UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

Prédiction des Troubles de Santé Mentale à partir des Habitudes de Vie : Une Approche par Apprentissage Machine

Auteur : De Meester Julien et de Nauw Doryan
Promoteur : Vande Kerckhove
Année académique 2023-2024
Master [120] : ingénieur de gestion

Declaration Regarding AI Tool Usage in Master's Thesis

We recognize that AI tools might be valuable aids during the master's thesis work, but they are not infallible. Remember that transparency fosters trust, and acknowledging AI's role enhances the credibility of your work.

Therefore, when deciding to use such a tool, you need to adhere to the following principles of responsible use of AI.

1. **Critical Evaluation :**
    - We critically assessed the AI-generated output, ensuring its alignment with our research objectives. ○ Any modifications or corrections were made based on our expertise and domain knowledge.
2. **Transparency :**
    - We acknowledge the use of chatgpt/consensus/co-pilot transparently, emphasizing that it contributed to our work but did not replace human judgment.
    - Our commitment to transparency ensures the integrity of this thesis.
3. **Ethical Considerations :**
    - We actively monitored for biases or unintended consequences introduced by the AI tool.
    - Our ethical responsibility guided our decisions throughout the research process.

Declaration **(This declaration is mandatory and must appear on the first page (after the title page) of the document.**

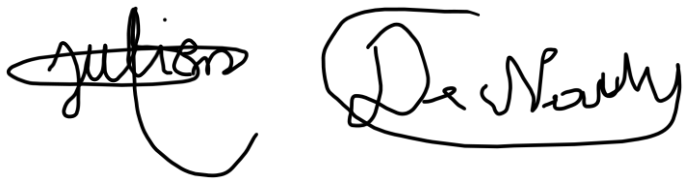
During the preparation of this master's thesis, the authors utilized chatgpt, consensus, github co-pilot for the following purpose:

1. Chatgpt to help us find mistakes in the written text and translate some text to help us understand them.
2. Consensus to help us find sources for the literature review.
3. Co-pilot to write the final piece of code especially lightgmboost and xgboost.

After using chatgpt/consensus/co-pilot, the authors diligently reviewed and edited the content produced by the tool. We take full responsibility for the final content presented in this thesis.

By signing this declaration, we affirm that the content of this master's thesis reflects our original work, augmented by the responsible use of AI.

07/08/2024



REMERCIMENTS

De Meester Julien

Ce mémoire, bien plus qu'un travail universitaire, est le reflet d'un cheminement, d'une collaboration, et surtout, d'un soutien indéfectible.

À ma famille,

Un merci tout particulier à ma mère et mes deux sœurs. Vos conseils, même si souvent ignoré (à tort je m'en suis bien rendu compte), ont toujours été une bouée de sauvetage. Votre patience et votre amour m'ont donné la force de persévérer.

À mon promoteur, Corentin Vande Kerckhove,

Votre passion contagieuse pour la recherche, votre expertise dans notre domaine, votre disponibilité sans faille et vos conseils avisés ont été inestimables. Vous nous avez poussés à dépasser nos limites, et pour cela, je vous suis profondément reconnaissant.

À l'UCLouvain,

Merci d'avoir mis à notre disposition un environnement d'apprentissage et de recherche aussi stimulant. Les ressources et la structure offertes ont été déterminantes dans la réalisation de ce mémoire.

Et enfin, à mon co-mémorant, Doryan,

Que dire ? Tu as été bien plus qu'un partenaire de travail. Ton esprit vif, ta capacité à déchiffrer les demandes les plus complexes (surtout celles de Corentin !) et ton soutien sans faille ont rendu cette aventure possible. Écrire, coder, débugger, rire (et parfois pleurer) à tes côtés restera un souvenir précieux. Sans toi, ce mémoire n'aurait tout simplement pas été le même.

De Nauw Doryan

Je tiens à exprimer mes plus sincères remerciements à toutes les personnes qui ont contribué à la réalisation de ce mémoire.

Tout d'abord, je souhaite remercier chaleureusement mon co-mémorant Julien. Ton soutien et la manière dont tu pensais pour résoudre certains problèmes et aboutir à certaines idées ont été une source inestimable de motivation et d'inspiration tout au long de ce projet. Ta présence et l'amitié qui nous unit ont permis de réaliser ce travail. Sans toi ce travail terminé ne m'aurait pas donné le sentiment de fierté que je ressens actuellement. Je te remercie pour tous ces moments en ta compagnie.

Je tiens également à remercier mon frère, dont la complicité et l'affection ont toujours été un pilier de ma vie. Ta compréhension et ton soutien constant ont grandement contribué à la réussite de ce projet.

Je suis profondément reconnaissant envers mes parents, dont l'amour et les sacrifices ont rendu tout cela possible. Votre confiance en moi et votre soutien m'ont donné la force de poursuivre mes rêves. Merci pour tout ce que vous avez fait et continuez de faire pour moi.

Je souhaite également exprimer ma gratitude à mon promoteur Corentin Vande Kerckhove. Votre expertise, vos conseils avisés et votre disponibilité ont été essentiels à la réalisation de ce mémoire. Merci pour votre accompagnement et votre soutien tout au long de ce parcours académique.

Un remerciement particulier à l'Université Laval pour les bases de données qu'elle nous a fournies. Sans ces ressources essentielles, ce travail n'aurait pas pu être réalisé. Votre contribution a été déterminante pour la réussite de ce projet.

Enfin, je remercie l'Université catholique de Louvain (UCL) pour l'opportunité et les ressources mises à disposition pour mener à bien ce projet.

À tous, je vous adresse mes remerciements les plus sincères.

Table des matières

INTRODUCTION

La santé mentale est un pilier essentiel du bien-être global, influençant non seulement la qualité de vie, mais également la productivité, les relations sociales et la capacité à faire face aux défis quotidiens de la vie. Cependant, dans la société actuelle, la santé mentale demeure stigmatisée, souvent délaissé par rapport à la santé physique. Malgré la prévalence croissante des problèmes de santé mentale, le domaine souffre d'une insuffisance des ressources allouées à la prévention, au diagnostic et aux traitements de ces différents troubles. (Sartorius, 2007)

Selon l'Organisation mondiale de la santé (OMS), en 2022, Une personne sur huit souffre d'un trouble de santé mentale. Les troubles mentaux tels que la dépression, l'anxiété, le stress post-traumatique et les troubles bipolaires, entre autres, constituent des défis majeurs pour les systèmes de santé publique à travers le monde. Ces troubles peuvent être déclenchés ou exacerbés par divers facteurs, notamment les conditions socio-économiques, les événements traumatiques, les antécédents familiaux et le stress professionnel(WHO, 2022)

Ce mémoire a pour objectif de contribuer à la détection précoce des personnes susceptibles d'être atteintes par un trouble de la santé mentale. Pour ce faire nous utiliserons des données issues de deux questionnaires administrés par l'université Laval. Les données issues de ces deux projet ULA et ESSAIM serviront, à terme, à créer un système de recommandation d'habitude de vie pour améliorer la santé mentale. Ce mémoire se positionne en amont de ce travail avec pour but d'aider à identifier les habitudes de vies essentielles permettant de créer ce système de recommandation.

La complexité des interactions entre différents facteurs rend la tâche d'identifier les éléments influençant fortement la santé mentale particulièrement ardue. C'est dans ce contexte que les approches basées sur les données et l'apprentissage machine offrent un potentiel significatif. L'analyse des réponses à des questionnaires bien conçus à l'aide de techniques avancées d'apprentissage machine peut révéler des schémas et des facteurs influents qui ne seraient pas évidents autrement. Cependant, pour que ces analyses soient utiles et applicables dans des contextes réels, il est crucial que les méthodes utilisées soient non seulement précises, mais également explicables.(Davenport & Kalakota, 2019)

Dans la section suivante, nous passerons en revue la littérature existante sur le sujet. Nous explorerons les méthodes de l'apprentissage machine couramment utilisées pour la classification et la prédiction, ainsi que les approches spécifiques à la détection des troubles de santé mentale. Nous examinerons également les défis liés à l'utilisation de données de questionnaires et les stratégies pour garantir la validité et la fiabilité des résultats.

Ensuite, nous présenterons nos résultats. Nous détaillerons les modèles d'apprentissage machine que nous avons développée pour prédire la présence de troubles de santé mentale en utilisant les données des questionnaires. Nous évaluerons les performances de ces modèles en termes de précision, de sensibilité et de spécificité. De plus, nous utiliserons des techniques d'interprétation pour comprendre les facteurs qui contribuent le plus à la prédiction, ce qui pourrait fournir des informations précieuses pour la prévention et le traitement.

Enfin, nous discuterons des limites de notre étude et des implications de nos résultats pour la recherche future et la pratique clinique. Nous soulignerons les défis à relever pour améliorer la détection précoce des troubles de santé mentale et proposerons des pistes pour l'intégration de nos modèles dans la pratique clinique.

PARTIE 1) REVUE DE LITTÉRATURE

Cette revue de littérature sera divisée en 5 grandes sections. Tout d'abord, nous essaierons de faire un tour d'horizon sur le sujet de la santé mentale. Nous tenterons de définir ce qu'est un problème de santé mentale. Ensuite, nous nous intéresserons aux manières de détecter un problème de santé mentale. Nous essaierons notamment de voir comment on peut détecter un problème de santé mentale à l'aide de variable quantitative, plus facile à intégrer dans des modèles d'apprentissage supervisé.

Pour continuer, nous explorerons les méthodes d'apprentissage machines. Ici, nous nous focaliserons sur des méthodes supervisées comme l'arbre de décision, la régression logistique. Nous détaillerons également les méthodes de sélection de caractéristiques qui seront utilisées dans ce mémoire. Nous définirons en général comment sont constituées ces méthodes et leur but. Puis, nous nous attarderons sur une série de méthodes précises que nous voulons utiliser dans le cadre de ce travail.

Ensuite, nous mettrons en avant les différentes manières dont l'IA a été utilisée pour prévenir la santé mentale.

De plus, nous définirons nos différentes questions de recherches en fonction de ce qui n'a pas été vu dans la revue de littérature et de ce que nous pouvons apporter en plus.

Enfin, nous ferons un résumé de ce qui a été vu dans cette revue de littérature.

1.1) TROUBLES DE SANTÉ MENTALE

1.1.1) INTRODUCTION ET DÉFINITION

Afin d'appliquer les méthodes de sciences des données à la santé mentale, il est essentiel de comprendre ces problèmes. Pour cette raison, il est essentiel de présenter une analyse détaillée des problèmes de santé mentale dans un premier temps. Ensuite nous verrons comment on détecte un problème de santé mentale, notamment via une approche quantitative, plus facile à intégrer aux modèles d'apprentissage supervisé. Cela établi, nous explorerons comment utiliser les habitudes de vie pour détecter un problème de santé mentale.

Selon l'organisation mondiale de la santé : « La santé mentale correspond à un état de bien-être mental qui nous permet d'affronter les sources de stress de la vie, de réaliser notre potentiel, de bien apprendre et de bien travailler, et de contribuer à la vie de la communauté »(WHO, 2024).Cette définition fait de l'état de bien être la clé pour définir la santé mentale. Plusieurs auteurs apportent cependant une nuance. Cet état de bien-être n'est pas suffisant pour définir la santé mentale. En effet, si une personne se sent euphorique en commettant une série de meurtre, l'on ne peut pas admettre que cette personne est dans un bon état mental. La littérature sur le sujet met donc 3 formes de bien être en avant : le bien être psychologique, le bien être émotionnelle et le bien être sociale. Ces 3 notions sont importantes pour conclure qu'une personne est en bonne santé mentale.(Galderisi et al., 2015; Keyes, 2013; Waterman, 1993)

Pour généraliser l'identification de trouble psychologique et l'objectiver il est utile de mettre au point des méthodes quantitative de mesure de la santé mentale. On retrouve 5 grandes méthodes permettant d'avoir des indicateurs quantitatifs.

1. Questionnaires standardisés : Les questionnaires auto-administrés ou administrés par un professionnel sont couramment utilisés pour mesurer divers aspects de la santé mentale.(Brazier et al., 2014)
2. Échelles de stress perçu : Des échelles comme le Perceived Stress Scale (PSS) évaluent le niveau de stress perçu par l'individu.(Brazier et al., 2014)
3. Données biométriques : Utilisation de données telles que la fréquence cardiaque, la variabilité de la fréquence cardiaque, et les niveaux de cortisol pour inférer l'état de stress et d'anxiété.(Shatte et al., 2019)

4. Analyses de données comportementales : Données provenant de dispositifs portables (wearables), comme les montres connectées, qui suivent l'activité physique, le sommeil, et d'autres indicateurs comportementaux pouvant être liés à la santé mentale.(Shatte et al., 2019)
5. Analyses de données de réseaux sociaux : Des algorithmes de traitement du langage naturel (NLP) peuvent analyser les posts sur les réseaux sociaux pour détecter des signes de détresse mentale.(Shatte et al., 2019)

Dans le cadre de ce mémoire on va particulièrement s'intéresser aux méthodes touchant aux « questionnaires standardisés » et aux données comportementales. Penchons-nous donc sur ces deux méthodes leurs points fort et leurs faiblesses.

1.1.2) EVALUER LES TROUBLES DE SANTÉ MENTALE A L'AIDE DE QUESTIONNAIRES STANDARDISÉES

Les questionnaires standardisés sont une méthode classique pour évaluer la santé mentale d'un patient. Il en existe de plusieurs types pour évaluer différents troubles de la santé mentale (dépression, anxiété, trouble du sommeil…). Ces questionnaires consistent en une série de questions administrées aux patients afin d'évaluer leur bien être par rapport à une norme. La validité de cette méthode pour évaluer la santé mentale a été prouvée empiriquement. Il existe plusieurs types de questionnaires utilisés pour évaluer différents aspects de la santé mentale :

Questionnaires de Dépistage : Ces questionnaires sont conçus pour identifier rapidement les individus susceptibles de présenter des troubles de santé mentale. Exemples courants :

1. PHQ-9 (Patient Health Questionnaire-9) : Utilisé pour dépister la dépression, il comprend neuf questions basées sur les critères de diagnostic du DSM-IV(Brazier et al., 2014).
2. GAD-7 (Generalized Anxiety Disorder-7) : Un outil de dépistage pour l'anxiété généralisée, comprenant sept questions(Brazier et al., 2014).

Inventaires de Symptômes : Ces questionnaires évaluent la gravité des symptômes spécifiques à un trouble mental donné. Exemples :

1. Beck Depression Inventory (BDI) : Évalue la gravité de la dépression à travers 21 items(Elovanio et al., 2020).
2. State-Trait Anxiety Inventory (STAI) : Mesure l'anxiété état (temporaire) et l'anxiété trait (durable) via deux sous-échelles distinctes(Newson et al., 2020).
3. Questionnaires de Qualité de Vie : Ces outils évaluent l'impact des troubles de santé mentale sur la qualité de vie globale de l'individu(Elovanio et al., 2020).

Les questionnaires auto-administrés présentent plusieurs avantages clés dans l'évaluation de la santé mentale. Tout d'abord ils sont faciles à administrer. Ces questionnaires peuvent être remplis par les individus eux-mêmes, ce qui réduit la nécessité d'une intervention directe d'un professionnel de santé. Ils peuvent être distribués à un large public via des formats en ligne, sur papier, ou via des applications mobiles, rendant l'évaluation accessible même dans des régions éloignées. Les répondants peuvent remplir les questionnaires dans un cadre privé, ce qui peut réduire l'effet de la désirabilité sociale et encourager des réponses honnêtes. Finalement les questionnaires sont standardisés, ce qui permet des comparaisons fiables entre différents groupes et populations(Brazier et al., 2014).
Malgré leurs avantages, les questionnaires auto-administrés présentent également certaines limitations :

1. Biais de Désirabilité Sociale : Les individus peuvent répondre de manière à paraître plus favorables ou à éviter la stigmatisation, ce qui peut fausser les résultats.(Huang et al., 1998)
2. Capacité d'Introspection : Tous les individus ne sont pas également capables de décrire avec précision leurs expériences internes. Certaines personnes peuvent sous-estimer ou surestimer leurs symptômes.(Huang et al., 1998)
3. Manque de Contexte Clinique : Les questionnaires ne prennent pas toujours en compte le contexte clinique complet de l'individu, ce qui peut limiter la profondeur de l'évaluation.(Wu et al., 1997)
4. Problèmes de Compréhension : Les variations dans la compréhension des questions peuvent entraîner des réponses inconsistantes, surtout si les questionnaires ne sont pas adaptés culturellement ou linguistiquement.(Logan et al., 2008)

5. Sensibilité au Moment de Remplissage : Les réponses peuvent varier en fonction de l'état d'esprit du moment, introduisant une variabilité qui peut affecter la fiabilité des résultats.(Logan et al., 2008)
6. Problème pour assurer la prévention : Pour détecter le problème de santé mentale la personne doit d'elle-même choisir d'aller voir un praticien qui lui administrera le questionnaire.(Logan et al., 2008)

Les questionnaires auto-administrés sont des outils pour l'évaluation de la santé mentale. Ils offrent une méthode accessible, efficace et standardisée pour dépister et suivre les troubles mentaux. Cependant, il est crucial de les utiliser en complément d'autres méthodes d'évaluation, telles que les entretiens cliniques et les données biométriques, pour obtenir une vue d'ensemble complète et précise de la santé mentale des individus.

Mais on le voit ces questionnaires souffrent de limites importantes. Une autre approche que nous avons découverte consiste a utilisé les habitudes de vie comme prédicteurs a l'apparition d'un problème de santé mentale. Intéressons-nous à ces habitudes et aux moyens de les mesurer.

1.1.3) LES HABITUDES DE VIE COMME PRÉDICTEURS DES TROUBLES DE SANTÉ MENTALE

Plusieurs habitudes de vie sont indicatives de la présence d'un problème de santé mentale. De nombreuses études ont relevé l'importance du sommeil pour être en bonne santé mentale. Une personne souffrant d'insomnie a en fonction des études, 10 à 17 fois plus de risque de développer des symptômes cliniques de dépression ou de trouble d'anxiété(Scott et al., 2021). De plus l'absence d'un sommeil de qualité est également associée au syndrome post traumatique(Harvey et al., 2003) aux troubles de l'alimentation(Lauer & Krieg, 2004) ou encore aux troubles psychotique(Reeve et al., 2015). Les problèmes de trouble du sommeil offrent aussi l'avantage de pouvoir facilement être observés en temps réels à l'aide de montre connecté ou de smartphone(Mechanick & Zhao, 2020).

L'activité physique joue un rôle crucial dans la santé mentale. Une pratique régulière d'exercice est associée à une diminution significative des symptômes de dépression et d'anxiété (Rebar et al., 2015). L'exercice physique stimule la production de neurotransmetteurs tels que les

endorphines, souvent appelées les "hormones du bonheur", qui améliorent l'humeur et procurent une sensation de bien-être (Dishman et al., 2006). De plus, l'activité physique régulière favorise la neurogenèse dans l'hippocampe, une région du cerveau impliquée dans la régulation de l'humeur et la cognition (Erickson et al., 2011).Les appareils de fitness et les applications mobiles permettent de suivre l'activité physique en temps réel, facilitant ainsi la détection de périodes d'inactivité prolongée, potentiellement indicative de troubles de la santé mentale (Mechanick & Zhao, 2020).

L'alimentation est un autre pilier fondamental de la santé mentale. Des études montrent que les régimes alimentaires riches en fruits, légumes, grains entiers, et poissons sont associés à une réduction du risque de dépression et d'anxiété(Lai et al., 2014). À l'inverse, une consommation élevée d'aliments transformés et sucrés est liée à une augmentation des symptômes dépressifs (Jacka et al., 2010).Les nutriments tels que les oméga-3, les vitamines B, le magnésium et le zinc jouent un rôle crucial dans le fonctionnement cérébral et la régulation de l'humeur(Rao et al., 2008). Les applications de suivi nutritionnel et les journaux alimentaires électroniques permettent de surveiller la qualité de l'alimentation, fournissant ainsi des indicateurs précieux sur les habitudes alimentaires pouvant influencer la santé mentale(Boushey et al., 2017).

La vie sociale est un facteur déterminant pour la santé mentale. Des relations sociales solides et un soutien social adéquat sont associés à une meilleure santé mentale et à une réduction des risques de dépression et d'anxiété(Cohen, 2004). L'isolement social, en revanche, est un facteur de risque majeur pour le développement de troubles mentaux(Holt-Lunstad et al., 2010). Participer à des activités communautaires, maintenir des contacts réguliers avec des amis et la famille, et s'engager dans des activités sociales peuvent aider à prévenir et à atténuer les symptômes de troubles mentaux (Haslam et al., 2014)Les réseaux sociaux et les plateformes de communication en ligne offrent des moyens de suivre et de renforcer les interactions sociales, contribuant ainsi à une meilleure santé mentale(Iacus et al., 2015).

La consommation d'alcool et de tabac a des impacts significatifs sur la santé mentale. Une consommation excessive d'alcool est associée à une augmentation du risque de dépression, d'anxiété, et de troubles bipolaires(Boden & Fergusson, 2011). De même, le tabagisme est lié à une plus grande prévalence de troubles mentaux, incluant la dépression et l'anxiété(Fluharty et al., 2017). Ces substances affectent le système nerveux central et altèrent les neurotransmetteurs, contribuant ainsi à des déséquilibres émotionnels et cognitifs(Gilman &

Abraham, 2001). Les applications de suivi de la consommation d'alcool et de tabac permettent de surveiller les habitudes de consommation et de fournir des alertes en cas de dépassement des seuils recommandés, aidant ainsi à prévenir les impacts négatifs sur la santé mentale(Alessi & Petry, 2013).

L'utilisation des habitudes de vie comme prédicteurs de la santé mentale est déjà une réalité. Des études ont démontré le potentiel de l'analyse des données de capteurs pour prédire les humeurs, le stress et la santé mentale future(Yu & Sano, 2020). De plus, des recherches ont exploré l'utilisation de graphes semi-supervisés pour inférer la santé mentale à partir des données comportementales(Dong et al., 2021). Ces avancées soulignent l'importance croissante de cette approche dans le domaine de la santé mentale et ouvrent la voie à de nouvelles possibilités pour la prévention et le traitement des troubles mentaux.

Contrairement aux questionnaires traditionnels, qui reposent sur l'auto-évaluation et peuvent être sujets à des biais, l'utilisation des habitudes de vie comme prédicteurs offre une approche plus objective et quantifiable. Les données collectées par les capteurs intelligents sont moins susceptibles d'être influencées par la désirabilité sociale ou la subjectivité des réponses, offrant ainsi une vision plus précise de l'état mental d'un individu.

De plus, cette approche permet une surveillance continue et en temps réel de la santé mentale, contrairement aux questionnaires qui ne fournissent qu'un aperçu ponctuel. Les données comportementales peuvent révéler des changements subtils dans les habitudes de vie qui pourraient être des signes précurseurs de troubles mentaux, permettant ainsi une intervention précoce et préventive.

Enfin, l'utilisation des habitudes de vie comme prédicteurs ouvre la voie à des interventions personnalisées et ciblées. En identifiant les facteurs de risque spécifiques à chaque individu, il est possible de proposer des recommandations adaptées pour améliorer leur bien-être mental. Par exemple, si les données révèlent un manque de sommeil chronique, des conseils pour améliorer l'hygiène du sommeil peuvent être prodigués.

1.2) APPRENTISSAGE MACHINE

1.2.1) INTRODUCTION

Après avoir exploré en profondeur les multiples facettes de la santé mentale et les différentes manières de la mesurer, notamment à travers les questionnaires standardisés et l'analyse des habitudes de vie, nous allons maintenant nous tourner vers les outils qui permettront d'exploiter ces données de manière optimale. Nous nous sommes demandé quelle méthode permettrait de prédire un événement Y (problème de santé mentale) à l'aide d'un ensemble de facteurs X (habitudes de vies) et nous nous sommes tournés vers l'apprentissage machine.

L'apprentissage machine, offre des méthodes puissantes pour identifier des patterns et des relations complexes au sein de vastes ensembles de données. En appliquant ces techniques aux données relatives à la santé mentale, nous espérons découvrir des modèles prédictifs qui permettront non seulement d'identifier les individus à risque, mais aussi de mieux comprendre les facteurs sous-jacents qui influencent leur bien-être.

Dans la section suivante, nous plongerons dans les détails des méthodes d'apprentissage supervisé, en nous concentrant sur celles qui se sont révélées particulièrement prometteuses dans le domaine de la santé mentale : la régression logistique et les arbres de décision. Nous examinerons comment ces algorithmes peuvent être utilisés pour modéliser la relation entre des causes x et une conséquence y, et donc comment ils peuvent contribuer à la détection précoce des troubles mentaux.

Généralement, l'apprentissage machine est présenté comme une branche de la science informatique qui vise à apprendre à partir de données afin d'améliorer les performances dans différentes tâches (prédiction, classification, etc.). Les méthodes de « machine learning » sont souvent regroupées en trois grandes familles : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage profond.(Jiang et al., 2020)

L'apprentissage supervisé s'applique sur des données dites étiquetées. Cela signifie que la machine apprend à partir de données annotées et peut construire des modèles de régression ou de classification pour prédire ou classer la variable cible. Ainsi, équipé de nouvelles données sans la variable cible, le modèle sera capable de prédire ou de classer cette variable.(Jiang et al., 2020)

L’apprentissage non supervisé vise à trouver des relations entre les données sans qu'il y ait de variable cible étiquetée. Cette technique est fréquemment utilisée dans les problèmes de classification où les algorithmes révèlent des groupes basés sur les caractéristiques des variables indépendantes. Ces groupes émergent sans qu'ils soient explicitement définis dans la base de données.(Jiang et al., 2020)

L'apprentissage profond, ou Deep learning, est une sous-catégorie du machine learning qui utilise des réseaux de neurones artificiels à plusieurs couches. Cette technique s'inspire du fonctionnement du cerveau humain et permet de modéliser des relations complexes et non linéaires entre les données. L'apprentissage profond est particulièrement efficace pour traiter de grands volumes de données et excelle dans des domaines tels que la reconnaissance d'images, la traduction automatique, et la génération de texte. Grâce à sa capacité d'apprentissage automatique à partir de caractéristiques brutes, il est souvent utilisé pour découvrir des patterns subtils que des approches moins complexes ne pourraient pas détecter.(Jiang et al., 2020)

Dans le cadre de notre mémoire nous chercherons à expliquer/prédire des données étiquetés. Nous allons donc nous concentrer sur les méthodes d’apprentissages supervisé. Nous mettrons en place plusieurs méthodes dans le cadre de notre analyse dont 2 principales : « la régression logistique, l’arbre de décision ».

1.2.2) REGRESSION LOGISTIQUE

La régression logistique est une méthode statistique et d'apprentissage automatique couramment utilisée pour modéliser et prédire des résultats binaires. Contrairement à la régression linéaire, qui est adaptée aux variables continues, la régression logistique convient aux variables dépendantes catégorielles, souvent binaire, c'est-à-dire ayant deux catégories.(Hosmer, 2013)

La régression logistique modélise la probabilité qu'un certain événement se produise en fonction de plusieurs variables indépendantes. Au lieu de prédire directement le résultat comme dans la régression linéaire, elle prédit la probabilité qu'un événement survienne, cette probabilité étant ensuite utilisée pour classer les observations en deux catégories : par exemple, "présence" ou "absence" d'un trouble. La régression logistique fonctionne en utilisant une

fonction logistique pour transformer une combinaison linéaire des variables prédictives en une probabilité. Cette fonction logistique produit une courbe en forme de S qui peut prendre des valeurs entre 0 et 1, rendant ainsi les prédictions de probabilités appropriées pour les résultats binaires.(Marill, 2004a, 2004b)

Pour évaluer la performance d'un modèle de régression logistique, plusieurs métriques peuvent être utilisées :

- **Matrice de confusion** : Elle permet de visualiser les performances de classification en termes de vrais positifs, faux positifs, vrais négatifs et faux négatifs.
- **Courbe ROC (Receiver Operating Characteristic) et AUC (Area Under the Curve)** : La courbe ROC permet de visualiser la performance du modèle à différents seuils de classification, tandis que l'AUC fournit une mesure globale de cette performance.
- **Coefficient de détermination (pseudo-R^2)** : Bien qu'il n'y ait pas de véritable R^2 en régression logistique comme en régression linéaire, des versions alternatives comme le pseudo-R^2 de McFadden peuvent être utilisées pour évaluer l'ajustement du modèle. (Hosmer, 2013)

La régression logistique est largement utilisée dans plusieurs domaines :

- **Médecine** : Pour prédire la présence ou l'absence de maladies, par exemple, la probabilité qu'un patient ait une maladie cardiaque en fonction de divers facteurs de risque. (Harrell, 2015)
- **Marketing** : Pour modéliser la probabilité qu'un client achète un produit donné en fonction de ses caractéristiques démographiques et comportementales.

Dans notre cas ce qui nous intéresse le plus est en lien avec la santé.

Limites de la régression logistique

Bien que puissante, la régression logistique a ses limites. Elle suppose une relation log-linéaire entre les variables indépendantes et la log-odds de la variable dépendante. En outre, elle peut être sensible aux valeurs aberrantes et a multicolinéarité. Pour pallier certaines de ces limitations, des variantes comme la régression logistique multinomiale (pour des variables dépendantes à plus de deux catégories) et la régression logistique pénalisée (pour gérer la multicolinéarité et les nombreux prédicteurs) peuvent être utilisées. (Menard, 2002)

1.2.3) ARBRE DE DECISION

Les arbres de décision sont des outils puissants utilisés en apprentissage automatique pour la classification et la prédiction. Ils fonctionnent en segmentant un ensemble de données en sous-ensembles plus petits et plus homogènes selon certains critères, jusqu'à ce qu'une décision ou une prédiction puisse être faite pour chaque segment final, ou feuille de l'arbre (Breiman et al., 2017). Un arbre de décision commence par un nœud racine, qui pose une question ou un test sur une caractéristique spécifique des données. En fonction de la réponse, les données sont divisées en deux ou plusieurs branches, et ce processus se répète jusqu'à ce que chaque branche aboutisse à une feuille, qui représente une classe ou une valeur prédite.

Les arbres de décision présentent plusieurs avantages qui en font des outils largement utilisés en apprentissage automatique. L'un de leurs points forts est leur interprétabilité : la structure en arbre permet une visualisation claire du processus de décision, ce qui est précieux dans de nombreux domaines où la compréhension du raisonnement est aussi importante que la prédiction elle-même (Quinlan, 1986).

Un autre avantage majeur des arbres de décision est leur capacité à gérer efficacement des données mixtes, c'est-à-dire à la fois numériques et catégorielles, sans nécessiter de prétraitement complexe (Loh, 2011). Cette flexibilité les rend particulièrement adaptés à une large gamme de problèmes du monde réel.

Les arbres de décision sont également capables de gérer les données manquantes, ce qui est fréquent dans de nombreux ensembles de données réels. Plusieurs stratégies ont été développées pour traiter ce problème, comme l'utilisation de substituts ou la création de branches spécifiques pour les valeurs manquantes (Twala et al., 2008).

Cependant, les arbres de décision ne sont pas sans inconvénients. L'un des problèmes majeurs est leur tendance au surapprentissage, en particulier lorsque l'arbre devient trop profond et capture du bruit dans les données plutôt que des motifs significatifs (Hastie et al., 2009). Pour atténuer ce problème, diverses techniques ont été développées :

1. L'élagage : Cette technique consiste à réduire la taille de l'arbre après sa construction initiale, en supprimant les branches qui n'améliorent pas significativement la performance prédictive (Quinlan, 1986).

2. Les ensembles d'arbres : Des méthodes comme les forêts aléatoires (Breiman, 2001) ou le boosting (Freund & Schapire, 1997)combinent plusieurs arbres pour améliorer la précision et réduire le surapprentissage.
3. Les arbres de décision obliques : Ces variantes utilisent des combinaisons linéaires de caractéristiques pour les splits, ce qui peut améliorer la performance sur certains types de données(Murthy et al., 1994).

De plus, les arbres de décision peuvent être instables, c'est-à-dire que de petits changements dans les données d'entraînement peuvent entraîner des changements importants dans la structure de l'arbre. Cette instabilité peut être problématique pour l'interprétabilité et la généralisation.

Malgré ces défis, les arbres de décision restent un outil fondamental en apprentissage automatique, servant de base à des algorithmes plus avancés et continuant à être utilisés dans une grande variété d'applications. Leur simplicité d'utilisation et d'interprétation quand maintenu peu profond peu représenter de grand atout dans le domaine de la santé mentale.

Une question se pose alors comment maintenir les arbres de décision ou les régressions logistique peu complexes avec un nombre limité d'éléments afin de garder leur interprétabilité ? Une branche de recherche concernant ce domaine existe la sélection de caractéristique.

1.2.4) SELECTION DE CARACTÉRISTIQUES

Introduction

La sélection de caractéristiques est une composante essentielle du processus de modélisation en apprentissage automatique, particulièrement dans des domaines avec des jeux de données à haute dimensionnalité comme la santé mentale. Cette revue de littérature examine les différentes techniques de sélection de caractéristiques, leur importance, et leur application spécifique dans le domaine de la santé mentale. Les progrès récents en apprentissage automatique ont permis des avancées significatives dans la sélection de caractéristiques, rendant cette revue cruciale pour les chercheurs et les praticiens.

Importance de la Sélection de Caractéristiques

La sélection de caractéristiques réduit la dimensionnalité des données, minimisant ainsi le bruit et améliorant la performance des modèles. Elle est cruciale pour éviter le "fléau de la dimensionnalité", qui peut diluer l'importance explicative des variables et augmenter le risque

de surajustement(Guyon & Elisseeff, 2003). De plus, elle facilite l'interprétation des modèles, ce qui est particulièrement important dans des domaines sensibles comme la santé mentale.(Chandrashekar & Sahin, 2014)

1. Réduction de la Dimensionnalité :

La réduction de la dimensionnalité permet de simplifier les modèles, réduisant ainsi le temps de calcul et les ressources nécessaires. Par exemple, dans l'analyse de grandes bases de données génétiques, la sélection de caractéristiques aide à identifier les gènes les plus pertinents sans analyser chaque gène individuellement. (Saeys et al., 2007)

2. Amélioration de la Précision :

En éliminant les caractéristiques non pertinentes ou redondantes, la précision des modèles prédictifs est souvent améliorée. Cela est particulièrement utile dans des applications comme la détection de fraudes, où des variables spécifiques peuvent fortement influencer la performance du modèle.(Bolón-Canedo et al., 2012)

3. Facilitation de l'Interprétation :

Les modèles simplifiés sont plus faciles à interpréter et à expliquer, ce qui est crucial dans des domaines comme la médecine, où les décisions doivent être justifiables.(Kim, 2017)

Catégories de Techniques de Sélection de Caractéristiques

Les techniques de sélection de caractéristiques se classent en trois grandes catégories :

1. Sélection Supervisée :

Utilisée pour des ensembles de données étiquetées, elle vise à améliorer les performances des modèles de classification et de régression en identifiant les variables les plus significatives. Les techniques courantes incluent la sélection par l'importance des coefficients dans les modèles linéaires et les méthodes basées sur les arbres de décision.(Liu & Yu, 2005)

- Exemple : Dans une étude sur la prédiction des maladies cardiaques, les caractéristiques comme l'âge, la pression artérielle et les niveaux de cholestérol ont été identifiées comme étant les plus importantes à partir de modèles de régression logistique.(Detrano et al., 1989)

2. Sélection Non Supervisée :

Appliquée aux données non étiquetées, elle utilise des méthodes statistiques telles que l'analyse de l'entropie et les coefficients de corrélation pour évaluer l'importance des variables. Les techniques incluent l'analyse en composantes principales (ACP) et l'analyse factorielle.(He et al., 2005)

- Exemple : Dans l'analyse des données d'images, l'ACP est souvent utilisé pour réduire le nombre de pixels à un ensemble de caractéristiques principales représentant la variance maximale dans les données. (Jolliffe, 2002)

3. Sélection Semi-Supervisée :

Cette approche combine les données étiquetées et non étiquetées, offrant une flexibilité accrue et souvent une meilleure performance globale. Elle est particulièrement utile lorsque les étiquettes sont coûteuses à obtenir.(Zhou, 2012)

- Exemple : Dans le domaine de la reconnaissance de la parole, les techniques semi-supervisées ont aidé à améliorer les modèles en utilisant à la fois des enregistrements de voix étiquetés et non étiquetés pour former les algorithmes (Zhu, 2005).

Stratégies de Sélection de Caractéristiques

La sélection de caractéristiques peut être réalisée par différentes stratégies :

1. Méthodes de Filtrage :

Basées sur des critères statistiques indépendants du modèle, ces méthodes sont efficaces pour un tri initial des caractéristiques mais peuvent manquer de précision. Les méthodes de filtrage incluent l'analyse de la variance (ANOVA), le test du chi-carré, et les coefficients de corrélation.(Yu & Liu, 2004)

- Études de Cas : Dans une étude sur la classification des textes, la méthode du chi-carré a été utilisée pour sélectionner les mots les plus significatifs, améliorant ainsi la performance des modèles de classification des documents.(Forman, 2003)

2. Méthodes Enveloppantes :

Elles évaluent des sous-ensembles de caractéristiques en fonction de leur performance sur un modèle prédictif, offrant une précision accrue mais à un coût computationnel plus élevé. Les techniques courantes incluent la recherche exhaustive, les méthodes séquentielles, et l'algorithme génétique.(Kohavi & John, 1997)

- Exemple : Une étude sur la détection des cancers utilisant des méthodes enveloppantes a montré que l'utilisation de l'algorithme génétique pour sélectionner les caractéristiques pouvait améliorer significativement la précision des diagnostics (Hancer, Ozyer, & Alhajj, 2018).

3. Méthodes Intégrées :

Intégrant la sélection de caractéristiques dans le processus de formation du modèle, ces méthodes offrent une approche simultanée et efficace. Les méthodes intégrées incluent les forêts aléatoires, les machines à vecteurs de support (SVM) avec sélection de caractéristiques intégrée, et les réseaux de neurones.(Louppe, 2014)

- Exemple : Dans une étude sur la prédiction des ventes, l'utilisation de forêts aléatoires a permis non seulement de construire un modèle robuste mais aussi de sélectionner automatiquement les caractéristiques les plus importantes pour les prédictions.(Breiman, 2001)

Techniques Avancées

Des techniques comme l'élimination récursive de caractéristiques (RFE), les valeurs de Shapley, la sélection par lasso offrent des améliorations significatives :

- Élimination Récursive de Caractéristiques (RFE):

Identifie progressivement le sous-ensemble optimal en éliminant les caractéristiques les moins performantes. Cette technique est couramment utilisée avec les SVM et les modèles de régression.(Guyon et al., 2002)

- Exemple : Une application de RFE sur des données de génétique a permis de réduire le nombre de gènes analysés tout en maintenant une haute précision de classification des types de cancers.(Ramaswamy et al., 2002)

- Valeurs de Shapley :

Utilisées dans des approches comme SHAP, elles quantifient l'importance de chaque caractéristique pour expliquer les prédictions des modèles, offrant une sélection robuste et interprétable.(Lundberg & Lee, 2017)

- Exemple : Dans une étude sur les prêts bancaires, les valeurs de Shapley ont été utilisées pour identifier les variables les plus influentes dans la prédiction du risque de défaut de paiement, aidant ainsi les banques à mieux évaluer les demandes de prêt. (Owusu et al., 2023)

- Méthode LASSO

La méthode LASSO (Least Absolute Shrinkage and Selection Operator) est une technique de régularisation qui sélectionne les caractéristiques en imposant une contrainte de somme absolue sur les coefficients de régression, forçant certains à zéro, donc effectuant une sélection automatique des variables. (Tibshirani, 1996)

1. *Présentation* : LASSO réduit la variance du modèle en rétrécissant les coefficients, ce qui améliore la précision des prédictions en présence de nombreuses caractéristiques.(Tibshirani, 1996)
2. *Applications* : Utilisée pour la sélection de gènes dans des études génétiques et pour l'amélioration des modèles prédictifs dans divers domaines (Corporate Finance Institute, 2023).
3. Méthode mRMR

La méthode mRMR (Maximum Relevance Minimum Redundancy) sélectionne les caractéristiques qui ont une pertinence maximale avec la variable cible tout en minimisant la redondance entre elles.(Hanchuan et al., 2005)

- *Présentation* : Cette méthode utilise des critères d'information mutuelle pour évaluer et sélectionner les caractéristiques les plus pertinentes tout en minimisant la redondance entre elles.(Hanchuan et al., 2005)

- *Applications* : Utilisée dans diverses applications, notamment l'analyse d'images médicales et la bioinformatique pour identifier les caractéristiques les plus informatives.(Hanchuan et al., 2005)

Conclusion

La sélection de caractéristiques est une étape critique dans la construction de modèles prédictifs efficaces, surtout dans des domaines comme la santé mentale. Les différentes approches et techniques, allant des méthodes de filtrage aux valeurs de Shapley, offrent des solutions pour réduire la dimensionnalité tout en maintenant la pertinence et la performance des modèles. Les recherches futures devraient continuer à explorer de nouvelles méthodes et à affiner les techniques existantes pour répondre aux défis posés par des données de plus en plus complexes. En fin de compte, la sélection de caractéristiques joue un rôle clé dans l'amélioration de l'exactitude, de l'interprétabilité et de l'efficacité des modèles prédictifs, contribuant ainsi de manière significative à l'avancement de la recherche et de la pratique en apprentissage automatique.

1.2.5) PRÉCISION VS. EXPLICABILITÉ

Un défi majeur dans l'application de l'apprentissage supervisé en santé mentale est de trouver un équilibre entre la précision des modèles et leur explicabilité. Les modèles complexes, tels que les réseaux de neurones profonds, offrent souvent une grande précision prédictive mais sont difficiles à interpréter. Rish et al. (2020) ont démontré comment les modèles linéaires généralisés peuvent être utilisés pour une réduction dimensionnelle supervisée, souvent plus explicable. De plus, Ajith et al (2024) ont proposé une approche d’apprentissage profond pour la prédiction de la qualité de la santé mentale, où les compromis entre précision et explicabilité peuvent être analysés.(Ajith et al., 2023; Rish et al., 2008)

1.3) L'UTILISATION DE L'IA POUR PRÉVENIR LA SANTÉ MENTALE

Après avoir détaillé ces méthodes d'apprentissage machine nous nous sommes demandé à quel point celles-ci avaient déjà été utilisé pour prédire des troubles de santé mentale. Nous avons découvert quelques études utilisant L'apprentissage supervisé pour prédire les problèmes de santé mentale, avec deux méthodes particulièrement populaires : la régression logistique et les arbres de décision.

La régression logistique est appréciée pour sa capacité à évaluer l'impact de plusieurs variables explicatives sur la probabilité d'un trouble mental. Elle a été utilisée dans diverses études pour prédire la dépression, l'insomnie, et d'autres problèmes de santé mentale. Par exemple, Lin et al. (2021) ont utilisé cette méthode pour identifier les facteurs de risque de la dépression chez les adultes, tandis que Gandaputra et al. (2021) ont examiné l'association entre le jeu vidéo et l'insomnie chez les adolescents.

Les arbres de décision, quant à eux, offrent une approche visuelle et intuitive pour la prédiction. Ils ont été employés dans des études sur la dépression et l'insomnie (Scott et al., 2021), les troubles de l'alimentation chez les adolescents (Lauer & Krieg, 2004), et le syndrome de stress post-traumatique chez les vétérans (Harvey et al., 2003).

L'étude COMPASS, menée au Canada, illustre particulièrement bien l'utilisation de ces deux méthodes. Cette étude a collecté des données auprès de plus de 74 000 élèves de 136 écoles, incluant des mesures de l'anxiété, de la dépression, et du bien-être psychologique, ainsi que 23 variables sociodémographiques et comportementales.

Dans le cadre de COMPASS, la régression logistique a été utilisée pour analyser l'impact de diverses variables sur les niveaux d'anxiété et de dépression. Les résultats ont montré que des facteurs tels que le sentiment d'appartenance à l'école et le soutien familial étaient des prédicteurs significatifs de la santé mentale des élèves (Battista et al., 2023).

Parallèlement, les arbres de décision, notamment les modèles CART (Classification and Regression Trees) et CTREE (Conditional Inference Trees), ont été appliqués aux données COMPASS. Ces modèles ont permis de segmenter la population en sous-groupes homogènes, identifiant ainsi les facteurs de risque majeurs et les sous-groupes à haut risque. Les arbres de décision ont non seulement confirmé l'importance des facteurs identifiés par la régression

logistique, mais ont aussi révélé des interactions complexes entre les variables, offrant une perspective plus nuancée (Battista et al., 2023).

L'étude COMPASS a démontré que les arbres de décision présentaient plusieurs avantages par rapport aux méthodes de régression classiques. Ils ont accordé une plus grande importance relative aux principaux prédicteurs et ont montré une capacité accrue à cerner les sous-groupes à risque élevé, ce qui est crucial pour cibler les efforts de prévention et d'intervention en santé mentale.

Cette recherche nous a poussé à nous demander si l'on pouvait récolter les données en temps réels. (Yu & Sano, 2020) ont exploré l'utilisation des données passives des capteurs pour prédire les humeurs futures, la santé et le stress à l'aide de techniques de deep learning. Leur recherche indique que l'analyse des comportements quotidiens peut fournir des indices précieux sur l'état mental futur des individus. De plus, Jiang et al. (2020) ont fourni une introduction aux techniques d'apprentissage supervisé appliquées aux données de santé mentale, soulignant la capacité de ces techniques à identifier des modèles prédictifs robustes basés sur des données comportementales et cliniques.(Jiang et al., 2020; Yu & Sano, 2020)

En conclusion, tant la régression logistique que les arbres de décision ont prouvé leur efficacité dans la prédiction des problèmes de santé mentale. Chaque méthode offre des avantages uniques : la régression logistique permet une interprétation claire de l'impact de chaque variable, tandis que les arbres de décision offrent une visualisation intuitive des interactions complexes entre les facteurs de risque. L'étude COMPASS illustre comment ces deux approches peuvent être utilisées de manière complémentaire pour obtenir une compréhension plus complète des facteurs influençant la santé mentale des jeunes.

L'une des applications les plus prometteuses de l'apprentissage supervisé est l'identification des facteurs de risque les plus significatifs parmi les habitudes de vie des individus. Dong et al. (2021) ont utilisé des graphes semi-supervisés pour inférer la santé mentale, permettant d'isoler les variables les plus influentes. Rish et al. (2008) ont exploré la réduction dimensionnelle supervisée pour déterminer les variables significatives dans les données de santé mentale, offrant des insights précieux sur les comportements et habitudes ayant le plus d'impact sur la santé mentale.(Dong et al., 2021; Rish et al., 2008)

1.4) QUESTION DE RECHERCHE

La littérature existante, bien que riche en études sur la santé mentale, les habitudes de vie et l'apprentissage automatique, présente une lacune notable. En effet, peu de recherches ont exploré de manière approfondie et quantitative la relation entre ces trois domaines. La plupart des études se sont concentrées sur l'un ou l'autre de ces aspects, sans véritablement les intégrer dans une approche globale.

Certaines études ont examiné l'impact des habitudes de vie sur la santé mentale, mais souvent de manière descriptive ou en se limitant à des analyses statistiques simples. D'autres ont appliqué l'apprentissage automatique pour prédire des troubles mentaux, mais en se basant principalement sur des données cliniques ou sociodémographiques, sans accorder une attention suffisante aux habitudes de vie.

Il manque donc des recherches qui combine ces trois éléments de manière systématique, en utilisant des techniques d'apprentissage automatique avancées pour analyser des données détaillées sur les habitudes de vie et leur lien avec la santé mentale. Ce manque nous amène à poser la question suivante

Comment les habitudes de vie influencent-elles la prédiction des troubles de santé mentale ?

Cette question de recherche peut en réalité être divisée en 4 sous questions auxquelles nous allons essayer de répondre durant ce mémoire :

1. **Dans quelle mesure peut-on prédire les troubles de santé mentale, à l'aide des habitudes de vie ?**
2. **À quel point pouvons-nous identifier les différents troubles ?**
3. **Parmi ces habitudes de vie, quels sont les facteurs qui ont le plus d'impact ?**
4. **À quel point le modèle perd-il en précision si nous voulons un modèle plus explicable ?**

Explications détaillées des questions

La capacité de prédire les troubles de santé mentale est une question cruciale dans le domaine de la psychologie et de la psychiatrie. Pour explorer cette capacité, nous devons nous concentrer sur les habitudes de vie des individus. En collectant des données détaillées via des

questionnaires, nous pouvons analyser les comportements quotidiens et les routines qui pourraient être liés à divers troubles mentaux. En appliquant des algorithmes d'apprentissage supervisé, notre objectif est d'identifier les variables les plus significatives et de comprendre comment ces facteurs influencent la prédiction des troubles.

Il est aussi essentiel de déterminer dans quelle mesure ces outils peuvent différencier les différents types de troubles de santé mentale, ce qui pourrait améliorer la précision des diagnostics et offrir des perspectives pour des interventions plus ciblées. Si le modèle ne permet que de prédire la présence d'un trouble quel qu'il soit, plus tard un système de recommandation ne pourra pas apporter un soutien ciblé sur les troubles effectivement subis.

Une autre question importante est de savoir quels facteurs de ces habitudes de vie ont le plus d'impact sur la prédiction des troubles de santé mentale. En examinant les données recueillies, nous cherchons à isoler les comportements les plus influents, tels que le sommeil, l'alimentation, l'exercice physique, et d'autres aspects du mode de vie. Les techniques d'apprentissage machine nous permettront d'évaluer l'importance de chaque variable et de comprendre comment elles interagissent pour affecter la santé mentale. Cette analyse détaillée nous aidera à identifier les cibles potentielles pour des interventions préventives et à élaborer des recommandations personnalisées pour les individus à risque, renforçant ainsi la prévention et le traitement des troubles mentaux.

Enfin, il est également crucial de considérer le compromis entre la précision des modèles d'apprentissage machine et leur explicabilité. Dans le domaine de la santé mentale, les professionnels de santé et les décideurs doivent pouvoir comprendre et utiliser les résultats des analyses prédictives. Pour répondre à cette exigence, nous explorerons des techniques telles que SHAP (SHapley Additive exPlanations), qui offrent des explications transparentes des prédictions des modèles. Nous analyserons l'impact de l'intégration de ces techniques sur la précision des modèles et discuterons des meilleures pratiques pour communiquer ces explications de manière claire et accessible. L'objectif est de trouver un équilibre optimal entre précision et explicabilité, afin d'améliorer la confiance et l'adoption des modèles prédictifs dans le domaine de la santé mentale. Pour améliorer l'explicabilité nous pouvons également faire appel à d'autres modèles plus explicables mais qui seront surement moins précis.

1.5) RÉSUMÉ REVUE DE LITTÉRATURE ET METHODOLOGIE

Notre analyse de la revue de littérature a fait remonter plusieurs éléments concernant la santé mentale et les habitudes de vies. Premièrement nous avons vu que la définition de santé mentale allait au-delà du simple bien être psychologique et encapsule en plus des dimensions sociales et émotionnelle.

Deuxièmement, on a noté qu'il y avait différentes méthodes pour identifier un trouble de santé mentale. La méthode la plus répandue consiste en l'utilisation de questionnaire standardiser administrer par un professionnel de la santé. L'utilisation de données comportementale à l'aide de capteur intelligent (montre connecté, smartphone etc…), ou de questionnaires a également montré son efficacité.

Troisièmement, nous nous sommes penchés sur les algorithmes qui permettent de détecter un trouble de la santé mentale. L'étude a fait ressortir particulièrement 2 algorithmes l'arbre de décision et la régression logistique. De plus pour essayer d'expliquer les décisions de ces algorithmes la méthode SHAP est ressorti comme un choix approprié.

Quatrièmement, plusieurs méthodes de sélections de caractéristiques ont été abordées. On a vu les avantages de celles-ci, permettant de réduire la dimensionalité du problème et de faciliter l'interprétabilité des modèles. On a vu qu'elles étaient particulièrement utilisées dans l'exploitation de données biométriques ou dans l'analyse des réseaux sociaux pour prédire les troubles mentaux.

En somme, cette revue de littérature met en lumière la complexité de la santé mentale, dépassant le cadre strictement psychologique pour englober des dimensions sociales et émotionnelles. Elle souligne également la diversité des approches pour identifier les troubles mentaux, allant des questionnaires standardisés aux données comportementales issues de capteurs intelligents. Fort de ces enseignements, nous mettrons en place une méthodologie qui capitalise sur ces avancées, en combinant des approches variées pour une compréhension plus holistique de la santé mentale et une détection plus précise des troubles associés.

PARTIE 2) PARTIE MÉTHODOLOGIE

2.1) PRESENTATION DES BASES DE DONNÉE

Pour répondre à ces questions nous nous sommes appuyés sur deux bases de données ULA et ESSAIM.

Le projet ULA vise à dresser un portrait de la santé mentale, des capacités d'autogestion et connaissances en santé mentale, et de l'environnement de travail pour les employés dans les PME canadiennes, et à mieux comprendre les interrelations entre ces facteurs. Ce faisant, le projet permettra aussi de mettre à l'essai un questionnaire englobant mesurant l'ensemble de ces aspects liés à la santé mentale auprès de cette population de travailleur.se. s. le questionnaire a été administré a 2500 travailleur.se.s. de PME canadienne. Les questionnaires incluent des questions sur le dépistage (comme l'âge et les heures de travail), la santé mentale en général, l'autogestion de la santé mentale, le bien-être au travail, les facteurs psychosociaux environnementaux, les programmes de santé offerts aux travailleurs, et enfin, les performances au travail.

La plupart de ces questions suivent une échelle ordinale, demandant de s'auto-évaluer sur une échelle. Au totale la base de données compte 154 questions. Pour donner une meilleure idée de la structure des questions nous en reprenons une ici :

« Au cours des deux dernières semaines, à quelle fréquence avez-vous été dérangé(e) par les problèmes suivants ? »

	Jamais	Plusieurs jours	Plus de la moitié des jours	Presque tous les jours
Q17. Sentiment de nervosité, d'anxiété ou de tension	○	○	○	○
Q18. Incapable d'arrêter de vous inquiéter ou de contrôler vos inquiétudes	○	○	○	○
Q19. Peu d'intérêt ou de plaisir à faire les choses	○	○	○	○
Q20. Se sentir triste, déprimé.e ou désespéré.e	○	○	○	○

Tableau 1: Exemple de question de notre base de données

Nous utiliserons deux types de variables cibles au sein de ce questionnaire. Tout d'abord la présence déclaré d'un trouble mental chez un travailleur. C'est cette variable que nous essaierons de prédire à l'aide des autres questions. Et ensuite on fera la même chose sur les différents troubles présents en nombre suffisant dans la base de données.

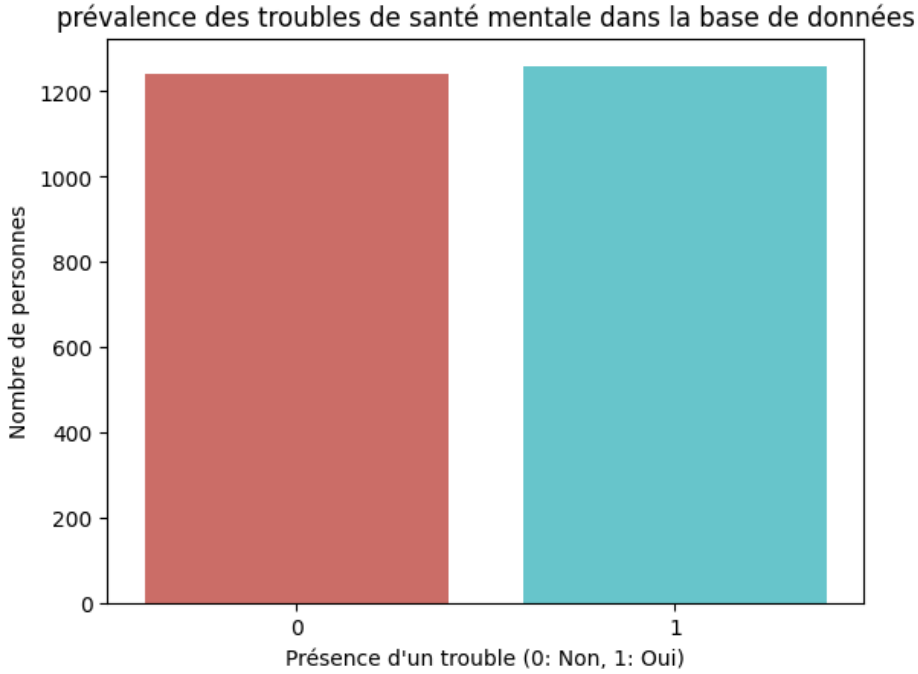


Figure 1: Présence d'un Trouble de la santé mentale

Sur ce graphique, on peut voir que les personnes avec et sans troubles se retrouve en nombre assez équilibré au sein de la base de données. Cependant beaucoup de troubles sont peu représenté dans notre base de données comme on peut le voir dans le tableau suivant :

Trouble	Nombre de cas	Proportion du trouble (%)	Proportion au sein de la base
Trouble anxieux	565	44.84	22.6
Épuisement professionnel	497	39.44	19.88
Dépression	384	30.48	15.36
Dépendance	195	15.48	7.8
TDH_A	181	14.37	7.24
Trouble alimentaire	135	10.71	5.4

Trouble post-traumatique	122	9.68	4.88
TOC	117	9.29	4.68
Jeu compulsif	56	4.44	2.24
Trouble bipolaire	38	3.02	1.52
Trouble de la personnalité	38	3.02	1.52
Spectre de l'autisme	34	2.70	1.36
Autre	33	2.62	1.32
Psychose	12	0.95	0.48
Schizophrénie	6	0.48	0.24

Tableau 2: Tableau récapitulatif des différents troubles dans la base de données ULA

On le voit parmi l'ensemble des données beaucoup de ces troubles sont très peu représentés. En effet quand les classes sont trop déséquilibrées et que certaines sont peu représentées, les modèles ont tendance à mal performer sur les classes minoritaires. (Figueroa et al., 2012)

C'est pourquoi nous avons décidé de prendre en compte uniquement dans notre recherche, les troubles représentant plus de 15% de notre base de données. Afin d'atténuer ce problème car lors de nos premières analyses nous avons constaté qu'en travaillant sur tous les troubles en dessous du seuil de 15% nous avions une précision et un rappel de 0% ou proche de celui-ci ce qui nous paraissait peu utile pour nos analyses tandis qu'au-dessus de ce seuil nous commencions à avoir des résultats analysables d'où notre choix. De plus, Si on veut que les modèles d'apprentissages puissent les prédire séparément il faudra utiliser des méthodes d'équilibrages de classes sinon nos modèles risquent de classifier tout le monde comme ne présentant pas de trouble tout en ayant une excellente précision. Par exemple, un modèle prédisant qu'une personne n'est pas dépressive aurait une exactitude de 84.64% mais un rappel et une précision de 0%. Pour approfondir, est disponible en annexe un tableau récapitulatif qui montre les prédictions des classes sous-représentées (voir annexe 4).

Le projet Essaim est de nature différente. Il collecte les données provenant des réponses à un questionnaire administré à 930 Canadiens, incluant des étudiants, des employés, des chômeurs,

etc. L'objectif de ce projet d'analyse de données est d'explorer les facteurs associés aux troubles de l'humeur et de l'anxiété, en se basant notamment sur les habitudes de vie. À cette fin, les données sont très multidimensionnelles, comprenant 727 variables mesurant différents aspects de la santé mentale ainsi que des habitudes de vie alimentaires, sociales et autres. Les principales sections de ce document sont divisées comme suit :

1. Général : Cette section présente des données socio-démographiques pour décrire les participants, incluant l'âge, l'occupation, le revenu, la situation familiale, etc.
2. Sommeil : Une série de questions relatives au sommeil, couvrant notamment la durée, la qualité et la satisfaction du sommeil.
3. Alimentation : Cette partie s'intéresse au régime alimentaire, examinant la consommation de produits laitiers, de viande, de végétaux, etc.
4. Activité physique : Cette section vise à comprendre la pratique sportive des participants, en particulier la nature de cette pratique (quel sport) et son intensité.
5. Santé psychologique : Cette section cherche à établir la santé psychologique des participants, en particulier leur niveau de stress et d'anxiété.
6. Bien-être : Cette section vise à établir le sentiment de bien-être des participants, notamment en ce qui concerne leur joie et leur tristesse.

On le voit la première base de données comporte de nombreuses données non liées aux habitudes de vies tandis que la seconde est plus concentrée sur celles-ci. Cela pourra nous permettre de comparer la précision des modèles dans les deux cas et de voir si les habitudes de vies apportent une grande plus-value par rapport aux méthodes classique d'utilisation d'échelle de stress et d'anxiété.

De plus, en ce qui concerne la base de données ULA, comme le montre le tableau récapitulatif du nombre de cas par trouble dans notre base de données et comme dit précédemment, il existe un déséquilibre significatif dans nos données. Nous allons donc utiliser dans ce travail des méthodes d'équilibrage de classes qui seront également détaillés ultérieurement (voir partie 2.3).

Maintenant que nous avons présenté ces bases de données, nous allons plutôt nous concentrer sur la manière dont nous allons essayer de répondre à nos différentes questions de recherches.

2.2) VUE D'ENSEMBLE DU PLAN EXPÉRIMENTAL

Compte tenu de nos deux bases de données et de nos différentes questions de recherche, nous avons opté pour une approche méthodologique permettant de répondre au mieux à nos interrogations. Notre démarche s'est articulée autour de nos sous-questions :

1.Dans quelle mesure peut-on prédire les troubles de santé mentale, à l'aide des habitudes de vie ?

Pour répondre à cette question nous avons besoins d'habitudes de vies et de troubles de la santé mentale. Nous allons utiliser les modèles d'apprentissage machine pour prédire les troubles à l'aide des habitudes de vies. Nous estimerons la précision de ces modèles prédictifs sur différents indicateurs pour voir leur performance à prédire un trouble de santé mentale à l'aide des habitudes de vies. Cela nous permettra de répondre à notre question.

2. À quel point pouvons-nous identifier les différents troubles ?

La capacité de nos modèles à distinguer entre différents troubles de santé mentale sera évaluée en les entraînant spécifiquement sur chaque trouble identifié dans nos bases de données (anxiété, dépression, insomnie). Les performances de chaque modèle seront mesurées à l'aide des mêmes indicateurs que ceux utilisés pour la prédiction générale de la présence d'un trouble (accuracy, précision, rappel, etc.).

Une comparaison des performances entre les différents troubles permettra d'identifier ceux qui sont les plus facilement prédits par les habitudes de vie. De plus, étant donné le déséquilibre potentiel dans la représentation de certains troubles, nous utiliserons des techniques d'équilibrage des classes telles que SMOTE et la pondération pour affiner nos modèles et améliorer leur capacité à détecter les troubles moins fréquents.

3.Parmi ces habitudes de vie, quels sont les facteurs qui ont le plus d'impact ?

Pour déterminer les habitudes de vie les plus influentes pour prédire un trouble de santé mentale, nous utiliserons des techniques d'interprétation des modèles, notamment les valeurs de Shapley (SHAP). Ces valeurs quantifient la contribution de chaque caractéristique à la prédiction du modèle, permettant ainsi d'identifier les habitudes de vie ayant le plus d'impact.

L’analyse des graphique SHAP nous permettra de visualiser l’importance relative de chaque habitude de vie, globalement et pour chaque trouble spécifique. Cette analyse comparative mettra en évidence les facteurs les plus déterminants dans la prédiction des troubles mentaux, offrant ainsi des pistes précieuses pour la prévention et les interventions ciblés.

4.À quel point le modèle perd-il en précision si nous voulons un modèle plus explicable ?

Le compromis entre précision et explicabilité sera évalué en comparant les performances de nos modèles explicables (arbre de décision et régression logistique) avec celles de modèles plus complexes, tels que les réseaux de neurones. La perte de précision sera mesurée à l’aide des mêmes indicateurs que vus précédemment

Cette analyse permettra de quantifier le coût, en termes de précision, de l’utilisation de modèles plus transparents et interprétables. Nous parlerons des implications de ce compromis pour l’application pratique des modèles dans le domaine de la santé mentale, en soulignant l’importance de trouver un équilibre entre performance et explicabilité pour favoriser leur adoption efficace par les professionnels de santé.

En ce qui concerne la base de données ESSAIM, nous avons identifié trois variables cibles et en avons créé une quatrième. Les trois variables cibles présentes dans la base de données concernent la présence d'un trouble dépressif, d'un trouble anxieux et d'un trouble du sommeil (insomnie). Ces données, issues d'un travail de l'Université Laval, établissent la présence de ces troubles à l'aide de questionnaires validés dans notre revue de littérature (PHQ-2 et PHQ-9 pour la dépression, GAD-2 pour l'anxiété). Nous avons également créé une quatrième variable cible non explicitement présente dans la base de données : la présence d'un trouble, quel qu'il soit. Cette variable est égale à 1 si un trouble est présent, 0 sinon. Elle sera utile pour déterminer si nous pouvons prédire la présence d'un trouble en général ou si nous sommes en mesure de différencier les différentes pathologies.

Pour notre base de données ULA, la méthode employée a été un peu différente étant donné que nos variables cibles ne sont pas exactement les mêmes. En effet si l’on regarde le tableau reprenant les troubles présents dans notre base de données (annexe …) nous avons considéré qu’un trouble sera pris en tant que variable cible uniquement si plus de 15% du trouble est recensé dans nos données. De plus, cette base de données étant fort axée sur le milieu du travail, celle-ci était moins concentrée sur les habitudes de vie. Cependant nous tenions quand même a

l’intégré pour deux raisons. Premièrement cela nous permet de voir a quel point les modèles se détériorent en l’absence de certaines habitudes de vie. Deuxièmement cette base de données aura des applications pour nos conseils managériaux.

Une fois ces cibles établies, nous avons retravaillé notre base de données afin qu'elle ne contienne plus que des habitudes de vie pour essaim. Pour ULA en plus des habitudes de vie des échelles de stress perçu ou de pratique d’entreprise ont été gardé. Nous nous sommes concentrés sur les questions permettant une mesure à l'aide d'appareils connectés (montre, smartphone, capteur cardiaque, etc.) afin de pouvoir suivre l'évolution de ces habitudes de vie en temps réel.

Enfin, nous avons entraîné nos modèles prédictifs (arbre de décision et régression logistique) avec et sans méthodes de sélection de caractéristiques. Nous avons également fait appel à la méthode de Shapley pour expliquer ces résultats.

2.3) EQUILIBRAGE DES DONNÉES.

Lorsque deux classes sont présentes de manière très déséquilibrée dans une base de données, il est parfois nécessaire d'équilibrer ces deux classes. En effet, si l'on ne rééquilibre pas, le modèle pourrait constamment prédire l'appartenance à la classe dominante, ce qui peut fausser la précision apparente du modèle. Il arrive que, paradoxalement, un modèle ne prédisant que l'appartenance à la classe dominante soit plus précis qu'un modèle cherchant à s'adapter aux données. Plusieurs techniques existent pour pallier ce problème de déséquilibre. (Nanni et al., 2015)

Le sur-échantillonnage de la classe minoritaire consiste à créer de nouvelles instances de cette classe afin d'obtenir un équilibre. Plusieurs méthodes existent, qui sont souvent des variations de la méthode SMOTE (Synthetic Minority Oversampling Technique). Cette méthode sera décrite en détail plus loin dans le texte. (Chawla et al., 2002)

On peut également utiliser une méthode d'apprentissage sensible aux coûts. Dans ce cas, un coût plus élevé est attribué à une mauvaise classification d'un membre de la classe minoritaire par rapport à celle de la classe majoritaire. Cela incite le modèle d'apprentissage à prendre davantage en compte les données sous-représentées. (Nanni et al., 2015)

Les deux méthodes citées ci-dessus seront celles utilisées dans le cadre de notre mémoire. Il existe cependant d'autres méthodes populaires, telles que les méthodes d'ensemble, incluant le « boosting » et le « bagging ». Ces méthodes consistent à utiliser la combinaison de plusieurs algorithmes d'apprentissage automatique pour obtenir de meilleurs résultats. (Galar et al., 2013; Li et al., 2008)

2.3.1) SMOTE

Comme mentionné précédemment, la méthode SMOTE consiste à créer de nouvelles instances de la classe minoritaire pour compenser le déséquilibre dans les données. Le principe de base de SMOTE n'est pas de générer ces instances de manière aléatoire, mais plutôt d'effectuer une interpolation entre des données voisines de la classe minoritaire. En introduisant ces exemples dans le voisinage des données déjà existantes, la méthode permet d'améliorer la performance de généralisation du modèle de classification..(Fernandez et al., 2018)

Il existe différentes implémentations de SMOTE mais la base du procédé est la suivante :

1. Choix d'un point : Sélectionnez un échantillon (x_i) de la classe minoritaire.
2. Trouver les voisins : Identifiez les k plus proches voisins de (x_i) qui appartiennent également à la classe minoritaire.
3. Générer un exemple synthétique : Choisissez aléatoirement un voisin (x_j) parmi les k voisins. Créez un nouvel exemple synthétique (x_{new}) par interpolation linéaire entre (x_i) et (x_j) :

$$x_{\text{new}} = x_i + \lambda \cdot (x_j - x_i)$$

Équation 1 : implémentation du Smote

La procédure mathématique ici décrite est considérée comme la base de SMOTE. Il existe cependant différentes implémentations qui font varier certaines propriétés essentielles de l'algorithme. Ces propriétés incluent : le choix des données à sur-échantillonner, l'intégration avec le sous-échantillonnage, la manière d'interpoler les nouvelles données, l'intégration avec les méthodes de changement de dimensionalité. (Chawla et al., 2002)

Nous ne développerons pas chacune de ces différentes méthodes, car nous n'utiliserons qu'une seule dans le cadre de ce mémoire. Nous nous contenterons d'une implémentation basique de la méthode SMOTE et ne la combinerons avec aucune autre méthode. Concernant le choix des données de la classe minoritaire, ainsi que pour la génération des nouvelles données, nous nous concentrerons sur le choix et la création de données à la frontière entre les deux classes.

2.3.2) PONDÉRATION

L'équilibrage des données en utilisant une pondération différente pour les erreurs de classification des membres de la classe sous-représentée est une technique efficace pour gérer les déséquilibres de classes dans les ensembles de données. Cette méthode attribue un poids plus élevé aux erreurs de classification des classes minoritaires, afin de compenser leur sous-représentation. (Sun et al., 2007)

En pratique, lors de l'entraînement d'un modèle, une pénalité plus sévère est appliquée pour les erreurs commises sur les instances des classes sous-représentées. Cela encourage le modèle à accorder plus d'attention à ces classes, améliorant ainsi la capacité de prédiction pour les classes

minoritaires. Cette approche est particulièrement utile dans les contextes où il est crucial de détecter les cas rares, tels que les problèmes de santé mentale ou la détection de fraudes. (Sun et al., 2007)

En résumé, l'équilibrage des données par pondération des classes est une stratégie efficace pour améliorer la performance des modèles sur des ensembles de données déséquilibrés, en réduisant le biais envers les classes majoritaires et en augmentant le rappel des classes minoritaires. (Sun et al., 2007)

2.4) EXPLICATION DES MODÈLES

2.4.1) MODÈLES UTILISÉS POUR PRÉDIRE

Commençons par expliquer les deux modèles de prédiction au cœur de notre travail l'arbre de décision et la régression logistique.

ARBRE DE DÉCISION

L'algorithme CART (Classification and Regression Trees) est un algorithme de décision binaire qui produit des arbres de décision pour les tâches de classification et de régression. Cet algorithme a été développé par Breiman et al. en 1984. Pour les tâches de classification, l'algorithme CART utilise la mesure d'impureté de Gini ou l'entropie pour évaluer la qualité des partitions.L'impureté de Gini pour un ensemble d'éléments ayant K classes possibles est définie comme :

$$G = \sum_{i=1}^{K} p_i(1 - p_i))$$

Équation 2 : Impureté de Gini

où p_i est la probabilité qu'un élément appartienne à la classe i. L'impureté de Gini mesure la probabilité qu'un élément choisi au hasard soit mal classé si il était étiqueté aléatoirement selon la distribution des classes dans l'ensemble. L'algorithme CART construit l'arbre de décision de manière récursive en utilisant une approche gloutonne pour diviser les données à chaque nœud(Breiman et al., 2017) :

1. **Sélection de la meilleure partition :** À chaque nœud, l'algorithme évalue toutes les partitions possibles des données en utilisant le critère choisi (impureté de Gini, entropie, ou MSE) et sélectionne la partition qui minimise ce critère.
2. **Division des données :** Les données sont divisées en deux sous-ensembles en fonction de la partition sélectionnée.
3. **Répétition :** Les étapes 1 et 2 sont répétées de manière récursive pour chaque sous-ensemble jusqu'à ce qu'un critère d'arrêt soit atteint (par exemple, une profondeur maximale de l'arbre ou un nombre minimum d'échantillons par feuille).

RÉGRESSION LOGISTIQUE

La régression logistique est une technique statistique utilisée pour modéliser la probabilité d'une variable binaire en fonction d'une ou plusieurs variables prédictives. Elle est couramment utilisée pour des problèmes de classification binaire, comme prédire si un courriel est un spam ou non, ou si un patient a une maladie donnée. (Team, 2024)

Dans la régression logistique, l'objectif est de trouver la relation entre les caractéristiques (variables indépendantes) et la probabilité que la variable dépendante soit égale à 1. Cette relation est modélisée en utilisant la fonction logistique, qui est une fonction sigmoïde. La fonction sigmoïde transforme n'importe quel nombre réel en une valeur entre 0 et 1, ce qui la rend appropriée pour modéliser des probabilités. (Team, 2024)
La formule générale de la régression logistique est la suivante :

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-((\beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn))}}$$

Équation 3: Formule générale de la régression logistique

- P(Y=1|X) est la probabilité conditionnelle que la variable dépendante Y soit égale à 1, étant donné les variables indépendantes X.
- β0, β1, β2, ..., βn sont les coefficients du modèle, qui sont estimés pendant le processus d'apprentissage du modèle.
- X1, X2, ..., Xn sont les variables prédictives.
- e est la base du logarithme naturel.

En résumé, la régression logistique utilise la fonction logistique pour lier la combinaison linéaire des caractéristiques d'entrée à la probabilité d'une issue binaire.

MODÈLES PLUS COMPLEXES

Nous développerons moins la théorie de ces modèles étant donné qu'ils sont moins centraux dans notre travail Voici une brève description de ces modèles supplémentaires :

- **Forêt aléatoire (Random Forest)** : Ce modèle combine plusieurs arbres de décision pour améliorer la précision et réduire le surapprentissage. Chaque arbre est construit sur un échantillon aléatoire des données et utilise un sous-ensemble aléatoire de caractéristiques, ce qui permet de diversifier les décisions et d'obtenir une prédiction plus robuste.(Jaiswal & Samikannu, 2017)

- **XGBoost et LightGBM** : Ces modèles sont des implémentations avancées de l'algorithme de boosting de gradient, qui combine plusieurs arbres de décision faibles pour créer un modèle fort. Ils sont connus pour leur grande efficacité et leur précision, notamment dans les compétitions de machine learning.(Kohavi & John, 1997)
- **SVM (Support Vector Machine)** : Ce modèle cherche à trouver un hyperplan qui sépare au mieux les différentes classes de données. Il est particulièrement efficace pour les problèmes de classification avec des données non linéaires, grâce à l'utilisation de fonctions noyaux.(Burges, 1998)
- **Réseaux de neurones** : Inspirés du fonctionnement du cerveau humain, les réseaux de neurones sont des modèles composés de nœuds interconnectés (neurones) organisés en couches. Ils excellent dans la modélisation de relations complexes et non linéaires au sein des données. Les réseaux de neurones profonds (Deep Neural Networks), avec de multiples couches cachées, sont particulièrement performants pour des tâches telles que la reconnaissance d'images, le traitement du langage naturel et la prédiction de séries temporelles.(Ajith et al., 2023)

ÉVALUATION DES MODÈLES

3 instruments statistiques seront utilisés majoritairement pour évaluer les différents modèles. Tout d'abord « l'accuracy » qui mesure simplement le taux de bonne réponse du modèle. Ensuite deux instruments plus complexe la précision et le rappel :

a) $$\text{Précision} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}} = \frac{TP}{TP+FP}$$

Équation 4 : Précision

b) $$\text{Rappel} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}} = \frac{TP}{TP+FN}$$

Équation 5 : Rappel

La précision mesure la part des personnes identifiées comme ayant un trouble et en ayant réellement un, tandis que le rappel mesure la part des gens ayant un trouble identifié comme tels. Ces deux valeurs sont liées aux erreurs de types 1 et 2. Les erreurs de type 1 consistent à juger quelqu'un malade alors qu'il ne l'est pas, et les erreurs de type 2 consistent à ne pas juger quelqu'un comme malade alors qu'il l'est.

2.4.2) MÉTHODES UTILISÉES POUR ÉVALUER L'IMPORTANCE DES CARACTÉRISTIQUES

Pour cette partie, nous allons présenter les trois méthodes que nous avons utilisées pour évaluer l'importance des caractéristiques sur nos différents modèles de prédiction.

VALEUR SHAPLEY

Les Valeurs de Shapley proviennent de la théorie des jeux coopératifs et sont utilisés pour attribuer de manière équitable la contribution de chaque joueur au résultat global. Dans le cadre des modèles d'apprentissage supervisés, les valeurs de Shapley servent à expliquer la contribution de chaque caractéristique à la prédiction d'un modèle.(Nicolas Peltre, 2021)

Pour calculer ces valeurs de Shapley, il faut procéder de la manière suivante :

1) Considérer toutes les combinaisons possibles de caractéristiques

Ce qui veut dire que pour chaque caractéristique toutes les combinaisons des autres caractéristiques sont prises en compte

2) Calculer la valeur marginale

La valeur marginale de l'ajout d'une caractéristique i à une combinaison S de caractéristiques est calculée comme suit :

$$\text{Valeur marginale} = v(S \cup \{i\}) - v(S)$$

Équation 6 : Valeur marginale

Où v est la fonction de la valeur, S est un sous ensemble de caractéristiques et $S \cup \{i\}$ est le sous-ensemble S avec l'ajout de la caractéristique i

3) Moyenne sur toutes les combinaisons

La valeur de Shapley pour une caractéristique i est la moyenne de toutes ses valeurs marginales sur toutes les combinaisons possibles des autres caractéristiques :

$$\phi_i(c) = \sum_{Z \subseteq N \setminus \{i\}} \frac{|Z|!(n - |Z| - 1)!}{n!} \times [c(Z \cup \{i\}) - c(Z)]$$

Équation 7 : Équation valeur de Shapley

Où N est l'ensemble de toutes les caractéristiques, et $\frac{|Z|!(n-|Z|-1)!}{n!}$ est un facteur de pondération basé sur le nombre de caractéristiques dans le sous-ensemble Z.(Nicolas Peltre, 2021)

Au plus les valeurs de Shapley sont élevées au plus la caractéristique est impactante pour le modèle.

LASSO

En ce qui concerne les méthodes intégrées nous avons dans ce mémoire prévu d'utiliser la régression lasso.

Le lasso, initialement développé pour les modèles de régression linéaire, est devenu une méthode courante pour la sélection de modèles et l'estimation de réduction. Dans un cadre de régression linéaire standard, avec une réponse continue Y ∈ ℝn, une matrice de conception n×p X et un vecteur de paramètres β ∈ ℝp, l'estimateur lasso est défini comme suit :

$$\widehat{\beta}_{\lambda} = \arg\min_{\beta} \left(\left\| \mathbf{Y} - X\beta \right\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

Équation 8 : Équation de Lasso

Pour des valeurs élevées de λ, certaines composantes de β sont réduites à zéro. Le lasso peut également être appliqué à d'autres modèles, comme la régression de Cox, la régression logistique ou la régression logistique multinomiale, en remplaçant la somme des carrés des résidus par la fonction de vraisemblance négative correspondante (Meier et al., 2008)

La régression LASSO (Least Absolute Shrinkage and Selection Operator) utilise la méthode des moindres carrés avec une contrainte ℓ1. Cela est réalisé en ajoutant une pénalité sous forme de la valeur absolue du coefficient de régression, multipliée par lambda, à la forme de Lagrange pour obtenir un estimateur de paramètre. La LASSO fait rétrécir les coefficients de régression vers zéro. Elle peut sélectionner les variables indépendantes qui expliquent la variable réponse, similaire à une régression pas à pas. (Rusyana et al., 2021)

RECURSIVE FEATURE ELIMINATION (RFE)

La méthode RFE fonctionne de manière itérative pour éliminer les caractéristiques moins importantes. Voici les étapes mathématiques :(Jeon & Oh, 2020)

1. **Entraînement initial :** Entraîner le modèle de régression logistique avec toutes les caractéristiques disponibles pour obtenir les poids **w**.

2. **Evaluation des importances des caractéristiques :**

   (a) Après avoir entrainé le modèle, nous obtenons le vecteur des poids = $[w_1,w_2,...,w_n]$ ou n est le nombre de caractéristiques.

   (b) La contribution d'une caractéristique x_j a la prédiction est représentée par le coefficient w_j. L'importance de cette caractéristique peut être mesurée par la magnitude du coefficient w_j, car un coefficient plus élevé (en valeur absolue) indique une plus grande influence sur la prédiction.

   (c) Ainsi, l'importance de la j-ème caractéristique est donnée par :

$$I_j = |w_j|$$

Équation 9 : Définition de l'importance

   (d) Calculer cette importance pour chaque caractéristique donne un vecteur d'importances :

$$\mathbf{I} = [|w_1|,|w_2|,...,|w_n|]$$

Équation 10: Vecteur d'importances

   (e) Classer ces importances de manière croissante pour identifier les caractéristiques les moins importantes.

3. **Elimination des caractéristiques :** Identifier la caractéristique avec l'importance la plus faible et l'éliminer. Supposons que k soit l'indice de cette caractéristique. La nouvelle matrice de caractéristiques $\mathbf{X}'$ est alors :

$$\mathbf{X}' = \mathbf{X} \setminus \mathbf{X}_k$$

Équation 11: Matrice de caractéristiques

4. **Itération :** Répéter les étapes 1 `a 3 jusqu'à ce que le nombre souhaité de caractéristiques soit atteint.(Jeon & Oh, 2020)

Une illustration graphique de ce procédé ressemblerait à ça

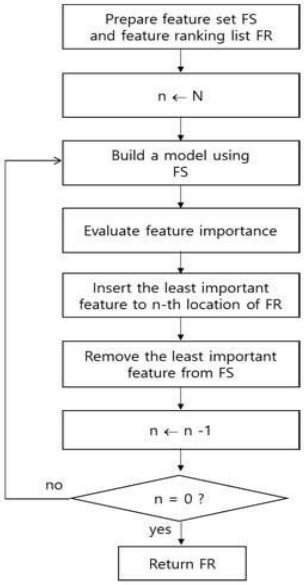


Figure 2: Le processus de l'élimination récursive des caractéristiques (RFE)

PARTIE 3) RÉSULTATS ET INTERPRÉTATIONS

3.1) PRÉDICTION DES TROUBLES DE SANTÉ MENTALE À PARTIR DES HABITUDES DE VIE

3.1.1) QUALITÉ DES MODELES DE BASE

Pour commencer, observons à quel point nous parvenons à prédire la présence d'un trouble mental dans nos deux bases de données. Commençons par utiliser l'ensemble des données dont nous disposons avant de nous concentrer uniquement sur les habitudes de vie. Les résultats sont résumés dans le tableau suivant :

Base de données	Variable cible	Accuracy	Méthodes
ULA	Présence d'un trouble	0.74	Arbre de décision
ULA	Présence d'un trouble	0.76	Régression logistique
ESSAIM	Présence d'un trouble	0.83	Arbre de décision
ESSAIM	Présence d'un trouble	0.84	Régression logistique

Tableau 3: tableaux des résultats de nos analyses

Nous pouvons voir que les modèles arrivent à prédire assez efficacement la présence d'un trouble dans les deux bases de données. La base de données ESSAIM offre une plus grande précision dans les deux cas, alors que la présence d'un trouble est plus rare dans cette base. Cela démontre que les différentes variables explicatives permettent mieux de cerner la présence d'un trouble dans ce cas-là.

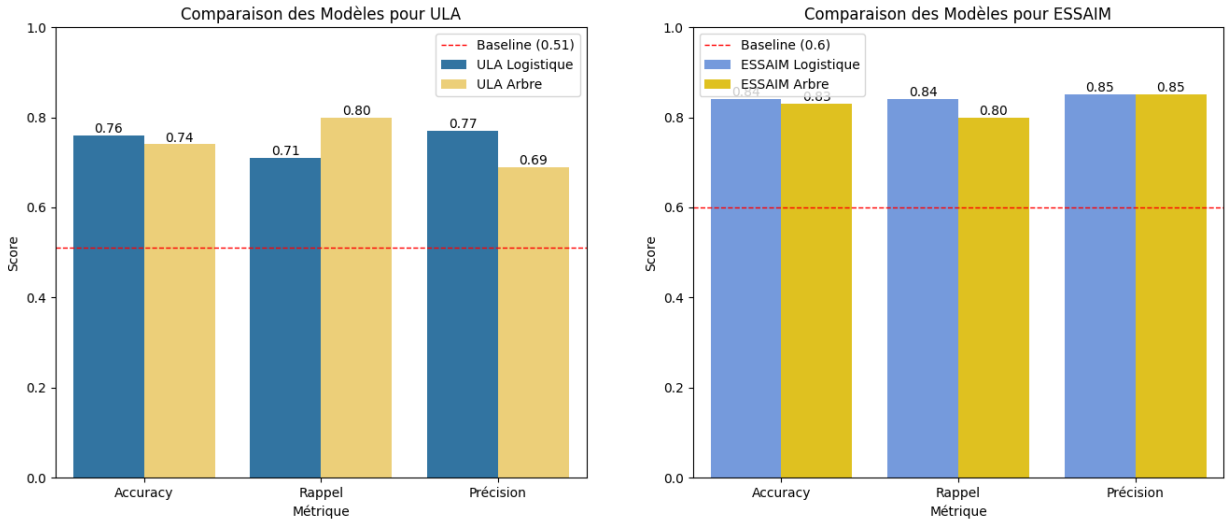


Figure 3: Comparaison des métriques en fonction du modèle utilisé

Comme le montre le graphique, les deux modèles sont proches d'être équivalents. Les différences d'accuracy sont minimes pour ESSAIM et pour ULA. De plus, quand un modèle semble avoir un avantage en rappel, l'autre obtient une meilleure précision, et inversement. Différencier ces modèles à l'aide de ces métriques est donc peu pertinent.

Cependant, ces valeurs sont intéressantes concernant les deux bases de données. Comme vu précédemment, ESSAIM a été épuré des questions ne portant pas sur les habitudes de vie. Or, cette base de données permet de prédire la présence d'un trouble avec une accuracy remarquable. En effet nous sommes entre 20 et 30 points de pourcentage au-dessus d'un modèle qui prédirai la présence d'un trouble en permanence, donc sur base de la classe dominante (Baseline 51% accuracy ULA 60% ESSAIM). On obtient un score de plus de 80 % pour l'accuracy et un rappel proche des 85 %, quel que soit le modèle. Cela signifie que 85 % des personnes possédant un trouble sont identifiées correctement uniquement grâce à leurs habitudes de vie, ce qui est très encourageant pour notre recherche.

3.1.2) MÉTHODES DE SÉLECTION DE CARACTÉRISTIQUES

Si l'on se penche sur les méthodes de sélection de caractéristique pour voir leur impact sur notre modèle on obtient le graphique suivant pour les données ESSAIM, avec l'utilisation de la régression logistique.

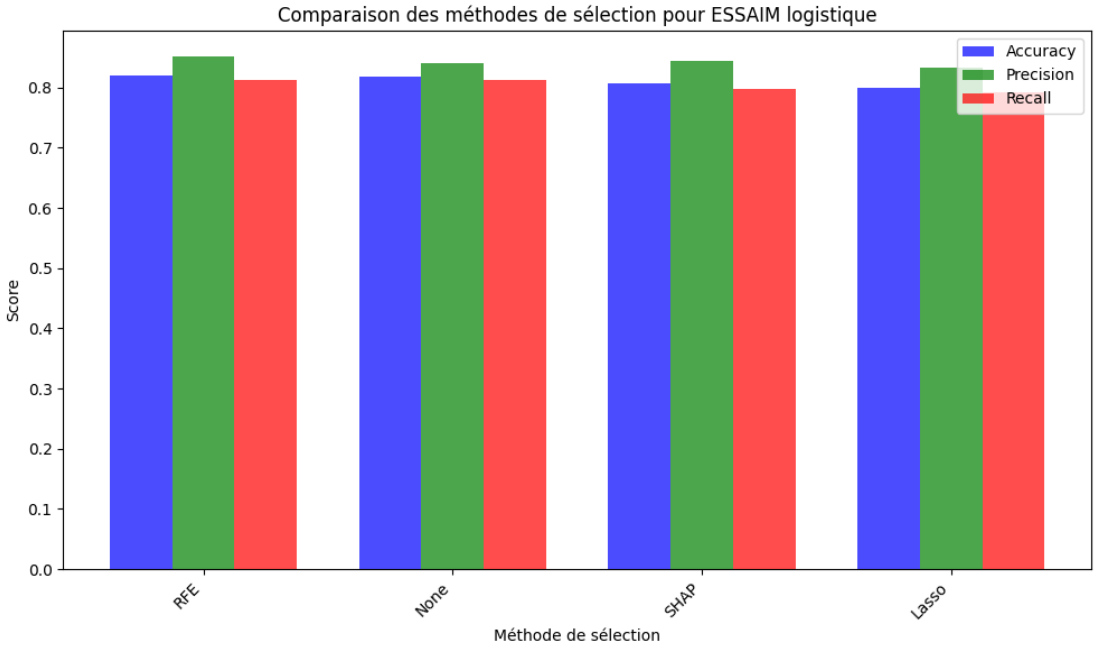


Figure 4: Score en fonction de la sélection de caractéristiques pour la régression logistique

On observe que, quelle que soit la méthode de sélection de caractéristiques utilisée, l'impact sur les performances du modèle reste très marginal. Les différences se jouent à chaque fois à quelques décimales de pourcentage près. Cette constatation soulève plusieurs points de réflexion :

1. Dimensionnalité des données : Nos données sont peut-être trop peu dimensionnées pour que la sélection de caractéristiques apporte une réelle plus-value à nos analyses. Dans des ensembles de données plus volumineux ou plus complexes, l'impact pourrait être plus significatif.
2. Qualité initiale des variables : Il est possible que les 107 variables observées soient déjà pertinentes pour le problème étudié, ce qui expliquerait le faible gain de performance après sélection.
3. Nature du problème : Certains problèmes de classification ou de régression peuvent être moins sensibles à la réduction de dimensionnalité, en fonction de la complexité des relations entre les variables et la variable cible.

Malgré ces observations, on peut mettre en avant deux points importants :

1. Réduction de la dimensionnalité : Les méthodes de sélection de caractéristiques permettent tout de même de réduire significativement la dimensionnalité du problème. Elles permettent de passer de 107 variables observées à un nombre beaucoup plus restreint (46 dans le cas de lasso). Cette réduction présente plusieurs avantages : a) Simplification du modèle : Un modèle plus simple est souvent plus facile à interpréter et à maintenir. b) Réduction du risque de surapprentissage : Moins de variables peut signifier moins de bruit et donc potentiellement une meilleure généralisation. c) Gain en temps de calcul : Moins de variables à traiter peut accélérer l'entraînement et l'inférence du modèle.
2. Focus sur les variables clés : La sélection de caractéristiques permet d'identifier les variables les plus informatives pour le modèle. Cela peut être particulièrement intéressant pour : a) Communication des résultats : On peut se concentrer sur ces quelques variables possédant beaucoup d'information pour le modèle lors de la présentation des résultats, rendant l'analyse plus accessible et percutante. b) Compréhension du phénomène étudié : Les variables sélectionnées peuvent offrir des insights sur les facteurs les plus influents dans le problème étudié. c) Orientation de futures collectes de données : Connaître les variables les plus importantes peut guider la conception de futures expériences ou collectes de données.

En conclusion, bien que l'impact sur les performances soit minime dans notre cas, la sélection de caractéristiques apporte une valeur ajoutée en termes de compréhension du modèle et de simplification du problème. Elle ouvre également des pistes pour de futures analyses plus ciblées.

3.2) IDENTIFICATION DES DIFFÉRENTS TROUBLES MENTAUX

3.2.1) DÉPRESSION ET INSOMNIE

Nous avons vu que nous arrivons à prédire avec une grande exactitude la présence d'un trouble de santé mentale grâce aux habitudes de vies. Mais il est probable que des habitudes de vies différentes soient liée à des troubles différents. De plus si plus tard nos observations veulent servir un système de recommandations d'habitudes de vie, différentes habitudes seraient recommandées en fonction des différentes pathologies rencontrés. Pour rappel trois troubles ont été sélectionné dans notre base de données l'insomnie, l'anxiété et la dépression. Concentrons-nous ici sur les données ESSAIM qui se focalisent plus sur les habitudes de vies et analysons tout d'abord le cas de la dépression.

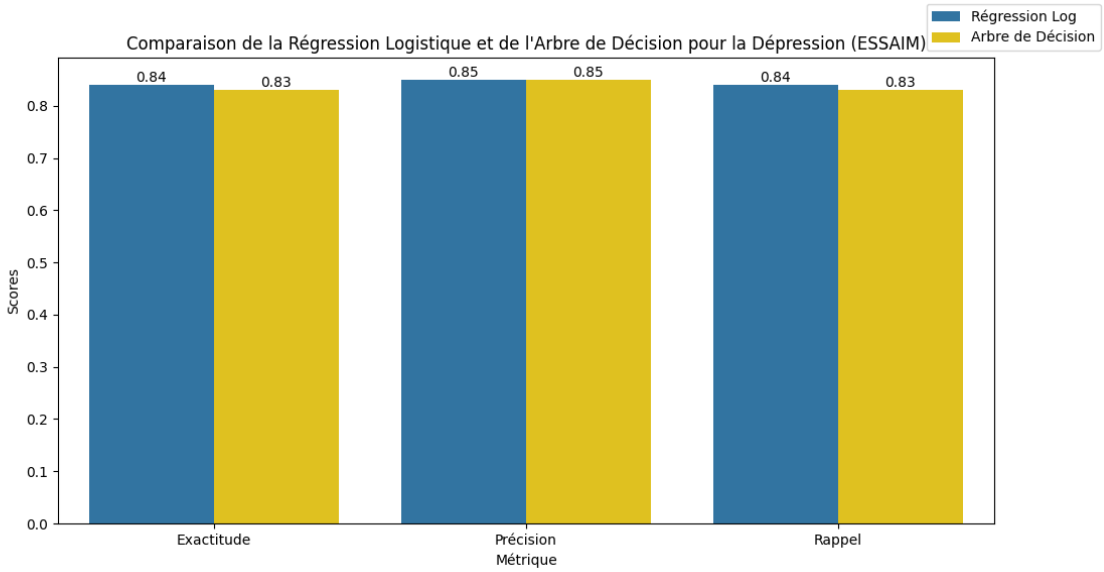


Figure 5:Comparaison de nos résultats pour la dépression en fonction du modèle utilisé

Nous avons constaté que nous arrivons à distinguer la dépression aussi bien que la présence d'un trouble mental en général. Cela est une nouvelle encourageante car nous pourrions, à terme, formuler des recommandations spécifiques aux problèmes de dépression, plutôt que de proposer des conseils généraux sur la santé mentale. Les graphiques relatifs à l'insomnie et à l'anxiété sont de même nature (voir annexes 2 et 3). L'anxiété, étant moins présente dans notre base de données, souffre d'un déficit de rappel (recall) et de précision (precision) par rapport aux deux autres modèles. Nous avons donc décidé d'explorer des méthodes d'équilibrage des classes pour remédier à ce problème.

3.2.2) CAS PARTICULIER DE L'ANXIÉTÉ

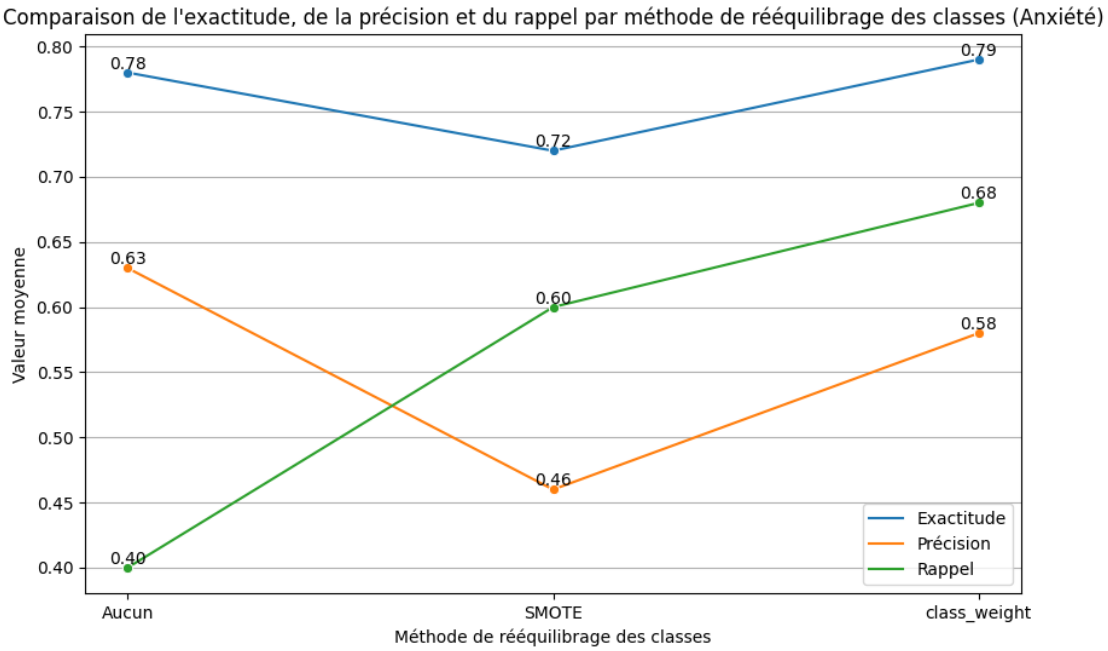


Figure 6: comparaison des métriques pour l'anxiété à l'aide de l'équilibrage des classes

Ce graphique est intéressant pour deux raisons principales. Tout d'abord, il démontre l'efficacité de la méthode d'équilibrage par pondération (class_weight). En appliquant cette méthode, nous avons pu obtenir une meilleure accuracy et un meilleur recall sans trop affecter notre précision. Dans notre cadre de travail, nous préférons surestimer le nombre de cas d'anxiété sévère plutôt que de les sous-estimer, afin de pouvoir apporter plus de conseils à un plus grand nombre de personnes dans le besoin. En apportant plus d'importance à la détection des personnes atteintes d'anxiété, nous parvenons à identifier près de 70 % d'entre elles.

L'application de la pondération des classes a ainsi permis d'améliorer le rappel (recall) de notre modèle pour l'anxiété, qui est un paramètre crucial dans notre contexte. En effet, l'objectif principal est de maximiser le nombre de cas identifiés, même si cela implique une légère diminution de la précision. Cette stratégie est particulièrement pertinente dans le domaine de la santé mentale, où le coût d'une non-détection est bien plus élevé que celui d'une fausse alerte. Le procédé d'équilibrage a été tester pour d'autres troubles minoritaires et offre des performances variables (annexes 4).

Ces résultats ouvrent la voie à la mise en place d'un système de recommandations personnalisées basé sur les troubles spécifiques. Pour la dépression, par exemple, des recommandations pourraient inclure des habitudes de vie telles que l'amélioration de l'hygiène de sommeil, la pratique régulière d'une activité physique modérée, et des techniques de gestion du stress. Pour l'anxiété, les recommandations pourraient inclure des techniques de relaxation, des exercices de respiration, et une réduction de la consommation de stimulants comme la caféine. Mais nous discuterons plus de cela dans la partie suivante

En conclusion, la capacité à distinguer les différents troubles mentaux grâce aux habitudes de vie, combinée à des méthodes d'équilibrage des classes pour améliorer la détection des troubles moins fréquents, constitue une avancée significative. Cela permettrait d'offrir des recommandations de vie plus ciblées et potentiellement plus efficaces, contribuant ainsi à une meilleure prise en charge des troubles mentaux.

3.3) IMPORTANCES DES DIFFÉRENTES HABITUDES DE VIES

Pour répondre à cette question, Nous devons nous concentrer en priorité sur la base de données Essaim car c'est dans cette base de données que nous avons le plus d'habitudes de vie.

3.3.1) HABITUDES DE VIE LIÉES A UN TROUBLE DE LA SANTÉ MENTALE

Ce que nous pouvons constater, c'est que pour la variable cible « obj », c'est-à-dire celle qui détermine si une personne est atteinte d'un trouble, l'habitude de vie la plus impactante est le sommeil. Peu importe la méthode utilisée pour sélectionner les caractéristiques ou si les données sont équilibrées ou non, les graphiques SHAP montrent que le sommeil reste toujours la caractéristique la plus influente. Les résultats des graphiques SHAP varient légèrement selon la méthode de sélection des caractéristiques et l'équilibrage des données, mais le sommeil reste l'habitude de vie qui ressort le plus.

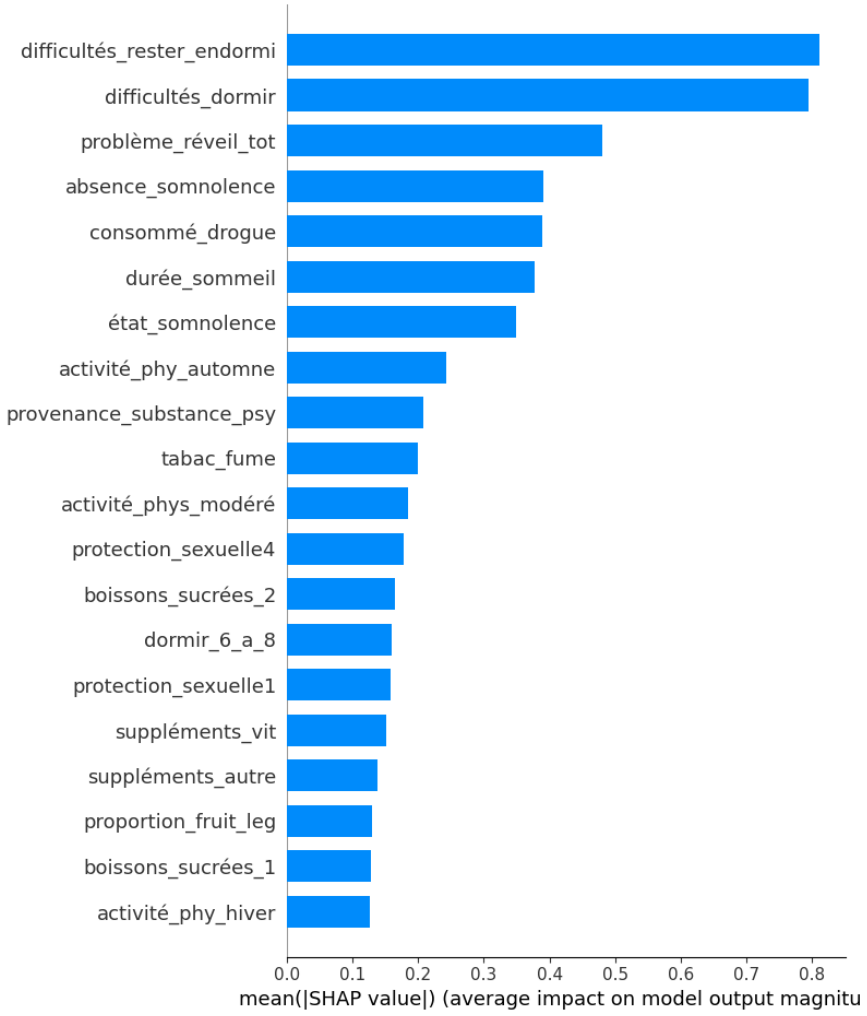


Figure 7: Importance des caractéristiques selon les valeurs Shap pour la variable cible "obj"

Le graphique ci-dessus (Figure 7), nous montre l'importance des caractéristiques dans un modèle prédictif, mesurée par la valeur moyenne des contributions de SHAP (SHapley Additive exPlanations).

Voici une explication sur le fait que le sommeil est l'habitude de vie la plus déterminante pour expliquer notre modèle :

1. Caractéristiques liées au sommeil : Les premières caractéristiques les plus influentes sont toutes des sous-ensembles de l'habitude de vie "sommeil". Par exemple :
   - difficultés_dormir: Difficultés à s'endormir.
   -diffcultés_rester_endormi: difficulté à rester endormi.
   - problème_reveil_tot : problème de réveil trop tôt le matin
2. Importance des valeurs de SHAP : Les valeurs de SHAP montrent l'impact moyen de chaque caractéristique sur le modèle. Les caractéristiques liées au sommeil ont des valeurs de SHAP plus élevées, indiquant qu'elles contribuent le plus aux variations des prédictions du modèle.

3. Rôle crucial du sommeil : Les caractéristiques liées au sommeil apparaissent en haut de la liste, ce qui signifie que le sommeil est un facteur crucial pour expliquer les prédictions de notre modèle. Les habitudes de sommeil affectent plusieurs aspects de la santé et du bien-être, ce qui peut expliquer leur importance élevée dans le modèle.

4. Autres caractéristiques : Bien que d'autres habitudes de vie, comme la consommation alimentaire (par exemple, alim_diff_arret) et le bien-être sexuel (par exemple, sbx_be_freq), soient présentes, elles ont un impact relativement moindre comparé aux caractéristiques liées au sommeil.

En conclusion, ce graphique montre clairement que le sommeil, par ses multiples dimensions, est l'habitude de vie la plus explicative dans notre modèle prédictif pour la variable cible « obj » Cela souligne l'importance du sommeil dans la santé et le bien-être général, ainsi que son rôle central dans la performance de notre modèle.

Nous pouvons maintenant nous concentrer sur les autres variables cibles que nous avons choisi, c’est-à-dire anxiété, dépression et insomnie.

3.3.2) HABITUDES DE VIE LIÉES A UN TROUBLE DE L'INSOMNIE

Si nous commençons par l'insomnie (Figure 8), nous remarquons bien évidemment que le sommeil est l'habitude de vie qui a le plus d'impact vu qu'ils sont étroitement liés mais si on ne tient pas compte du sommeil on constate que 2 habitudes de vies viennent impacter notre modèle.

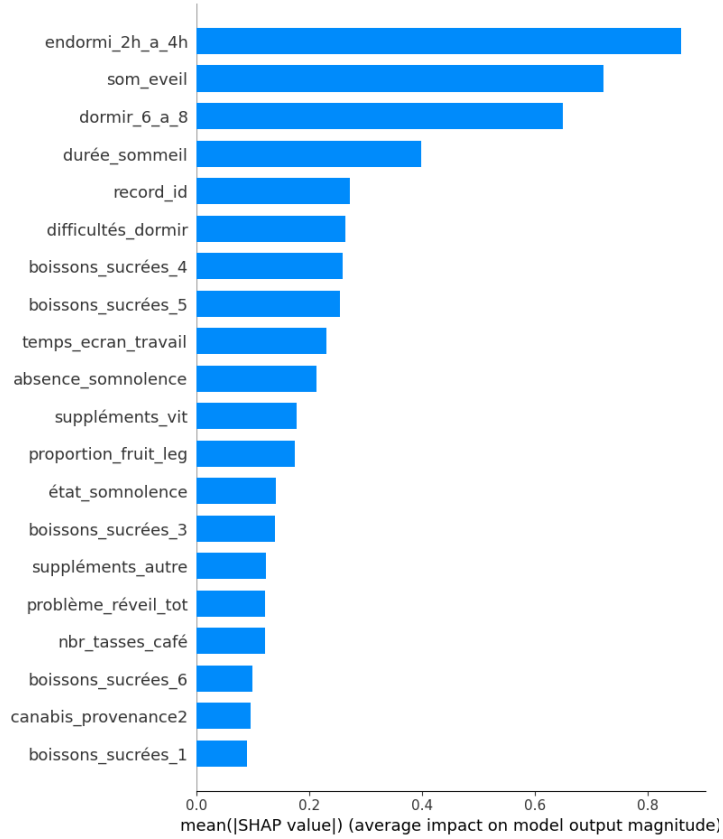


Figure 8:Importance des caractéristiques selon les valeurs Shap pour la variable cible "Insomnie"

Ces 2 habitudes de vies sont l'alimentation et l'activité physique. Ces résultats sont logiques car une mauvaise alimentation(boisson sucrée, trop de café avant de dormir,…) et activité physique inadéquate(les personnes qui ne font pas vraiment de sport qui restent derrière leur écrans, …) peuvent toutes deux perturber le sommeil, entraînant des difficultés à dormir et des insomnies. Tandis qu'une bonne alimentation et une bonne activité physique peuvent améliorer la facilité à s'endormir et la qualité générale du sommeil, réduisant ainsi le risque d'insomnie.

Sur les graphiques en annexe(annexes 6 à 9) vous pouvez également voir que pour l'insomnie, les habitudes de consommation de substances peuvent également impacter cette variable.

3.3.3) HABITUDES DE VIE LIÉES A UN TROUBLE De DÉPRESSION

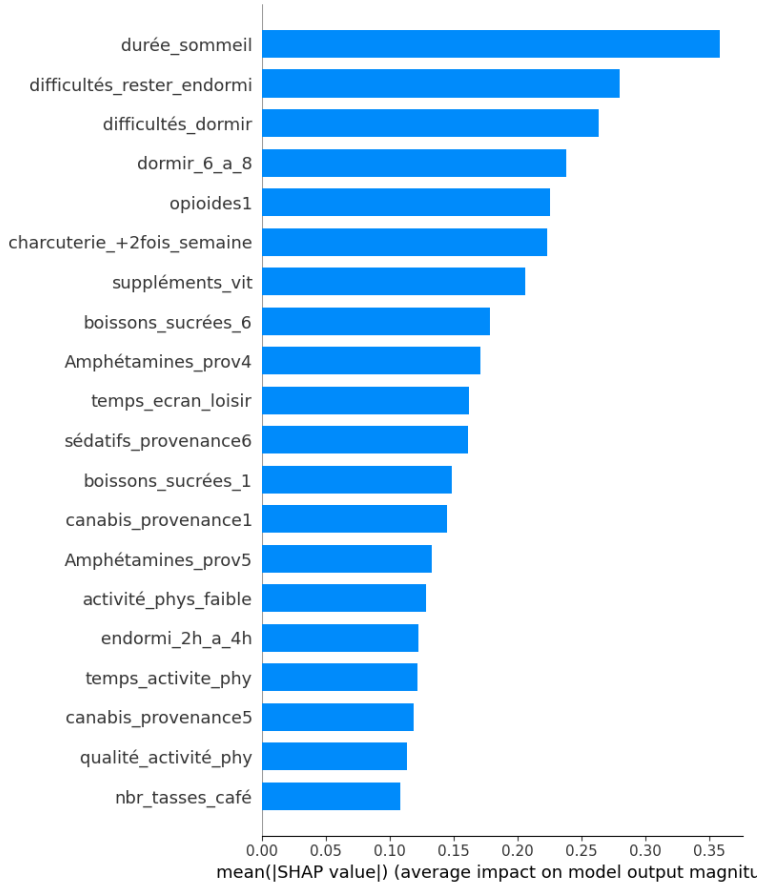


Figure 9: Importance des caractéristiques selon les valeurs Shap pour la variable cible "Dépression"

Sur le graphique ci-dessus (Figure 9), nous remarquons qu'encore une fois le sommeil est l'habitude de vie qui a le plus d'impact pour prédire si une personne souffre de dépression. Cependant, on constate également que les habitudes d'alimentation ainsi que les habitudes de consommation (drogue et médicaments) ont un impact sur la prédiction de la dépression.

3.3.4) HABITUDES DE VIE LIÉES A UN TROUBLE D'ANXIÉTÉ

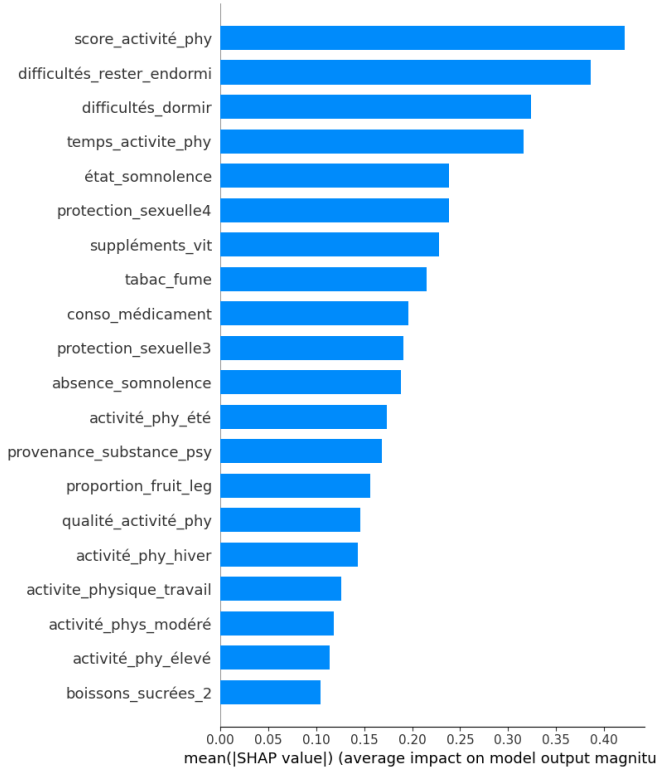


Figure 10:Importance des caractéristiques selon les valeurs Shap pour la variable cible "Anxiété"

En ce qui concerne la prédiction de l'anxiété le graphique ci-dessus (Figure 10), nous remarquons que l'activité physique est très importante en effet le score d'activité physique est l'élément le plus impactant pour déterminer si une personne est anxieuse. Ensuite nous avons encore les habitudes de sommeil, suivi des habitudes d'alimentations et de consommations de tabac d'alcool, etc.

En conclusion, comme vous pouvez le voir sur le tableau ci-dessous,

Trouble	Objectif (avoir un trouble sans distinction)	Insomnie	Dépression	Anxiété
Habitude la plus impactante	Sommeil	Sommeil	Sommeil	Activité physique
2ème habitude	Consommation	Alimentation	Consommation	Sommeil
3ème habitude	Activité physique	Activité physique	Alimentation	Consommation
4ème habitude	Alimentation	Consommation	Activité physique	Alimentation

Les quatre habitudes de vie retenues sont celles du sommeil, de l'activité physique, de l'alimentation et de la consommation. Parmi elles, les habitudes de sommeil sont les plus impactantes pour détecter si une personne souffre d'un trouble de santé mentale. Bien que certaines différences existent pour identifier le type de trouble spécifique, les habitudes de sommeil restent constamment parmi les caractéristiques les plus influentes (1ère ou 2ème position) pour chacune de nos variables cibles.

3.4) COMPROMIS ENTRE EXPLICABILITÉ ET PRÉCISION

Explication des Performances et de la Préférence pour des Modèles Explicables

Dans notre étude, nous avons comparé les performances des arbres de décision et des régressions logistiques pour prédire des résultats dans deux bases de données différentes : ULA et ESSAIM. De plus, nous avons décidé d'utiliser des modèles plus explicables pour plusieurs raisons importantes :

- Interprétabilité : Les modèles comme les régressions logistiques et les arbres de décision sont beaucoup plus faciles à interpréter. Cela permet de comprendre quels facteurs influencent les prédictions, ce qui est essentiel dans des domaines comme la santé mentale.
- Transparence : Dans notre étude, il est crucial que les utilisateurs finaux (les cliniciens et les chercheurs en santé mentale) puissent comprendre et faire confiance aux modèles. Les décisions basées sur des modèles explicables sont plus faciles à justifier.

Cependant, nous avons également tenté d'implémenter des modèles plus complexes comme les réseaux de neurones, forêt aléatoire, le xgmboost/lightgmboost et le modèle SVM. Le but étant de voir si cela pouvait améliorer significativement nos résultats. Voici une analyse détaillée des résultats obtenus dans notre étude.

Contexte de Notre Étude

Nous avons utilisé deux types de modèles explicables : les arbres de décision et les régressions logistiques, pour prédire des résultats dans les bases de données ULA et ESSAIM. Les variables cibles étaient "avoir un trouble" pour ULA et "obj" pour ESSAIM.

1. Base de Données ULA

Modèle	Exactitude	Précision	Rappel
Régression Logistique	0.76	0.78	0.73
Arbre de décision	0.74	0.70	0.80
Foret aléatoire	0.75	0.72	0.78
SVM	0.75	0.72	0.78
Lightgmboost	0.75	0.72	0.77
Xgboost	0.74	0.70	0.80

Réseaux de neurones	0.75	0.73	0.74

Tableau 4: Résultats ULA

2. Base de Données ESSAIM

Modèle	Exactitude	Précision	Rappel
Régression Logistique	0.82	0.85	0.84
Arbre de décision	0.80	0.86	0.80
Foret aléatoire	0.78	0.79	0.84
SVM	0.76	0.79	0.81
lightgmboost	0.79	0.78	0.89
Xgboost	0.77	0.74	0.94
Réseaux de neurones	0.77	0.75	0.82

Tableau4 : Résultat ESSAIM

Comparaison avec des Modèles Complexes

On le voit avec le tableau les modèles plus complexes ne semblent pas mieux prédire la présence d'un trouble de santé mentale. Pour les données issues de ULA tous les modèles ont des performances extrêmement proches, et sont indifférenciables. En ce qui concerne les données essaim Les modèles plus complexes lightgmboost et Xgboost se distingue par leur rappel. Ils arrivent à recouvrir entre 89% et 94% des personnes présentant un trouble aux prix d'une perte de précision. Rappelons que la base de données ESSAIM est plus fournie en habitude de vie ce qui peut expliquer une plus grande différenciation entre les modèles. Si l'on recueille beaucoup d'habitude de vie dans l'application finale ces deux modèles peuvent avoir un intérêt pour garantir de recouvrir la quasi-totalité des personnes présentant un trouble de la santé mentale. (Neal, 2019)

Nous nous sommes demandé pourquoi nos modèles plus complexes n'ont pas réussi à obtenir de meilleures performances que les modèles plus simples. Nous pensons avoir atteint un point d'équilibre en variance et biais par nos différents modèles. C'est-à-dire qu'en augmentant la complexité des modèles on arrive à faire baisser le biais mais le modèle se sur adapte aux données d'entrainement et donc se généralise moins bien. Étant donné cette équilibre l'utilisation de modèles plus explicables comme l'arbre de décision nous parait pertinent.(Mitliagkas, 2019)

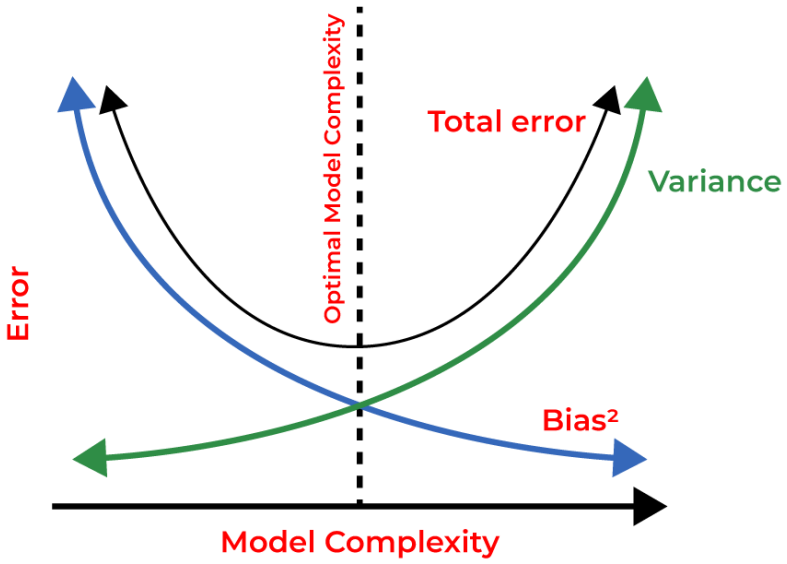


Figure 11: Compromis entre Biais-Variance

L'arbre de décision se distingue particulièrement par son explicabilité. Sa structure visuelle intuitive, sous forme d'arbre, permet de suivre le cheminement de la décision, de la racine aux feuilles, rendant le modèle compréhensible même pour les non-spécialistes (voir annexe 12). De plus, chaque branche de l'arbre peut être traduite en une règle de décision claire et explicite, facilitant ainsi l'interprétation et la communication des résultats. Enfin, l'arbre de décision permet de quantifier l'importance de chaque variable dans la prédiction, ce qui aide à identifier les facteurs les plus influents sur la santé mentale.

Cette explicabilité est cruciale dans le contexte de la santé mentale pour plusieurs raisons. Elle favorise la confiance et l'adoption du modèle par les professionnels de santé, qui doivent pouvoir comprendre le raisonnement derrière les prédictions pour prendre des décisions éclairées. Elle permet également une communication transparente avec les patients, en leur expliquant clairement les facteurs qui ont conduit à un diagnostic ou une recommandation.

Enfin, l'explicabilité est une question d'éthique et de responsabilité, permettant de justifier les décisions prises par le modèle, notamment dans un domaine aussi sensible que la santé mentale.

En ce qui concerne, la régression logistique elle est souvent considérée comme plus complexe à interpréter que les arbres de décision. En régression logistique, les coefficients représentent les log-odds d'appartenir à une classe, ce qui nécessite une compréhension approfondie des log-odds et des probabilités pour les interpréter correctement. De plus, la visualisation des effets des variables indépendantes sur la variable dépendante implique des graphiques complexes, rendant l'interprétation moins intuitive. Les interactions entre les variables doivent être spécifiées explicitement, ajoutant une couche de complexité. Tout cela rend les prédictions de ce modèle plus complexes à expliquer, en comparaison avec l’arbre de décision.

En conclusion, au vu des résultats obtenus, il est évident que nous préférons nous concentrer sur des modèles simples et explicables, comme les arbres de décision. Ceux-ci offrent des performances similaires aux modèles plus complexes, tout en étant plus faciles à interpréter et à mettre en œuvre.

PARTIE 4) LIMITES ET CONCLUSION

4.1) ANALYSE MANAGÉRIALE

Nous nous sommes peu concentrés sur cet aspect pendant le reste du mémoire, mais plusieurs implications managériales importantes peuvent être tirées de notre étude. Tout d'abord, nous partons du constat qu'il est utile pour n'importe quelle organisation d'avoir ses membres en bonne santé, notamment mentale, pour les maintenir performants. De plus, comme observé dans la base de données ULA, il n'est pas rare que des structures professionnelles mettent en place des cellules pour s'occuper de la santé mentale de leurs employés. Nous avons donc plusieurs conseils à donner à ces cellules et aux organisations en général.

Premièrement, mesurer la santé mentale des individus peut être difficile, notamment dans un cadre professionnel. Peu de gens risquent de mettre en avant un fort sentiment d'anxiété ou de dépression si on leur fait passer des questionnaires directs. Il peut donc être intéressant de mesurer la santé mentale à l'aide de proxys. En effet, un questionnaire touchant aux habitudes de vie (sport, sommeil, alimentation, etc.) a plus de chance de voir des réponses sincères. Utiliser ces proxys permettrait ainsi de repérer des personnes à l'état potentiellement anxieux ou déprimées et d'apporter des conseils adaptés, sans les stigmatiser. Cette approche indirecte pourrait être complétée par un suivi longitudinal, permettant de détecter les évolutions et tendances au fil du temps.

Deuxièmement, si les habitudes de vie permettent de repérer les problèmes de santé mentale, elles peuvent aussi apporter des améliorations à celle-ci. Une organisation professionnelle pourrait donc faire plusieurs choix pour améliorer la santé mentale de son personnel sans paraître intrusive. Nous avons vu que la grande consommation de sucre ou de café peut causer des problèmes similaires, et donc la cantine professionnelle pourrait adapter ses menus. Les déplacements à pied ou à vélo ont aussi un impact positif et pourraient donc être encouragés, par exemple en mettant en place des infrastructures adaptées ou des incitations financières. Au-delà de l'alimentation et de l'activité physique, les organisations pourraient également agir sur d'autres aspects des habitudes de vie, comme le sommeil, en sensibilisant à son importance et en adaptant éventuellement les horaires de travail pour favoriser un meilleur rythme circadien. L'environnement de travail lui-même joue un rôle crucial dans la santé mentale des employés. Les organisations devraient donc veiller à créer des espaces de travail favorisant le bien-être, avec par exemple de la lumière naturelle, des zones de détente et des espaces verts. La gestion

de la charge de travail est également primordiale, avec une attention particulière à une répartition équilibrée des tâches et à la prévention du surmenage. Les relations interpersonnelles et la reconnaissance du travail accompli sont d'autres facteurs importants sur lesquels les organisations peuvent agir pour promouvoir un climat de travail positif et valorisant.

La formation et la sensibilisation constituent un autre axe d'action essentiel. Les managers devraient être formés à la détection des signes de mal-être et à la promotion du bien-être mental. Des campagnes de sensibilisation sur la santé mentale pourraient être organisées pour réduire la stigmatisation, et des ressources comme des ateliers, des webinaires ou de la documentation sur la gestion du stress et la santé mentale pourraient être mises à disposition des employés.
Enfin, pour s'assurer de l'efficacité des mesures mises en place, il est crucial de mettre en œuvre un suivi et une évaluation réguliers. Cela peut se faire à travers des indicateurs comme le taux d'absentéisme, le turnover, ou la satisfaction des employés, ainsi que par des enquêtes régulières sur le bien-être au travail. Les programmes devraient être ajustés en fonction des résultats obtenus, dans une démarche d'amélioration continue.

En conclusion, la prise en compte de la santé mentale dans le milieu professionnel est un enjeu majeur pour les organisations modernes. Notre étude met en lumière l'importance des habitudes de vie comme indicateurs et leviers d'amélioration de la santé mentale. En adoptant une approche holistique qui englobe la mesure indirecte, l'amélioration des habitudes de vie, la création d'un environnement favorable, la formation et le suivi, les organisations peuvent significativement améliorer le bien-être mental de leurs employés. Cela se traduira non seulement par une meilleure santé individuelle, mais aussi par une performance organisationnelle accrue, une réduction de l'absentéisme et une meilleure rétention des talents. Il est important de noter que ces recommandations doivent être adaptées au contexte spécifique de chaque organisation et mises en œuvre de manière éthique, en respectant la vie privée et l'autonomie des employés. La santé mentale au travail est un domaine en constante évolution, et les organisations devront rester à l'écoute des dernières recherches et des besoins changeants de leur personnel pour maintenir un environnement de travail sain et productif.

4.2) LIMITES ET CRITIQUES DU MÉMOIRE

Au cours de la réalisation de notre mémoire, plusieurs limites et critiques sont apparues de manière assez claire. Nous allons ici nous concentrer sur cinq aspects de notre mémoire qui auraient pu être améliorés et étendus.

1) Connaissance du domaine

Pour ce mémoire, nous nous sommes penchés sur l'identification des facteurs permettant de repérer les personnes souffrant de troubles psychologiques. Cependant, pour réaliser notre travail, nous n'avions pas l'expertise d'un spécialiste de la santé mentale. Avec cette expertise, nous aurions peut-être pu nous rendre compte rapidement des limites de notre base de données. De plus, cela nous aurait peut-être permis d'identifier rapidement quelles pathologies qui risquaient de poser un défi particulier à identifier. Cela nous aurait également permis d'analyser plus en profondeur les questions et réponses que notre analyse informatique a identifiées.

2) Limites de nos bases de données

Pour ce mémoire, nous avons exploité deux bases de données fournies par l'Université Laval au Québec, issues des projets ULA et ESSAIM. La base de données ULA se concentre sur la présence de troubles mentaux chez les employés de PME au Québec et contient des données sociodémographiques, des mesures de la santé mentale, des capacités d'autogestion et de connaissances en santé mentale, et de l'environnement de travail. La base de données ESSAIM, quant à elle, est issue d'un questionnaire administré à 930 Canadiens et se concentre sur les habitudes de vie (sommeil, alimentation, activité physique, etc.) et leur lien avec la santé mentale. Nous avons constaté que ces questions permettaient d'identifier avec une précision satisfaisante la présence d'un trouble mental à l'aide de nos deux modèles d'apprentissage automatique (arbre de décision et régression logistique). Cependant, pour les données ULA, le manque d'habitudes de vie et la présence très minoritaire de certains troubles a limité nos analyses. De manière générale il aurait été intéressant de disposer de plus de données sur les différents troubles afin de rendre plus performant des modèles d'apprentissage profonds (réseaux de neurones)

3) Modèles utilisés

Nous aurions pu aller plus loin dans la diversité des modèles utilisés. Nous nous sommes appuyés sur deux modèles principaux (arbre de décision et régression logistique) avec des variantes (sélection de caractéristiques et rééquilibrage de classes). Cependant, il aurait été

intéressant de mettre en place des modèles de deep learning ou d'autres modèles moins sophistiqués pour évaluer leurs performances. Pour être totalement transparents, nous avons également entraîné un réseau de neurones une forêt aléatoire et deux dérivées de la forêt aléatoire, mais faute de place, et de performance particulièrement intéressante, nous ne les avons que peu intégrés dans notre analyse (point 3.4 du travail). L'utilisation de plus de modèles aurait peut-être permis d'identifier des caractéristiques communes aux pathologies sous-représentées.

4) Collaboration avec le travail en aval

Dans le cadre de ce mémoire, nous n'avons pas eu de contact direct avec l'équipe chargée de développer le système de recommandation. Nos différentes analyses pourraient donc ne pas correspondre entièrement à leurs attentes. En collaborant avec eux, nous aurions peut-être pu leur fournir des éléments plus précis pour les aider à bâtir leur système de recommandations. De plus, la collaboration dans la recherche permet souvent de faire surgir des idées nouvelles et d'améliorer l'ensemble du travail. Une meilleure organisation de notre part pour faciliter cette collaboration aurait pu grandement bénéficier à notre projet.

5) Fiabilité des données auto-déclarées

Les données sur la présence de troubles mentaux sont auto-déclarées par les participants, ce qui peut introduire des biais de déclaration. Les participants peuvent sous-déclarer ou sur-déclarer leurs problèmes de santé mentale en fonction de divers facteurs, tels que la stigmatisation ou la perception personnelle de leurs symptômes. De plus les habitudes de vies ont également été auto-déclaré. Il aurait été très intéressant d'avoir des données fournies par des capteurs et des données auto déclarée. On aurait ainsi pu comparer ces deux types de données pour tester la fiabilité des capteurs et ou de la déclaration d'habitudes de vies.

Ces cinq points semblent constituer les principales limites de notre mémoire. Évidemment, nous pouvons penser à d'autres limites, telles que la limitation de l'échantillon, la limitation des méthodes de sélection de caractéristiques et l'évaluation des modèles.

4.3) CONCLUSION

Notre mémoire a exploré l'impact des habitudes de vie sur la prédiction des troubles de santé mentale, en utilisant deux bases de données distinctes : ULA, axée sur les travailleurs de PME au Québec, et ESSAIM, plus générale sur les habitudes de vie des Canadiens. Grâce à des modèles d'apprentissage automatique, notamment des arbres de décision et des régressions logistiques, nous avons pu identifier les facteurs clés influençant la santé mentale et évaluer leur impact.

Dans quelle mesure peut-on prédire les troubles de santé mentale, à l'aide des habitudes de vie ?

Nos résultats ont mis en lumière la capacité remarquable de ces modèles à prédire la présence d'un trouble de santé mentale en se basant uniquement sur les habitudes de vie. En effet, nos modèles ont atteint une précision de plus de 80 % pour identifier les individus souffrant d'un trouble mental, ce qui démontre le potentiel considérable de cette approche pour le dépistage précoce. De plus, les habitudes de vie se sont montrées plus efficaces pour prédire un trouble de santé mentale que les données socio-démographiques ou professionnelles. Enfin on a pu voir qu'en sélectionnant un nombre restreint d'habitude de vie (40 pour la régression logistique une 20aine pour les arbres de décision) on arrive quand même à prédire les troubles avec de bons résultats.

À quel point pouvons-nous identifier les différents troubles ?

Notre étude a également mis en évidence la capacité des modèles à différencier certains troubles spécifiques, comme la dépression et l'insomnie, en fonction des habitudes de vie. Cependant, nous avons rencontré des défis pour identifier des troubles moins fréquents dans nos données, tels que l'anxiété. L'utilisation de techniques d'équilibrage des classes a permis d'améliorer la part de personne présentant ce trouble d'être repérées mais en échange d'une perte de précision. Cela souligne l'importance de prendre en compte la prévalence des différents troubles lors de la construction de modèles prédictifs et dans la construction de base de données pour la recherche.

Parmi ces habitudes de vie, quels sont les facteurs qui ont le plus d'impact ?

Les habitudes de sommeil jouent un rôle prépondérant dans la prédiction des troubles de santé mentale, quel que soit le trouble spécifique (dépression, anxiété, insomnie). Cette découverte souligne l'importance cruciale du sommeil pour le bien-être mental et ouvre la voie à des

interventions ciblées pour améliorer la qualité du sommeil et prévenir les troubles mentaux. D'autres habitudes, comme l'activité physique et l'alimentation, ont également montré une influence significative, bien que variable selon le trouble considéré. Par exemple, l'alimentation et l'activité physique sont ressortis comme les plus importantes pour détecter l'insomnie.

À quel point le modèle perd-il en précision si nous voulons un modèle plus explicable ?

L'explicabilité des modèles a été un autre aspect clé de notre travail. Grâce à des techniques telles que les valeurs de Shapley (SHAP), nous avons pu comprendre les raisons derrière les prédictions des modèles, ce qui est essentiel pour leur application dans le domaine de la santé mentale. Cette transparence permet aux professionnels de santé de mieux interpréter les résultats et de prendre des décisions éclairées pour la prévention et le traitement des troubles mentaux. De plus, nous avons démontré l'intérêt d'utiliser des modèles plus simples et explicables, tels que les arbres de décision et les régressions logistiques, qui ont surpassé en performance un modèle de réseau de neurones dans notre contexte d'étude.

Observations Générales et Perspectives

Ce mémoire ouvre des perspectives prometteuses pour l'utilisation de la science des données dans la prévention, le diagnostic et le traitement des troubles de santé mentale. Il souligne l'importance d'intégrer les habitudes de vie dans l'évaluation de la santé mentale et met en évidence le potentiel des modèles d'apprentissage automatique pour identifier les individus à risque et personnaliser les interventions. Les chercheurs de l'Université Laval, qui ont fourni les bases de données, pourraient utiliser nos résultats pour concevoir des algorithmes de recommandation d'habitudes de vie personnalisés. Ces algorithmes pourraient être intégrés dans des applications mobiles ou des plateformes en ligne, offrant ainsi des conseils adaptés aux besoins spécifiques de chaque individu en fonction de leurs habitudes de vie et de leur risque de développer un trouble mental.

Cependant, notre étude présente certaines limites, notamment la difficulté à distinguer précisément entre différents troubles mentaux en raison de la nature des données disponibles. De futures recherches pourraient s'appuyer sur des ensembles de données plus riches et diversifiés, ainsi que sur des modèles plus sophistiqués, pour améliorer la précision des prédictions et la personnalisation des recommandations.

En somme, ce mémoire contribue à une meilleure compréhension des liens complexes entre les habitudes de vie et la santé mentale, et ouvre la voie à des applications concrètes de la science des données pour améliorer le bien-être mental des individus. Il souligne l'importance d'une approche holistique et personnalisée de la santé mentale, où les habitudes de vie jouent un rôle central.

SOURCES

Ajith, M., Aycock, D. M., Tone, E. B., Liu, J., Misiura, M. B., Ellis, R., King, T. Z., Dotson, V. M., & Calhoun, V. (2023). A Deep Learning Approach for Mental Health Quality Prediction Using Functional Network Connectivity and Assessment Data. *bioRxiv*, 2023.2006.2001.543257. https://doi.org/10.1101/2023.06.01.543257

Alessi, S. M., & Petry, N. M. (2013). A randomized study of cellphone technology to reinforce alcohol abstinence in the natural environment. *Addiction*, *108*(5), 900-909. https://doi.org/10.1111/add.12093

Battista, K., Diao, L., Patte, K. A., Dubin, J. A., & Leatherdale, S. T. (2023). Examining the use of decision trees in population health surveillance research: an application to youth mental health survey data in the COMPASS study. *Health promotion and chronic disease prevention in Canada : research, policy and practice*, *43 2*, 73-86.

Boden, J. M., & Fergusson, D. M. (2011). Alcohol and depression. *Addiction*, *106*(5), 906-914. https://doi.org/10.1111/j.1360-0443.2010.03351.x

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2012). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*. https://doi.org/10.1007/s10115-012-0487-8

Boushey, C. J., Spoden, M., Zhu, F. M., Delp, E. J., & Kerr, D. A. (2017). New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. *Proc Nutr Soc*, *76*(3), 283-294. https://doi.org/10.1017/s0029665116002913

Brazier, J., Connell, J., Papaioannou, D., Mukuria, C., Mulhern, B., Peasgood, T., Lloyd Jones, M., Paisley, S., O'Cathain, A., Barkham, M., Knapp, M., Byford, S., Gilbody, S., & Parry, G. (2014). A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technol Assess*, *18*(34). https://doi.org/10.3310/hta18340

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010950718922

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification And Regression Trees. https://doi.org/10.1201/9781315139470

Burges, C. J. C. (1998). *Data Mining and Knowledge Discovery*, *2*(2), 121-167. https://doi.org/10.1023/a:1009715923555

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28. https://doi.org/https://doi.org/10.1016/j.compeleceng.2013.11.024

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Cohen, S. (2004). Social relationships and health. *American psychologist*, *59*(8), 676.

Darzi, P. (2023). Could Artificial Intelligence be a Therapeutic for Mental Issues? *Science Insights*, *43*(5), 1111-1113. https://doi.org/10.15354/si.23.co132

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthc J*, *6*(2), 94-98. https://doi.org/10.7861/futurehosp.6-2-94

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am J Cardiol*, *64*(5), 304-310. https://doi.org/10.1016/0002-9149(89)90524-9

Dishman, R. K., Berthoud, H. R., Booth, F. W., Cotman, C. W., Edgerton, V. R., Fleshner, M. R., Gandevia, S. C., Gomez-Pinilla, F., Greenwood, B. N., Hillman, C. H., Kramer, A. F., Levin, B. E., Moran, T. H., Russo-Neustadt, A. A., Salamone, J. D., Van Hoomissen, J. D., Wade, C. E., York, D. A., & Zigmond, M. J. (2006). Neurobiology of exercise. *Obesity (Silver Spring)*, *14*(3), 345-356. https://doi.org/10.1038/oby.2006.46

Dong, G., Tang, M., Cai, L., Barnes, L., & Boukhechba, M. (2021). *Semi-supervised Graph Instance Transformer for Mental Health Inference*. https://doi.org/10.1109/ICMLA52953.2021.00198

Dwyer, D., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Elovanio, M., Hakulinen, C., Pulkki-Råback, L., Aalto, A.-M., Virtanen, M., Partonen, T., & Suvisaari, J. (2020). General Health Questionnaire (GHQ-12), Beck Depression Inventory (BDI-6), and Mental Health Index (MHI-5): psychometric and predictive properties in a Finnish population-based sample. *Psychiatry Research*, *289*, 112973. https://doi.org/https://doi.org/10.1016/j.psychres.2020.112973

Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., Kim, J. S., Heo, S., Alves, H., White, S. M., Wojcicki, T. R., Mailey, E., Vieira, V. J., Martin, S. A., Pence, B. D., Woods, J. A., McAuley, E., & Kramer, A. F. (2011). Exercise training increases size of hippocampus and improves memory. *Proc Natl Acad Sci U S A*, *108*(7), 3017-3022. https://doi.org/10.1073/pnas.1015950108

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of artificial intelligence research*, *61*, 863-905. https://doi.org/10.1613/jair.1.11192

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10.1186/1472-6947-12-8

Fluharty, M., Taylor, A. E., Grabski, M., & Munafò, M. R. (2017). The Association of Cigarette Smoking With Depression and Anxiety: A Systematic Review. *Nicotine Tob Res*, *19*(1), 3-13. https://doi.org/10.1093/ntr/ntw140

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification [J]. *Journal of Machine Learning Research - JMLR*, *3*.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139. https://doi.org/https://doi.org/10.1006/jcss.1997.1504

Galderisi, S., Heinz, A., Kastrup, M., Beezhold, J., & Sartorius, N. (2015). Toward a new definition of mental health. *World Psychiatry*, *14*(2), 231-233. https://doi.org/https://doi.org/10.1002/wps.20231

Gandaputra, S., Waluyo, I., Efendi, F., & Wang, J.-Y. (2021). Insomnia Status of Middle School Students in Indonesia and Its Association with Playing Games before Sleep: Gender Difference. *International Journal of Environmental Research and Public Health*, *18*, 691. https://doi.org/10.3390/ijerph18020691

Gilman, S. E., & Abraham, H. D. (2001). A longitudinal study of the order of onset of alcohol dependence and major depression. *Drug and Alcohol Dependence*, *63*(3), 277-286. https://doi.org/10.1016/S0376-8716(00)00216-7

Guyon, I., & Elisseeff, A. (2003). An Introduction of Variable and Feature Selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, *3*, 1157-1182. https://doi.org/10.1162/153244303322753616

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, *46*, 389-422. https://doi.org/10.1023/A:1012487302797

Hanchuan, P., Fuhui, L., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

Harrell, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. https://doi.org/10.1007/978-3-319-19425-7

Harvey, A. G., Jones, C., & Schmidt, D. A. (2003). Sleep and posttraumatic stress disorder: a review. *Clinical Psychology Review*, *23*(3), 377-407. https://doi.org/https://doi.org/10.1016/S0272-7358(03)00032-1

Haslam, C., Cruwys, T., & Haslam, S. A. (2014). "The we's have it": evidence for the distinctive benefits of group engagement in enhancing cognitive health in aging. *Soc Sci Med*, *120*, 57-66. https://doi.org/10.1016/j.socscimed.2014.08.037

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. *Springer Series in Statistics*. https://doi.org/10.1007/978-0-387-84858-7

He, X., Cai, D., & Niyogi, P. (2005). *Laplacian Score for Feature Selection* (Vol. Vol. 18).

Holt-Lunstad, J., Smith, T. B., & Layton, J. B. (2010). Social Relationships and Mortality Risk: A Meta-analytic Review. *PLOS Medicine*, *7*(7), e1000316. https://doi.org/10.1371/journal.pmed.1000316

Hosmer, D. W. L., Stanley; Sturdivant, Rodney X. (2013). *Applied Logistic Regression* (3rd Edition ed.). Wiley. https://doi.org/10.1002/9781118548387

Huang, C.-y., Liao, H.-y., & Chang, S.-H. (1998). Social desirability and the clinical self-report inventory: Methodological reconsideration. *Journal of Clinical Psychology*, *54*(4), 517-528. https://doi.org/https://doi.org/10.1002/(SICI)1097-4679(199806)54:4<517::AID-JCLP13>3.0.CO;2-I

Iacus, S. M., Porro, G., Salini, S., & Siletti, E. (2015). Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. *arXiv preprint arXiv:1512.01569*.

Jacka, F. N., Pasco, J. A., Mykletun, A., Williams, L. J., Hodge, A. M., O'Reilly, S. L., Nicholson, G. C., Kotowicz, M. A., & Berk, M. (2010). Association of Western and traditional diets with depression and anxiety in women. *Am J Psychiatry*, *167*(3), 305-311. https://doi.org/10.1176/appi.ajp.2009.09060881

Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. https://doi.org/10.1109/wccct.2016.25

Jeon, H., & Oh, S. (2020). Hybrid-Recursive Feature Elimination for Efficient Feature Selection. *Applied Sciences*, *10*(9), 3211. https://www.mdpi.com/2076-3417/10/9/3211

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675-687. https://doi.org/https://doi.org/10.1016/j.beth.2020.05.002

Jolliffe, I. (2002). Principal Components Analysis.

Keyes, C. (2013). Mental Health as a Complete State: How the Salutogenic Perspective Completes the Picture. *Bridging Occupational, Organizational and Public Health: A Transdisciplinary Approach*, 179-192. https://doi.org/10.1007/978-94-007-5640-3-11

Kim, F. D.-V. e. B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* ([eprint] 1702.08608). https://arxiv.org/abs/1702.08608

Kohavi, R., & John, G. (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence*, *97*, 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X
Korgaonkar, M. S., Fornito, A., Williams, L. M., & Grieve, S. M. (2014). Abnormal structural networks characterize major depressive disorder: a connectome analysis. *Biol Psychiatry*, *76*(7), 567-574. https://doi.org/10.1016/j.biopsych.2014.02.018
Lai, J. S., Hiles, S., Bisquera, A., Hure, A. J., McEvoy, M., & Attia, J. (2014). A systematic review and meta-analysis of dietary patterns and depression in community-dwelling adults. *Am J Clin Nutr*, *99*(1), 181-197. https://doi.org/10.3945/ajcn.113.069880
Lauer, C. J., & Krieg, J.-C. (2004). Sleep in eating disorders. *Sleep Medicine Reviews*, *8*(2), 109-118. https://doi.org/https://doi.org/10.1016/S1087-0792(02)00122-3
Lin, S., Wu, Y., & Fang, Y. (2021). Comparison of Regression and Machine Learning Methods in Depression Forecasting Among Home-Based Elderly Chinese: A Community Based Study. *Front Psychiatry*, *12*, 764806. https://doi.org/10.3389/fpsyt.2021.764806
Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 491-502.
Logan, D. E., Claar, R. L., & Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *PAIN*, *136*(3), 366-372. https://doi.org/10.1016/j.pain.2007.07.015
Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14-23. https://doi.org/https://doi.org/10.1002/widm.8
Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*
Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.
Marill, K. A. (2004a). Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. *Academic Emergency Medicine*, *11*(1), 94-102. https://doi.org/https://doi.org/10.1197/j.aem.2003.09.006
Marill, K. A. (2004b). Advanced Statistics: Linear Regression,Part I: Simple Linear Regression. *Academic Emergency Medicine*, *11*(1), 87-93. https://doi.org/https://doi.org/10.1197/j.aem.2003.09.005
Mechanick, J. I., & Zhao, S. (2020). Wearable Technologies in Lifestyle Medicine. In J. I. Mechanick & R. F. Kushner (Eds.), *Creating a Lifestyle Medicine Center: From Concept to Clinical Practice* (pp. 133-143). Springer International Publishing. https://doi.org/10.1007/978-3-030-48088-2_13
Menard, S. (2002). *Applied Logistic Regression Analysis* https://doi.org/10.4135/9781412983433
Mitliagkas, B. N. a. S. M. a. A. B. a. V. T. a. M. S. a. S. L.-J. a. I. (2019). A Modern Take on the Bias-Variance Tradeoff in Neural Networks. https://arxiv.org/abs/1810.08591
Murthy, S. K., Kasif, S., & Salzberg, S. (1994). A system for induction of oblique decision trees. *Journal of artificial intelligence research*, *2*, 1-32.
Neal, B. (2019). On the Bias-Variance Tradeoff: Textbooks Need an Update. *CoRR*, *abs/1912.08286*. http://arxiv.org/abs/1912.08286
Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.00076
Nicolas Peltre, T. S. (2021). Les valeurs de Shapley en intelligibilité locale des modèles. https://www.quantmetry.com/blog/valeurs-de-shapley/
Owusu, E., Quainoo, R., Mensah, S., & Appati, J. (2023). A Deep Learning Approach for Loan Default Prediction Using Imbalanced Dataset. *International Journal of Intelligent Information Technologies*, *19*, 1-16. https://doi.org/10.4018/IJIIT.318672
Praagman, J. (1985). Classification and regression trees: Leo BREIMAN, Jerome H. FRIEDMAN, Richard A. OLSHEN and Charles J. STONE The Wadsworth

Statistics/Probability Series, Wadsworth, Belmont, 1984, x+ 358 pages. In: North-Holland.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81-106.
Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., & Golub, T. (2002). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, *98*. https://doi.org/10.1073/pnas.211566398
Rao, T. S., Asha, M. R., Ramesh, B. N., & Rao, K. S. (2008). Understanding nutrition, depression and mental illnesses. *Indian J Psychiatry*, *50*(2), 77-82. https://doi.org/10.4103/0019-5545.42391
Rebar, A. L., Stanton, R., Geard, D., Short, C., Duncan, M. J., & Vandelanotte, C. (2015). A meta-meta-analysis of the effect of physical activity on depression and anxiety in non-clinical adult populations. *Health Psychol Rev*, *9*(3), 366-378. https://doi.org/10.1080/17437199.2015.1022901
Reeve, S., Sheaves, B., & Freeman, D. (2015). The role of sleep dysfunction in the occurrence of delusions and hallucinations: A systematic review. *Clinical Psychology Review*, *42*, 96-115. https://doi.org/https://doi.org/10.1016/j.cpr.2015.09.001
Rish, I., Grabarnik, G., Cecchi, G., Pereira, F., & Gordon, G. (2008). *Closed-form supervised dimensionality reduction with generalized linear models*. https://doi.org/10.1145/1390156.1390261
Saeb, S., Zhang, M., Kwasny, M. M., Karr, C. J., Kording, K., & Mohr, D. C. (2015). The Relationship between Clinical, Momentary, and Sensor-based Assessment of Depression. *International Conference on Pervasive Computing Technologies for Healthcare : [proceedings]. International Conference on Pervasive Computing Technologies for Healthcare*, *2015*. https://doi.org/10.4108/icst.pervasivehealth.2015.259034
Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. https://doi.org/10.1093/bioinformatics/btm344
Sartorius, N. (2007). Stigma and mental health. *The Lancet*, *370*(9590), 810-811. https://doi.org/https://doi.org/10.1016/S0140-6736(07)61245-8
Scott, A. J., Webb, T. L., Martyn-St James, M., Rowse, G., & Weich, S. (2021). Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. *Sleep Medicine Reviews*, *60*, 101556. https://doi.org/https://doi.org/10.1016/j.smrv.2021.101556
Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426-1448. https://doi.org/10.1017/S0033291719000151
Shimada, K. (2023). The Role of Artificial Intelligence in Mental Health: A Review. *Science Insights*, *43*(5), 1119-1127. https://doi.org/10.15354/si.23.re820
Stein, D. J., Phillips, K. A., Bolton, D., Fulford, K. W., Sadler, J. Z., & Kendler, K. S. (2010). What is a mental/psychiatric disorder? From DSM-IV to DSM-V. *Psychol Med*, *40*(11), 1759-1765. https://doi.org/10.1017/s0033291709992261
Team, D. (2024). Logistic Regression - Simply explained. https://datatab.net/tutorial/logistic-regression
Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288. http://www.jstor.org/stable/2346178

Twala, B. E. T. H., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950-956. https://doi.org/https://doi.org/10.1016/j.patrec.2008.01.010

Waterman, A. (1993). Two Conceptions of Happiness: Contrasts of Personal Expressiveness (Eudaimonia) and Hedonic Enjoyment. *Journal of Personality and Social Psychology*, *64*. https://doi.org/10.1037/0022-3514.64.4.678

WHO. (2024). *Santé mentale : renforcer notre action*. Retrieved from https://www.who.int/fr/news-room/fact-sheets/detail/mental-health-strengthening-our-response

Wu, A. W., Jacobson, D. L., Berzon, R. A., Revicki, D. A., van der Horst, C., Fichtenbaum, C. J., Saag, M. S., Lynn, L., Hardy, D., & Feinberg, J. (1997). The effect of mode of administration on Medical Outcomes Study health ratings and EuroQol scores in AIDS. *Quality of Life Research*, *6*(1), 3-10. https://doi.org/10.1023/A:1026471020698

Yu, H., & Sano, A. (2020). Passive Sensor Data Based Future Mood, Health, and Stress Prediction: User Adaptation Using Deep Learning. *Annu Int Conf IEEE Eng Med Biol Soc*, *2020*, 5884-5887. https://doi.org/10.1109/embc44109.2020.9176242

Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.*, *5*, 1205–1224.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms* (Vol. 14). https://doi.org/10.1201/b12207

LISTE DES FIGURES

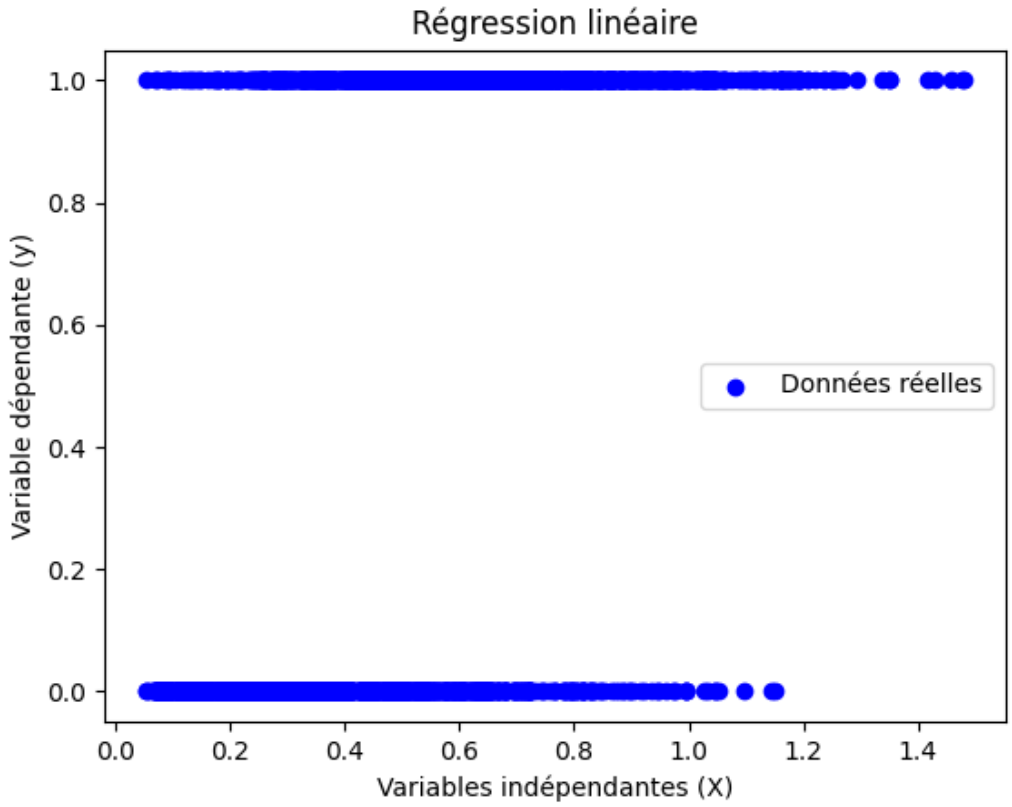
LISTE DES TABLES

LISTE DES ÉQUATIONS

ANNEXES

Annexe 1 : régression linéaire

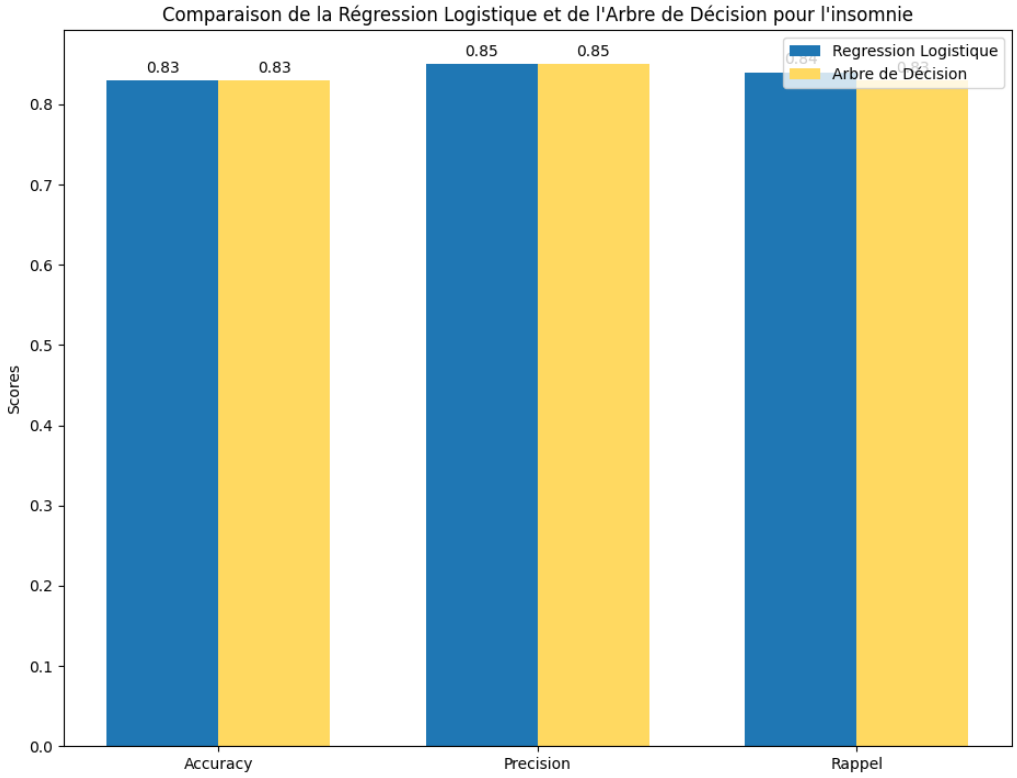
À la base nous avons testé si la régression linéaire pouvait être un bon modèle pour traiter nos données. Étant donné que nos données sont principalement des données catégorielles et que nos variables cibles sont des variables binaires, il est évident que la régression linéaire n'est pas le modèle approprié pour ce genre de données. Vu que celle-ci considère les données comme étant continues.



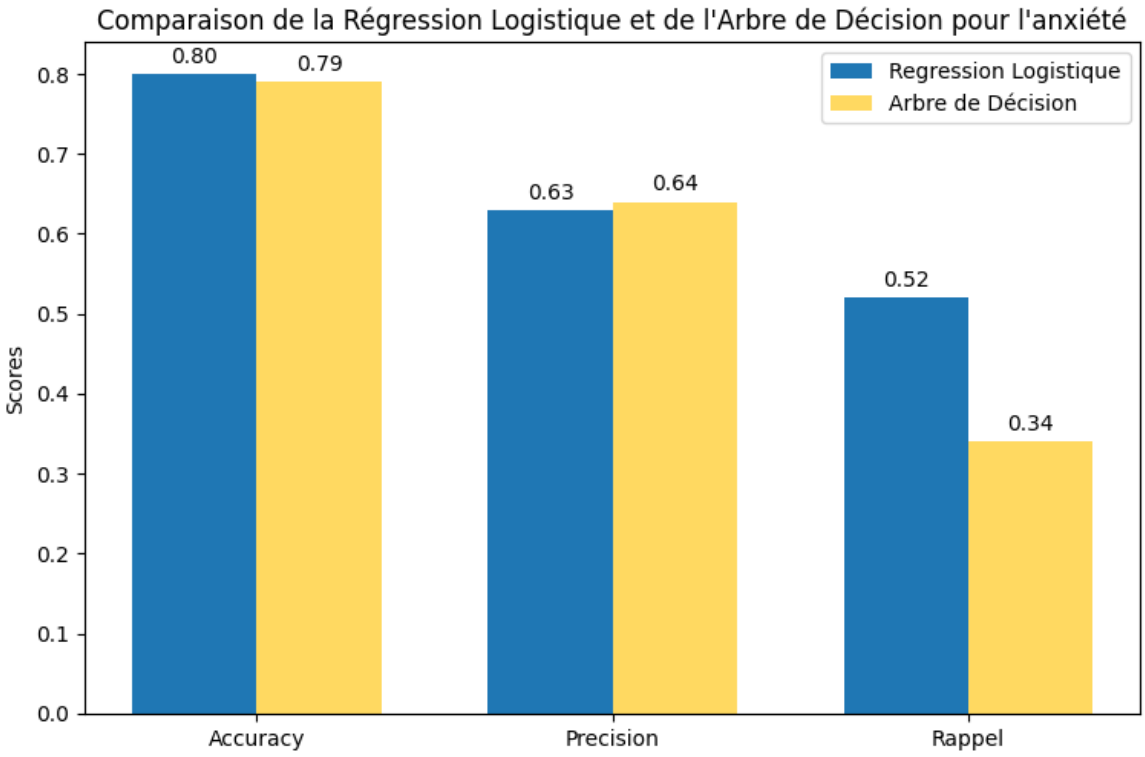
La régression linéaire révèle que notre variable cible, représentant la présence ou l'absence de troubles mentaux, est binaire, comme attendu. Le graphique illustre clairement que la variable cible peut prendre les valeurs 1 ou 0. Cependant, en termes d'explication et de résultats significatifs pour notre jeu de données, le graphique seul ne suffit pas.

En conclusion, la régression linéaire ne semble pas être le modèle approprié sur lequel nous devrions nous concentrer pour la suite de nos analyses.

Annexe 2 : comparaison des modèles pour l'insomnie



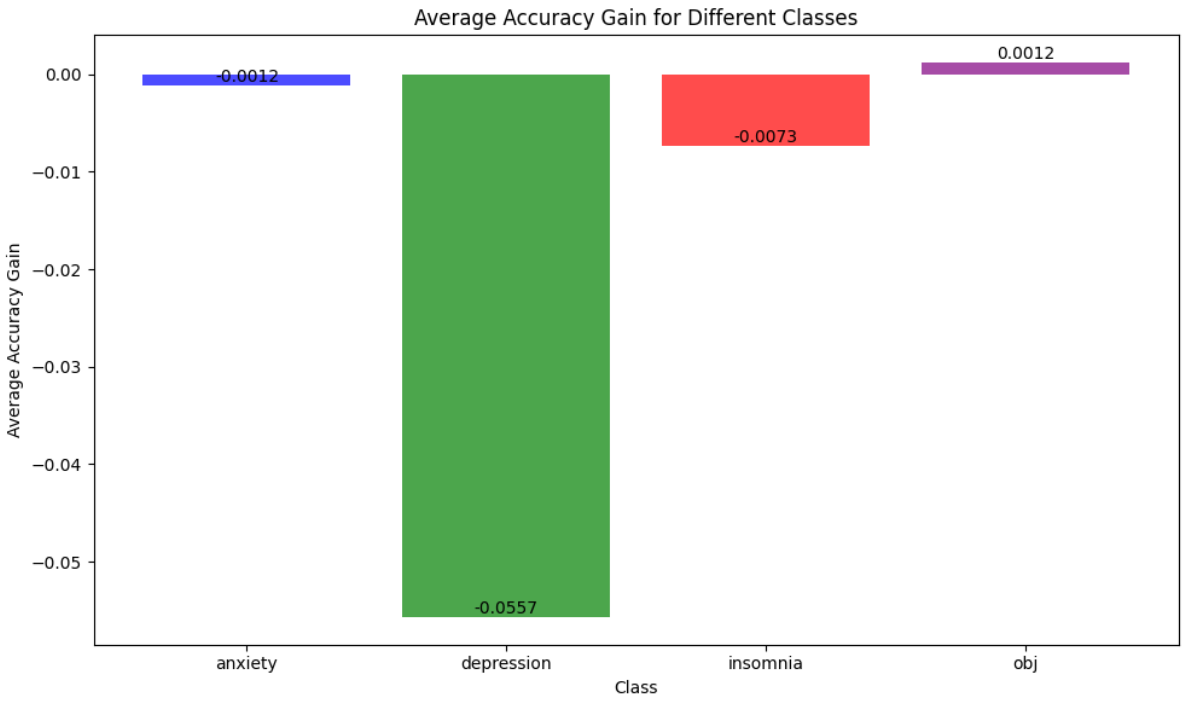
Annexe 3 : comparaison des modèles pour l'anxiété



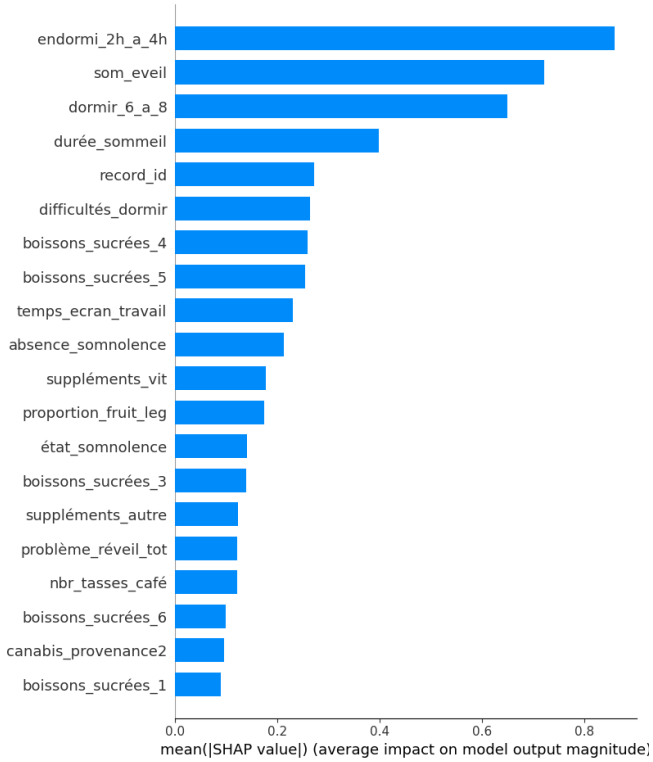
Annexe 4 : Tableau explication problème de sous-représentativités

obj	accuracy	precision	recall	f1	proportion	accuracy_si0
AvoirUnTrouble_1oui0non	0.74044266	0.6962963	0.8	0.74455446	49.80%	50.20%
Troubleanxieux	0.80482897	0.55555556	0.1980198	0.2919708	22.60%	77.40%
Epuisementprofessionnel	0.82696177	0.61111111	0.23404255	0.33846154	19.88%	80.12%
Depression	0.84909457	0.44827586	0.18055556	0.25742574	15.36%	84.64%
Dependance	0.92957746	0	0	0	7.80%	92.20%
TDH_A	0.93561368	0.375	0.1	0.15789474	7.24%	92.76%
Troublealimentaire	0.93561368	0.25	0.03333333	0.05882353	5.40%	94.60%
Troubleposttraumatique	0.93561368	0.44444444	0.12903226	0.2	4.88%	95.12%
TOC	0.94567404	0	0	0	4.68%	95.32%
Jeucompulsif	0.96780684	0	0	0	2.24%	97.76%
Troublebipolaire	0.97987928	0.33333333	0.11111111	0.16666667	1.52%	98.48%
Troublepersonnalite	0.97585513	0	0	0	1.52%	98.48%
Spectreautisme	0.98189135	0	0	0	1.36%	98.64%
Autre	0.97987928	0	0	0	1.32%	98.68%
Psychose	0.99798793	1	0.5	0.66666667	0.48%	99.52%
Schizophrenie	0.99798793	1	0.5	0.66666667	0.24%	99.76%

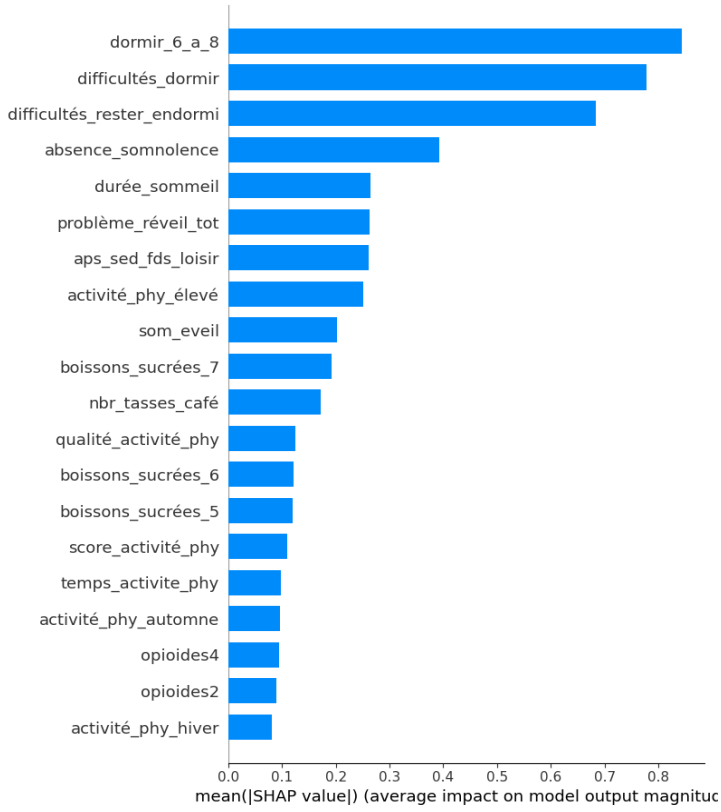
Annexe 5 : équilibrage des données



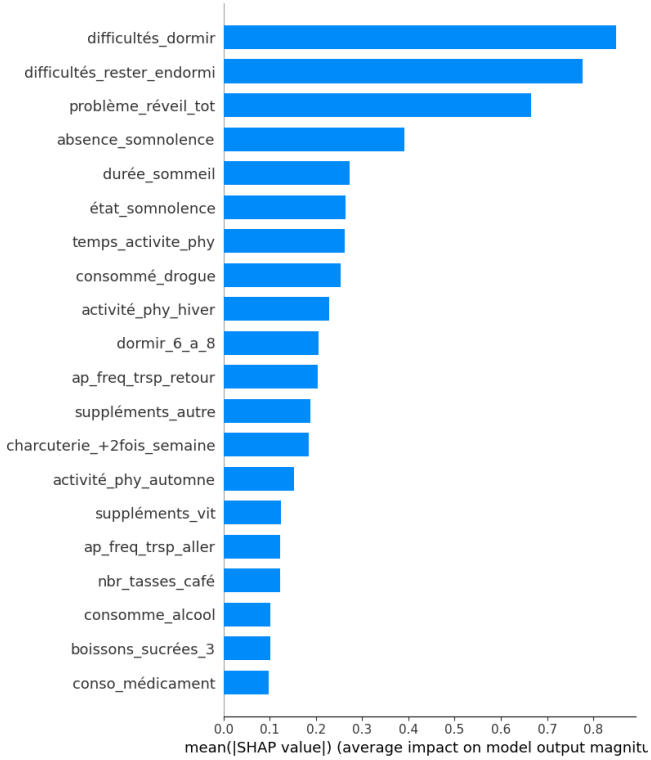
Annexe 6: Shap RFE Balanced Weight “Insomnie”



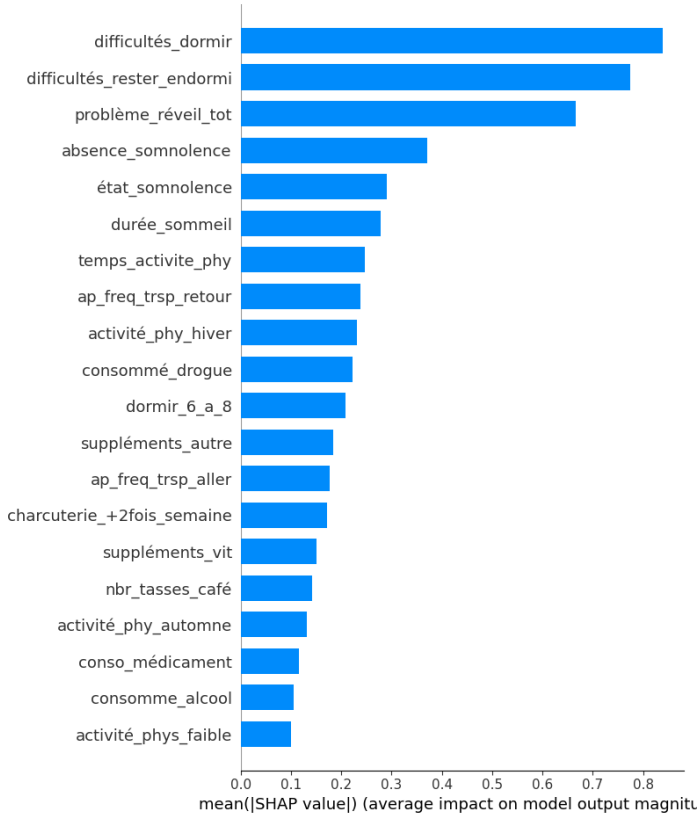
Annexe 7: Shap Lasso balanced weight “Insomnie”



Annexe 8: Shap None balanced weight “Insomnie”



Annexe 9: Shap Shap Balanced weight “insomnie”



Annexe 10 : Tableau contenant tous les résultats de nos analyses

Variable	Feature Selection	Class Balancing	Accuracy	Precision	Recall	ROC AUC	f1 score	database	methode
depression	RFE	None	0.88133333	0.58666667	0.43137255	0.69176652	0.49717514	ULA	LOGISTIQUE
depression	shap	None	0.87931034	0.6	0.16216216		0.25531915	ESSAIM	ARBRE
depression	None	None	0.87333333	0.53932584	0.47058824	0.70365832	0.5026178	ULA	LOGISTIQUE
depression	Lasso	None	0.87333333	0.54022989	0.46078431	0.69952796	0.4973545	ULA	LOGISTIQUE
depression	SHAP	None	0.872	0.53571429	0.44117647	0.69049564	0.48387097	ULA	LOGISTIQUE
depression	lasso	None	0.85714286	0.52941176	0.125		0.20224719	ULA	ARBRE
depression	rfe	None	0.84965035	0.55555556	0.11363636		0.18867925	ESSAIM	ARBRE
depression	None	None	0.84965035	0.55555556	0.11363636		0.18867925	ESSAIM	ARBRE
depression	None	None	0.84909457	0.44827586	0.18055556		0.25742574	ULA	ARBRE
depression	None	None	0.84792627	0.4375	0.22580645	0.58870968	0.29787234	ESSAIM	LOGISTIQUE
depression	RFE	None	0.84792627	0.42307692	0.17741935	0.56854839	0.25	ESSAIM	LOGISTIQUE
depression	shap	None	0.84708249	0.46153846	0.33333333		0.38709677	ULA	ARBRE
depression	lasso	None	0.84615385	0.5	0.11363636		0.18518519	ESSAIM	ARBRE
depression	Lasso	None	0.84562212	0.4137931	0.19354839	0.57392473	0.26373626	ESSAIM	LOGISTIQUE
depression	rfe	None	0.84305835	0.39285714	0.15277778		0.22	ULA	ARBRE
depression	SHAP	None	0.83870968	0.36666667	0.17741935	0.56317204	0.23913043	ESSAIM	LOGISTIQUE
insomnia	shap	smote	0.82758621	0.84768212	0.82580645		0.83660131	ESSAIM	ARBRE
insomnia	shap	class_weight	0.82758621	0.84768212	0.82580645		0.83660131	ESSAIM	ARBRE
insomnia	shap	None	0.82758621	0.84768212	0.82580645		0.83660131	ESSAIM	ARBRE
Epuisementprofessionnel	None	None	0.82696177	0.61111111	0.23404255		0.33846154	ULA	ARBRE
anxiety	Lasso	None	0.824	0.62096774	0.47530864	0.69768834	0.53846154	ULA	LOGISTIQUE
Epuisementprofessionnel	shap	None	0.82293763	0.57142857	0.25531915		0.35294118	ULA	ARBRE
depression	Lasso	Smote	0.82	0.40462428	0.68627451	0.76366195	0.50909091	ULA	LOGISTIQUE
anxiety	None	None	0.81866667	0.6015625	0.47530864	0.69428697	0.53103448	ULA	LOGISTIQUE
anxiety	SHAP	None	0.81866667	0.60655738	0.45679012	0.68757874	0.52112676	ULA	LOGISTIQUE
depression	SHAP	Smote	0.81866667	0.39880952	0.65686275	0.75049927	0.4962963	ULA	LOGISTIQUE

depression	None	Smote	0.816	0.3988764	0.69607843	0.76547749	0.50714286	ULA	LOGISTIQUE
obj	shap	smote	0.8137931	0.89189189	0.77647059		0.83018868	ESSAIM	ARBRE
anxiety	RFE	None	0.81333333	0.60185185	0.40123457	0.66405266	0.48148148	ULA	LOGISTIQUE
obj	shap	class_weight	0.81034483	0.88590604	0.77647059		0.82758621	ESSAIM	ARBRE
depression	lasso	None	0.80769231	0.34285714	0.27272727		0.30379747	ESSAIM	ARBRE
anxiety	None	None	0.80482897	0.55555556	0.1980198		0.2919708	ULA	ARBRE
anxiety	shap	None	0.80482897	0.55555556	0.1980198		0.2919708	ULA	ARBRE
insomnia	RFE	None	0.80414747	0.82300885	0.80519481	0.80407524	0.81400438	ESSAIM	LOGISTIQUE
obj	shap	None	0.80344828	0.85534591	0.8		0.82674772	ESSAIM	ARBRE
anxiety	lasso	None	0.8028169	0.54054054	0.1980198		0.28985507	ULA	ARBRE
depression	Lasso	Balanced_weight	0.80266667	0.37765957	0.69607843	0.75776144	0.48965517	ULA	LOGISTIQUE
insomnia	Lasso	None	0.80184332	0.8111588	0.81818182	0.80071652	0.81465517	ESSAIM	LOGISTIQUE
insomnia	RFE	Smote	0.80184332	0.82511211	0.7965368	0.80220928	0.81057269	ESSAIM	LOGISTIQUE
depression	SHAP	Balanced_weight	0.80133333	0.37823834	0.71568628	0.76525055	0.49491525	ULA	LOGISTIQUE
insomnia	lasso	class_weight	0.8006993	0.82644628	0.73529412		0.77821012	ESSAIM	ARBRE
insomnia	lasso	None	0.8006993	0.82644628	0.73529412		0.77821012	ESSAIM	ARBRE
depression	None	Balanced_weight	0.8	0.37755102	0.7254902	0.7686093	0.4966443	ULA	LOGISTIQUE
insomnia	RFE	Balanced_weight	0.79953917	0.82727273	0.78787879	0.80034333	0.80709534	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	rfe	None	0.79879276	0.35	0.07446809		0.12280702	ULA	ARBRE
anxiety	rfe	None	0.79678068	0.5	0.1980198		0.28368794	ULA	ARBRE
depression	RFE	Balanced_weight	0.796	0.37681159	0.76470588	0.7828159	0.50485437	ULA	LOGISTIQUE
depression	RFE	Smote	0.796	0.36923077	0.70588235	0.75803377	0.48484849	ULA	LOGISTIQUE
insomnia	SHAP	Balanced_weight	0.79493088	0.81981982	0.78787879	0.79541723	0.80353201	ESSAIM	LOGISTIQUE
anxiety	shap	class_weight	0.79310345	0.58139535	0.67567568		0.625	ESSAIM	ARBRE
insomnia	None	None	0.79262673	0.80257511	0.80952381	0.79146141	0.80603448	ESSAIM	LOGISTIQUE
insomnia	Lasso	Smote	0.79262673	0.81057269	0.7965368	0.79235707	0.80349345	ESSAIM	LOGISTIQUE
insomnia	Lasso	Balanced_weight	0.79032258	0.8125	0.78787879	0.79049112	0.8	ESSAIM	LOGISTIQUE
insomnia	None	None	0.79020979	0.796875	0.75		0.77272727	ESSAIM	ARBRE
Epuisementprofessionnel	lasso	None	0.78873239	0.39622642	0.22340426		0.28571429	ULA	ARBRE
insomnia	None	Balanced_weight	0.78801843	0.8061674	0.79220779	0.78772951	0.79912664	ESSAIM	LOGISTIQUE

insomnia	lasso	smote	0.78671329	0.80487805	0.72794118		0.76447876	ESSAIM	ARBRE
obj	None	smote	0.78671329	0.82089552	0.74829932		0.78291815	ESSAIM	ARBRE
anxiety	shap	None	0.7862069	0.63636364	0.37837838		0.47457627	ESSAIM	ARBRE
insomnia	None	Smote	0.78571429	0.80263158	0.79220779	0.78526646	0.79738562	ESSAIM	LOGISTIQUE
insomnia	SHAP	None	0.78571429	0.79487179	0.80519481	0.7843708	0.8	ESSAIM	LOGISTIQUE
insomnia	SHAP	Smote	0.78571429	0.80530973	0.78787879	0.78556501	0.79649891	ESSAIM	LOGISTIQUE
obj	rfe	smote	0.78321678	0.81481482	0.74829932		0.78014184	ESSAIM	ARBRE
insomnia	None	class_weight	0.78321678	0.7890625	0.74264706		0.76515152	ESSAIM	ARBRE
anxiety	SHAP	None	0.78110599	0.6056338	0.39090909	0.65224467	0.47513812	ESSAIM	LOGISTIQUE
obj	RFE	Smote	0.78110599	0.80916031	0.82490272	0.77120842	0.81695568	ESSAIM	LOGISTIQUE
obj	SHAP	None	0.78110599	0.79136691	0.85603113	0.76417376	0.82242991	ESSAIM	LOGISTIQUE
depression	lasso	smote	0.77972028	0.33333333	0.43181818		0.37623762	ESSAIM	ARBRE
obj	lasso	smote	0.77972028	0.80882353	0.74829932		0.77738516	ESSAIM	ARBRE
insomnia	lasso	smote	0.77972028	0.80672269	0.70588235		0.75294118	ESSAIM	ARBRE
insomnia	rfe	None	0.77972028	0.792	0.72794118		0.75862069	ESSAIM	ARBRE
depression	shap	smote	0.77931034	0.32911392	0.7027027		0.44827586	ESSAIM	ARBRE
depression	rfe	smote	0.77665996	0.36170213	0.70833333		0.47887324	ULA	ARBRE
anxiety	None	None	0.7764977	0.5890411	0.39090909	0.64915825	0.46994536	ESSAIM	LOGISTIQUE
anxiety	Lasso	None	0.7764977	0.5915493	0.38181818	0.646156	0.4640884	ESSAIM	LOGISTIQUE
obj	RFE	None	0.7764977	0.78776978	0.85214008	0.75940337	0.81869159	ESSAIM	LOGISTIQUE
obj	lasso	smote	0.77622378	0.80291971	0.74829932		0.77464789	ESSAIM	ARBRE
insomnia	lasso	None	0.77622378	0.79032258	0.72058824		0.75384615	ESSAIM	ARBRE
obj	None	None	0.77419355	0.78494624	0.85214008	0.75657851	0.81716418	ESSAIM	LOGISTIQUE
obj	SHAP	Smote	0.77419355	0.81176471	0.80544747	0.76713052	0.80859375	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	None	None	0.77333333	0.45918367	0.27777778	0.59382086	0.34615385	ULA	LOGISTIQUE
Epuisementprofessionnel	Lasso	None	0.77333333	0.45918367	0.27777778	0.59382086	0.34615385	ULA	LOGISTIQUE
Epuisementprofessionnel	RFE	None	0.77333333	0.45	0.22222222	0.57369615	0.29752066	ULA	LOGISTIQUE
insomnia	lasso	class_weight	0.77272727	0.784	0.72058824		0.75095785	ESSAIM	ARBRE
insomnia	rfe	smote	0.77272727	0.79831933	0.69852941		0.74509804	ESSAIM	ARBRE
insomnia	rfe	class_weight	0.77272727	0.784	0.72058824		0.75095785	ESSAIM	ARBRE

insomnia	None	smote	0.77272727	0.79831933	0.69852941		0.74509804	ESSAIM	ARBRE
obj	None	Smote	0.7718894	0.80859375	0.80544747	0.76430566	0.80701754	ESSAIM	LOGISTIQUE
obj	RFE	Balanced_weight	0.7718894	0.81349206	0.79766537	0.76606432	0.80550098	ESSAIM	LOGISTIQUE
obj	Lasso	None	0.76958525	0.77935943	0.85214008	0.7509288	0.81412639	ESSAIM	LOGISTIQUE
obj	SHAP	Balanced_weight	0.76958525	0.81526104	0.78988327	0.76499813	0.80237154	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	SHAP	None	0.76933333	0.44554455	0.27777778	0.59126984	0.34220532	ULA	LOGISTIQUE
depression	lasso	smote	0.76923077	0.30357143	0.38636364		0.34	ESSAIM	ARBRE
depression	rfe	smote	0.76923077	0.30357143	0.38636364		0.34	ESSAIM	ARBRE
depression	None	smote	0.76923077	0.30357143	0.38636364		0.34	ESSAIM	ARBRE
obj	None	Balanced_weight	0.76728111	0.812	0.78988327	0.76217327	0.80078895	ESSAIM	LOGISTIQUE
anxiety	lasso	smote	0.7665996	0.43902439	0.53465347		0.48214286	ULA	ARBRE
depression	None	smote	0.7665996	0.35897436	0.77777778		0.49122807	ULA	ARBRE
depression	shap	smote	0.7665996	0.35897436	0.77777778		0.49122807	ULA	ARBRE
anxiety	lasso	None	0.76573427	0.49056604	0.39393939		0.43697479	ESSAIM	ARBRE
anxiety	rfe	smote	0.76573427	0.49275362	0.51515152		0.5037037	ESSAIM	ARBRE
anxiety	None	None	0.76573427	0.49056604	0.39393939		0.43697479	ESSAIM	ARBRE
anxiety	RFE	None	0.76497696	0.55882353	0.34545455	0.62643098	0.42696629	ESSAIM	LOGISTIQUE
depression	SHAP	Balanced_weight	0.76267281	0.33333333	0.66129032	0.72043011	0.44324324	ESSAIM	LOGISTIQUE
obj	Lasso	Smote	0.76036866	0.8	0.79377432	0.75281936	0.796875	ESSAIM	LOGISTIQUE
obj	Lasso	Balanced_weight	0.76036866	0.80237154	0.78988327	0.7536987	0.79607843	ESSAIM	LOGISTIQUE
obj	RFE	None	0.75866667	0.77653631	0.73350924	0.75893791	0.75440977	ULA	LOGISTIQUE
obj	RFE	Balanced_weight	0.75866667	0.77966102	0.72823219	0.7589948	0.75306958	ULA	LOGISTIQUE
obj	RFE	Smote	0.756	0.77683616	0.72559367	0.75632783	0.75034106	ULA	LOGISTIQUE
depression	RFE	Balanced_weight	0.75576037	0.31967213	0.62903226	0.70295699	0.42391304	ESSAIM	LOGISTIQUE
anxiety	Lasso	Smote	0.75466667	0.456	0.7037037	0.73620559	0.55339806	ULA	LOGISTIQUE
anxiety	rfe	smote	0.75452716	0.41860465	0.53465347		0.46956522	ULA	ARBRE
anxiety	RFE	Balanced_weight	0.75333333	0.45724907	0.75925926	0.75547997	0.57076566	ULA	LOGISTIQUE
obj	Lasso	None	0.752	0.77337111	0.72031662	0.7523416	0.74590164	ULA	LOGISTIQUE
anxiety	SHAP	Balanced_weight	0.752	0.453125	0.71604938	0.73897707	0.55502392	ULA	LOGISTIQUE
obj	lasso	None	0.75174825	0.81147541	0.67346939		0.73605948	ESSAIM	ARBRE

depression	None	Balanced_weight	0.75115207	0.30508475	0.58064516	0.68010753	0.4	ESSAIM	LOGISTIQUE
depression	Lasso	Balanced_weight	0.75115207	0.30833333	0.59677419	0.68682796	0.40659341	ESSAIM	LOGISTIQUE
obj	None	Smote	0.75066667	0.77118644	0.72031662	0.75099389	0.74488404	ULA	LOGISTIQUE
Epuisementprofessionnel	rfe	class_weight	0.75050302	0.4025974	0.65957447		0.5	ULA	ARBRE
obj	None	None	0.74933333	0.77053824	0.7176781	0.74967463	0.7431694	ULA	LOGISTIQUE
obj	None	Balanced_weight	0.74933333	0.77053824	0.7176781	0.74967463	0.7431694	ULA	LOGISTIQUE
obj	Lasso	Smote	0.74933333	0.77207977	0.71503958	0.74970308	0.74246575	ULA	LOGISTIQUE
obj	Lasso	Balanced_weight	0.74933333	0.77207977	0.71503958	0.74970308	0.74246575	ULA	LOGISTIQUE
anxiety	None	Smote	0.74933333	0.45038168	0.72839506	0.74174855	0.55660377	ULA	LOGISTIQUE
anxiety	None	smote	0.74849095	0.40322581	0.4950495		0.44444444	ULA	ARBRE
depression	None	class_weight	0.74849095	0.33333333	0.73611111		0.45887446	ULA	ARBRE
anxiety	shap	smote	0.74849095	0.40322581	0.4950495		0.44444444	ULA	ARBRE
depression	shap	class_weight	0.74849095	0.33333333	0.73611111		0.45887446	ULA	ARBRE
obj	lasso	class_weight	0.74825175	0.85046729	0.61904762		0.71653543	ESSAIM	ARBRE
obj	lasso	None	0.74825175	0.80991736	0.66666667		0.73134328	ESSAIM	ARBRE
anxiety	lasso	None	0.74825175	0.4516129	0.42424242		0.4375	ESSAIM	ARBRE
obj	rfe	None	0.74825175	0.80991736	0.66666667		0.73134328	ESSAIM	ARBRE
anxiety	rfe	None	0.74825175	0.4516129	0.42424242		0.4375	ESSAIM	ARBRE
obj	None	None	0.74825175	0.80991736	0.66666667		0.73134328	ESSAIM	ARBRE
anxiety	None	Balanced_weight	0.748	0.44827586	0.72222222	0.73866213	0.55319149	ULA	LOGISTIQUE
anxiety	RFE	Smote	0.748	0.44827586	0.72222222	0.73866213	0.55319149	ULA	LOGISTIQUE
obj	SHAP	None	0.74666667	0.76770538	0.71503958	0.74700766	0.74043716	ULA	LOGISTIQUE
anxiety	SHAP	Smote	0.74654378	0.5	0.69090909	0.72817059	0.58015267	ESSAIM	LOGISTIQUE
obj	SHAP	Smote	0.74533333	0.76704546	0.71240106	0.7456884	0.73871409	ULA	LOGISTIQUE
obj	SHAP	Balanced_weight	0.74533333	0.76704546	0.71240106	0.7456884	0.73871409	ULA	LOGISTIQUE
Epuisementprofessionnel	lasso	smote	0.7444668	0.38926174	0.61702128		0.47736626	ULA	ARBRE
obj	rfe	class_weight	0.7444668	0.70454545	0.79148936		0.74549098	ULA	ARBRE
depression	SHAP	Smote	0.74423963	0.30081301	0.59677419	0.6827957	0.4	ESSAIM	LOGISTIQUE
anxiety	Lasso	Balanced_weight	0.744	0.44274809	0.71604938	0.73387503	0.54716981	ULA	LOGISTIQUE
anxiety	SHAP	Smote	0.744	0.44230769	0.70987654	0.73163895	0.5450237	ULA	LOGISTIQUE

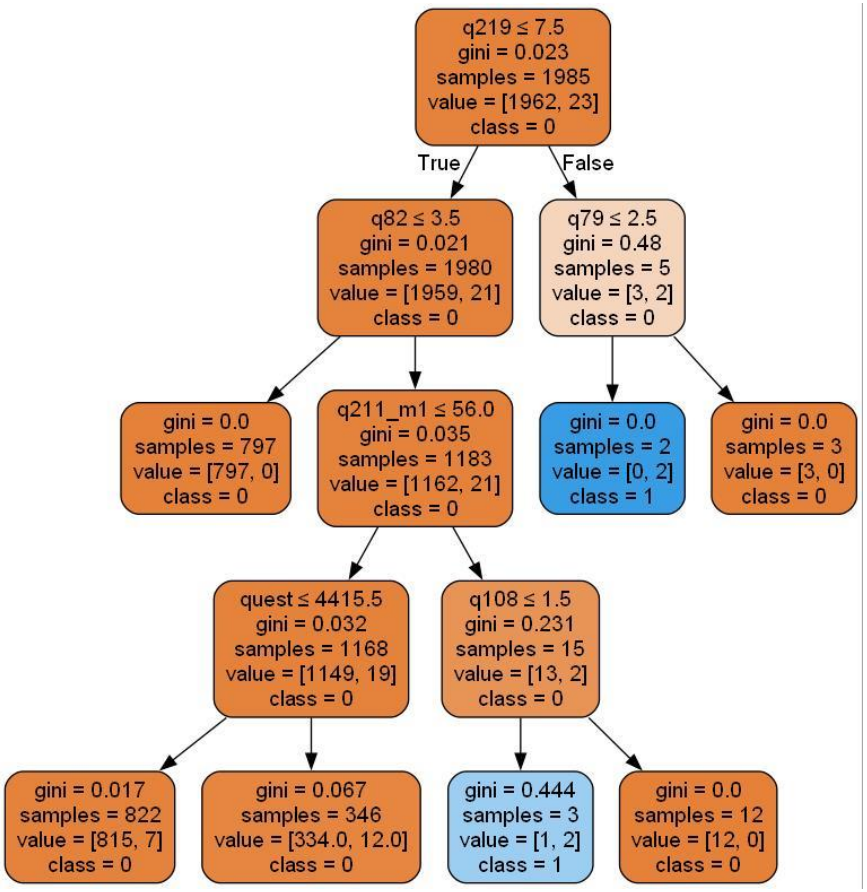
obj	rfe	smote	0.74245473	0.69741697	0.80425532		0.74703557	ULA	ARBRE
obj	rfe	None	0.74245473	0.69741697	0.80425532		0.74703557	ULA	ARBRE
obj	None	smote	0.74044266	0.6962963	0.8		0.74455446	ULA	ARBRE
obj	None	class_weight	0.74044266	0.6962963	0.8		0.74455446	ULA	ARBRE
obj	None	None	0.74044266	0.6962963	0.8		0.74455446	ULA	ARBRE
obj	shap	smote	0.74044266	0.69776119	0.79574468		0.74353877	ULA	ARBRE
obj	shap	class_weight	0.74044266	0.6962963	0.8		0.74455446	ULA	ARBRE
obj	shap	None	0.74044266	0.6962963	0.8		0.74455446	ULA	ARBRE
obj	lasso	smote	0.73843058	0.69516729	0.79574468		0.74206349	ULA	ARBRE
obj	lasso	class_weight	0.73843058	0.69516729	0.79574468		0.74206349	ULA	ARBRE
obj	lasso	None	0.73843058	0.69516729	0.79574468		0.74206349	ULA	ARBRE
anxiety	lasso	class_weight	0.73776224	0.44444444	0.54545455		0.48979592	ESSAIM	ARBRE
anxiety	lasso	class_weight	0.73776224	0.44444444	0.54545455		0.48979592	ESSAIM	ARBRE
anxiety	rfe	class_weight	0.73776224	0.44444444	0.54545455		0.48979592	ESSAIM	ARBRE
anxiety	None	class_weight	0.73776224	0.44444444	0.54545455		0.48979592	ESSAIM	ARBRE
depression	Lasso	Smote	0.73732719	0.29365079	0.59677419	0.67876344	0.39361702	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	None	smote	0.73641851	0.35433071	0.4787234		0.40723982	ULA	ARBRE
Epuisementprofessionnel	shap	smote	0.73641851	0.35433071	0.4787234		0.40723982	ULA	ARBRE
depression	None	Smote	0.73502304	0.29133858	0.59677419	0.67741935	0.39153439	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	rfe	smote	0.73440644	0.3515625	0.4787234		0.40540541	ULA	ARBRE
anxiety	lasso	smote	0.73426573	0.4375	0.53030303		0.47945206	ESSAIM	ARBRE
anxiety	None	smote	0.73426573	0.42647059	0.43939394		0.43283582	ESSAIM	ARBRE
anxiety	Lasso	Smote	0.73271889	0.48076923	0.68181818	0.71590909	0.56390977	ESSAIM	LOGISTIQUE
anxiety	RFE	Smote	0.73271889	0.48051948	0.67272727	0.71290685	0.56060606	ESSAIM	LOGISTIQUE
anxiety	RFE	Balanced_weight	0.73271889	0.48051948	0.67272727	0.71290685	0.56060606	ESSAIM	LOGISTIQUE
depression	rfe	class_weight	0.73239437	0.31952663	0.75		0.44813278	ULA	ARBRE
Epuisementprofessionnel	None	class_weight	0.73038229	0.37951807	0.67021277		0.48461538	ULA	ARBRE
Epuisementprofessionnel	shap	class_weight	0.73038229	0.37951807	0.67021277		0.48461538	ULA	ARBRE
depression	RFE	Smote	0.72580645	0.28244275	0.59677419	0.67204301	0.38341969	ESSAIM	LOGISTIQUE
anxiety	None	Balanced_weight	0.72580645	0.47169811	0.68181818	0.71127946	0.55762082	ESSAIM	LOGISTIQUE

anxiety	Lasso	Balanced_weight	0.72580645	0.47133758	0.67272727	0.70827722	0.55430712	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	Lasso	Smote	0.72266667	0.408	0.62962963	0.68896448	0.49514563	ULA	LOGISTIQUE
anxiety	SHAP	Balanced_weight	0.72119816	0.46540881	0.67272727	0.7051908	0.55018587	ESSAIM	LOGISTIQUE
anxiety	lasso	smote	0.72027972	0.40277778	0.43939394		0.42028986	ESSAIM	ARBRE
obj	lasso	class_weight	0.72027972	0.80733945	0.59863946		0.6875	ESSAIM	ARBRE
obj	rfe	class_weight	0.72027972	0.80733945	0.59863946		0.6875	ESSAIM	ARBRE
obj	None	class_weight	0.72027972	0.80733945	0.59863946		0.6875	ESSAIM	ARBRE
Epuisementprofessionnel	Lasso	Balanced_weight	0.72	0.40322581	0.61728395	0.68279164	0.48780488	ULA	LOGISTIQUE
Epuisementprofessionnel	SHAP	Balanced_weight	0.71866667	0.40466926	0.64197531	0.69088561	0.49642005	ULA	LOGISTIQUE
Epuisementprofessionnel	None	Balanced_weight	0.71866667	0.4	0.60493827	0.67746914	0.48157248	ULA	LOGISTIQUE
depression	lasso	smote	0.71830986	0.28205128	0.61111111		0.38596491	ULA	ARBRE
Epuisementprofessionnel	None	Smote	0.71733333	0.39754098	0.59876543	0.67438272	0.47783251	ULA	LOGISTIQUE
anxiety	None	Smote	0.71198157	0.45398773	0.67272727	0.69901796	0.54212454	ESSAIM	LOGISTIQUE
Epuisementprofessionnel	RFE	Smote	0.71066667	0.39130435	0.61111111	0.67460318	0.47710843	ULA	LOGISTIQUE
Epuisementprofessionnel	RFE	Balanced_weight	0.70933333	0.39147287	0.62345679	0.67822499	0.48095238	ULA	LOGISTIQUE
Epuisementprofessionnel	SHAP	Smote	0.70933333	0.388	0.59876543	0.66928068	0.47087379	ULA	LOGISTIQUE
depression	lasso	class_weight	0.7082495	0.30687831	0.80555556		0.44444444	ULA	ARBRE
anxiety	shap	smote	0.70689655	0.44329897	0.58108108		0.50292398	ESSAIM	ARBRE
Epuisementprofessionnel	lasso	class_weight	0.70422535	0.34857143	0.64893617		0.4535316	ULA	ARBRE
depression	shap	class_weight	0.67931034	0.24545455	0.72972973		0.36734694	ESSAIM	ARBRE
anxiety	shap	class_weight	0.67605634	0.36842105	0.83168317		0.5106383	ULA	ARBRE
anxiety	rfe	class_weight	0.67404427	0.36681223	0.83168317		0.50909091	ULA	ARBRE
depression	None	class_weight	0.66433566	0.25471698	0.61363636		0.36	ESSAIM	ARBRE
depression	lasso	class_weight	0.66083916	0.25233645	0.61363636		0.35761589	ESSAIM	ARBRE
anxiety	None	class_weight	0.65794769	0.35564854	0.84158416		0.5	ULA	ARBRE
depression	rfe	class_weight	0.65734266	0.25	0.61363636		0.35526316	ESSAIM	ARBRE
depression	lasso	class_weight	0.63986014	0.24347826	0.63636364		0.35220126	ESSAIM	ARBRE
anxiety	lasso	class_weight	0.63782696	0.348659	0.9009901		0.50276243	ULA	ARBRE

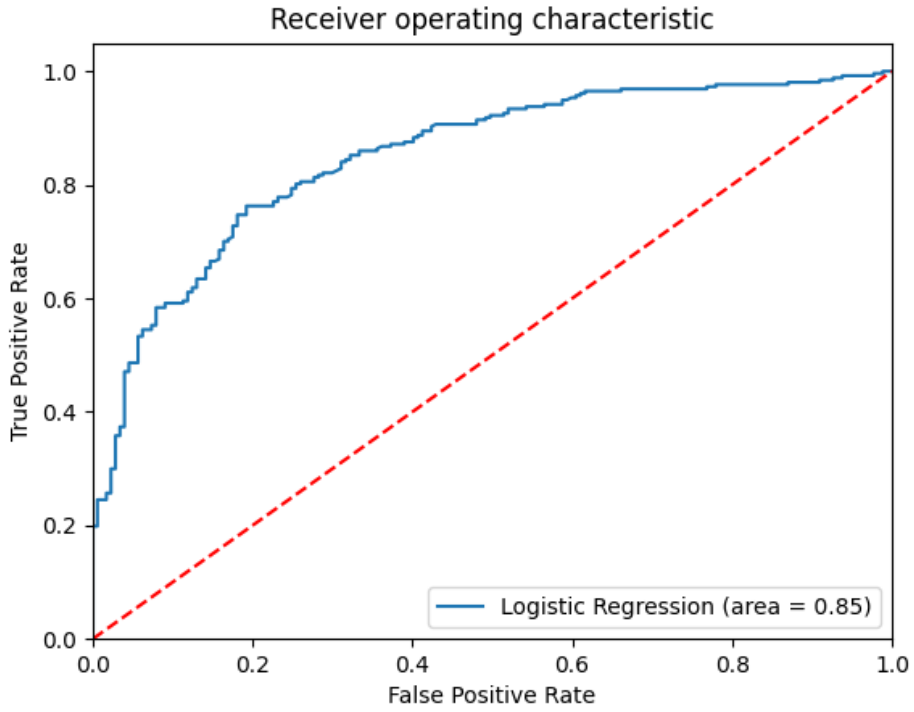
Annexe 11 : tableau performance réseaux de neuronnes

Variables	Accuracy	Precision	Recall	F1-score
obj	0.76573427	0.74691358	0.82312925	0.78317152
depression	0.82517483	0.44	0.5	0.46808511
anxiety	0.80769231	0.60377358	0.48484848	0.53781513
insomnia	0.77972028	0.74496644	0.81617647	0.77894737
Depression	0.83299799	0.41538462	0.375	0.39416058
Epuisementprofessionnel	0.83702213	0.63829787	0.31914894	0.42553191
Troubleanxieux	0.79678068	0.5	0.33663366	0.40236686
AvoirUnTrouble_1oui0non	0.74647887	0.72614108	0.74468085	0.73529412

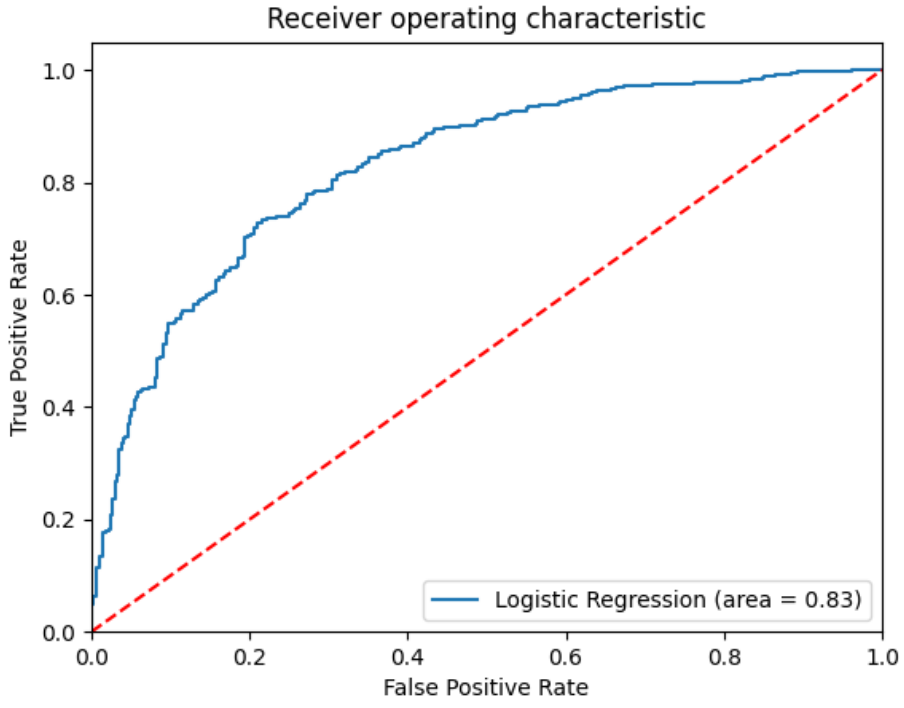
Annexe 12 : exemple arbre de décision



Annexe 13 : Courbe Roc « OBJ » (database Essaim)



Annexe 14 : Courbe Roc « Avoir un Trouble oui ou non » (database ULA)



Annexe 15 : Revue de littérature non utilisé mais intéressante

Utilisation IA dans la santé mentale

L'utilisation de l'intelligence artificielle et du machine learning peuvent transformer le domaine de la santé mentale. Ces technologies permettent de détecter et de traiter les troubles mentaux de manière plus efficace et personnalisée. Voici les applications de ces technologies qui sont ressorti pour améliorer la santé mentale :(Shimada, 2023)

1. Identification et Prévention Précoce : En analysant des données variées comme les publications sur les réseaux sociaux, les comportements en ligne et les mesures physiologiques, les systèmes de machine learning peuvent détecter des signes précoces de troubles mentaux, permettant ainsi une intervention rapide.(Shimada, 2023)

2. Interventions Thérapeutiques Accessibles : Les chatbots et autres systèmes basés sur l'IA peuvent fournir une assistance thérapeutique 24h/24, aidant les personnes en détresse ou sur liste d'attente à accéder à des soins psychologiques en temps réel.(Shimada, 2023)

3. Prévention du Suicide : Les algorithmes de machine learning peuvent analyser des données de diverses sources (réseaux sociaux, appels téléphoniques, chats en ligne) pour identifier les personnes à risque de suicide et permettre une intervention proactive.(Shimada, 2023)

4. Plans de Traitement Personnalisés : En utilisant des données cliniques, génétiques et comportementales, l'IA aide à développer des plans de traitement personnalisés, améliorant ainsi l'engagement et l'efficacité des soins pour les patients.(Shimada, 2023)

5. Considérations Éthiques : L'article souligne l'importance de traiter les questions éthiques liées à l'utilisation de l'IA en santé mentale, telles que la confidentialité, les biais dans les diagnostics, et l'importance de maintenir un contact humain empathique dans les soins.(Shimada, 2023)

6. Assistance Supplémentaire aux Praticiens : L'IA peut compléter les approches thérapeutiques traditionnelles en fournissant des informations précieuses aux praticiens de la santé mentale, basées sur l'analyse des schémas comportementaux des patients, améliorant ainsi les résultats thérapeutiques.(Darzi, 2023)

7. Réduction de la Stigmatisation : L'IA, en tant que fournisseur impartial de soutien, offre un environnement sécurisé pour que les individus puissent exprimer leurs pensées et émotions sans crainte de jugement ou de désapprobation sociale.(Darzi, 2023)

8. Surveillance Continue : Les dispositifs portables équipés de capacités IA peuvent détecter des changements physiologiques ou des signes avant-coureurs et notifier instantanément les utilisateurs ou les praticiens de santé, permettant une intervention proactive.(Darzi, 2023)

9. Précision Diagnostique : Grâce à une approche basée sur les données, l'IA peut améliorer la précision des diagnostics de santé mentale en analysant un large éventail de déterminants biologiques, psychologiques et sociaux, facilitant ainsi des stratégies de traitement fondées sur des preuves.(Darzi, 2023)

L'intégration de l'IA dans les soins de santé mentale promet d'améliorer l'accessibilité, la personnalisation des soins et la surveillance continue, tout en soulevant des considérations éthiques importantes. Il est crucial de maintenir un équilibre entre la thérapie assistée par l'IA et le soutien thérapeutique conventionnel pour garantir des résultats optimaux en matière de santé mentale.(Darzi, 2023; Shimada, 2023)

Applications en Santé Mentale sélection de caractéristique

Dans le domaine de la santé mentale, la sélection de caractéristiques est vitale pour la réduction de la dimensionnalité des données complexes et variées. Elle permet l'identification de biomarqueurs et améliore les stratégies de traitement personnalisé. Par exemple, des études ont montré que des méthodes de sélection de caractéristiques peuvent identifier des variables clés associées aux troubles dépressifs et anxieux, aidant ainsi à développer des interventions plus ciblées.(Dwyer et al., 2018)

1. Identification de Biomarqueurs :

La sélection de caractéristiques a permis d'identifier des biomarqueurs potentiels dans le cerveau lié à la dépression, grâce à l'analyse des images IRM fonctionnelles (fMRI) .(Korgaonkar et al., 2014)

2. Traitement Personnalisé :

En sélectionnant les caractéristiques les plus pertinentes parmi les données cliniques et génétiques, les chercheurs ont pu améliorer les stratégies de traitement pour les patients souffrant de troubles anxieux, en personnalisant les interventions thérapeutiques.(Stein et al., 2010)

3. Analyse des Comportements :

Les données de capteurs et d'applications mobiles ont été utilisées pour surveiller les comportements des patients en temps réel, où la sélection de caractéristiques a aidé à extraire les indicateurs les plus pertinents pour la détection précoce des épisodes dépressifs.(Saeb et al., 2015)

Utilisation en Santé Mentale Regression logistique

La régression logistique est particulièrement utile dans le domaine de la santé mentale pour prédire des issues binaires telles que la présence ou l'absence d'un trouble mental. Elle permet d'évaluer l'impact de plusieurs variables explicatives sur la probabilité d'un trouble, ce qui aide à mieux cibler les interventions. Vous trouverez ci-dessous plusieurs exemples.

Étude sur la Dépression

Une étude récente a utilisé la régression logistique pour analyser les facteurs de risque associés à la dépression chez les adultes. L'étude a collecté des données sur divers facteurs sociodémographiques, cliniques et de style de vie, y compris l'âge, le sexe, les antécédents familiaux de dépression, le niveau de stress perçu, et les habitudes de sommeil.

Les résultats de cette étude ont montré que plusieurs variables étaient significativement associées à un risque accru de dépression. Par exemple, un niveau de stress perçu élevé et des antécédents familiaux de dépression ont été identifiés comme des prédicteurs majeurs. La régression logistique a permis de quantifier l'impact de chaque facteur, fournissant des coefficients qui indiquent l'augmentation de la probabilité de dépression associée à chaque variable.(Lin et al., 2021)

Etude sur l’insomnie

Une étude réalisée en Indonésie a examiné l'association entre le jeu vidéo avant le coucher et l'insomnie chez les élèves de collège, en mettant un accent particulier sur les différences entre les genres. La régression logistique a permis de mettre en lumière des différences importantes entre les genres quant à l'impact du jeu vidéo sur l'insomnie. Ces résultats suggèrent que les interventions de santé publique visant à réduire les problèmes de sommeil chez les adolescents devraient tenir compte des différences de genre. Les filles semblent plus vulnérables aux effets négatifs des jeux vidéo sur le sommeil, nécessitant des approches ciblées pour minimiser les risques d'insomnie.(Gandaputra et al., 2021)

Impact des facteurs socio-démographiques

Dans l'étude COMPASS menée au Canada, qui a collecté des données auprès de plus de 74 000 élèves de 136 écoles, la régression logistique a été utilisée pour analyser l'impact de diverses variables sociodémographiques et comportementales sur les niveaux d'anxiété et de dépression. Les résultats ont montré que des variables telles que le sentiment d'appartenance à l'école et le soutien familial étaient des prédicteurs significatifs de la santé mentale des élèves (Battista et al., 2023)

Ces différents exemples permettent de mettre en avant certains avantages de la régression logistique.

1. Interprétation des Résultats : Chaque variable indépendante peut être interprétée quant à son effet sur la probabilité de l'événement, ce qui permet une compréhension claire des facteurs influençant le résultat.

2. Gestion des Variables Multicatégorielles : La régression logistique peut gérer à la fois des variables continues et catégorielles, permettant ainsi d'analyser des ensembles de données complexes.

3. Estimation des Probabilités : Elle permet d'estimer directement les probabilités d’occurrence d'un événement, facilitant ainsi la prise de décisions basées sur ces probabilités.

4. Évaluation de l'Impact des Variables : La régression logistique permet d'évaluer l'importance relative de chaque variable explicative, aidant ainsi à identifier les facteurs de risque clés.

En conclusion, la régression logistique est un outil puissant pour la prédiction des résultats binaires dans le domaine de la santé mentale. Son utilisation permet d'identifier et de quantifier l'impact des différents facteurs de risque, facilitant ainsi l'intervention précoce et ciblée. En raison de sa capacité à gérer des données complexes et à fournir des probabilités claires, elle reste une méthode de choix pour de nombreuses études en épidémiologie et en santé publique.

Arbre de décision appliqué

Pendant notre recherche nous avons découvert que les arbres de décision peuvent être utilisés de manière efficace pour prédire les problèmes de santé en analysant des ensembles de données complexes comprenant diverses variables. Ces variables peuvent inclure des aspects comportementaux, biologiques et environnementaux qui influencent la santé mentale. Les arbres de décisions sont utilisés aussi bien dans le domaine de la santé que dans celui de la santé mentale et plusieurs exemples vont être donné.

Étude sur la Dépression et l'Insomnie : Une étude a utilisé des arbres de décision pour analyser les données de patients souffrant d'insomnie et a identifié des facteurs clés prédictifs de la dépression. Les résultats ont montré que la combinaison d'une mauvaise qualité de sommeil et de facteurs de stress élevés augmentait considérablement le risque de dépression. Cependant l'arbre de décision est minoritaire dans l'études qui utilise beaucoup le test du chi2(Scott et al., 2021).

Analyse des Troubles de l'Alimentation : Un autre cas d'utilisation impliquait l'utilisation d'arbres de décision pour prédire les troubles de l'alimentation chez les adolescents. Les variables analysées comprenaient les habitudes alimentaires, les niveaux d'activité physique et les interactions sociales. Les arbres de décision ont permis d'identifier des sous-groupes à haut risque, facilitant ainsi une intervention précoce (Lauer & Krieg, 2004).

Prédiction du Syndrome de Stress Post-Traumatique (SSPT) : Les arbres de décision ont été utilisés pour prédire le SSPT chez les vétérans de guerre. En analysant des données telles que l'exposition à des événements traumatiques, la qualité du sommeil et les antécédents de santé mentale, les modèles ont pu prédire avec précision les individus à risque de développer le SSPT(Harvey et al., 2003).

Arrêt cardiaque : Au Centre médical de l'Université de Californie à San Diego, les patients admis pour une crise cardiaque se voient mesurer 19 variables au cours des premières 24 heures, incluant la pression artérielle et l'âge. L'objectif de l'étude était d'identifier les patients à haut risque, c'est-à-dire ceux susceptibles de ne pas survivre au-delà de 30 jours, en se basant sur ces données initiales. L'arbre de décision créé à partir de cette étude classifiait les patients selon des réponses à trois questions clés sur la pression artérielle, l'âge et la présence de tachycardie sinusale. Les résultats ont montré que cette méthode pouvait produire des prédictions précises tout en restant relativement simple et facile à interpréter(Praagman, 1985).

En plus de ces 4 cas une étude a particulièrement retenu notre attention. L'étude COMPASS, menée au Canada, a recueilli des données auprès de plus de 74 000 élèves de 136 écoles, incluant des mesures de l'anxiété, de la dépression et du bien-être psychologique, ainsi que 23 variables sociodémographiques et comportementales. Les arbres de décision, notamment les modèles CART (Classification and Regression Trees) et CTREE (Conditional Inference Trees), ont permis de segmenter la population en sous-groupes homogènes en fonction des interactions complexes entre les variables, identifiant ainsi les facteurs de risque majeurs et les sous-groupes à haut risque(Battista et al., 2023).

L'utilisation des arbres de décision dans cette étude a démontré plusieurs avantages par rapport aux méthodes de régression classiques. Les arbres de décision ont non seulement permis d'identifier les mêmes facteurs de prédiction importants que les modèles de régression, mais ont également accordé une plus grande importance relative aux principaux prédicteurs. Par exemple, le sentiment d'appartenance à l'école et une vie familiale heureuse ont été systématiquement classés parmi les facteurs les plus importants pour prédire les niveaux de dépression et d'anxiété. De plus, les arbres de décision ont montré une capacité accrue à cerner les sous-groupes à risque élevé, ce qui est crucial pour cibler les efforts de prévention et d'intervention en santé mentale(Battista et al., 2023).

Notre travail pousse à mettre en avant trois avantages des arbres de décision concernant les problèmes de santé mentale.

1. Interprétabilité. Contrairement à d'autres modèles d'apprentissage automatique plus complexes, les arbres de décision sont faciles à interpréter. Chaque branche de l'arbre représente une règle de décision claire, permettant de comprendre facilement pourquoi une certaine prédiction a été faite.

2. Flexibilité. Les arbres de décision peuvent gérer à la fois des données numériques et catégorielles, ce qui les rend polyvalents pour diverses applications.
3. Manipulation des Données Manquantes. Ils peuvent traiter les données manquantes efficacement en utilisant des techniques de substitution ou en segmentant les données en fonction de leur disponibilité.

En conclusion, les arbres de décision sont des outils puissants et polyvalents pour la prédiction des problèmes de santé mentale. Ils offrent une interprétation claire et une capacité de traitement des données diverses, ce qui les rend particulièrement utiles pour des applications pratiques dans le domaine de la santé mentale.

Annexe 16 : lien GitHub du projet

https://github.com/julienDM2001/m-moire.git