Rapport d'analyse prédictive sur la santé mentale des jeunes

UE Data Santé

DAMMAK Iyed SORO Yiré

28 mars 2025

Lien GitHub du projet : https://github.com/DAMMAK26/Data-sante.git
Dossier partagé Google Drive :
https://drive.google.com/drive/folders/1Ng$_9$$ZJAGbY34HKTRKZx9S5AVnzJsX7Jb?usp=sharing$

1. Contexte

La santé mentale des jeunes représente aujourd'hui une préoccupation majeure en matière de santé publique. En Nouvelle-Calédonie, les données récentes montrent que plus de 500 étudiants sur 3 500 ont envisagé des idées suicidaires au cours des 12 derniers mois, tandis que plus de 1 000 ont fait état d'une période prolongée de tristesse ou de désespoir. Ces constats alarmants soulignent l'urgence de développer des outils d'évaluation prédictive capables d'identifier les jeunes en situation de détresse psychologique, tout en évitant toute forme de stigmatisation.

2. Objectifs

Ce projet a pour objectif de concevoir un dispositif d'analyse prédictive visant à évaluer l'état de santé mentale des jeunes à partir des réponses fournies à un questionnaire. Il poursuit les finalités suivantes :

— **Détecter précocement** différents types de troubles psychologiques (absence de bien-être, isolement, mal-être persistant, idées suicidaires) en s'appuyant sur des modèles d'apprentissage automatique.

— **Identifier les facteurs explicatifs** les plus pertinents associés à ces états, en examinant les influences des dimensions individuelles, familiales, scolaires et comportementales.

— **Favoriser une démarche préventive**, en permettant une identification discrète et proactive des jeunes à risque, grâce à une application interactive dédiée à la détection.

Ce travail s'inscrit à la fois dans une démarche scientifique et dans une perspective sociétale, en contribuant à outiller les professionnels de l'éducation et de la santé dans leur mission d'accompagnement des adolescents.

3. Données

Les données utilisées proviennent du *Baromètre Santé Jeunes 2019*, une enquête menée en Nouvelle-Calédonie. Le questionnaire complet comprend 238 variables couvrant un large éventail de thématiques :

1. Caractéristiques socio-démographiques
2. Vie familiale
3. Vie scolaire
4. État de santé
5. Accès aux soins
6. Santé bucco-dentaire
7. Alimentation
8. Activité physique
9. Sédentarité
10. Santé mentale
11. Comportements à risque dans les véhicules
12. Violence et harcèlement
13. Tabac
14. Alcool
15. Cannabis
16. Kava et autres drogues
17. Sexualité
18. Perception des risques

Pour notre analyse, un sous-ensemble de 58 variables a été sélectionné, incluant notamment :

- Informations socio-démographiques
- Indicateurs de vie familiale et scolaire
- Variables liées à l'accès aux soins, à l'alimentation, à l'hygiène de vie, à la sédentarité
- Expériences de violence et comportements à risque
- Cibles de santé mentale :
  - **SM1** : Bien-être général subjectif
  - **SM2a** : Fréquence du sentiment de solitude
  - **SM3** : Épisodes prolongés de tristesse
  - **SM6** : Idées suicidaires

4. Choix des modèles prédictifs

SM1 : Bien-être général subjectif

Pour la variable SM1, quatre modèles ont été comparés : la régression logistique, le SVM, la forêt aléatoire (Random Forest) et XGBoost. La forêt aléatoire obtient la meilleure exactitude globale (environ 79,7%), mais son rappel sur la classe minoritaire est très faible (environ 10%), avec un F1-score limité (17%). Les modèles XGBoost et SVM atteignent des performances similaires en termes d'accuracy (environ 79%), mais échouent également à bien détecter les cas minoritaires (rappel inférieur à 15%, F1 autour de 0,20), le SVM ne détectant même aucun positif.

À l'inverse, la régression logistique sacrifie légèrement la performance globale (accuracy de 75,6%) pour offrir un meilleur rappel (26,3%) et un F1-score de 34% sur la classe rare. En conclusion, le modèle retenu pour SM1 est la régression logistique, car il maximise la détection des cas minoritaires tout en conservant une précision acceptable.

SM2a : Sentiment de solitude

La variable SM2a a d'abord été abordée comme une classification multiclasse, mais les performances initiales étaient faibles (accuracy entre 38% et 44%) avec des classes rares totalement ignorées. En regroupant les réponses en deux classes (présence ou absence de solitude), les résultats se sont nettement améliorés pour tous les modèles, atteignant une accuracy entre 85% et 88% et un F1-score global d'environ 0,92.

Cependant, un déséquilibre persiste dans la détection de la classe minoritaire. Le modèle Random Forest, par exemple, ne détecte aucun cas de solitude (0% de rappel), tandis que la régression logistique n'en détecte que 3% malgré un bon F1 global (0,94). Les modèles SVM et XGBoost parviennent à identifier environ 7 à 8% des jeunes effectivement isolés. XGBoost a finalement été retenu pour SM2a, car il offre le meilleur compromis entre une performance globale élevée (accuracy de 85%) et une meilleure détection des cas positifs que les autres modèles.

SM3 : Épisodes prolongés de tristesse

Pour SM3, tous les modèles montrent des performances proches, avec une accuracy comprise entre 70% et 73% et un F1-score autour de 0,81–0,82. La régression logistique obtient l'accuracy la plus élevée (73,0%) et le meilleur F1 global (0,819), mais son rappel sur la classe minoritaire reste limité (24%).

XGBoost et SVM présentent des profils similaires avec une faible détection des cas rares (rappel de 21% environ). Le modèle Random Forest offre ici un meilleur équilibre : avec une accuracy de 71,8%, un F1 de 0,817 et un rappel de 29% sur la classe minoritaire. En conséquence, Random Forest a été choisi pour SM3, car il permet une meilleure détection des élèves en détresse prolongée tout en maintenant une bonne précision globale.

SM6 : Idées suicidaires

Pour la variable SM6, les quatre modèles évalués (régression logistique, SVM, RF et XGBoost) affichent d'excellentes performances globales. Le SVM, Random Forest et XGBoost atteignent chacun une accuracy d'environ 91,9%, avec des scores très proches en précision (92,6%), rappel (98,7%) et F1-score (0,955). La régression logistique se situe légèrement en retrait avec une accuracy de 90,0% et un F1 de 0,942.

Néanmoins, tous ces modèles présentent une faiblesse commune : bien qu'ils détectent très bien la classe majoritaire (« Non »), ils peinent à reconnaître la classe minoritaire (« Oui »), avec un rappel d'environ 46–48%. Aucun algorithme ne corrige complètement ce déséquilibre, bien que le SVM conserve une légère avance en performance globale.

Par conséquent, le SVM a été retenu pour SM6, car il présente la meilleure balance entre précision, rappel et F1-score, tout en maintenant des performances comparables à celles des modèles RF et XGBoost.

5. Résultats clés

SM1 : Bien-être général subjectif

Le modèle sélectionné pour la prédiction du bien-être général subjectif (SM1) est la régression logistique. Il permet d'identifier les élèves « heureux » versus ceux qui déclarent un mal-être. Ce modèle révèle que plusieurs attributs du mode de vie et du contexte familial sont fortement associés à un bon équilibre mental :

— **Santé perçue excellente** : les jeunes en bonne santé physique se déclarent en général plus heureux.

— **Confort matériel et financier** : une situation économique aisée réduit le stress et favorise le bien-être.

— **Image corporelle positive** : l'acceptation de son apparence agit comme un facteur protecteur important.

— **Habitudes alimentaires saines** : une alimentation équilibrée et le fait de prendre le petit-déjeuner sont liés à un meilleur bien-être.

— **Activité physique régulière** : le sport est bénéfique pour le moral et la gestion du stress.

En revanche, les indicateurs de vulnérabilité incluent :

— **Parcours scolaire spécifique** (voie professionnelle/technologique), souvent lié à un mal-être accru.

— **Présence de maladies ou handicaps**, qui fragilisent l'équilibre mental.

— **Difficultés familiales ou scolaires**, telles que les conflits ou une mauvaise ambiance à l'école.

Interprétation : Le bien-être mental des jeunes dépend fortement de leur hygiène de vie, du contexte familial et du climat scolaire. Un soutien stable et un mode de vie sain favorisent un état mental équilibré.

SM2a : Sentiment de solitude, anxiété et troubles du sommeil

Le modèle SVM a été retenu pour la prédiction des sentiments de solitude et des troubles émotionnels (SM2a). Il oppose efficacement les élèves souffrant fréquemment de ces troubles à ceux qui n'en rapportent pas.

Les principaux facteurs identifiés sont :

— **Climat scolaire** : un environnement violent ou insécurisant à l'école est fortement associé à la solitude et à l'anxiété.

— **Soutien familial et social** : les élèves dont les parents sont peu présents ou peu à l'écoute se sentent plus souvent seuls.

— **Habitudes de vie** : sédentarité excessive, mauvaise image corporelle et santé perçue dégradée augmentent le risque de troubles du sommeil et d'anxiété.

Interprétation : Les troubles émotionnels sont liés à un cumul de facteurs sociaux et comportementaux. Les élèves à risque cumulent souvent un isolement familial, une insécurité scolaire et de mauvaises habitudes de vie.

SM3 : Épisodes prolongés de tristesse

Pour la variable SM3 (mal-être psychologique prolongé), la régression logistique est le modèle retenu. Elle met en évidence les éléments suivants comme prédicteurs significatifs :

— **Niveau scolaire (1re générale)** : période critique marquée par la pression académique.

— **Insécurité scolaire** : les jeunes qui ont peur à l'école présentent davantage de symptômes dépressifs.

— **Santé fragile** : antécédents médicaux comme l'épilepsie ou consultations médicales récentes sont associés à un mal-être accru.

Les facteurs protecteurs incluent :

— **Âge avancé (Terminale)** : effet de maturité émotionnelle.

— **Bonne santé perçue et image corporelle positive**.

— **Confiance scolaire** : estime de soi académique réduit le risque de mal-être prolongé.

Interprétation : L'équilibre psychologique est influencé par l'environnement scolaire, la santé physique, et la perception de soi. Les élèves fragilisés cumulent souvent insécurité, maladies et faible estime de soi.

SM6 : Idées suicidaires

La variable SM6 concerne les idées suicidaires. Le modèle SVM a été retenu comme légèrement plus performant pour prédire cette dimension sensible. Plusieurs facteurs de risque majeurs ressortent :

— **Violences subies** : harcèlement, abus, violences physiques ou sexuelles augmentent fortement le risque suicidaire.

— **Manque de soutien familial** : adolescents livrés à eux-mêmes ou mal compris par leurs parents sont plus vulnérables.

— **Habitudes à risque** : consommation d'alcool, de tabac ou de cannabis est souvent corrélée aux idées suicidaires.

— **Mal-être prolongé non pris en charge** : SM3 et SM6 étant liés, une détresse non détectée peut évoluer vers des pensées suicidaires.

Interprétation : Le soutien affectif, la sécurité scolaire et la prise en charge précoce du mal-être sont essentiels pour prévenir les pensées suicidaires. Le genre joue également un rôle : les filles sont deux fois plus nombreuses à exprimer ces pensées, même si les garçons passent plus fréquemment à l'acte.

6. Conclusion

Ce projet a permis de créer des modèles de prédiction de la santé mentale intégrés dans une application interactive. Il démontre que certains profils de jeunes peuvent être identifiés comme à risque à partir de leurs conditions de vie et ressentis. Bien que les modèles restent imparfaits, ils constituent une base précieuse pour un dépistage non intrusif et préventif. Des perspectives incluent l'amélioration de la prédiction de la classe minoritaire et l'utilisation de techniques d'équilibrage ou de données supplémentaires.

Appendix

A. Estimations Globales pour SM6

Métriques Globales

Modèle	Accuracy	Precision	Recall	F1-score
Random Forest (RF)	0.919	0.926	0.987	0.955
XGBoost	0.919	0.926	0.987	0.955
SVM	0.919	0.926	0.987	0.955
Logistic Regression	0.900	0.915	0.972	0.942

Accuracy par classe pour SM6

Modèle	Classe	Correct	Total	Accuracy
RF	1	569	583	0.976
RF	2	42	90	0.467
XGB	1	569	583	0.976
XGB	2	42	90	0.467
SVM	1	570	583	0.978
SVM	2	42	90	0.467
Log Reg	1	551	571	0.965
Log Reg	2	49	102	0.480

B. Prédiction de la Fréquence de Solitude (SM2a)

Expériences Multiclasse

Exp.	Modèle	Accuracy	F-mesure	Commentaire
1	SVM	0.382	0.319	-
1	XGBoost	0.382	0.346	-
2	XGBoost	0.444	0.366	-
2.2	XGBoost	0.424	0.383	-
2.3	Tous	0.850	0.919	-

Binaire (Exp2.3)

Modèle	Accuracy	F-mesure	Recall
SVM, XGB, RF	0.850	0.919	1.000
Log Reg	0.882	0.937	0.979

C. Analyse SM2b

Performances Globales

Modèle	Accuracy	F-mesure	Recall
Random Forest	0.859	0.924	0.994
XGBoost	0.862	0.926	1.000
SVM	0.862	0.926	1.000
Log Reg	0.858	0.922	0.974

Accuracy par classe

Modèle	Classe	Correct	Total	Accuracy
RF	1	571	576	0.991
RF	2	3	97	0.031
XGB	1	576	576	1.000
XGB	2	0	97	0.000
SVM	1	575	576	0.998
SVM	2	0	97	0.000

D. Analyse SM2c

Performances Globales

Modèle	Accuracy	F1-score	Recall
Random Forest	0.845	0.916	1.000
XGBoost	0.842	0.913	0.982
SVM	0.848	0.916	0.988
Log Reg	0.830	0.903	0.959

Accuracy par classe

Modèle	Classe 1	Classe 2
RF	1.000	0.000
XGB	0.975	0.078
SVM	0.982	0.068

E. Analyse SM1

Modèle	Accuracy	Precision	Recall	F1-score
RF	0.797	0.576	0.100	0.171
XGB	0.789	0.482	0.149	0.228
SVM	0.791	NA	0.000	NA
Log Reg	0.756	0.484	0.263	0.341

F. Analyse SM3

Modèle	Accuracy	F1-score	Precision	Recall
RF	0.718	0.817	0.741	0.910
XGBoost	0.704	0.813	0.722	0.930
SVM	0.692	0.805	0.716	0.919
Log Reg	0.730	0.819	0.754	0.898

Modèle	Classe 1	Classe 2
RF	0.908	0.294
XGB	0.938	0.211
SVM	0.934	0.206