Rapport intérimaire pour évaluation formative

STT-7335 Méthodes d'analyse de données

José Manuel Rodríguez Caballero

April 10, 2025

1 Introduction

La présente étude vise à déterminer dans quelle mesure on peut distinguer les poètes suicidaires des poètes non suicidaires, à partir de différentes variables démographiques (par exemple le sexe ou l'orientation sexuelle) et d'indicateurs émotionnels (colère, joie, confiance, etc.). Dans ce cadre, nous avons constitué un jeu de données comprenant 16 poètes, 3 poèmes par poète et 3443 vers, chacun associé à plusieurs observations des scores émotionnels. Des techniques statistiques variées — allant du rééchantillonnage *bootstrap* à la réduction de dimension par *Analyse en Composantes Principales* (ACP), en passant par des algorithmes de classification (régression logistique, XGBoost, mélanges gaussiens) — ont été mobilisées pour explorer la robustesse de la distinction suicidaire *vs.* non suicidaire. L'objectif global est d'évaluer la fiabilité de ces approches, tant sur le plan méthodologique (gestion de données manquantes, validation *leave-one-poet-out*) que sur le plan pratique (comparaison des performances et limites inhérentes au faible effectif).

2 Description du jeu de données

Le présent travail repose sur un corpus de 3443 vers annotés, chacun étant associé à un certain nombre d'informations concernant l'auteur ou l'autrice (*poète*) et à plusieurs scores émotionnels. Les variables collectées couvrent différents aspects : caractéristiques biographiques (identifiant unique, nom, sexe, date de naissance, date de décès, etc.), période d'écriture, indicateurs booléens (orientation sexuelle, statut suicidaire), ainsi que des mesures quantitatives telles que le nombre de mots par vers ou des scores de sentiment et d'émotion (*colère*, *peur*, *joie*, etc.).

2.1 Structure des variables

2.1.1 Niveau : poète

Le nombre total de poètes est de 16.

- **poet_id** : Identifiant unique propre à chaque poète.
- **poet** : Nom de l'auteur ou de l'autrice.
- **suicidal** : Variable binaire (indicateur booléen) précisant si le poète s'est effectivement suicidé (`TRUE`) ou non (`FALSE`). **Il s'agit de la variable réponse principale** de la présente étude.
- **sex** : Sexe du poète (`Male` ou `Female`).
- **heterosexual** : Variable binaire (`TRUE` pour un poète hétérosexuel, `FALSE` dans les autres cas).
- **date_of_birth** : Date de naissance de l'auteur ou de l'autrice.
- **date_of_death** : Date de décès (si disponible).
- **country_of_birth** : Pays de naissance.

2.1.2 Niveau : poème

Pour chaque poète, il y a trois poèmes, chacun correspondant à trois périodes.

- **period** : Période d'écriture du poème (variable catégorielle : *Early*, *Middle*, *Later*).
- **poem_title** : Titre du poème auquel le vers appartient.
- **n_words** : Nombre de mots présents dans le vers.

2.1.3 Niveau : vers

- **anger**, **disgust**, **fear**, **joy**, **sadness**, **surprise**, **trust**, **anticipation** : Scores numériques représentant l'intensité de diverses émotions dans le vers.
- **negative** : Score global regroupant des émotions à valence négative.
- **positive** : Score global regroupant des émotions à valence positive.
- **verse** : Position normalisée du vers dans le poème (0 pour le premier vers, 1 pour le dernier).

Provenance et préparation

Les poèmes retenus proviennent de diverses sources en ligne, leurs métadonnées (dates de naissance et de décès, pays d'origine, orientation sexuelle, statut suicidaire, etc.) ayant principalement été extraites de Wikipédia et recherches en Google. Pour l'évaluation des émotions, on s'est appuyé sur un module d'analyse lexicale (*syuzhet*) permettant de calculer automatiquement des scores de sentiment dans chaque vers. Pour l'un des poètes, aucune preuve claire n'a pu être trouvée, conduisant à une valeur manquante (`NA`).

3 Imputation des données manquantes

Soit X la variable aléatoire représentant le *score moyen d'émotions positives* associé à un poète donné, c'est-à-dire la moyenne arithmétique des scores d'émotion positive pour chaque verset associé au poète donné. Soit Z la variable aléatoire binaire indiquant si ce poète est *hétérosexuel* ($Z = 1$) ou non ($Z = 0$). Dans l'échantillon, une seule valeur de Z est manquante, correspondant à la poétesse Edith Sitwell. Bien que l'ampleur de ce problème soit très réduite, il importe de disposer d'un jeu de données complet pour conduire les analyses ultérieures, notamment celles où l'on souhaite étudier la relation entre la variable suicidaire Y et les autres caractéristiques. On suppose que le méchanisme de la donnée manquante est MCAR.

3.1 Visualisation des valeurs manquantes

La Figure 1 illustre la distribution de X (score d'émotions positives) en fonction de Z (hétérosexuel ou non), en signalant l'observation pour laquelle Z est manquant par une ligne horizontale. On observe ainsi que cette observation présente un score $X \approx 0{,}17$.

Figure 1: Distribution de X (score des émotions positives) selon Z (hétérosexuel ou non). La ligne horizontale indique l'observation pour laquelle Z est manquant.

3.2 Modélisation logistique pour l'imputation

Pour estimer la probabilité $\mathbb{P}(Z = 1 \mid X = x)$, on ajuste un modèle de régression logistique binaire sur toutes les observations complètes (celles pour lesquelles la valeur de Z est connue). Plus précisément, on suppose que :

$$\mathbb{P}(Z = 1 \mid X = x) \;=\; \frac{1}{1 + \exp\big(-(\beta_0 + \beta_1 x)\big)},$$

où β_0 et β_1 sont des paramètres inconnus à estimer par la méthode de la vraisemblance maximale.

Une fois ce modèle estimé, la probabilité d'appartenir à la classe $Z = 1$ est calculée pour la poétesse dont la donnée est manquante. Afin de produire une imputation binaire, on convertit cette probabilité en classe $\hat{z} \in \{0, 1\}$ selon la règle de décision suivante :

$$\hat{z} = \begin{cases} 1 & \text{si } \mathbb{P}(Z = 1 \mid X = x) > 0{,}5, \\ 0 & \text{sinon.} \end{cases}$$

Dans le présent cas, la probabilité prédite est supérieure à 0,5, ce qui conduit à imputer la valeur $\hat{z} = 1$ à la poétesse concernée (c'est-à-dire, à la considérer comme hétérosexuelle).

Après ce traitement, la variable Z ne contient plus aucune valeur manquante, permettant ainsi l'intégration de l'ensemble des poètes (et notamment celui pour lequel l'absence

de renseignement posait initialement problème) dans les analyses descriptives et inférentielles. En somme, cette imputation, bien qu'artificielle, facilite l'exploitation globale des données et demeure conforme aux principes de la régression logistique binaire.

4 Statistiques descriptives

4.1 Variables démographiques

Tableaux de fréquences

Les tableaux de fréquences ci-après (Table 1 et Table 2) présentent la distribution de la variable suicidaire selon le sexe (femme vs. homme) et selon la variable hétérosexuel (non vs. oui). Chaque cellule indique le nombre d'individus appartenant à la catégorie en ligne et en colonne correspondante. Ces comptes permettent d'apprécier la répartition conjointe de deux caractéristiques (p. ex. sexe et suicidaire) et d'identifier d'éventuelles différences de fréquence entre les groupes.

Table 1: Distribution de la variable suicidaire selon le sexe.

		Suicidaire	
		Non	Oui
Sexe	Femme	4	4
	Homme	4	4

Table 2: Distribution de la variable suicidaire selon la hétérosexualité.

		Suicidaire	
		Non	Oui
Hétérosexualité	Non	3	2
	Oui	5	6

Moyennes de la variable suicidaire

En complément de ces tableaux de fréquences, la proportion d'individus suicidaires a été calculé dans différents sous-groupes de la population étudiée : hommes, femmes, hétérosexuels et non-hétérosexuels. Les résultats, présentés dans le tableau 3, indiquent que :

- Le taux de suicidaires est de 50% tant chez les hommes que chez les femmes.
- Les hétérosexuels présentent un taux moyen légèrement plus élevé (environ 55%) que les non-hétérosexuels (environ 40%).

Table 3: Proportion moyenne de la variable suicidaire selon le sexe et l'hétérosexualité.

Sous-groupe	**Moyenne**
Hommes	0.50
Femmes	0.50
Hétérosexuels	0.55
Non-hétérosexuels	0.40

4.2 Variables émotionnelles

Afin d'évaluer les différences d'intensité émotionnelle entre individus «non suicidaires » et «suicidaires », dix variables ont été analysées : *colère, anticipation, dégoût, peur, joie, tristesse, surprise, confiance, négatif et positif.* Deux approches ont été mises en œuvre : l'examen direct des valeurs observées (données originales) et l'estimation des moyennes via rééchantillonnage bootstrap.

4.2.1 Approche avec données originales.

Dans un premier temps, chaque émotion a été représentée sous forme de diagramme en boîtes (boxplot) pour les groupes «non suicidaires » et «suicidaires ». Cette visualisation permet de comparer la médiane, l'étendue interquartile et la présence éventuelle de valeurs atypiques pour chacune des dix émotions, comme le montre la Figure 2.

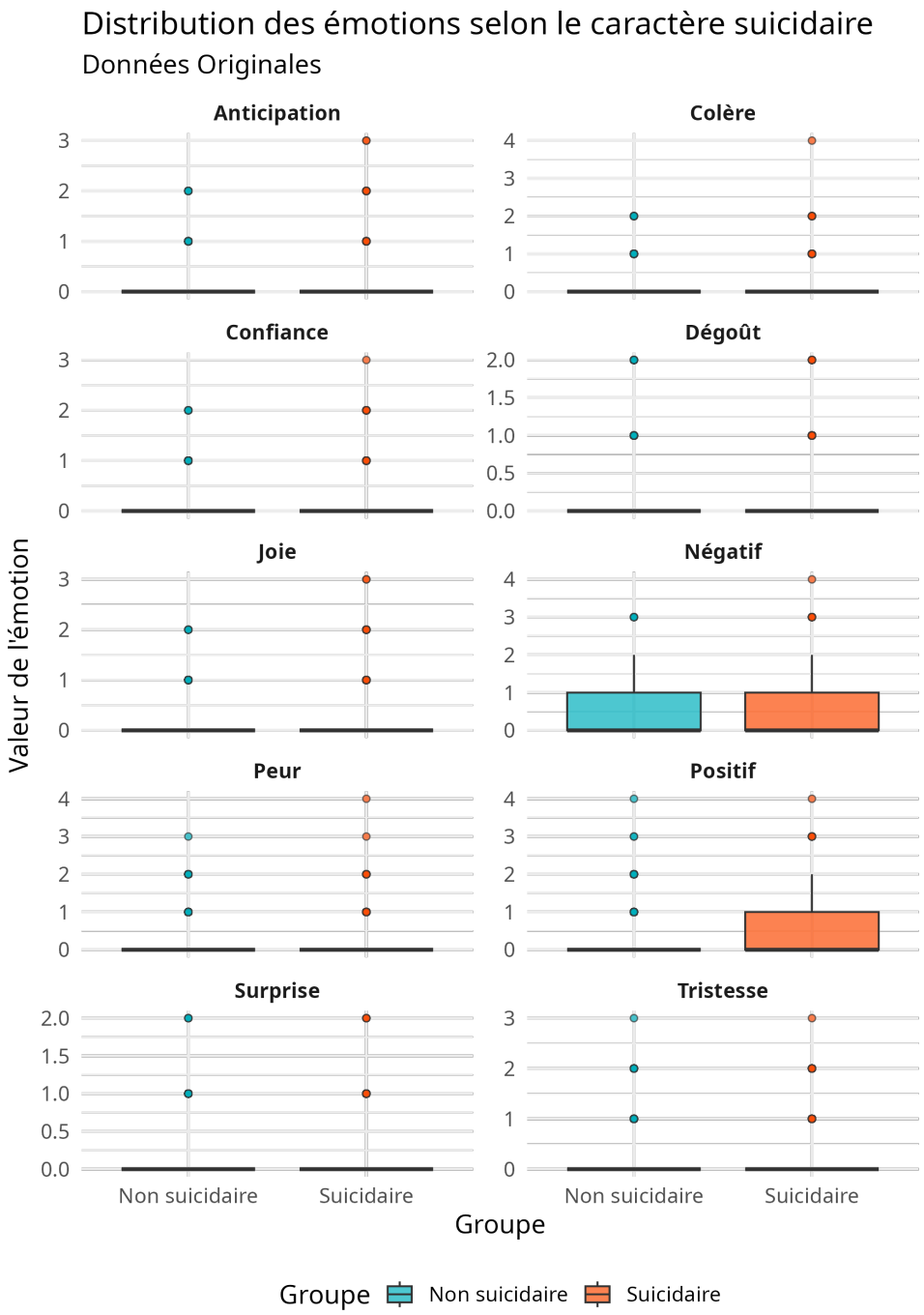

Figure 2: Distribution des émotions selon le caractère suicidaire (*données originales*).

4.2.2 Approche par rééchantillonnage Bootstrap.

Pour compléter cette analyse et apprécier la robustesse des estimations, un rééchantillonnage bootstrap a été réalisé à 10 000 reprises. À chaque itération, un échantillon de même taille que l'échantillon initial est constitué «avec remise ». Les moyennes des émotions sont alors recalculées séparément dans les deux groupes. Les distributions de ces moyennes (pour chaque émotion) sont illustrées par des diagrammes en boîtes dans la Figure 3.

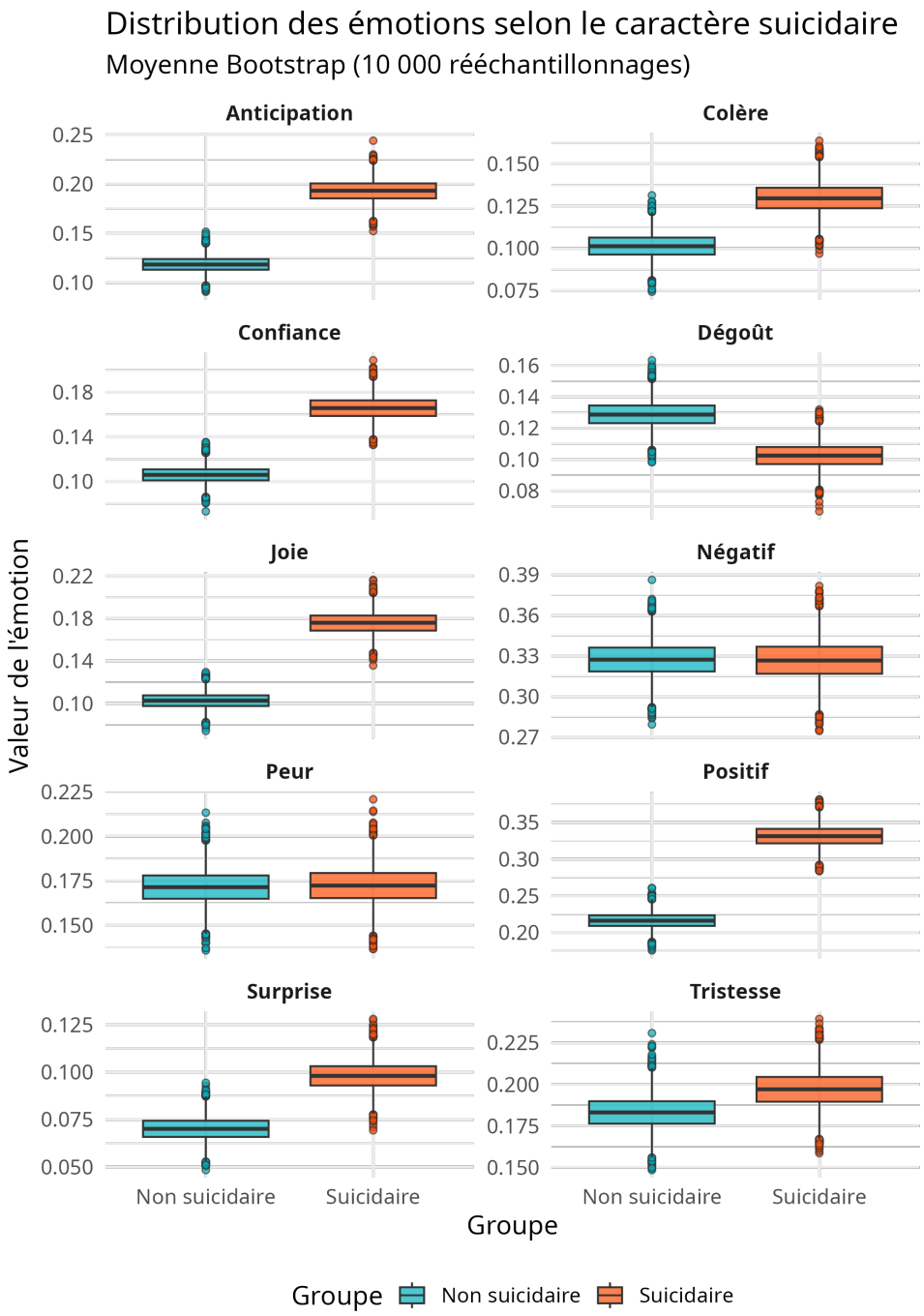

Figure 3: Distribution des émotions selon le caractère suicidaire (*moyennes bootstrapées, 10 000 rééchantillonnages*).

4.2.3 Projection des données sur les deux premiers axes (ACP)

Pour évaluer l'éventuelle différenciation globale entre poètes suicidaires et non suicidaires, une analyse en composantes principales (ACP) a été réalisée sur les variables numériques (émotions principales). La Figure 4 illustre la distribution des individus sur le premier plan factoriel (PC1 et PC2).

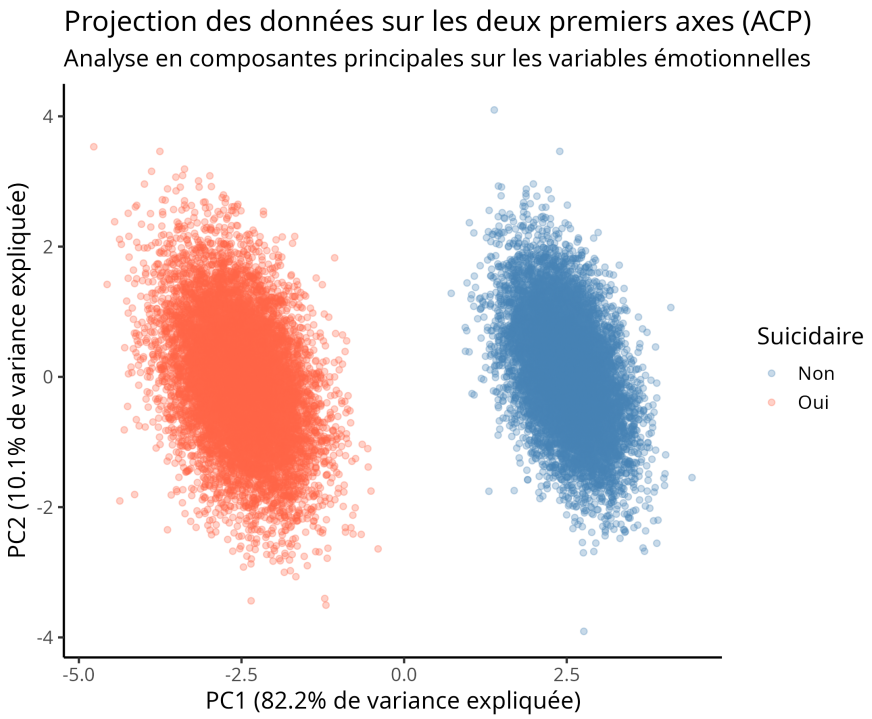

Figure 4: Projection des données sur les deux premiers axes (ACP). Les individus sont coloriés selon la variable `suicidal` (« Oui » pour suicidaire, « Non » pour non suicidaire).

5 Comparaison des trois modèles de classification

Dans cette section, nous comparons trois techniques de classification afin de prédire si un poète appartient à la classe « suicidaire » ou à la classe « non suicidaire ». Les méthodes considérées sont :

- **XGBoost**, qui repose sur un algorithme de *gradient boosting*,
- **Régression logistique**, associée à un modèle linéaire généralisé (lien logit),
- **Modèle gaussien**, où chaque classe suit une loi normale multivariée (analyse discriminante).

5.1 Procédure d'évaluation

Un ensemble de données, constitué de poètes « suicidaires » et « non suicidaires », est subdivisé en un échantillon d'apprentissage et un échantillon de test. Pour renforcer la robustesse des estimations, l'échantillon d'apprentissage est rééchantillonné de multiples fois (*bootstrap*), et chacun des trois algorithmes de classification est ajusté sur ces rééchantillonnages. Les poètes conservés hors de l'apprentissage servent alors à tester la capacité de généralisation de chaque méthode. En répétant ce processus plusieurs fois, on agrège

l'ensemble des prédictions afin d'obtenir une *matrice de confusion globale* pour chaque modèle, accompagnée des principales mesures de performance.

Pour avoir une **vue d'ensemble** des résultats, nous proposons de présenter séparément les **matrices de confusion** (une par modèle) et **les mesures de performance** (les indicateurs usuels) dans deux séries de tableaux.

Table 4: Matrice de confusion pour **XGBoost**.

		Prédit	
		FAUX	**VRAI**
Observé	**FAUX**	30	17
	VRAI	10	43

Table 5: Matrice de confusion pour la **Régression Logistique**.

		Prédit	
		FAUX	**VRAI**
Observé	**FAUX**	36	17
	VRAI	6	41

Table 6: Matrice de confusion pour le **Modèle Gaussien**.

		Prédit	
		FAUX	**VRAI**
Observé	**FAUX**	24	21
	VRAI	0	55

Les **Tables 4, 5 et 6** montrent le nombre d'occurrences pour chacune des quatre cellules (vrais négatifs, faux positifs, faux négatifs, vrais positifs) en fonction du modèle considéré. Rappelons que **FAUX** et **VRAI** désignent respectivement «non suicidaire » et «suicidaire », et que la «classe positive » est donc **VRAI**.

Table 7: Principales mesures de performance pour chacun des trois modèles.

Indicateur	**XGBoost**	**Logistique**	**Gaussien**
Exactitude (*Accuracy*)	0,73	0,77	0,79
IC à 95 % (Exactitude)	[0,632 – 0,8139]	[0,6751 – 0,8483]	[0,6971 – 0,8651]
Sensibilité (*Sensitivity*)	0,7167	0,7069	0,7237
Spécificité (*Specificity*)	0,7500	0,8571	1,0000
Valeur préd. positive (*VPP*)	0,8113	0,8723	1,0000
Valeur préd. négative (*VPN*)	0,6383	0,6792	0,5333
Kappa	0,4534	0,5444	0,557
p-value (Test de McNemar)	0,2482	0,03706	$1{,}275\times10^{-5}$

Remarque : la **classe positive** est **VRAI** (suicidaire).

6 Modélisation par mélanges gaussiens et ACP

L'objectif de ce chapitre est d'évaluer la capacité d'un **modèle de mélanges gaussiens** (`MclustDA`) à distinguer les poètes **suicidaires** (« VRAI ») des **non suicidaires** (« FAUX »), en s'appuyant sur différentes variables émotionnelles et démographiques. Pour renforcer la robustesse du modèle, on applique notamment un **rééchantillonnage bootstrap** suivi d'une **Analyse en Composantes Principales (ACP)**. Les prédictions sont finalement validées à l'aide d'une procédure « *leave-one-poet-out* ».

6.1 Protocole général

1. **Lecture et préparation des données.** Les observations portent sur 16 poètes, pour lesquels on dispose de :

   - leurs *caractéristiques émotionnelles* (par exemple `anticipation`, `colère`, `joie`, `positive`, etc.),
   - leur statut « suicidaire » ou « non suicidaire », encodé en **VRAI**/**FAUX**.

   Après la lecture du fichier `clean_data_2.csv`, les colonnes appropriées sont sélectionnées et converties aux formats voulus (facteurs, booléens, etc.).

2. **Création d'échantillons bootstrap.** Pour chaque itération, un nouveau *jeu d'entraînement* est généré *avec remise* (10 000 rééchantillonnages), et la moyenne des variables numériques est calculée séparément pour les poètes suicidaires (VRAI) et non suicidaires (FAUX). Cette étape vise à lisser la variabilité, surtout lorsque l'échantillon est restreint.

3. **Réduction de dimension par ACP.** Sur chaque jeu d'entraînement bootstrapé, on applique une **analyse en composantes principales**. Les variables émotionnelles, centrées et réduites, sont projetées dans un nouvel espace de dimensions potentiellement plus faibles. Cela limite les effets de la colinéarité et facilite l'ajustement du modèle gaussien.

4. **Ajustement du modèle `MclustDA`.** Les scores sur les composantes principales (ACP) et la variable cible `suicidal` (VRAI/FAUX) servent de base pour estimer un **modèle de mélanges gaussiens discriminants**. Pour un nouvel individu, le modèle alloue la classe (VRAI ou FAUX) qui maximise la probabilité a posteriori.

5. **Validation « leave-one-poet-out».** À chaque itération, on *exclut* un poète afin de constituer l'échantillon de test. Le reste des données est utilisé pour :

   - générer le bootstrap et ajuster le modèle (`MclustDA`) après ACP,
   - projeter le poète exclu dans l'espace des composantes principales, afin de prédire son statut suicidaire (VRAI/FAUX).

   Cette procédure est répétée pour les 16 poètes, chacun étant tour à tour l'exclu.

6.2 Résultats de la validation leave-one-poet-out

Le Tableau 8 détaille, pour chaque poète, la classe **observée** (VRAI = suicidaire, FAUX = non suicidaire) et la classe **prédite** (VRAI/FAUX) par le modèle après application du protocole. Les poètes dont l'observation et la prédiction ne coïncident pas apparaissent en gras.

Table 8: Comparaison entre la classe prédite et la classe observée pour chaque poète (validation leave-one-poet-out).

Poète	**Observé (suicidal)**	**Prédit (suicidal)**
Sylvia Plath	VRAI	VRAI
Denise Levertov	FAUX	FAUX
Anne Sexton	VRAI	VRAI
Adrienne Rich	FAUX	VRAI
Randall Jarrell	VRAI	VRAI
Robert Lowell	FAUX	FAUX
John Berryman	VRAI	VRAI
Lawrence Ferlinghetti	FAUX	FAUX
Hart Crane	VRAI	VRAI
William Carlos Williams	FAUX	FAUX
Sara Teasdale	VRAI	VRAI
Edna St. Vincent Millay	FAUX	VRAI
Charlotte Mew	VRAI	VRAI
Edith Sitwell	FAUX	FAUX
John Davidson	VRAI	VRAI
Alfred Edward Housman	FAUX	VRAI

6.3 Matrice de confusion et indicateurs de performance

En agrégeant l'ensemble des prédictions issues du tableau précédent, on obtient la **matrice de confusion globale** (Tableau 9) :

Table 9: Matrice de confusion globale issue du protocole *leave-one-poet-out*.

	Observé : Non-suicidaire	**Observé : Suicidaire**
Prédit : Non-suicidaire	5	0
Prédit : Suicidaire	3	8

La Table 10 regroupe plusieurs indicateurs dérivés de cette matrice :

Table 10: Statistiques de performance du modèle `MclustDA` (validation leave-one-poet-out).

Mesure	**Valeur estimée**
Exactitude (Accuracy)	81.25%
Intervalle de confiance (95 %)	[54.35%, 95.95%]
Kappa	0.625
Sensibilité (Recall)	1.00
Spécificité	0.625
Exactitude équilibrée (Balanced Accuracy)	0.8125

Faits saillants :

- **Sensibilité = 1,00** : le modèle détecte tous les poètes suicidaires (aucun *faux négatif*).
- **Spécificité = 0,625** : près d'un tiers des poètes non suicidaires sont étiquetés à tort comme suicidaires (*faux positifs*).
- **Exactitude globale** (Accuracy) : environ 81 %, avec un intervalle de confiance assez large, reflétant la petite taille de l'échantillon.
- **Kappa = 0,625** : indique une concordance modérée entre la classification prédite et la classification observée, au-delà du pur hasard.

7 Conclusion

Les différentes analyses menées dans ce travail avaient pour objet d'examiner dans quelle mesure les variables émotionnelles et certaines caractéristiques démographiques peuvent aider à distinguer les poètes suicidaires des non-suicidaires. Nous avons successivement abordé plusieurs aspects : la gestion d'une donnée manquante, la description statistique du corpus, puis diverses stratégies de modélisation (avec ou sans ACP), assorties de protocoles de validation. Les enseignements principaux sont les suivants :

1. Gestion des données manquantes.

- Malgré un **faible nombre de données manquantes** (une seule observation), la mise en place d'un modèle de régression logistique pour imputer la variable d'orientation sexuelle illustre la cohérence d'une approche statistique rigoureuse.
- Le choix d'un **unique prédicteur** (le score moyen d'émotions positives) pour estimer cette variable reposait sur une hypothèse empirique, non universelle, visant essentiellement à démontrer la faisabilité d'une imputation fondée sur un modèle explicatif.
- Cette **imputation** ne prétend pas refléter la réalité biographique de la poétesse concernée ; il s'agit avant tout d'obtenir un jeu de données complet, avec la nécessité de rappeler qu'il s'agit d'une hypothèse artificielle lors de toute analyse ultérieure.

2. Statistiques descriptives des variables démographiques.

- Les premières observations suggèrent qu'**aucune différence notable** n'existe entre hommes et femmes en ce qui concerne la probabilité d'être classés comme suicidaires.
- Une **légère disparité** est toutefois remarquée entre hétérosexuels et non-hétérosexuels, mais la petite taille de l'échantillon et de possibles biais de mesure invitent à la prudence.

- Ces constats descriptifs constituent un **point de départ** pour des analyses plus approfondies, plutôt qu'une conclusion définitive.

3. Statistiques descriptives des variables émotionnelles.

- Les diagrammes en boîtes sur les données brutes ne révélaient pas de patterns aisément interprétables, justifiant le recours à une **moyenne bootstrap** pour extraire le signal sous-jacent et atténuer le bruit.

- La **variabilité** des estimations, visualisée par le rééchantillonnage (10 000 itérations), permet de distinguer des émotions affichant une différence plus stable entre groupes suicidaires et non suicidaires (p. ex. colère, joie, confiance, surprise, etc.) d'autres, comme la peur ou les émotions négatives, pour lesquelles les différences moyennes demeurent plus ténues ou inconsistantes.

- Le recours à l'**ACP** suggère également la possibilité de regrouper certaines émotions corrélées et de projeter les individus (poètes) dans un plan factoriel, offrant un aperçu visuel de leur proximité émotionnelle. Bien que les deux premiers axes puissent expliquer une partie substantielle de la variance, leurs valeurs exactes (82,2 % et 10,1 % de variance expliquée) restent modestes et nécessitent prudence dans l'interprétation.

4. Comparaison de modèles de classification (sans ACP).

- Nous avons examiné trois algorithmes (régression logistique, XGBoost et modèle gaussien) sur le même jeu de données, constatant des **performances globalement similaires** (de 73 % à 79 % d'exactitude).

- Les **intervalles de confiance** se recouvrent, ce qui empêche de désigner formellement un algorithme comme « meilleur ». Les différences observées (p. ex. sensibilité ou spécificité plus élevée pour l'un) s'expliquent en partie par la **faible taille de l'échantillon** et le nombre de rééchantillonnages.

- De surcroît, lorsque la spécificité ou la valeur prédictive positive atteignent 1 pour le modèle gaussien, la valeur prédictive négative s'avère plus faible ; cela illustre les **compromis classiques** (faux positifs *vs.* faux négatifs) dans un contexte de classification binaire.

5. Modélisation par mélanges gaussiens et ACP (validation leave-one-poet-out).

- En passant par une **phase de réduction de dimension** (ACP) et un **modèle de mélanges gaussiens** ajusté sur des données bootstrapées, la validation *leave-one-poet-out* indique une **excellente sensibilité** (aucun poète suicidaire n'a été manqué), mais une **spécificité plus modeste** (environ 62,5 %).

- L'**exactitude globale** avoisine 81 %, ce qui est appréciable compte tenu du faible nombre total de poètes (16). Toutefois, chaque erreur influence fortement les indicateurs de performance, et l'intervalle de confiance relatif à l'exactitude demeure large (54,35 % – 95,95 %).

- Ces résultats confirment la **capacité du modèle** à reconnaître quasi systématiquement les poètes suicidaires, mais au prix de quelques *faux positifs*. Ils illustrent le **dilemme** entre éliminer les faux négatifs (taux de détection maximal) et restreindre les faux positifs.

En synthèse, qu'a-t-on appris ?

- **Valeur du rééchantillonnage et de l'ACP** : Les techniques de bootstrap et de réduction dimensionnelle (ACP) se révèlent pertinentes pour des données hétérogènes et de taille réduite, permettant de stabiliser les estimations et de visualiser les proximités émotionnelles.

- **Caractère sensible du statut suicidaire** : La distinction suicidaire *vs.* non suicidaire requiert une attention particulière aux biais potentiels (biais documentaire, ambiguïtés historiques, etc.), ainsi qu'une réflexion éthique (pas de conclusions hâtives sur la « réalité » du statut).

- **Fiabilité limitée des conclusions** : Les statistiques demeurent fragiles, car l'échantillon se compose de seulement 16 poètes, dont les contributions individuelles pèsent fortement sur les résultats. Les intervalles de confiance larges et les risques de sur-ajustement imposent de nuancer l'interprétation.

- **Comparaison et compromis entre modèles** : Aucun algorithme n'est largement supérieur, chaque modèle exhibant des avantages (sensibilité, spécificité) et des inconvénients (faux positifs, faux négatifs). Dans l'ensemble, l'équilibre global (exactitude) avoisine les 70–80 %, ce qui peut être jugé correct dans ce contexte exploratoire.

References

[1] Stirman, Shannon Wiltsey, and James W. Pennebaker. "Word use in the poetry of suicidal and nonsuicidal poets." *Biopsychosocial Science and Medicine* 63, no. 4 (2001): 517–522.