

# MACHINE LEARNING :

Un outil de détection de la dépression chez  
les soignants pendant une crise sanitaire



# Table des matières

Introduction

I- Comprendre le problème : l'histoire derrière nos données

II- Choisir les bons outils : qu'est ce qu'on optimise et comment ?

III- Quel algorithme pour résoudre ce problème ?

IV- Améliorer : peut-on réduire le nombre de variables du modèle ?

Conclusion

# Introduction

Pendant la pandémie de Covid-19 :  
**augmentation plus importante des troubles mentaux chez les soignants** (40 %-50 %) que dans la population générale.

En 2020, 56% des soignants français montraient des signes de **détresse psychologique**, 21% des troubles de **stress post-traumatique**

(Données Santé Publique France)

D'où l'importance de prédire et prévenir ces troubles mentaux lors de futures crises sanitaires, à l'aide du machine learning



# I- Comprendre le problème : l'histoire derrière nos données

- Étude réalisée sur des **professionnel.le.s de la santé** d'Asie Centrale entre juillet et novembre 2022
- **2685** réponses à l'enquête
- Types de données :

**Données socio-démographiques** et professionnelles : sexe, âge, emploi (médecins ou infirmières), situation familiale, antécédents de travail en première ligne pendant la pandémie de COVID-19

**Données de santé mentale** : prévalence, niveau et gravité de la dépression, de l'anxiété et du stress, calculés à partir du questionnaire DASS-21

**Impact du COVID-19** sur la vie personnelle et professionnelle

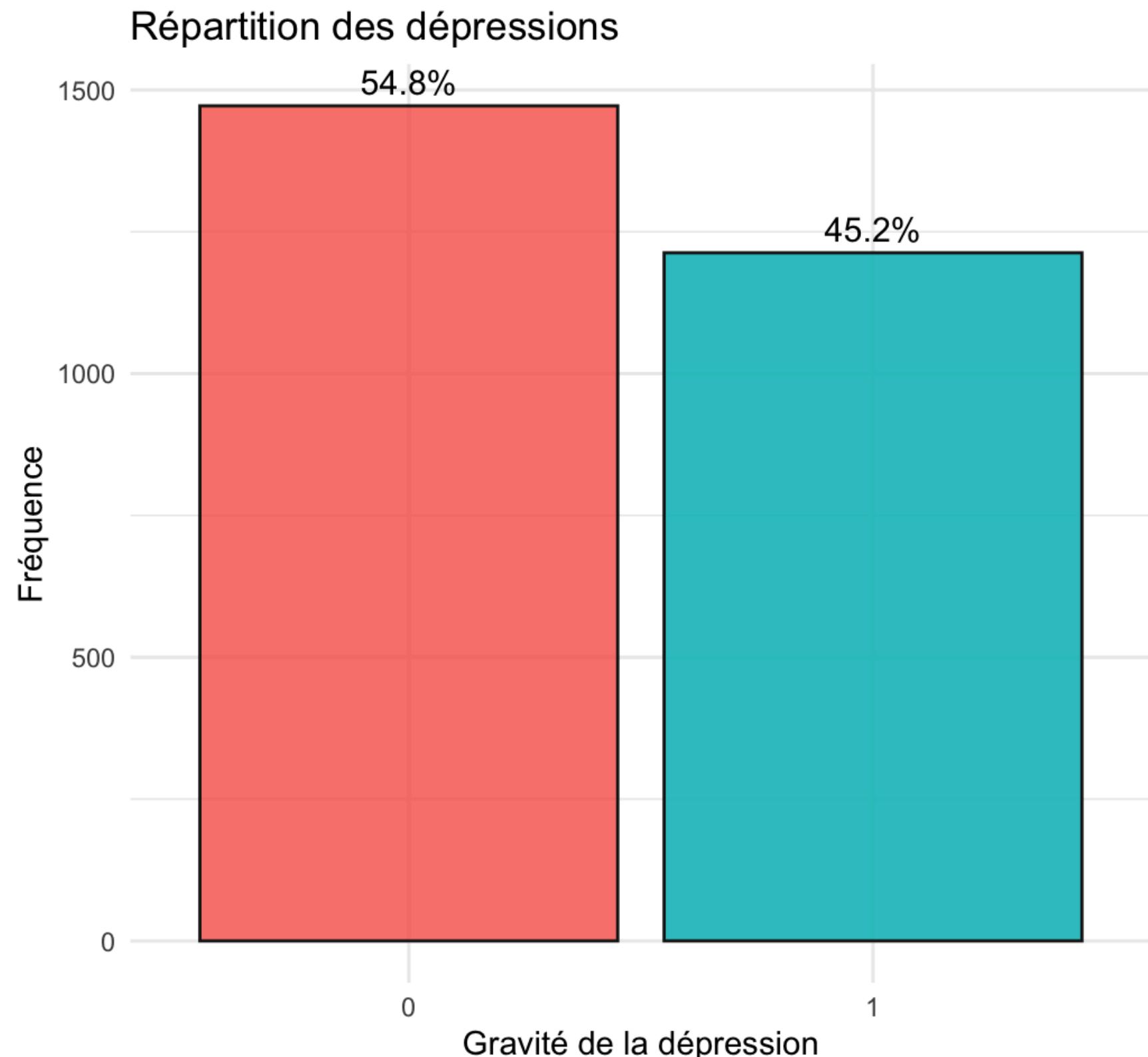
## II- Choisir les bons outils : qu'est ce que l'on optimise et comment ?

2 mesures dans le jeu de données :

- **score** au DASS-21 :
  - variable **quantitative**
  - de 0 à 21
- **sévérité** de la dépression
  - variable **qualitative**
  - de 0 (pas de dépression) à 4 (dépression très sévère)

On binarise la variable de sévérité :

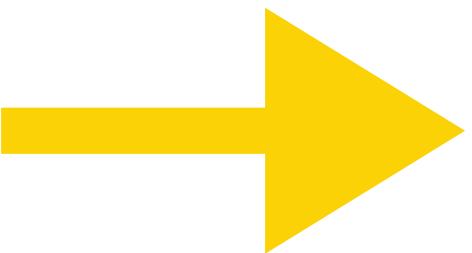
- sévérité de 0 : classe 0
- sévérité  $\geq 1$  : classe 1 → **Il faut surveiller ce patient**



## II- Choisir les bons outils : qu'est ce que l'on optimise et comment ?

On teste **5 méthodes** de prédictions différentes :

- Régression linéaire multiple (GLM)
- Régression linéaire multiple pénalisée (GLMnet)
- Les K plus proches voisins (KNN)
- Forêts aléatoires (RF)
- Support-Vector-Machine (SVM)



Laquelle est la meilleure ?

- 1) Test de plusieurs hyperparamètres, dont le **seuil de décision**
- 2) Comparaison des **F1-scores**, particulièrement adaptés aux données déséquilibrées

## II- Choisir les bons outils : qu'est ce que l'on optimise et comment ?

### Démarche générale

- 1 ) Division des données : 80% apprentissage, 20% test
- 2) **Optimisation des hyperparamètres** (validation croisée 10 plis, maximisation du F1-score) sur les données d'apprentissage
- 3) **Évaluation des performances** du meilleur modèle sur les données de test

## II- Choisir les bons outils : qu'est ce que l'on optimise et comment ?

	Hyperparamètres	Plage de valeurs
GLM	Aucun	
KNN	k (nb voisins)	1 à 50 (entiers)
GLMnet	lambda (régularisation) alpha (mix Ridge/Lasso)	1e-4 à 1e-1 (5 valeurs) 0 à 1 (5 valeurs)
SVM	C (régularisation)	0.1 à 2 (5 valeurs)
Random Forest	mtry (nb variables/split) min.node.size	2 à 10 (entiers) 2 à 10 (entiers)

+ **seuil** en  
hyperparamètre

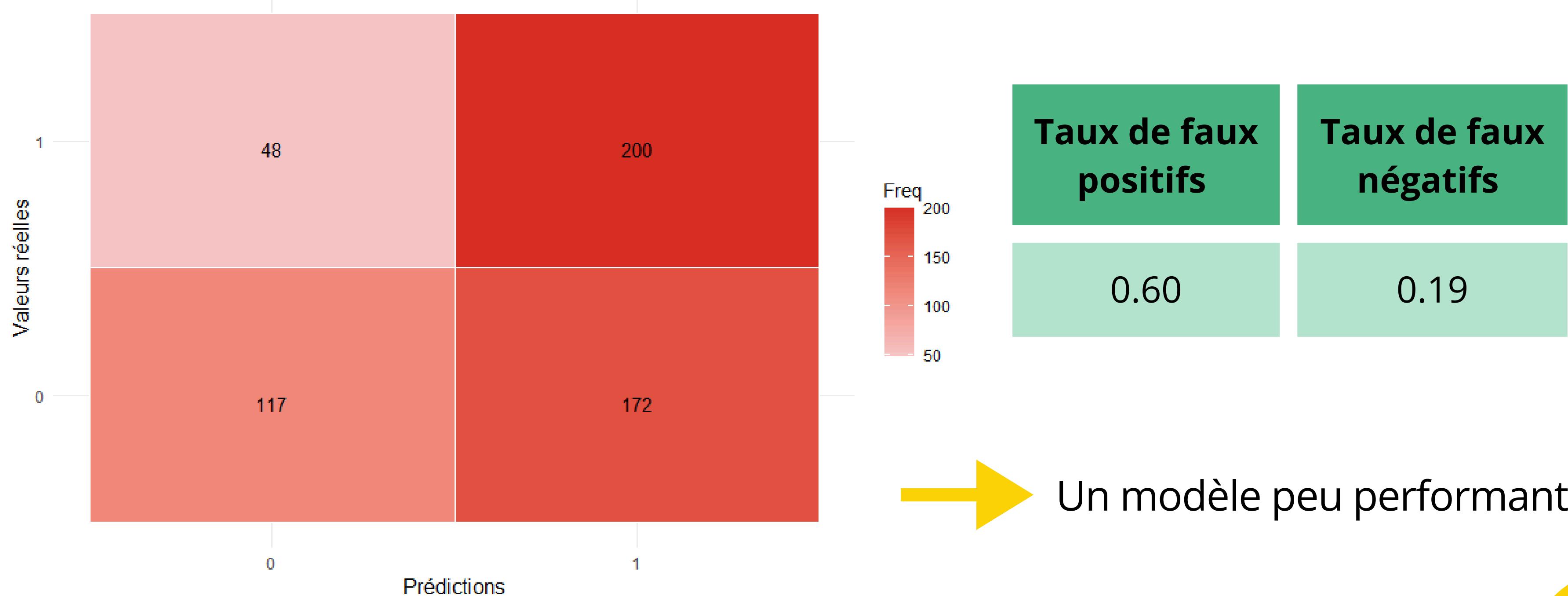
Plage de valeurs :  
0.05 à 0.5 (10 valeurs)

# III- Quel algorithme pour résoudre ce problème ?

	GLMnet	GLMnet	GLMnet	GLMnet	GLMnet
Hyper paramètres	Seuil : 0.35 Alpha : 0.25 Lambda : 0.050050	Seuil : 0.35 Alpha : 0.50 Lambda : 0.025075	Seuil : 0.35 Alpha : 0.75 Lambda : 0.025075	Seuil : 0.35 Alpha : 0.25 Lambda : 0.025075	Seuil : 0.35 Alpha : 1.00 Lambda : 0.025075
F1-score	0.6620968	0.6582278	0.6575672	0.6574885	0.6569630

# III- Quel algorithme pour résoudre ce problème ?

Matrice de confusion GLMnet

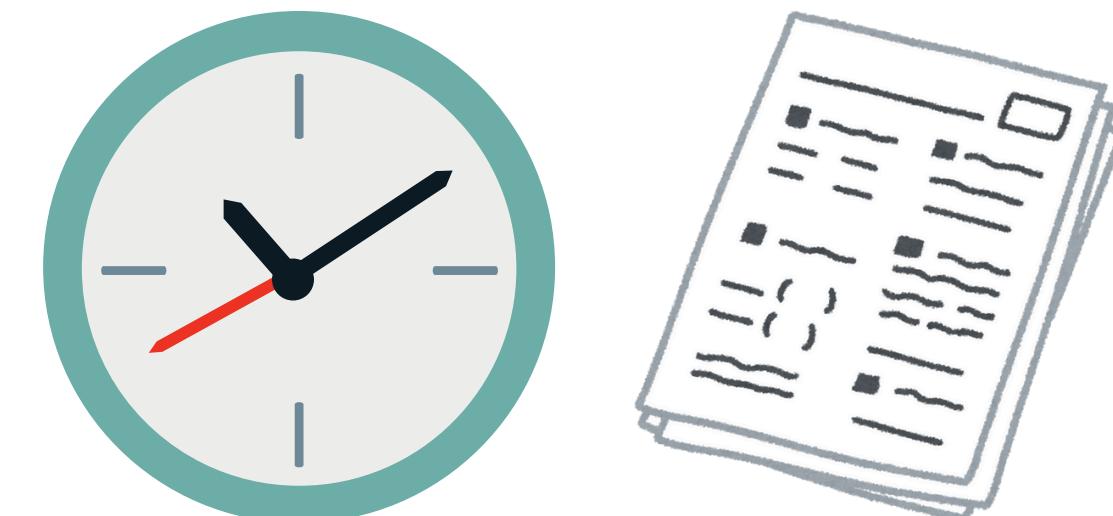


## IV- Améliorer : peut-on réduire le nombre de variables du modèle ?

Dans la réalité du terrain : il faudrait avoir accès au **minimum de données** possible sur le soignant (par exemple à travers un **questionnaire rapide**) pour **déetecter une dépression éventuelle**.



De 21 questions



À ? questions

# IV- Améliorer : peut-on réduire le nombre de variables du modèle ?

```
Depression_severity ~ Gender + Age + Job + Manager_position +  
COVID_frontline + Family_status + Children + more_work +  
additional_workload + overtime + work_stress + work_conflicts +  
afraid_family + people_avoid_me + afraid_others + people_avoid_family +  
working_attitude + insufficient_employees + appreciation_employer +  
appreciation_society + appreciation_govt
```

**Modèle  
complet  
GLM**

**Dans le modèle réduit GLMNet, les variables suivantes ont été supprimées :**

Soignant en première ligne

Statut familial

Travail supplémentaire

J'évite de parler aux autres de la nature de mon travail

Manque de personnel

# IV- Améliorer : peut-on réduire le nombre de variables du modèle ?

Il existe d'autres modèles moins performants au niveau du F1-Score, mais qui suppriment 2 à 3 variables par rapport au meilleur modèle.

## Variables du modèle GLMNet

Genre  
Heures supplémentaires  
Peur de dire à la famille les risques liés aux Covid  
On m'évite à cause de mon travail  
On évite ma famille à cause de mon travail  
J'évite de parler de mon travail  
Manque de reconnaissance par le gouvernement

## Variables du modèle de l'article revu

Pays  
Statut marital

Tableau: Variables existantes dans notre modèle qui n'existent pas dans celui de l'article dont est issu le jeu de données, et inversement

# Conclusion

Notre modèle laisse assez **peu de personnes dépressives sous les radars**, 2 personnes sur 10 dépressives ne seront pas diagnostiquées correctement.

Cependant, notre modèle a tendance à **prédir plus** de personnes dépressives **qu'il y a en réalité**.



Notre modèle se place **en complément d'un diagnostic plus approfondi**, le plus important étant de détecter précocement les personnes atteintes de maladies mentales.

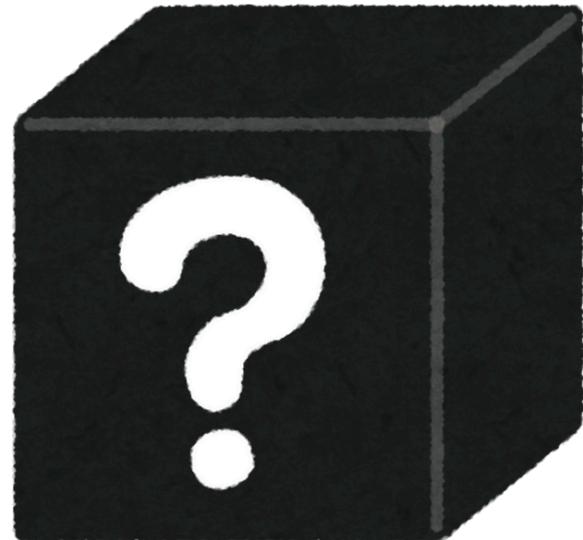
# Conclusion

**Une analyse des coûts pourrait être intéressante**, entre les coûts engendrés par les personnes non dépressives diagnostiquées qui font des examens plus poussés, et les bénéfices réalisés sur une prise en charge plus précoce des malades grâce au diagnostic.

Notre modèle, qui s'appuie initialement sur le DASS-21, **ne prédit pas exactement la probabilité d'être en dépression, mais plutôt de traverser un épisode dépressif**. Il existe en effet d'autres maladies mentales qui sont à l'origine de symptômes dépressifs, comme le trouble bipolaire.



# Conclusion



Les épisodes dépressifs ont des origines complexes et multi-causes, avec **des facteurs qui n'ont pas pu être évalués** pendant cette étude.

**Manque de variables importantes** telles que :

Facteurs environnementaux

Habitudes de vie

État émotionnel préexistant

Génétique

# Conclusion

**Manque de variables importantes** telles que :

Facteurs environnementaux

Habitudes de vie

État émotionnel préexistant

Génétique

merci