PROJET DE BUSINESS INTELLIGENCE

AI-JE DES RISQUES DE DÉVELOPPER UNE MALADIE MENTALE ?

Cécile GIANG

Célina KHALFAT

I - Introduction	2
II- Données et modélisations	3
III - La dépression et l'anxiété dans le monde	4
IV - Comment mon environnement affecte-t-il ma santé mentale ?	9
Liberté	9
Sécurité	11
V - Ma santé physique a-t-elle un impact sur ma santé mentale ?	15
VI - Les aides médicales proposées dans ma zone géographique suffisent-elles	s pour
m'accompagner ?	18
Dépenses liées à la santé	18
Ressources humaines : Professionnels de santé	18
Aménagements médicaux : Hôpitaux psychiatriques	18
VII - Quel est l'impact du travail sur mon bien-être ?	18
VIII - Conclusion	21

I - Introduction

L'OMS décrit la santé mentale comme "un état de bien-être qui permet à chacun de réaliser son potentiel, de faire face aux difficultés normales de la vie, de travailler avec succès et de manière productive et d'être en mesure d'apporter une contribution à la communauté". Dans ce projet, lorsque nous mentionnerons les "maladies mentales" ou les "troubles mentaux", nous parlerons en particulier de l'anxiété et la dépression. Nous allons chercher à identifier les différents paramètres à prendre en compte pour essayer d'anticiper l'apparition de troubles de l'anxiété et de dépression à travers le monde. Pour cela, nous allons répondre aux différentes sous-problématiques suivantes :

- Comment mon environnement affecte-t-il ma santé mentale ?
- Ma santé physique a-t-elle un impact sur ma santé mentale?
- Les aides médicales proposées dans ma zone géographique suffisent-elles pour m'accompagner?
- Quel est l'impact du travail sur mon bien-être?

Toutes ces sous-problématiques vont nous permettre de répondre à une problématique plus globale : Ai-je des risques de développer une maladie mentale ?

II- Données et modélisations

Tous nos datasets, excepté celui de la géographie, viennent de *Kaggle* et de *Our World in Data*. Celui de la géographie, qui au départ contient les continents et les pays, vient lui de *data.world*. Nous les détaillerons dans les parties qui suivent.

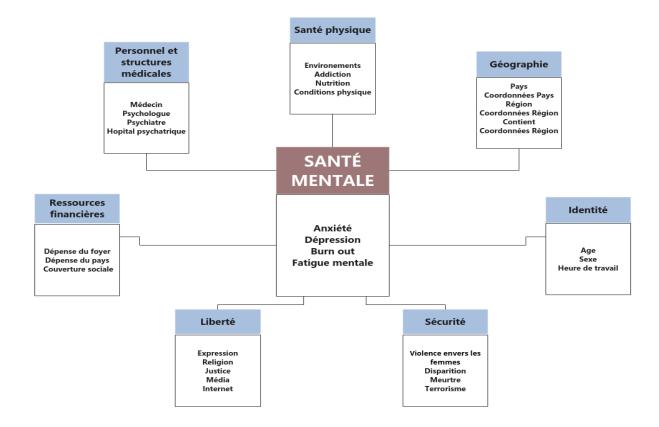
Le fait Santé mentale est exprimé en fonction des dimensions :

- <u>Santé Physique</u>: Nous avions au départ les informations suivantes: le nombre de morts, par pays et par années (de 1990 à 2015) dues à la pollution, à l'eau, à l'alcool, à la drogue, au tabagisme, à des régimes faibles en fruits, légumes, viandes rouges ou différentes sortes de graines, à une faible activité physique et à un IMC trop élevé. À partir de toutes ces catégories, nous en formons 4 : Environment, Addiction, Nutrition et Condition physique, qui correspondent à la proportion du nombre moyen de morts par rapport à la population pour chacune de ces catégories. Nous faisons également une moyenne sur les 25 années.
- <u>Géographie</u>: Nous avons les coordonnées, latitude et longitude, associé à chaque pays, région du monde et continent. Les pays ont la granularité la plus fine. Par ailleurs, il faut noter que deux pays appartenant à la même région, peuvent appartenir à deux continents différents.
- <u>Identité</u> : Nous prenons en compte le sexe, l'age et le nombre d'heures travaillées par jour pour une personne.
- <u>Sécurité</u>: Les attributs disparition, meurtre et terrorisme correspondent chacun à un réel compris entre 0 et 10, avec 0 correspondant à un pays sûr. La violence envers la Femme correspond à la proportion des femmes qui ont déjà subit des violences physiques ou sexuelles dans chaque pays.

- Liberté: Nous regardons différents aspects de la liberté: la liberté d'expression, la liberté de religion, l'idée d'équité en justice, la liberté des médias ainsi que la liberté d'avoir accès à internet sans être contrôlé ou bridé. Un réel entre 0 et 10 est attribué à chacun de ces aspects, avec 0 une note où la liberté est un droit respecté.
- Ressources financières: Dans cette dimension nous recensons pour chaque pays le pourcentage des foyers dépensant plus de 10% de leur revenu dans la santé, le pourcentage du PIB dépensé dans le secteur de la santé et enfin le pourcentage de la population couverte par une assurance santé.
- <u>Personnel et structures médicales</u> : Nous avons le nombre de médecins, de psychologues, psychiatres et hôpitaux psychiatriques pour 100 000 habitants par pays.

Plusieurs traitements ont été faits sur les datasets. Le premier traitement consistait à garder seulement les colonnes qui nous intéressaient. Les datasets contenant des informations sur plusieurs années ont été agrégés pour garder seulement une moyenne par pays. Les pays n'ayant pas de code de pays ont été supprimés : ils n'étaient pas reconnus comme étant des pays. Les datasets ont été normalisés de sorte à ce que les pays aient le même nom que dans le dataset Géographie. Pour obtenir les régions, nous avons joint le dataset Géographie avec un autre dataset qui, lui, contenait ces informations. Certains pays n'étaient associés à aucune région, nous ajouté manuellement la région correspondante à l'aide d'un notebook. Pour obtenir les coordonnées GPS des régions et des continents, nous les avons également ajoutées à l'aide d'un notebook. Les coordonnées GPS des pays ont quant à elles été générées par Geocoder.

Voici le schéma en étoile correspondant à nos données :



III - La dépression et l'anxiété dans le monde

Les datasets sur la dépression et l'anxiété ne sont au départ pas labellisés. Ils proviennent chacun de la jointure de deux autres datasets qui listent la part de la population par pays atteinte de dépression ou d'anxiété selon le sexe (7 colonnes dont 2 sur les tranches d'âge), puis selon la tranche d'âge (15 colonnes dont 12 sur les tranches d'âge). Ces données étaient au départ données pour chaque pays sur plusieurs années; nous avons décidé de les agréger (moyenne) et nous passons ainsi de plus de 10 000 lignes à seulement 137 (une ligne par pays). Afin de pouvoir également accéder à ces informations à des granularités moins fines (échelle des régions ou des continents), nous faisons également la jointure de ces 2 datasets avec un dataset de géographie. Nous souhaitons maintenant mettre en évidence :

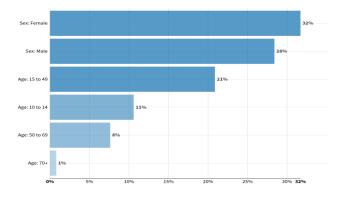
- les pays où l'anxiété (respectivement la dépression) est important
- les pays où l'anxiété (respectivement la dépression) est relativement faible

Pour cela, nous avons essayé d'entraîner plusieurs algorithmes de clustering sur nos données : la mixture gaussienne, le K-means (classique ou mini-batch), le clustering spectral et le regroupement hiérarchique. Pour chaque algorithme, les performances ont été testées en prenant comme métrique de référence la silhouette, avec et sans réduction de dimension (ACP) en gardant au moins 90% de variance. La réduction de dimension permet de réduire le nombre de variables en regroupant toutes les variables corrélées. Dans notre cas, c'est le K-Means avec une réduction ACP qui a été le plus performant : nous l'avons ainsi utilisé sur nos données pour construire les clusters.

Voici les résultats des différents algorithmes :

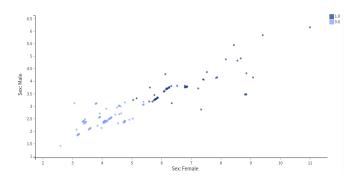
Gaussian Mixture (k=2)	0.496	☆
■ KMeans (k=2)	0.564	☆
■ KMeans (k=2)	? 0.598	☆
Spectral clustering	0.567	☆
Spectral clustering	0.559	☆
■ Mini-batch KMeans (k=2)	? 0.598	☆
Mini-batch KMeans (k=2)	0.564	☆
Agglomerative Clustering (K=2)	0.576	; ☆
Agglomerative Clustering (K=2)	0.545	☆
Gaussian Mixture (k=2)	0.473	☆

En appliquant l'algorithme de clustering sur nos données, 65% d'entre elles se retrouvent dans le cluster 1 qui correspond à l'ensemble des pays où l'anxiété est présente, et les 35% restants se retrouvent dans le cluster 0, qui correspond aux pays où elle est faiblement présente. L'ACP a également montré que les décisions sur les clusters se basaient principalement sur les données sur le sexe.



L'importance de chaque variable dans la prédiction des clusters pour l'anxiété

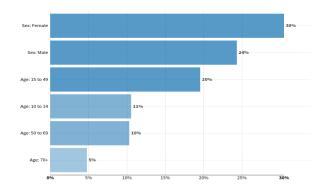
Nous le voyons très bien sur le graphique suivant : le cluster prédit est 0 lorsque le pourcentage des hommes et des femmes atteint d'anxiété est faible, et 1 lorsque le pourcentage est plus élevé.



Le pourcentage d'hommes et de femmes atteints d'anxiété

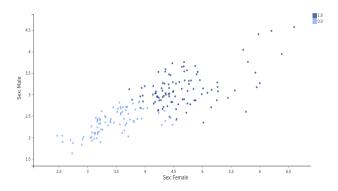
Nous pouvons remarquer que le cluster 1 correspond majoritairement à des pays d'Amérique du Nord et d'Europe de l'Ouest. Dans le cluster 0, nous retrouvons principalement les pays d'Afrique du Nord, d'Europe de l'Ouest, d'Asie du Sud, d'Amérique du Nord, Centrale et du Sud-Est.

La même stratégie de clustering a été appliquée pour les données sur la dépression. Comme pour l'anxiété, pour la dépression, ce sont les variables *Female* et *Male* qui contiennent la majorité des informations.



L'importance de chaque variable dans la prédiction des clusters pour l'anxiété

Le cluster 1 correspond toujours aux pays qui sont beaucoup touchés par la dépression, et le cluster 0, aux pays qui le sont moins.



Le pourcentage d'hommes et de femmes atteints de dépression

Comme pour l'anxiété, l'Amérique du Nord, l'Amérique Centrale, l'Europe de l'Ouest et l'Asie du Sud sont les régions les plus touchées par la dépression. La dépression touche également une très grande partie de l'Afrique.

IV - Comment mon environnement affecte-t-il ma santé mentale?

A) Liberté

L'article 1 de la Déclaration des droits de l'Homme et du citoyen stipule que "Les Hommes naissent et demeurent libres et égaux en droits". Plus globalement, la liberté représente la possibilité de penser et d'agir selon sa propre volonté, sans contrainte extérieure.

Nos données sur la liberté proviennent d'un dataset sur 123 colonnes fournissant pour chaque pays (162 en tout) et région du monde un index allant de 0 à 10 sur 116 critères de liberté (par exemple la liberté d'association ou la liberté pour une femme d'hériter des biens de sa famille entre autres). Un indice proche de 10 indiquait un bon respect de la liberté sur un critère. Nous avons choisi de ne sélectionner que 5 critères: les index liés à l'équité en justice, à la liberté de religion, à la liberté d'expression, au contrôle du gouvernement sur les médias et sur l'accès à internet. Bien entendu, ce dataset a été joint au dataset sur la géographie afin de fournir également les informations sur les continents.

Tout d'abord, nous pouvons relever que tous les critères de liberté ne sont pas significatifs : une mauvaise justice procédurale semble être un facteur déterminant pour discriminer les régions peu ou moyennement développées où les maladies mentales sont très présentes, tandis qu'un contrôle des médias et de l'internet par le gouvernement ne semble pas affecter la santé mentale d'une population. Les régions du monde où la justice est la moins présente sont l'Afrique et l'Amérique centrale : deux zones modestes ou peu développées du globe où l'on avait déjà remarqué une forte présence de dépression et d'anxiété. Il semblerait que le développement de troubles mentaux soit implicitement lié à l'absence de liberté dans certaines régions pauvres du monde. Ainsi, le critère de liberté, et en particulier celui de l'équité face à la justice, ne semble pas applicable sur les régions riches du monde.

B) Sécurité

Nos données proviennent du même dataset qui a fourni les informations sur les indices de liberté de chaque pays/région du monde, mais nous avons cette fois-ci sélectionné 5 critères différents afin de générer un nouveau dataset sur la sécurité: les index liés au nombre d'homicides, de disparitions, de morts ou blessés par attaques terroristes. A ce dataset vient s'ajouter un autre exclusivement dédié à proportion dans chaque pays de femmes victimes de violences physiques ou sexuelles. Ces datasets ont été joints au dataset sur la géographie afin de fournir les mêmes informations à l'échelle des continents. Les régions du monde les moins sûres sont l'Amérique Centrale, l'Afrique et l'Asie du Sud. Nous remarquons que les parties du monde où les crimes sont omniprésents sont également des régions où le taux d'anxiété et de dépression est fort. Cependant, le critère seul de sécurité ne suffit pas à englober l'intégrité des pays où ces troubles sont prédominants. Ce constat semble vrai pour les régions économiquement pauvres ou modestes, comme l'Amérique Centrale et l'Afrique du Nord, ou politiquement instable, tandis que les pays riches et développés ne semblent pas être exposés à ce manque de sécurité. La sécurité est donc bien corrélée au développement de troubles mentaux sans en être un facteur déterminant.

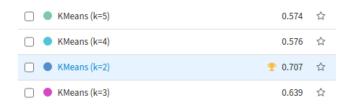
Violence envers les femmes

La violence physique et sexuelle est surtout présente en Afrique Subsaharienne, en Asie du Sud et en Océanie. Or, la dépression et l'anxiété ne sont pas prédominantes dans ces régions du monde. À première vue, on pourrait penser que les violences envers les femmes ne semblent pas être corrélées au risque de développer une maladie mentale de type dépression ou anxiété. Il faut cependant bien prendre en considération le fait que même si le taux de violences est plus faible dans les autres régions du monde comparé à l'Afrique, cette violence est toujours présente. Par exemple, en France, "En 2019, selon l'étude nationale relative aux morts violentes au sein du couple, le nombre de féminicides a augmenté de 21 % en France avec 146 femmes tuées par leur partenaire ou ex-partenaire 84 % des victimes de mort violente au sein d'un couple sont des femmes. Le collectif Féminicides par compagnon ou ex compte lui 152 cas", d'après la page Wikipédia "Féminicide". Il est ainsi compliqué d'utiliser cette carte pour faire un lien entre les violences envers les femmes et la santé mentale.

V - Ma santé physique a-t-elle un impact sur ma santé mentale?

Nous partons d'un dataset listant pour chaque pays le nombre de morts liées à 35 causes différentes (manque d'eau potable, manque d'hygiène, etc...). Ces données étant fournies pour plusieurs années, et les années étant différentes d'un pays à l'autre, nous avons simplement décidé de les agréger (moyenne). Nous nous retrouvons ainsi avec 231 lignes, chacune correspondant à un pays. Nous avons également choisi de ne garder que 15 parmi les causes de mort possibles, que nous agrégeons entre elles afin de créer un dataset sur la santé physique est associé à 4 critères : l'environnement, l'addiction, la condition physique et la nutrition. Nous joignons ce dataset au dataset sur la géographie ainsi qu'à un dataset contenant la population de chaque pays: nous souhaitons ainsi pondérer le nombre de morts par pays à sa population et ne garder qu'une proportion. Nous allons analyser chaque critère indépendamment. Nous rappelons que c'est le nombre de morts moyens qui sera représenté. Comme

pour l'anxiété et la dépression, nous allons essayer de séparer nos données en plusieurs clusters. Nous allons développer la partie machine learning sur l'addiction, mais la même stratégie a été adoptée pour les 3 autres critères. Contrairement au clustering sur l'anxiété et la dépression, nous n'allons pas fixer le nombre de clusters à 2 : nous allons essayer d'optimiser le nombre de clusters en utilisant K-means comme algorithme.



Performance de K-means sur plusieurs tailles de cluster avec la silhouette comme métrique

Nous avons ainsi décidé que nous allions séparer nos données en deux clusters. Une analyse rapide des labels nous informe que le cluster de label 1 correspond aux pays où le nombre de morts dues à une addiction est considéré comme élevé, on y trouve environ 20% des pays.

Il s'avère que pour toutes les catégories, le nombre de clusters qui maximisent la silhouette est 2. Ainsi, comme pour les morts dues à des addictions, nous séparons les données en deux classes pour les autres critères..

Dans cette partie, nous aimerions savoir s'il existe une corrélation entre la santé physique et la santé mentale. Nous avons été amenées à nous demander si un trouble mental pourrait engendrer de mauvaises habitudes quotidiennes qui affecteraient la santé physique, ou inversement, si de mauvaises conditions physiques peuvent affecter la santé mentale. Dans les différents graphiques, nous observons qu'en Afrique et en Europe de l'Est, l'environnement est la cause d'une partie non négligeable des morts. Les morts dues aux addictions, à une mauvaise alimentation et à de mauvaises conditions physiques se situent principalement en Europe de l'Est et en Amérique du Nord. Ces observations ne nous permettent pas de déceler une quelconque corrélation.

VI - Les aides médicales proposées dans ma zone géographique suffisent-elles pour m'accompagner ?

Cette problématique pose la question de l'accès aux soins dans chaque pays. Nous avons décidé d'y répondre en se basant sur trois axes d'analyses: les dépenses liées à la santé, l'accès à des aménagements médicaux et enfin l'accès aux services de professionnels de la santé.

A) Dépenses liées à la santé

Ce critère d'analyse fait appel à deux datasets: un dataset fournissant pour chaque pays le pourcentage de la population couverte par une assurance santé, l'autre indiquant pour chaque pays la part du PIB réservé aux dépenses en santé. Ces datasets sont joints à celui sur la géographie.

Ce sont dans les régions riches et développées que les dépenses de l'Etat sont les plus conséquentes en matière de Santé. Cela reflète les moyens financiers de ces régions mais peut-être aussi un besoin particulier de dispositifs médicaux pour leur population. Par ailleurs, ce n'est pas parce qu'une population est bien couverte par une assurance ou que le gouvernement dépense une grosse partie de son PIB dans le secteur de la santé que les risques de développer de l'anxiété ou d'être atteint de dépression sont plus faibles : les consultations chez le psychologues ne sont peut-être pas pris en charge. Les régions où le moins de foyers dépensent plus de 10% de leur revenu dans la santé sont l'Asie et l'Afrique, qui sont aussi les régions où les dépenses de l'Etat en matière de santé sont les plus faibles.

B) Ressources humaines : Professionnels de santé

Nous nous intéressons maintenant aux services humains mis à disposition par chaque pays pour leur population en matière de santé. Pour analyser ce critère que nous appellerons "ressources humaines" et qui fait référence aux professionnels de la santé, nous partons de 2 datasets différents, qui donnent respectivement pour chaque pays le nombre de médecins disponibles pour 10 000 habitants et le nombre de psychiatres, psychologues, infirmières et employés sociaux pour 100 000 habitants. Nous décidons de ne garder que les informations sur les docteurs, psychiatres et psychologues que nous mettons à une échelle de 100 000 habitants. Ce dataset final est joint à la géographie.

L'analyse de ces datasets nous permet de nous rendre compte que les pays riches et développés qui sont très touchés par la dépression et l'anxiété sont ceux qui mettent le plus de moyens pour lutter contre les maladies mentales: l'Amérique du Cenrale et du Nord et l'Europe de l'Ouest, qui sont les pays riches les plus touchés, arrivent en tête du nombre de médecins et psychologues pour 100 000 habitants, tandis que l'Afrique du Nord, qui est tout aussi touchée, en fournit aussi peu que des régions développées du monde où l'anxiété et la dépression sont moins présentes telles que l'Asie de l'Est. Ainsi nous émettons la possibilité que parmi les régions développées du monde, celles qui mettent le plus grand nombre de professionnels de santé (et en particulier des professionnels en troubles mentaux) sont les pays où les maladies mentales sont les moins rares.

C) Aménagements médicaux : Hôpitaux psychiatriques

Nous nous sommes intéressés aux ressources humaines spécialisées dans la santé mises à disposition pour chaque pays et avons montré qu'il existait bien une corrélation entre leur présence et le risque de troubles mentaux dans les pays développés seulement. Qu'en est-t-il maintenant des aménagements médicaux ? Nous partons d'un dataset indiquant le nombre de 7 types d'établissements médicaux pour 100 000 habitants dans différents pays du monde. Comme il manque pour certains pays les données sur quelques types d'établissement mais que les données remplies sont en nombre assez

conséquent, nous nous proposons de prédire les données manquantes par différents algorithmes d'apprentissage supervisé (Régression linéaire L2, SVM, forêts aléatoires, SGD et arbres de décision). L'arbre de décision nous permettant d'avoir le plus petit coût aux moindres carrés, nous décidons d'entraîner notre modèle avec cet algorithme et remplissons les cellules vides. Enfin, nous ne gardons que les informations sur le nombre d'hôpitaux mentaux pour 100 000 habitants que nous joignons au dataset sur la géographie afin de pouvoir accéder à ces informations à différentes échelles de granularité (région et continent en plus du pays).

Comme pour les professionnels de santé, les pays développés et touchés par l'anxiété et la dépression mettent à disposition de leur population beaucoup plus d'établissements spécialisés pour lutter contre les maladies mentales que les pays pauvres touchés. Une observation qui nous a surprises est le fait que le Japon et la Corée du Sud font également partie de ces pays où l'on retrouve beaucoup d'hôpitaux mentaux, ce qui reflèterait un besoin de dispositifs dans le secteur de la santé mentale, et pourtant très peu de personnes sont diagnostiquées comme atteintes d'anxiété ou de dépression comme nous l'avons vu en analysant les régions touchées par ces troubles mentaux... Nous émettons l'hypothèse, que nous devrons vérifier, que la santé mentale est très probablement un tabou dans les pays d'Asie de l'Est, c'est pourquoi peu de personnes déclareraient en être atteintes ou iraient se faire diagnostiquer.

VII - Quel est l'impact du travail sur mon bien-être ?

Dans cette dernière partie, nous allons nous concentrer sur les conséquences du nombre d'heures travaillées sur la santé mentale des employées. Pour mesurer la santé mentale des employées, nous allons nous focaliser sur le burnout et la fatigue mentale. Le burn out est défini comme un état de fatigue intense et de grande détresse causé par le stress au travail. La fatigue mentale est souvent la conséquence d'une vie un peu trop agitée : excès de stress, surmenage, difficulté à concilier vie professionnelle et vie familiale par exemple.

Notre analyse se base sur un dataset indiquant, pour une certain nombre d'heures travaillées par jours et pour chaque sexe, le score de fatigue mentale ressentie par un employé ainsi que son taux de saturation. Ce dataset a été créé en se basant sur un dataset fournissant pour 22 750 actifs des informations sur leur vie active (date d'arrivée dans leur compagnie, type de compagnie entre autres). Nous nous intéressons en particuliers aux colonnes correspondant au nombre d'heures travaillées pour chaque employé ainsi que leur taux de saturation et leur score de fatigue. Puisque nous souhaitons au final obtenir pour chaque nombre d'heures travaillées un seul score de fatigue et un seul taux de saturation, nous moyennons chaque critère sur chaque tranche d'heures. Nous pouvons ainsi créer un diagramme du score de fatigue mentale ou du taux de saturation en fonction du nombre d'heures travaillées.

Les diagrammes, sur le burn out et la fatigue mentale, montrent clairement un corrélation entre le nombre d'heures travaillées par jour d'un employé avec le burn out et la fatigue mentale : plus un homme, ou une femme, travaille dans une journée, plus cette personne est susceptible de subir un burn out ou une fatigue mentale. Ainsi, par extension, le nombre d'heures travaillées a également un impact direct sur la santé mentale d'un employé. La carte suivante montre le nombre d'heures travaillées en moyenne par an

dans le monde. Ces données proviennent de Wikipédia, elles ont été récupérées par web scraping. Nous les affichons ensuite sur une carte du monde.



On peut par exemple retrouver la Corée du Sud qui travaille un peu plus de 2 000 heures par an, alors qu'on a trouvé aucun signe de dépression ou d'anxiété dans les analyses précédentes. Cela peut nous permettre de prendre du recul par rapport aux données : chaque pays du monde ne parle pas de l'anxiété et de la dépression de manière ouverte. Dans certains pays, certaines maladies mentales, voire toutes les maladies mentales peuvent être considérées comme tabou. Ainsi les personnes peuvent avoir honte de s'identifier comme

dépressive, par exemple, ou même de se permettre de se l'avouer.

VIII - Conclusion

Pour conclure, les zones où le risque de développer une maladie mentale est Europe de l'Ouest, en Afrique du Nord, en Amérique du Nord et Centrale et en Asie du Sud. Ainsi, vivre dans ces régions du monde augmente la probabilité d'être atteint d'anxiété ou de dépression. Bien que les différentes analyses ont montré qu'il n'existait aucun critère absolu, ces mêmes analyses ont fait ressortir certains points. Dans les pays peu développés ou en développement, l'anxiété et la dépression sont causées par les conflits politiques ou militaires et le manque de sécurité et de liberté. Pour les pays de l'Asie du Sud Est, c'est la saturation au travail la raison principale des développements de maladies mentales. Concernant les pays développés, aucun critère n'a été relevé dans ce projet, mais la mise en évidence des ressources mise en place par les pays démontrent un intérêt pour l'amélioration des conditions mentales des habitants. De plus, la santé physique ne joue pas de rôle direct dans le développement de troubles mentaux. Pour finir, les régions les mieux aptes, que ce soit en personnels soignants ou en structures, pour traiter une personne anxieuse ou dépressive est en Amérique du Nord et en Europe de l'Ouest.