\text{Distance Euclidienne} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.10}$$

$$\text{Distance de Minkowski} = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}} \tag{2.11}$$

SVM : Support Vector Machines (SVM) est un modèle de classification qui vise à trouver la ligne ou l'hyperplan optimal pour séparer deux classes dans un ensemble de données qui sont linéairement séparables (Bisong, 2019) comme illustré dans la figure 2.5. Il extrait des principes généraux à partir d'exemples observés pour faire des prédictions basées sur un objectif de prédiction spécifique (Bzdok *et al.*, 2018).

Abre de décision : Un arbre de décision peut être décrit comme un modèle à base de règles qui opère en utilisant un processus de raisonnement inductif. Ce modèle analyse des exemples de données fournis, apprend des règles implicites à partir de ces exemples, et utilise ces règles pour prendre des décisions sur de nouvelles données comme le montre la figure 2.6 (Quinlan, 1986; Safavian et Landgrebe, 1991). L'attribut le plus crucial est initialement identifié en utilisant des méthodes telles que l'entropie et l'indice de

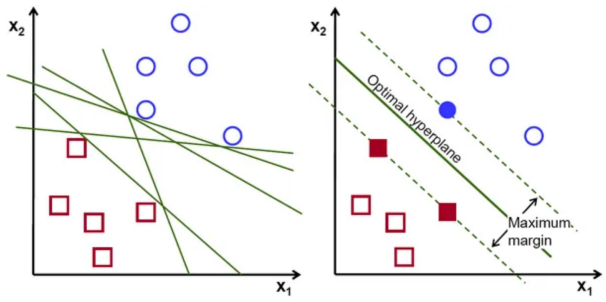

Figure 2.5 – Fonctionnement du SVM (gauche :hyper-plans non optimaux. droite : hyper-plan donne par le SVM : Source :(Gandhi, 2018)

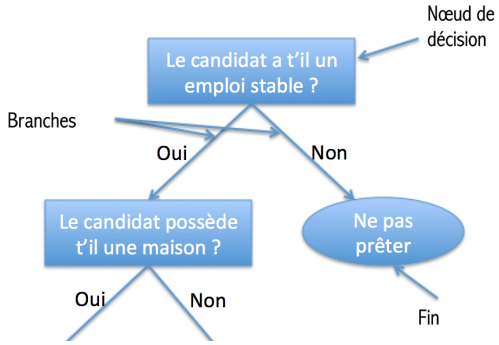

Figure 2.6 – Exemple d'arbre de décision. Source :(M@XCode, 2016)

Gini (Raileanu et Stoffel, 2004). Ces métriques permettent de mesurer la pureté des divisions potentielles, aidant ainsi à déterminer la variable la plus discriminante pour le partitionnement des données.

Soit un nœud n et p(j/n) la fréquence relative de la classe j au nœud n alors l'entropie et l'indice de gini sont définis comme suit.

$$Entropie(n) = -\sum_{j} p(j|n) \log_2 p(j|n) \tag{2.12}$$

$$Gini(n) = 1 - \sum_{j} [p(j|n)]^2 \tag{2.13}$$

Forêt aléatoire : Le Random Forest est un algorithme d'ensemble utilisé dans des tâches de classification et de régression. Il repose sur l'intégration de multiples arbres de décision afin de générer des prédictions précises. Chaque arbre de décision est construit en utilisant un sous-ensemble des données d'entraînement ainsi qu'une sélection aléatoire de caractéristiques. Ensuite, l'algorithme fusionne les prédictions de tous ces arbres de décision pour produire la prédiction finale comme l'illustre la figure 2.7. Le Random Forest

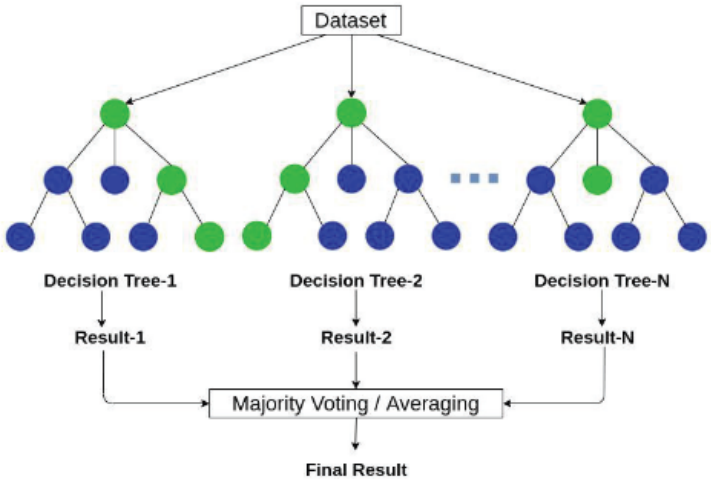

Figure 2.7 – Fonctionnement de l'algorithme Random Forest Source :(Paul *et al.*, 2022)

est renommé pour sa haute précision, sa robustesse et sa capacité à gérer les données manquantes. Il résout le problème de surajustement en moyennant les prédictions de l'ensemble des arbres. De plus, il est relativement rapide en termes de temps de calcul par rapport à d'autres techniques qui n'utilisent que des données structurées (Paul *et al.*, 2022).

Réseaux de neurones : Les réseaux de neurones artificiels sont des modèles d'apprentissage automatique conçus pour reproduire le fonctionnement des réseaux de neurones naturels du cerveau humain. Ils peuvent présenter une grande complexité en utilisant plusieurs couches successives de neurones (Sarle, 1994). Les réseaux de neurones artificiels sont un genre d'algorithme d'apprentissage supervisé qui se base sur des fonctions d'entrée et d'activation, ainsi que sur la structure du réseau et l'importance des liaisons entre les entrées (Osisanwo *et al.*, 2017). Le réseau de neurones artificiel le plus simple est le perceptron (Sanger et Baljekar, 1958). Le perceptron se constitue d'un unique élément de traitement, appelé neurone, qui reçoit des entrées pondérées et génère une sortie en fonction d'une fonction d'activation préalablement définie.

Dans la figure 2.8 le perceptron (représenté par un cercle) prend en entrée des données x1, x2 et x3 et produit une valeur de sortie.

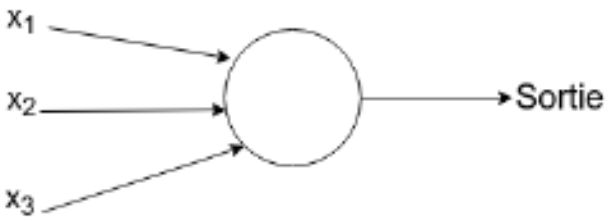

Figure 2.8 – Exemple de perceptron

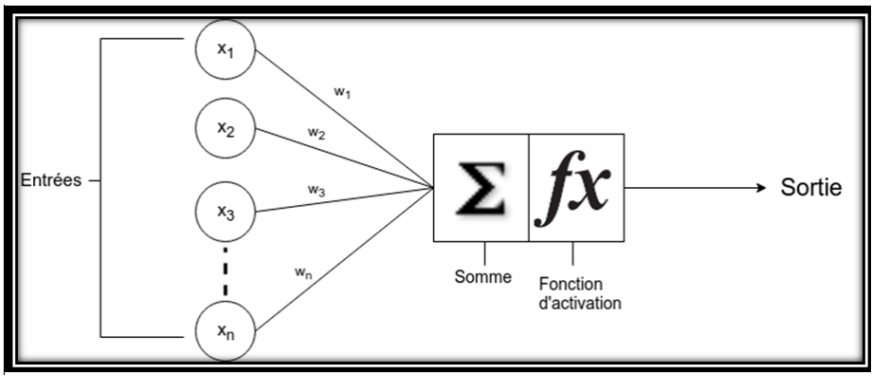

Figure 2.9 – Traitement à l'intérieur du neurone

.

Le neurone effectue deux étapes distinctes. La première étape consiste à effectuer une somme pondérée des informations reçues, ce qui implique d'additionner les valeurs d'entrée après les avoir multipliées par leurs poids respectifs. Ensuite, une fonction appelée fonction d'activation est appliquée à cette somme. En fonction de la valeur obtenue, le neurone peut s'activer et générer une valeur en sortie, comme illustré dans la Figure 2.9 .

Il existe de nombreuses fonctions d'activation qui peuvent être utilisées. À titre d'exemple, une fonction d'activation simple pourrait être la fonction "malade". Cette fonction renvoie 0 en absence de maladie et 1 en présence de maladie.

Un réseau de neurones est composé de plusieurs perceptrons, ce qui équivaut à plusieurs neurones et plusieurs couches de neurones. Pour obtenir une valeur de sortie dans un réseau de neurones, l'information pénètre par la couche d'entrée, et chaque neurone de chaque couche génère une valeur qui est transmise aux neurones des couches suivantes. Ce processus se répète jusqu'à atteindre la dernière couche, qui peut

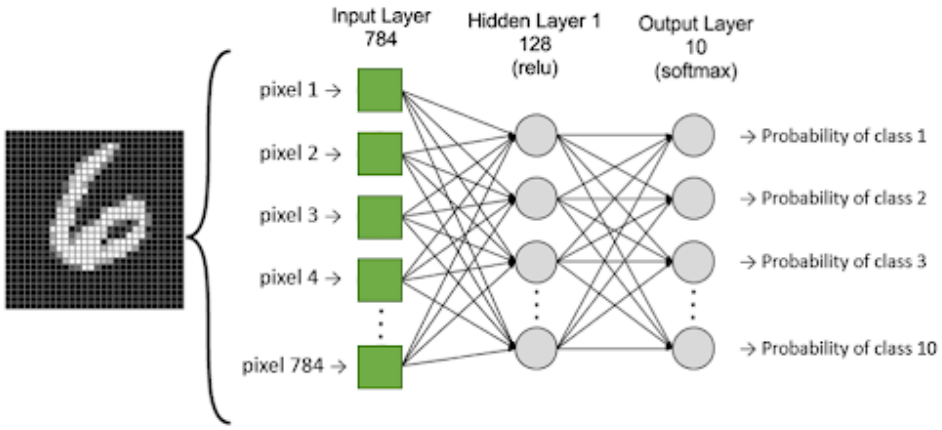

Figure 2.10 – Un exemple de réseau de neurones à trois couches dont une couche cachée. Source :(Kassel, 2020)

être composée d'un ou de plusieurs neurones, et elle fournit la valeur de sortie, comme illustré dans la fgure 2.10. Les différentes couches jouent un rôle essentiel dans la transformation de l'information avant d'aboutir à la sortie finale du réseau.

2.3.2 Apprentissage non-supervisé

L'apprentissage non supervisé constitue une catégorie d'apprentissage automatique où l'algorithme acquiert des modèles et identifie des relations entre les données sans recourir à des exemples étiquetés ni à des directives fournies par un superviseur humain (Sammut et Webb, 2017).

Voici les algorithmes les plus importants en apprentissage non supervisé (Aurélien *et al.*, 2017).

— Partition (k-moyennes, DBSCAN, Partition hiérarchique)
— Détection d'anomalies et détection de nouveautés (SVM à une classe, Forêt d'isolation)
— Visualisation et réduction de dimension (Analyse en composantes principales, Analyse en composantes principales à noyaux, Méthode t-SNE)
— Apprentissage de règles d'association (Apriori, Eclat)

2.3.3 Evaluation de performance

Afin de déterminer l'efficacité d'un modèle de classification, c'est-à-dire sa capacité à bien classifier un ensemble de données indépendant de celui utilisé pour l'entraînement, il est essentiel d'appliquer des mé-

thodes et des mesures d'évaluation appropriées. Les métriques d'évaluation des classifieurs sont des outils essentiels pour évaluer l'efficacité d'un modèle de classification. Il est important de noter que l'évaluation d'un classifieur est généralement plus complexe que celle d'un régresseur. (Aurélien *et al.*, 2017).

2.3.3.1 Méthodes d'évaluation

Validation (Entraînement / Test) : Cette méthode implique de diviser un ensemble de données initial en deux groupes distincts. L'un de ces sous-groupes est utilisé pour entraîner un modèle d'apprentissage, appelé ensemble d'entraînement, tandis que l'autre est réservé pour tester la performance du modèle, appelé ensemble de test.

Validation croisée k : La validation croisée est une méthode statistique plus rigoureuse et approfondie que la simple division en ensembles d'entraînement et de test. Elle implique de diviser l'ensemble des données en k partitions disjointes approximativement équivalentes. Ensuite, des cycles d'entraînement et de test sont effectués k fois de manière alternative. À chaque itération, k-1 partitions sont utilisées pour former un modèle d'entraînement, tandis que la dernière partition est réservée comme ensemble de test.

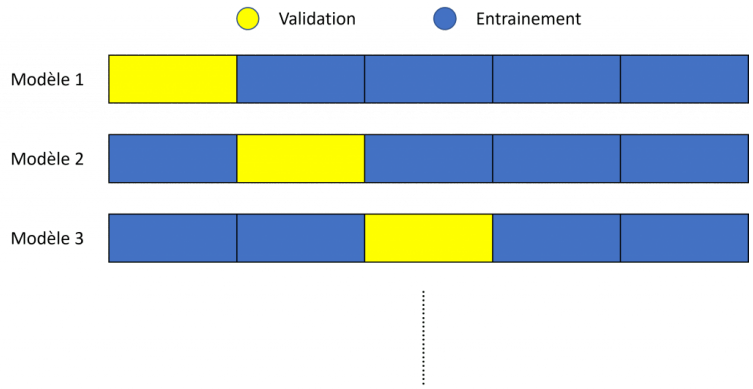

Figure 2.11 – Schéma de validation croisée.Source :(kwamimayeden, 2022)

Bootstrap : Cette méthode consiste à prélever des échantillons de données de manière aléatoire à partir d'un ensemble initial, en permettant le remplacement des observations prélevées. Cette opération est répétée n fois pour créer plusieurs ensembles de données échantillonnées. À chaque itération i, un ensemble bootstrap $boot_i$ est créé, sur lequel un modèle est entraîné, et un ensemble de test $test_i$ est constitué pour évaluer le modèle d'entraînement. Cette méthode permet ainsi d'avoir des objets similaires présents à la fois dans l'ensemble d'entraînement et dans l'ensemble de test.

Jackknife : Cette méthode est semblable au Bootstrap, à ceci près qu'aucun échantillonnage avec remplacement n'est effectué lors de la constitution des ensembles d'entraînement et de test à chaque itération.

Par conséquent, il n'est pas possible d'avoir des instances identiques à la fois dans l'ensemble d'entraînement et dans l'ensemble de test.

2.3.3.2 Métriques d'évaluation des performances

Pour illustrer diverses mesures d'évaluation des performances de la classification, considérons un ensemble de données réparti en deux classes : une classe '1' et une classe '0'. Les diverses prédictions qu'un modèle pourrait faire sur ce type de données incluraient :

Matrice de confusion : La matrice de confusion constitue un instrument d'évaluation de l'efficacité des modèles de classification, lorsqu'ils traitent deux classes ou plus. Chaque colonne de la matrice représente une classe réelle, tandis que chaque ligne représente une classe prédite (Aurélien *et al.*, 2017).

	Classe 1	Classe 0
1	Vrai négatif (VN)	Faux négatif (FN)
0	Faux positif (FP)	Vrai positif (VP)

Table 2.2 – Matrice de confusion

Taux de faux Positifs (TFP) : C'est la proportion d'exemples négatifs incorrectement classés. Il se calcule en divisant le nombre de faux positifs (FP) par la somme des vrais négatifs (VN) et des faux positifs (FP).

$$TFP = \frac{FP}{VN + FP} \tag{2.14}$$

Taux de faux négatifs (TFN) : C'est la proportion d'exemples positifs incorrectement classés. Il se calcule en divisant le nombre de faux négatifs (FN) par la somme des vrais positifs (VP) et des faux négatifs (FN)

$$TFN = \frac{FN}{VP + FN} \tag{2.15}$$

Taux de vrais positifs (TVP) : C'est la proportion d'exemples positifs correctement classifiés, également connue sous le nom de sensibilité. Il se calcule en divisant le nombre de vrais positifs (VP) par la somme des vrais positifs (VP) et des faux négatifs (FN)

$$TVP = \frac{VP}{VP + FN} \tag{2.16}$$

Taux de vrais négatifs (TVN) : C'est la proportion d'exemples négatifs correctement classifiés, également connue sous le nom de spécificité. Il se calcule en divisant le nombre de vrais négatifs (VN) par la somme des vrais négatifs (VN) et des faux positifs (FP)

$$TVN = \frac{VN}{VN + FP} \tag{2.17}$$

Rappel : Le rappel évalue la capacité d'un modèle de classification à détecter toutes les instances pertinentes d'un ensemble de données. Il représente la proportion entre les vrais positifs (VP) et la somme des vrais positifs (VP) et des faux négatifs (FN) (Bellet *et al.*, 2015).

$$\text{Rappel} = \frac{\text{VP}}{\text{VP + FN}} \tag{2.18}$$

Précision La précision, en revanche, évalue la capacité d'un modèle de classification à identifier exclusivement les instances pertinentes. Elle correspond à la proportion entre les vrais positifs (VP) et la somme des vrais positifs(VP)et des faux positifs (FP) (Bellet *et al.*, 2015).

$$\text{Précision} = \frac{\text{VP}}{\text{VP + FP}} \tag{2.19}$$

F-mesure : La F-mesure est la moyenne harmonique de la précision et du rappel. Il s'agit d'une mesure qui résiste aux variations dans la distribution des classes dans le jeu de données (présence de maladie ou non dans notre cas) (Sasaki, 2007). La F-mesure repose sur la précision ainsi que sur le rappel (Buckland et Gey, 1994).

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision + Rappel}} \tag{2.20}$$

Exactitude (Accuracy) : La mesure de l'exactitude (accuracy) est généralement utilisée pour évaluer la performance globale du modèle sur l'ensemble des classes. Elle s'avère utile lorsque toutes les classes ont la même importance. Son calcul repose sur le rapport entre le nombre de prédictions correctes et le nombre total de prédictions (Bellet *et al.*, 2015).

$$\text{Exactitude} = \frac{\text{VP + VN}}{\text{VP + VN + FP + FN}} \tag{2.21}$$

La courbe ROC : L'analyse ROC (Receiver Operating Characteristic) évalue la précision des prévisions d'un modèle en illustrant la sensibilité par rapport au taux de faux positifs d'un test de classification (Aurélien

et al., 2017). Un moyen de comparer les classificateurs est de quantifier l'étendue sous la courbe (Area Under the Curve ou AUC). Un classificateur idéal aurait une aire sous la courbe ROC (ROC AUC) égale à 1, alors qu'un classificateur complètement aléatoire aurait une ROC AUC de 0,5 (Aurélien *et al.*, 2017). La courbe ROC, très similaire à la courbe de précision/rappel (ou courbe PR), peut susciter la question de savoir laquelle choisir. En général, la courbe PR est préférable lorsque la classe positive est rare, ou si vous accordez une plus grande importance aux faux positifs qu'aux faux négatifs. En revanche, la courbe ROC est plus adaptée dans le cas contraire (Aurélien *et al.*, 2017). La figure 2.12 montre que la courbe ROC (ligne orange)

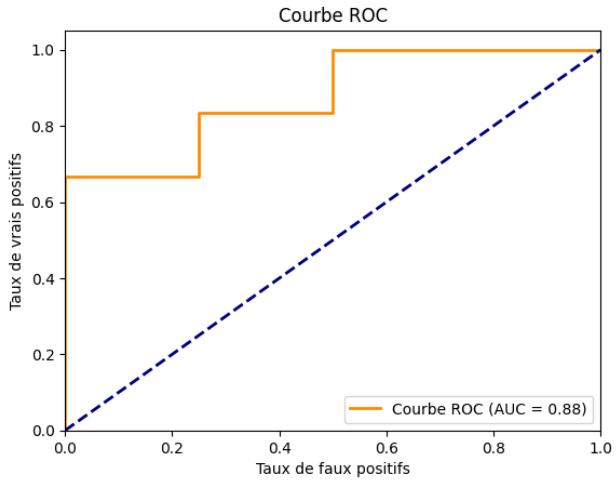

Figure 2.12 – Courbe ROC

montre une augmentation rapide du taux de vrais positifs (TVP) avec une augmentation minimale du taux de faux positifs (TFP). Cela indique que le modèle a une bonne capacité à séparer les classes positives des classes négatives. L'AUC est proche de 1 (0.96), ce qui indique une performance élevée du modèle. Cela suggère que le modèle est capable de bien classer les exemples positifs et négatifs dans cet ensemble de données.

Dans ce chapitre, nous avons mis en lumière les concepts fondamentaux de la distribution statistique et des méthodes d'analyse, avant d'aborder l'ingénierie des caractéristiques et les principaux concepts de l'apprentissage automatique. Ces outils et notions constituent une base méthodologique pour structurer et analyser les données de manière efficace. Dans le chapitre suivant, nous situerons notre travail dans le contexte de l'état de l'art en explorant les recherches existantes et en précisant la problématique abordée. Cette transition permettra de positionner notre approche par rapport aux défis actuels et aux lacunes identifiées dans le domaine.

CHAPITRE 3
ÉTAT DE L'ART ET PROBLÉMATIQUE

Dans cette section, nous aborderons les revues de littérature sur la recherche portant sur les techniques d'ingénierie des caractéristiques, les approches basées sur le profilage métabolique et les méthodes d'apprentissage automatique pour la prédiction des maladies métaboliques. Enfin, nous définirons notre problématique de recherche.

3.1 Ingénierie des caractéristiques

Des études telles que celle menée par (Dineva et Atanasova, 2023) ont exploré diverses techniques d'ingénierie des caractéristiques. Ces techniques comprennent la conversion d'objets en types de données spécifiques, le traitement des valeurs manquantes et aberrantes, le regroupement, l'équilibrage des données utilisant la technique SMOTE, ainsi que l'encodage catégorique (qui ont déjà été introduites dans le chapitre 2.

(Ozella *et al.*, 2023) a utilisé les données de température ruminale pour développer des modèles prédictifs, identifiant des sous-groupes de vaches présentant des variations de la température du prépartum. L'étude menée par (Vázquez-Diosdado *et al.*, 2023) exploite les données comportementales et physiologiques recueillies à partir de capteurs ruminaux chez les vaches laitières pour anticiper le moment du vêlage. En analysant des paramètres tels que la température et l'activité, les chercheurs ont développé cinq modèles distincts capables de prédire le vêlage jusqu'à 5 jours à l'avance, avec une précision atteignant 87,81%. Les événements de consommation ont été identifiés à partir des enregistrements quotidiens de température, tandis que neuf caractéristiques ont été évaluées chaque jour, comprenant la médiane et l'écart type de la température et de l'activité, ainsi que des mesures relatives aux baisses de température. Les baisses de température ont été utilisées pour calculer les caractéristiques liées à la consommation, comme détaillé dans l'étude de (Vázquez-Diosdado *et al.*, 2019).

L'étude sur le taux de respiration, en tant qu'activité physiologique fondamentale des vaches, reflète non seulement leur santé respiratoire, mais sert également d'indicateur crucial du bien-être pour évaluer le stress thermique pendant les saisons chaudes (Yan *et al.*, 2024). Les données utilisées ont été traitées comme suit : les valeurs manquantes dans les variables catégorielles ont été remplacées par le mode, tandis

que celles dans les variables numériques ont été comblées avec la médiane. Les enregistrements avec des taux de respiration manquants ont été directement exclus de l'analyse et de la modélisation. Après avoir traité les valeurs manquantes, la méthode du score z a été utilisée pour identifier quantitativement les valeurs aberrantes.

Les algorithmes d'apprentissage automatique travaillent avec des données numériques, mais il est crucial d'avoir des méthodes de transformation pour gérer des données hétérogènes. Pour sélectionner les caractéristiques les plus importantes,(Huang *et al.*, 2023) a utilisé la méthode d'élimination de caractéristiques récursives (RFE) pour choisir celles qui sont fortement corrélées. Des recherches telles que celle menées par (Pan *et al.*, 2022) ont révélé que les caractéristiques catégorielles surpassent souvent les caractéristiques numériques et combinées pour la prédiction des maladies cardio-vasculaires.

Le choix de la méthode de traitement des valeurs manquantes dépend du pourcentage de données manquantes et de la nature des données. Il est essentiel d'identifier et de traiter les valeurs aberrantes pour éviter des résultats biaisés lors de l'analyse et de la modélisation. De plus, l'équilibrage des données dans les problèmes de classification avec des classes déséquilibrées peut entraîner une perte d'informations ou un surajustement, modifiant également l'environnement réel. Certaines transformations des données, comme l'encodage one-hot, peuvent augmenter considérablement la dimensionnalité de l'espace des caractéristiques, ce qui entraîne des coûts plus élevés en termes de mémoire et de temps d'entraînement, surtout lorsque le nombre de catégories est élevé (Kuhn et Johnson, 2019). En abordant ces défis de manière appropriée, les techniques d'ingénierie de caractéristiques peuvent améliorer la qualité des données et optimiser les performances des modèles d'apprentissage automatique.

3.2 Méthodes d'utilisation des signatures de biomarqueurs métaboliques

Dans les sections suivantes, nous aborderons les approches basées sur le profilage métabolique. Ensuite, nous explorerons l'apprentissage automatique qui est apparu comme l'approche de modélisation prédominante. La plupart des études visaient à détecter les problèmes de santé des vaches, en particulier la mammite (Ozella *et al.*, 2023).

3.2.1 Approches basées sur le profilage métabolique

Les signatures de biomarqueurs métaboliques ont été utilisées dans le secteur agricole et laitier pour diagnostiquer et détecter divers troubles métaboliques chez les vaches laitières. (Zhang *et al.*, 2022) a identifié des biomarqueurs de métabolites pour diagnostiquer la stéatose hépatique chez les vaches en utilisant des profilages métabolomique. De même, (Dervishi *et al.*, 2018) ont employé le profilage métabolomique pour détecter les métabolites associés à la métrite chez les vaches, identifiant ainsi des signes précoces de la maladie. D'autres recherches, telles que celle de (He *et al.*, 2022), ont développé des méthodes de profilage métabolique pour identifier les signatures métaboliques liées à la boiterie chez les vaches, offrant ainsi des possibilités de diagnostic et de traitement précoces. Dans une étude sur la santé cardiovasculaire, (Yun *et al.*, 2022) ont découvert des signatures lipidomiques associées aux produits laitiers en utilisant du profilage métabolique, tandis que (Tata *et al.*, 2022) ont utilisé du profilage métabolique pour distinguer les échantillons de lait provenant de différents systèmes laitiers. Des techniques de profilage métabolomique sont utilisées pour analyser les profils métaboliques chez les bovins infectés par certaines maladies, comme la MAP (Mycobacterium avium sous-espèce paratuberculosis) (Massaro *et al.*, 2023), et pour caractériser les métabolites dans le plasma sanguin de veaux laitiers infectés par le virus respiratoire syncytial bovin (Abdullah Basoglu et Gulersoy, 2020b). De plus, des biomarqueurs tels que la lactate déshydrogénase (LDH) ont été utilisés pour détecter la mammite (Motohashi *et al.*, 2020).

Ces études démontrent l'importance cruciale de l'analyse des profils métaboliques dans la compréhension des mécanismes sous-jacents aux maladies métaboliques chez les vaches laitières, et détecter rapidement toute anomalie physiologique, ce qui permet un suivi efficace de l'état de santé des animaux et une mise en place précoce de traitements appropriés. Le profilage métabolique a permis de détecter plusieurs maladies et de comprendre leurs mécanismes. Des exemples incluent la boiterie (Zhang *et al.*, 2020), mammite (**?**; Zhang *et al.*, 2020), et la métrite (Hailemariam *et al.*, 2018; Zhang *et al.*, 2017).

En outre, l'utilisation de biomarqueurs contribue à améliorer le bien-être global des animaux d'élevage en permettant une gestion proactive de leur santé. La mise en œuvre des signatures des biomarqueurs métaboliques dans le secteur agricole et laitier est confrontée à divers défis et limitations. L'analyse du profil métabolique sanguin est coûteuse et stressante pour les vaches (Mota *et al.*, 2023), notamment lorsque cela nécessite l'utilisation d'équipements spécialisés ou des analyses de laboratoire sophistiquées (Les Producteurs laitiers du Canada, 2023). De plus, l'interprétation des données générées à partir des profils de

biomarqueurs peut être complexe, nécessitant des compétences interdisciplinaires pour extraire des informations pertinentes (Bao et Xie, 2022). La variabilité naturelle des profils de biomarqueurs constitue également un défi, car ceux-ci peuvent être influencés par un large éventail de facteurs tels que la génétique, l'environnement et l'alimentation, rendant difficile la distinction entre les effets spécifiques (Dervishi *et al.*, 2021; Schwegler *et al.*, 2013). Enfin, la standardisation des protocoles d'échantillonnage, d'extraction et d'analyse est cruciale pour garantir la fiabilité et la reproductibilité des résultats, mais cela représente également un défi logistique et méthodologique. Ces défis soulignent la nécessité d'une approche multidisciplinaire et collaborative pour surmonter les limitations actuelles et exploiter pleinement le potentiel dans l'agriculture et le secteur laitier (Ezanno *et al.*, 2020).

Ces études démontrent le potentiel des signatures de biomarqueurs métaboliques pour améliorer le diagnostic et la gestion dans le secteur agricole et laitier. De plus, des données cliniques telles que les scores de santé, les historiques médicaux et les performances laitières sont souvent incluses pour enrichir les modèles prédictifs.

3.2.2 Approches d'apprentissage automatique pour la prédiction des maladies métaboliques

Les approches d'apprentissage automatique utilisées dans la prédiction des maladies métaboliques chez les vaches laitières comprennent une variété de techniques, notamment les réseaux de neurones artificiels, les machines à vecteurs de support (SVM), les méthodes ensemblistes telles que le Random Forest (Ghaffari *et al.*, 2019; Tanyildizl et Yildirim, 2019), les arbres de décision, les K plus proches voisins (Tanyildizl et Yildirim, 2019), et les méthodes de régression(Trevor Hastie, 2009; Matyka et Koza, 2012). Ces approches impliquent l'utilisation des divers attributs, comme les informations génomiques et métaboliques, les enregistrements de performance laitière et d'autres paramètres pertinents (Hyde *et al.*, 2020) pour prédire les maladies métaboliques chez les vaches laitières.

L'étude de (Motohashi *et al.*, 2020) a utilisé la conductivité électrique du lait comme indicateur pour prédire la mammite avec des modèles basés sur les machines à vecteurs de support (SVM) et les forêts aléatoires, atteignant une sensibilité de 81% et une précision de 46%. De plus, l'étude de (Tanyildizl et Yildirim, 2019) souligne l'importance de la détection précoce pour la survie des animaux et la prévention des pertes. En comparant différents algorithmes de classification, ils ont constaté que l'algorithme J48 obtenait les meilleurs résultats, surpassant SVM et Naïve Bayes. Une autre étude menée par (Panchal *et al.*, 2015) a

appliqué des modèles connexionnistes anticipés pour classer les vaches Sahiwal en bonne santé et atteintes de mammite en utilisant divers paramètres tels que le pH, la conductivité électrique, la température, les cellules somatiques du lait. L'approche connexionniste (EBP) s'est révélée la plus efficace. Pour classer les vaches Sahiwal en bonne santé et celles atteintes de mammite. Les réseaux neuronaux, en particulier les modèles de perceptron multicouche (MLP), ont été utilisés pour prédire et classer la cétose subclinique chez les vaches laitières avec une précision améliorée (Bauer et Jagusiak, 2022). Les réseaux de neurones sont souvent privilégiés en raison de leur capacité à traiter des données complexes et non linéaires, tandis que les SVM sont couramment utilisés pour leur aptitude à gérer des ensembles de données de grande taille et à déceler des relations non linéaires (Trevor Hastie, 2009; Matyka et Koza, 2012). Pour évaluer les performances des algorithmes, des mesures telles que la précision, la spécificité et le rappel, ainsi que des matrices de confusion ont été utilisées pour analyser les résultats (Hyde *et al.*, 2020; Motohashi *et al.*, 2020).

Ces approches d'apprentissage automatique ont donné des résultats prometteurs en matière de prédiction précise des maladies métaboliques chez les vaches laitières, fournissant ainsi des informations précieuses pour les stratégies de détection précoce et de prévention (Tanyildizl et Yildirim, 2019). L'un des défis l'accès à des ensembles de données de grande taille et de haute qualité reste un défi, car les données métaboliques peuvent être coûteuses à collecter et nécessitent une expertise spécialisée pour leur interprétation. Comme le démontre (Giannuzzi *et al.*, 2022), des données d'entraînement médiocres peuvent entraîner des performances insuffisantes des algorithmes d'apprentissage automatique. Un autre défi est l'intégration de plusieurs sources de données, telles que les données metabolomiques, les données sur les exploitations agricoles et les informations génomiques, afin d'améliorer la précision des prévisions (He *et al.*, 2022). L'accessibilité des données publiques constitue un défi (Cockburn, 2020) ainsi que le besoin de technologies modernes et de précision dans l'industrie laitière (Bauer et Jagusiak, 2022). De plus, la variabilité naturelle des profils métaboliques entre les vaches et les troupeaux peut rendre la prédiction des maladies plus complexe (Foroutan *et al.*, 2020; Dervishi *et al.*, 2021). Cependant, il existe toujours un manque de méthodologie robuste pour utiliser les techniques d'apprentissage automatique dans ce domaine, et il a également été observé que la distribution des classes (labels) dans les jeux de données sur les maladies animales peut être souvent non-equilibrée, conduisant à des populations non équilibrées lors de la prédiction des problèmes de santé (Ozella *et al.*, 2023).

Dans l'ensemble, bien que l'apprentissage automatique soit devenu un outil courant dans la recherche laitière, il reste encore des défis à surmonter pour prédire efficacement les maladies métaboliques chez les vaches laitières. Des collaborations interdisciplinaires entre les scientifiques des données, les vétérinaires et les producteurs sont essentielles pour développer des modèles prédictifs robustes et adaptés à l'environnement de production laitière (Ezanno *et al.*, 2020).

3.3 Définition du problème

Cette recherche a pour but de concevoir une approche pour identifier des signatures de métabolites spécifiques aux maladies à déclarations obligatoires à partir d'un échantillon de dosage de cinq métabolites clés. Ces signatures doivent être discriminantes pour la prédiction des maladies métaboliques.

Ainsi, le problème peut être défini de la façon suivante : comment pouvons-nous, à partir d'une liste (L) de n dosages métaboliques associée à x maladies, identifier des signatures minimales S présentes dans L associées aux maladies qui pourront être intègrées dans un modèle de classification ?

Dans ce chapitre, nous avons passé en revue la littérature existante sur les techniques d'ingénierie des caractéristiques, les approches basées sur le profilage métabolique et les méthodes d'apprentissage automatique utilisées pour prédire les maladies métaboliques. Cette analyse critique a permis de mettre en évidence les avancées et les limites dans ce domaine, tout en justifiant l'importance de notre travail. Enfin, nous avons défini notre problématique de recherche, qui constitue le point de départ de notre démarche scientifique. Dans le chapitre suivant, nous décrirons en détail la méthodologie adoptée pour répondre à cette problématique, en expliquant les choix des outils et des approches analytiques. Cette transition marque le passage de la réflexion théorique à la mise en œuvre pratique de notre recherche.

CHAPITRE 4
MÉTHODOLOGIE

Dans ce chapitre, nous présentons l'approche utilisée pour identifier les signatures métaboliques. Ce chapitre est divisé en trois blocs principaux. Nous commençons par présenter les données et les étapes de pré-traitement associées à celles-ci. Ensuite, nous exposons l'approche proposée pour extraire les signatures métaboliques. Enfin, nous décrivons comment les différents algorithmes sont utilisés pour prédire la présence ou l'absence de maladies chez les vaches laitières.

Nous avons utilisé plusieurs outils et technologies pour atteindre nos objectifs, présentés ci-dessous :

- **Langage de programmation** : Python (version 3.8.10).
- **Frameworks et bibliothèques** : NumPy (1.23.5) pour les calculs mathématiques, Pandas (2.0.1) pour l'analyse des données, Scikit-learn (1.3.0) pour les modèles d'apprentissage automatique, et Matplotlib (3.7.1) pour la visualisation.
- **Environnement de développement** : Jupyter Notebook, facilitant le développement interactif et la documentation.

4.1 Description des données

Les données exploitées dans cette recherche sont issues du Dossier de Santé Animale (DSA), un système de gestion informatisé du dossier médical des animaux auquel les médecins vétérinaires québécois participent depuis plus de 30 ans.

Il est important de noter que nous n'utilisons pas la base de données complète de la DSA. Le jeu de données du projet comprend 1200 vaches provenant de 50 fermes québécoises rassemble des informations diverses sur la santé des animaux, y compris la composition du lait, les caractéristiques propres à chaque animal, et des détails pertinents sur la production laitière. Ces données détaillées permettent une analyse approfondie de la qualité nutritionnelle du lait en explorant ses composants essentiels. De plus, elles offrent une perspective individualisée sur les performances de chaque animal, guidant ainsi les producteurs dans la prise de décisions éclairées concernant la santé, la reproduction, et le suivi de chaque membre du troupeau.

Notre jeu de données comprend des caractéristiques cruciales pour l'analyse de la santé métabolique des vaches laitières. Ces caractéristiques, ou "features", ont été soigneusement sélectionnées en raison de leur pertinence dans l'évaluation de la santé métabolique. Elles comprennent des mesures telles que la quantité de matières grasses dans le lait (Fat), le Beta-hydroxybutyrate (BHB) pour évaluer le statut énergétique, le Milk Urea Nitrogen (MUN) pour des informations sur l'équilibre nutritionnel, et les protéines brutes dans le lait (Crude Protein) pour évaluer la qualité nutritionnelle. De plus, des informations spécifiques à chaque animal, telles que l'identifiant unique (Animal ID) et la date d'observation, sont également incluses.

Les caractéristiques liées à la santé, comme le nombre de cellules somatiques (SCC), sont utilisées dans notre étude. Le SCC, mesurant les cellules somatiques dans le lait, est un indicateur important de la santé des vaches laitières. La variable binaire "Label" indique si l'animal est classé comme malade ou non malade. La distribution des labels au sein de notre ensemble de données apporte des éclairages essentiels sur la répartition des animaux malades et non malades. Sur un total de 18 714 instances, 1171 animaux ont été assignés comme malades, tandis que 17543 animaux ont été identifiés comme non malades.

4.1.1 Analyse statistique des données

D'abord, une analyse de la distribution des valeurs manquantes est réalisée pour évaluer la complétude de nos données et guider les décisions liées au traitement des valeurs manquantes. Le tableaux 4.1 ci-dessous présente le pourcentage de valeurs manquantes pour les différentes caractéristiques. Pour garantir la qualité et la pertinence de notre analyse, nous avons mis en œuvre le processus de pré-traitement suivant. Chaque observation, représentée par une instance i, a subi une transformation temporelle, en considérant une fenêtre de -60 jours par rapport l'occurrence d'un contrôle qualité du lait. La fenêtre de -60 jours est choisie pour couvrir la période de collecte d'échantillons, généralement perpétuée dans -45 jours avant le début d'une lactation.

$$\text{ObservationDate-60} \ = \text{Date des contrôles qualité du lait à l'indice } i \text{ à la date d'observation} - 60 \text{ jours.} \tag{4.1}$$

Le jeu de données utilisé dans ce projet est composé de 18 714 instances pour 623 vaches. Cette taille d'échantillon optimisée forme la base solide sur laquelle repose notre analyse statistique, comme l'illustre la table4.1. Le jeu de données était segmenté de cinq façons.

— **Générale :** les propriétés des variables sont évaluées indépendamment des identifiants d'animal ou de ferme. Cette première étape nous offre une vue d'ensemble de la distribution des caractéristiques

Caractéristique	% valeurs manquantes
Fat	3,02%
Crude Protein	2,79%
MUN	18,25%
BHB	84,23%
SCC	2,79%
Sampledate	0%
AnimalId	0%
ObservationDate	0%
ObservationDate-60	0%
healthCD	0%
Label	0%
Farmid	0%

Table 4.1 – Taux de valeurs manquants pour les attributs du jeux de données

dans notre ensemble de données.

— **Troupeaux :** calcul des moyennes d'attributs pour chaque ferme, fournissant des indications sur la santé globale des troupeaux

— **Animal :** calcul des moyennes d'attributs pour chaque animal sur différentes dates d'échantillon.

— **Ensemble :** inclus toutes les instances des contrôles de qualité du lait

— **Complet :** inclus les contrôles de qualité du lait pour lesquels l'ensemble des attributs n'a aucune valeur manquante.

Pour une meilleure compréhension de notre jeu de données, veuillez consulter le tableau suivant4.2

Contrôles qualité du lait	Ensemble	Complet
Nombre de Contrôles qualité du lait	18714	4567
Animaux Distinctifs	623	313
Troupeaux distinctifs	42	27

Table 4.2 – Présente des statistiques sur les deux représentations du jeux de donnés.

À travers ces différentes stratégies de segmentation, notre objectif est de dévoiler des tendances significatives et d'approfondir notre compréhension des interactions dans notre ensemble de données.

4.1.2 Évaluation de la distribution des données

Avant d'approfondir notre analyse, une étape cruciale consiste à évaluer la distribution des données métaboliques à travers deux méthodes complémentaires : le Kernel Density Estimation (KDE) gaussien (Kristan

et al., 2011) et le test de normalité de Shapiro (Rakotomalala, 2008) .

Ces approches nous permettent d'explorer la forme générale de la distribution ainsi que de quantifier la normalité statistique des données. Pour ce faire, nous avons remplacé les valeurs manquantes par la moyenne, une stratégie courante dans le traitement des données manquantes (Kunwar, 2020) . Le test de normalité de Shapiro a été utilisé avec un seuil de significativité de 0.05. Ainsi, la décision d'accepter ou de rejeter l'hypothèse de normalité dépendra de la valeur-p résultante. Si la valeur-p est inférieure à 0.05, nous rejetterons l'hypothèse de normalité ; sinon, nous l'accepterons. Les résultats de ces analyses seront présentés 5.1.2. , apportant des éclaircissements essentiels sur la nature des distributions de nos variables métaboliques. Ces informations préliminaires sont cruciales pour orienter notre choix d'outils statistiques et interpréter de manière appropriée les résultats ultérieurs de notre étude.

4.2 Approche méthodologique

4.2.1 Catégorisation des caractéristiques

Après avoir calculé la moyenne et l'écart-type pour chaque attribut de notre $D_{\text{controle de qualité du lait}}$, nous déterminons des seuils significatifs en utilisant différents seuils par rapport a l'écart-type. Ces seuils sont ensuite utilisés comme base pour la catégorisation des observations au sein de notre ensemble de données.

Les valeurs de chaque observation sont comparées aux seuils préalablement déterminés, comme le montre l'équation 4.3. En fonction de cette comparaison, chaque observation est attribuée à une catégorie pertinente, comme illustré par l'équation 4.4.

$$seuil\_inf = moyenne - n \times \text{écart-type} \tag{4.2}$$

$$seuil\_sup = moyenne + n \times \text{écart-type} \tag{4.3}$$

Soit x la valeur de l'observation et seuil_inf et seuil_sup les seuils inférieur et supérieur respectivement, la catégorie peut être déterminée comme suit :

$$\text{f}(x) = \begin{cases} \text{low} & \text{si } x < \text{seuil\_inf} \\ \text{medium} & \text{si seuil\_inf} \leq x \leq \text{seuil\_sup} \\ \text{high} & \text{si } x > \text{seuil\_sup} \end{cases} \tag{4.4}$$

Premièrement, la catégorisation a pour but de faciliter l'analyse statistique (Miola et Miot, 2022), de déterminer l'indépendance des catégories (Li *et al.*, 2020), et de gérer la sparcité des données en identifiant des valeurs plus discriminantes à travers des catégories définies (Grandini *et al.*, 2020). Enfin l'utilisation de termes intuitifs comme **low (faible)**, **high (élevé)** ou **medium (normal)** facilite grandement la communication avec les éleveurs, permettant une transmission claire et compréhensible des informations sur la santé métabolique des animaux. En effet, traduire des résultats en termes de catégories simplifiées offre aux éleveurs une compréhension immédiate de l'état de leurs troupeaux, sans nécessiter une expertise approfondie en statistiques. Cette approche favorise une communication efficace entre les professionnels de la santé animale et les éleveurs, renforçant ainsi la valeur pratique de notre méthodologie dans un contexte d'élevage. La facilité d'interprétation des résultats contribue à une utilisation plus étendue de nos conclusions dans la prise de décision quotidienne des éleveurs, soulignant ainsi l'aspect appliqué et concret de notre approche.

Dans le cadre de notre méthodologie, nous avons entrepris une série d'expérimentations en vue de déterminer le nombre d'écart-types optimal pour la catégorisation de nos données, explorant ainsi divers scénarios pour évaluer leur impact sur la classification des données.

Au total, nous avons conduit 18 expériences, testant trois valeurs (n= 0.1, n=0.3, n=2) de multiples de l'écart-type pour définir les seuils de catégorisation. Ces investigations ont été menées sur les cinq segmentation distinctes : **Général**, **Animal** et **Troupeaux**, **Ensemble** et **Complete**. Les seuils choisis ont été soigneusement appliqués à chaque variable. Les 18 scénarios expérimentaux qui sont présents dans le tableau suivant 4.3.

Segmentation	Threshold	Contrôle qualite du lait
Général	$2 \times$ std	Ensemble
Général	$0.3 \times$ std	Ensemble
Général	$0.1 \times$ std	Ensemble
Général	$2 \times$ std	Complet
Général	$0.3 \times$ std	Complet
Général	$0.1 \times$ std	Complet
Animal	$2 \times$ std	Ensemble
Animal	$0.3 \times$ std	Ensemble
Animal	$0.1 \times$ std	Ensemble
Animal	$2 \times$ std	Complet
Animal	$0.3 \times$ std	Complet
Animal	$0.1 \times$ std	Complet
Troupeaux	$2 \times$ std	Ensemble
Troupeaux	$0.3 \times$ std	Ensemble
Troupeaux	$0.1 \times$ std	Ensemble
Troupeaux	$2 \times$ std	Complet
Troupeaux	$0.3 \times$ std	Complet
Troupeaux	$0.1 \times$ std	Complet

Table 4.3 – Paramètres pour les 18 expériences avec différentes segmentations et contrôles qualité du lait

Voici les valeurs de n choisies et leur impact dans la catégorisation et l'analyse subséquente.

— $2 \times$ **std :** Ce seuil englobe l'écart-type à 2, soit $2 \times$ std, s'inscrit dans une démarche stratégique visant à englober la majorité des données au sein de la catégorie principale tout en identifiant les valeurs qui s'écartent significativement de la distribution normale. Cette approche a pour objectif de capturer approximativement 95% de nos données dans la plage moyenne, conformément à la règle empirique de la distribution normale (Investir Sorcier, 2021).

— $0.1 \times$ **std :** Le choix d'utiliser 0.1 écart-type vise à maintenir les seuils autour de la moyenne.

— $0.3 \times$ **std :** Ce seuil est utilisé à titre comparatif, permettant de mettre en lumière les différences d'interprétation qui pourraient découler de l'utilisation de valeurs multiples plus faibles.

Cependant, après une évaluation approfondie des résultats, une conclusion émerge de manière claire et pertinente : la valeur de $0.1 \times$ std se distingue comme étant la plus adaptée pour notre analyse. Ce choix repose sur plusieurs considérations cruciales, garantissant ainsi la pertinence et la fiabilité de notre approche.

Premièrement, le choix du seuil 0.1 suit la recommandation des experts en science animale lié à l'observation que nos données suivent une distribution normale. En utilisant un multiple de l'écart-type aussi faible,

nous maintenons une fenêtre relativement proche de la moyenne, ce qui est particulièrement adapté aux données distribuées normalement.

Ainsi, le choix de 0.1 pour le multiple de l'écart-type résulte d'une approche stratégique, combinant la nature des données, la proximité à la moyenne, et les recommandations éclairées des experts du domaine. Cette décision éclaire la robustesse de notre méthodologie, positionnant notre catégorisation comme une étape cruciale dans la compréhension fine des variations métaboliques au sein de notre ensemble de données.

4.2.2 Représentation des signatures des biomarqueurs

À partir de cette partie, toutes les transformations présentées découleront de la segmentation générale et de l'ensemble des contrôles qualité du lait, comme décrit dans la section 4.1.1. Dans le cadre de notre méthodologie d'analyse, nous avons introduit le concept de signature métabolique pour synthétiser les informations cruciales contenues dans différentes caractéristiques métaboliques. La signature métabolique est définie comme des tuples (a, b), où a représente les caractéristiques étudiées (BHB, SCC, MUN) et b représente les catégories (low, medium, high), agissant comme une représentation agrégée des multiples dimensions métaboliques que nous évaluons. En consolidant ces caractéristiques, la signature offre une vue d'ensemble, permettant de capturer de manière efficace et concise les nuances complexes des données de chaque individu dans notre jeu de données. L'utilisation de cette formule s'inscrit dans notre volonté de simplifier la complexité des données tout en préservant leur richesse d'information, facilitant ainsi l'interprétation et l'analyse ultérieure de nos résultats. Les signatures des biomarqueurs peuvent être représentées suivant une fonction comme définie par l'equation 4.4

$$f : \mathbb{A} \rightarrow \{\text{low}, \text{medium}, \text{high}\} \tag{4.5}$$

4.2.3 Signatures des biomarqueurs discriminantes

Après avoir généré nos signatures des biomarqueurs, l'étape suivante de notre procédure implique une évaluation visant à déterminer si certaines de ces signatures présentent des caractéristiques discriminantes significatives. Pour ce faire, nous avons établi les termes :

— Both : Les signatures dont la p-value est inférieure à 0,05 avec le test de Fisher, et la valeur 1 n'appartient pas à l'intervalle de confiance.
— OneP : Les signatures dont la p-value est inférieure à 0,05 avec le test de Fisher, et la valeur 1 appartient à l'intervalle de confiance.
— OneCI : Les signatures dont la p-value est supérieure à 0,05 avec le test de Fisher, et la valeur 1 n'appartient pas à l'intervalle de confiance.

Ces catégories visent à structurer l'analyse statistique, permettant d'identifier les signatures métaboliques qui présentent des différences significatives et informatives dans notre ensemble de données.

4.3 Analyse des signatures métaboliques basée sur l'apprentissage automatique

Pour la mise en place de notre modèle, nous avons utilisé notre jeu de données de différentes manières : Nous avons utilise l'ensemble des signatures, puis nous avons ensuite sélectionné les signatures discriminantes en fonction des tests statistiques définis précédemment(Both, CI, OneP). En suite, une attention particulière a été accordée à la surreprésentation de la classe minoritaire en raison d'un déséquilibre significatif entre les classes. En effet, le jeu de données présente une disparité marquée, la classe majoritaire représentant 93,74% des instances tandis que la classe minoritaire ne compte que pour 6,25%. Cette disparité dans la distribution des classes peut induire des biais dans les modèles de classification, favorisant la prédiction de la classe majoritaire au détriment de la classe minoritaire.

Enfin, étant donné que dans notre jeu de données, la classe minoritaire représente 6,25% (1171 instances), nous avons adopté une méthodologie spécifique en termes d'échantillonnage. Dans cette configuration, nous avons entraîné un modèle en sélectionnant rigoureusement 800 signatures pour chaque classe (1600 signatures au total), tandis que 320 signatures ont été mises de côté pour le test, réparties équitablement entre les classes (640 signatures au total).

L'encodage one-hot a été uniformément adopté comme méthode d'encodage pour toutes les configurations de modèles. L'encodage One-Hot était utilisé pour convertir les variables non-ordinales sous une forme qui pourrait être fournie aux algorithmes d'apprentissage machine.

Pour les expériences, le jeu de données est divisé entre train/test, avec 70% pour l'entraînement et 30% pour le test a été adopté. Cette approche a été adoptée pour explorer diverses configurations d'échantillon-

nage afin de mieux comprendre l'impact des déséquilibres de classe sur les performances des modèles de classification. Ces expérimentations contribuent à une évaluation plus complète et nuancée des capacités des modèles dans des contextes variés.

Dans cette section dédiée à l'analyse des signatures des biomarqueurs, plusieurs algorithmes d'apprentissage automatique ont été choisis pour répondre aux défis particuliers de ce domaine. Le choix des algorithmes repose sur les conclusions de l'analyse comparative des algorithmes de classification supervisée réalisée par les différentes études (Ghaffari *et al.*, 2019; Tanyildizl et Yildirim, 2019; Trevor Hastie, 2009; Matyka et Koza, 2012), qui ont démontré leur performance, leur flexibilité, leur adaptabilité ainsi que leur capacité à gérer des données de grande dimension, comme illustré dans la section 3.2.2 .le choix de SVM, Random Forest et régression logistique pour notre classifieur binaire est justifié par leur robustesse, leur capacité à gérer des données complexes et déséquilibrées, ainsi que leur interprétabilité. Chacun de ces algorithmes apporte des avantages uniques qui peuvent être exploités selon les caractéristiques spécifiques de nos données

- Random Forest :(Criterion = gini, Number of Estimators = 50, Max Depth = 10, Max Features = `n_features`)
- SVM :(Kernel = linear, C = 1.0)
- Régression Logistique : `LogisticRegression(random_state=16)`

Les métriques incluent l'exactitude, la précision, le rappel, et le F-mesure, chacune de ces métriques d'évaluation, permet de capturer à la fois la justesse globale du modèle et sa capacité à gérer les classes minoritaires de manière robuste.

Dans ce chapitre, nous avons détaillé l'approche méthodologique adoptée pour identifier les signatures métaboliques. Nous avons d'abord décrit les données utilisées ainsi que les étapes essentielles de prétraitement. Ensuite, nous avons présenté l'approche proposée pour extraire ces signatures et, enfin, montré comment différents algorithmes permettent de prédire la présence ou l'absence de maladies métaboliques chez les vaches laitières. Dans le chapitre suivant, nous analyserons les résultats obtenus à partir de cette méthodologie. Cette transition nous permettra d'évaluer la pertinence des signatures métaboliques identifiées, d'examiner l'efficacité des algorithmes appliqués et de discuter des implications de ces résultats en lien avec la problématique de recherche. Cette discussion servira également à mettre en perspective les forces et les limites de notre approche.

CHAPITRE 5

RESULTATS ET DISCUSSION

Dans ce chapitre, nous présenterons les résultats statistiques et l'évaluation de la normalité des données. Ensuite, nous exposerons nos résultats concernant la catégorisation et les signatures obtenues. Enfin, nous présenterons les résultats de la prédiction des maladies chez les vaches laitières basée sur l'apprentissage automatique.

5.1 Analyse des attributs du contrôle qualité du lait et évaluation de leur normalité

5.1.1 Profil statistique des attributs en fonction du contrôle qualité du lait

Avant de plonger dans les détails spécifiques, il est essentiel de conduire une analyse statistique des résultats de nos contrôles de qualité du lait. Cette étape cruciale nous permet de comprendre la distribution générale des valeurs, offrant un aperçu de la tendance centrale, de la dispersion, ainsi que des valeurs extrêmes dans nos données. Les mesures statistiques telles que la moyenne, l'écart-type, le minimum et le maximum seront présentées dans la table 5.1. Ces indicateurs fournissent une vision cohérente de la variabilité des données médicales et facilitent l'interprétation des résultats spécifiques à chaque attribut.

	Ensemble				Partiel				Complete			
Attribute	Moyenne	Ecart-type	Min	Max	Moyenne	Ecart-type	Min	Max	Moyenne	Ecart-type	Min	Max
Fat(%/ Kg)	3.79	1.40	0.0	10.18	3.86	1.53	0.0	10.18	4.15	0.87	1.49	9.15
CrudeProtein (%/ Kg)	3.04	1.00	0.0	6.44	2.98	1.13	0.0	6.40	3.23	0.41	2.40	6.44
MUN(mg/dl)	11.32	3.49	1.0	24.79	11.46	3.42	1.0	24.79	11.00	3.61	1.70	23.39
SCC(count/kg)	324.61	829.42	0.0	9999.0	351.60	877.31	0.0	9999.0	249.01	671.82	3.0	9999.0
BHB(mmol/L)	0.11	0.12	0.0	4.51	0.14	0.30	0.009	4.51	0.10	0.05	0.0	0.5

Table 5.1 – Profil statistique des attributs en fonction du contrôle qualité du lait

D’après le tableau 5.1 : l’attribut "Fat" montre une légère diminution de la moyenne (passant de 3.86 à 3.76) pour l’ensemble des contrôles qualité du lait par rapport au contrôle qualité du lait partiel, puis une augmentation plus marquée du contrôle qualité du lait complet au contrôle qualité du lait partiel (passant de 3.86 à 4.15). De même, l’attribut "CrudeProtein" présente une légère diminution de la moyenne (passant de 3.04 à 2.98) pour l’ensemble des contrôles qualité du lait par rapport au contrôle qualité du lait partiel, suivie d’une augmentation plus significative du contrôle qualité du lait complet au contrôle qualité du lait partiel. En ce qui concerne les autres attributs tels que "MUN", "SCC" et "BHB", les variations entre les contrôles qualité du lait Ensemble, Partiel et Complet sont moins importantes. Cela indique que les vaches soumises à un contrôle qualité du lait complet présentent un pourcentage de matière grasse légèrement plus élevé.

Pour Fat, on observe une légère augmentation de l’écart-type pour l’ensemble des contrôles qualité du lait par rapport au contrôle qualité du lait partiel (1.4 a 1.53), puis une diminution de l’écart-type du contrôle qualité du lait partiel au contrôle médical complet(1.53 a 0.87). Cela suggère que les valeurs sont moins dispersées autour de la moyenne dans la catégorie Contrôle qualité du lait complet. Pour CrudeProtein (%/Kg), l’écart-type augmente légèrement (1 a 1.13) de l’ensemble du contrôle qualité aux du contrôle qualité partiels, puis diminue davantage des du contrôle qualité partiels aux du contrôle qualité complets (1.13 a 0.41). Cela indique également une dispersion moindre des valeurs dans la catégorie Contrôle qualité complet. Pour les attributs MUN, SCC et BHB, les variations d’écart-type entre les catégories sont moins importante. Le contrôle qualité du lait complet semble présenter des caractéristiques plus homogènes et des valeurs moyennes plus stables pour plusieurs attributs, notamment la teneur en matière grasse (Fat) et la concentration de BHB.

5.1.2 Évaluation de la normalité des données

Nous commençons par présenter les résultats de l’évaluation de la distribution des données métaboliques. Cette évaluation a été réalisée en utilisant deux méthodes complémentaires : le Kernel Density Estimation (KDE) gaussien et le test de normalité de Shapiro, comme expliqué dans la sous-section 4.1.2. Les courbes des distributions de la Figure 5.1 présentent une forme caractéristique en cloche avec une symétrie par rapport à la moyenne et une dispersion réduite des données autour de la moyenne. Il est possible de conclure que les données semblent suivre une distribution normale.

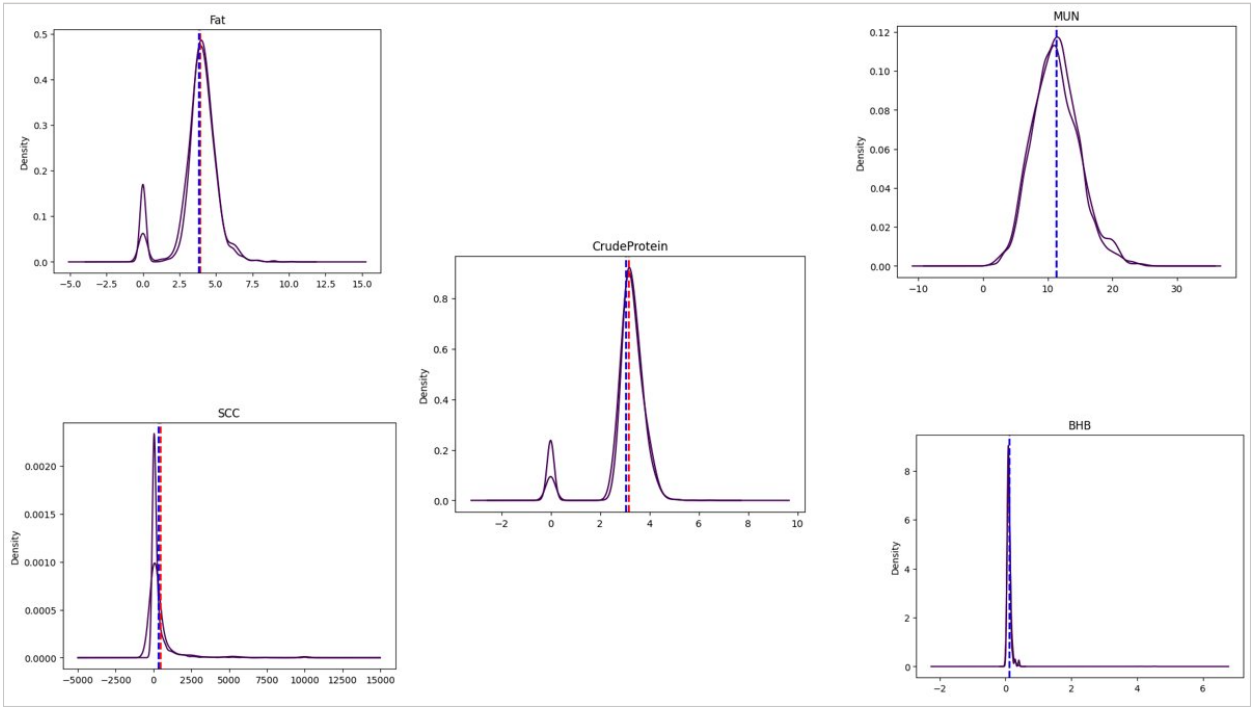

Figure 5.1 – Plot distribution KDE (gaussienne)

En particulier, les courbes correspondant aux concentrations de SCC et de BHB sont relativement étroites, suggérant une faible dispersion des données autour de leurs moyennes respectives. Ces observations sont cohérentes avec les résultats de l'analyse KDE gaussienne, confirmant ainsi la stabilité et la cohérence des données dans ces domaines spécifiques.

Cette évaluation a été réalisée à l'aide du test de normalité de Shapiro, comme expliqué dans la sous-section 4.1.2. Pour toutes les variables (Fat, CrudeProtein, MUN, SCC, BHB) l'hypothèse nulle selon laquelle les données suivent une distribution normale est acceptée. Après avoir utilisé à la fois le Kernel Density Estimation (KDE) gaussien et le test de normalité de Shapiro pour évaluer la distribution des données métaboliques, nous pouvons conclure que les données semblent suivre une distribution normale. Cette conclusion est étayée par les résultats des deux méthodes.

5.2 Résultats de l'approche

Dans cette section, nous allons présenter les résultats de la catégorisation et des signatures de biomarqueurs obtenues.

Biomarqueur	Segmentation		Moyenne	Ecartype	Seuil_inf	Seuil_sup	Low(occ)	Medium(occ)	High(occ)
Fat	Ensemble	General	3.79	1.35	3.66	3.93	3.17%	20%	**49.21%**
		Troupeaux	4.15	0.80	4.07	4.24	**48%**	14.79%	37.24%
		Animal	4.06	0.85	3.97	4.14	48.21%	12.5%	39.28%
	Complete	General	4.15	0.87	4.06	4.24	**51.06%**	8.88%	40.04%
		Troupeaux	4.29	0.75	4.21	4.37	**52.08%**	6.25%	41,66%
		Animal	4.07	0.82	3.98	4.15	**50%**	3.84%	46,15%
CrudeProtein	Ensemble	General	3.04	0.96	2.95	3.14	21.31%	24.34%	**54.54%**
		Troupeaux	3.30	0.36	3.26	3.33	**49.48%**	10.20%	40.30
		Animal	3.30	0.42	3.26	3.34	**54.46%**	11.60%	33.92%
	Complete	General	3.23	0.41	3.18	3.27	**54.97%**	9.74%	35.29%
		Troupeaux	3.20	0.32	3.17	3.23	**50%**	10.41%	39.58%
		Animal	3.21	0.37	3..17	3.24	**53.84%**	7.69%	38..46%
MUN	Ensemble	General	11.32	2.98	11.02	11.62	**35.23%**	32.05%	32.71%
		Troupeaux	11.26	2.58	11	11.52	**41.83%**	24.48%	33.67
		Animal	11.60	2.59	11.34	11.86	**39.28%**	24.10%	36.60%
	complete	General	11.04	3.61	10.67	11.40	**48,5%**	10.9%	41,44%
		Troupeaux	10.69	2.32	10.45	10.92	**50%**	6.25%	43.75%
		Animal	10.48	2.10	10.27	10.69	42.30%	11.53%	**46.15%**
SCC	Ensemble	General	324.61	798.87	224.72	404.50	**68.68%**	14.64%	16.66%
		Troupeaux	215.03	610.42	154	276.07	**73.46%**	15.30%	11,22%
		Animal	183.05	543.23	128.73	237.37	**72.32%**	18.75%	8.92%
	Complete	General	249.01	671.82	181.82	316.19	**74.57%**	8.36%	17.05%
		Troupeaux	126.95	375.96	89.36	164.55	**79.16%**	8.33%	12.50%
		Animal	84.19	129.13	71.27	97.10	**65.38%**	19.23%	15.38%
BHB	Ensemble	General	0.11	0.06	0.10	0.11	16.25%	**73.74%**	10%
		Troupeaux	0.10	0.02	0.10	0.105	16.32%	**72.44%**	11.22%
		Animal	0.08	0.02	0.08	0.11	15.17%	**72.32%**	12.5%
	Complete	General	0.10	0.05	0.09	0.10	**50.22%**	9.48%	40.28%
		Troupeaux	0.10	0.048	0.09	0.10	**50%**	10.41%	39.58%
		Animal	0.08	0.04	0.079	0.08	**46.15%**	11.53%	42,3%

Table 5.2 – Statistiques trouvées par différentes segmentations du jeux de donnés

5.2.1 Catégorisation

Dans la table 5.2 La plupart des données dans chaque mesure sont classées dans la catégorie "low", ce qui signifie qu'elles sont inférieures à la valeur de référence (4.3).

Les données proviennent de différentes segmentations et concernent divers biomarqueurs tels que Fat, CrudeProtein, MUN, SCC et BHB. Les statistiques comprennent la moyenne, l'écart-type, les seuils inférieur et supérieur, ainsi que la répartition en pourcentage des catégories low, medium et high pour chaque biomarqueur et segmentation. Les résultats montrent des variations selon les types de segmentation et les biomarqueurs, avec des tendances distinctes observées. Par exemple, la catégorie "high" pour le biomarqueur Fat est plus fréquente chez les troupeaux et les animaux par rapport à la segmentation générale, tandis que pour le CrudeProtein, elle est plus élevée dans la segmentation animal que dans la segmentation troupeaux. En ce qui concerne le MUN, il est plus fréquent dans la segmentation générale. Pour SCC, la catégorie "low" est la plus présente dans les cinq segmentations. De plus, des variations significatives sont observées dans la répartition des niveaux low, medium et high selon les différentes segmentations et les biomarqueurs analysés.

5.2.2 Analyse de la composition des signatures

Dans cette section, nous présentons les résultats de notre analyse des signatures pour les vaches malades et non malades.

Nous avons identifié un total de 208 signatures distinctes. Toutes les 208 signatures étaient présentes chez les vaches saines, tandis que 131 étaient associées aux vaches malades, comme l'illustre la heatmap 5.2 qui montre les variations dans les signatures.

La table 5.3 révèle des tendances distinctes entre les vaches malades et non malades. En ce qui concerne la graisse (Fat), on remarque que la catégorie la plus représentative pour les vaches malades est "High", avec une proportion de 38.17%, tandis que pour les vaches non malades, c'est "Low" avec 36.06%. Concernant la protéine brute (Crudeprotein), les niveaux restent relativement similaires entre les deux groupes, sans différence notable. Cependant, pour l'urée du lait (MUN), les vaches malades montrent une proportion plus élevée, en particulier dans la catégorie "Low" avec 38.93%, comparée à 35.10% chez les non malades. Les cellules somatiques du lait (SCC) présentent une différence significative, avec une proportion plus élevée

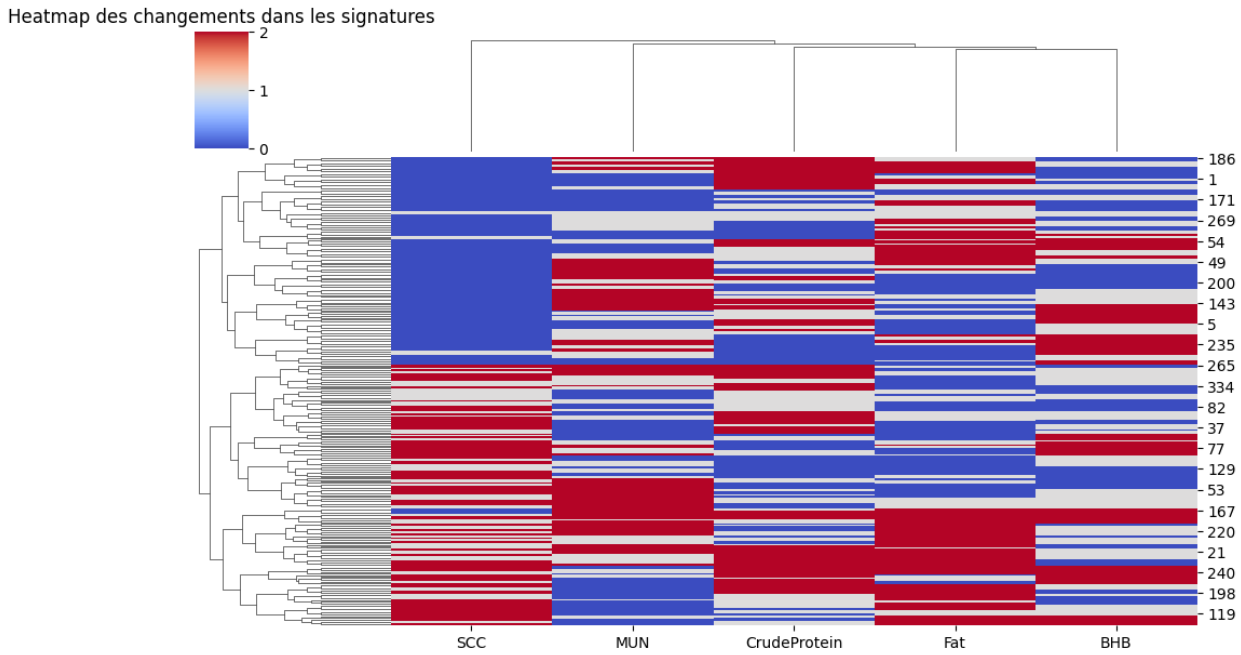

Figure 5.2 – Variation dans les signatures métaboliques

.

chez les vaches malades, surtout dans la catégorie "Low" (50.38% contre 40.65%). De même, pour BHB, la catégorie "Low" est la plus représentative pour les vaches malades avec 48.85%, tandis que pour les non malades, elle est à 37.50%.

Une analyse de l'homogénéité des biomarqueurs dans les signatures des vaches malades par rapport à celles des vaches non malades souligne des variations distinctes, indiquant des différences potentielles dans leur santé et leur condition physiologique. Les chiffres présentés dans la Table 5.4 indiquent qu'il existe une différence notable dans la distribution de l'homogénéité des signatures entre les deux groupes.

Sur les 131 signatures associées aux vaches malades, 67,15% présentent une homogénéité de plus de 50% (c'est-à-dire que la même catégorie se répète au moins 3 fois pour 5 les attributs), tandis que sur les 208 signatures des vaches non malades, seulement 62,5% affichent une telle homogénéité. Ces résultats soulignent l'importance de cette caractéristique pour distinguer avec précision les deux groupes.

Les 131 signatures associées aux vaches malades, ont été retrouvée chez les vaches non malades, ce qui indique une absence de spécificité des signatures pour les animaux malades dans notre analyse. Cette observation pourrait suggérer la nécessité de recherches plus approfondies pour identifier des signatures

Biomarqueur	Label	Low(occ)	Medium(occ)	High(occ)
Fat	Malade	37.40%	24.43%	**38.17%**
	nonMalade	**36.06%**	28.85%	35.10%
Crudeprotein	Malade	32.82%	**33.59%**	**33.59%**
	nonMalade	31.25%	34.13%	**34.62%**
MUN	Malade	**38.93%**	30.53%	30.53%
	nonMalade	**35.10%**	32.21%	32.69%
SCC	Malade	**50.38%**	18.32%	31.30%
	nonMalade	**40.65%**	26.64%	32.71%
BHB	Malade	**48.85%**	27.48%	23.66%
	nonMalade	**37.50%**	33.65%	28.85%

Table 5.3 – Statistiques trouvés pour les differentes categories entre les vaches malade et non malade

Label	compte_Homogénéité
malade	88
non-malade	130

Table 5.4 – Nombre de signatures homogènes

spécifiques à la maladie.

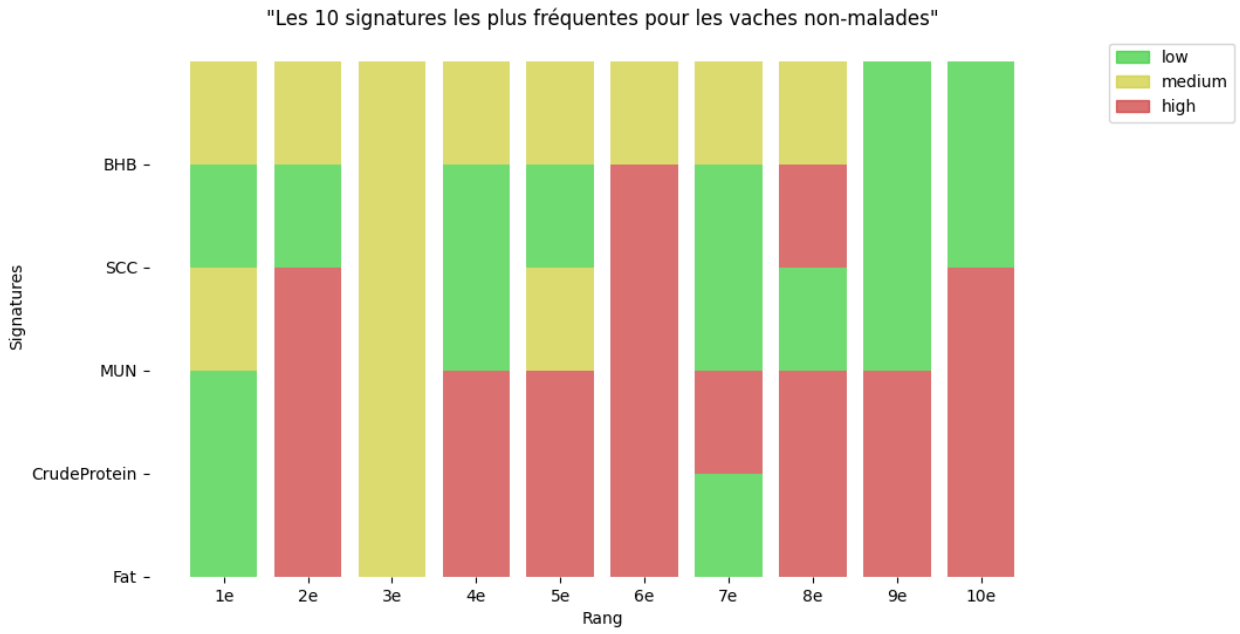

Figure 5.3 – Les 10 signatures les plus fréquentes pour les vaches non malades

Les figures 5.4 5.3 suivantes présenteront les 10 signatures les plus fréquentes pour les vaches, qu'elles soient malades ou non. En comparant les deux figures, nous pouvons observer certaines similitudes ainsi que des différences notables. Sur les 10 signatures les plus fréquentes associées aux vaches malades, 70%

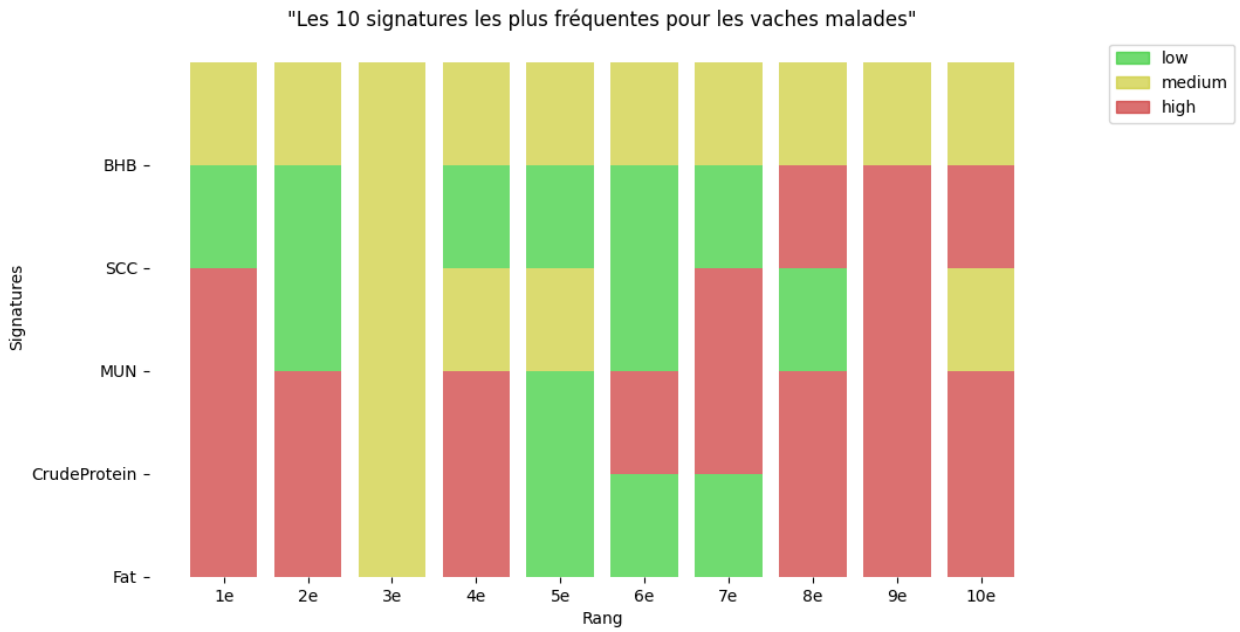

Figure 5.4 – Les 10 signatures les plus fréquentes pour les vaches malades

présentent une homogénéité de plus de 50%, tandis que pour les vaches non malades, ce chiffre est de 80%. Cela suggère que certaines caractéristiques sont partagées entre les deux groupes. Une observation notable est que les troisième et huitième signatures apparaissent dans les deux groupes et occupent la même position. En outre, la signature la plus fréquente dans le groupe "Non-malade" (1ère position) correspond à la cinquième position dans le groupe "Malade".

Les signatures les plus distinctives semblent être celles avec des combinaisons de valeurs "low" et "high". Par exemple, les signatures comportant plusieurs occurrences de "high" sont plus fréquentes chez les vaches malades, tandis que les signatures avec des occurrences de "low" sont plus courantes chez les vaches non malades.

Cette analyse met en lumière l'importance de comprendre les schémas de biomarqueurs pour distinguer entre les vaches malades et non malades. Elle souligne également la complexité des maladies chez les vaches et la nécessité d'une approche nuancée dans leur diagnostic et leur traitement.

Les différences observées dans les signatures les plus fréquentes entre les vaches malades et non malades suggèrent qu'il existe des caractéristiques distinctives associées à chaque groupe. Cependant, la présence de signatures communes souligne également la complexité de la relation entre les signatures et l'état

de santé des vaches.

5.2.3 Analyse des signatures discriminantes

Dans ce qui suit, nous présentons les résultats de notre recherche de signatures discriminantes entre les vaches malades et non malades, basée sur deux tests statistiques : la valeur de p-value et l'intervalle de confiance comme décrit dans cette section4.2.3.

Type de Signature	Nombre de Signatures
Total	131
Both	96
OneCI	56
OneP	23

Table 5.5 – Répartition des signatures discriminantes

La table 5.5 montrent que la majorité des signatures (96 sur 131) passe avec succès les deux tests statistiques, ce qui indique une forte corrélation entre ces signatures et l'état de santé des vaches.

D'autre part, 56 signatures ont passé uniquement le test de l'intervalle de confiance, ce qui suggère une relation significative entre ces signatures et l'état de santé des vaches, même si leur valeur de p n'était pas statistiquement significative ($p > 0.05$).

Enfin, 23 signatures ont passé le test de p-value mais ont échoué à l'intervalle de confiance, indiquant une corrélation statistiquement significative avec l'état de santé des vaches, mais une incertitude quant à la précision de cette corrélation.

En regardant ces données la Table 5.6 montre que : Les signatures ont des valeurs de p inférieures à 0,05, ce qui indique une association statistiquement significative avec le résultat (malade/non-malade). Les valeurs d'odds ratio peuvent être utilisées pour interpréter la force de l'association entre les variables et les résultats. Les colonnes "1 dans CI" indiquent si l'association est statistiquement significative.

Si on regarde les deux premières signatures de la Table5.6, on note que le rapport de cotes est de 0,692, ce qui signifie que les vaches avec la Signature 1 ont environ 0,692 fois plus de chances d'avoir la maladie par rapport à celles avec la signature 2.

Signatures	Fat	CrudeProtein	MUN	SCC	BHB	Malade(occ)	Non-malade(occ)	odds	P-value	1 dans CI
1	high	high	high	low	medium	111	1496	0,692	0,029	vrai
2	high	high	medium	low	medium	62	579			
3	high	high	low	low	medium	105	1206	4,54	0	vrai
4	high	high	low	medium	medium	4	209			
5	medium	medium	medim	medium	medium	82	1278	0,54	0,001	vrai
6	low	low	medium	low	medium	53	1516			
7	high	high	low	high	medium	35	422	0,95	0,92	faux
8	high	high	low	low	medium	105	1206			

Table 5.6 – Exemple de signatures discriminante

La valeur de p-value est de 0,029, ce qui est inférieur à 0,05, indiquant qu'il existe suffisamment de preuves statistiques pour conclure de manière significative que la signature 1 est significativement différente de la signature 2 en termes d'association avec la maladie.

5.3 Résultats des signatures métabolique basées sur l'apprentissage automatique

À partir de cette partie, tous les résultats qui seront présentés découleront de la segmentation **générale** et de l'ensemble de nos contrôles qualité du lait (**ensemble**).

- **d1 :** Ensemble des signatures.
- **d2 :** Signatures ayant passé le test p-value et l'intervalle de confiance.
- **d3 :** Signatures ayant seulement passé l'intervalle de confiance.
- **d4 :** Signatures ayant seulement passé le test p-value .
- **d5 :** 800 signatures pour l'entraînement et 320 signatures pour le test pour chaque classe.

La Table 5.7 présente les résultats de classification de l'algorithme SVM. La première colonne fournit les informations relatives aux ensembles de données utilisés. Les autres colonnes affichent les différentes métriques de performance (précision, rappel, F-mesure) obtenues lors de l'évaluation. Pour les jeux de données d1, d2, d3 et d4, le modèle obtient une précision de 0,94 pour la classe "nonMalade", ce qui signifie qu'il prédit correctement 94% des exemples non malades parmi tous ceux qu'il a étiquetés comme non malades. Cependant, la précision pour la classe "Malade" est de 0, ce qui indique que le modèle n'a pas correctement identifié d'exemples malades dans ces jeux de données. Le rappel, la F-mesure et l'exactitude pour la classe "Malade" sont également de 0 dans ces cas.

En ce qui concerne le jeu de données d5, le modèle affiche des performances légèrement meilleures

jeu de données	Classe	Précision	Rappel	F-mesure	Exactitude	Moyenne pondère F-mesure
d1	Malade	0	0	0	0,94	0,91
	nonMalade	0,94	1	0,97		
d2	Malade	0	0	0	0,94	0,91
	nonMalade	0,94	1	0,97		
d3	Malade	0	0	0	0,91	0,88
	nonMalade	0,92	1	0		
d4	Malade	0	0	0	0,92	0,88
	nonMalade	0,92	1	0,96		
d5	Malade	0,50	0,74	0,60	0,54	0,52
	nonMalade	0,61	0,36	0,46		

Table 5.7 – SVM : Performance de classification

jeu de données	Classe	Précision	Rappel	F-mesure	Exactitude	Moyenne pondère F-mesure
d1	nonMalade	0,60	0,40	0,48	0,56	0,55
	Malade	0,55	0,73	0,63		
d2	nonMalade	0,60	0,30	0,40	0,55	0,52
	Malade	0,54	0,80	0,65		
d3	nonMalade	0	0	0	0,50	0,33
	Malade	0,63	1	0,67		
d4	nonMalade	0,61	0,48	0,54	0,58	0,58
	Malade	0,50	0,68	0,61		

Table 5.8 – SVM : Performance de classification avec suréchantillonnage la classe minoritaire

pour la classe "Malade", avec une précision de 0,50, un rappel de 0,74 et une F-mesure de 0,60. Cependant, la précision, le rappel et la F-mesure pour la classe "nonMalade" sont plus faibles, ce qui indique une performance globalement moins satisfaisante sur ce jeu de données. La "Moyenne pondérée F-mesure" reste assez stable à travers les différents jeux de données, variant entre 0,88 et 0,52. Cela suggère que, bien que le modèle puisse avoir des difficultés à classifier correctement certaines classes individuelles (comme observé dans les précédentes analyses pour la classe "Malade"), il parvient néanmoins à maintenir une performance globalement acceptable lorsqu'on considère toutes les classes ensemble.

La table 5.8 suivant présente les résultats de classification de l'algorithme SVM avec surechantillonnage de la classe minoritaire : Pour les jeux de données d1, d2 et d4, nous observons une amélioration significative dans la performance de la classe "Malade" par rapport aux résultats précédents. La précision, le rappel et la F-mesure pour la classe "Malade" ont tous augmenté, indiquant une meilleure capacité du modèle à identifier correctement les exemples de cette classe. La "Moyenne pondérée F-mesure" varie entre 0.33 et 0.58, ce qui suggère une amélioration globale de la performance du modèle par rapport aux résultats

jeu de données	Classe	Précision	Rappel	F-mesure	Exactitude	Moyenne pondère F-mesure
d1	nonMalade	0,94	1	0,97	0,94	0,91
	Malade	0	0	0		
d2	nonMalade	0,94	1	0,97	0,93	0,91
	Malade	0	0	0		
d3	nonMalade	0,92	1	0,96	0,91	0,88
	Malade	0	0	0		
d4	nonMalade	0,92	1	0,96	0,92	0,88
	Malade	0	0	0		
d5	nonMalade	0,59	0,42	0,49	0,53	0,53
	Malade	0,50	0,66	0,57		

Table 5.9 – Random Forest : Performance de classification

précédents. Cependant, il est important de noter que les performances peuvent varier d'un jeu de données à l'autre.

La table 5.9 présente les résultats de classification de l'algorithme Random Forest. La première colonne fournit les informations relatives aux ensembles de données utilisés. Les autres colonnes affichent les différentes métriques de performance (précision, rappel, F-mesure) obtenues lors de l'évaluation. Il convient de noter que le modèle de Random Forest parvient à détecter une classe particulière, à savoir les vaches non malades, à travers les ensembles de données d1, d2, d3 et d4, comme en témoignent les valeurs de précision, de rappel et de score F-mesure associées à cette classe (voir table5.9).

Dans l'ensemble, le modèle affiche des performances solides en termes de précision et de rappel, bien que ces performances puissent varier d'un ensemble de données à un autre. Les valeurs d'exactitude oscillent entre 91% et 94%. Cependant, pour l'ensemble des données d5, l'exactitude chute à 53%. Malgré cela, le modèle parvient à identifier les deux classes en termes de moyenne pondérée F1-score avec 53%, bien que ce soit légèrement moins performant que le SVM. Cela suggère qu'il peut encore détecter un nombre considérable de vaches malades malgré la faible exactitude globale.

La table 5.10 présente les résultats de classification de l'algorithme Random Forest après avoir appliqué une technique de sur-échantillonnage à la classe minoritaire. La première colonne fournit les informations relatives aux ensembles de données utilisés. Les autres colonnes affichent les différentes métriques de performance (précision, rappel, F-mesure) obtenues lors de l'évaluation. Nous avons constaté une légère diminution de l'exactitude, oscillant entre 50% et 64%, toutefois restant supérieure à celle du SVM. Il est cependant crucial de souligner que malgré cette baisse, notre modèle parvient toujours à identifier effica-

jeu de données	Classe	Précision	Rappel	F-mesure	Exactitude	Moyenne pondère F-mesure
d1	nonMalade	0,59	0,39	0,47	0,64	0,54
	Malade	0,54	0,73	0,62		
d2	nonMalade	0,59	0,45	0,51	0,63	0,62
	Malade	0,51	0,70	0,62		
d3	nonMalade	0	0	0	0,50	0,33
	Malade	0,50	1	0,67		
d4	nonMalade	0,64	0,38	0,47	0,60	0,57
	Malade	0,54	0,78	0,64		

Table 5.10 – Random Forest : Performance de classification avec suréchantillonnage la classe minoritaire

jeu de données	Classe	Précision	Rappel	F-mesure	Exactitude	Moyenne pondère F-mesure
d1	nonMalade	0,94	1	0,97	0,94	0,91
	Malade	0	0	0		
d2	nonMalade	0,94	1	0,97	0,93	0,91
	Malade	0	0	0		
d3	nonMalade	0,92	1	0,96	0,91	0,88
	Malade	0	0	0		
d4	nonMalade	0,92	1	0,96	0,92	0,88
	Malade	0	0	0		
d5	nonMalade	0,59	0,44	0,50	0,53	0,53
	Malade	0,50	0,64	0,56		

Table 5.11 – Régression logistique : Performance de classification

cement les deux classes, démontrant ainsi sa capacité à traiter des situations plus complexes. De manière prometteuse, nous avons atteint la meilleure performance en termes de moyenne pondéré F-mesure, avec une valeur de 62%, en utilisant des signatures provenant de l'ensemble de données d2.

La table 5.11 présente les résultats de classification de l'algorithme régression logistique. La première colonne fournit les informations relatives aux ensembles de données utilisés. Comme pour les deux modèles précédents, la régression logistique arrive a identifier, une classe spécifique, celle des vaches non malades, à travers les ensembles de données d1, d2, d3 et d4, comme le confirment les valeurs de précision, de rappel et de F-mesure associées à cette classe . Globalement, le modèle affiche des performances robustes en termes de précision et de rappel, bien que ces performances puissent varier d'un ensemble de données à un autre. Les valeurs d'exactitude se situent entre 92% et 94%. Cependant, pour l'ensemble de données d5, l'exactitude diminue à 53%. Malgré cela, le modèle parvient à identifier les deux classes avec une moyenne pondérée de F-mesure de 53%, bien que légèrement moins performant que le SVM et le Random Forest.

La table 5.12 expose les résultats de la classification pour la régression logistique suite à l'application

jeu de données	Classe	Précision	Rappel	F-mesure	Exactitude	Moyenne pondère F-mesure
d1	nonMalade	0,59	0,39	0,47	0,56	0,55
	Malade	0,54	0,73	0,62		
d2	nonMalade	0,59	0,45	0,51	0,57	0,55
	Malade	0,51	0,70	0,62		
d3	nonMalade	0	0	0	0,50	0,33
	Malade	0,50	1	0,67		
d4	nonMalade	0,64	0,38	0,47	0,57	0,57
	Malade	0,54	0,78	0,64		

Table 5.12 – Régression logistique : Performance de classification avec suréchantillonnage la classe minoritaire

Modèle	données	Rappel
RandomForest	d2	0,81
Régression Logistique	d4	0,78
SVM	d5	0,74

Table 5.13 – Analyse comparative du rappel des algorithmes RandomForest, Régression Logistique et SVM

d'une technique de sur-échantillonnage à la classe minoritaire. Nous avons remarqué une légère réduction de l'exactitude, variant entre 50% et 57%, tout en restant supérieure à celle du SVM. Cependant, il est essentiel de noter que malgré cette baisse, nous avons obtenu la meilleure performance en termes de F score, avec une valeur de 57%, en utilisant des caractéristiques discriminantes issues de l'ensemble de données d2.

La table 5.13 résume les résultats de la classification pour différents algorithmes presentes ici. La première colonne fournit les informations relatives aux modèles, la deuxième colonne fournit les informations relatives aux ensembles de données utilisés, et la dernière colonne fournit le métrique. Dans cette analyse, nous avons privilégié le rappel par rapport à la précision, car dans ce contexte, il est crucial de détecter toutes les vaches malades, même au risque de quelques faux positifs. En utilisant le sur-échantillonnage de la classe minoritaire, RandomForest a atteint une performance remarquable avec un rappel de 81% pour les signatures ayant passé les deux tests statistiques, démontrant ainsi le meilleur résultat global. Par contre, la régression logistique a démontré sa supériorité pour les signatures ayant uniquement réussi le test de la valeur p, avec un rappel de 78%. SVM, quant à lui, a obtenu le meilleur résultat avec un rappel de 74% sur un jeu de données équilibré.

5.4 Discussion

L'objectif de cette recherche a pour but de concevoir une approche pour identifier des signatures de métabolites spécifiques aux maladies chez les vaches laitieres. Ces signatures doivent être discriminantes pour la prédiction des maladies métaboliques. Nos recherches offrent des perspectives intéressantes sur les signatures métaboliques associées à la santé des vaches. Nous avons identifié un total de 208 signatures sur 243 possibles, soit 85% des signatures possibles, ce qui est très élevé. Parmi ces 208 signatures, toutes étaient présentes dans les vaches non malades, tandis que seulement 131 étaient présentes dans les vaches malades. En utilisant une méthode de feature engineering pour discrétiser les mesures en classes low, medium et hight avec un seuil de 0,3 écart-type, nous avons identifié une signature présente chez les vaches malades et absente chez les vaches saines. Cela montre que le choix du seuil d'écart-type influence nos résultats. Il serait judicieux de définir une approche pour trouver le bon seuil, par exemple en utilisant des techniques de data mining, que nous n'avons pas employées ici, car l'objectif était de déterminer si une signature était présente ou non. On note que chez les vaches malades, 67,15 % des signatures présentent une homogénéité de plus de 50 % (c'est-à-dire que la même catégorie se répète au moins 3 fois sur 5 attributs), tandis que 62,5 % affichent une telle homogénéité chez les vaches saines.

En examinant les dix signatures les plus fréquentes associées aux vaches malades, nous avons observé que 70 % d'entre elles présentent une homogénéité de plus de 50 %, tandis que ce chiffre est de 80 % pour les vaches non malades. Ces observations suggèrent une similarité métabolique entre les deux groupes. Bien que des différences significatives subsistent, cette superposition met en évidence la complexité de la distinction entre ces deux groupes. Une analyse approfondie des 131 signatures communes a révélé que 96 d'entre elles ont passé à la fois les deux statistiques, ce qui représente nos signatures discriminantes. En regardant la table 5.14 , on note une disparité qui mérite une méthode plus concise pour déterminer si une signature est discriminante ou non. Par exemple, la plus petite valeur d'une signature dans la catégorie malade est 1 et 4 chez les non-malades, tandis que la plus grande valeur est de 40 contre 1516. Il serait judicieux de déterminer un seuil minimal de différence pour qualifier une signature de discriminante.

L'utilisation de l'apprentissage automatique s'est avéré être la méthode prédominante dans la prédiction des problèmes chez les vaches laitières, représentant 63% des études, tandis qu'une proportion significative de 82% des études examinées se concentrent principalement sur la détection des problèmes de santé des vaches, en mettant particulièrement l'accent sur la mammite (Ozella *et al.*, 2023). Il est bien établi que

des marqueurs de santé et métaboliques tels que les cellules somatiques (SCC), le BHB et le MUN sont des indicateurs fiables de la santé chez les vaches laitières (Gonçalves Frasco *et al.*, 2020). Par exemple, les vaches présentant un SCC variant entre 200 000 et 500 000 par ml sont catégorisées comme des vaches atteintes de mammite subclinique, tandis que celles ayant un SCC de lait supérieur à 500 000/ml sont considérées comme relevant de la catégorie de mammite clinique. L'étude (Panchal *et al.*, 2015) a démontré que le nombre limite acceptable de SCC est de 400 000 au Canada. Notre jeu de données montre une moyenne du SCC de 364 000, ce qui facilite son intégration dans un modèle de classification selon ces catégories.

L'utilisation des signatures pour prédire les maladies chez les vaches laitières reste un domaine à explorer. Dans notre revue littéraire, nous n'avons pas trouvé d'étude utilisant les mêmes attributs pour extraire les signatures. Par conséquent, il ne serait pas très judicieux de tirer des conclusions telles que les signatures sont meilleures que les métabolites comme attributs pour la mise en place d'un classifieur.

Comparativement à une étude menée par (Lasser *et al.*, 2021) portant sur la prédiction des vaches malades en utilisant le SCC avec d'autres attributs, notre modèle présente un score de F-mesure variant entre 0,47 et 0,91, supérieur à leur score compris entre 0,5 et 0,72. La revue de la littérature met en évidence l'importance du SCC et des biomarqueurs alternatifs dans la prédiction et le diagnostic de maladies telles que la mammite chez les vaches laitières, soulignant la nécessité de méthodologies et de critères diagnostiques standardisés (Darbaz *et al.*, 2023; C, 2023). En termes d'exactitude, l'étude (Tanyildizl et Yildirim, 2019) a obtenu un score de 97% avec l'algorithme de Random Forest et de 96% avec le SVM en utilisant le SCC, le pH du lait et la conductivité électrique. Bien que notre modèle ait atteint une précision de 94% malgré un déséquilibre entre les classes, une collecte de données équilibrée entre vaches malades et saines pourrait améliorer davantage sa capacité à distinguer correctement les vaches malades des vaches saines.

Comme l'a soulevé (Ozella *et al.*, 2023), la question de pourquoi les méthodes d'apprentissage automatique ne sont pas davantage utilisées pour améliorer les stratégies de production. Une explication possible est l'absence de jeux de données complets contenant diverses informations sur la qualité et la quantité du lait(Ozella *et al.*, 2023). Des informations précises et de grande qualité, qui englobent divers éléments, pourraient contribuer à la création d'algorithmes visant à optimiser la gestion de la production laitière.

Fat	CrudeProteine	MUN	SCC	BHB	Malade	nonMalade
high	high	high	high	high	3	42
high	high	high	high	low	1	25
high	high	high	high	medium	30	439
high	high	high	low	high	3	152
high	high	high	low	low	8	339
high	high	high	low	medium	11	1496
high	high	high	medium	high	3	37
high	high	high	medium	medium	12	253
high	high	low	high	high	3	63
high	high	low	high	low	2	55
high	high	low	high	medium	35	422
high	high	low	low	high	2	174
high	high	low	low	low	7	411
high	high	low	low	medium	105	1206
high	high	low	medium	medium	4	209
high	high	medium	high	high	2	35
high	high	medium	high	low	2	26
high	high	medium	high	medium	25	324
high	high	medium	low	high	4	77
high	high	medium	low	low	2	90
high	high	medium	low	medium	62	579
high	high	medium	medium	medium	8	117
medium	low	high	medium	medium	3	4
medium	low	medium	low	low	3	13
medium	medium	low	low	low	9	60
medium	low	high	high	medium	5	22

Fat	CrudeProteine	MUN	SCC	BHB	Malade	nonMalade
high	low	high	high	medium	1	35
high	low	high	low	high	4	73
high	low	high	low	low	3	42
high	low	high	low	medium	5	123
high	low	low	high	medium	3	43
high	low	low	low	high	3	78
high	low	low	low	low	8	69
high	low	low	low	medium	7	85
high	low	medium	low	high	3	75
high	low	medium	low	medium	5	48
high	medium	high	high	medium	3	58
high	medium	high	low	high	6	106
high	medium	high	low	medium	17	228
high	medium	high	medium	high	2	10
high	medium	low	high	low	6	13
high	medium	low	high	medium	5	65
high	medium	low	low	high	9	112
high	medium	low	low	low	5	106
high	medium	low	low	medium	8	155
high	medium	medium	low	high	2	60
high	medium	medium	low	low	3	14
high	medium	medium	low	medium	19	70
low	low	medium	high	high	3	28
low	low	medium	high	medium	5	35
low	low	medium	low	low	2	38

Fat	CrudeProteine	MUN	SCC	BHB	Malade	nonMalade
low	high	high	high	medium	5	71
low	high	high	low	high	1	22
low	high	high	low	low	1	106
low	high	high	low	medium	36	275
low	high	low	high	low	1	26
low	high	low	high	medium	6	78
low	high	low	low	high	3	38
low	high	low	low	low	3	130
low	high	low	low	medium	40	428
low	high	low	medium	medium	2	73
low	high	medium	high	medium	10	83
low	high	medium	low	medium	17	145
low	low	medium	low	medium	1	33
low	low	high	high	medium	4	22
low	low	low	high	low	1	27
low	low	high	low	low	11	143
low	low	high	low	medium	19	176
low	low	high	medium	low	2	7
low	low	low	high	medium	5	60
low	low	low	low	high	6	45
low	low	low	low	low	22	122
low	low	low	low	medium	22	251
low	medium	low	low	low	3	101
low	medium	low	low	medium	22	232
low	medium	low	medium	low	3	39

Fat	CrudeProteine	MUN	SCC	BHB	Malade	nonMalade
medium	high	low	high	medium	3	75
medium	high	low	low	low	1	112
medium	high	low	low	medium	21	252
medium	high	medium	high	medium	4	55
medium	low	high	low	low	1	27
medium	high	medium	low	low	2	42
medium	high	medium	low	medium	10	125
low	medium	low	high	medium	15	95
low	low	medium	medium	medium	3	7
low	medium	high	high	medium	4	49
low	medium	high	low	low	4	100
low	medium	high	low	medium	7	198
low	medium	medium	high	medium	8	38
low	medium	medium	low	low	3	52
low	medium	medium	low	medium	6	85
medium	high	high	high	medium	6	59
medium	high	high	low	medium	20	254
low	medium	low	medium	medium	1	26
medium	low	high	low	medium	1	25
low	low	medium	low	medium	53	1516

Table 5.14 – Présentation de la table des Signatures Métaboliques discriminantes en Fonction de l'état de Santé des Vaches

CONCLUSION

La solution développée a permis d'acquérir et de consolider des compétences variées en informatique, apprentissage automatique et santé animale.Cette combinaison de connaissances nous a permis de concevoir une approche pour extraire des signatures métaboliques et les utiliser dans un modèle de classification visant à détecter les vaches malades parmi les vaches en bonne santé. Notre étude visait à concevoir une approche pour identifier des signatures métaboliques spécifiques aux maladies chez les vaches laitières. Nous avons démontré que ces signatures peuvent être discriminantes pour la prédiction des maladies métaboliques, offrant ainsi des perspectives intéressantes pour l'amélioration de la santé animale et de la productivité des exploitations laitières.

Afin d'évaluer de manière exhaustive notre approche, nous avons constitué plusieurs ensembles de données. Ces ensembles ont été soumis à diverses évaluations et comparaisons, permettant ainsi d'extraire des signatures discriminantes grâce à des tests de Fisher et des intervalles de confiance. Les résultats de notre étude ont mis en évidence l'efficacité de notre approche dans la classification des maladies métaboliques. Nos résultats ont révélé l'existence de 208 signatures au total. L'ensemble de ces signatures sont présentes chez les vaches non malades, tandis que 131 sont présentes chez les vaches malades. Parmi ces signatures, une proportion significative présentait une homogénéité de plus de 50% entre les groupes de vaches malades et non malades. Une analyse approfondie des signatures communes a également souligné l'importance d'une approche multifactorielle dans la sélection des signatures métaboliques pour la prédiction des maladies. Les marqueurs métaboliques, tels que les cellules somatiques (SCC), le BHB et le MUN, ont été cruciaux dans notre démarche, confirmant leur rôle essentiel dans le diagnostic des maladies chez les bovins laitières.

Cependant, notre recherche a également révélé des défis méthodologiques significatifs, notamment en ce qui concerne la sensibilité de la catégorisation basée sur l'écart-type et le déséquilibre des données entre les classes de vaches malades et non malades. En conclusion, notre étude a contribué à la transition vers une agriculture plus durable et efficace en utilisant les avancées technologiques pour améliorer la santé des bovins laitières. En envisageant l'avenir, une extension prometteuse de notre méthodologie serait le développement d'un classifieur multi-classes pour identifier des signatures spécifiques à chaque maladie métabolique. Cette approche plus ciblée pourrait permettre une gestion plus précise et proactive de la santé des bovins laitiers, contribuant ainsi à améliorer la productivité et le bien-être global du troupeau.

Toutefois, cela nécessiterait une quantité accrue de données pour garantir la robustesse et la précision du modèle. En somme, notre étude marque une avancée significative dans le domaine de l'élevage de précision et de la santé animale. En continuant à explorer ces domaines, nous aspirons à façonner un avenir où la technologie et l'agriculture travaillent ensemble pour relever les défis alimentaires mondiaux tout en préservant la santé et le bien-être des animaux et de l'environnement.

BIBLIOGRAPHIE

Abdullah Basoglu, Nuri Baspinar, L. T. C. L. et Gulersoy, E. (2020a). Nuclear magnetic resonance (nmr)-based metabolome profile evaluation in dairy cows with and without displaced abomasum. *Veterinary Quarterly*, *40*(1), 1–15. http://dx.doi.org/10.1080/01652176.2019.1707907

Abdullah Basoglu, Nuri Baspinar, L. T. C. L. et Gulersoy, E. (2020b). Nuclear magnetic resonance (nmr)-based metabolome profile evaluation in dairy cows with and without displaced abomasum. *Veterinary Quarterly*, *40*(1), 1–15. http://dx.doi.org/10.1080/01652176.2019.1707907

Aggarwal, S., Banerjee, N., Parihari, S., Roy, J., Bojak, K. et Shah, R. (2022). *Metabolomics : Role in pathobiology and therapeutics of COVID-19* (1st éd.). CRC Press.

Ahmed, H. A., Muhammad Ali, P. J., Faeq, A. K. et Abdullah, S. M. (2022). An investigation on disparity responds of machine learning algorithms to data normalization method. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, *10*(2), 29–37. http://dx.doi.org/10.14500/aro.10970

Alam, R. (2020). Standardization and normalization | towards data science. *Medium ; Towards Data Science*. Récupéré de https://towardsdatascience.com/normalization-vs-standardization-explained-209e84d0f81e.

Aurélien, G., Anne, B. et Géron, A. (DL 2017). *Machine learning avec Scikit-Learn / Aurélien Géron ; traduit de l'anglais par Anne Bohy*. Malakoff : Dunod.

Bao, J. et Xie, Q. (2022). Artificial intelligence in animal farming : A systematic literature review. *Journal of Cleaner Production*, *331*, 129956. http://dx.doi.org/https://doi.org/10.1016/j.jclepro.2021.129956.

Bauer, E. A. et Jagusiak, W. (2022). The use of multilayer perceptron artificial neural networks to detect dairy cows at risk of ketosis. *Animals*, *12*(3). http://dx.doi.org/10.3390/ani12030332

Bellet, A., Habrard, A. et Sebban, M. (2015). *Metric Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers. http://dx.doi.org/10.2200/S00626ED1V01Y201501AIM030.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. et Milanesi, L. (2016). Methods for the integration of multi-omics data : mathematical aspects. *BMC Bioinformatics*, *17*(2), S15. http://dx.doi.org/10.1186/s12859-015-0857-9.

Bewley, J. (2010). Precision dairy farming : Advanced analysis solutions for future profitability. *Proc. 1st North Am. Conf. Precis. Dairy Manag*.

Bisong, E. (2019). Support vector machines. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Récupéré de https://api.semanticscholar.org/CorpusID:204088138

Bohnsack, K. S., Kaden, M., Abel, J. et Villmann, T. (2023). Alignment-free sequence comparison : A systematic survey from a machine learning perspective. *IEEE/ACM Transactions on*

Computational Biology and Bioinformatics, *20*(1), 119–135. http://dx.doi.org/10.1109/TCBB.2022.3140873

Bors, A. et Pitas, I. (1999). Object classification in 3-d images using alpha-trimmed mean radial basis function network. *IEEE Transactions on Image Processing*, *8*(12), 1744–1756. http://dx.doi.org/10.1109/83.806620.

BROLIS HerdLine (2024). Valeur diagnostique du lactose dans le lait. Retrieved from https://brolisherdline.com/fr/lactose-dans-le-lait/.

Buckland, M. et Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, *45*(1), 12–19. http://dx.doi.org/https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.

Bzdok, D., Krzywinski, M. et Altman, N. (2018). Machine learning : supervised methods. *Nature Methods*, *15*(1), 5–6. http://dx.doi.org/10.1038/nmeth.4551.

C, R. B. (2023). Cattle disease prediction using artificial intelligence. *International Journal For Science Technology And Engineering*, *11*(4), 2184–2189. http://dx.doi.org/10.22214/ijraset.2023.50535.

Caja, G., Castro-Costa, A. et Knight, C. H. (2016). Engineering to support wellbeing of dairy animals. *The Journal of dairy research*, *83*(2), 136–147. http://dx.doi.org/10.1017/S0022029916000261.

Carr, J. E., Austin, J., Hatfield, D. B. et Bailey, J. S. (1996). The standard deviation as an informative measure of variability in reporting interobserver agreement means. *Journal of Behavior Therapy and Experimental Psychiatry*, *27*(3), 263–267. http://dx.doi.org/https://doi.org/10.1016/S0005-7916(96)00024-9.

Carvalho, M. R., Penagaricano, F., Santos, J. E. P., DeVries, T. J., McBride, B. W. et Ribeiro, E. S. (2019). Long-term effects of postpartum clinical disease on milk production, reproduction, and culling of dairy cows. *Journal of Dairy Science*. http://dx.doi.org/10.3168/jds.2019-17025.

Chantal, P. (2022). Mammite et avortements. *Journal of Dairy Science*, *82*(08), 1684.

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). Smote : Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Clinique Vétérinaire de l'Aérodrome Saint Romain de Colbosc (2024). Clinique vétérinaire de l'aérodrome saint romain de colbosc. Récupéré de https://www.cliniqueveterinairesaintromain.fr/AccueilClinique.aspx?code=12603&parent=12603./

Cockburn, M. (2020). Review : Application and prospective discussion of machine learning for the management of dairy farms. *Animals*, *10*(9). http://dx.doi.org/10.3390/ani10091690. Récupéré de https://www.mdpi.com/2076-2615/10/9/1690

Darbaz, I., Ulusoy, B., Darbaz, T., Hecer, C. et Aslan, S. (2023). The importance of somatic cell count in dairy technology. *Mljekarstvo*, *73*(2), 75–84. Récupéré de https://doi.org/10.15567/mljekarstvo.2023.0201.

Das, R., Sailo, L., Verma, N., Bharti, P., Saikia, J., Imtiwati et Kumar, R. (2016). Impact of heat stress on health and performance of dairy animals : A review. *Veterinary World*, *9*, 260 – 268. http://dx.doi.org/10.14202/vetworld.2016.260-268.

Dervishi, E., Plastow, G., Hoff, B. et Colazo, M. (2021). Common and specific mineral and metabolic features in dairy cows with clinical metritis, hypocalcaemia or ketosis. *Research in Veterinary Science*, *135*, 335–342.

Dervishi, E., Zhang, G., Dunn, S. M., Mandal, R., Wishart, D. S. et Ametaj, B. N. (2017). Gc–ms metabolomics identifies metabolite alterations that precede subclinical mastitis in the blood of transition dairy cows. *Journal of Proteome Research*, *16*(2), 433–446. PMID : 28152597, http://dx.doi.org/10.1021/acs.jproteome.6b00538.

Dervishi, E., Zhang, G., Hailemariam, D., Mandal, R., Wishart, D. S. et Ametaj, B. N. (2018). Urine metabolic fingerprinting can be used to predict the risk of metritis and highlight the pathobiology of the disease in dairy cows. *Metabolomics*, *14*(6), 83. http://dx.doi.org/10.1007/s11306-018-1379-z.

Dillon, E. J. et Hennessy, T. (2012). Measuring the impact of improved animal health practices on the economic efficiency of irish dairy farms. http://dx.doi.org/10.22004/AG.ECON.158706.

Dineva, K. et Atanasova, T. (2023). Health status classification for cows using machine learning and data management on aws cloud. *Animals*, *13*(20). http://dx.doi.org/10.3390/ani13203254.

Educative (2015). Data science in 5 minutes : What is one hot encoding ? Récupéré de https://www.educative.io/blog/one-hot-encoding#what.

Etienne (2020). Quel est l'impact de l'industrie laitière sur l'environnement ? https://www.planete-durable.com/quel-est-limpact-de-lindustrie-laitiere.

Ezanno, P., Picault, S., Winter, N., Beaunée, G., Monod, H. et Guégan, J.-F. (2020). Intelligence artificielle et santé animale. *INRAE Productions Animales*, *33*. http://dx.doi.org/10.20870/productions-animales.2020.33.2.3572.

Fadul-Pacheco, L., Delgado, H. et Cabrera, V. E. (2021). Exploring machine learning algorithms for early prediction of clinical mastitis. *International Dairy Journal*, *119*, 105051. http://dx.doi.org/https://doi.org/10.1016/j.idairyj.2021.105051.

Fao, F. (2024). La production laitière et les produits laitiers : La santé animale. Retrieved from https://www.fao.org/dairy-production-products/production/animal-health/fr/.

Favole, A., Testori, C., Bergagna, S., Gennero, M. S., Ingravalle, F., Costa, B., Barresi, S., Curti, P., Barberis, F., Ganio, S., Orusa, R., Vallino Costassa, E., Berrone, E., Vernè, M., Scaglia, M., Palmitessa, C., Gallo, M., Tessarolo, C., Pederiva, S., Ferrari, A., Lorenzi, V., Fusi, F., Brunelli, L., Pastorelli, R., Cagnotti, G., Casalone, C., Caramelli, M. et Corona, C. (2023). Brain-derived neurotrophic factor, kynurenine pathway, and lipid-profiling alterations as potential animal welfare indicators in dairy cattle. *Animals*, *13*(7). http://dx.doi.org/10.3390/ani13071167.

Filzinger, T. (2023). Ingénierie des fonctionnalités : des données brutes à l'ensemble de formation.

Konfuzio. Récupéré de https://konfuzio.com/fr/feature-engineering/#:~:text=de%20mani%C3%A8re%20discr%C3%A8te.-,Mise%20%C3%A0%20l'%C3%A9chelle,et%20donc%20%C3%A0%20les%20normaliser.

Foroutan, A., Fitzsimmons, C., Mandal, R., Piri-Moghadam, H., Zheng, J., Guo, A., Li, C., Guan, L. L. et Wishart, D. S. (2020). The bovine metabolome. *Metabolites*, *10*(6). http://dx.doi.org/10.3390/metabo10060233.

Gandhi, R. (2018). Support vector machine — introduction to machine learning algorithms. *Medium ; Towards Data Science*. Récupéré de https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

Gao, Q., Praticò, G., Scalbert, A., Vergères, G., Kolehmainen, M., Manach, C., Brennan, L., Afman, L., Wishart, D., Andres-Lacueva, C., Garcia-Aloy, M., Verhagen, H., Feskens, E. et Dragsted, L. (2017). A scheme for a flexible classification of dietary and health biomarkers. *Genes Nutrition*, *12*. http://dx.doi.org/10.1186/s12263-017-0587-x

Garzon, A., Habing, G., Lima, F., del Rio, N. S., Samah, F. et Pereira, R. (2022). Defining clinical diagnosis and treatment of puerperal metritis in dairy cows : A scoping review. *Journal of Dairy Science*, *105*(4), 3440–3452. http://dx.doi.org/https://doi.org/10.3168/jds.2021-21203.

Genesis (2018). Pros and cons of k-nearest neighbors. From The GENESIS. Récupéré de https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/

Ghaffari, M. H., Jahanbekam, A., Sadri, H., Schuh, K., Dusel, G., Prehn, C., Adamski, J., Koch, C. et Sauerwein, H. (2019). Metabolomics meets machine learning : Longitudinal metabolite profiling in serum of normal versus overconditioned cows and pathway analysis. *Journal of Dairy Science*, *102*(12), 11561–11585.

Ghoddusi, H., Creamer, G. G. et Rafizadeh, N. (2019). Machine learning in energy economics and finance : A review. *Energy Economics*, *81*, 709–727. http://dx.doi.org/https://doi.org/10.1016/j.eneco.2019.05.006.

Giannuzzi, D., Mota, L. F. M., Pegolo, S., Gallo, L., Schiavon, S., Tagliapietra, F., Katz, G., Fainboym, D., Minuti, A., Trevisi, E. et Cecchinato, A. (2022). In-line near-infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. *Scientific Reports*, *12*(1), 8058. http://dx.doi.org/10.1038/s41598-022-11799-0.

Gonçalves Frasco, C., Radmacher, M., Lacroix, R., Cue, R., Valtchev, P., Robert, C., Boukadoum, M., Sirard, M.-A. et Diallo, A. B. (2020). Towards an effective decision-making system based on cow profitability using deep learning. 949–958. http://dx.doi.org/10.5220/0009174809490958.

Gouvernement du Québec (2019). Production laitière (lait de vache). Retrieved from https://www.quebec.ca/agriculture-environnement-et-ressources-naturelles/agriculture/industrie-agricole-au-quebec/productions-agricoles/production-lait-vache.

Grandini, M., Bagli, E. et Visani, G. (2020). Metrics for multi-class classification : an overview. *ArXiv*, *abs/2008.05756*. Récupéré de https://api.semanticscholar.org/CorpusID:221112671.

Grelet, C., Vanden Dries, V., Leblois, J., Wavreille, J., Mirabito, L., Soyeurt, H., Franceschini, S., Gengler, N., Brostaux, Y., HappyMoo Consortium et Dehareng, F. (2022). Identification of chronic stress biomarkers in dairy cows. *animal*, *16*(5), 100502. http://dx.doi.org/https://doi.org/10.1016/j.animal.2022.100502.

Géron, A. (2020). *Deep Learning avec Keras et TensorFlow : Mise en œuvre et cas concrets*. DUNOD.

Hailemariam, D., Zhang, G., Mandal, R., Wishart, D. S. et Ametaj, B. N. (2018). Identification of serum metabolites associated with the risk of metritis in transition dairy cows. *Canadian Journal of Animal Science*, *98*(3), 525–537. http://dx.doi.org/10.1139/cjas-2017-0069.

He, W., Cardoso, A. S., Hyde, R. M., Green, M. J., Scurr, D. J., Griffiths, R. L., Randall, L. V. et Kim, D.-H. (2022). Metabolic alterations in dairy cattle with lameness revealed by untargeted metabolomics of dried milk spots using direct infusion-tandem mass spectrometry and the triangulation of multiple machine learning models. *Analyst*, *147*, 5537–5545. http://dx.doi.org/10.1039/D2AN01520J.

Huang, H. N., Chen, H. M., Lin, W. W., Huang, C. J., Chen, Y. C., Wang, Y. H. et Yang, C. T. (2023). Employing feature engineering strategies to improve the performance of machine learning algorithms on echocardiogram dataset. *Digital health*, *9*, 20552076231207589. http://dx.doi.org/10.1177/20552076231207589.

Hung, H. M. J., O'Neill, R. T., Bauer, P. et Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, *53*(1), 11–22. Récupéré le 2024-01-23 de http://www.jstor.org/stable/2533093.

Hyde, R. M., Down, P. M., Bradley, A. J., Breen, J. E., Hudson, C., Leach, K. A. et Green, M. J. (2020). Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Scientific Reports*, *10*(1), 4289. http://dx.doi.org/10.1038/s41598-020-61126-8. Récupéré de https://doi.org/10.1038/s41598-020-61126-8

Häring, A. M. (2003). Organic dairy farms in the eu : Production systems, economics and future development. *Livestock Production Science*, *80*(1), 89–97. Organic Livestock Production, http://dx.doi.org/https://doi.org/10.1016/S0301-6226(02)00308-1.

Institut National de Santé Publique Du Québec (2023). Les risques à la santé publique associés aux activités de production animale. https://www.inspq.qc.ca/bise/les-risques-la-sante-publique-associes-aux-activites-de-production-animale.

Investir Sorcier (2021). Définition de la règle empirique. https://www.investirsorcier.com/definition-de-la-regle-empirique/#google_vignette.

Jain, K. K. (2008). *Role of Biomarkers in Personalized Medicine*. New York, NY : Humana Press. http://dx.doi.org/10.1007/978-1-4419-0769-1_3.

Johnson, N. L. (1969). *Discrete Distributions*. Houghton Mifflin Company.

Kares, M. (2022). Influence of dairy farming practices on milk production. a critical literature review. *Animal Health Journal*, *3*(1), 1–15. http://dx.doi.org/10.47941/ahj.771.

Karoui, Y., Boatswain Jacques, A. A., Diallo, A. B., Shepley, E. et Vasseur, E. (2021). A deep learning framework for improving lameness identification in dairy cattle. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15811–15812. http://dx.doi.org/10.1609/aaai.v35i18.17902.

Kassel, R. (2020). Réseau de neurones : définition et fonctionnement. *Formation Data Science | DataScientest.com*. Récupéré de https://datascientest.com/fonctionnement-des-reseaux-neurones.

Kristan, M., Leonardis, A. et Skočaj, D. (2011). Multivariate online kernel density estimation with gaussian kernels. *Pattern Recognition*, *44*(10), 2630–2642. http://dx.doi.org/https://doi.org/10.1016/j.patcog.2011.03.019.

Kuhn, M. et Johnson, K. (2019). *Feature Engineering and Selection : A Practical Approach for Predictive Models* (1st éd.). Chapman and Hall/CRC. Récupéré de https://doi.org/10.1201/9781315108230.

Kunwar, A. (2020). Missing data imputation in feature engineering. Medium. Récupéré de https://medium.com/swlh/missing-data-imputation-in-feature-engineering-aeefd03ba58d.

kwamimayeden (2022). La validation croisée en machine learning. Kwami Mayeden. Récupéré de https://kwamimayeden.com/la-validation-croisee-en-machine-learning/.

La Santé Des Ruminants (2023). Des biomarqueurs pour diagnostiquer acidose et cétose subcliniques. https://www.la-sante-des-ruminants.fr/le-lait/metabolique-peripartum/des-biomarqueurs-pour-diagnostiquer-acidose-et-cetose-subcliniques/.

Lasser, J., Matzhold, C., Egger-Danner, C., Fuerst-Waltl, B., Steininger, F., Wittek, T. et Klimek, P. (2021). Integrating diverse data sources to predict disease risk in dairy cattle. http://dx.doi.org/10.1101/2021.03.25.436798.

LEBLANC, S. (2010). Monitoring metabolic health of dairy cattle in the transition period. *Journal of Reproduction and Development*, *56*(S), S29–S35. http://dx.doi.org/10.1262/jrd.1056S29.

Les Producteurs laitiers du Canada, P. l. d. C. Les clés de la prévention des maladies et les impacts économiques si elles ne sont pas maîtrisées. PDF. Récupéré de https://www.producteurslaitiers.ca/Media/Files/proaction/Cles_prevention_maladies_impacts.pdf.

Les Producteurs laitiers du Canada, P. l. d. C. (2023). Manuel de référence juillet 2023.

Li, Z., Zhang, J., Gong, Y., Yao, Y. et Wu, Q. (2020). Field-wise learning for multi-field categorical data.

Limpert, E., Stahel, W. A. et Abbt, M. (2001). Log-normal Distributions across the Sciences : Keys and Clues : On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal : That is the question. *BioScience*, *51*(5), 341–352.

http://dx.doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.

Loffler-Laurian, A.-M. (1996). *La traduction automatique*. Presses Univ. Septentrion.

Lopez, P. D. (2021). La cétose de la vache laitière. https://www.lacompagniedesanimaux.com/conseil-veterinaire/la-cetose-de-la-vache-laitiere.html.

Macmillan, K., Gobikrushanth, M., Behrouzi, A., López-Helguera, I., Cook, N., Hoff, B. et Colazo, M. (2020). The association of circulating prepartum metabolites, minerals, cytokines and hormones with postpartum health status in dairy cattle. *Research in Veterinary Science*, *130*, 126–132. http://dx.doi.org/https://doi.org/10.1016/j.rvsc.2020.03.011.

Mansour, U. M., Belal, H. E. S. et Dohreig, R. M. (2021). Biomarkers for negative energy balance and fertility in early lactating dairy cows. *German journal of veterinary research*, *2*(2), 11–16. http://dx.doi.org/10.51585/gjvr.2022.2.0031.

Marchand, C. R., Farshidfar, F., Rattner, J. et Bathe, O. F. (2018). A framework for development of useful metabolomic biomarkers and their effective knowledge translation. *Metabolites*, *8*(4). http://dx.doi.org/10.3390/metabo8040059.

Massaro, A., Tata, A., Pallante, I., Bertazzo, V., Bottazzari, M., Paganini, L., Dall'Ava, B., Stefani, A., De Buck, J., Piro, R. et Pozzato, N. (2023). Metabolic signature of mycobacterium avium subsp. paratuberculosis infected and infectious dairy cattle by integrating nuclear magnetic resonance analysis and blood indices. *Frontiers in Veterinary Science*, *10*. http://dx.doi.org/10.3389/fvets.2023.1146626.

Mathworld (2024). Binomial distribution. Wolfram.com. Récupéré de https://mathworld.wolfram.com/BinomialDistribution.html.

Matyka, M. et Koza, Z. (2012). How to calculate tortuosity easily ? *AIP Conference Proceedings*, *1453*, 17–22. http://dx.doi.org/10.1063/1.4711147.

Ministère de l'Agriculture et de la Souveraineté Alimentaire (2019). Le bien-être et la protection des vaches laitières. https://agriculture.gouv.fr/le-bien-etre-et-la-protection-des-vaches-laitieres.

Miola, A. C. et Miot, H. A. (2022). Comparing categorical variables in clinical and experimental studies. *Jornal Vascular Brasileiro*, *21*, e20210225. http://dx.doi.org/10.1590/1677-5449.20210225.

Mota, L. F. M., Giannuzzi, D., Pegolo, S., Trevisi, E., Ajmone-Marsan, P. et Cecchinato, A. (2023). Integrating on-farm and genomic information improves the predictive ability of milk infrared prediction of blood indicators of metabolic disorders in dairy cows. *Genetics Selection Evolution*, *55*(1), 23. http://dx.doi.org/10.1186/s12711-023-00795-1.

Motohashi, H., Ohwada, H. et Kubota, C. (2020). Early detection method for subclinical mastitis in auto milking systems using machine learning. Dans *2020 IEEE 19th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, 76–83. http://dx.doi.org/10.1109/ICCICC50026.2020.9450258.

M@XCode (2016). Machine learning : classification à l'aide des arbres de décisions : fonctionnement et application en nodejs. M@XCode. Récupéré de :maximilienandile.github.io/2016/10/17/.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, *32*, 685 – 694. http://dx.doi.org/10.1080/02664760500079464.

Naghashi, V., Dallago, G., Diallo, A. et Boukadoum, M. (2023). Univariate and multivariate time-series methods to forecast dairy income.

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B. et Turaga, D. S. (2017). Learning feature engineering for classification. Dans *IJCAI*, volume 17, 2529–2535.

Nejati, A., Shepley, E., Dallago, G. et Vasseur, E. (2024). Investigating the impact of 1h daily outdoor access on the gait and hoof health of non-clinically lame cows housed in a movement restricted environment. *JDS Communications*. http://dx.doi.org/https://doi.org/10.3168/jdsc.2023-0498.

Ojo, O. E., Hajek, L., Johanns, S., Pacífico, C., Sener-Aydemir, A., Ricci, S., Rivera-Chacon, R., Castillo-Lopez, E., Reisinger, N., Zebeli, Q. et Kreuzer-Redmer, S. (2023). Evaluation of circulating microrna profiles in blood as potential candidate biomarkers in a subacute ruminal acidosis cow model - a pilot study. *BMC Genomics*, *24*(1), 333. http://dx.doi.org/10.1186/s12864-023-09433-y.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O. et Akinjobi, J. (2017). International journal of computer trends and technology. *48*(3), 128–136. Récupéré de https://doi.org/10.14445/22312803/IJCTT-V48P126.

Ozella, L., Brotto Rebuli, K., Forte, C. et Giacobini, M. (2023). A literature review of modeling approaches applied to data collected in automatic milking systems. *Animals*, *13*(12). http://dx.doi.org/10.3390/ani13121916.

Pan, C., Poddar, A., Mukherjee, R. et Ray, A. K. (2022). Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction. *Biomedical Signal Processing and Control*, *76*, 103666. http://dx.doi.org/https://doi.org/10.1016/j.bspc.2022.103666.

Panchal, I., Sawhney, I. K. et Sharma, A. K. (2015). Identifying healthy and mastitis sahiwal cows using electro-chemical properties : A connectionist approach. Dans *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1185–1188.

Paul, S., Ranjan, P., Kumar, S. et Kumar, A. (2022). Disease predictor using random forest classifier. Dans *2022 International Conference for Advancement in Technology (ICONAT)*, 1–4. http://dx.doi.org/10.1109/ICONAT53423.2022.9726023.

Popov, A. (2023). 1 - feature engineering methods. In K. Pal, S. Ari, A. Bit, et S. Bhattacharyya (dir.), *Advanced Methods in Biomedical Signal Processing and Analysis* 1–29. Academic Press

Producteurs de Lait du Québec (2014). La qualité du lait. https://lait.org/la-ferme-en-action/la-qualite-du-lait/.

Pyorala, S. (2003). Indicators of inflammation in the diagnosis of mastitis. *Veterinary research*, *34 5*, 565–78. http://dx.doi.org/10.1051/VETRES:2003026.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. Récupéré de https://api.semanticscholar.org/CorpusID:13252401.

Raileanu, L. E. et Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*, 77–93. Récupéré de https://api.semanticscholar.org/CorpusID:207658568.

Rakotomalala, R. (2008). Tests de normalité. *Université Lumière Lyon*.

Roy, R., Baral, M. M., Pal, S. K., Kumar, S., Mukherjee, S. et Jana, B. (2022). Discussing the present, past, and future of machine learning techniques in livestock farming : A systematic literature review. Dans *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, volume 1, 179–183. http://dx.doi.org/10.1109/COM-IT-CON54601.2022.9850749.

Safavian, S. et Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674. http://dx.doi.org/10.1109/21.97458.

Sammut, C. et Webb, G. I. (2017). *Unsupervised Learning*. Boston, MA : Springer US. http://dx.doi.org/10.1007/978-1-4899-7687-1_976.

Sanger, T. et Baljekar, P. N. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, *65 6*, 386–408. Récupéré de https://api.semanticscholar.org/CorpusID:12781225.

Sarle, W. (1994). Neural networks and statistical models. Récupéré de https://api.semanticscholar.org/CorpusID:2562349.

Sasaki, Y. (2007). The truth of the f-measure. *Teach Tutor Mater*.

Schwegler, Elizabeth, S., Augusto, M. et Paula, Acosta, D. A. V. (2013). Predictive value of prepartum serum metabolites for incidence of clinical and subclinical mastitis in grazing primiparous holstein cows. *Tropical Animal Health and Production*, *16*(2), 1549–1555. http://dx.doi.org/10.1007/s11250-013-0398-z

Sinharay, S. (2023). Discrete probability distributions. In R. J. Tierney, F. Rizvi, et K. Ercikan (dir.), *International Encyclopedia of Education (Fourth Edition)* 718–722. Oxford : Elsevier, (fourth edition éd.)

Skelly, D. (2023). Fièvre du lait chez les vaches - causes, symptômes et traitement. Récupéré de https://www.moocall.com/fr/milk-fever-in-cows/.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*,*3*.

Tanyildizl, E. et Yildirim, G. (2019). Performance comparison of classification algorithms for the diagnosis

of mastitis disease in dairy animals. Dans *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, 1–4. http://dx.doi.org/10.1109/ISDFS.2019.8757469.

Tata, A., Massaro, A., Riuzzi, G., Lanza, I., Bragolusi, M., Negro, A., Novelli, E., Piro, R., Gottardo, F. et Segato, S. (2022). Ambient mass spectrometry for rapid authentication of milk from alpine or lowland forage. *Scientific Reports*, *12*(1), 7360. http://dx.doi.org/10.1038/s41598-022-11178-9.

Tremblay, C. (2022). Smote : comprendre l'algorithme et connaître les 5 règles pour mieux l'utiliser. Kobia. Récupéré de https://kobia.fr/imbalanced-data-smote/#:~:text=Le%20SMOTE%2C%20acronyme%20pour%20Synthetic,de%20sur%C3%A9chantillonnage%20des%20observations%20minoritaires.

Trevor Hastie, Robert Tibshirani, J. F. (2009). The elements of statistical learning.

Van Nuffel, A., Zwertvaegher, I., Pluym, L., Van Weyenberg, S., Thorup, V. M., Pastell, M., Sonck, B. et Saeys, W. (2015). Lameness detection in dairy cows : Part 1. how to distinguish between non-lame and lame cows based on differences in locomotion or behavior. *Animals*, *5*(3), 838–860. http://dx.doi.org/10.3390/ani5030387.

Vilanova, C. et Porcar, M. (2016). Are multi-omics enough ? *Nature Microbiology*, *1*(8), 16101. http://dx.doi.org/10.1038/nmicrobiol.2016.101.

Vázquez-Diosdado, J., Miguel-Pacheco, G., Plant, B., Dottorini, T., Green, M. et Kaler, J. (2019). Developing and evaluating threshold-based algorithms to detect drinking behavior in dairy cows using reticulorumen temperature. *Journal of Dairy Science*, *102*(11), 10471–10482. http://dx.doi.org/https://doi.org/10.3168/jds.2019-16442. Récupéré de https://www.sciencedirect.com/science/article/pii/S0022030219307246

Vázquez-Diosdado, J. A., Gruhier, J., Miguel-Pacheco, G., Green, M., Dottorini, T. et Kaler, J. (2023). Accurate prediction of calving in dairy cows by applying feature engineering and machine learning. *Preventive Veterinary Medicine*, *219*, 106007. http://dx.doi.org/https://doi.org/10.1016/j.prevetmed.2023.106007.

Wagner, N., Antoine, V., Koko, J., Mialon, M.-M., Lardy, R. et Veissier, I. (2021). Comparaison de méthodes d'apprentissage automatique pour détecter les anomalies dans l'activité des animaux. Dans *EGC*, 521–522.

Wang, G.-Q., Zheng, H.-Y., Hou, J.-L., Wang, C., Sun, H.-L. et Wang, L. (2023). The role of leukotriene b in cow metritis. *Journal of Veterinary Research*, *67*(1), 99–104. http://dx.doi.org/doi:10.2478/jvetres-2023-0011.

Wang, J., Zhang, Y., Wang, J., Zhao, K., Li, X. et Liu, B. (2022). Using machine-learning technique for estrus onset detection in dairy cows from acceleration and location data acquired by a neck-tag. *Biosystems Engineering*, *214*, 193–206.

Wells, S., Ott, S. et Seitzinger, A. H. (1998). Key health issues for dairy cattle—new and old. *Journal of Dairy Science*, *81*(11), 3029–3035.

Yan, G., Zhao, W., Wang, C., Shi, Z., Li, H., Yu, Z., Jiao, H. et Lin, H. (2024). A comparative study of machine learning models for respiration rate prediction in dairy cows : Exploring algorithms, feature engineering, and model interpretation. *Biosystems Engineering*, *239*, 207–230. http://dx.doi.org/https://doi.org/10.1016/j.biosystemseng.2024.01.010.

Ye, Q., Yang, X., Chen, C. et Wang, J. (2019). River water quality parameters prediction method based on lstm-rnn model. Dans *2019 Chinese Control And Decision Conference (CCDC)*, 3024–3028. http://dx.doi.org/10.1109/CCDC.2019.8832885

Yun, H., Sun, L., Wu, Q., Luo, Y., Qi, Q., Li, H., Gu, W., Wang, J., Ning, G., Zeng, R., Zong, G. et Lin, X. (2022). Lipidomic signatures of dairy consumption and associated changes in blood pressure and other cardiovascular risk factors among chinese adults. *Hypertension*, *79*(8), 1617–1628. http://dx.doi.org/10.1161/HYPERTENSIONAHA.122.18981. Récupéré de https://doi.org/10.1161/HYPERTENSIONAHA.122.18981

Zhang, G., Q, D., Mandal, R., Wishart, D. et Ametaj, B. (2017). Metabolic profiling for identification of early predictive serum biomarkers of metritis in transition dairy cows. http://dx.doi.org/10.1021/acs.jafc.7b02000.

Zhang, G., Zwierzchowski, G., Mandal, R., Wishart, D. et Ametaj, B. (2020). Serum metabolomics identifies metabolite panels that differentiate lame dairy cows from healthy ones. *Metabolomics*, *16*(6). http://dx.doi.org/10.1007/s11306-020-01693-z.

Zhang, X., Liu, T., Hou, X., Hu, C., Zhang, L., Wang, S., Zhang, Q. et Shi, K. (2022). Multi-channel metabolomics analysis identifies novel metabolite biomarkers for the early detection of fatty liver disease in dairy cows. *Cells*, *11*(18). http://dx.doi.org/10.3390/cells11182883.

Zheng, A. et Casari, A. (2018). *Feature engineering for machine learning : principles and techniques for data scientists*. " O'Reilly Media, Inc.".

Éleveurs de demain - Le blog (2020). Éleveurs de demain. https://www.eleveursdedemain.fr/.