Rapport du projet

Mental Health Viz

Suicide, dépression, addictions et automutilation

Projet:	Mental Health Viz
Matiere:	Visualisation des données
Sujet:	OMS + IHME GBD
Auteur:	Nawfal RAZOUK
Encadrant:	Nabila ZRIRA
Formation:	3A Génie Informatique - Option : IData
Institution:	ENSMR

Rabat
2 janvier 2026

Résumé

Ce rapport présente une plateforme d'analyse et de visualisation autour de la santé mentale, avec un focus sur le suicide, la dépression, les addictions et l'automutilation. Les données proviennent de l'OMS (suicide 2021) et de l'IHME GBD (2023) et sont structurées en trois versions : v0 (visualisations statiques), v1 (tableau de bord principal sur données réelles) et v2/v3 (analyses avancées et estimation de risque). Le pipeline inclut un inventaire des sources, un mapping ISO3, un nettoyage par fichier, puis des fusions pour construire une table ML cohérente. Le tableau de bord v1 fournit des cartes, des classements et des comparaisons par sexe et âge, ainsi que des pages de contexte (tendances toutes causes, grandes catégories, probabilité de décès). La partie analytique propose des corrélations, une démonstration ML (Ridge, RandomForest) et des techniques de data mining dans v2. Le rapport insiste sur la reproductibilité, les limites de fusion inter-années et les risques d'interprétation. L'objectif est de fournir une lecture claire, professionnelle et exploitable des indicateurs de santé mentale à l'échelle mondiale.

Abstract

This report presents a mental health analytics and visualization platform focused on suicide, depression, addictions, and self-harm. Data sources combine WHO suicide statistics (2021) and IHME GBD indicators (2023), organized into three project versions : v0 (static visuals), v1 (main dashboard on real data), and v2/v3 (advanced analytics and risk estimation). The pipeline covers inventorying, ISO3 mapping, per-file cleaning, and merges to build a consistent ML table. The v1 dashboard delivers maps, rankings, and sex/age comparisons, plus contextual pages on all-cause trends, big categories, and probability of death. The analytics layer adds correlations, an ML demo (Ridge, RandomForest), and data mining techniques in v2. The report highlights reproducibility, cross-year merge limits, and interpretation risks. The goal is to provide a clear, professional, and actionable view of global mental health indicators.

Mots-clés / Keywords

Français	**English**
santé mentale	mental health
suicide	suicide
dépression	depression
addictions	addictions
automutilation	self-harm
OMS	WHO
IHME GBD	IHME GBD
visualisation	visualization

Table des matières

Table des figures

Chapitre 1 Introduction

1.1 Motivation

La santé mentale est un enjeu mondial majeur, à la fois médical, social et économique. Les troubles dépressifs, les addictions et l'automutilation affectent des millions de personnes et fragilisent les systèmes de santé. En parallèle, le suicide reste une cause importante de mortalité évitable. Disposer d'une lecture claire, comparative et accessible de ces indicateurs est essentiel pour orienter la prévention et les politiques publiques.

1.2 Pourquoi le suicide et les troubles mentaux comptent

Le suicide représente un indicateur critique de détresse psychologique et sociale. Les troubles mentaux et les troubles liés aux substances constituent des facteurs de risque majeurs et contribuent aussi à la charge de morbidité (DALYs). Comprendre comment ces dimensions évoluent dans l'espace (pays, régions) et selon les caractéristiques démographiques (sexe, âge) permet d'identifier des zones à risque, des populations vulnérables et des pistes d'action.

1.3 Périmètre et public cible

Ce rapport présente une analyse multi-couches combinée de données OMS et IHME GBD, avec trois versions du projet : v0 (visualisations statiques), v1 (tableaux de bord sur données réelles) et v2/v3 (analyses avancées et outils d'aide à la décision). Le public cible est double : d'une part les enseignants et évaluateurs, d'autre part les utilisateurs non techniques qui ont besoin d'une lecture intuitive des tendances et des relations entre indicateurs.

Chapitre 2 Objectifs et questions de recherche

2.1 Objectifs

— O1 : Décrire les profils du suicide à l'échelle mondiale et régionale.
— O2 : Comparer les charges liées à la dépression, aux addictions et à l'automutilation.
— O3 : Explorer les relations entre indicateurs et construire des modèles prédictifs de base.
— O4 : Démontrer des techniques de BI et de data mining sur des données de santé publique.
— O5 : Garantir la reproductibilité avec des pipelines et des versions clairement séparées.

2.2 Questions de recherche

— Q1 : Quelles sont les différences géographiques majeures des taux de suicide et comment se distribuent-elles par région ?
— Q2 : Comment se positionnent les pays sur les indicateurs de dépression (DALYs), d'addictions et d'automutilation ?
— Q3 : Observe-t-on des relations cohérentes entre les indicateurs (suicide vs dépression, addictions, automutilation) ?
— Q4 : Un modèle prédictif simple peut-il estimer un risque relatif à partir des indicateurs disponibles ?
— Q5 : Quelles techniques avancées (clustering, détection d'anomalies) apportent une valeur ajoutée en analyse exploratoire ?

Chapitre 3 Sources de données et couverture

3.1 Catalogue des jeux de données

Le catalogue automatique `v1/data_clean/dataset_catalog.csv` recense 14 fichiers bruts : 7 fichiers OMS (suicide global + régions) et 7 fichiers IHME GBD. Pour chaque fichier, le catalogue fournit le nombre de lignes, les colonnes disponibles, les années, les sexes, les âges, les causes, ainsi que les mesures et métriques associées. La matrice d'usage `v1/report/dataset_usage_matrix.md` documente l'utilisation de chaque fichier dans les pages du tableau de bord et dans la table ML.

3.2 Données OMS sur le suicide (2021)

La source OMS fournit une vue 2021 du suicide par pays [1].

— `who_global_master.csv` (549 lignes) : variables de base (pays, sexe, groupe de revenu, région), taux brut et taux standardisé par âge (2021), et nombre total de suicides.

— Fichiers régionaux (drilldown) : `who_africa_region_full.csv` (141), `who_americas_region_ful` (99), `who_emro_region_full.csv` (63), `who_europe_region_full.csv` (150), `who_searo_regio` (33), `who_wpro_region_full.csv` (63).

3.3 Jeux de données GBD centraux (v1)

La version v1 s'appuie sur un noyau de fichiers GBD pour la partie analytique et la table ML [2] :

— `IHME-GBD_2023_DATA-dalys-causes-1.csv` (69 768 lignes) : DALYs, métrique Rate/Number/Percent, année 2023, sexe Both, âges `<20`, `20-24`, `25+`. Utilisé pour la dépression.

— `IHME-GBD_2023_DATA-deaths-mental-substance-violence-1.csv` (88 128 lignes) : Deaths, métrique Rate/Number/Percent, année 2023, sexes Male/Female, âges `<20`, `20-24`, `25+`. Utilisé pour l'automutilation et les addictions.

— `IHME-GBD_2023_DATA-age-standardized-death-rate-1.csv` (30 615 lignes) : Deaths, métrique Rate, année 2023, âge standardisé. Utilisé pour les comparaisons par cause (addictions).

Ces fichiers sont combinés avec `who_global_master.csv` pour constituer la table ML (taux de suicide + indicateurs GBD).

3.4 Jeux de données GBD de contexte

Des jeux de données additionnels enrichissent les pages de contexte :

— `IHME-GBD_2023_DATA-all-cause-burden-all-ages-1.csv` (63 585 lignes) : 2021–2023, mesures Deaths/DALYs/YLLs, métriques Number/Percent/Rate, âge `All ages`, cause `All causes`.

— `IHME-GBD_2023_DATA-probability-of-death-1.csv` (93 480 lignes) : métrique `Probability of death`, année 2023, sexes Both/Female/Male, âges `<20`, `20-24`, `25+`, `All ages`.

— `IHME-GBD_2023_DATA-risk-factor-burden-1.csv` (51 480 lignes) : mesures Deaths/-DALYs, métriques Number/Percent/Rate, année 2023, détail par facteurs de risque.

3.5 Jeux de données additionnels pour v0

La version v0 (visualisations statiques) exploite l'ensemble des fichiers bruts pour maximiser la variété de graphiques. Un fichier supplémentaire est spécifique à v0 :

— `IHME-GBD_2023_DATA-anemia-prevalence-ylds-1.csv` (56 520 lignes) : 2020–2023, mesures Prevalence/YLDs, métriques Number/Percent/Rate, âge `All ages`.

3.6 Notes de couverture

La couverture temporelle est hétérogène : OMS est figée sur 2021, tandis que GBD est principalement 2023, avec des séries courtes pour les tendances (2021–2023) et l'anémie (2020–2023). Les métriques `Rate` sont des taux (par 100 000 habitants), `Number` des volumes bruts, `Percent` des parts relatives, et `Probability of death` une probabilité entre 0 et 1. Les variables de sexe (Both/Male/Female) et d'âge (`<20`, `20-24`, `25+`, `All ages`, `Age-standardized`) sont harmonisées dans la phase de nettoyage pour assurer la cohérence des comparaisons.

Chapitre 4 Préparation des données et pipeline

4.1 Inventaire et profilage

Le script `src/00_inventory.py` parcourt `data_raw/` et produit un catalogue complet des fichiers avec le nombre de lignes, la liste des colonnes et les valeurs uniques pour `year`, `sex_name`, `age_name`, `cause_name`, `metric_name` et `measure_name`. Le résultat est écrit dans `v1/data_clean/dataset_catalog.csv` et sert de base à l'audit des sources.

4.2 Cartographie des pays et ISO3

Le script `src/01_country_mapping.py` harmonise les noms de pays OMS et GBD vers un code ISO3 unique, en combinant un mapping automatique (`pycountry`) et des exceptions manuelles. Deux fichiers sont générés :

- `data_clean/country_iso3_mapping.csv` : mapping final (source OMS/GBD, type de match).
- `data_clean/country_iso3_unmatched.csv` : liste des pays non résolus pour correction.

4.3 Nettoyage OMS

Le script `src/02_clean_who.py` lit `who_global_master.xlsx` ou `who_global_master.csv`, normalise les noms de colonnes, ajoute `iso3` et fixe l'année à 2021. Les sorties principales sont :

- `data_clean/who_2021_clean.csv` (global).
- `data_clean/who__clean.csv` (régions, si présentes).

4.4 Nettoyage GBD

Le script `src/03_clean_gbd.py` applique des filtres par fichier (cause, mesure, métrique) et insère le code ISO3. Les sorties standardisées sont :

- `data_clean/gbd_addiction_clean.csv`
- `data_clean/gbd_selfharm_clean.csv`
- `data_clean/gbd_depression_dalys_clean.csv`
- `data_clean/gbd_prob_death_clean.csv`
- `data_clean/gbd_allcauses_clean.csv`

— `data_clean/gbd_big_categories_clean.csv`

4.5 Fusions et tables analytiques

Le script `src/04_merge_ml.py` fusionne OMS et GBD sur `iso3` pour construire la table ML (suicide 2021 + indicateurs GBD). Le script `src/05_merge_context.py` produit des tables de contexte prêtes à visualiser (tendances toutes causes, grandes catégories, probabilité de décès) dans `data_clean/context_tables/`.

4.6 Table ML et rapport de baseline

Le script `src/06_ml_baseline.py` construit une table de features, entraîne des modèles simples (Ridge, RandomForest) et écrit :

— `data_clean/ml_baseline_features.csv`
— `data_clean/ml_baseline_predictions.csv`
— `report/ml_baseline_results.csv`
— `report/ml_baseline_cv.csv`
— `report/ml_feature_importance.csv`
— `report/ml_baseline.md`

4.7 Pipelines versionnées

La variable d'environnement `MHP_VERSION` oriente toutes les sorties vers `v1/`, `v2/` ou `v3/`. Des scripts de pipeline sont fournis :

— `scripts/run_v1_pipeline.py` (v1, données réelles).
— `scripts/run_v2_pipeline.py` (v2, analytics avancées).
— `scripts/run_v3_pipeline.py` (v3, risk estimator).
— v0 utilise `src/v0_visuals.py` pour les exports statiques.

Chapitre 5 Modèle BI et indicateurs (KPIs)

5.1 Schéma en étoile

Le modèle BI est construit autour d'une table de faits principale issue de la fusion OMS+GBD, et de dimensions simples qui permettent la navigation par pays, temps, sexe et âge.

— **Table de faits** : `FACT_MentalHealth_CountryYearAge` (`v1/data_clean/merged_ml_country.c`
— **Grain** : Pays × groupe d'âge × année (OMS 2021) avec année GBD stockée à part.

Dimensions principales

— `DIM_Country` : `iso3`, `location_name`, `region_name`, `income_group`
— `DIM_Time` : `year` (OMS), `gbd_year` (GBD)
— `DIM_AgeGroup` : `age_name`
— `DIM_Sex` : `sex_name` (stocké par mesure dans la table ML)

Tables de contexte (couche analytique)

— `FACT_Context_AllCause` : `v1/data_clean/context_tables/context_allcauses_trend.csv`
— `FACT_Context_ProbDeath` : `v1/data_clean/context_tables/context_probdeath_2023.csv`
— `FACT_Context_BigCategories` : `v1/data_clean/context_tables/context_big_categories_`

5.2 Définitions des KPIs

Les indicateurs clés sont définis de façon cohérente avec les sources OMS/GBD :

— **Taux de suicide (OMS)** : `age_standardized_suicide_rate_2021` (pour 100 000).
— **Taux brut de suicide** : `crude_suicide_rate_2021` (pour 100 000).
— **Dépression (GBD)** : `gbd_depression_dalys_rate_both` (DALYs rate / 100 000).
— **Addictions (GBD)** : `gbd_addiction_death_rate_both/female/male` (Deaths rate / 100 000).
— **Automutilation (GBD)** : `gbd_selfharm_death_rate_female/male` (Deaths rate / 100 000).
— **Probabilité de décès** : `metric_name = Probability of death` (valeurs entre 0 et 1).

5.3 Dictionnaire des données

Le dictionnaire détaillé des colonnes est disponible dans `v1/report/data_dictionary.md`. Il précise pour chaque variable : définition, unité, et remarques de qualité. Les champs clés couvrent :

— iso3, location_name, region_name, income_group
— year, gbd_year, age_name, sex_name
— measure_name, metric_name, cause_name
— val, upper, lower

Chapitre 6 Versions du projet

6.1 v0 : Visualisations statiques

La version v0 est une vitrine visuelle. Elle utilise les fichiers bruts avec des transformations minimales et génère des figures PNG/HTML pour couvrir un maximum de types de graphiques (cartes, violons, heatmaps, treemaps, small multiples). Elle sert à démontrer la richesse des sources et la capacité de restitution sans logique applicative complexe.

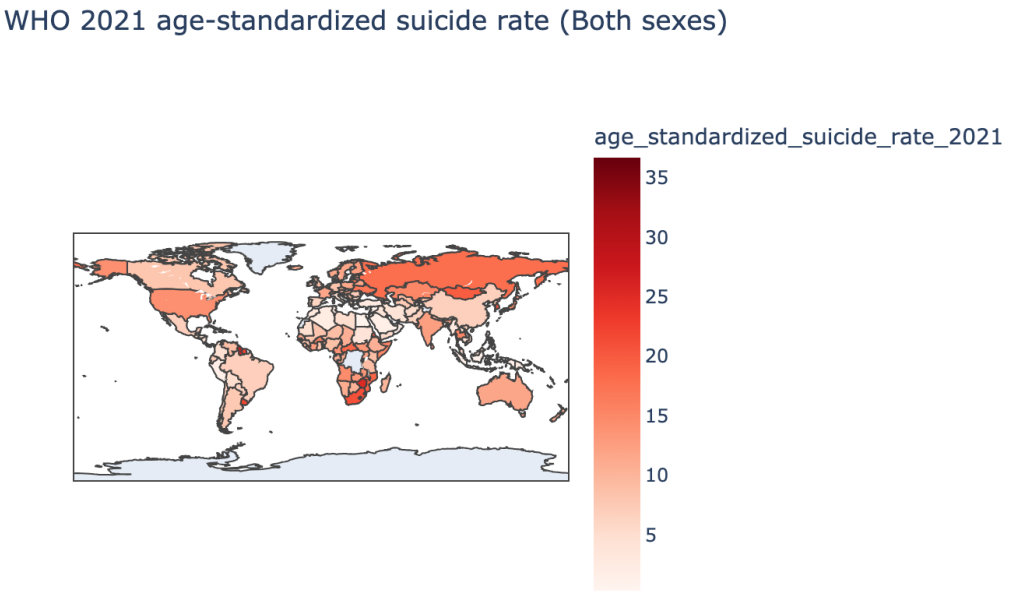

FIGURE 6.1 – Exemple v0 : carte mondiale du taux de suicide âge-standardisé (OMS 2021).

6.2 v1 : Tableau de bord central (données réelles)

La version v1 constitue le cœur analytique. Elle s'appuie sur les données OMS et GBD nettoyées, produit une table ML cohérente, et propose un tableau de bord complet avec pages thématiques (suicide, dépression, addictions, automutilation, contexte). Elle inclut

une baseline ML (Ridge, RandomForest) et une documentation BI (modèle en étoile, dictionnaire, qualité).

6.3 v2 : Analyses avancées (données synthétiques)

La version v2 permet d'explorer des techniques avancées de data mining et de BI à partir d'un jeu synthétique. On y trouve la segmentation, la détection d'anomalies, les forecasts, les intervalles de prédiction, l'explicabilité, les graphes de similarité, les règles d'association, ainsi que des rapports de qualité et des tests. Cette version met l'accent sur l'innovation et la démonstration méthodologique.

6.4 v3 : Estimateur de risque

La version v3 propose un module interactif d'estimation de risque (probabilité de haut risque) à partir d'inputs utilisateur. Elle intègre la calibration, des explications simplifiées et des scénarios contrefactuels pour illustrer l'usage décisionnel des modèles.

6.5 Tableau comparatif

Version	Points forts
v0	Visualisations statiques, couverture maximale des jeux bruts, export PNG/HTML.
v1	Tableau de bord principal, données réelles, table ML, BI documentation.
v2	Analytics avancées, data mining, qualité, forecasting, synthetic data.
v3	Module interactif, estimation de risque, calibration et scénarios.

Chapitre 7 Résultats et tableau de bord (v1)

7.1 Vue d'ensemble

La page d'accueil consolide les indicateurs principaux et fournit une lecture rapide :
— Cartes KPI : taux de suicide âge-standardisé, taux brut, et points de repère globaux.
— Carte choropleth 2021 pour situer la charge par pays.
— Tendance régionale : évolution du taux par région pour situer les dynamiques.
L'objectif est de donner une première lecture, puis d'orienter l'exploration vers les pages thématiques.

7.2 WHO Suicide Explorer

Cette page se concentre sur les données OMS 2021 :
— Carte des taux âge-standardisés par pays.
— Classements (top/bottom) des pays.
— Comparaison par sexe (Male/Female/Both) pour visualiser les écarts.
— Dispersion taux brut vs taux standardisé pour détecter les effets de structure d'âge.

7.3 Dépression (DALYs)

La page dépression s'appuie sur les DALYs (GBD 2023) :
— Carte des DALYs rate par pays.
— Top 20 des pays les plus élevés.
— Comparaison par groupes d'âge (ex. <20, 20-24, 25+).
Cette vue met en avant les différences d'âge et les zones de forte charge.

7.4 Addictions

La page addictions explore les causes liées aux troubles de l'usage de substances :
— Sélection de cause (alcohol, drug, substance use disorders).
— Carte et top 20 par taux de décès.
— Comparaison par sexe.

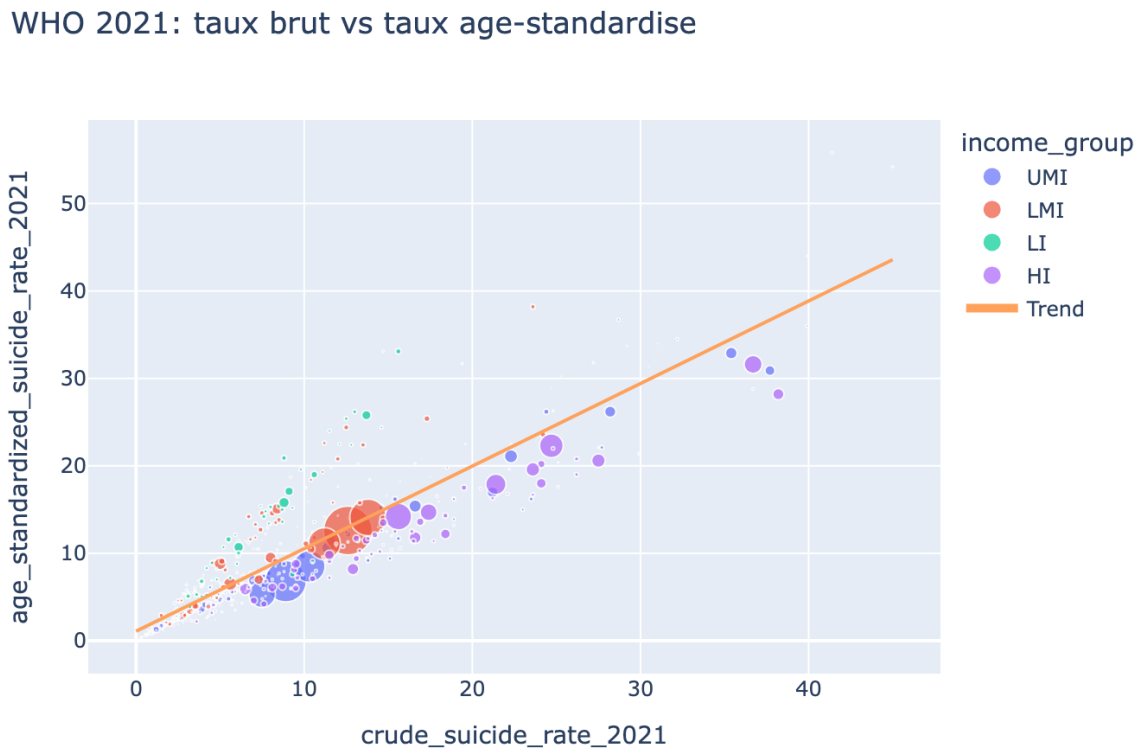

Figure 7.1 – v1 : comparaison taux brut vs taux âge-standardisé (OMS 2021).

7.5 Automutilation

La page self-harm présente :
— Carte du taux de décès par automutilation.
— Comparaison Male/Female.
— Optionnel : distribution par méthodes si disponible.

7.6 Probabilité de décès

Cette page utilise la métrique `Probability of death` (GBD) :
— Carte mondiale pour un couple cause/âge sélectionné.
— Classement des pays.
La lecture se fait en probabilité (0 à 1), différente d'un taux par 100 000.

7.7 Tendances toutes causes

Les tendances globales 2021–2023 sont présentées par métrique :
— Courbes `Number` et `Rate` par sexe.
— Comparaison entre évolution des volumes et évolution des taux.
Cette page sert de contexte macro pour les indicateurs mentaux.

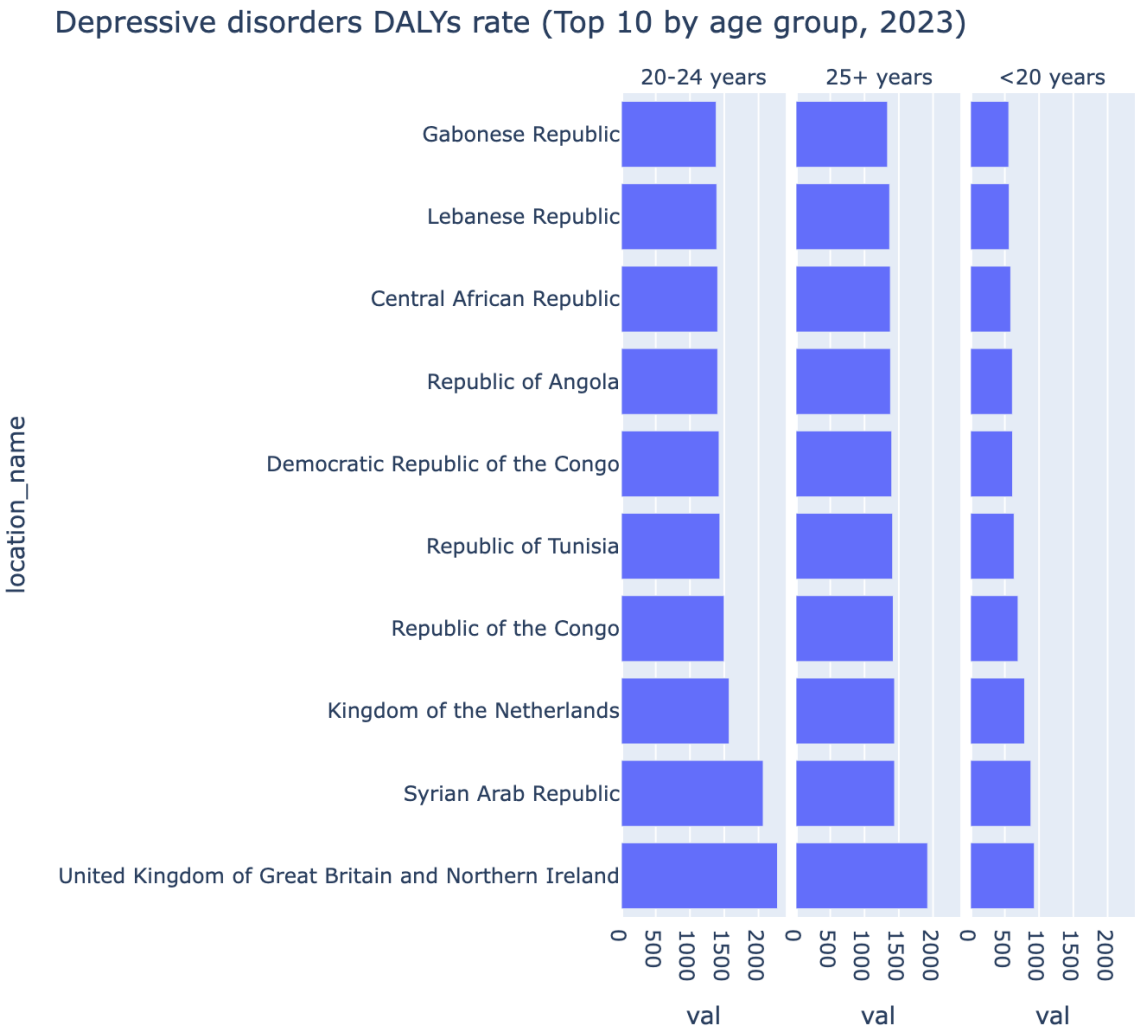

Figure 7.2 – v1 : top 10 des DALYs de dépression par groupe d'âge (2023).

7.8 Grandes catégories

La page "Big categories" résume la structure des causes (GBD) via un treemap :
- Niveau 0 : All causes.
- Niveau 1 : Communicable, Non-communicable, Injuries.
- Niveau 2 : Substance use disorders dans Non-communicable.
- Niveau 3 : Alcohol use disorders et Drug use disorders.

Cette hiérarchie met en perspective la place des troubles liés aux substances.

7.9 Relations entre indicateurs

La page "Relationships" montre les liens entre suicide et autres charges :
- Scatter suicide vs dépression.
- Scatter suicide vs addictions.
- Heatmap de corrélations.

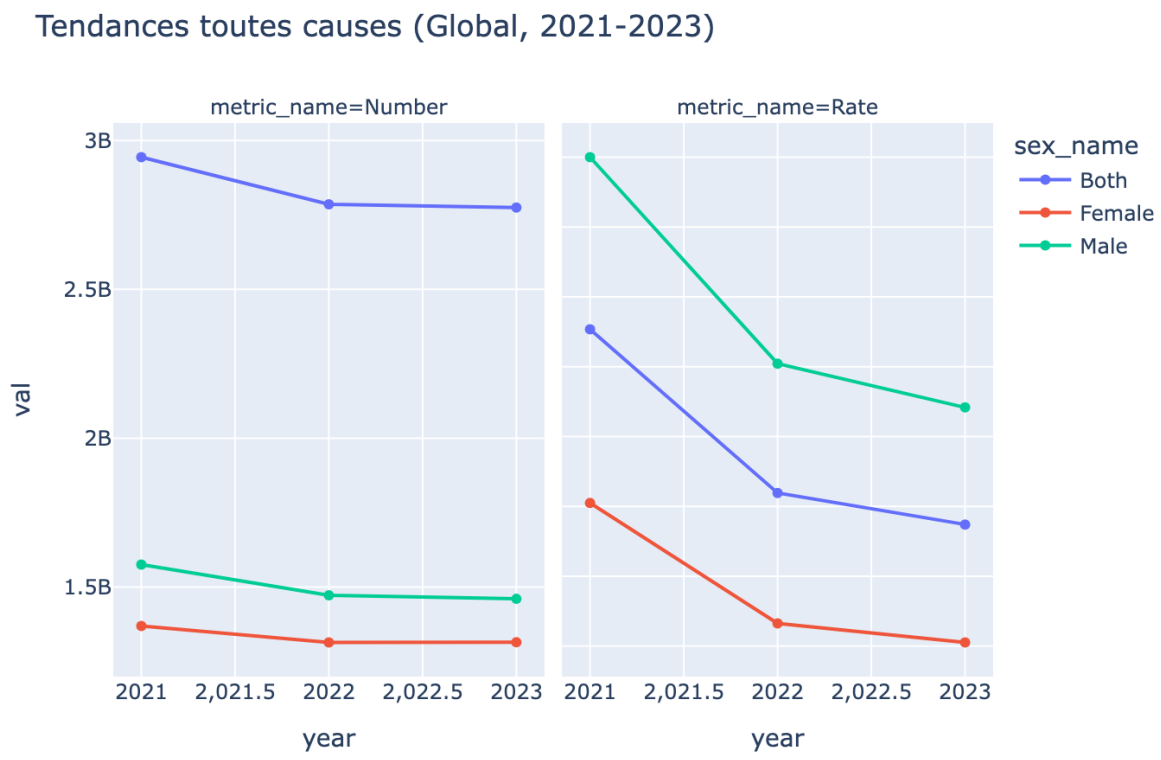

Figure 7.3 – v1 : tendances toutes causes (Global, Number et Rate).

7.10 Démonstration ML

La page ML illustre une baseline prédictive :
— Modèles Ridge et RandomForest.
— Métriques MAE et R2 (holdout + validation croisée).
— Importance des variables pour interprétabilité.

7.11 Méthodes et limites

La page finale rappelle les principaux points de prudence :
— Décalage temporel OMS 2021 vs GBD 2023.
— Données agrégées et risque d'erreur écologique.
— Données manquantes ou pays non appariés en ISO3.

Figure 7.4 – v1 : treemap des grandes catégories (hiérarchie GBD).

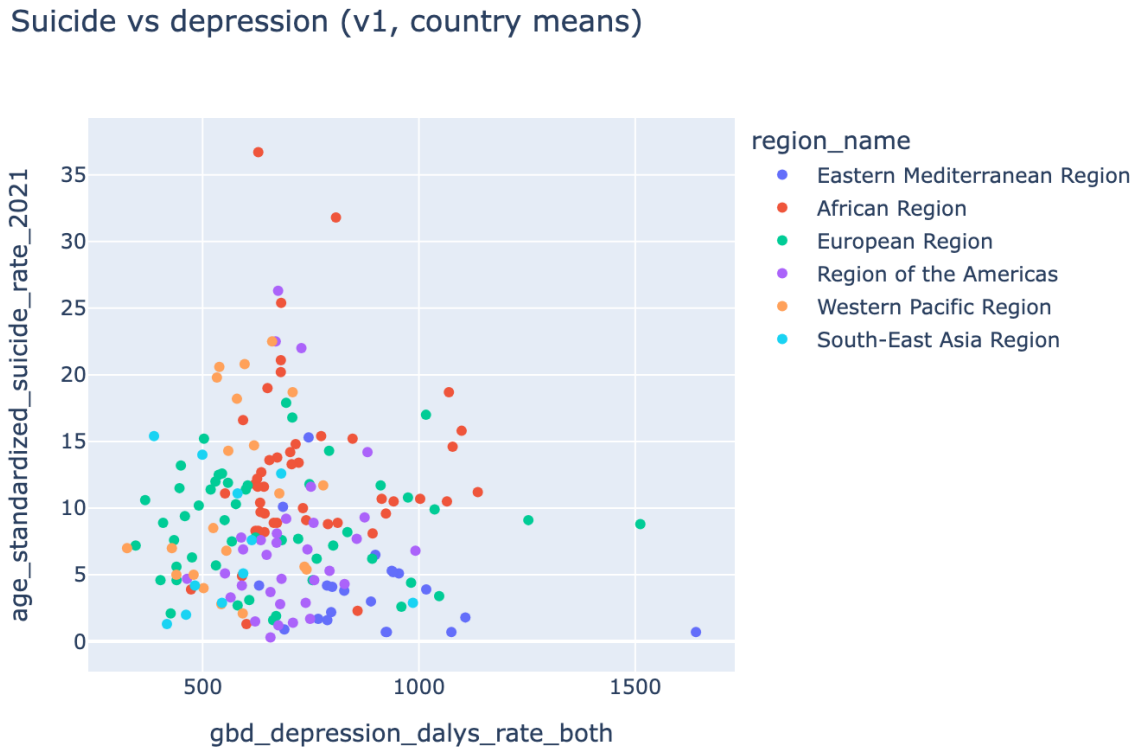

Figure 7.5 – v1 : relation entre taux de suicide et DALYs de dépression (moyenne pays).

Chapitre 8 Analyses avancées (v2)

8.1 Clustering et segmentation

La version v2 introduit des profils de pays par clustering sur des indicateurs synthétiques (suicide, dépression, addictions, automutilation). L'objectif est d'identifier des groupes de pays aux profils similaires et d'offrir une lecture comparative (clusters faibles, moyens, élevés).

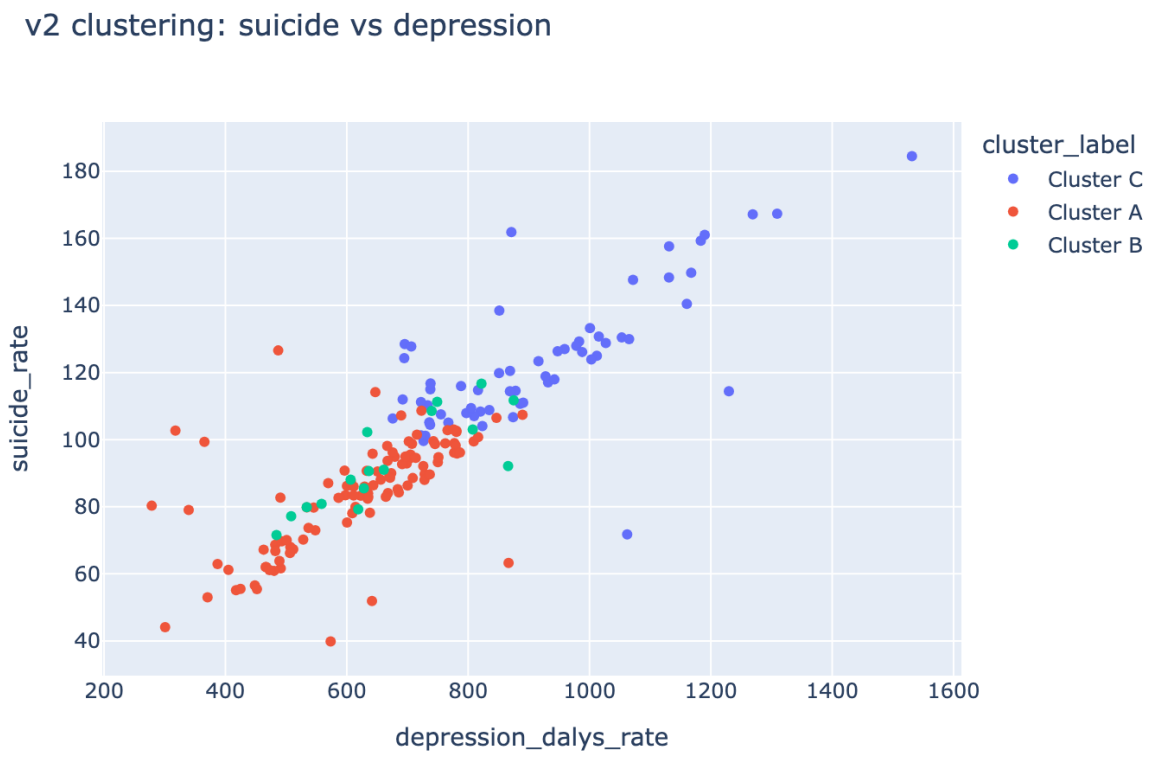

FIGURE 8.1 – v2 : dispersion des pays par cluster (suicide vs dépression).

8.2 Anomalies et alertes

Une détection d'anomalies met en évidence des pays atypiques (valeurs extrêmes ou combinaisons incohérentes). Ces résultats sont présentés sous forme de tableaux et de nuages de points afin de guider des analyses qualitatives.

8.3 Forecasts et backtest

Les tendances régionales sont modélisées avec des approches classiques et une variante deep learning. Un backtest évalue la performance via MAE/RMSE et met en comparaison les trajectoires prédites vs observées.

8.4 Régression quantile et intervalles

La régression quantile permet de produire des intervalles de prédiction (ex. quantiles 0.1, 0.5, 0.9) afin d'exprimer l'incertitude et de ne pas se limiter à une valeur ponctuelle.

8.5 Explicabilité

Deux techniques sont mobilisées :
— Importance par permutation pour classer les variables selon leur impact.
— Dépendances partielles pour visualiser l'effet marginal des features majeures.

8.6 Graphes de similarité et règles d'association

Un graphe de similarité (distance cosinus) produit des communautés de pays, et des règles d'association sont extraites après discrétisation (low/med/high). Ces outils renforcent l'aspect data mining et l'interprétation exploratoire.

8.7 Scénario lab

Le module "what-if" simule l'effet de variations des indicateurs (ex. baisse des addictions) sur des mesures cibles. Il sert à illustrer l'usage décisionnel des modèles.

8.8 Synthèse qualité

La qualité des données v2 est suivie via un rapport Great Expectations et un résumé de tests (types, plages, valeurs manquantes). Cela permet d'encadrer l'usage des données synthétiques par des contrôles explicites.

Chapitre 9 Estimateur de risque (v3)

9.1 Problématique et seuil de risque

La v3 formalise un score de *haut risque* à partir de la probabilité prédite. Le seuil (cutoff) est sélectionné dans l'interface et permet d'adapter l'analyse à des contextes plus ou moins stricts (ex. 0,60 ou 0,70). Ce choix est explicite et ajustable afin de garder une interprétation transparente.

9.2 Entrées et sorties du modèle

L'utilisateur fournit des valeurs pour les variables clés (ex. dépression, addictions, automutilation), ainsi que le pays/région lorsqu'ils sont disponibles. Le modèle retourne :

- une probabilité estimée de haut risque ;
- une décision binaire (au-dessus ou en dessous du seuil) ;
- des éléments d'explication visuels pour comprendre le résultat.

9.3 Calibration et fiabilité

La calibration isotone est appliquée pour aligner les probabilités prédites avec les fréquences observées. Une courbe de fiabilité permet de vérifier que les prédictions sont interprétables comme des probabilités cohérentes.

9.4 Pistes contrefactuelles

Un module de simulation propose des variations simples (ex. baisse de 10% de l'indicateur d'automutilation) et montre l'effet attendu sur la probabilité de risque. Cela transforme le modèle en outil d'aide à la décision.

9.5 Drivers des prédictions

Des barres d'explication de type SHAP simplifiées permettent d'identifier les variables qui poussent le plus le risque vers le haut ou vers le bas. L'objectif est de rendre le modèle interprétable pour un public non technique.

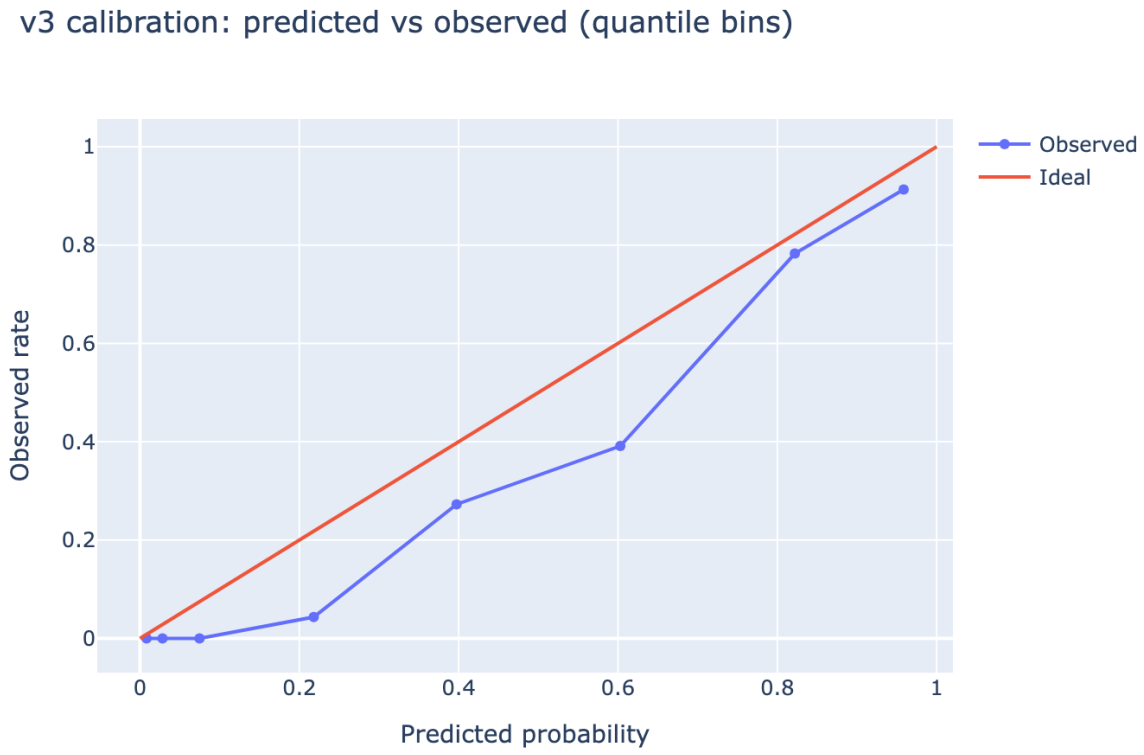

Figure 9.1 – v3 : calibration ou distribution du risque selon le seuil.

Chapitre 10 Qualité des données et limites

10.1 Manquants et trous de couverture

Le scorecard de qualité (`v1/report/data_quality_scorecard.md`) montre que plusieurs fichiers GBD comportent des lignes sans ISO3 (valeurs agrégées GBD). Les fichiers contextuels ont donc une proportion élevée d'iso3 manquants (ex. `context_big_categories_2023` proche de 99%), ce qui est attendu pour des agrégats. La table ML finale (`merged_ml_country`) ne contient pas de manquants ISO3. Le rapport `data_quality_iso3_unmatched.csv` indique 574 pays/entités GBD non appariés lors du mapping ISO3, principalement des agrégats et des sous-régions.

10.2 Fusion inter-années

La fusion principale combine un outcome OMS 2021 avec des features GBD 2023. Cela introduit un décalage temporel explicite : les relations estimées ne sont pas causales et peuvent varier entre les périodes. Cette limitation est affichée dans la page méthodes du tableau de bord et rappelle que les analyses ML sont exploratoires.

10.3 Risque d'erreur écologique

Les analyses sont conduites à l'échelle des pays et régions. Les corrélations observées ne doivent pas être interprétées comme des relations individuelles. Cette erreur écologique est une limite classique des analyses macroscopiques et doit être mentionnée dans toute interprétation.

10.4 Limites liées aux données synthétiques

Les versions v2/v3 reposent sur des données synthétiques pour permettre des démonstrations avancées (clustering, forecasting, scenario lab). Ces résultats ont une valeur pédagogique et méthodologique, mais ne doivent pas être interprétés comme des faits empiriques. Cette distinction est indiquée dans la documentation et dans les pages de méthodes.

Chapitre 11 Reproductibilité

11.1 Dépendances

Les dépendances sont séparées en deux niveaux :
— `requirements.txt` : dépendances cœur (pandas, numpy, plotly, streamlit, scikit-learn, etc.).
— `requirements-v2.txt` : dépendances avancées (torch, tslearn, great_expectations, mlxtend, networkx, streamlit-plotly-events).

Installation recommandée :

```
pip install -r requirements.txt
pip install -r requirements-v2.txt
```

11.2 Exécuter chaque version

Les commandes sont centralisées dans `VERSIONS.md`. Résumé :
— v0 (exports statiques) : `MHP_VERSION=v0 python src/v0_visuals.py`
— v1 (pipeline complet) : `python scripts/run_v1_pipeline.py`
— v2 (synthétique/avancé) : `python scripts/run_v2_pipeline.py`
— v3 (risk estimator) : `python scripts/run_v3_pipeline.py`
— Dashboard : `python scripts/run_app.py -version vX`

11.3 Sorties attendues

v0 :
— `v0/assets/` : PNG et HTML
— `v0/assets/manifest.csv`
v1 :
— `v1/data_clean/` : fichiers nettoyés + `merged_ml_country.csv`
— `v1/data_clean/context_tables/` : tables de contexte
— `v1/report/` : qualité, dictionnaire, ML baseline
v2 :
— `v2/data_clean/` : données synthétiques (long/country/region) + clusters
— `v2/report/` : forecasting, clustering, backtest, qualité
v3 :
— `v3/data_clean/` : tables de features v1/v2
— `v3/report/` : résumé des features et rapports légers

Chapitre 12 Conclusion et perspectives

12.1 Constats principaux

Ce projet met en évidence la forte hétérogénéité des taux de suicide et des indicateurs associés selon les pays, les régions, le sexe et l'âge. Les tableaux de bord v1 offrent une lecture claire des tendances et des comparaisons internationales, tandis que les couches analytiques (v2/v3) illustrent la valeur d'approches avancées pour la segmentation, la détection d'anomalies et l'estimation de risque.

12.2 Perspectives

Plusieurs extensions sont envisageables :

— Intégrer des années supplémentaires (séries longues) pour consolider l'analyse temporelle.
— Enrichir les facteurs explicatifs (variables socio-économiques, services de santé, indicateurs d'accessibilité).
— Valider les modèles sur des données externes ou des études de cas nationales.
— Ajouter des mécanismes de gouvernance des données (audit automatique, versioning avancé).

Chapitre A Annexes

A.1 Catalogue des datasets

Le catalogue complet est fourni dans `v1/data_clean/dataset_catalog.csv`. Il liste chaque fichier, ses colonnes et ses valeurs uniques par dimension (année, sexe, âge, cause, mesure, métrique).

A.2 Matrice d'usage

La matrice d'usage est disponible dans `v1/report/dataset_usage_matrix.md`. Elle justifie l'utilisation de chaque dataset dans les pages du tableau de bord et dans la table ML.

A.3 Rapport de qualité de fusion

Le rapport de fusion est disponible dans `v1/report/merge_quality.md`. Il documente les tailles avant/après, les filtres appliqués et la note de fusion inter-années.

Bibliographie

[1] World Health Organization, “Suicide worldwide in 2021,” Dataset, 2021, wHO suicide data (global and regional extracts).

[2] Institute for Health Metrics and Evaluation, “Global burden of disease study 2023 results,” Dataset, 2023, iHME GBD 2023 data files used for depression, addiction, self-harm, all-cause trends, probability of death, and risk factors.