

Master 2 Data Science

Prédiction de signes neuropsychiques chez les salariés des Industries Electriques et Gazières suivis dans l'observatoire santé-travail EVREST.

Analyse exploratoire des expositions psychosociales et modélisation prédictive

Tuteurs :

Jean Phan Van, Thierry Calvez, Aleksandra Piotrowski (EDF),
Cyril Dalmasso, Vincent Runge (UEVE)

Auteur :

Manal Derghal

5 septembre 2025

Table des matières

1 Contexte et objectifs	3
1.1 Présentation de l'entreprise et de l'équipe	3
1.2 Observatoire EVREST	3
1.2.1 Salariés concernés	4
1.2.2 Qui remplit le questionnaire ?	4
1.2.3 Détails par section	4
1.3 Industries Électriques et Gazières et SPST	5
1.4 Risques psychosociaux	6
1.5 Objectifs	7
2 Méthodes générales	8
2.1 Données disponibles	8
2.2 Outils et librairies	9
3 Description de la population	10
3.1 Participation et sélection des visites	10
3.2 Caractéristiques socio-démographiques de l'échantillon	11
4 Objectif 1 : dimensions des expositions psychosociales	13
4.1 Méthodes	13
4.1.1 Correspondances des axes du rapport Gollac avec EVREST	13
4.1.2 Analyse Factorielle Exploratoire de variables ordinaires	14
4.1.3 Rotation Orthogonale Varimax	14
4.1.4 Corrélation polychoriques et tétrachoriques	15
4.1.5 Sélection des salariés et des visites	15
4.1.6 Indice de Kaiser-Meyer-Olkin (KMO)	16
4.1.7 Test de Sphéricité de Bartlett	17
4.1.8 Déterminant de la Matrice de Corrélations Polychoriques	17
4.1.9 Fiabilité des Facteurs (Alpha de Cronbach)	18
4.2 Résultats	18
4.2.1 Choix et Préparation des Variables	18
4.2.2 Vérification de l'adéquation des données	19
4.2.3 Méthodologie de l'Extraction et de la Rotation Factorielle	19
4.2.4 Qualité de l'Ajustement du Modèle et Interprétation des Facteurs .	20
5 Objectif 2 : prédiction des signes neuropsychiques (à un temps donné)	23
5.1 Méthodes	23
5.1.1 Sélection et Préparation des Données	23
5.1.2 Définition de la Variable Cible (PNP)	23
5.1.3 Gestion des Valeurs Manquantes et Préparation Finale	24
5.1.4 Harmonisation des Caractéristiques	24
5.1.5 Modèles de classification	24

5.1.6	Optimisation des hyperparamètres	26
5.1.7	Gestion du léger déséquilibre des classes	27
5.2	Résultats	27
5.2.1	Observations Préliminaires	27
5.2.2	Aperçu des Performances des Modèles	31
6	Objectif 3 : prédition des signes neuropsychiques (analyse longitudinale)	34
6.1	Méthodes	34
6.1.1	Préparation des données et nettoyage	35
6.2	Résultats	36
6.2.1	Courbe ROC des différents modèles :	36
6.2.2	Performances des modèles de classification	36
7	Discussion	38
7.1	Résumé des résultats	38
7.2	Forces et limites	38
7.2.1	Points positifs de l'étude	38
7.2.2	Limites à considérer	39
7.3	Perspectives	40
8	Bibliographie	41
Annexe A :Questionnaire EVREST		42
Annexe B :Variables retenues pour l'objectif 1		45
Annexe C :Variables retenues pour l'objectif 2 et 3		46
Annexe D :Dépôt GitHub du Projet		48

1 Contexte et objectifs

1.1 Présentation de l'entreprise et de l'équipe

EDF, un acteur majeur du secteur de l'énergie, est très impliqué dans la santé et la sécurité de ses salariés. J'effectue mon alternance au sein de la **Direction des Ressources Humaines (DRH)**, plus précisément à la **Direction Prévention Santé Sécurité Groupe (DP2SG)**.

Je fais partie du **Pôle Outils et Données (POD) Santé Sécurité Toxicologie**, dont le rôle est de concevoir et gérer les outils permettant l'analyse des données de santé et de sécurité au travail (absentéisme, accidents, ...). Dans ce cadre, ma mission principale se concentre sur l'observatoire EVREST (Évolutions et relations en santé au travail). Cet observatoire collecte des données via un questionnaire proposé de façon répétée aux salariés. Mon travail consiste à analyser plus profondément ces données, avec un accent particulier sur la compréhension et la gestion des risques psychosociaux (RPS).

Ma problématique scientifique découle directement de cette mission : "Prédiction de signes neuropsychiques chez les salariés des Industries Électriques et Gazières suivis dans l'observatoire santé-travail EVREST : analyse exploratoire des expositions psychosociales et modélisation prédictive". L'objectif est d'utiliser ces données pour identifier comment les conditions de travail influencent la santé mentale. Je cherche à prédire l'apparition de signes neuropsychiques (comme l'anxiété, les troubles du sommeil ou la lassitude) grâce à des modèles de machine learning. L'intention est de fournir à EDF des analyses précises pour affiner ses stratégies de prévention des RPS et améliorer le bien-être au travail de ses équipes.

J'évolue au sein d'une équipe multidisciplinaire, encadrée par :

- Thierry Calvez, épidémiologiste, qui assure le suivi et l'encadrement régulier de mon travail.
- Jean Phan Van, médecin référent EVREST-IEG et délégué santé du Groupe EDF.
- Aleksandra Piotrowski, responsable du POD, spécialisée en toxicologie.

1.2 Observatoire EVREST

L'observatoire EVREST est un dispositif national français dédié à la surveillance épidémiologique des conditions de travail et de la santé des salariés. Mis en place à la fin des années 2000 par des médecins du travail et des chercheurs, il est géré par un Groupement d'Intérêt Scientifique (GIS) associant des partenaires publics (agences nationales, universités) et privés, dont EDF SA. L'objectif principal d'EVREST est de passer d'une approche individuelle de la santé au travail à une vision collective, permettant de quantifier les expositions professionnelles déclarées par les salariés et leur état de santé pour alimenter le dialogue social et les actions de prévention au sein des entreprises ou des

branches professionnelles.

Le dispositif repose sur un questionnaire standardisé relativement court (trois pages) cf. Annexe A, proposé aux salariés par les équipes des Services de Prévention et de Santé au Travail (SPST) lors des visites médicales périodiques ou des entretiens santé-travail infirmiers. Les données recueillies sont ensuite pseudonymisées grâce à une clé unique par salarié, permettant potentiellement des suivis longitudinaux tout en préservant la confidentialité.

1.2.1 Salariés concernés

Le questionnaire EVREST est proposé lors des entretiens santé-travail systématiques aux salariés. Pour les Visites d'Information et de Prévention (VIP) d'embauche, le salarié doit être au poste depuis au moins deux mois. Les visites de reprise après arrêt de travail sont incluses si elles sont associées à une visite périodique. Tous les types de contrats de travail sont concernés, y compris les intérimaires (avec questions sur les deux derniers mois de mission), mais excluant les contrats très courts. Les équipes de santé au travail peuvent aussi utiliser le questionnaire pour des études ciblées sur l'ensemble des salariés d'une entreprise.

1.2.2 Qui remplit le questionnaire ?

Le remplissage du questionnaire est une tâche collaborative. La **partie administrative** est généralement complétée par le/la **secrétaire** du service de santé au travail. Les sections sur les **conditions de travail, la formation et le mode de vie** sont renseignées par le **salarié** via un auto-questionnaire complété le plus souvent seul par le salarié. Un réexamen systématique des réponses avec le salarié pendant l'entretien est impératif pour validation, compréhension et discussion afin d'enrichir les données chiffrées par la clinique. Le remplissage de la partie **état de santé actuel** est réservé au **médecin ou l'infirmier/ère**, en concertation avec le salarié.

1.2.3 Détails par section

Le questionnaire EVREST se trouve en annexe A. Le guide de remplissage est détaillé sur le site web d'EVREST [1].

Partie Administrative Contient les données d'identité (nom, prénom, sexe, date de naissance) utilisées pour générer une **clé d'anonymat**, ainsi que des informations comme la PCS-ESE [2], la date de visite, le médecin, l'entreprise et l'effectif.

Partie Conditions de travail Renseigne sur les **horaires**, les **contraintes temporelles**, les **ressentis** (autonomie, soutien, reconnaissance), la **charge physique** et les **expositions**. L'entretien permet de valider ces éléments.

Formation – Encadrement – Parcours Porte sur la formation reçue, les fonctions de formateur ou d'encadrant, les mobilités professionnelles, et la capacité perçue à poursuivre son emploi.

Mode de vie Traite de l'activité physique, des consommations de tabac et d'alcool et des trajets domicile-travail.

État de santé actuel Évalue les plaintes et signes perçus la semaine précédant l'entretien, leur gêne au travail et les soins en cours. Le professionnel de santé précise sa fonction et la date de l'entretien.

Partie Facultative Jusqu'à 10 questions libres peuvent être ajoutées après le questionnaire standard, sur une page distincte. Pour EDF SA, 5 d'entre elles sont consacrées aux RPS, 1 sur le changement climatique, et 4 sur la qualité du sommeil.

1.3 Industries Électriques et Gazières et SPST

Les entreprises du secteur des Industries Électriques et Gazières (IEG) du Tableau 1 regroupent de nombreux acteurs majeurs sur le territoire français :

TABLE 1 – Nombre approximatif de salariés en France par entreprise

Entreprise	Nombre approximatif de salariés en France
EDF SA	66 000
Enedis	40 000
Réseau de Transport d'Électricité (RTE)	10 000
Production Électrique Insulaire (PEI)	400
Groupe Électricité de Strasbourg (ÉS)	1100
Engie	4000
Gaz Réseau Distribution France (GRDF)	11 000
NaTran (ex GRTgaz)	3000
Elengy	400
Storengy	600

le nombre total et approximatif de salariés dans le secteur des IEG est d'environ **136 500**. Un groupe de travail spécifique "EVREST-IEG" a été constitué, animé par des médecins référents, afin d'analyser les données propres à ce secteur et de produire des bilans annuels. Ces bilans, destinés aux acteurs de la prévention, agrègent généralement les données sur deux années glissantes. L'implication du secteur IEG dans EVREST de novembre 2008 jusqu'à février 2025 se traduit par **119 567** questionnaires enregistrés,

correspondant à près de **61 293** salariés distincts. Parmi eux, **26 816** ont participé au moins deux fois, illustrant la possibilité de réaliser des analyses longitudinales pour suivre l'évolution des conditions de travail et de la santé au sein de ce collectif. L'outil EVREST permet ainsi, au-delà des bilans annuels, des analyses ciblées ou approfondies pour répondre à des problématiques spécifiques d'une entreprise ou d'un métier. [3]

1.4 Risques psychosociaux

L'évaluation et le suivi des Risques PsychoSociaux (RPS) en France s'appuient largement sur les travaux fondateurs du Collège d'expertise sur le suivi statistique des risques psychosociaux au travail, dont le rapport est fréquemment associé au nom de son président, Michel Gollac [4]. Ce rapport de référence a structuré l'analyse des facteurs de RPS autour de six dimensions principales, permettant une approche systémique de ces risques au travail. Ces six axes sont définis comme suit :

Intensité et temps de travail : Cet axe regroupe les facteurs liés à la charge de travail quantitative et aux contraintes temporelles. Il inclut les contraintes de rythme imposées, l'existence d'objectifs de travail perçus comme irréalistes ou manquant de clarté, les exigences de polyvalence non maîtrisée, le poids des responsabilités assumées, la présence éventuelle d'instructions contradictoires, la fréquence des interruptions d'activités non préparées et l'exigence de compétences particulièrement élevées pour réaliser le travail demandé.

Exigences émotionnelles : Celles-ci renvoient à la nécessité pour le travailleur de maîtriser, voire de modifier ou de masquer, ses propres émotions et sentiments dans le cadre professionnel. Cette exigence est particulièrement prégnante dans les métiers de service ou en contact avec des personnes (clients, patients, usagers, élèves...), où il faut souvent gérer les émotions d'autrui. Le fait de devoir cacher ses propres émotions est en soi une charge psychologique.

Autonomie et marges de manœuvre : Cet axe concerne la possibilité pour le travailleur d'avoir une prise sur son travail et son environnement professionnel. Il englobe les marges de manœuvre opérationnelles (comment faire le travail ?), la participation aux décisions concernant son travail ou son service, ainsi que la possibilité d'utiliser ses compétences actuelles et d'en développer de nouvelles. L'autonomie est ici comprise comme un facteur permettant non seulement l'efficacité mais aussi le développement personnel et le plaisir au travail.

Rapports sociaux et reconnaissance au travail : Cette dimension intègre la qualité des relations professionnelles, que ce soit avec les collègues (soutien social, coopération) ou avec la hiérarchie (style de management, justice organisationnelle). Elle inclut également la reconnaissance obtenue pour le travail effectué, qui peut prendre diverses formes : rémunération, perspectives de carrière, adéquation entre la tâche et les compétences, clarté

et justice des procédures d'évaluation, et attention portée par l'employeur au bien-être des salariés. Les formes pathologiques de rapports sociaux, comme le harcèlement moral, sont également incluses.

Conflits de valeurs : Cette dimension se manifeste lorsqu'une personne est confrontée à l'obligation d'agir d'une manière qui entre en conflit avec ses valeurs professionnelles (le "travail bien fait"), sociales ou personnelles. Ce conflit peut naître du but même du travail ou de ses conséquences, s'ils heurtent les convictions du travailleur. Il peut aussi provenir de l'obligation de travailler d'une façon jugée non conforme à sa conscience professionnelle (ex : manque de moyens pour faire un travail de qualité, devoir mentir, etc.)

Insécurité de la situation de travail : Cet axe fait référence à l'insécurité socio-économique liée au travail. Elle peut découler du risque de perdre son emploi (précarité du contrat, restructurations...), du risque de voir son revenu baisser, ou de l'absence de perspectives de carrière jugées "normales". Des conditions de travail perçues comme non soutenables à long terme peuvent aussi générer un sentiment d'insécurité. De plus, les incertitudes pesant sur l'avenir du métier ou l'évolution des conditions de travail, parfois alimentées par des changements organisationnels incessants ou mal communiqués, contribuent à ce sentiment.

1.5 Objectifs

L'objectif général de cette étude est d'analyser les expositions psychosociales (EPS) mesurées par le questionnaire EVREST, et de développer des modèles prédictifs de plaintes neuropsychiques.

Plus spécifiquement, les objectifs de recherche sont les suivants :

1. **Caractérisation et construction des dimensions d'expositions psychosociales (EPS) :** réaliser une analyse factorielle exploratoire (AFE) pour identifier les principales dimensions sous-jacentes des risques psychosociaux (RPS) dans le questionnaire EVREST et évaluer leur congruence avec les axes théoriques du rapport Gollac.
2. **Développement de modèles prédictifs transversaux :** explorer et évaluer la capacité de différents modèles à prédire l'existence ou non de plaintes neuropsychiques en fonction des expositions psychosociales et des caractéristiques individuelles à un temps donné.
3. **Développement de modèles prédictifs longitudinaux :** explorer et évaluer la capacité de différents modèles à prédire l'existence ou non de plaintes neuropsychiques lors de la prochaine visite d'un même salarié, en se basant sur ses expositions

psychosociales et ses caractéristiques individuelles observées à la visite précédente.

2 Méthodes générales

2.1 Données disponibles

Les données issues du questionnaire EVREST utilisées pour cette analyse proviennent des questionnaires collectés jusqu'en **février 2025**. Initialement, cette base comportait **119 567 enregistrements** (lignes), correspondant aux questionnaires remplis par les salariés entre **novembre 2008 et février 2025**, et comprenait **238 variables** (colonnes).

Les données issues du questionnaire EVREST utilisées pour cette analyse comportent différents types de variables. Une caractéristique importante est que de nombreuses questions, en particulier celles évaluant la fréquence de certaines situations ou le degré d'accord avec des affirmations relatives au travail et à la santé, sont mesurées sur des échelles **ordinaires**. Ces échelles classent les réponses selon un ordre logique. Les principales échelles ordinaires rencontrées dans nos données sont :

- **Échelles de fréquence (4 niveaux)** : utilisées pour des questions comme "En raison de la charge de travail, vous arrive-t-il de...", elles comportent typiquement les niveaux : $0=Jamais$, $1=Rarement$, $2=Assez souvent$, $3=Souvent$.
- **Échelles d'accord (4 niveaux)** : employées pour des questions d'appréciation ("Votre travail est reconnu...") ou de perception ("Vous travaillez avec la peur..."), les niveaux sont généralement : $0=Non pas du tout$, $1=Plutôt non$, $2=Plutôt oui$, $3=Oui tout à fait$.
- **Échelles numériques (ordinaires ou quasi-intervalles)** : certaines questions demandent une cotation numérique sur une échelle, comme la difficulté liée à la pression temporelle (0 à 10) ou l'ambiance de travail (0 à 10).
- **Autres échelles spécifiques** : des codages ordonnés spécifiques existent pour des variables comme la consommation de tabac, la fréquence ou la quantité d'alcool, ou encore la date du dernier entretien santé-travail.

La nature ordinaire de ces variables guide le choix des méthodes statistiques appropriées pour leur description et leur analyse, en reconnaissant l'ordre des catégories. Les variables binaires (codées 0/1 pour Non/Oui) sont également fréquentes pour indiquer la présence ou l'absence d'une condition ou d'une exposition.

2.2 Outils et librairies

Pour la réalisation de l'analyse factorielle exploratoire (AFE) spécifiquement, le logiciel statistique **R** (**version 4.4.2**) a été mobilisé.



Les principales librairies R utilisées pour cette tâche spécifique sont :

- **polycor** : Pour l'analyse factorielle de données ordinaires. Cette librairie calcule les corrélations polychoriques.
- **psych** : Pour effectuer les analyses factorielles exploratoires : analyse en axes principaux, déterminer le nombre de facteurs à retenir, examiner la structure factorielle, et évaluer la cohérence interne des dimensions identifiées (calcul de l'alpha de Cronbach).



Python

Python (**version 3.11.0**) constitue l'outil principal pour la partie prédition. Les principales librairies Python mobilisées pour ce projet sont :

- **pandas** et **numpy** : pour l'ensemble des opérations de manipulation, de nettoyage et de prétraitement des données, ainsi que pour les calculs numériques complexes.
- **matplotlib** et **seaborn** : pour la visualisation des données, que ce soit pour l'analyse exploratoire ou la présentation des résultats.
- **scikit-learn** : pour le prétraitement des données (normalisation, division en ensembles d'entraînement/test), la construction de divers modèles de classification (Régression Logistique, Arbres de Décision, Forêts Aléatoires), la validation croisée et l'optimisation des hyperparamètres.
- **lightgbm** : spécifiquement utilisée pour le modèle LightGBM, un algorithme de boosting.

3 Description de la population

3.1 Participation et sélection des visites

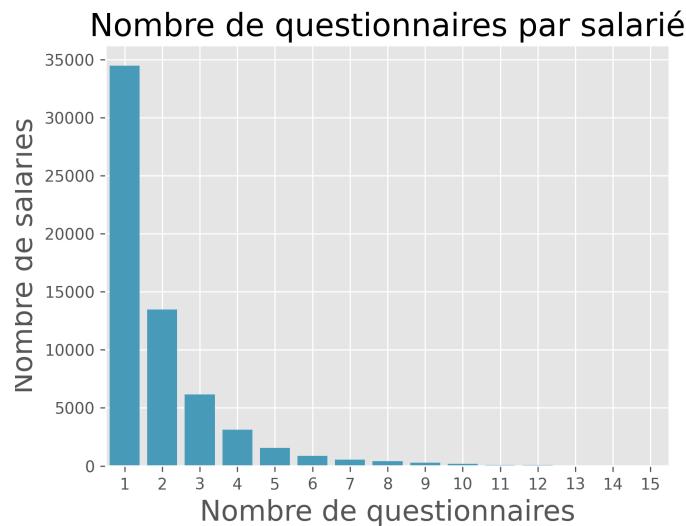


FIGURE 1 – Distribution du nombre de questionnaires EVREST remplis par salarié sur la période 2008-2025.

La Figure 1 illustre la répartition du nombre de participations par salarié. On observe qu'une majorité de salariés (34 475) n'a rempli le questionnaire qu'une seule fois. Cependant, un nombre important de participants l'a rempli plus de deux fois (26 818), confirmant le potentiel pour des analyses longitudinales. Le nombre moyen de visites par personne est de 1,95.

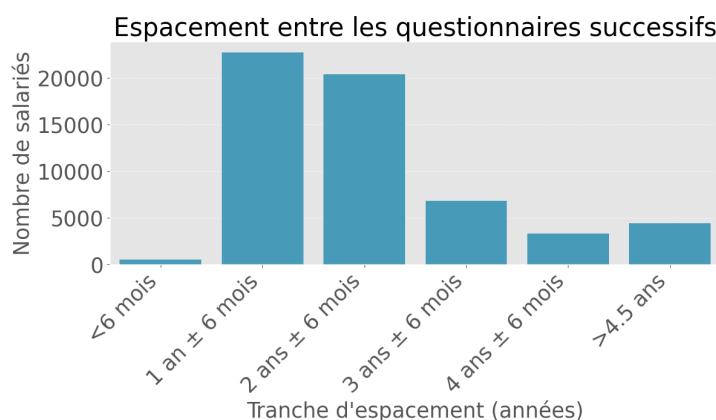


FIGURE 2 – Distribution de l'intervalle de temps (en années) entre les questionnaires consécutifs pour les salariés ayant participé plusieurs fois.

Pour les 26 818 salariés ayant participé au moins deux fois, la Figure 2 détaille l’espacement temporel entre leurs visites successives. On retrouve un pic marqué autour de 1 an \pm 6 mois (22 752 salariés) et autour de 2 ans \pm 6 mois (20 401 salariés). L’espacement moyen est de 2,17 ans.

3.2 Caractéristiques socio-démographiques de l’échantillon

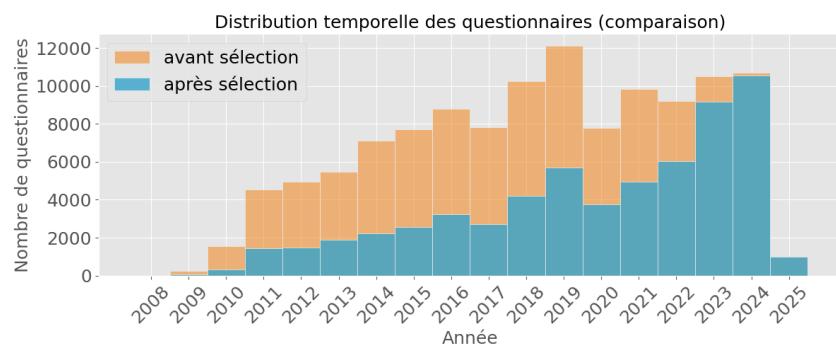


FIGURE 3 – Comparaison de la distribution temporelle des questionnaires : total enregistré versus échantillon utilisé pour l’analyse (sélection du plus récent pour les salariés ayant plusieurs questionnaires).

La Figure 3 compare la répartition temporelle des questionnaires bruts (en orange) à celle de l’échantillon final (en bleu). Cet échantillon a été construit en ne retenant que le questionnaire le plus récent pour chaque salarié, afin de ne prendre en compte qu’une seule observation par individu pour les analyses. La concentration de l’échantillon final sur les années les plus récentes est une conséquence directe de cette méthode de sélection.

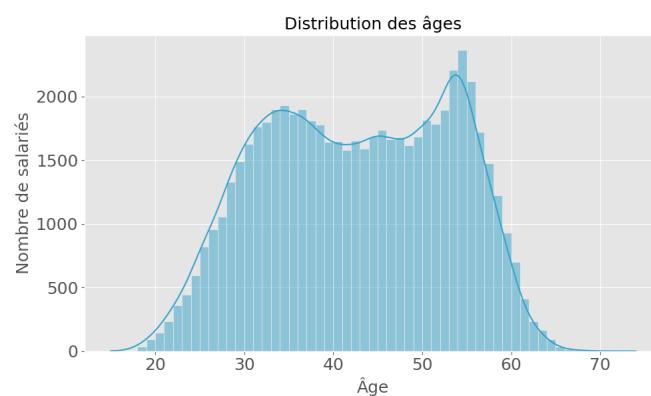


FIGURE 4 – Distribution de l’âge des salariés au moment de leur dernier questionnaire.

La Figure 4 présente la distribution de l’âge des salariés de l’échantillon. La distribution des âges s’étend principalement de 20 à 65 ans, avec une moyenne d’âge de 42 ans.

Enfin, pour décrire l'échantillon, nous utilisons les **Professions et Catégories Socio-professionnelles (PCS)** [2]. Les principales catégories incluent les Cadres, Professions intermédiaires, Employés et Ouvriers. Après avoir sélectionné le dernier questionnaire rempli par chaque salarié, et en excluant les valeurs de PCS erronées, notre échantillon final compte 61 284 salariés.

TABLE 2 – Résumé des caractéristiques démographiques par catégorie socioprofessionnelle (PCS).

Catégorie	Nombre de salariés	% Total	Age moyen	% Femmes
Cadre	24 181	39.5	45	28.0
Prof.Interm	28 852	47.1	43	25.2
Employé	2152	3.5	42	56.3
Ouvrier	6099	10.0	37	9.4
Total	61 284	100	42	25.8

Le Tableau 2 illustre les principales caractéristiques de l'échantillon réparties par catégorie socioprofessionnelle. Les catégories les plus représentées sont les Professions intermédiaires avec 28 852 personnes et les Cadres avec 24 181 personnes.

En termes de composition par sexe, une forte prédominance masculine est observée chez les Ouvriers, les Cadres et les Professions intermédiaires. Concernant l'âge moyen, les Cadres (45 ans) et les Professions intermédiaires (43 ans) affichent des âges légèrement plus élevés que les Employés (42 ans) et surtout les Ouvriers (37 ans).

4 Objectif 1 : dimensions des expositions psychosociales

4.1 Méthodes

4.1.1 Correspondances des axes du rapport Gollac avec EVREST

Le questionnaire EVREST, bien que couvrant un champ plus large que les seuls risques psychosociaux (RPS), contient plusieurs questions qui peuvent être mises en correspondance avec les six axes définis par le rapport Gollac. Cette approche permet une lecture opérationnelle des résultats d'EVREST sous l'angle des RPS, en utilisant certaines questions comme des indicateurs indirects de l'exposition potentielle des salariés aux différentes dimensions du risque. Le tableau ci-dessous présente la correspondance dans ce cadre interprétatif.

TABLE 3 – Correspondances des axes du rapport Gollac avec EVREST

Axe défini dans le rapport Gollac	Questions EVREST
Intensité et temps de travail	<ul style="list-style-type: none"> – Difficultés liées à la pression temporelle (score 0–10) – Interruptions d'activité perturbant le travail – Travail de nuit – Horaires irréguliers ou alternés – Dépassement d'horaires, sauts de pauses/repas, traitement trop rapide
Exigences émotionnelles	<ul style="list-style-type: none"> – Contact avec le public – Exposition à la pression psychologique
Autonomie et marges de manœuvre	<ul style="list-style-type: none"> – Faible liberté dans la manière de travailler – Pas de possibilité d'apprentissage
Rapports sociaux et reconnaissance au travail	<ul style="list-style-type: none"> – Travail peu reconnu – Manque d'entraide et coopération
Conflits de valeurs	<ul style="list-style-type: none"> – Faire des choses que l'on désapprouve – Manque de moyens pour un travail de qualité
Insécurité de la situation de travail	<ul style="list-style-type: none"> – Peur de perdre son emploi

Cette correspondance interprétative du Tableau 3 fournit une base pour identifier les grandes dimensions des RPS à partir des réponses au questionnaire EVREST. Dans la suite, nous analyserons cette structure en détail à l'aide d'une analyse factorielle afin d'explorer les dimensions sous-jacentes aux réponses des salariés.

4.1.2 Analyse Factorielle Exploratoire de variables ordinales

L'Analyse Factorielle Exploratoire (AFE) repose sur l'hypothèse d'une structure latente et de relations linéaires entre les variables et les facteurs. Le modèle décompose les variables observées en une combinaison linéaire de facteurs communs non observables et d'un facteur unique à chaque variable.

Contrairement à l'Analyse en Composantes Principales (ACP) qui forme des composantes comme une combinaison linéaire des variables observées, l'AFE modélise les variables observées comme étant causées par des facteurs latents. L'influence de chaque facteur latent sur une variable mesurée est appelée **charge factorielle** [5].

Formulation du modèle : Le modèle factoriel s'écrit sous la forme matricielle suivante :

$$X = \mu + \Lambda F + \epsilon$$

Où :

- $X \in \mathbb{R}^p$: Vecteur des **variables observées**.
- $\mu \in \mathbb{R}^p$: Vecteur des **moyennes** des variables.
- $\Lambda \in \mathbb{R}^{p \times m}$: **Matrice des charges factorielles**. Chaque λ_{ij} représente l'impact du facteur j sur la variable i .
- $F \in \mathbb{R}^m$: Vecteur des **facteurs communs** non observables.
- $\epsilon \in \mathbb{R}^p$: Vecteur des **facteurs uniques** ou erreurs.

4.1.3 Rotation Orthogonale Varimax

Après avoir extrait les facteurs, on les fait pivoter pour mieux les interpréter. La rotation Varimax est une technique qui vise à simplifier la structure des chargements factoriels. Son objectif est de rendre chaque variable très liée à un seul facteur et très peu liée à tous les autres.

La rotation Varimax est une **rotation orthogonale** [6], ce qui signifie que les facteurs restent non corrélés entre eux (leurs axes restent à 90° les uns des autres).

Mathématiquement, le processus consiste à trouver une matrice de rotation orthogonale \mathbf{T} qui, appliquée à la matrice des chargements initiale Λ , donne une nouvelle matrice de chargements rotatée Λ^* plus simple :

$$\Lambda^* = \Lambda \mathbf{T}$$

où $\mathbf{T}'\mathbf{T} = \mathbf{I}$. L'algorithme Varimax recherche l'orientation des facteurs qui maximise la variance des chargements au carré pour chaque facteur.

4.1.4 Corrélation polychoriques et tétrachoriques

Étant donné la nature majoritairement ordinaire des variables du questionnaire EVREST, des **corrélations polychoriques** et **tétrachoriques** ont été utilisées pour estimer leurs relations.

La corrélation polychorique estime la corrélation de Pearson entre deux variables continues et non observées qui sous-tendent les variables ordinaires [7]. Une variable ordinaire, Y , peut être vue comme une discrétisation d'une variable **variable latente** continue sous-jacente, X .

La corrélation polychorique suppose que ces variables latentes suivent une distribution normale bivariée avec une corrélation ρ . La corrélation polychorique est alors définie comme l'estimation de ρ . Autrement dit, lorsque l'on dit que la corrélation polychorique entre deux variables ordinaires, Y_1 et Y_2 , est r , cela signifie que r est une estimation de la corrélation de Pearson entre deux variables latentes, X_1 et X_2 , qui sont inférées à partir de Y_1 et Y_2 . La **corrélation tétrachorique** est un cas spécifique utilisé lorsque les deux variables sont dichotomiques.

Ce processus complexe a été automatisé en R grâce à la fonction `auto.correlate` du package `polycor`. Celle-ci détecte automatiquement le type de variable et applique la méthode de corrélation appropriée, qu'elle soit polychorique, tétrachorique, ou une autre, pour produire une matrice de corrélation fiable pour l'analyse factorielle. L'adéquation de l'échantillonnage pour l'analyse a ensuite été vérifiée afin de s'assurer que les données sont appropriées pour cette technique statistique [8, 9].

4.1.5 Sélection des salariés et des visites

Conformément aux objectifs de l'étude, une première étape a consisté à sélectionner un sous-ensemble de variables pertinentes pour l'analyse des expositions psychosociales et des indicateurs de santé. Nous avons constitué un échantillon en ne conservant que le **questionnaire le plus récent pour chaque salarié**. Cette opération permet d'éviter de compter plusieurs fois le même individu dans les descriptions de population. L'échantillon final est composé de **61 284** salariés distincts. Pour l'obtenir, nous avons exclu les valeurs de PCS erronées.

Une première analyse de la matrice des corrélations entre les 28 variables initiales, visualisée sur la Figure 5, a été menée où les corrélations positives sont en vert, les négatives en rouge. Sur la base de la force de ces corrélations, les variables jugées trop redondantes (trop corrélées) ou pas assez pertinentes (pas assez corrélées au reste du jeu) ont été écartées. Cette étape de sélection a permis de réduire le nombre de variables à **13**, qui ont été conservées pour la suite de l'analyse factorielle. **La description détaillée des libellés, codages et traitements de ces variables est présentée en Annexe B.** Plusieurs variables ont été inversées afin que des scores élevés indiquent systématiquement

une exposition accrue au risque.

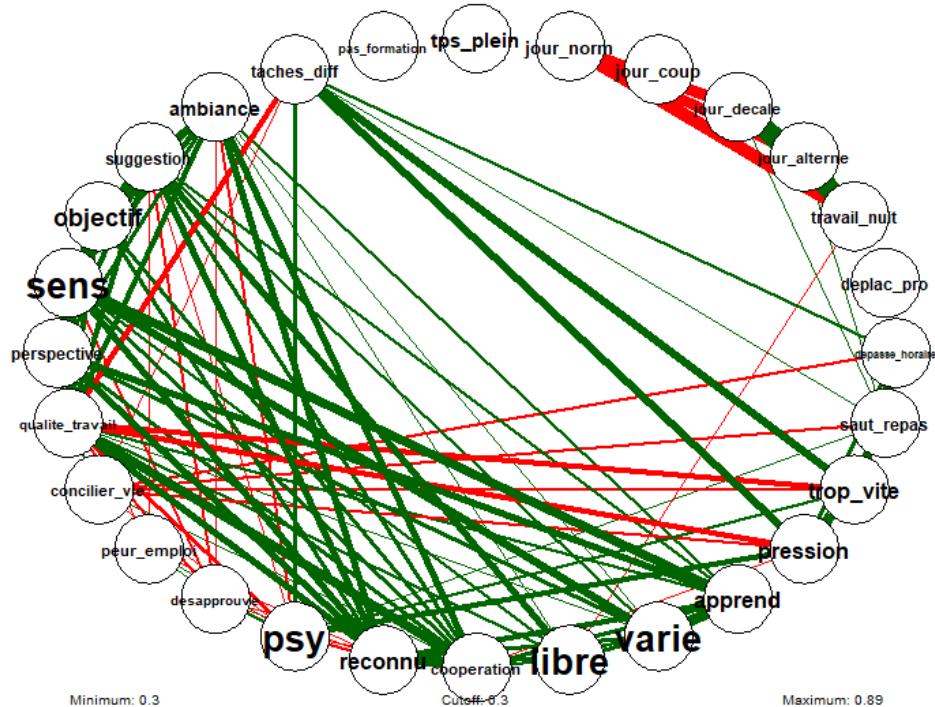


FIGURE 5 – Visualisation des corrélations polychoriques

4.1.6 Indice de Kaiser-Meyer-Olkin (KMO)

L’indice KMO est une mesure statistique qui évalue la proportion de la variance dans nos variables pouvant être expliquée par des facteurs sous-jacents communs. Il détermine si les corrélations polychoriques observées entre les variables sont suffisamment fortes et cohérentes pour justifier une analyse factorielle. L’indice KMO global, pour un ensemble de variables, est défini par la formule suivante :

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2}$$

Où :

- r_{ij} représente le **coefficent de corrélation polychorique** entre la variable ordinaire i et la variable ordinaire j .
- p_{ij} représente le **coefficent de corrélation partielle** entre la variable i et la variable j , calculé à partir de la matrice de corrélations polychoriques, en contrôlant pour l’influence de toutes les autres variables de l’ensemble.

La mesure Kaiser-Meyer-Olkin (KMO) évalue l'adéquation d'un ensemble de variables à une analyse factorielle en quantifiant la proportion de variance commune. Ses valeurs varient de 0 à 1, une valeur élevée (proche de 1) indiquant que les variables partagent une variance commune significative et que les corrélations partielles sont faibles, ce qui est favorable à l'analyse factorielle. Inversement, une valeur faible suggère des corrélations trop faibles pour extraire une structure factorielle pertinente.

Kaiser (1974) considère les valeurs inférieures à 0.50 comme inacceptables. Toutefois, d'autres spécialistes recommandent un minimum de 0.60, avec une préférence pour des valeurs supérieures ou égales à 0.70. [9]

4.1.7 Test de Sphéricité de Bartlett

Le test de sphéricité de Bartlett est un test statistique qui examine si la matrice de corrélation des variables est significativement différente d'une matrice identité. Une matrice identité signifie que toutes les corrélations entre les variables sont nulles (ou très proches de zéro), ce qui impliquerait qu'il n'y a pas de relations significatives entre les variables et donc qu'aucune structure factorielle ne peut être extraite.

L'hypothèse nulle (H_0) de ce test est que la matrice de corrélation est une matrice identité (c'est-à-dire que les variables sont non corrélées dans la population). L'hypothèse alternative (H_1) est que la matrice de corrélation n'est pas une matrice identité, ce qui signifie qu'il existe des corrélations significatives entre au moins certaines variables.

La statistique du test de Bartlett, notée χ^2 , est calculée par la formule suivante :

$$\chi^2 = - \left[N - 1 - \frac{(2p + 5)}{6} \right] \ln(|\mathbf{R}|)$$

Où :

- N est la taille de l'échantillon (nombre d'observations).
- p est le nombre de variables.
- $|\mathbf{R}|$ est le déterminant de la matrice de corrélation observée.

Cette statistique suit approximativement une loi du χ^2 à $\frac{p(p-1)}{2}$ degrés de liberté.

Une valeur p associée à ce test, généralement inférieure à un seuil de signification de 0.05, permet de rejeter l'hypothèse nulle. [9]

4.1.8 Déterminant de la Matrice de Corrélations Polychoriques

Le déterminant de la matrice de corrélations est une mesure de la multicolinéarité parmi les variables. En analyse factorielle, il est préférable d'éviter une multicolinéarité sévère, c'est-à-dire des variables qui sont excessivement corrélées entre elles au point d'être redondantes.

Un déterminant très proche de zéro (inférieur à 0.00001) indiquerait une multicolinéarité sévère, ce qui peut rendre les calculs instables et les résultats de l'analyse factorielle peu fiables. Un déterminant supérieur à ce seuil suggère que les variables ne sont pas linéairement dépendantes au point de poser problème, et que chaque variable apporte une information unique à la structure factorielle. [9] [10]

4.1.9 Fiabilité des Facteurs (Alpha de Cronbach)

L'**alpha de Cronbach** (α) mesure la cohérence interne d'un ensemble d'items, c'est-à-dire le degré auquel les items censés mesurer le même facteur sont corrélés entre eux. En d'autres termes, un alpha élevé indique que les items d'une échelle mesurent bien la même dimension latente.

L'alpha de Cronbach peut s'exprimer en fonction de la corrélation moyenne inter-items (\bar{r}) :

$$\alpha = \frac{k \cdot \bar{r}}{1 + (k - 1)\bar{r}}$$

Où :

- k est le nombre total d'items dans le facteur (ou l'échelle).
- \bar{r} est la corrélation moyenne entre les items.

On utilise ici la **corrélation polychorique** pour estimer \bar{r} . Une valeur de α varie généralement entre 0 et 1. Un alpha de 0 indique une absence de cohérence interne, tandis qu'un alpha de 1 traduit une cohérence interne parfaite. En règle générale, une valeur de α supérieure à 0.70 est considérée comme acceptable pour la recherche. [9]

4.2 Résultats

4.2.1 Choix et Préparation des Variables

L'analyse factorielle exploratoire (AFE) a été conduite dans le but d'identifier les structures latentes sous-jacentes aux risques psychosociaux au travail, tels qu'évalués par 13 items du questionnaire EVREST.

Population et Échantillon : L'analyse a porté sur un échantillon de $N = 61\,284$ participants, une taille considérée comme excellente pour une AFE robuste. La Figure 6 illustre la faible proportion de valeurs manquantes (environ 2%).

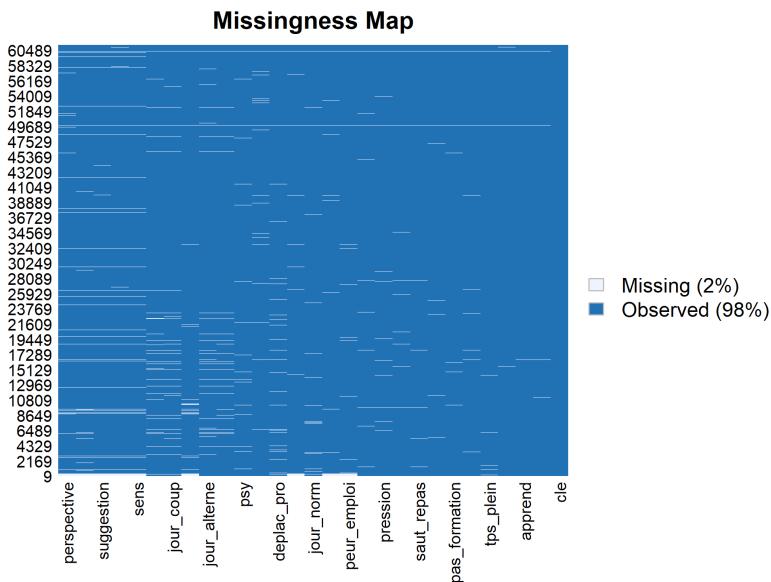


FIGURE 6 – Visualisation des valeurs manquantes (Missmap)

Les données manquantes ont été supprimées pour simplifier le jeu de données. Cette suppression nous ramène à un échantillon final de $N = 50\,562$ participants, une taille qui reste **très grande et amplement suffisante** pour l'analyse factorielle exploratoire (AFE).

4.2.2 Vérification de l'adéquation des données

L'adéquation des données à une analyse factorielle a été rigoureusement examinée. Les résultats confirment que l'échantillon est approprié :

- Un **indice de Kaiser-Meyer-Olkin (KMO)** de **0.89**, jugé excellent, indique que les corrélations partielles entre les variables sont faibles.
- Le **test de sphéricité de Bartlett** est hautement significatif ($\chi^2(78) = 288741.6$, $p < 0.001$), confirmant que les variables sont suffisamment corrélées entre elles pour être structurées en facteurs.
- Le **déterminant de la matrice des corrélations polychoriques** (0.003341982) est supérieur au seuil critique de 0.00001, ce qui suggère l'absence de multicolinéarité sévère.

4.2.3 Méthodologie de l'Extraction et de la Rotation Factorielle

L'extraction des facteurs a été réalisée en utilisant l'**analyse en axes principaux (PA)**. Cette méthode a permis de capturer **51% de la variance totale** des données.

Pour déterminer le nombre optimal de facteurs à retenir, nous nous sommes basés sur le **graphique des éboulis** (Figure 7). Cette analyse a suggéré une solution à **deux facteurs**. Une **rotation orthogonale varimax** a ensuite été appliquée.

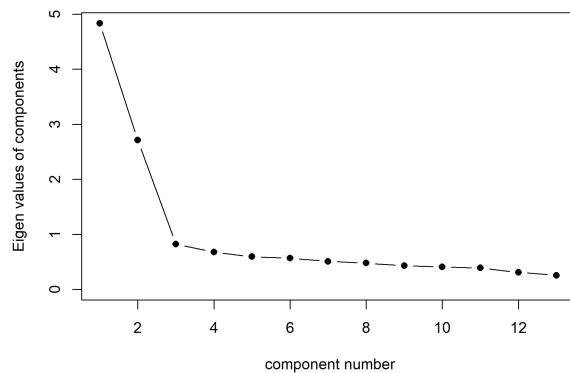


FIGURE 7 – Graphique des éboulis

4.2.4 Qualité de l’Ajustement du Modèle et Interprétation des Facteurs

Le graphique des résidus de corrélation (Figure 8) montre une majorité de résidus très faibles ($RMSR = 0.04$), confirmant le bon ajustement du modèle. La complexité moyenne des items est de 1.2. Dans le cadre d’une analyse factorielle, la complexité d’un item fait référence au nombre de facteurs sur lesquels cet item présente une saturation significative, c'est-à-dire une forte corrélation (>0.3).

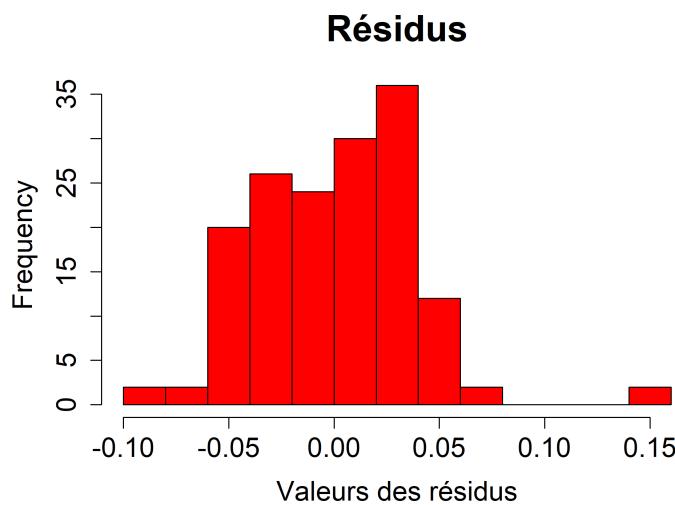


FIGURE 8 – Distribution des résidus de corrélation

Le diagramme factoriel (Figure 9) visualise la structure des saturations des items et les corrélations entre les **deux facteurs extraits**.

La qualité de l'estimation des scores factoriels est évaluée par le carré du coefficient de corrélation multiple (R^2) des scores avec les facteurs latents. Ces valeurs sont très élevées pour les facteurs (PA1 : **0.94**, PA2 : **0.91**), indiquant que les scores factoriels estimés à partir des items sont de très bons représentants des facteurs latents sous-jacents.

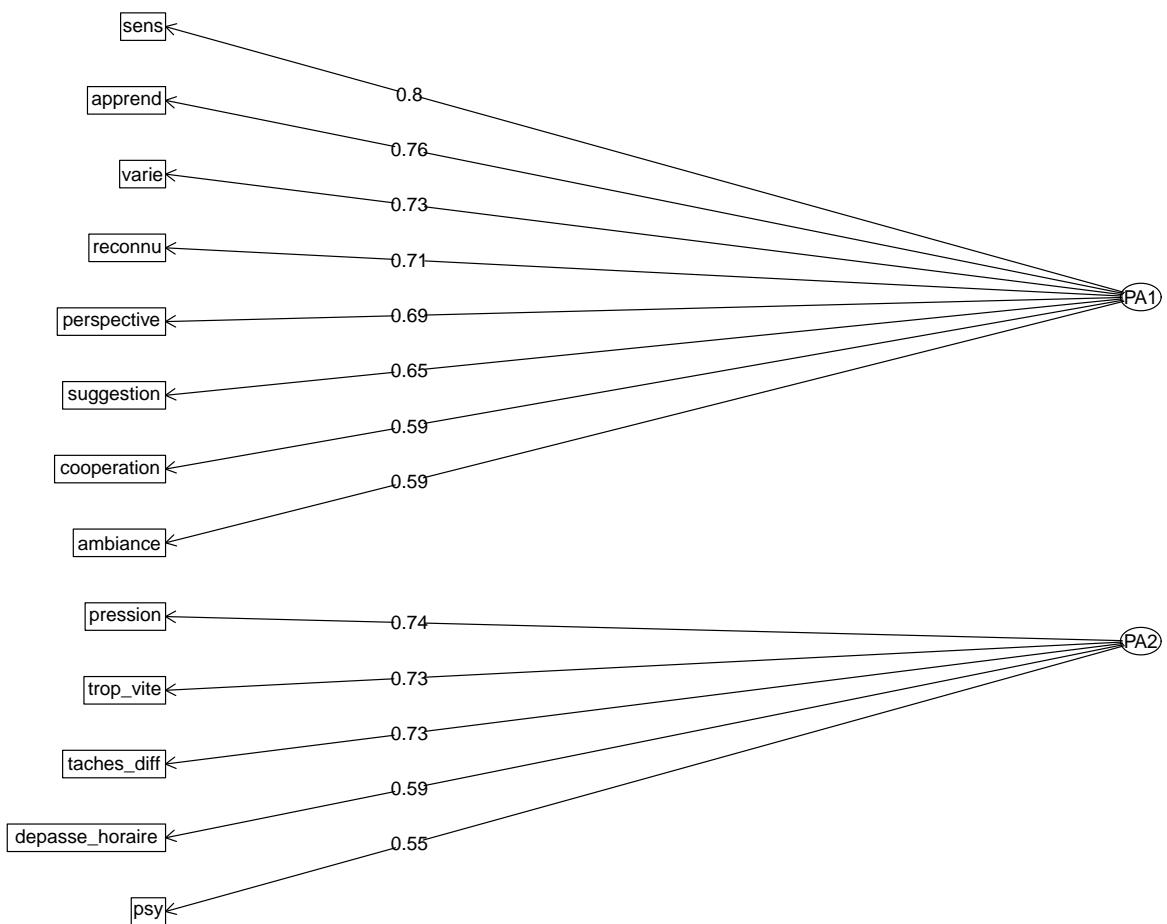


FIGURE 9 – Diagramme factoriel (solution à 2 facteurs, rotation varimax)

L’interprétation des deux facteurs retenus, basée sur les items présentant les plus fortes saturations, a permis d’identifier les dimensions clés suivantes des risques psychosociaux.

Facteur 1 : Satisfaction et Opportunités de Développement. Ce facteur regroupe les items fortement liés à la qualité de l’environnement de travail, à la reconnaissance et aux perspectives d’évolution. Ses chargements les plus élevés sont sur les items **sens** (0,797), **apprend** (0,759), **varie** (0,726), **reconnu** (0,706), **perspective** (0,691), **suggestion** (0,649), **cooperation** (0,591), et **ambiance** (0,589). Ce facteur met en évidence une **dynamique de travail positive** et des **opportunités de croissance professionnelle**.

Facteur 2 : Pression et Charge de Travail. Ce facteur est constitué d’items reflétant une forte charge de travail et des contraintes temporelles importantes. Les items qui le définissent le mieux sont **pression** (0,741), **trop_vite** (0,733), **taches_diff** (0,728), **depasse_horaire** (0,586), et **psy** (0,551). Il reflète l’impact d’une **charge de travail élevée** sur le bien-être et l’équilibre vie professionnelle/personnelle.

Cohérence Interne des Facteurs Les valeurs d’alpha de Cronbach, un indicateur de la cohérence interne des items de chaque facteur, sont tout à fait satisfaisantes.

- **Facteur 1** (8 items) : $\alpha = 0,881$
- **Facteur 2** (5 items) : $\alpha = 0,804$

Toutes les valeurs d’alpha sont bien supérieures au seuil généralement accepté de 0,70. Cela confirme que les items au sein de chaque facteur mesurent un même concept sous-jacent avec une bonne fidélité.

Conclusion En synthèse, cette analyse factorielle exploratoire a permis d’identifier deux facteurs clés des risques psychosociaux : **Satisfaction et Opportunités de Développement** et **Pression et Charge de Travail**. Ces dimensions, ainsi que leur interprétation, sont cohérentes avec les axes définis par le rapport Gollac.

5 Objectif 2 : prédition des signes neuropsychiques (à un temps donné)

5.1 Méthodes

Pour ce second objectif, notre démarche vise à construire et évaluer des modèles de Machine Learning capables de prédire la présence ou non de plaintes neuropsychiques chez les salariés à un instant donné, en s'appuyant sur leurs expositions psychosociales et leurs caractéristiques individuelles.

5.1.1 Sélection et Préparation des Données

Dans un premier temps, nous avons procédé à une sélection des variables. Le jeu de données initial, comportant 119 567 enregistrements et 238 variables, a été réduit pour ne retenir que les colonnes pertinentes pour notre problématique. Nous avons écarté les variables non essentielles (ex : code entreprise, atelier service, taille, poids), celles présentant un taux élevé de valeurs manquantes (notamment issues de questions récemment ajoutées ou retirées du questionnaire, ou liées à la période COVID), ainsi que celles jugées non directement liées aux risques psychosociaux (ex : exposition aux intempéries, plaintes digestives ou dorsales). **Les 40 variables finalement retenues pour l'analyse sont détaillées en Annexe C.**

Étant donné que certains salariés ont répondu au questionnaire plusieurs fois, nous avons filtré les données pour ne conserver que la dernière réponse de chaque salarié. L'échantillon final est composé de 61 226 salariés distincts. Pour l'obtenir, nous avons d'abord exclu les valeurs de PCS erronées. De plus, nous avons choisi de conserver uniquement les données collectées à partir de 2010, car les enregistrements antérieurs à cette date concernaient un nombre très limité de salariés.

5.1.2 Définition de la Variable Cible (PNP)

La variable cible a été nommée ‘**pnp**’ (Plainte NeuroPsychique). Nous avons exploré deux options basées sur les plaintes neuropsychiques :

- Une association d'au moins une plainte parmi l'anxiété (**s_anxiete_plainte**), la fatigue/lassitude (**s_fatigue_plainte**), ou les troubles du sommeil (**s_sommeil_plainte**).
- Une association où les trois plaintes (anxiété, fatigue, sommeil) sont présentes simultanément.

Nous avons opté pour la première option, considérant la présence d'au moins une de ces plaintes comme indicateur de PNP. Cette décision a abouti à une variable **pnp** composée de **32 766 observations (68,7 %)** pour la classe 0 (« pas de plainte ») et de **14 926 observations (31,3 %)** pour la classe 1 (« plainte »), après la gestion des valeurs manquantes.

5.1.3 Gestion des Valeurs Manquantes et Préparation Finale

Les observations comportant des valeurs manquantes ont été supprimées ce qui a entraîné la suppression de 13 534 lignes, ramenant la taille finale de notre jeu de données à **47 692 lignes**.

Les données ont été divisées en ensembles d'entraînement `X_train`, `y_train` (80 %) et de test `X_test`, `y_test` (20 %) pour évaluer la capacité de généralisation des modèles.

5.1.4 Harmonisation des Caractéristiques

Conformément à l'objectif 1, les variables binaires et ordinaires ont été harmonisées de sorte que la valeur 1 (ou une valeur plus élevée pour les échelles) indique une augmentation de l'exposition aux plaintes.

5.1.5 Modèles de classification

Nous avons évalué plusieurs modèles de classification :

- **Arbre de Décision** : Un arbre de décision partitionne récursivement l'espace des variables explicatives en sous-régions homogènes en termes de classe cible. Chaque nœud est scindé selon la variable qui maximise un critère de pureté, tel que le gain d'information et l'entropie utilisée par la suite. (Algorithme glouton).

- Entropie :

$$H(S) = - \sum_{k=1}^2 p_k \log(p_k),$$

où p_k est la proportion d'observations de la classe k dans un nœud.

- Gain d'information : L'algorithme de l'arbre de décision choisit toujours la division qui maximise cette valeur.

$$IG(S, A) = H(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} H(S_i),$$

où $IG(S, A)$ est le gain d'information obtenu en scindant l'ensemble S sur la caractéristique A . $|S|$ est le nombre total d'échantillons dans le noeud parent S . $|S_i|$ est le nombre d'échantillons dans le nœud enfant S_i . $H(S)$ est l'entropie du nœud parent. $H(S_i)$ est l'entropie du nœud enfant S_i .

- **Forêt Aléatoire (Random Forest)** : La forêt aléatoire est un *ensemble d'arbres de décision* construits sur des sous-échantillons de données et de variables. Chaque

arbre vote pour une classe, et la prédiction finale correspond à une agrégation (vote majoritaire). Ce procédé réduit la variance et améliore la généralisation.

- **LightGBM** : LightGBM (Light Gradient Boosting Machine) est un algorithme de *gradient boosting* qui construit séquentiellement des arbres de décision en minimisant une fonction de perte par descente de gradient. Chaque nouvel arbre est entraîné à approximer le gradient négatif de la fonction de perte par rapport aux prédictions précédentes.

Si $F_m(x)$ désigne le modèle à l'itération m , alors l'itération suivante est définie par :

$$F_{m+1}(x) = F_m(x) + \eta h_m(x),$$

où $h_m(x)$ est un arbre de décision ajusté sur le gradient négatif, et η le taux d'apprentissage.

Dans le cas d'une classification binaire, la fonction de perte couramment utilisée est la **log-loss**, définie par :

$$L(y_i, p_i) = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)].$$

- y_i est l'étiquette de classe réelle (0 ou 1).
- p_i est la probabilité prédite par le modèle pour la classe 1 de l'observation i .

- **Régression Logistique** : La régression logistique est un modèle linéaire probabiliste qui relie les variables explicatives x à la probabilité d'appartenir à la classe positive ($y = 1$) via une fonction sigmoïde :

$$p(x) = \frac{1}{1 + e^{-(w^\top x + b)}}.$$

- w est le vecteur des **coefficients de poids**.
- x est le vecteur des **variables explicatives** (caractéristiques) d'une observation.
- b est le **biais** (ou ordonnée à l'origine)

La fonction de perte à utiliser est la log-loss :

$$L(y_i, p_i) = y_i \log(p_i) + (1 - y_i) \log(1 - p_i).$$

- y_i est l'étiquette de classe réelle (0 ou 1).
- p_i est la probabilité prédite par le modèle pour la classe 1 de l'observation i .

Régularisation : Pénalisations L1 et L2 Pour éviter le sur-apprentissage, on ajoute un terme de **régularisation** à la fonction de perte. Cette technique pénalise les coefficients de poids trop élevés.

Pénalisation L2 (Ridge) La pénalisation **L2** ajoute un terme de pénalité proportionnel à la somme des carrés des coefficients de poids. La fonction de perte devient :

$$\text{Minimiser} \quad L(y, p) + \alpha \sum_j w_j^2$$

Cette pénalisation force les coefficients à rester petits mais ne les annule généralement pas. Elle est très utile pour gérer la **multicolinéarité** et pour stabiliser le modèle.

Pénalisation L1 (Lasso) La pénalisation **L1** ajoute un terme de pénalité proportionnel à la somme des valeurs absolues des coefficients de poids. La fonction de perte devient :

$$\text{Minimiser} \quad L(y, p) + \alpha \sum_j |w_j|$$

Cette pénalisation a la propriété de pouvoir réduire certains coefficients à **zéro**. Elle est donc très efficace pour la **sélection de variables**, créant des modèles plus faciles à interpréter.

5.1.6 Optimisation des hyperparamètres

L'optimisation des hyperparamètres de chaque modèle a été réalisée via une recherche par grille aléatoire (`RandomizedSearchCV`). Cette méthode, couplée à une validation croisée stratifiée (`StratifiedKFold`), permet de trouver les meilleures configurations tout en évitant le surapprentissage. La métrique d'évaluation principale a été le **rappel** (recall) pour la classe 1, car l'objectif est de maximiser la détection des salariés présentant une plainte neuropsychique :

$$\text{Rappel} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}}$$

D'autres métriques ont également été observées pour une évaluation complète des performances du modèle, telles que :

Précision : Mesure la proportion de prédictions positives qui sont réellement correctes.

$$\text{Précision} = \frac{\text{Vrais Positifs (VP)}}{\text{Vrais Positifs (VP)} + \text{Faux Positifs (FP)}}$$

Accuracy : L'accuracy mesure la proportion de l'ensemble des prédictions qui sont correctes.

$$\text{Accuracy} = \frac{\text{Vrais Positifs (VP)} + \text{Vrais Négatifs (VN)}}{\text{Vrais Positifs (VP)} + \text{Vrais Négatifs (VN)} + \text{Faux Positifs (FP)} + \text{Faux Négatifs (FN)}}$$

Le choix de la métrique principale est important, car elle privilégie la détection maximale des cas positifs, considérant qu'un faux négatif (un cas non détecté) a des conséquences plus importantes qu'un faux positif (un cas mal classé).

5.1.7 Gestion du léger déséquilibre des classes

Le jeu de données présentait un déséquilibre entre la classe majoritaire et la classe minoritaire. La variable cible `pnp` se composait de (68,7 %) observations pour la classe 0 (« pas de plainte ») et de (31,3 %) observations pour la classe 1 (« plainte »).

Bien que cet écart ne soit pas extrême, il est suffisant pour risquer de biaiser l'apprentissage du modèle en faveur de la classe la plus fréquente. Pour contrer ce phénomène, une stratégie de pondération a été mise en œuvre en utilisant le paramètre `class_weight='balanced'` de Scikit-learn. Cette approche ajuste le poids de chaque classe dans la fonction de coût de l'algorithme selon la formule suivante :

$$w_j = \frac{n}{k \times n_j}$$

Où : n est le nombre total d'échantillons , k est le nombre total de classes , n_j est le nombre d'échantillons appartenant à la classe j .

Application avec la variable `pnp`

Avec un total de $n = 47\,692$ échantillons répartis en $k = 2$ classes, les poids sont calculés comme suit :

- Pour la classe 0 (majoritaire, $n_0 = 32\,766$) :

$$w_0 = \frac{47692}{2 \times 32766} \approx 0,73$$

- Pour la classe 1 (minoritaire, $n_1 = 14\,926$) :

$$w_1 = \frac{47692}{2 \times 14926} \approx 1,60$$

Lors de l'entraînement, une erreur de classification sur un échantillon de la classe « plainte » (poids de 1,60) sera pénalisée environ **2,2 fois plus** qu'une erreur sur un échantillon « pas de plainte » (poids de 0,73).

5.2 Résultats

5.2.1 Observations Préliminaires

Avant de plonger dans les performances des modèles, une analyse exploratoire des données prétraitées a été menée pour mieux comprendre la distribution des caractéristiques et leurs associations avec la variable cible `pnp`.

Les figures suivantes mettent en lumière les associations entre chaque caractéristique et la variable cible `pnp` pour différents groupes de facteurs, ce qui est essentiel pour identifier visuellement les facteurs potentiellement influents sur la présence de signes neuropsychiques.



FIGURE 10 – Associations entre les variables et la variable cible PNP

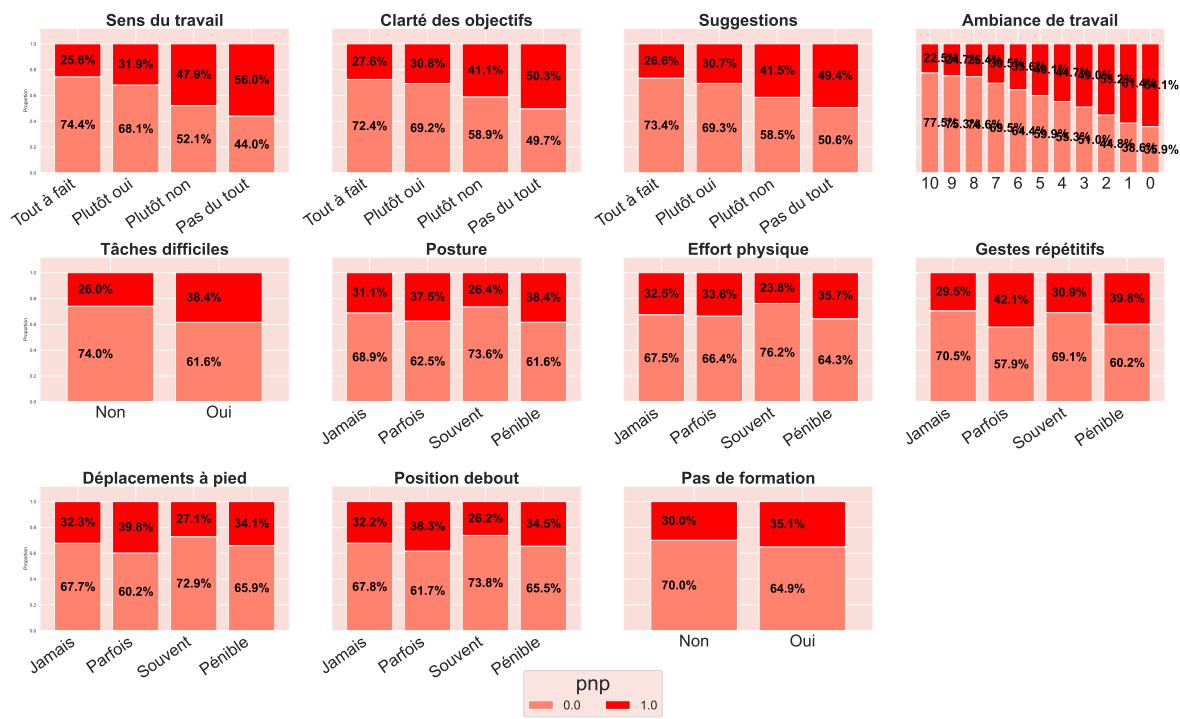


FIGURE 11 – Associations entre les variables et la variable cible PNP - Suite

L'analyse visuelle des figures 10 et 11 met en évidence une augmentation de PNP dans plusieurs situations et chez certains groupes de salariés. On observe ainsi que les femmes présentent une PNP plus élevé, tout comme les individus âgés de plus de 40 ans. Les plaintes sont également plus élevées chez ceux qui ne travaillent pas à temps plein, les personnes dépassant fréquemment leurs horaires normaux, ou celles qui sautent souvent un repas et ne prennent pas de pause. Une augmentation de PNP est également liée aux situations où les opérations sont traitées trop vite, ainsi qu'à une pression temporelle grandissante. Concernant la qualité et le développement professionnel, le manque de moyens pour réaliser un travail de qualité et le fait de n'apprendre rien dans son travail sont des facteurs associés à des PNP plus élevés. De même, un manque de travail varié, un déficit de liberté dans la façon de procéder, et un manque de coopération et de reconnaissance sont associés à une augmentation de plaintes. Une mauvaise ambiance de travail, l'absence de perspectives claires et d'objectifs, et la non-prise en compte des suggestions constituent également des facteurs de PNP. Enfin, la présence de pression psychologique, le fait de devoir faire des choses que l'on désapprouve (conflit de valeurs), la peur de perdre son emploi (insécurité), un manque de conciliation entre vie professionnelle et personnelle, et l'abandon de tâches perçues comme perturbantes sont autant de situations où les PNP sont accrues.

Par ailleurs, une projection t-SNE (t-distributed Stochastic Neighbor Embedding) a été réalisée afin de visualiser la structure des données en deux dimensions et d'apprécier la séparabilité des classes (Figure 12).

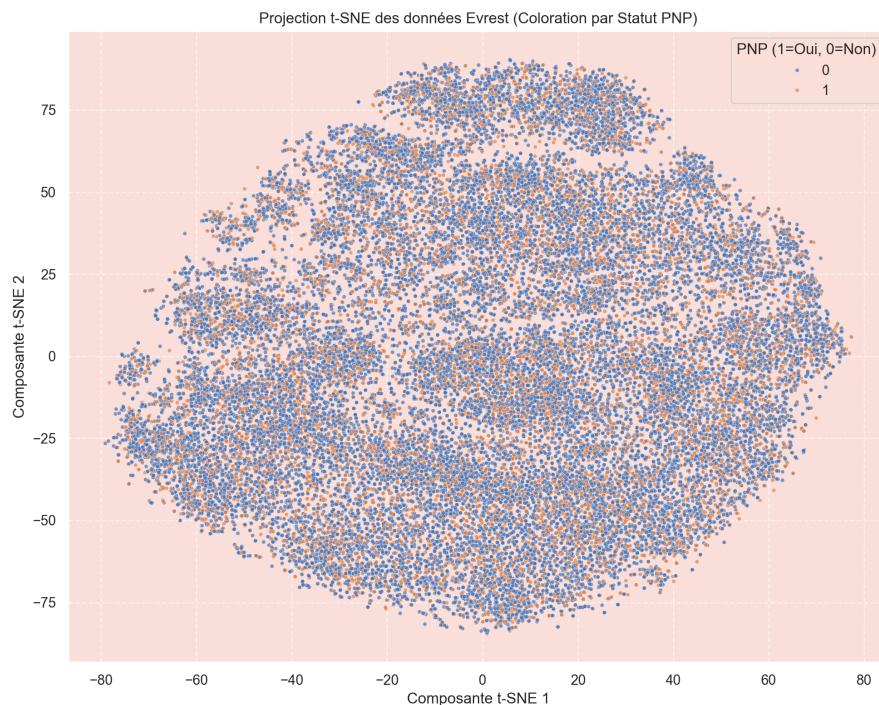


FIGURE 12 – Projection t-SNE des données EVREST (Coloration par Statut PNP).

L'examen de la projection t-SNE révèle une forte imbrication des deux classes ($\text{PNP}=0$ et $\text{PNP}=1$) dans l'espace réduit. Les points correspondant aux salariés avec plainte (classe 1) et ceux correspondant aux salariés sans plainte (classe 0) sont largement entremêlés, sans former de clusters distincts. Cette absence de séparation claire suggère que les salariés présentant des PNP ne sont pas facilement discriminables des autres salariés sur la base des caractéristiques disponibles. Cela indique une complexité intrinsèque à la détection de la classe " $\text{PNP}=1$ " et met en lumière le défi que les modèles de classification devront relever pour isoler cette classe minoritaire.

5.2.2 Aperçu des Performances des Modèles

Courbe ROC des différents modèles : La figure 13 représente des **courbes ROC** (Receiver Operating Characteristic) pour les différents modèles de classification. Les modèles ont été entraînés et optimisés sur l'ensemble d'entraînement (38 165 observations), puis évalués sur l'ensemble de test (9542 observations) pour évaluer leur capacité de généralisation.

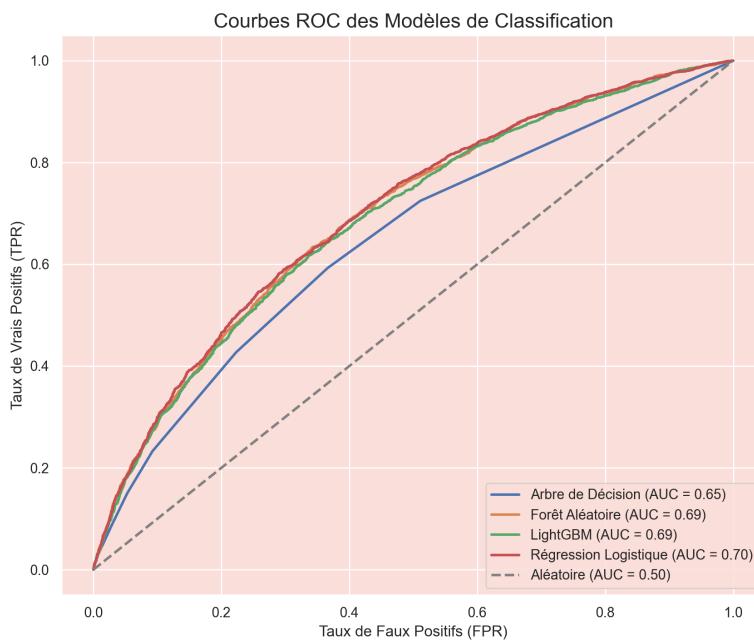


FIGURE 13 – Analyse des Courbes ROC.

- **Axe des X (Taux de Faux Positifs)** : Représente la proportion de cas négatifs qui ont été incorrectement classés comme positifs.
- **Axe des Y (Taux de Vrais Positifs)** : Représente la proportion de cas positifs qui ont été correctement identifiés.

L'**Aire Sous la Courbe (AUC)** est la métrique clé à observer. Elle quantifie la capacité d'un modèle à distinguer les classes positives des classes négatives. Une AUC de 1.0 est un modèle parfait, tandis qu'une AUC de 0.50 est un modèle qui ne fait pas mieux que le hasard.

- **Régression Logistique (AUC = 0.70)** : Ce modèle affiche la meilleure performance. Leurs courbes sont très proches et se situent bien au-dessus de la ligne du hasard, indiquant une bonne capacité de discrimination.

- **Forêt Aléatoire (AUC = 0.69)** et **LightGBM (AUC = 0.69)** : Ces modèles sont légèrement moins performant que le premier, mais la performance reste très proche. La différence d'AUC de 0.01 est minime.
- **Arbre de Décision (AUC = 0.65)** : Ce modèle a la performance la plus faible parmi tous et suggère qu'il a plus de difficulté à différencier correctement les classes.

Le tableau 4 présente les métriques de performance clés pour chaque modèle sur les données de test :

TABLE 4 – Performances des modèles de classification

Modèle	P_0	P_1	R_0	R_1	Accuracy
Arbre de Décision	0.8	0.39	0.49	0.72	0.56
Gradient Boosting	0.79	0.45	0.64	0.63	0.64
Régression Logistique	0.79	0.46	0.67	0.62	0.65
Forêt Aléatoire	0.79	0.47	0.70	0.59	0.66

Dans le tableau 4, les indices se réfèrent aux classes spécifiques :

- **P_0 : précision pour la classe 0** (Absence de plaintes neuropsychiques). C'est la proportion d'individus prédits comme n'ayant pas de plainte qui, en réalité, n'en ont pas.
- **P_1 : précision pour la classe 1** (Présence de plaintes neuropsychiques). C'est la proportion d'individus prédits comme ayant des plaintes qui, en réalité, en ont.
- **R_0 : rappel pour la classe 0** (Absence de plaintes neuropsychiques). C'est la proportion d'individus qui n'ont réellement pas de plaintes et qui ont été correctement identifiés comme tels.
- **R_1 : rappel pour la classe 1** (Présence de plaintes neuropsychiques). C'est la proportion d'individus qui ont réellement des plaintes et qui ont été correctement identifiés comme tels par le modèle.
- **Accuracy** : L'accuracy pour l'ensemble du modèle est calculée en considérant à la fois les prédictions correctes pour les deux classes.

L'analyse des résultats du tableau 4 met en évidence plusieurs points importants :

- **Rappel de la Classe 1 (R_1)** : On observe un rappel pour la classe 1 (salariés avec PNP) qui varie selon les modèles. L'**Arbre de Décision (AD)** se distingue particulièrement avec un rappel de **0.72**, ce qui signifie qu'il identifie correctement 72% des vrais cas de PNP. **Gradient Boosting (GB)** suit avec **0.63**, et la **Régression Logistique (RL)** avec **0.62**. La **Forêt Aléatoire (FA)** est un peu en retrait avec **0.59**.

- **Précision de la Classe 1 (P₁) :** La précision pour la classe 1 reste un point à surveiller. Elle varie de **0.39 (AD)** à **0.47 (FA)**. Ces valeurs indiquent que lorsqu'un modèle prédit qu'un salarié est à risque, cette prédiction est correcte dans moins de la moitié des cas (39% à 47%). Il y a donc un nombre substantiel de faux positifs.
- **Accuracy :** L'accuracy globale varie entre **0.56 (AD)** et **0.66 (FA)**. La Forêt Aléatoire et la Régression Logistique affichent les meilleures performances globales, avec des taux de prédictions correctes de 66% et 65% respectivement. L'Arbre de Décision a la plus faible accuracy, à 56%.
- **Comparaison des Modèles :**
 - L'**Arbre de Décision**, bien que le plus simple, affiche le meilleur rappel pour la classe 1 (**0.72**), ce qui est un atout majeur pour la détection des cas à risque. Cependant, sa précision (0.39) est la plus faible, conduisant à un nombre plus élevé de faux positifs.
 - La **Régression Logistique** présente la meilleure précision pour la classe 1 (**0.47**).
 - Le **Gradient Boosting** offre de bonnes performances, avec un rappel de la classe 1 à **0.63**.
 - La **Forêt Aléatoire** est comparable à la RL en termes de précision (0.46), mais son rappel pour la classe 1 (**0.59**) est légèrement inférieur à celui de LR et GB.

Ces premiers résultats sont prometteurs pour la capacité à identifier une partie des salariés à risque, avec une performance particulièrement notable du DT sur le rappel.

6 Objectif 3 : prédition des signes neuropsychiques (analyse longitudinale)

6.1 Méthodes

Notre troisième objectif est de construire et d'évaluer des modèles de *Machine Learning* pour prédire l'apparition ou non de plaintes neuropsychiques chez les salariés. Plus précisément, nous cherchons à déterminer la présence de ces plaintes à un instant $t + k$ (avec $6 \text{ mois} < k < 3,5 \text{ ans}$) en nous basant sur les expositions psychosociales et les caractéristiques individuelles du salarié observées à l'instant t .

Afin de préserver la cohérence méthodologique, nous avons repris la même sélection et les mêmes transformations de variables que pour l'objectif précédent. Après avoir retiré toutes les lignes contenant des valeurs manquantes, la taille finale de notre jeu de données s'élève à **90 518 lignes**.

Par la suite, nous avons défini de nouvelles variables que nous allons expliquer à l'aide d'un exemple :

TABLE 5 – Exemple de données

id	Année	pnp	pnp_next	next_year	gap
1	2016	0	1	2018	2.3
1	2018	1	0	2021	3.2
1	2021	0	NA	NA	NA
2	2018	1	NA	NA	NA
3	2015	0	0	2016	1.0
3	2016	0	0	2021	5.4
3	2021	0	NA	NA	NA

Les nouvelles variables sont :

- **pnp_next** : Pour un salarié ayant été vu à t , cette variable enregistre le statut pnp pour le suivi suivant. S'il n'y a pas de suivi ultérieur, la valeur est NA. C'est cette variable que l'on va chercher à prédire.
- **next_year** : Cette variable enregistre pour un salarié la date de sa prochaine visite de suivi. NA sinon.
- **gap** : Cette variable enregistre le temps, en années, entre le suivi à t et la prochaine visite. NA sinon.

6.1.1 Préparation des données et nettoyage

Nous avons sélectionné les observations où l'écart entre deux visites consécutives (**gap**) est compris entre 6 mois et 3,5 ans, tout en supprimant les lignes contenant des valeurs manquantes. Ces étapes de filtrage ont permis de ramener la taille finale de notre jeu de données à **31 845 observations**. Le tableau 6 montre l'effet de ce nettoyage du tableau 5.

TABLE 6 – Exemple de données

id	Année	pnp	pnp_next	next_year	gap
1	2016	0	1	2018	2.3
1	2018	1	0	2021	3.2
<hr/>					
3	2015	0	0	2016	1.0

Ainsi pour la variable **pnp**, nous observons 27,5% de plaintes et 72,5% sans plaintes. De même pour **pnp_next**, nous observons 27,8% de plaintes et 72,2% sans plaintes. Pour la variable **gap**, 52,6% des écarts sont compris entre 1 an \pm 6 mois, 39,3% entre 2 ans \pm 6 mois et 8,1% entre 3 ans \pm 6 mois.

Les données ont été divisées en ensembles d'entraînement **X_train**, **y_train** (80 % soit 25 701 observations) et de test **X_test**, **y_test** (20 % soit 6 369 observations) pour évaluer la capacité de généralisation des modèles. Les modèles sont les mêmes que dans l'objectif 2.

6.2 Résultats

6.2.1 Courbe ROC des différents modèles :

La Figure 14 illustre les performances de nos modèles de classification via les courbes ROC.

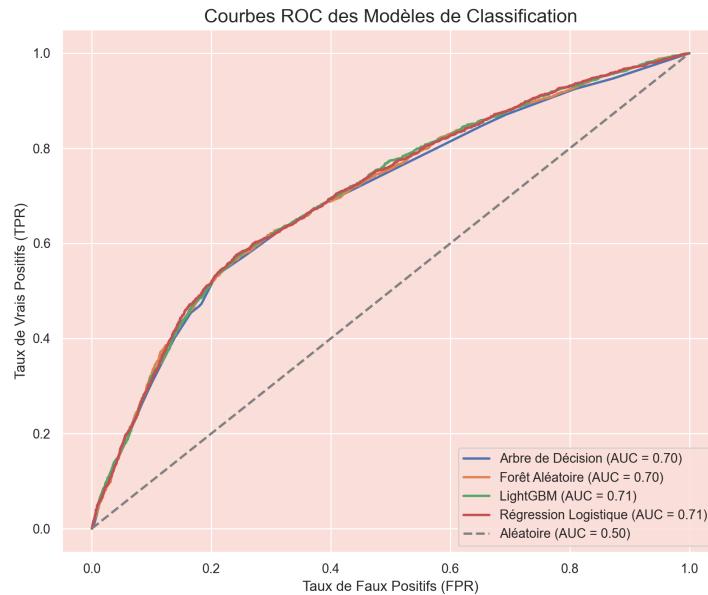


FIGURE 14 – Analyse des Courbes ROC.

Régression Logistique (AUC = 0.71), LightGBM (AUC = 0.71), Arbre de Décision (AUC = 0.70), Forêt Aléatoire (AUC = 0.70) : Leurs courbes se trouvent nettement au-dessus de la ligne aléatoire (AUC = 0.50), ce qui indique qu'ils sont bien plus performants que le hasard.

6.2.2 Performances des modèles de classification

Le tableau 7 présente les métriques de performance clés pour chaque modèle sur les données de test :

TABLE 7 – Performances des modèles de classification

Modèle	P_0	P_1	R_0	R_1	Accuracy
Gradient Boosting	0.82	0.45	0.72	0.60	0.69
Forêt Aléatoire	0.82	0.46	0.73	0.60	0.69
Régression Logistique	0.82	0.47	0.75	0.58	0.70
Arbre de Décision	0.82	0.46	0.74	0.57	0.70

L'analyse des résultats du tableau 7 met en évidence plusieurs points importants, en particulier concernant l'équilibre entre la détection des cas à risque et le nombre de fausses alertes :

- **Rappel de la Classe 1 (R₁)** : Le **Gradient Boosting** et La **Forêt Aléatoire** obtiennent le meilleur rappel avec **0,6**, ce qui signifie qu'ils détectent près de 60 % des cas réels de PNP. La **Régression Logistique** suit de près avec **0,58** et l'**Arbre de Décision** avec **0,57**.
- **Précision de la Classe 1 (P₁)** : La précision pour la classe 1 reste un défi pour tous les modèles, avec des valeurs allant de **0,45 à 0,47**. Cela signifie que lorsqu'un modèle prédit qu'un salarié est à risque, cette prédiction est correcte dans moins de la moitié des cas.
- **Accuracy** : L'Accuracy globale est relativement élevée pour tous les modèles (de **0,69 à 0,70**, ce qui s'explique en partie par la prédominance de la classe 0 (pas de plainte)).

En conclusion, les modèles présentés dans le second tableau affichent une **Accuracy** globale supérieure, signifiant une meilleure performance pour l'ensemble des prédictions. Cette amélioration est principalement due à leur grande efficacité à identifier les cas sans plainte (Classe 0), comme le montrent les hauts scores de **Précision (P₀)** et de **Rappel (R₀)**. Cependant, le **Rappel (R₁)** pour les cas à risque est moins bon que dans le premier tableau, et la **Précision (P₁)** reste faible. Cela indique que, malgré une meilleure performance globale, les modèles peinent toujours à identifier correctement et sans fausse alarme les cas de plaintes réelles.

7 Discussion

7.1 Résumé des résultats

Cette étude visait à explorer les liens entre les expositions psychosociales (EPS) et la santé mentale des salariés du secteur des Industries Électriques et Gazières (IEG), en se basant sur les données de l'observatoire EVREST. L'analyse a permis d'aborder les trois objectifs de recherche fixés initialement.

Premièrement, l'analyse factorielle exploratoire a montré que la structure des risques psychosociaux (RPS) dans notre population semble se regrouper en deux dimensions principales. Ces deux dimensions, la **Satisfaction et les Opportunités de Développement** et la **Pression et la Charge de Travail**, suggèrent la pertinence des items du questionnaire EVREST pour une évaluation des RPS dans ce contexte.

Deuxièmement, les modèles de machine learning mis en place pour la prédiction transversale des plaintes neuropsychiques ont montré des capacités de détection qui, bien que modestes, sont instructives. L'**Arbre de Décision** a affiché le meilleur **rappel (72%)** pour la classe des salariés avec plainte, ce qui le rend potentiellement utile pour identifier une proportion non négligeable de cas. Toutefois, la précision d'ensemble des modèles est restée faible, ce qui indique que l'on pourrait se retrouver avec un nombre important de faux positifs.

Troisièmement, la prédiction longitudinale a donné des résultats comparables à l'analyse transversale, avec une précision légèrement supérieure. Le **Gradient Boosting** et la **Forêt Aléatoire** ont atteint un **rappel de 60%**, un résultat qui met en lumière la difficulté de prédire à moyen terme un phénomène aussi complexe que les plaintes neuropsychiques, influencé par une multitude de facteurs externes (professionnels et extra-professionnels) et personnels fluctuants.

En résumé, ces modèles prédictifs peuvent être considérés comme un premier pas vers un outil de détection des salariés les plus à risque de souffrance au travail.

7.2 Forces et limites

7.2.1 Points positifs de l'étude

- **Étendue des données :** L'utilisation de la base EVREST-IEG, avec un grand nombre de questionnaires sur une longue période, a permis d'obtenir un échantillon substantiel. Cela offre une base solide pour les analyses statistiques et pour l'entraînement des modèles de machine learning.
- **Cohérence des dimensions identifiées :** L'analyse factorielle a révélé deux facteurs psychométriques fiables, confirmant que les items du questionnaire EVREST

mesurent bien des concepts cohérents. La fiabilité interne de ces facteurs, mesurée par un alpha de Cronbach supérieur à 0.80, constitue un point de confiance pour la suite de l'analyse.

- **Potentiel de la dimension longitudinale** : La structure des données, qui permet le suivi des mêmes salariés au fil du temps, est un atout majeur. Cela nous a offert la possibilité de dépasser la simple analyse transversale et d'explorer les évolutions, même si la prédiction à long terme reste complexe.

7.2.2 Limites à considérer

- **Nature des données cliniques** : Les plaintes neuropsychiques sont basées sur les échanges avec les personnels des SPST (médecins ou infirmiers) et ne constituent pas des diagnostics médicaux validés. Elles n'apportent pas d'informations sur la gravité ou la persistance des troubles.
- **Généralité des variables** : Certaines variables souffrent d'un codage très général (ex : "Exposition à la pression psychologique"), ce qui a pu limiter la précision des modèles. Des variables plus spécifiques sur les sources de stress pourraient permettre des modèles plus performants.
- **Construction de la variable cible** : Le questionnaire EVREST n'inclut pas d'outil de mesure de souffrance psychique (dépression, anxiété, etc.) validé. Nous avons dû construire la variable cible "plaintes neuropsychiques" à partir d'une combinaison de trois variables (anxiété, fatigue/lassitude, troubles du sommeil). Ce choix peut potentiellement biaiser l'analyse en capturant incomplètement les différentes facettes de la souffrance psychique.
- **Association plutôt que causalité** : Les modèles prédictifs mettent en lumière des associations entre les expositions psychosociales et l'apparition de plaintes, mais ils ne prouvent en aucun cas une relation de cause à effet directe. Il est important de reconnaître que d'autres facteurs non mesurés par le questionnaire **EVREST**, tels que des aspects de la vie personnelle, peuvent avoir une influence significative. De plus, il est possible d'imaginer que des déterminants communs, comme la susceptibilité individuelle, soient à l'origine à la fois des contraintes psychosociales vécues et des signes neuropsychiques. Une autre hypothèse serait que la souffrance psychique elle-même rende le vécu des contraintes psychosociales plus douloureux.

7.3 Perspectives

Cette étude a ouvert plusieurs pistes de réflexion pour de futurs travaux de recherche :

- **Analyse factorielle par sous-groupes** : Pour affiner l'analyse, une exploration de la structure factorielle séparément pour les hommes et les femmes pourrait révéler des dimensions de risques psychosociaux distinctes. Cette approche permettrait d'observer si les facteurs d'exposition diffèrent en fonction du genre.
- **Amélioration des modèles de prédiction** :
 - **Ingénierie de variables** : Intégrer les scores factoriels issus de la première partie de l'étude dans les modèles prédictifs pourrait être une piste. Ces nouvelles variables, plus représentatives des dimensions sous-jacentes des RPS, pourraient potentiellement améliorer les performances et l'interprétabilité des modèles.
 - **Test d'autres algorithmes** : L'exploration de modèles plus complexes, tels que les réseaux de neurones, pourrait être envisagée pour voir s'ils sont capables de capturer des interactions plus fines entre les variables et d'améliorer la précision.
 - **Définition alternative de la variable cible** : La variable 'pnp' que nous avons construite pourrait être définie de manière différente. On pourrait par exemple utiliser une approche basée sur des échelles ou des scores combinés, afin d'explorer si cela permettrait d'améliorer la performance des modèles.
 - **Gestion du déséquilibre des classes** : La difficulté des modèles à identifier les cas de plaintes pourrait être en partie liée au déséquilibre des données. Bien que la pondération des classes ait été utilisée, d'autres techniques de rééquilibrage pourraient être explorées pour améliorer la détection des cas de plaintes. Des méthodes telles que le suréchantillonnage de la classe minoritaire (oversampling, ex : SMOTE) ou le sous-échantillonnage de la classe majoritaire (undersampling) pourraient permettre aux modèles d'apprendre plus efficacement [11].
- **Approfondissement des analyses longitudinales** :
 - **Modèles de transition d'état** : Il serait pertinent d'explorer des modèles capables de prédire le passage d'un état de non-plainte à un état de plainte neuropsychique. On pourrait également s'intéresser au temps nécessaire pour qu'une telle transition se produise en fonction des expositions.
 - **Analyse des interactions** : Il serait intéressant d'étudier comment certains facteurs personnels ou professionnels (par exemple, le soutien social ou le type de management) peuvent moduler l'impact des expositions psychosociales sur la santé mentale des salariés.

8 Bibliographie

- [1] ÉQUIPE PROJET NATIONALE EVREST. *Guide de remplissage du questionnaire Evrest Année 2025*. Guide applicable pour l’année 2025. 2024. URL : https://evrest.istnf.fr/_docs/Fichier/2024/4-241216025139.pdf.
- [2] INSEE. *Nomenclatures des professions et catégories socioprofessionnelles des emplois salariés des employeurs privés et publics*. <https://www.insee.fr/fr/information/2497958>. 2003.
- [3] ÉQUIPE PROJET NATIONALE EVREST. *Réaliser un “Evrest en entreprise” : Guide méthodologique*. Coordination : Céline Mardon et Marie Murcia. Mars 2017. URL : https://evrest.istnf.fr/_docs/Fichier/2017/4-170411062747.pdf.
- [4] COLLÈGE D’EXPERTISE SUR LE SUIVI DES RISQUES PSYCHOSOCIAUX AU TRAVAIL. *Mesurer les facteurs psychosociaux de risque au travail pour les maîtriser*. Rapp. tech. Rapport présidé par Michel Gollac. Ministère du Travail, de l’Emploi et de la Santé, 2011. URL : https://travail-emploi.gouv.fr/sites/travail-emploi/files/files-spip/pdf/rapport_SR PST_definitif_rectifie_11_05_10.pdf.
- [5] THE PENNSYLVANIA STATE UNIVERSITY. *Module 7.3 : The Factor Analysis Model*. <https://online.stat.psu.edu/stat505/lesson/12>.
- [6] Henry F KAISER. « The varimax criterion for analytic rotation in factor analysis ». In : *Psychometrika* (1958).
- [7] Francisco P HOLGADO-TELLO et al. « Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis with ordinal variables ». In : *Quality Quantity* (2010).
- [8] Peter SAMUELS. *Advice on Exploratory Factor Analysis*. Rapp. tech. University of Wolverhampton, juin 2017. URL : https://www.researchgate.net/publication/319165677_Advice_on_Exploratory_Factor_Analysis.
- [9] Marley W. WATKINS. *Exploratory Factor Analysis : A Step-by-Step Guide Using R*. Routledge, 2021.
- [10] Amy S. BEAVERS et al. *Practical Considerations for Using Exploratory Factor Analysis in Educational Research*. 2013.
- [11] Marini SYSTEMS. *Unbalanced classes (Machine Learning)*. URL : <https://marini.systems/en/glossary/unbalanced-classes-machine-learning/>.

Annexe A : Questionnaire EVREST

En Orange ce qui concerne l'objectif 1 seulement. En Orange et Jaune l'objectif 2.

Date du jour : ___ / ___ / ___ Nom du Médecin ou Infirmier: _____ SST: _____ SAISIE:



EVREST IEG 2024

Nom naiss[] Prénom[] Sexe M/F [] Date naiss[]
 Dép. naissance [] Salarié [] Contrat : CDI ou assimilé Autre PCS-ESE
 Entreprise [] NAF2008 [] Nb salariés []
 Etablissement de type : Privé Public Suivi individuel renforcé : oui non
 Atelier (facultatif) [] Champ libre (facultatif) []
 Entretien réalisé en présentiel [] distanciel [] Echantillon : représentatif [] ciblé []

Conditions de travail

En remplissant ce questionnaire, je reconnaiss avoir pris connaissance et accepter les termes de la note d'information sur le dispositif Evrest.

1. Travaillez-vous à temps plein ? Oui, Non,

2. Habituellement, travaillez-vous en journée normale ? Oui, Non,

- Avez-vous régulièrement :
- Des coupures de plus de 2 heures
 - Des horaires décalés (tôt le matin, tard le soir)
 - Des horaires irréguliers ou alternés
 - Du travail de nuit (entre 0h et 5h)

Faites-vous régulièrement des déplacements professionnels de plus de 24h ? Oui, Non,

Actuellement, vous travaillez : Sur site exclusivement, Sur site et en télétravail, En télétravail exclusivement,

3. Contrainte de temps :

- a) En raison de la charge de travail, vous arrive-t-il de :
- Dépasser vos horaires normaux
 - Sauter ou écouter un repas, ne pas prendre de pause
 - Traiter trop vite une opération qui demanderait davantage de soin
 - Travailler chez vous sur vos temps de repos, de congés

Jamais ₀	Rarement ₁	Assez souvent ₂	Très souvent ₃
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

b) Pouvez-vous coter les difficultés liées à la pression temporelle (devoir se dépêcher, faire tout très vite, ...)

Pas difficile 0 1 2 3 4 5 6 7 8 9 10 Très difficile (Entourer un chiffre)

c) Devez-vous fréquemment abandonner une tâche que vous êtes en train de faire pour une autre non prévue ?

Oui, Non,

Si oui, diriez-vous que cette interruption d'activité : - perturbe votre travail Oui, Non,
 - est un aspect positif de votre travail Oui, Non,

4. Appréciations sur votre travail :

- | | Non pas du tout ₀ | Plutôt Non ₁ | Plutôt oui ₂ | Oui tout à fait ₃ |
|---|------------------------------|--------------------------|--------------------------|------------------------------|
| - Votre travail vous permet d'apprendre des choses | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Votre travail est varié | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Vous pouvez choisir vous-même la façon de procéder | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Vous avez des possibilités suffisantes d'entraide, de coopération | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Vous avez les moyens de faire un travail de bonne qualité | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Votre travail est reconnu par votre entourage professionnel | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Vous devez faire des choses que vous désapprouvez | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Vous travaillez avec la peur de perdre votre emploi | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| - Vous arrivez à concilier vie professionnelle et vie hors-travail | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

5. Charge physique du poste de travail : votre poste de travail présente-t-il les caractéristiques suivantes ?

	Non jamais ₀	Oui parfois ₁	Oui souvent ₂	Si oui, est-ce difficile ou pénible ?
Postures contraignantes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Si OUI→ Oui <input type="checkbox"/> Non <input type="checkbox"/>
Effort, Port de charges lourdes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Si OUI→ Oui <input type="checkbox"/> Non <input type="checkbox"/>
Gestes répétitifs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Si OUI→ Oui <input type="checkbox"/> Non <input type="checkbox"/>
Importants déplacements à pied	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Si OUI→ Oui <input type="checkbox"/> Non <input type="checkbox"/>
Station debout prolongée	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Si OUI→ Oui <input type="checkbox"/> Non <input type="checkbox"/>

6. Etes-vous exposé à :

Oui <input type="checkbox"/>	Non <input type="checkbox"/>										
Produits chimiques	<input type="checkbox"/>	<input type="checkbox"/>	Gêne sonore	<input type="checkbox"/>	<input type="checkbox"/>	Chaleur intense	<input type="checkbox"/>	<input type="checkbox"/>	Risque infectieux	<input type="checkbox"/>	<input type="checkbox"/>
Poussières, fumées	<input type="checkbox"/>	<input type="checkbox"/>	Bruit > 80db	<input type="checkbox"/>	<input type="checkbox"/>	Froid intense	<input type="checkbox"/>	<input type="checkbox"/>	Contact avec le public (usagers, patients, clients, élèves....)	<input type="checkbox"/>	<input type="checkbox"/>
Ray. ionisants	<input type="checkbox"/>	<input type="checkbox"/>	Contrainte visuelle	<input type="checkbox"/>	<input type="checkbox"/>	Intempéries	<input type="checkbox"/>	<input type="checkbox"/>			
Vibrations	<input type="checkbox"/>	<input type="checkbox"/>	Conduite routière prolongée	<input type="checkbox"/>	<input type="checkbox"/>	Pression psychologique	<input type="checkbox"/>	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>

Formation – Encadrement – Parcours professionnel
1. Depuis 1 an, avez-vous eu une formation ?

Si oui, était-ce : en rapport avec votre travail actuel
en rapport avec un futur poste

Oui Non

Oui Non

2. Avez-vous un rôle de formateur, de tuteur ?

Oui Non

3. Avez-vous un ou plusieurs salariés sous vos ordres ou votre autorité ?

Oui Non

4. Depuis 2 ans, avez-vous changé de travail ?

Si oui, était-ce pour raison médicale ?

Oui Non

Oui Non

5. Pensez-vous que dans 2 ans votre état de santé vous permettrait d'effectuer votre travail actuel ?

Non, sans doute pas Ce n'est pas sûr Oui, c'est à peu près certain

Mode de vie
1. Faites-vous de façon régulière (au moins 1 fois/semaine) une activité physique ou sportive ?

Oui Non

2. Consommation usuelle :

- Tabac (nb de cig/jour) Non fumeur Ancien fumeur < 5 cig 5 à 15 cig > 15 cig

- A quelle fréquence vous arrive-t-il de consommer des boissons contenant de l'alcool ?

Jamais ou 1 x / mois 2 à 4 x / mois 2 à 3 x / semaine 4 x / semaine ou plus

- Combien de verres standards buvez-vous au cours d'une journée ordinaire où vous buvez de l'alcool ?

Non concerné (non buveur) 1 ou 2 3 ou 4 5 ou 6 7 à 9 10 ou plus

3. Avez-vous des trajets domicile/travail longs ou pénibles ?

Oui Non

État de santé actuel = les 7 derniers jours (à remplir par le médecin ou l'infirmier·e)

Questionnaire renseigné par : le médecin, l'infirmier(e), Nom IdEST _____

Dernier entretien santé-travail (hors reprise, à la demande, ...) il y a : ⚡ 1 an 2 ans 3 ans 4 ans 5 ans ou + jamais

Lien avec le travail actuel : 0 = aucun lien / 1 = lien peu probable / 2 = lien probable / 3 = lien très probable

Poids : ____ kg Taille : ____ cm

		Plaîntes ou signes cliniques au cours des 7 derniers j	Est-ce une gêne dans le travail ?	Traitemennt ou autre soin	Lien avec le travail actuel
RAS <input type="checkbox"/>	Cardio-respiratoire	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- appareil respiratoire	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- appareil cardio-vasculaire	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- HTA	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	Neuro-psychique				
RAS <input type="checkbox"/>	- fatigue, lassitude	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- anxiété, nervosité, irritabilité	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- troubles du sommeil	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	Digestif	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	Ostéo-articulaire				
RAS <input type="checkbox"/>	- épaule	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- coude	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- poignet / main	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- membres inférieurs	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- vertèbres cervicales	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	- vertèbres dorso-lombaires	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	Dermatologie	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>
RAS <input type="checkbox"/>	Troubles de l'audition	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	Oui, <input type="checkbox"/> Non <input type="checkbox"/>	<u>1_1</u>

Date du jour : ___ / ___ / ___ Nom du Médecin: _____ SST : _____ SAISIE

N° Salarié : _____

Questionnaire complémentaire IEG (à remplir par le salarié)

Pour les questions suivantes, saisir le chiffre correspondant le mieux à votre réponse :

0 = Non pas du tout / 1 = Plutôt non / 2 = Plutôt oui / 3 = Oui tout à fait

(Q1) Avez-vous le sentiment d'avoir des perspectives dans votre métier ou dans votre entreprise ?|_|

(Q2) Trouvez-vous du sens et de l'intérêt dans le travail que vous effectuez ?|_|

(Q3) Pour effectuer votre travail, avez-vous des objectifs et des consignes clairement définis ?|_|

(Q4) Pouvez-vous agir sur votre organisation de travail en exprimant votre point de vue et en faisant des suggestions ?|_|

(Q5) Pouvez-vous coter votre ambiance de travail ?



(Q6)|_|_|

(Q7)|_|_|

(Q8)|_|_|

(Q9)|_|_|

(Q10)|_|_|

Annexe B : Variables retenues pour l'objectif 1

Nom de variable	Libellé	Codage
depasse_horaire	Fréquence dépassement horaires	0 = Jamais → 3 = Souvent
trop_vite	Fréquence travail sous pression (traitement trop rapide)	0 = Jamais → 3 = Souvent
pression	Difficultés liées à la pression temporelle	Score de 0 à 10. <i>Recodée en 4 catégories basées sur les quartiles pour harmoniser l'échelle.</i>
tache_diff	Abandon de tâche pour imprévu	1 = Oui / 0 = Non
psy	Pression psychologique liée au travail	1 = Oui / 0 = Non
apprend	Manque d'opportunités d'apprendre	0 = Permet d'apprendre → 3 = Ne permet pas. <i>Variable inversé.</i>
varie	Manque de variété dans le travail (Monotonie)	0 = Varié → 3 = Pas varié. <i>Variable inversée.</i>
coop	Manque d'entraide / coopération	0 = Suffisante → 3 = Insuffisante. <i>Variable inversée.</i>
reconnu	Manque de reconnaissance	0 = Reconnu → 3 = Pas reconnu. <i>Variable inversée.</i>
perspective	Sentiment d'avoir des perspectives dans son métier ou entreprise (Développement des compétences)	0 = Non pas du tout → 3 = Oui tout à fait
sens	Trouver du sens et de l'intérêt dans le travail (Qualité du travail)	0 = Non pas du tout → 3 = Oui tout à fait
suggestion	Possibilité d'exprimer son point de vue et faire des suggestions (Autonomie et participation)	0 = Non pas du tout → 3 = Oui tout à fait
ambiance	Ambiance de travail (Rapports sociaux)	0 = Très mauvaise → 10 = Très bonne

Annexe C : Variables retenues pour l'objectif 2 et 3

Similaire à L'AFE avec les variables supplémentaires suivantes :

Variable	Description	Codage
sexe	Sexe	1 = Homme / 2 = Femme
age	Age	0 : < 30 ans / 1 : entre 30 et 40 ans / 2 : entre 40 et 50 ans / 3 : > 50 ans
pcs	Code PCS-ESE 2003 - niveau 1	3 : Cadres et professions intellectuelles supérieures / 4 : Profession intermédiaires / 5 : Employés / 6 : Ouvriers
tps_plein	Travaillez-vous à temps plein ?	1 = Oui / 0 = Non
jour_norm	Habituellement travaillez-vous en journée normale ?	1 = Oui / 0 = Non
jour_coup	Des coupures de plus de 2 heures	1 = Oui / 0 = Non
jour_decale	Des horaires décalés (tôt le matin, tard le soir)	1 = Oui / 0 = Non
jour.Alterne	Des horaires irréguliers ou alternés	1 = Oui / 0 = Non
travail_nuit	Du travail de nuit (entre 0h et 5h)	1 = Oui / 0 = Non
deplac_pro	Faites-vous régulièrement des déplacements professionnels de plus de 24 heures ?	1 = Oui / 0 = Non
saut_repas	Fréquence saut de repas	0 = Jamais → 3 = Souvent
concilier	Difficulté à concilier vie pro/perso	0 = Pas de difficulté → 3 = Difficulté importante. <i>Variable inversé. Score élevé = Difficulté.</i>
libre	Manque de liberté/autonomie	0 = Autonomie → 3 = Pas d'autonomie. <i>Variable inversée.</i>
objectif	Objectifs et consignes clairement définis (Clarté du travail)	0 = Non pas du tout → 3 = Oui tout à fait
peur_emploi	Peur de perdre son emploi (Insécurité de la situation de travail)	0 = Non pas du tout → 3 = Oui tout à fait

Suite à la page suivante

Table 9 – suite de la page précédente

Variable	Description	Codage
qualite_travail	Manque de moyens pour un travail de qualité	0 = Moyens suffisants → 3 = Pas de moyens. <i>Variable inversée.</i>
desapprouve	Devoir faire des tâches désapprouvées (Conflit éthique)	0 = Non → 3 = Oui (variable déjà orientée risque)
sport	Activité physique ou sportive régulière (au moins 1 fois/semaine)	1 = Oui / 0 = Non
trajet	Trajets domicile/travail longs ou pénibles	1 = Oui / 0 = Non
pnp	Plainte NeuroPsychique (PNP) : présence d'au moins l'une des trois plaintes neuropsychiques (anxiété, fatigue ou troubles du sommeil)	1 = Oui / 0 = Non
posture	Postures contraignantes & pénibilité	1 = Oui / 0 = Non
effort	Effort, port de charges lourdes & pénibilité	1 = Oui / 0 = Non
repetitif	Gestes répétitifs & pénibilité	1 = Oui / 0 = Non
pieton	Importants déplacements à pied & pénibilité	1 = Oui / 0 = Non
debout	Station debout prolongée & pénibilité	1 = Oui / 0 = Non
pas_formation	Absence de formation depuis 1 an	1 = Oui / 0 = Non
tabac	Consommation de tabac	1 = Oui / 0 = Non

Annexe D : Dépôt GitHub du Projet

Le code source incluant le fichier `R Markdown` pour l'analyse factorielle (Objectif 1), ainsi que les `notebooks Python` (pour la description de la population, l'Objectif 2 et l'Objectif 3), sont disponibles publiquement sur le dépôt GitHub suivant :

<https://github.com/derghalmanal/Analyse-EVREST-IEG>

Pour des raisons de confidentialité et de protection des données des salariés, le fichier de données source n'a pas été inclus.