Université de Lille

PROJET DE FÉVRIER
Identification de profils et parcours type pour faciliter la prise en charge et la réintégration professionnelle des nouveaux arrivants dans l'entreprise de réinsertion VitamineT.

Étudiants
DANÈS Élorine
DELBERGUE Clémence
WATIOTIENNE Henry
MASTER M2 MA ISN

Professeurs :
Madame DUVAL Céline -
Monsieur Wicker Nicolas
Année 2023-2024

21 Février 2024

Remerciements

Un grand merci à nos référents, Madame Céline Duval et Monsieur Nicolas Wicker, pour leur aide et leur disponibilité tout au long de notre projet. Leurs conseils ont été d'une grande aide dans sa réalisation et ont grandement contribué à l'aboutissement de notre travail.

Nous souhaitons également remercier chaleureusement Monsieur Fabrice Denoual et Monsieur Tanguy Desauw, qui ont pris de leur temps pour nous faire une visite passionnante de l'entreprise et nous expliquer son fonctionnement. Nous les remercions de plus, ainsi que toutes les autres personnes rencontrées, pour nous avoir partagé leurs expériences et répondus à nos nombreuses questions. Cette expérience nous a permis de saisir des aspects pratiques importants qui ont éclairé notre compréhension du sujet et alimenté notre réflexion.

Introduction

Dans le cadre de notre formation en Master, nous avons acquis une variété de techniques avancées en traitement de données, notamment en matière de classification et de régression. Ces compétences nous ont permis de maîtriser les fondamentaux de l'analyse et de la manipulation de données structurées (comme les nombres concernant l'âge ou encore les données catégorielles comme la civilité...). Cependant, ce projet représente une opportunité unique de mettre en pratique ces connaissances dans un contexte professionnel réel, mais également de relever le défi de traiter des données non structurées (c'est-à-dire du texte libre comme les colonnes commentaire par exemple), un domaine jusqu'alors peu exploré dans notre cursus.

L'objectif principal de notre travail est de concevoir, en collaboration avec nos professeurs et à travers nos propres recherches, une stratégie méthodique pour passer d'une problématique métier complexe à une solution implémentable. Pour cela, nous allons présenter la problématique métier ainsi que l'entreprise concernée, identifier le problème spécifique à résoudre et définir la solution demandée. Notre objectif sera alors de simplifier cette problématique pour en faciliter la résolution.

Le sujet de notre projet porte sur l'identification de profils et parcours types pour faciliter la prise en charge et la réintégration professionnelle des nouveaux arrivants dans l'entreprise de réinsertion VitamineT. Nous explorerons les différentes étapes nécessaires à cette démarche, en mettant en avant les défis spécifiques rencontrés dans ce contexte.

Nous débuterons par une analyse approfondie des données à notre disposition. Cette étape comprendra la présentation de ces données, leur nettoyage global et une étude approfondie des différentes tables les contenant. Ensuite, nous procéderons à la modélisation, en présentant nos démarches avec leurs résultats concrets, tel que la prédiction de sorties ou encore la recherche de profils types et des actions associées, puis en évaluant les possibilités d'amélioration.

Nous aborderons au passage la phase de prédiction, où nous explorerons les capacités de notre modèle à anticiper des résultats futurs. Enfin, nous discuterons des limites de notre approche, notamment en termes de fiabilité et de généralisation, avant de conclure sur les enseignements tirés de ce projet et les perspectives d'amélioration futures.

Ainsi, ce plan nous permettra de guider notre réflexion et notre travail tout au long de ce projet, en nous aidant à structurer nos analyses et nos actions pour répondre de la manière la plus efficace et pertinente à la problématique métier qui nous est posée.

Table des matières

Chapitre 1

PRÉSENTATION DE LA PROBLÉMATIQUE MÉTIER

Dans le cadre d'un projet basé sur les données, notre objectif principal est d'utiliser ces informations pour répondre à des questions spécifiques ou résoudre des problèmes relatifs au secteur d'activité de l'entreprise concernée. Pour mener à bien un tel projet, il nous est important de bien cibler la problématique métier, c'est-à-dire la question à laquelle il nous faut répondre. Cela permet de comprendre le besoin, de choisir les outils et les méthodes adaptées, de mobiliser les acteurs impliqués si besoin et de définir les critères de succès.

1.1 Présentation de l'entreprise

VitamineT est une entreprise sociale créée en 1978, qui se positionne comme le premier acteur français de l'insertion par l'activité économique. Son objectif est d'accompagner les personnes qui ont du mal à accéder au marché du travail en leur offrant un parcours de retour à l'emploi ou à la formation. Pour cela, elle réalise un diagnostic préalable pour identifier les difficultés rencontrées, et met en place un suivi personnalisé par des conseillers et des encadrants. Chaque personne est alors encadrée pendant deux ans maximum.

VitamineT intervient dans quatre régions, notamment les Hauts-de-France, l'Île-de-France, le Grand-Est et la Bourgogne-Franche-Comté. Avec pour slogan *L'inclusion par l'action*, l'entreprise affirme sa volonté de favoriser l'insertion professionnelle des personnes en situation de précarité.

Elle leur offre des programmes d'accompagnement sur mesure ainsi qu'une formation adaptée à leurs besoins et aux demandes du marché de l'emploi. L'entreprise dispose d'ailleurs de près de 31 filiales, qui couvrent plusieurs domaines d'activités, tels que les services, l'industrie et le recyclage, l'alimentation, ainsi que les solutions en ressources humaines.

Reconnue pour son engagement en faveur de l'inclusion, son ancrage territorial et son innovation, VitamineT joue un rôle essentiel dans la création de débouchés professionnels pour les publics vulnérables, participant ainsi à l'édification d'une société plus équitable et solidaire.

1.2 Problème posé

Grâce à l'action de ses nombreux encadrants et conseillers, l'entreprise affiche aujourd'hui un taux de 66% de sorties positives, c'est-à-dire de sorties qui permettent aux individus d'accéder à un emploi ou une solution durable, et cela, dans un délai maximum de deux ans. Pour chaque profil, des solutions adaptées sont mises

en place à partir d'un diagnostic initial. Toutes ces démarches, et notamment chaque étape du parcours d'un individu, sont enregistrées dans des bases de données.

L'entreprise souhaite exploiter ces données, qui ont jusqu'ici une valeur purement administrative, dans le but de comprendre pourquoi la réinsertion échoue pour un individu sur trois, mais aussi de valoriser le travail accompli par les encadrants.

Si certaines de ces données sont structurées, comme celles qui concernent les critères d'état (critères d'entrée et de sortie), et sont déjà utilisées pour des analyses statistiques descriptives, il n'en est pas de même pour les nombreuses données textuelles saisies par les encadrants, qui décrivent les différents parcours entre l'entrée et la sortie. Ces données sont une source d'information précieuse, mais elles ne sont pas exploitées, car leur absence de structuration les rend difficiles à traiter.

1.3 Solution Demandée

Afin d'approfondir notre compréhension des enjeux de l'entreprise et du rôle de ses différentes entités, nous avons échangé avec divers collaborateurs du pôle de Lesquin. Ces discussions ont permis de déterminer les objectifs à fixer pour notre projet.

Dans ce cadre, VitamineT a souligné l'importance de classifier les données non structurées, en mettant particulièrement l'accent sur les données textuelles décrivant les actions mises en place, leurs objectifs et leurs résultats. L'objectif est de transformer ces données textuelles en données catégorielles pour rendre ces données facilement exploitables à l'avenir.

En outre, il nous a bien sûr été demandé de mettre en place un modèle permettant d'identifier les profils type ainsi que les parcours d'accompagnement conduisant à une réintégration professionnelle durable. Aucune directive précise n'a toutefois été donnée quant à la forme du rendu final ou au niveau de détail requis pour établir ces parcours. Nous bénéficions ainsi d'une liberté totale dans notre démarche.

1.4 Notre Objectif

Conformément à la demande de l'entreprise, nous souhaitions que la priorité soit donnée à l'organisation des données textuelles, notamment en ce qui concerne les résultats des actions. Nous pensions que cela constituerait un bon point de départ pour pouvoir ensuite sélectionner celles qui sont les plus prometteuses pour différents profils d'individus.
Cependant, face à la diversité et à la qualité des données, ainsi qu'aux défis techniques du traitement du langage naturel, nous avons dû adapter notre objectif pour mieux aborder cette problématique et ces données complexes.

Pour ce projet, nos objectifs finaux auront donc d'abord été de nous assurer de comprendre pleinement les données à notre disposition et leurs liens, puis de parvenir à créer un premier modèle à partir des données structurées à notre disposition et enfin chercher à l'affiner pour intégrer les données non structurées.

Ce sont ces différentes étapes de notre réflexion vont constituer la structure même de ce rapport : partir du plus simple et aller vers le plus complexe afin de nous rapprocher le plus possible du résultat souhaité par l'entreprise.

Chapitre 2

ANALYSE DES DONNÉES À NOTRE DISPOSITION

L'analyse descriptive représente une étape incontournable dans le bon déroulement d'un projet axé sur les données. Cette phase nécessite une exécution méthodique pour assurer une préparation optimale des données en vue de leur utilisation dans un modèle. Cela permet une évaluation rigoureuse de la qualité et de la pertinence des données collectées, la détermination du modèle analytique le plus approprié pour les interpréter et la décision quant à l'éventuelle exigence d'un nettoyage préliminaire des données.

2.1 Présentation des données à notre disposition

Afin de réaliser l'objectif demandé, un document `donnees.xlsx` contenant plusieurs feuilles Excel nous a été fourni par l'entreprise. Ces feuilles représentent les différentes avancées du suivi d'une personne, c'est-à-dire les différentes étapes de leur parcours. Il y en a cinq : `difficulté_export`, `action_export`, `formation_export`, `projet_export` et `sortie_export`. Ainsi, ces données, anonymes, fournissent des informations sur le profil des individus, les difficultés rencontrées, les solutions mises en place (actions ou formations suivies), ainsi que l'évaluation de leur situation à la sortie du programme.

2.1.1 Feuille 1 : `difficultes_export`

Cette feuille dresse un inventaire des difficultés rencontrées par les individus suivis dans le programme. Chaque ligne correspond à un enregistrement de difficulté distinct, lié à une personne par une clé unique.

Cette table de données est composée de 14852 lignes et de 16 colonnes. Ces lignes, identifiées par une clé, ne représentent pas chacune un individu différent car plusieurs clés de cette table ne sont pas uniques : au total, nous obtenons 6489 clés (et donc individus). Cela s'explique par le fait qu'un individu peut avoir plusieurs difficultés.

Les colonnes incluent des informations suivantes : la clé, la civilité, la situation, l'âge, la société, le code postal, le numéro et le nom du département, le nom de la région, la catégorie de la difficulté (il y en a 6 explicitées), la date (qui est majoritairement vide), le libellé (donnant des informations sur la difficulté), l'avancement (identifié/en cours/partiellement résolu...), un commentaire (qui est du texte libre donnant des informations supplémentaires sur les difficultés et leurs résolutions), la résolution (si elle est en interne/externe ...) et la date de la difficulté.

Les commentaires libres fournissent un contexte essentiel pour comprendre la nature des problèmes. Les métadonnées telles que les dates de signalement et les catégories de difficultés sont également présentes pour permettre une analyse temporelle et thématique des données.
Cette table nous permet donc de déterminer les différents profils.

2.1.2 Feuille 2 : `action_export`

Cette feuille regroupe les actions entreprises pour chaque cas répertorié dans la première feuille. Les données sont également organisées par clé unique correspondant à chaque individu. La table est composée de 35863 lignes et de 18 colonnes.

Les différentes colonnes de cette table sont : la clé, la civilité, la situation familiale, l'âge, la société, le code postal, le numéro et le nom du département, le nom de la région, la catégorie de l'action, le type d'action, le

libellé du type d'action, l'objectif, le résultat, le commentaire, la date de début, la date de fin et la durée. Les colonnes *objectif*, *résultat* et *commentaires* représentent du texte libre détaillant l'action effectuée.

Ces différentes colonnes pourraient nous aider à contextualiser l'intervention. En effet, cette table nous permet de visualiser quelles sont les actions qui ont été mises en place pour accompagner chaque individu.
À noter que parmi les 35863 lignes, seulement 3995 clés sont uniques. Ce qui suggère que plusieurs actions ont pu être mises en place pour un seul individu.

2.1.3 Feuille 3 : `formation_export`

Dans cette table, sont explicitées les données se concentrant sur les parcours de formation des personnes accompagnées. Cette table est de 2044 lignes et de 11 colonnes, et parmi tous ces individus, 937 sont uniques.

Les informations couvrent la clé, le type d'action, le libellé de la formation, le but de la formation, la qualification (qui est un texte libre et caractérise les qualifications à l'issue de la formation), le commentaire concernant la formation (texte libre également, commentant éventuellement la formation), le résultat de la formation (qui semble totalement vide), le centre de la formation, la date de la formation, la durée de la formation et un commentaire sur le centre de formation (texte libre avec éventuellement l'adresse et/ou le contact du centre).
Cela fournit une base pour analyser l'adéquation entre les formations proposées et les besoins des participants, ainsi que le succès de ces initiatives.

2.1.4 Feuille 4 : `projet_export`

La quatrième feuille renseigne les projets des individus pour leur sortie. Elle prend en compte la clé, le type de projet (personnel/professionnel/formation), le nom du projet (mots-clés libres qualifiant le projet), un commentaire (texte libre pour décrire le projet) et la date du projet. Elle est composée de 7847 lignes et de 5 colonnes. Concernant l'unicité, nous avons, parmi ces 7847 lignes, 6919 individus différents dans cette table. Un individu peut donc avoir exprimé plusieurs projets.

Ces données permettent de suivre la progression des projets et d'évaluer leur impact sur les parcours des personnes accompagnées.

2.1.5 Feuille 5 : `sortie_export`

La dernière feuille trace les sorties par individu, avec une clé unique par personne toujours. Cette table est composée de 4988 lignes. Les 6 colonnes de cette table décrivent la clé, le type de sortie (professionnel/personnel/formation), le nom du métier, le statut de la sortie (positive, à retirer/autre/durable/transition), la date de la sortie et le motif de sortie.
Cette table est intéressante pour mesurer les résultats finaux du programme et pour comprendre les motifs de sortie, qu'ils soient positifs (emploi, formation) ou négatifs (sans nouvelle, difficultés non résolues).

2.2 Nettoyage Global

Pour assurer l'intégrité et la fiabilité de notre analyse, un processus de nettoyage des données a été mené. Cela a impliqué, entre autres, la recherche de doublons et la suppression des enregistrements incomplets qui pourraient fausser nos conclusions. Par exemple, dans la table `sortie_export`, nous avons éliminé les lignes où le `statut_sortie` n'était pas renseigné. Cette suppression est importante selon nous, car le statut de sortie est un indicateur déterminant de l'issue du parcours d'accompagnement des individus. En retirant ces lignes, nous nous assurons que chaque donnée prise en compte dans notre étude est à la fois pertinente et complète, ce qui renforce la précision de notre analyse et la pertinence de nos recommandations basées sur ces données nettoyées.

Afin d'effectuer ensuite une analyse descriptive de ces tables, nous avons ensuite cherché à homogénéiser notre mappage afin de pouvoir analyser les différentes tables, mais avec les mêmes regroupements de données. En effet, comme nous le verrons ensuite dans les différentes représentations, nous avons rassemblé les données de cette manière :

- Nous avons rassembler les âges par tranches : < 18, $18 - 25$, $25 - 35$, $35 - 45$, > 45
- Les civilités (dont nous avons uniformisé l'appellation qui n'était pas toujours la même) comme étant M pour monsieur et Mme pour madame
- Le statut Familial : célibataire (regroupant célibataire, séparé, divorcé, veuf), marié (regroupant marié, pacsé et vie maritale) et non précisé (ne souhaite pas répondre, vide et non géré)

2.3 Étude statistique des différentes tables

Afin d'étudier de manière plus approfondie chaque table, nous avons effectué des statistiques descriptives pour visualiser un peu mieux ce que représentent ces données. Ceci dans l'optique de voir si certaines données nécessitaient une autre transformation, mais aussi d'avoir une première idée de l'ensemble des profils qu'il est possible de considérer. Nous nous sommes dit que cela pourrait constituer un point de départ intéressant pour notre étude et la création d'un modèle répondant à cette problématique.

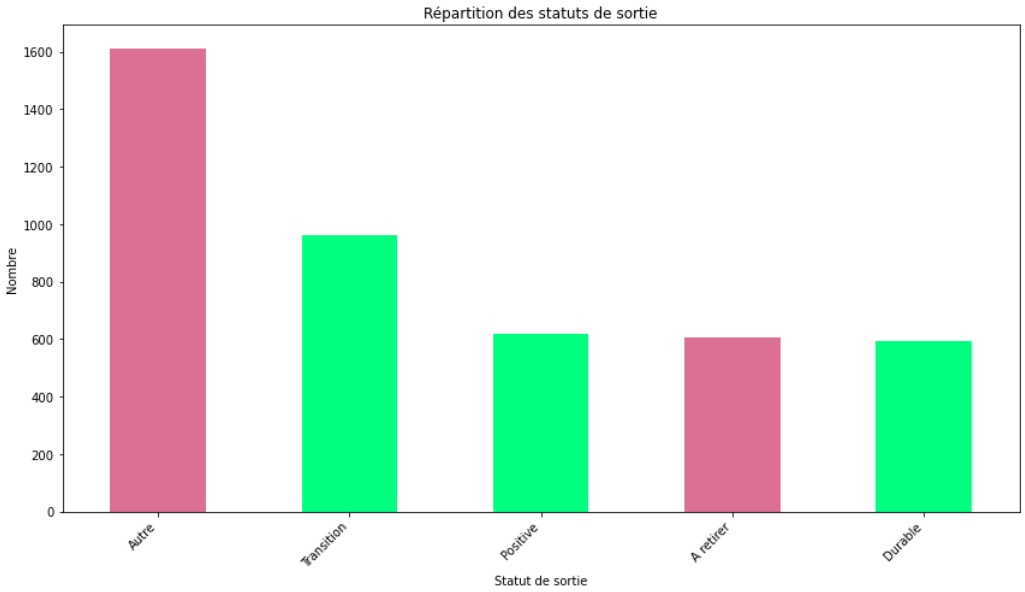
2.3.1 `sortie_export`

Nous avons commencé par la table concernant les sorties puisque nous nous en sommes ensuite servis comme point de départ pour effectuer les statistiques descriptives sur les autres tables. L'objectif étant de comprendre dans un premier temps comment VitamineT caractérise ses sorties, afin de déterminer comment distinguer une sortie réussie d'une sortie non réussie. Nous rappelons que cette table possède 4988 lignes pour 4312 individus différents.

Concernant les caractéristiques à disposition, nous allons regarder les statuts de sorties, les types de sortie, mais aussi les métiers. Nous mettons de côté la date de sortie, car nous ne considérons pas la dimension temporelle, ainsi que le motif de sortie trop diversifié et complexe.

En moyenne, nous observons entre 1 et 6 sorties renseignées par individu, ce qui nous donne une moyenne individuelle à 1.18 (1.21 en cas de sortie négative et 1.15 en cas de sortie positive). Cette première information nous permet de voir que la majorité des individus semblent n'avoir qu'une seule sortie indiquée en général, c'est à dire un seul accomplissement notifié qui permet d'acter la fin de l'accompagnement. Cela nous permet, aussi, de constater que ce nombre d'accomplissement ne semble pas être discriminant vis à vis des sorties négatives et positives.

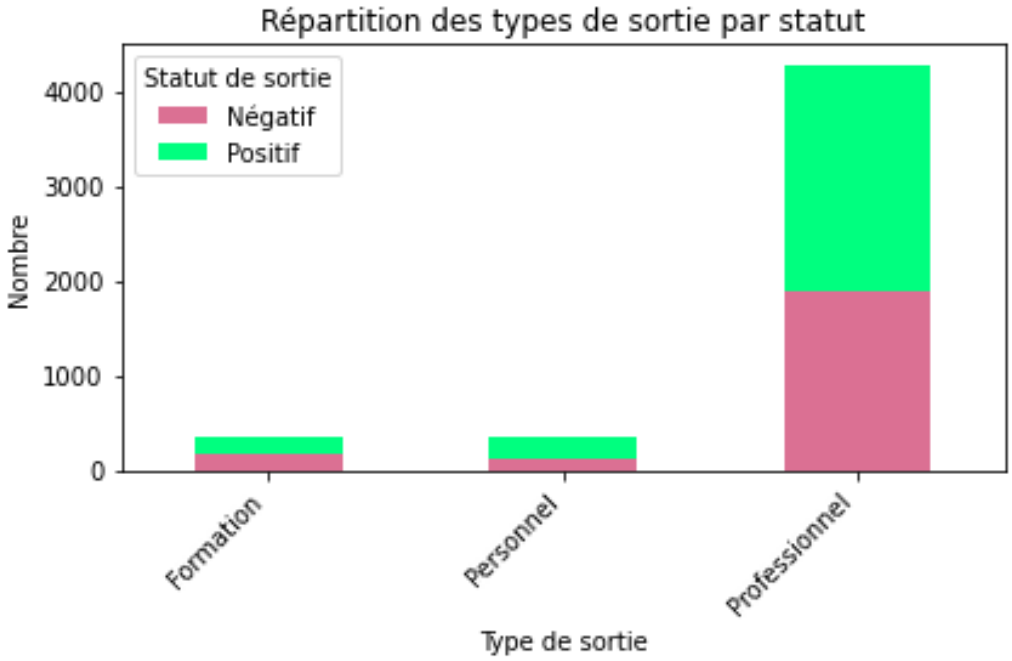
Nous avons commencé par étudier ce que nous allons considérer comme étant une sortie positive ou négative. En effet, en se basant sur les données, nous aurions les différentes sorties suivantes.



En ce qui concerne le statut de la sortie, nous obtenons un pourcentage de 36.7% de 'Autre' représentant les individus dont le motif de sortie est 'sans nouvelle' ou encore 'au chômage'. Cette catégorie semble symboliser quelque chose de plutôt négatif. Les 13.5% d'individus dont le statut des 'durable' représente les individus dont le motif de sortie est un emploi d'une durée de plus de six mois. De la même manière que les 21.9% d'individus considérés en 'transition' (dont les individus ont obtenu un emploi d'une durée de moins de six mois) et les 14.1% d'individus ayant une sortie positive, ces catégories peuvent donc être considérées comme positives.
Ainsi, nous obtenons globalement un taux de positif de 56% et en ne tenant pas compte de la catégorie 'Autre', nous retrouvons le pourcentage de 66% (plus précisément 65.5%) précisé par l'entreprise.

De plus, nous avons envisagé la possibilité qu'une personne ayant plusieurs sorties enregistrées, de types différents, puisse obtenir des résultats variés. Cela nécessiterait alors de réfléchir à la manière de qualifier plus finement la sortie "globale" du programme d'accompagnement de l'individu. Cependant, nous avons rapidement

constaté, à travers des filtrages et des comptages rapides, que dans le cas de sorties multiples, toutes les sorties présentaient le même statut.

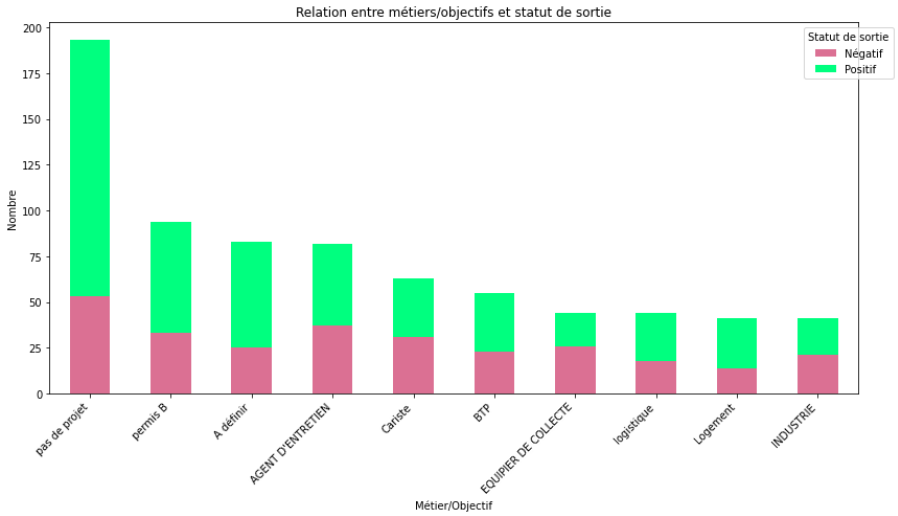


Nous observons que le type de sortie (qu'elle soit positive ou négative) correspond principalement à une sortie professionnelle (plus de 4000 individus concernés, soit 83.9%) alors que les taux de sortie personnelle et de formation sont, eux, plutôt similaires, d'environ 500 (soit d'environ 8%).
Concernant le pourcentage de sorties positives observées, nous obtenons le tableau suivant :

Type de sortie	**Pourcentage de positifs**
Formation	53.74%
Personnel	63.79%
Professionnel	56.04%

TABLE 2.1 – Pourcentage de sorties positives par type de sortie

Ainsi, en général, nous constatons que pour tout type de sortie, nous obtenons une sortie principalement positive.



Finalement, lorsque nous regardons ce dernier graphique obtenu à partir des données plutôt structurées des métiers mentionnées dans la feuille de sortie, nous voyons apparaître les différents métiers/objectifs des individus renseignés dans la sortie de chaque individu. Autrement dit, ce sont les objectifs de projets réalisés par les individus suivis. Nous pouvons y remarquer que le taux de personnes n'ayant pas de projet (que nous avons ajusté en assimilant avec la catégorie 'NON' de ce même tableau) est alors très élevé. C'est ensuite le 'permis B' qui figure parmi les objectifs les plus cités.
Quand nous observons ce graphique regardant les statuts de sortie, nous pouvons déduire que les personnes n'ayant pas de projet sont majoritairement des personnes qui obtiennent un résultat négatif ('chômage', 'sans nouvelle'). Pour les autres barres de ce graphique, nous obtenons des proportions plutôt similaires reflétant celles des métiers.

Cette table nous fournit donc les éléments clé pour décrire la sortie des individus accompagnés par VitamineT. À ce stade, nous retenons que la multitude de sorties pour certains individus ne pose pas de problème, mais aussi qu'une simplification de la caractérisation des résultats est envisageable en regroupant les statuts de sortie 'Positif', 'Transition' et 'Durable' sous l'appellation 'Positif', tandis que 'A retirer' serait considéré comme 'Négatif'.
Ainsi, nous nous baserons sur cette approche pour la suite de notre étude, avec, pour objectif de déterminer les caractéristiques plus à même d'obtenir un résultat positif.
Par ailleurs, nous constatons déjà ici que la formulation d'un projet semble fortement corrélée à une sortie positive. Nous allons donc voir dans les autres tables quels autres champs méritent de retenir notre intérêt, mais aussi comment intégrer cette formulation de projet à notre modélisation.

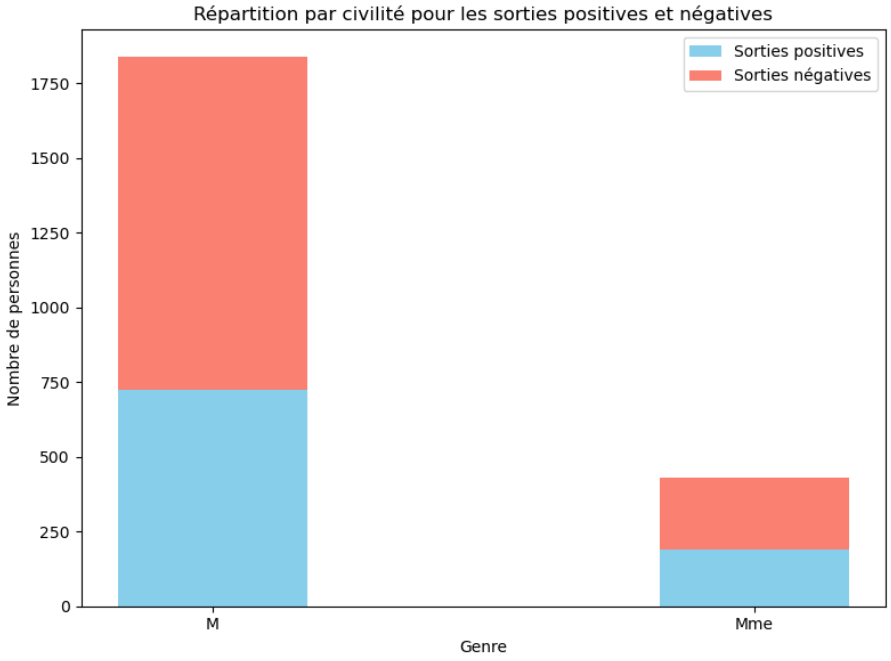
2.3.2 `difficultes_export`

Nous allons maintenant traiter de la table `difficultes_export`. Cette table est particulièrement importante, car, comme dit plus haut, elle contient les informations relatives au diagnostic d'entrée des individus et donc ce qui va caractériser leur profil en tant que personne à accompagner. Notre objectif est de comprendre quels profils sont pris en charge par VitamineT, mais aussi de déterminer si certaines caractéristiques influent sur la sortie de l'individu et, implicitement, sur son accompagnement.

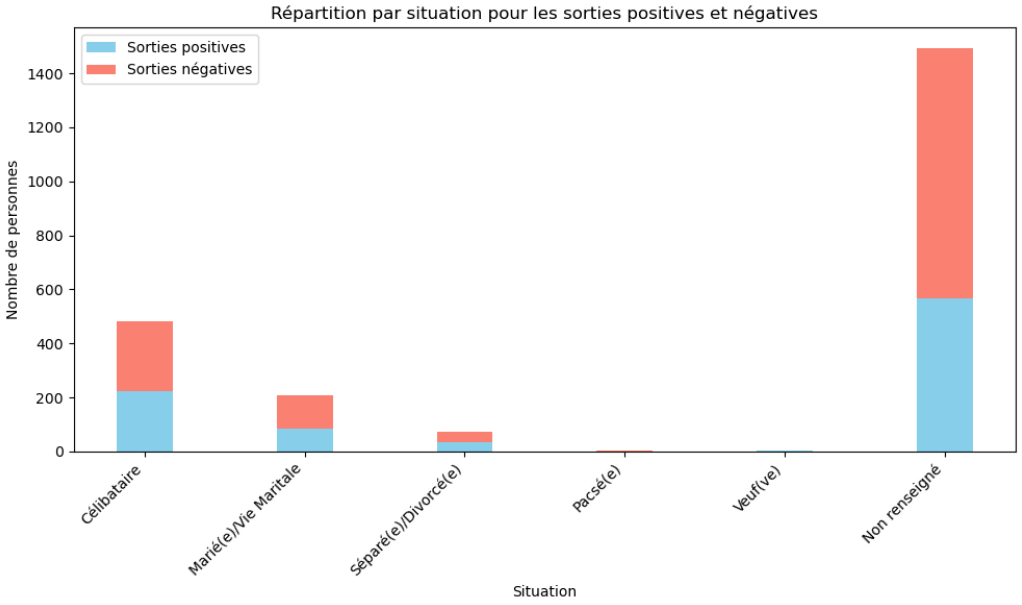
Pour simplifier notre analyse, nous regrouperons les sorties 'Durable', 'Transition' et 'Positive' sous le terme de sorties positives, et 'A retirer' ainsi que 'Autres' sous celui de sorties négatives, comme évoqué précédemment.

Nous rappelons que cette table contient 14852 lignes pour 6489 individus différents. Parmi ces individus, 911 possèdent une sortie dite positive, 1355 une sortie négative et 4221 n'ont pas de sortie renseignée. Nous choisissons donc d'étudier plus en détails le profil des 2266 personnes pour lesquelles la sortie est renseignée, les autres ne contribuant pas à notre analyse.

En ce qui concerne les colonnes et donc les caractéristiques, nous faisons le choix de nous concentrer sur la civilité, la situation, l'âge, le département, la région, les catégories de difficultés et les libellés des catégories. Nous écartons l'avancement du diagnostic, les commentaires textuels, le type de résolution et la date du diagnostic qui sont plus complexes à étudier et qui ne nous semblent pas pertinents pour définir un profil.

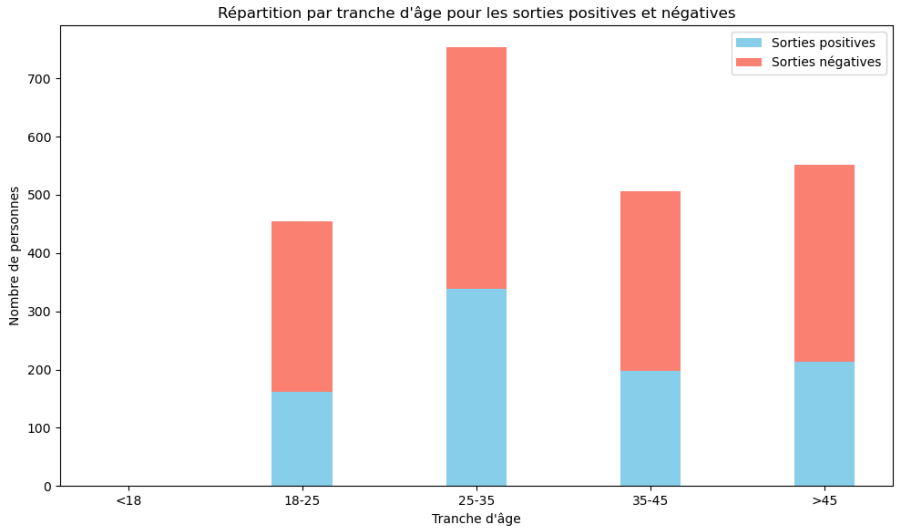


Globalement, nous observons une majorité d'hommes à accompagner par rapport aux femmes. La répartition est d'environ 81% d'hommes et de 19% de femmes. Cette première caractéristique naturelle pourrait être un point pour définir nos profils type, toutefois, il est important de souligner que la différence est légère : 44% des femmes ont une sortie positive et 40% des hommes également. Ainsi, la civilité ne semble pas pertinente pour différencier une sortie réussie d'une sortie qui ne l'est pas. D'un autre côté, nous pouvons penser qu'une telle variable pourrait peut-être avoir un impact sur le type d'actions à mettre en place. Nous reviendrons ensuite dessus en étudiant la table correspondante.



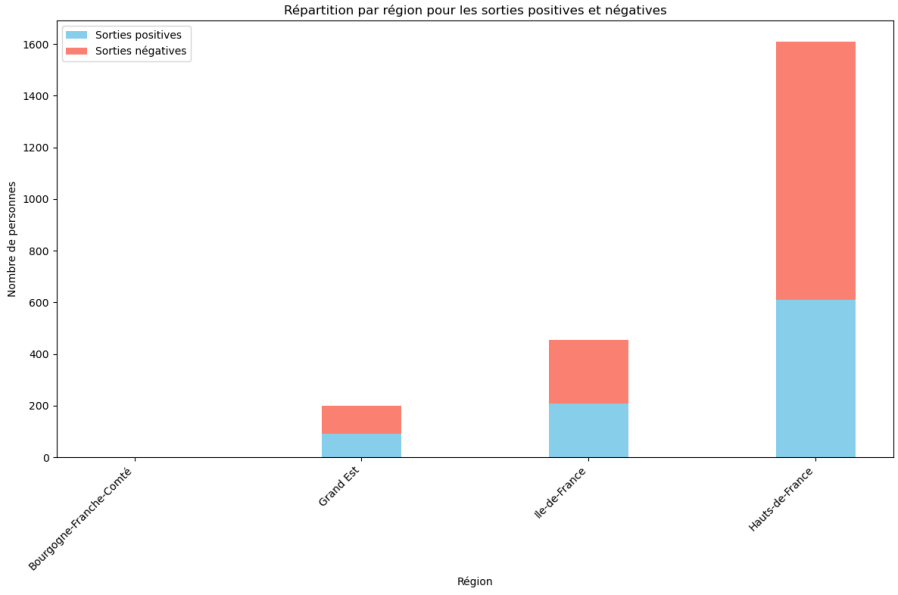
En ce qui concerne la situation familiale de ces individus, il apparaît que la plupart des individus n'ont pas renseigné leur statut. Ceci ne semble donc pas être une variable pertinente pour notre étude. Nous pouvons l'écarter.

Concernant l'âge, les données recueillies varient entre 18 et 70 ans. La moyenne tourne autour de $35,67$ ans avec un écart-type d'environ 11 ans. La classe majoritaire est la tranche $25 - 35$.

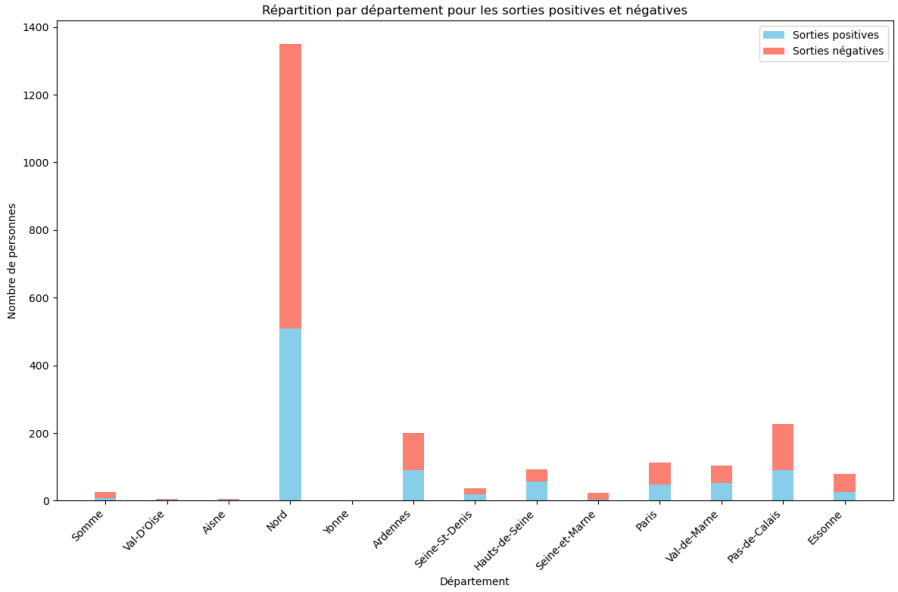


Néanmoins, il ne semble pas avoir de différences notables entre les tranches d'âge en ce qui concerne la réussite de la réinsertion. Il nous semble tout de même important de conserver cette variable, car nous devons garder en tête que nous allons établir des profils dans le but de déterminer en premier lieu des actions à mettre en place : nous pouvons imaginer que l'accompagnement ne sera pas le même pour un jeune de 18 ans que pour une personne plus âgée de 60 ans.

Un aspect supplémentaire à prendre en compte est la localisation géographique des individus, car il est raisonnable de supposer que leur lieu de résidence peut influencer à la fois leur réinsertion professionnelle et leur accompagnement.

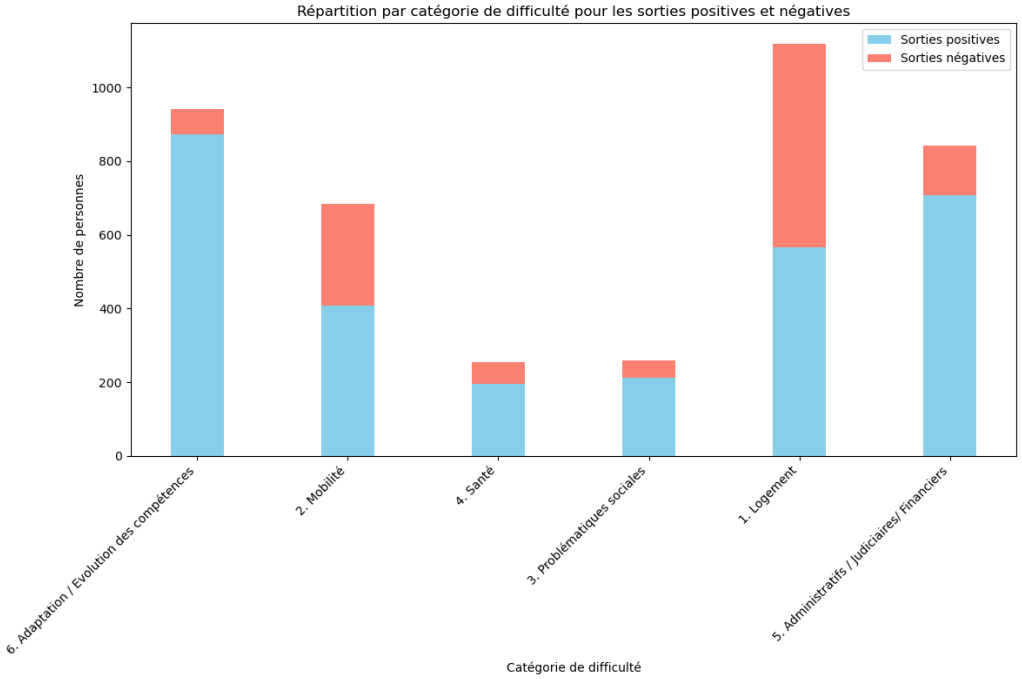


En examinant les régions d'origine des individus, nous constatons tout d'abord que les données sur les personnes originaires de Bourgogne-Franche-Comté sont très limitées (seulement 2 individus), tandis que nous disposons de données abondantes pour les Hauts-de-France. En ce qui concerne le taux de réussite de la réinsertion par région, il semble équilibré pour le Grand Est et l'Île-de-France, mais les réinsertions non abouties sont plus fréquentes dans les Hauts-de-France. Ces observations renforcent ainsi l'importance de conserver cette variable dans notre analyse.
Pour aller plus loin, nous pouvons plus précisément regarder le département de provenance :



Nous constatons, cependant, une répartition très inégale entre les différents départements, avec une grande majorité d'individus provenant du département du Nord (plus de 1200), tandis que les autres départements comptent environ 200 individus chacun. Cette disparité suggère qu'il n'est pas pertinent de considérer les départements pour établir nos profils, car cela ajouterait de la complexité sans fournir nécessairement d'informations supplémentaires utiles, surtout compte tenu de l'information déjà disponible sur la région, et cela risquerait d'être influencé par la classe très majoritaire du département du Nord (berceau de l'entreprise). Par conséquent, nous n'envisagerons pas d'examiner les codes postaux dans notre analyse.

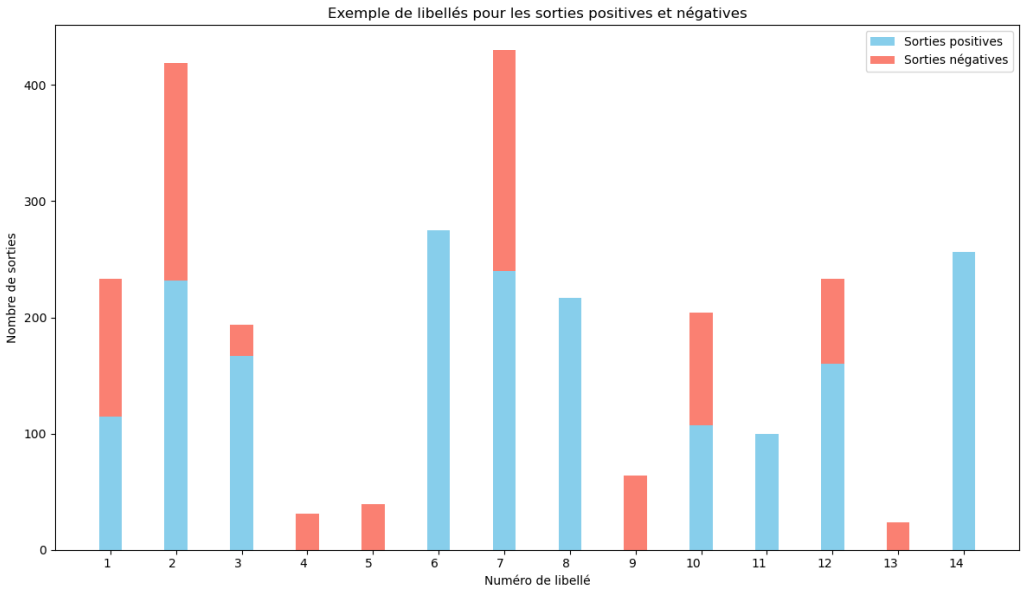
Nous allons à présent nous pencher sur l'élément central du diagnostic : les difficultés identifiées. Nous nous doutons que ça soit un élément clé pour la réussite et, avant cela, pour les actions à mettre en place.
Il convient de rappeler qu'un individu peut présenter plusieurs difficultés identifiées. En moyenne, ces difficultés se chiffrent à $2,61$ par individu ($2,59$ pour les sorties positives et $2,63$ pour les sorties négatives), avec une fourchette variant de 1 à 6 difficultés par personne.



Nous pouvons constater que certaines difficultés sont plus récurrentes que d'autres, notamment la numéro 1, à savoir 'Logement'. Quoi qu'il en soit, nous pouvons également voir qu'il peut y avoir de légères disparités en ce qui concerne les taux de réussite de l'accompagnement. Cela ne fait que renforcer ce dont nous nous doutions. Remarquons que les classes positives dominent dans ce graphique. Ceci peut paraître surprenant au premier

abord, néanmoins, les graphiques précédents incorporaient les individus pour lesquels aucune difficulté n'a été diagnostiquée et de plus, généralement, les sorties correspondent à plusieurs difficultés simultanément. Nous pouvons donc imaginer que le fait d'identifier précisément les diverses catégories de difficultés rencontrées aide à mettre en place un meilleur accompagnement et donc une sortie positive. Nous n'avons pas tenu compte de cela ici, car l'idée était juste d'avoir une idée de la part représentée par chaque difficulté ainsi que de la tendance globale des sorties associées.

Ces difficultés, généralement définies selon les critères de l'état, sont identifiées par le biais de libellés structurés. Au total, nous en répertorions 67 différents.
En moyenne, chaque individu présente $1,80$ difficultés précises identifiées ($1,71$ pour les sorties positives et $1,87$), avec une fourchette variant entre 0 et 11. Voici les 14 premières :



Numéro	Libellé de difficultés	pourcentage de positifs (%)
1	Problème d'accès au logement (dont 1er logement)	49.35
2	Absence de permis (non obtenu / suspendu / perdu)	55.36
3	Autre (3)	86.08
4	Difficultés de maintien dans le logement	0
5	S Mobilité : absence de permis (non obtenu, suspendu ou perD	0
6	Autre (6)	100
7	Autre (1)	55.81
8	Absence de projet professionnel réalisable	100
9	Logement inadapté/vétuste/insalubre	0
10	S Logement : Problème d'accès au logement	52.45
11	Autre (4)	100
12	S Administratif/judiciaire/financier : Problèmes financiers	68.66
13	Autre (2)	0
14	Autre (5)	100

TABLE 2.2 – Correspondance des numéros aux libellés de difficultés

Bien que certains libellés soient bien plus fréquents que d'autres, nous estimons qu'il est très pertinent de prendre en considération toutes ces variables, car elles permettent de préciser les problèmes diagnostiqués. En effet, nous pouvons envisager qu'un problème administratif, par exemple, nécessite un type d'accompagnement

différent en fonction de sa nature. De plus, les proportions des résultats de sorties semblent très différentes selon ces difficultés. Certaines ne sont connotées qu'à des sorties négatives et d'autres qu'à des sorties positives.

Il convient de noter que certains libellés sont différents alors qu'ils expriment une même difficulté, un aspect qui sera discuté ultérieurement dans ce rapport.

En tenant compte de ces différentes caractéristiques (civilité, âge, région, catégorie de difficulté et libellé), cela nous offre près de 16080 profils différents possibles. Cela nous semble être une base assez solide pour établir les profils type souhaités.

2.3.3 `action_export`

À présent, il s'agit de traiter de la feuille concernant les actions. Nous rappelons que cette table correspond aux 35861 actions mises en place pour 3995 personnes accompagnées.

Le nombre moyen d'actions par individu (toujours par rapport aux clés correspondant à la table `sortie_export`) est de 20.31. Ainsi, en moyenne 20 actions sont nécessaires pour se retrouver dans la table action et dans la table sortie. Cette moyenne se situe dans une fourchette allant de 1 à 106 actions par personne. Cependant, ceci ne nous donne pas d'information quant à l'importance du nombre d'action pour avoir une sortie positive. C'est pourquoi nous pouvons également mettre ce taux en lien avec le type de sortie. Nous obtenons que, pour obtenir une sortie positive, l'individu doit d'effectuer en moyenne 18.69 actions, contre 22.27 pour les individus ayant une sortie négative.
Nous pouvons déjà interpréter cela par le fait que, plus un individu a d'actions, moins il a de chance d'obtenir une sortie positive, même si cela reste très léger puisque l'écart est assez minime.
Ceci nous indique que le nombre d'actions mises en place pourrait être une variable discriminante qui mérite notre intérêt.

Certaines informations contenues dans cette table `action` sont redondantes par rapport aux autres tables, comme la civilité, l'âge, la situation familiale... Nous supposons que les personnes pour qui on met en place des actions ont déjà été évaluées au début. Donc, nous pensons que leurs résultats seront similaires à ceux observés dans l'étude sur les difficultés. Nous tenterons cependant de visualiser ces données en fonction des actions afin de pouvoir juger de leur pertinence.

Nous y trouvons également de nouvelles informations telles que *cat_action*, *type_action*, *lib_type_action*, la *duree*... Ces dernières pourront alors nous aider à déduire de nouvelles interprétations utiles concernant le statut de sortie basées sur les actions mises en place lors de l'accompagnement des personnes en difficultés.

Nous écarterons comme précédemment les données temporelles à savoir *date_debut* et $date\_fin$, mais aussi l'analyse des colonnes $objectif$, *commentaire* et *resultat* car il s'agit de texte libre. Ne pouvant pas être traitées aussi facilement, nous les traiterons dans le *chapitre* 3.
Subséquemment, nous ne prendrons pas en compte les champs *cat_action* et *type_action*, qui sont trop généraux pour identifier des actions spécifiques contrairement à leur libellé (*lib_type_action*).

À présent, nous pouvons alors nous concentrer sur l'analyse descriptive concernant *lib_type_action*, ceci nous permettra certainement d'avoir un aperçu de l'impact du choix des actions sur la sortie d'un individu.

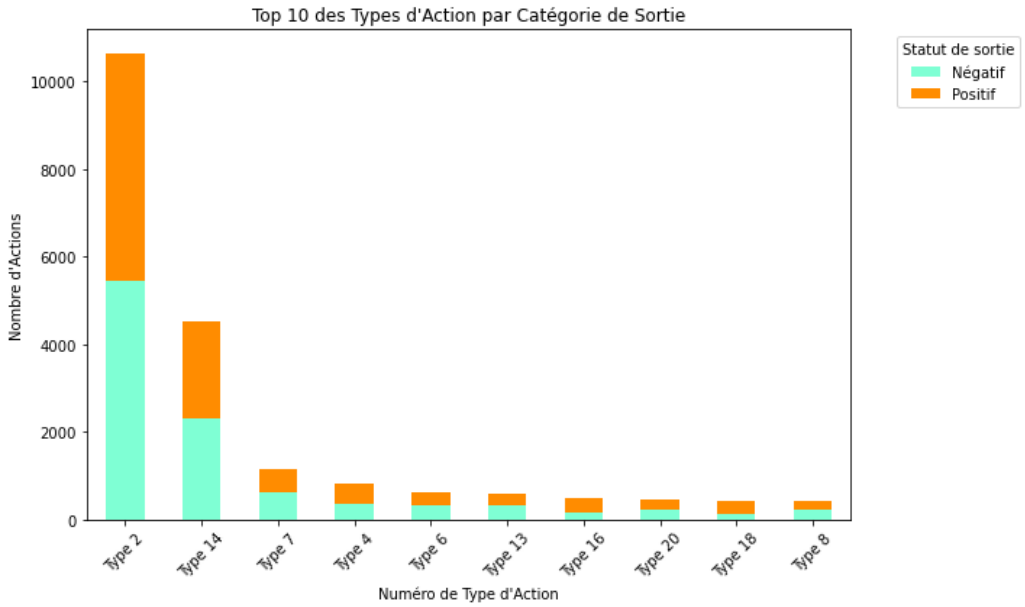


FIGURE 2.1 – Répartition par civilité et statut de sortie

Ce que nous pouvons observer ici est que le taux de sorties positives varie selon le libellé du type de l'action. Bien que ce taux de positif reste plutôt autour des 50% pour la plupart, nous pouvons remarquer que le suivi social obtient un taux de sortie positif bien supérieur à cette moyenne. De même, les personnes accompagnées dont l'entreprise est parvenue à effectuer un suivi de parcours obtiennent des résultats plutôt positifs, et ceci est logique : les personnes dont il n'y a pas de suivi ont plus de chance de ne pas réussir à résoudre leurs problématiques. Au contraire, nous avons que la définition d'un projet professionnel n'est pas propice à une sortie positive. Nous pouvons imaginer que cette action est envisagée lorsque la personne suivie n'arrive pas à définir un projet professionnel d'elle-même. Cela montre donc bien que le choix des actions à mettre en place joue un rôle fondamental dans le résultat de sortie.

Continuons donc avec les informations jugées pertinentes précédemment dans l'idée qu'elles pourraient influencer le suivi. Nous allons donc visualiser ces différentes données, mais cette fois-ci en jugeant de leur influence sur les actions mises en place.
Par souci d'interprétation, nous n'avons pas jugé utile de visualiser les graphiques en lien avec les régions et les départements, étant donné qu'après quelques calculs, nous obtenons que plus de 80% des informations de cette table proviennent d'individus du Nord.
Commençons par regarder le nombre d'action en fonction de la civilité.

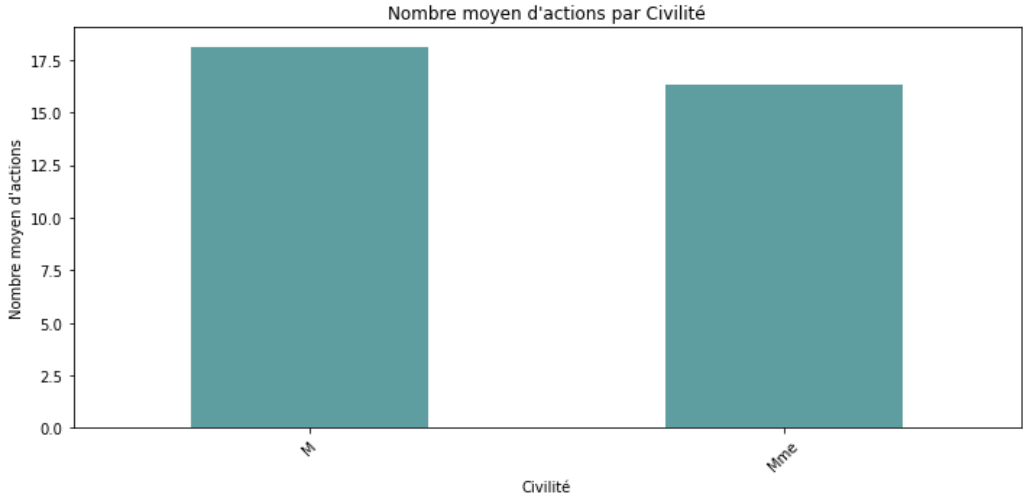


FIGURE 2.2 – Nombre moyen d'action en fonction de la civilité

En y regardant de plus près, nous voyons, pour commencer, que le nombre d'actions moyen entre les hommes

et les femmes n'a pas de différence significative.

Or, nous pouvons tout de même tenter de ressortir les actions qui reviennent le plus, pour les hommes et pour les femmes, pour vérifier si la civilité a un impact sur l'accompagnement mis en place. C'est ce que nous pouvons observer au niveau des deux dernières colonnes du tableau suivant :

Type de l'action	Libellé du type de l'action	pourcentage de positifs (%)	femmes	hommes
Type 2	Entretien et bilan régulier	49	47.5	49.5
Type 4	Diagnostic socio-professionnel	49.1	50	57.4
Type 6	Passerelle vers les entreprises classique	45.7	58.1	41.6
Type 7	Solution aux problématiques sociales	55.7	41.5	52.9
Type 8	Bilan de fin de parcours	46.1	41.5	52.9
Type 13	Définition d'un projet professionnel	20.8	51.3	37.3
Type 14	Suivi social	75.7	49.7	48.9
Type 16	Affectation référent social	48.4	68.1	63.9
Type 18	Suivi parcours : Entretien de renouvellement	68.7		68.9
Type 20	Technique de recherche d'emploi	49.5	45.7	49.8

TABLE 2.3 – Répartition du taux de positifs selon le libellé de l'action

Nous constatons alors une différence au niveau de l'importance des actions effectuées. Effectivement, les pourcentages de sorties positives pour les hommes et les femmes selon les libellés d'actions semblent se différencier, notamment en ce qui concerne la définition d'un projet professionnel qui semble être une action mieux réussie pour les femmes (51.3%) que pour les hommes (37.3%).

La valeur manquante exprime le fait que l'action 'suivi' 'parcours : Entretien de renouvellement' n'est pas présente dans les 10 principales actions effectuées. C'est l'action 'appel telephonique' qui remplace l'action manquante dans de top 10, avec près de 60% de réussite, ce qui est plutôt satisfaisant.

Ainsi, comme supposé précédemment, nous avons bien une différence de suivi selon la civilité de la personne se présentant à l'entreprise.

Nous allons donc étudier, de la même façon, l'impact de l'âge sur les actions.

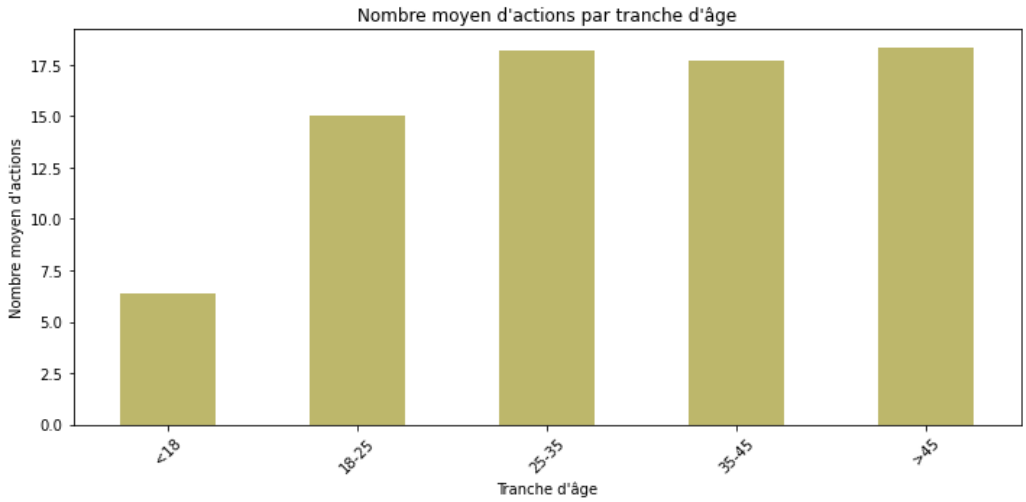


FIGURE 2.3 – Nombre moyen d'action en fonction de la tranche d'âge

Nous observons une nette différence du nombre moyen d'action en fonction de la tranche d'âge, ce qui nous fait penser que l'âge joue également un rôle important pour le suivi et renforce le fait que cette variable nous sera utile dans la modélisation. Pour conforter cette idée, nous allons, comme pour les civilités, expliciter ces différentes tranches d'âge en fonction des libellés explicités au début de cette sous-partie :

À noter que les individus dont l'âge ne dépasse pas 18 ans dans la table `action` ne sont pas présents dans la tale sortie et ne permettent donc pas d'en ressortir des pourcentages.

Type de l'action	Libellé du type de l'action	positifs (%) $18-25$	$25-35$	$35-45$	>45
Type 2	Entretien et bilan régulier	43.3	39.5	56.1	56
Type 4	Diagnostic socio-professionnel	52.9	48.5	63.1	57.2
Type 6	Passerelle vers les entreprises classique	41.4	48.6	65.6	37
Type 7	Solution aux problématiques sociales	57.9	40.1	47.6	60.2
Type 8	Bilan de fin de parcours	42.6	48.4	52.2	
Type 13	Définition d'un projet professionnel	59.1	42.5	49.6	26.9
Type 14	Suivi social	78.7	40.4	52.3	53.9
Type 16	Affectation référent social	31.7	67.7	69.9	78.1
Type 18	Suivi parcours : Entretien de renouvellement	51.5	64.9	78.9	69
Type 20	Technique de recherche d'emploi	62.3	39.2	62.5	42

TABLE 2.4 – Répartition du taux de positifs selon le libellé de l'action et l'âge

Cependant, nous remarquons une nette différence à propos des taux de sorties positive en fonction de l'âge pour quelques types d'action. Dans l'ensemble, le suivi de parcours semble donner un bien meilleur taux de sorties positives pour les $35-45$ ans.

Le 'suivi social' lui, semble avoir un impact positif sur la sortie surtout pour les $18-25$ ans mais un peu moins pour les $25-35$ ans.
Nous voyons également que l'action 'Affectation référent social' témoigne d'une réussite croissante en fonction de l'âge de la personne : plus la personne encadrée est âgée, plus elle a tendance à obtenir une sortie positive pour cette action.

En ce qui concerne les plus de 45 ans, nous avons également dans les 10 principales actions notées l'action 'appel telephonique' et qui présente un taux de sortie positive de seulement 28.4%, qui n'a donc pas fonctionné.

Enfin, nous voyons par exemple que l'action relative à la définition d'un projet professionnel ne semble pas faire aboutir les personnes de plus de 45 ans à une sortie fructueuse puisque le taux de réussite pour ces personnes se trouve de n'être seulement de 26.9%.
Ainsi, comme nous le pressentions, cette analyse témoigne de la pertinence de la variable âge pour nos modélisations.

L'examen de cette table nous a permis de confirmer un aspect essentiel sur lequel repose notre problématique : le choix des actions, ainsi que leur quantité, ont une influence significative sur la réinsertion des individus accompagnés. De plus, cette analyse nous a permis de ne constater que des caractéristiques propres aux individus telles que la civilité ou l'âge ont bien leur importance dans nos modélisations.

2.3.4 `formation_export`

Ensuite, nous allons donc traiter de la table concernant les formations.
À nouveau, le texte libre ne pourra pas être traité dans ces statistiques descriptives, que des informations relatives à la $qualification$, $commentaire\_formation$ et $resultat\_formation$ autour de la formation ou de $commentaire\_centre\_de\_formation$ qui rassemble principalement des adresses ce qui ne semble pas non plus pertinent dans notre étude. Nous ne traiterons pas non plus de la *date*, car nous partons toujours du principe que l'on souhaite trouver un résultat valable dans le temps. De même, nous écarterons le type d'action : il s'agit toujours d'une formation en interne donc il semble inutile de la comparer avec la sortie.
Cependant, nous pouvons souligner le fait que cette table est moins "grande" que les autres, puisque nous le rappelons, seulement 937 individus uniques sont dans cette table (qui est par exemple bien moins élevé des près de 4000 individus unique de la table action traitée juste au dessus).
Parmi ces 937 individus (2044 lignes au total), seulement 338 sont en commun avec la table sortie (1294 formations au total corroborent avec les lignes de la sortie, toujours en termes de clé). Cela peut tout de même nous permettre de regarder quelques statistiques en rapport avec la table de sortie.

Tout d'abord, observons les informations traitant de la durée de la formation. Pour cela, nous allons à nouveau ne conserver que les formation dont les clés sont présentes dans la table `sortie_export`. Cependant, nous ne conserverons pas une formation par clé, car un individu pourrait effectuer plusieurs formations. Ils

seront traités ici comme s'il s'agissait d'individus différents et nous n'étudierons qu'à la fin de cette partie le nombre de formations par individu.

Penchons-nous sur les statistiques descriptives de cette colonne : nous observons que la durée d'une formation peut varier de 0 (on suppose que cela pourrait être parce que l'individu était inscrit à une formation, mais ne s'y est pas présenté par exemple) à 1360 heures (l'unité n'est pas précisée, mais il s'agit de valeurs décimales, et il semble courant que le temps d'une formation soit comptabilisé en heure d'après ce que nous pourrions observer pour le même type de structure ; 1360 heures pour une journée de 8 heures nous donne 170 jours donc environ 8 mois et demi, ce qui semble être un temps raisonnable pour le maximum d'une formation).

Le nombre moyen d'heures de formation est de 28 heures, nous allons donc tenter d'étudier comment le nombre d'heures de formation influe sur les sorties positives ou négatives de la personne se présentant au sein de la structure.

Pour cela, nous allons découper en tranches la durée de formation, comme nous avons fait pour l'âge, pour pouvoir les observer dans un graphique. Au vu des statistiques descriptives (moyennes, quantiles, variance) nous allons découper cette durée en 5 tranches : $0-7$, $7-16$, $16-35$, $35-71$ et >71. Cela nous permet d'avoir une proportion équivalente pour chaque tranche. Nous obtenons alors le graphique suivant :

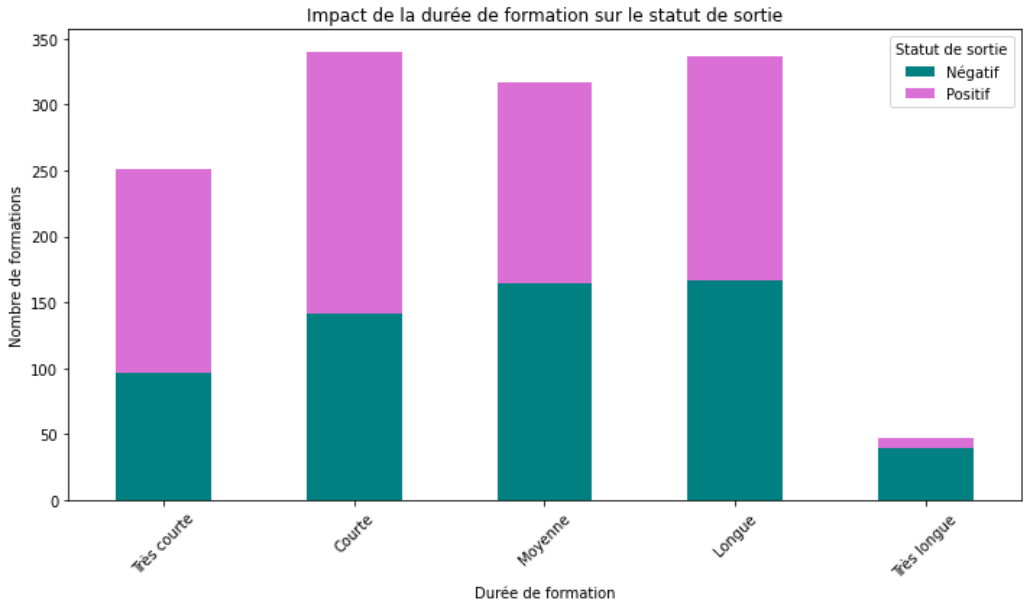


FIGURE 2.4 – Répartition par durée et statut de sortie

Ici, il nous semble intéressant de notifier que, plus on augmente le nombre d'heures de formation, moins la sortie semble positive. En effet, nous observons le pourcentage de positifs suivant :

Durée de formation	Très courte	Courte	Moyenne	Longue	Très longue
Taux de positif	61%	58%	47.9%	50.4%	14%

TABLE 2.5 – Répartition du taux de positif selon la durée de formation

Une chose flagrante et sur laquelle il semble pertinent d'insister est que, pour les formations 'très longues' (qui on le rappelle concernent les formations de plus de 71 heures), nous avons un taux très faible de sorties positives, soit 7 personnes sur les 47 ayant suivi une formation longue.
Cette variable a donc un impact significatif sur la sortie.

Étudions à présent la colonne traitant du libellé de la formation. Nous allons chercher à interpréter les mêmes résultats que pour la colonne concernant la durée. Ici encore, nous n'allons traiter que de la formation en elle-même (et non pas des formations pour un individu traité uniquement) pour les mêmes raisons que précédemment.

Nous obtenons le tableau suivant :

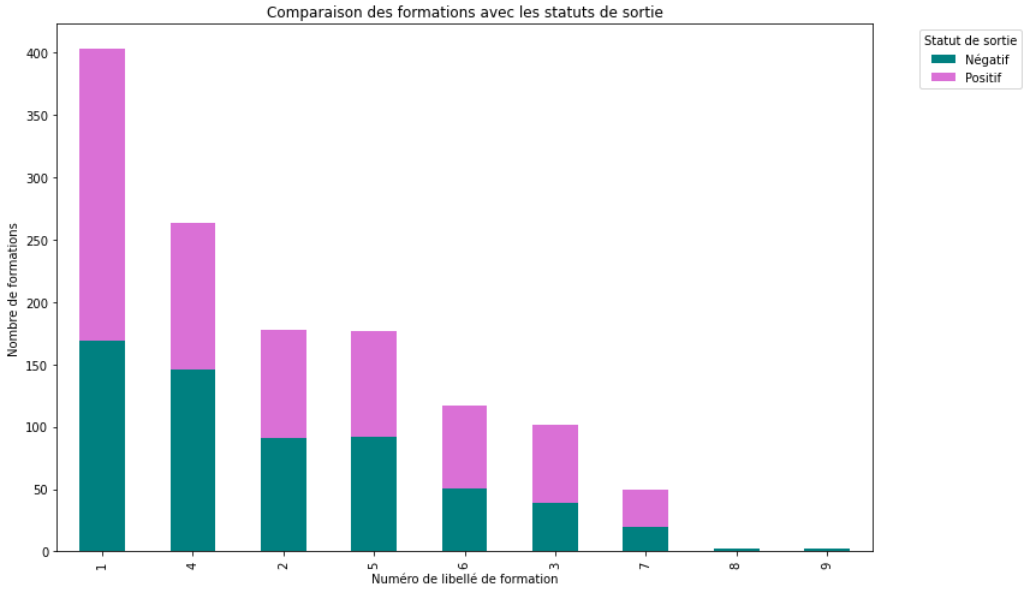


FIGURE 2.5 – Répartition par durée et statut de sortie

Les numéros de 1 à 9 correspondent aux libellés suivants :

Numéro	Libellé de formation	pourcentage de positifs (%)
1	Acquisition, entretien ou perfectionnement des connaissances	58.1
2	Adaptation et développement des compétences	48.9
3	Autres	61.8
4	Lutte contre l'illettrisme et apprentissage de la langue française	44.7
5	Préformation et préparation à la vie professionnelle	48
6	Aucune formation explicitée	56
7	Prévention	60
8	Conversion	0
9	Promotion professionnelle	0

TABLE 2.6 – Correspondance des numéros aux libellés de formation

Aucun résultat probant n'a été observé pour les 4 formations dont le libellé est 'Conversion' ou 'Promotion professionnelle'. Il y a cependant trop peu de formations de ce type pour affirmer qu'elles représentent un frein à l'obtention d'une sortie positive.

Les taux 'Acquisition, entretien ou perfectionnement des connaissances', 'Prévention' et 'Autre' sont des taux qui semblent être les libellés favorisant le plus le pourcentage de sorties positives. Dans l'ensemble, chacun des taux semble assez éloignés de 50% et reflètent l'importance d'inclure cette table dans notre modélisation.

Une chose surprenante à remarquer également est que le pourcentage pour ceux n'ayant pas précisé de libellé de formation est aussi élevé. Cela pourrait simplement s'expliquer par le fait que les formations ont pu être faites, mais pas précisées, et pas que la formation n'ait pas aboutie.

Ensuite, traitons du but de la formation. Cette étude peut nous aider à connaître l'importance de cette variable dans le sens où nous pourrions cibler quel(s) objectif(s) de formation mène(nt) à une sortie positive.

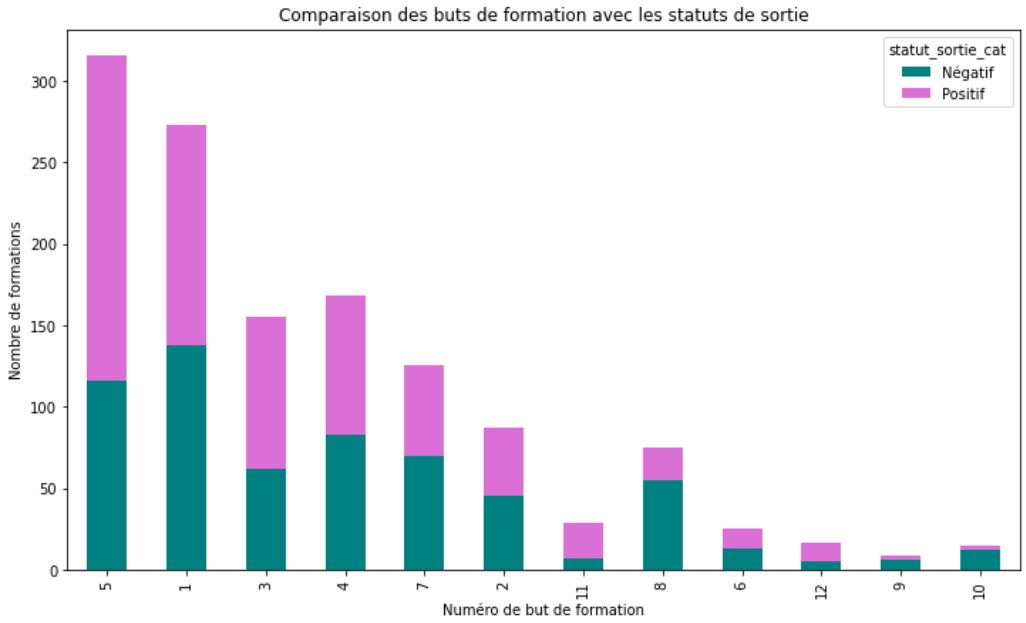


FIGURE 2.6 – Répartition par durée et statut de sortie

Numéro	But de formation	Pourcentage de positifs (%)
1	Acquisition de compétences transverses	49.5
2	Hygiène-sécurité	48.3
3	Sans but explicité	56.2
4	Maîtrise des Savoirs de base	50.6
5	Professionnalisation	63.3
6	Qualification / diplôme (CCP, CQP)	48
7	FLE	44.4
8	Préqualification (non diplômant, non qualifiant)	26.7
9	Remise à niveau / bureautique	75
10	Alphabétisation	20
11	Gestes et postures / PRAP	75
12	Lutte contre l'illettrisme	70.6

TABLE 2.7 – Correspondance des numéros aux buts de formation

Voici alors ce que nous pouvons déduire de ces résultats : nous remarquons une nette différence de résultats positifs selon le but de la formation. Alors que certaines formations amènent à des types de sorties plutôt défavorables pour la personne accompagnée (par exemple le but 'Alphabétisation' ou encore 'Préqualification' ont des taux vraiment faibles), d'autres favorisent vraiment les sorties positives (comme par exemple les 'Remise à niveau/bureautique', les 'Gestes et postures/PRAP' ou encore 'Lutte contre l'illettrisme'). Nous pourrions supposer que ces résultats sont dûs au fait que les buts qui n'ont pas de liens directs avec une formation (mais seulement en tant que 'pré-formation') ne permettent pas de résoudre le problème en lui-même, mais doivent être un chemin à passer pour tenter ensuite d'obtenir une sortie positive. Ainsi, le but de la formation semble influencer de manière significative sur le type de sortie que l'on obtient.

Concernant ces deux derniers résultats sur les libellés et les buts des formations explicités dans les données, maintenant que nous avons justifié leur importance, nous pourrons tenter de nous en servir par la suite dans notre modélisation. Cependant, nous avons éludé un aspect pendant cette analyse, parce que nous n'avons pas pris en compte tout ce qui est relatif au nombre de formations par individu. Nous allons donc le voir ci-après.

Nous avons observé une moyenne de 3.83 formations par individu. En d'autres termes, quand un individu est présent dans la table de sortie, il a donc eu besoin d'environ 4 formations pour obtenir une sortie (qu'elle soit positive ou négative). Cela pourrait alors s'expliquer par le fait qu'une personne qui suit une formation pourrait avoir besoin d'une autre pour compléter la demande entière pour un poste ou encore que l'individu s'est vu devoir tester plusieurs formations pour en trouver une qui lui convient.
Il nous reste alors à comprendre si ce nombre de formations ont une influence positive ou non sur la sortie. Tout d'abord, après calculs, les individus ayant une sortie positive contiennent un nombre moyen de 3.68 formations nécessaires alors que les personnes ayant une sortie négative ont eux suivi en moyenne 4 formations. Même si ces résultats restent très proches, ils pourraient s'interpréter comme le fait que, les personnes ayant eu une sortie négative, mais ayant tout de même suivi des formations ont dû tester plus de formations qui ne leurs convenaient pas et ne sont pas parvenus à trouver un emploi (ou autre selon la problématique, nous parlons ici de sortie positive) par la suite.

Nous pouvons alors étudier les proportions des formations en fonction des sorties. Nous trouvons que 52.74% des individus ayant suivi une ou plusieurs formations ont une sortie positive et 47.26% une sortie négative : nous pouvons donc conjecturer la pertinence de conserver ces différentes informations dans les modélisations suivantes, puisqu'elles semblent influer sur le type de sortie.

2.3.5 `projet_export`

Enfin, nous allons nous axer sur la table projet. Nous rappelons que cette table possède 7847 entrées pour 6919 individus uniques, et contient 5 variables.

Nous pouvons souligner que cette table possède moins de colonnes et qu'elles ne sont pas toutes exploitables dans cette sous-partie. En effet, les colonnes *nom_projet* et *commentaire* sont des colonnes avec du texte libre et ne sont donc pas traitables ici. Nous ne traiterons pas non plus de la colonne *date_projet* ne nous indiquant rien de pertinent pour la réalisation de notre analyse.

Avançons à présent dans l'analyse des statistiques concernant le type de projet.

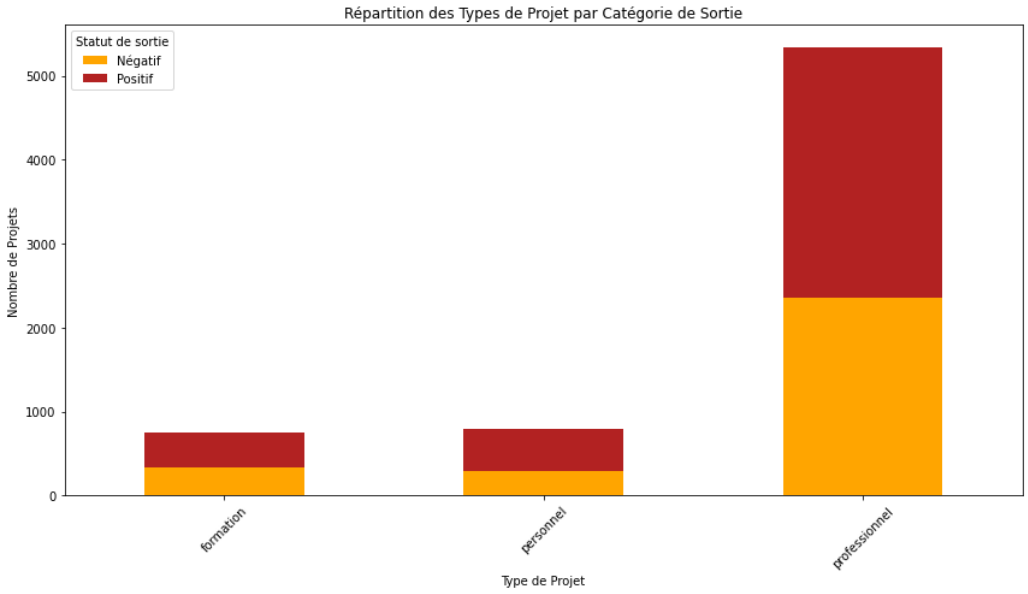


FIGURE 2.7 – Répartition par durée et statut de sortie

Ce que nous pouvons voir ici, c'est que la majorité des projets sont de types professionnels. En les comparant avec les réponses de sortie, nous pouvons constater que le pourcentage de personnes possédant une sortie positive ou négative est à peu près le même pour tous les types de projets (58.8% pour la formation, 63.8% pour les projets personnels et 55.8% pour les projets professionnels), cette variable ne semble donc pas influer sur le type de sortie.

Ensuite, nous allons étudier de manière plus approfondie le rapport entre les projets et les individus en explicitant le nombre de projets par individu, tout cela toujours en regardant les résultats par rapport aux individus qui se trouvent également dans la table `sortie_export`. Le nombre moyen de projet par individu est à peine plus grande que 1 (1.15 pour être exact), et cela ne semble pas influer sur la positivité du résultat de sortie puisque le nombre moyen est également d'environ 1 que l'on prenne les individus ayant une sortie positive ou négative.

Au premier abord, il ne semble donc pas réellement pertinent de prendre en compte cette table dans notre modélisation.

En analysant les différents métiers envisagés, nous observons des taux de réussite différents entre les individus. Nous avons procédé à un traitement de la colonne contenant le libellé des métiers afin de la restructurer ; cet algorithme est détaillé (dans la partie 3.3). Nous étudions uniquement les métiers ayant au moins 10 occurrences et examinons le ratio de sorties positives par rapport au nombre total de sorties pour les individus du même métier.

Voici les métiers qui donnent les meilleurs ratios de sorties positives :

Métier	**Pourcentage de sorties positives**
Maraîchage	91,7%
Infirmier / Infirmière	81,8%
Mécanicien	70.5%
Secrétaire	64.3%
Retraite	62.2%
Ambulancier / Ambulancière	61.8%

TABLE 2.8 – Meilleurs ratios de sorties positives par métier

Voici les métiers qui donnent les pires ratios de sorties positives :

Métier	**Ratios de sorties positives**
Soudeur / Soudeuse	29.6%
Maçon / Maçonne	28.6%
Pas de projet	28.0%
Gardin immeuble	27.3%
Employé / Employée de restaurant	25.0%
Taxi	25.0%
Cuisinier / Cuisinière	23.1%
Facteur / Factrice	22.2%
Transport	21.6%
Couture	19.4%
Vendeur	14.8%

TABLE 2.9 – Pires ratios de sorties positives par métier

Cette approche nous permet d'identifier les métiers offrant les meilleures chances de succès et ceux qui semblent être associés à des taux de réussite inférieurs. Nous pouvons notamment remarquer que si l'individu n'a pas de projet clairement défini alors il aura moins de chance d'avoir une sortie positive.

En conclusion, cette table révèle des informations pertinentes pour nos futures modélisations. Les métiers envisagés semblent avoir un impact sur le statut de sortie, mettant en évidence l'importance de considérer ces données dans l'élaboration de nos modèles prédictifs.

2.4 Lien entre les différentes tables

Lors de l'analyse préliminaire des différentes tables, nous avons constaté que celles-ci ne contenaient pas le même nombre d'individus, ni même les mêmes individus. En effet, chaque table contient un nombre différent

de clés et ceci pourrait nous poser problème pour réaliser notre objectif, dans le sens où nous avons besoin de connaître le chemin complet d'un individu pour savoir ce qui a fonctionné.

Ainsi, nous pouvons récapituler le nombre d'individus qui se retrouvent d'une table à l'autre dans le tableau suivant :

	Difficultés	Action	Formation	Projet	Sortie
Difficultés	6489	1689	439	3290	2272
Action		3995	868	1653	1176
Formation			937	405	339
Projet				6919	4312
Sortie					4312

TABLE 2.10 – Correspondance entre les individus des différentes feuilles de données

De plus, nous aurions besoin de connaître plus particulièrement cette correspondance pour les tables `difficultes_export` - `action_export` - `sortie_export` car il s'agit des trois tables qui, comme nous avons pu le voir au dessus, marqueront une réelle importance dans notre modélisation. Ce nombre est de 895 individus communs, ce qui paraît peu élevé par rapport à ce que contiennent chacune des feuilles, mais ceci nous assure ici qu'en traitant ces personnes accompagnées, présentent dans la table `difficultes_export`, le seront également dans les tables `action_export` et `sortie_export`.

Afin de nous assurer que ne considérer que les 3416 individus, écartés pour notre étude, ne posera pas de perte d'information, nous avons étudié quelles étaient les sorties liées aux individus écartés. En effet, si, par exemple, la plupart des personnes écartées possédaient une sortie positive (ou a contrario négative), juste les supprimer nous introduirait un biais pour nos résultats de modélisations et nécessiteraient d'être étudiés plus en détails d'une façon différente.

Voici les résultats obtenus :

	Pourcentage d'individus concernés
Sorties positives	$24,12\%$
Sorties négatives	$10,07\%$
Sorties non renseignées	$65,81\%$

TABLE 2.11 – Sorties associées aux individus mis de côté

Ces résultats nous montrent que les individus que nous pensions retirer, dûs aux données manquantes pour certaines étapes clé de leur parcours, peuvent l'être sans que cela ne nécessite de précaution. En effet, la plupart de ces individus ne possédaient pas de sortie renseignée et étaient donc inexploitables de notre point de vue.

En conclusion, cette analyse descriptive nous a permis de mieux comprendre les leviers favorisant le retour vers l'emploi et donc d'identifier des variables clés discriminantes entre les sorties positives et négatives. Combinées à une modélisation pertinente, ces variables offrent un potentiel prometteur pour identifier les profils pris en charge et les parcours types associés. En outre, en ayant clarifié les approches pour exploiter efficacement les données de l'entreprise, nous sommes désormais prêts à les intégrer dans nos futures modélisations.

Chapitre 3

Traitement des Données Textuelles

3.1 Remarque sur le traitement du texte libre

Nous avons remarqué que bon nombre de commentaires utilisent des abréviations ou des acronymes, ce qui peut compliquer la compréhension du texte ou son traitement par nos algorithmes. Pour pallier ce problème, nous avons sollicité l'entreprise un glossaire des abréviations typiques à leur domaine d'activité, auquel nous avons ajouté diverses abréviations courantes utilisées dans la prise de notes. Par la suite, nous avons procédé au remplacement des abréviations présentes dans les textes par leur désignation complète. Cette démarche vise à améliorer la clarté des textes et à faciliter leur traitement.

3.2 Traitement du texte libre de la table `action_export`

Dans la table `action`, les colonnes 'Objectif', 'Résultat' et 'Commentaire' contiennent du texte libre. Par conséquent, nous devons le traiter avant d'appliquer nos algorithmes. Initialement, nous avons envisagé une classification non supervisée pour former des clusters de commentaires similaires, puis analyser ces clusters manuellement. Cependant, cet algorithme s'est avéré peu performant en raison de la diversité des commentaires en termes de contenus et de longueur. Nous avons constaté un grand nombre de clusters contenant un seul élément. En ce qui concerne les longs commentaires, nous avions pensé à utiliser un algorithme de traitement du langage qui permettrait de les résumer. Certains sont proposés dans la bibliothèque open-source d'Hugging Face.
Les résultats n'étant pas très probants pour la plupart des commentaires, nous avons abandonné l'idée.

Nous avons donc exploré une autre piste qui s'est avérée plus concluante. Il s'agit de la mise en place d'un algorithme de traitement du langage naturel (NLP) dans le but de déterminer si le commentaire était positif ou négatif. Ainsi, nous pouvons déduire si l'action a été positive ou négative. Pour cela, nous utilisons un algorithme pré-entraîné, ce qui signifie qu'il a été formé sur un grand ensemble de données avant d'être intégré à la bibliothèque. Les données sur lesquelles l'algorithme a été entraîné sont des données textuelles regroupant un grand nombre de commentaires comme par exemple ceux du tableau 3.1. En général, ces ensembles de données contiennent des milliers, voire des millions de phrases ou de documents, chacun étant marqué avec son sentiment (positif, négatif, neutre). Ce sont souvent des commentaires de films ou des analyses de tweets. Les pré-traitements nécessaires tels que la tokenisation (division du texte en unités plus petites (en mots généralement), et la vectorisation (conversion du texte en vecteurs numériques) sont gérés directement par la bibliothèque. Ainsi, après avoir concaténé les colonnes 'Objectif' 'Résultat' et 'Commentaire', nous appliquons l'algorithme pré-entraîné présent dans la librairie TextBlob.

Analysons les résultats. Voici quelques exemples de commentaires classés négatifs :

Exemple de commentaire	Polarité
Problème de retards	Négatif 0.9
Atelier commencé 20 minutes en retard, l'animateur ne s'est pas réveillé, ne lit pas ses mails	Négatif 0.9
Problèmes sur les sites, remontées des clients, conflits, retards...	Négatif 0.9

Table 3.1 – Quelques exemples de commentaires négatifs

Il s'agit de commentaires présentant une forte probabilité d'être négatifs. Nous remarquons que le mot "retard" est commun à ces commentaires, ce qui suggère qu'il joue un rôle important dans l'algorithme de classification.

Exemple de commentaire	Polarité
Evegeny a un super niveau enmaths, la difficulté est surtout dans la conjugaison et l'orthographe	Positif 0.5
OK	Positif 0.5
Christophe compose avec ses difficultés à l'écrit pour parer également à ses problèmes à l'oral, un certain manque de confiance en lui l'empêche d'exploiter pleinement ses capacités de se valoriser. Il progresse donc c'est très bien !	Positif 0.3

Table 3.2 – Quelques exemples de commentaires positifs

Pour les commentaires positifs, nous retrouvons des qualificatifs ou mots tels qu' 'excellent', 'très bien', 'parfait', 'OK'... Cependant, dans la grande majorité des cas, notre algorithme labellise les commentaires comme neutre. Voici la répartition des labels :

Catégorie	Nombre	Pourcentage
Négatif	1207	3.37%
Neutre	28699	80.03%
Positif	5955	16.61%

Table 3.3 – Répartition des labels avec pourcentages

Les commentaires neutres sont sur-représentés en raison du grand nombre de commentaires vides ou contenants des informations simples sur le contenu de l'action, justifiant ainsi le statut neutre. Pourtant, l'algorithme commet parfois des erreurs en attribuant cette étiquette, et il est difficile d'en quantifier la proportion, car nous n'avons pas accès aux vraies étiquettes des commentaires pour effectuer des comparaisons. Voici quelques exemples où l'algorithme commet des erreurs :

Exemple de commentaire	Polarité
Etat civil : demande de naturalisation Ne s'est pas présenté au RDV.	Neutre 0
Découverte du monde du travail (codes, attitudes, droits, devoirs en entreprise, marché de l'emploi, métier porteurs... Malika a une bonne connaissance du monde de l'entreprise. Un Très bon savoir-être. Elle a participé activement. Elle souhaite poursuivre dans le domaine de l'administratif.	Neutre 0
7 Séances bonne participation AUCUNE ABSENCE	Négatif 0.1

Table 3.4 – Quelques exemples d'erreurs de l'algorithme

Malheureusement, il est compliqué d'évaluer les performances de notre algorithme. Nous pourrions envisager de labelliser manuellement une petite partie des commentaires pour évaluer les performances de notre algorithme sur cet extrait du jeu de données. Par contre, nous n'avons pas réalisé ces tests, car ils sont assez longs à mettre en œuvre. De plus, en se basant sur les proportions des labels positifs, neutres et négatifs, nous pouvons supposer que la précision du modèle est relativement faible, car il ne classe pas beaucoup de commentaires en dehors de la catégorie neutre. Cependant, ceux qui sont classés en dehors de neutre semblent être plutôt bien classés.

L'algorithme mis en place demeure donc assez basique. Il est envisageable d'améliorer considérablement sa performance en entraînant un algorithme sur ces mêmes commentaires, préalablement étiquetés à la main. Cette

méthode permettrait d'obtenir un algorithme plus performant. Toutefois, comme nous l'avons dit précédemment, cette approche ne sera pas explorée en raison du temps considérable requis pour l'étiquetage manuel de ces commentaires.
Par ailleurs, nous aurions souhaité comparer les résultats de cet algorithme avec ceux d'un autre provenant d'une autre bibliothèque telle que BERT. Malheureusement, nous n'avons pas réussi à le mettre en œuvre et avons préféré ne pas nous attarder davantage sur cette alternative.

Les résultats suivants montrent le lien entre la polarité de l'action et la sortie. Ici, nous pouvons donc observer les moyennes des scores de polarité pour nos différents statuts de sortie :

Statut de sortie	Moyenne des scores de polarité
Négatif	0.034345
Positif	0.041804

TABLE 3.5 – Moyennes des scores de polarité par statut de sortie

En examinant ces moyennes, nous constatons que les sorties qualifiées de positives affichent des moyennes supérieures pour le score de polarité, suggérant qu'en moyenne, lorsque les actions reçoivent des commentaires positifs, l'individu a davantage de chances d'avoir une issue favorable. Or, en prenant en compte les intervalles de confiance, cette différence n'apparaît plus comme significative comme nous pouvons le voir sur les représentations des boîtes à moustaches ci-dessous où les intervalles de confiance se chevauchent :

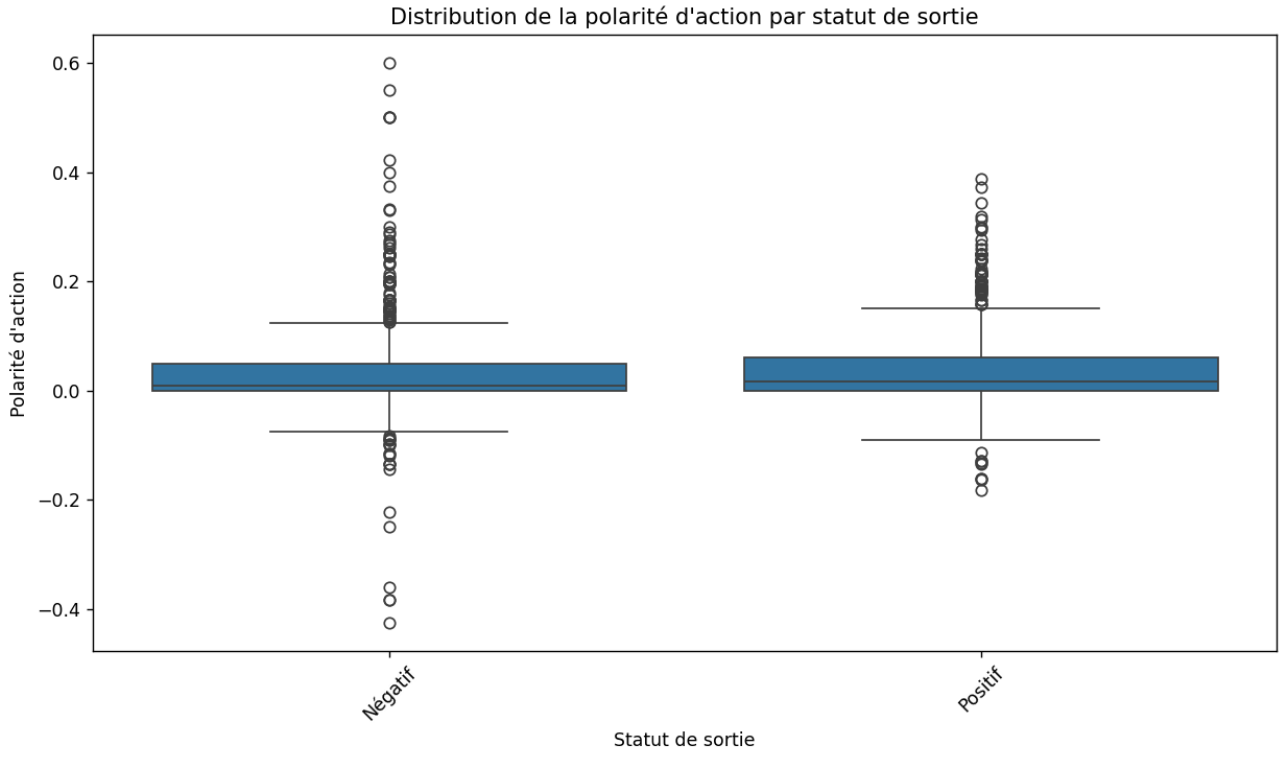


FIGURE 3.1 – Distribution du score de polarité en fonction du statut de sortie.

Bien que les moyennes ne soient pas statistiquement significatives une fois les intervalles de confiance pris en compte, nous observons une dispersion plus importante des valeurs pour les sorties négatives. Ainsi, les scores les plus bas correspondent majoritairement à des issues négatives, suggérant que les actions ayant des commentaires négatifs vont être associées à des sorties négatives.

Avoir cette information, nous permet donc d'avoir une caractéristique supplémentaire plus précise à prendre en compte dans nos modélisations.

3.3 Traitement du texte libre de la table `projet_export`

Pour rendre exploitable la colonne 'libelle_metier' qui contient actuellement des noms de métiers envisagés pour les individus en format de texte libre, notre objectif a été de structurer ces noms.
L'idée consiste à comparer les noms de métiers mentionnés avec ceux présents dans la base de données de l'ONISEP. Le premier objectif de cet algorithme est d'identifier clairement ceux qui n'ont pas de projet professionnel, car actuellement, pour les indicateurs d'absence de projet, on retrouve diverses formulations telles que : 'pas de projet', 'pas de projet professionnel', 'à définir', 'ne sait pas', ainsi que toutes les variantes avec des majuscules, sans ou avec accents, et parfois avec des fautes de frappe. Cela permet également d'uniformiser certains métiers écrits au masculin ou au féminin. Voyons donc comment l'algorithme s'en sort pour corriger ce texte. À noter que nous avons ajouté à la base de métiers de l'ONISEP des libellés qui revenaient souvent et qui n'étaient pas présents, comme 'pas de projet', 'à définir', 'permis B', etc.

Pour illustrer la diversité des expressions indiquant l'absence de projet professionnel, voici un tableau regroupant quelques exemples :

Table 3.6 – Correspondance entre les libellés métiers originaux et les métiers identifiés

Libellé métier original	Métier identifié
pas de projet	pas de projet
a definir	à définir
pas de projet professionnel concret	pas de projet professionnel
pas de projet professionnel pour le moment	pas de projet
ne sait pas	Ne sait pas
doit définir projet professionnel	pas de projet
pas de projet pro à ce jour	pas de projet
pas de projet définit quelques pistes envisageait bâtiment plomberie	pas de projet
pas de projet défini ne sait pas dans quel domaine il peut travailler	pas de projet
pas de projet defini avant sa mise en retraite	pas de projet
na pas de projet professionnel clairement établi souhaite seulement travailler	pas de projet
pas de projet professionnel a long terme	pas de projet professionnel
pas de projet clairement défini	pas de projet
a definir eboueur conducteur d engins tp	à définir
projet pro à définir peut etre le numerique	pas de projet
pas de projet professionnel précis aide ménagère domicile ou structure logistique travail de nuit	pas de projet
pp pas didee	pas de projet professionnel
pp à définir	à definir

Ces exemples illustrent comment notre algorithme peut normaliser les données, les rendant ainsi exploitables pour de futures analyses. Il reste simplement à fusionner les catégories pertinentes pour identifier les individus sans projet professionnel. Il est crucial de remplacer les abréviations, comme 'pp' qui est l'abréviation de 'projet professionnel'. Sans cette étape, notre algorithme aurait probablement attribué un label incorrect.

Regardons d'autres exemples pour les labels identifiés comme 'boulangerie', 'BTP', 'aide soignante' ou 'agent de sécurité' :

Table 3.7 – Correspondance entre les libellés métiers originaux et les métiers identifiés

Libellé métier original	Métier identifié
finition dans le bâtiment	bâtiment
aide soignante	aide-soignant / aide-soignante
cqp agent de sécurité sécurité incendie	agent de sécurité
aide à domicile	aide-soignant / aide-soignante
carrière dans le btp	BTP
soignante animaux	aide-soignant / aide-soignante
btp	BTP
domaine btp	BTP
aide boiseur	aide-soignant / aide-soignante
boulangerie	boulangerie
mécanique ou boulangerie	boulangerie
formation en boulangerie	boulangerie
espace vert, bâtiment, gardiennage, boulangerie, pâtisserie	boulangerie
fin de scolarité avant bac pro commerce, agent immo ou agent de sécurité	agent de sécurité
ouvrier du bâtiment	bâtiment
souhaite travailler dans le secteur du bâtiment	bâtiment
formation agent de sécurité à terme en attendant de reprendre une activité professionnelle dans différents domaines	agent de sécurité
manœuvre, agent polyvalent bâtiment	bâtiment
souhait de se former en boulangerie pâtisserie	boulangerie
mi-temps bâtiments, vente boulangerie	bâtiment
aide ménagère	aide-soignant / aide-soignante
agent de sécurité ou veilleur de nuit	agent de sécurité

Malheureusement, il est difficile d'estimer un taux d'erreurs ou la précision de cet algorithme. Une approche possible pour évaluer les performances serait d'étiqueter manuellement un extrait du jeu de données, de manière similaire au traitement du texte libre. Cependant, même cette méthode d'étiquetage serait fastidieuse à réaliser manuellement.

Pour améliorer cet algorithme, nous pourrions envisager de retravailler la base de données contenant les métiers. Nous avons constaté qu'il y avait des métiers très précis avec de longs intitulés qui sont souvent identifiés à tort. Supprimer ces métiers trop précis pour garder un intitulé plus général améliorerait les résultats. Or, l'objectif premier était de regrouper les libellés caractérisant une absence de projet, et dans ce cas, l'algorithme a été performant.

Ainsi, grâce à ces deux traitements, nous avons réussi à convertir des données non structurées en données catégorielles, facilitant ainsi leur exploitation tant d'un point de vue analytique qu'algorithmique.

Bien que nous puissions critiquer la précision des résultats obtenus, il convient de souligner que cela représente néanmoins un excellent point de départ, d'autant plus qu'aucun modèle existant (à notre connaissance du moins) n'a été entraîné sur des données similaires.

Notre analyse s'est principalement concentrée sur le traitement des données relatives aux métiers renseignés dans la table `projet_export`, ainsi que sur les commentaires, les résultats et les objectifs associés aux actions mises en place. Cependant, cette approche pourrait également être étendue aux données non structurées d'autres tables, ce qui pourrait fournir des informations supplémentaires sur l'efficacité des actions entreprises et leurs résultats. De même, une analyse similaire pourrait être appliquée aux projets eux-mêmes.

Chapitre 4

Modélisation

L'objectif de cette étape de modélisation est de passer de l'exploration descriptive à la création de modèles prédictifs qui peuvent nous aider pour comprendre et à anticiper les tendances et les comportements des personnes à accompagner.
Pour cela, nous avons choisi d'adopter deux points de vue différents mais également complémentaires. D'une part, nous détaillerons notre démarche pour prédire le statut de sortie des individus en fonction de divers critères. Cela nous permettra de pouvoir regarder si un individu est plus à même d'avoir une sortie positive que d'autres, mais aussi de pouvoir juger de la pertinence d'actions mises en place. D'autre part, nous chercherons à identifier les profils type en mettant en lumière les actions clé qui contribuent à une sortie positive pour eux dans le but de pouvoir accompagner de nouveaux individus aux profils similaires.

4.1 Prédiction du statut de sortie

L'objectif est de développer un algorithme utilisant la méthode de la forêt aléatoire (*Random Forest*) pour prédire le statut de sortie des individus. Les détails concernant le principe de fonctionnement de cet algorithme sont expliqués en *Annexe* 1. Nous imaginons que l'algorithme que nous allons détailler par la suite peut être employé de deux manières différentes :

1. **Identification précoce des individus à risque :** En utilisant les informations du bilan d'entrée et de `projet_export` il permet de détecter les individus les plus susceptibles d'avoir une sortie négative. Cette approche pourrait potentiellement permettre une prise en charge plus adaptée dès le début.
2. **Évaluation de l'efficacité de l'accompagnement :** Toujours dans le but de prédire le statut de sortie, en intégrant également les données des actions et formations (tables `action_export` et `formation_export`), nous pouvons imaginer que cette approche pourrait fournir une estimation de l'efficacité de l'accompagnement. En cas de prédiction d'une sortie non positive, cela pourrait indiquer la nécessité de prolonger l'accompagnement.

4.1.1 Traitement des variables et des données

En plus du pré-traitement préliminaire sur les données expliqué précédemment, nous effectuons un traitement sur les variables pour gérer les enregistrements multiples par individu. En effet, comme nous l'avons vu dans le *Chapitre* 2, pour la table `action_export` une ligne représente une action or, pour faire marcher notre algorithme, nous avons besoin d'une ligne par individu. Pour obtenir cela, nous avons utilisé une approche de consolidation en transformant les variables catégorielles en variables indicatrices et en agrégeant les valeurs pour obtenir un unique enregistrement par personne. De plus, certaines colonnes jugées redondantes ou non pertinentes pour notre modèle ont été éliminées comme discuté dans le chapitre précédent :

— Les variables relatives aux dates ont été éliminées, considérées comme non informatives.
— Certaines colonnes redondantes présentes dans différentes tables ont été supprimées.
— Les commentaires détaillant les difficultés ont été ignorés, car la colonne 'libelle', offre une synthèse normalisée de ces informations.
— Les informations concernant les centres de formation (adresse, nom) ont été exclues de l'analyse.

Il convient de souligner que nous avons délibérément conservé certaines variables qui semblaient initialement non pertinentes pour notre analyse descriptive, dans le but de nous assurer de ne pas négliger un élément important.

En complément, les traitements effectués dans le *Chapitre* 3 ont été appliqués et intégrés. Ainsi, un score mesurant la réussite d'une action, *objectif*, *resultat*, et *commentaire*, a été introduit, remplaçant ainsi ces trois colonnes par une nouvelle colonne *score*.

En parallèle, la colonne *nom_projet* de la table `projet_export` a elle aussi été nettoyée afin de normaliser au mieux nos données.

En accord avec notre analyse descriptive, nous avons intégré des colonnes supplémentaires, comme le nombre d'actions, pour enrichir nos données. De même, une variable indicatrice a été insérée pour repérer les individus ayant suivi une formation, permettant ainsi d'englober ceux absents de cette table spécifique. Cela vise à maximiser le nombre d'individus considérés lors de l'agrégation de nos différentes tables.

Ces pré-traitements effectués, nous pouvons maintenant construire notre modèle.

4.1.2 Modélisation

Nous avons choisi la forêt aléatoire pour sa robustesse et sa capacité à gérer les ensembles de données complexes. Sa facilité d'implémentation et son efficacité en font un choix idéal. De plus, la forêt aléatoire permet une interprétation facilitée des variables importantes, ce qui est essentiel pour comprendre les facteurs influençant les résultats.

Pour contrer le déséquilibre des classes positives et négatives, nous utiliserons SMOTE, un outil de sur-échantillonnage, qui génère des exemples synthétiques pour les classes minoritaires en interpolant les données existantes, visant ainsi à équilibrer les erreurs de prédiction. L'impact de SMOTE sera évalué à travers les résultats. Nous avons pareillement exploré le sous-échantillonnage, mais cette technique s'est révélée inadaptée à notre contexte, principalement en raison de la perte significative de données qu'elle implique. Les performances obtenues avec le sous-échantillonnage étaient inférieures à celles obtenues sans son utilisation.

Nous avons également tenté d'appliquer la régression logistique, mais les résultats n'ont pas été concluants, le modèle ayant tendance à classer tous les individus dans la classe majoritaire. C'est la raison pour laquelle nous allons détailler uniquement les résultats obtenus avec le modèle de *Random Forest*.

La classification des statuts de sortie nécessite une approche réfléchie pour capturer au mieux ces issues. Pour cela, nous avons testé plusieurs choix :

- **Choix n°1** : Ne pas toucher aux labels existants.
- **Choix n°2** : Regroupement des statuts indiquant une issue favorable ('durable', 'positif', 'transition vers l'emploi') sous une nouvelle catégorie 'Positif'.
- **Choix n°3** : Regroupement des statuts 'Autre' et 'A retirer' sous le label 'Négatif'.
- **Choix n°4** : Suppression ciblée des sorties que le modèle ne peut prédire telles que les décès ou les congés de longue durée (maternité, maladie).
- **Choix n°5** : Suppression de la catégorie 'Autre'.

Les choix mentionnés précédemment sont justifiés de la manière suivante. Le **choix n°2** est cohérent, car il regroupe sous le même label des statuts traduisant une issue positive. La différence entre ces trois labels réside principalement dans le type de contrat (CDI/CDD) et la durée (plus ou moins de six mois). Il en va de même pour le regroupement (**choix n°3**) de 'Autre' et 'A retirer', car aucun d'eux n'indique une sortie positive. Cependant, cela inclut des motifs de sortie dont le modèle n'a aucune information pour les prédire avec nos données, comme c'est le cas pour les congés de longue durée ou les décès. C'est pourquoi l'option de les retirer (**choix n°4**) peut être envisagée. Le dernier choix, plus radical, consiste à supprimer la catégorie 'Autre' (**choix n°5**), car elle sera difficile à prédire avec notre modèle.

Une fois les choix établis pour les labels, nous pouvons procéder à la division de notre jeu de données en ensembles d'entraînement et de test. Nous choisirons, pour l'ensemble des modèles, 80% pour l'entraînement et 20% pour le test. L'ensemble d'entraînement est utilisé pour construire et ajuster le modèle, tandis que l'ensemble de test, qui n'a pas été utilisé pendant l'entraînement, permet d'évaluer sa capacité à généraliser sur de nouvelles données. Cette séparation aide à prévenir le sur-apprentissage (ce qui impliquerait une réduction de ses capacités à généraliser correctement ses résultats sur de nouvelles données non vues auparavant).

4.1.3 Résultats pour l'identification préalable des individus à risque

Puisque notre intérêt se porte sur l'identification précoce des individus ayant le moins de chances d'obtenir une issue positive, en nous basant uniquement sur leur bilan d'entrée, nous utilisons les variables disponibles dans les tables `difficultes_export` et `projet_export`.

Nous décidons de ne pas modifier les statuts de sortie et obtenons la répartition suivante :

Statut de sortie	Proportion
Autre	44.02%
A retirer	15.51%
Positive	13.86%
Durable	13.74%
Transition	12.84%

TABLE 4.1 – Proportions des statuts de sortie avec 5 classes.

Nous observons une classe majoritaire, 'Autre', tandis que les autres classes sont similaires en termes de nombre d'individus. Cela nous amène à envisager la possibilité d'un ré-échantillonnage, par exemple avec l'utilisation de SMOTE.

Analysons les résultats obtenus par la validation croisée basée sur 30 simulations. Nous observons une précision globale du modèle de forêt aléatoire de 45.8%, avec un intervalle de confiance de 95% allant de 42.50% à 48.88%. Cette précision suggère que notre modèle a une capacité très limitée à prédire correctement le statut de sortie, la précision étant relativement faible, surtout en comparaison avec la proportion majoritaire de la catégorie 'Autre', comme en témoigne la répartition des erreurs avec la matrice de confusion :

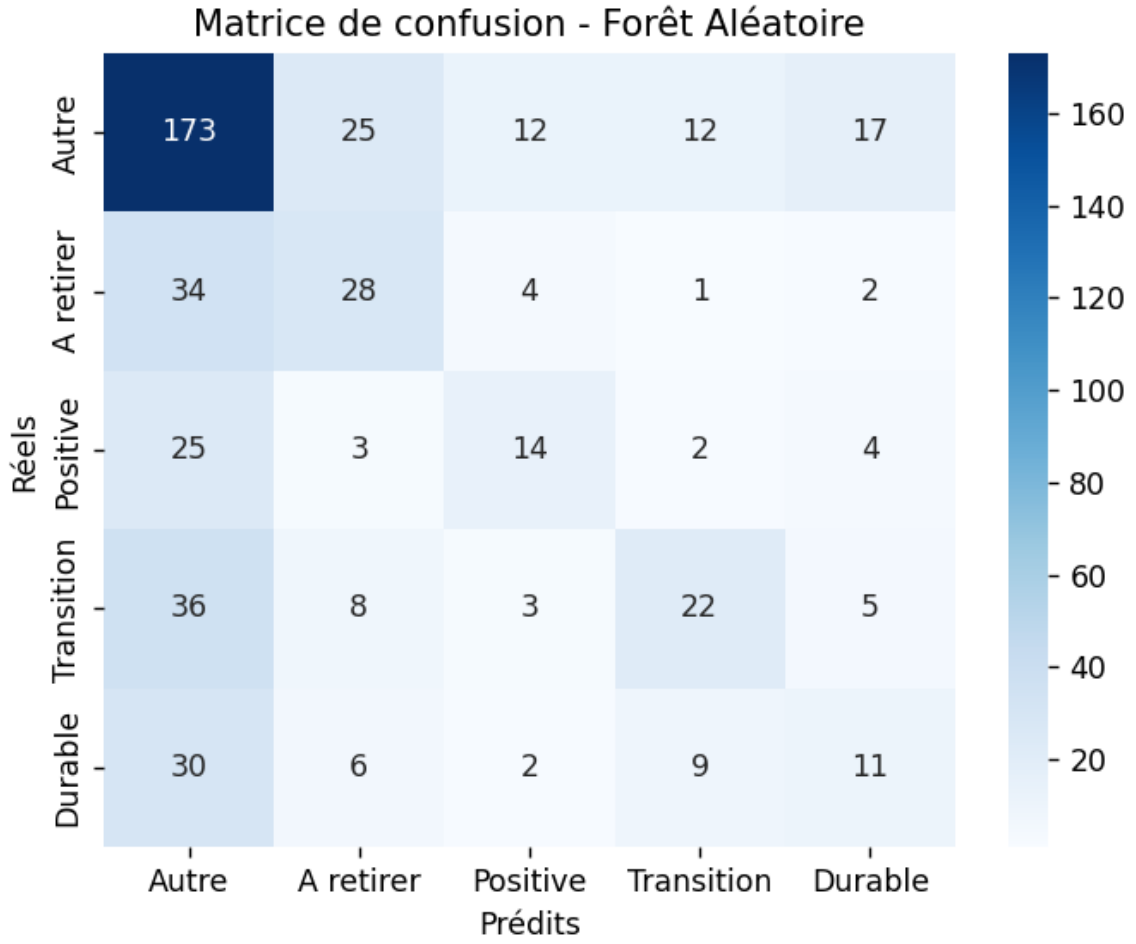


FIGURE 4.1 – Matrice de confusion

La classe 'Autre' affiche le plus grand nombre de prédictions correctes, mais en contrepartie, elle présente également un nombre relativement élevé de faux positifs, correspondant aux cas où le modèle a prédit 'Autre' alors que la classe réelle était différente. Cette observation suggère que le modèle a tendance à classifier les échantillons dans la catégorie 'Autre' plus fréquemment qu'il ne le devrait, ce déséquilibre pouvant être attribué à l'asymétrie dans les données d'entraînement. En revanche, pour les autres classes ('A retirer', 'Positive',

'Transition' et 'Durable'), les erreurs sont plus équitablement réparties entre les faux positifs et les faux négatifs.

Intéressons-nous aux variables les plus influentes dans ce modèle :

1. **Age**
2. **Code Postal**
3. **Nombre de Problèmes**
4. **Avancement Identifié** : Si les problèmes ont été identifiés
5. **Projet Identifié Autre** : les projets étiquetés 'Autre'.
6. **Administratifs / Judiciaires / Financiers** : Une catégorie de problème
7. **1. Logement** : Une catégorie de problème
8. **Résolution Interne et Externe** Le type de résolution des problèmes
9. **Numéro de Département**
10. **civilité M**

Globalement, ce modèle se base donc sur des variables qui ont intuitivement du sens pour prédire le statut de sortie, garantissant ainsi sa cohérence. Nous y voyons quand même que certaines variables que nous ne soupçonnions pas apparaissent tels que le code postal ou l'avancement.

Nous allons tenter d'améliorer les résultats en appliquant SMOTE. Les précisions obtenues sont les suivantes : une précision moyenne de 43.92% avec un intervalle de confiance entre 39.95% et 47.65%. Malheureusement, les résultats restent insatisfaisants, car notre modèle est en moyenne moins précis qu'un algorithme qui classerait tous les individus dans la classe majoritaire. La matrice de confusion n'est donc pas impactée significativement par ce ré-échantillonnage.

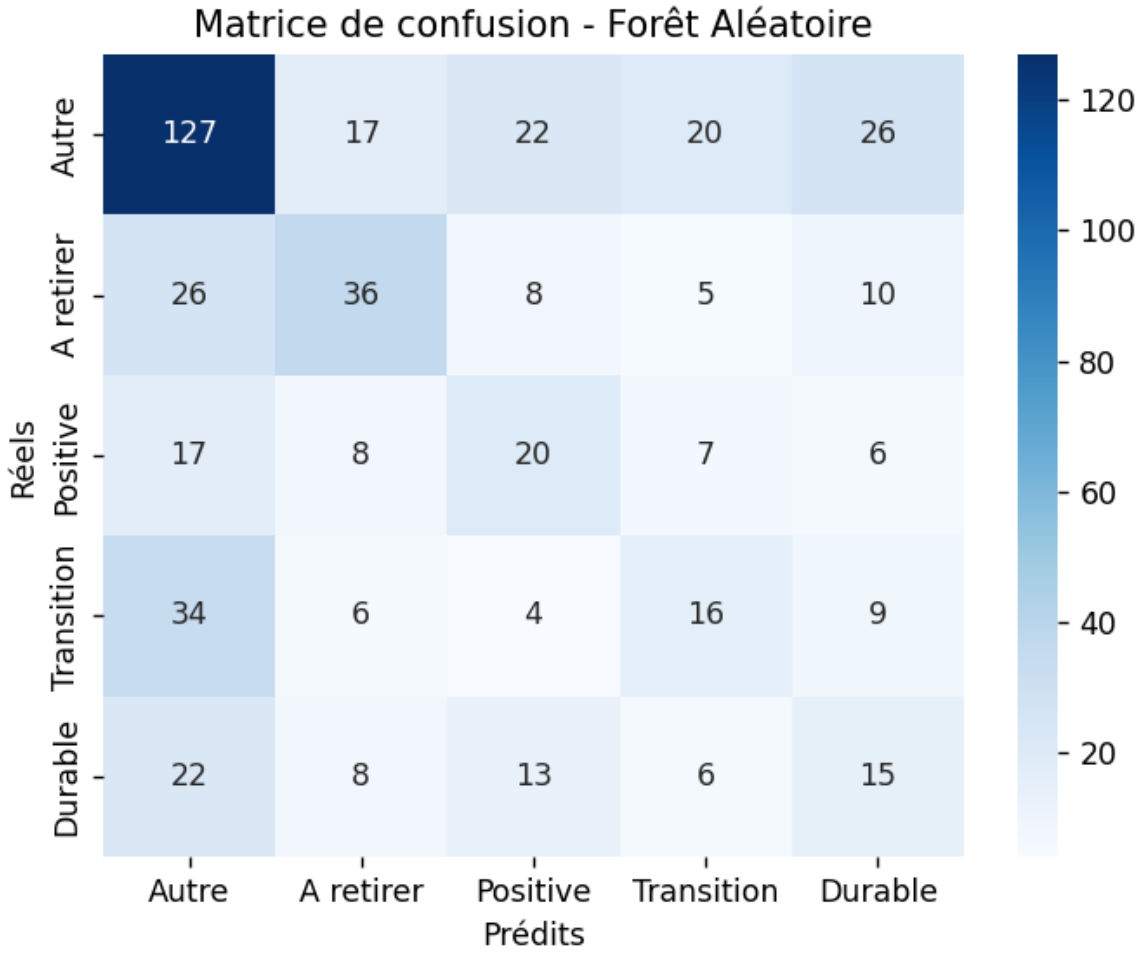


FIGURE 4.2 – Matrice de confusion après *SMOTE*

Certaines catégories sont mieux prédites au détriment d'autres, par exemple, la performance de prédiction de la catégorie 'A retirer' s'améliore aux dépens de la catégorie 'Durable'. Les types d'erreurs apparaissent plus équilibrés. Cependant, les performances de l'algorithme ne semblent pas suffisamment fiables pour être utilisées en pratique.

Nous avons donc tenté d'améliorer le traitement des labels en combinant les choix évoqués dans la partie 4.1.2 afin d'obtenir de meilleurs résultats. Cependant, comme détaillé dans la section 4.1.5, les résultats obtenus n'étaient pas plus satisfaisants : le taux de réussite du modèle étant souvent très proche de la proportion de la classe majoritaire.
Nous ne les détaillerons donc pas ici.

Ce problème de classification semble donc très difficile à résoudre en nous basant uniquement sur les données du bilan d'entrée. Nous verrons que l'ajout de plus de variables, notamment les actions entreprises pour chaque individu, nous permettra d'améliorer de manière significative les prédictions du modèle, les rendant ainsi plus satisfaisantes et utilisables en pratique.

4.1.4 Évaluation de l'efficacité de l'accompagnement

Comme mentionné dans l'introduction de cette section, nous allons désormais utiliser nos quatre tables pour prédire le statut de sortie. En nous concentrant sur les individus présents dans chacune des quatre tables, nous réduisons notre jeu de données à 895 individus.

4.1.4.1 1^{re} Classification : Nous ne touchons pas aux labels

Commençons par étudier la répartition des classes :

Statut de sortie	**Proportion**
Autre	34.18%
A retirer	22.69%
Positive	17.80%
Durable	12.81%
Transition	12.52%

TABLE 4.2 – Proportion des statuts de sortie avec 5 classes.

Étant donné le déséquilibre des classes, nous évaluerons ultérieurement l'impact de l'utilisation de SMOTE sur notre modèle.
La précision moyenne du modèle *Random Forest* basé sur ces informations simples, évaluée par validation croisée, s'élève à 52.96%. L'intervalle de confiance à 95% pour ces scores de précision se situe entre 48.83% et 57.32%. Il convient de noter que cet intervalle est assez large en raison de la taille réduite de l'échantillon de test.

Bien que ces résultats puissent sembler modestes à première vue, il est important de noter que certains motifs de sortie demeurent imprévisibles avec nos données. Cela inclut les congés de longue durée (maternité, maladie), les décès, ou les cas de personnes sans nouvelles, regroupés dans les statuts 'A retirer' ou 'Autre'. Par conséquent, il est attendu que notre modèle rencontre des difficultés à prédire ces statuts de sortie spécifiques. En revanche, les statuts de sortie 'Positive', 'Durable' et 'Transition' devraient être plus aisément prévisibles. Nous allons procéder à une vérification de cette hypothèse à travers l'analyse de la matrice de confusion.

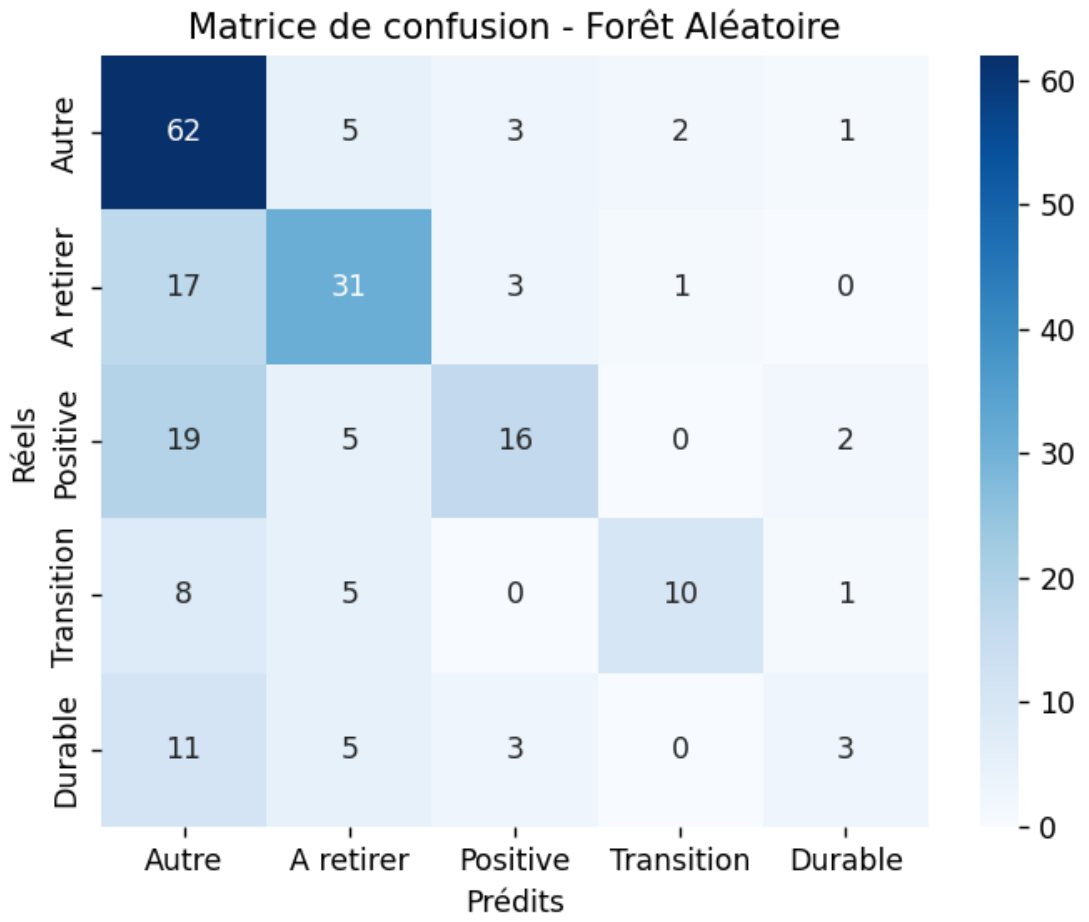


FIGURE 4.3 – Matrice de confusion

Effectivement, nous observons que le label 'Autre' est particulièrement compliqué à prédire. En revanche, les résultats sont très satisfaisants pour les statuts de sortie à connotation positive.

En effet, les labels 'Positive' 'Transition' et 'Durable', représentent tous les trois une sortie positive, mais la différence se fait au niveau du type de contrat (un contrat à durée déterminée (CDD) ou à durée indéterminée (CDI) de plus ou moins de six mois). Il est donc à noter que la confusion entre ces catégories lors de la classification peut être considérée comme peu importante, étant donné que toutes représentent des issues positives pour les individus concernés.

Nous allons désormais analyser les variables d'influence, c'est-à-dire les facteurs qui jouent un rôle significatif dans la prédiction du statut de sortie par notre modèle. L'identification de ces variables nous permettra de comprendre quels aspects des données contribuent le plus à la décision du modèle.

Les 10 variables les plus influentes dans la prédiction du statut de sortie, selon notre modèle de forêt aléatoire, sont les suivantes :

1. **Durée** : Le temps consacré aux actions.
2. **Nombre d'actions** : La quantité d'action.
3. **Polarité de l'action** : La nature des actions (positive ou négative).
4. **Âge**
5. **Libellé de l'action - Entretien et bilan régulier** Les Actions portant le libelle Entretien et bilan régulier
6. **Nombre de projets**
7. **Nombre de problèmes**
8. **Avancement - Identifié** : L'état d'avancement des difficultés
9. **Code postal**
10. **Libellé de l'action - Suivi social** : Les Actions portant le Suivi social.

Les variables identifiées comme influentes dans notre modèle sont en adéquation avec nos attentes concernant les facteurs déterminants du statut de sortie. Nous y retrouvons des caractéristiques propres au profil, mais aussi principalement d'autres liées aux actions entreprises. Le fait que les actions régulières, telles que les entretiens et les bilans réguliers, ressortent parmi les variables les plus importantes montre l'efficacité d'un suivi continu.

Étant donné le déséquilibre des classes, l'utilisation de SMOTE peut également être pertinente dans ce cas, cependant, les résultats obtenus sont moins satisfaisants, avec une moyenne de 50.9%. L'intervalle de confiance à 95% pour ces scores de précision se situe entre 45.6% et 55.3%. Comme précédemment, ces résultats sont évalués après validation croisée.

4.1.4.2 2^{eme} Classification : Regroupement des positifs et suppression du statut 'Autre'

Nous avons observé que notre modèle rencontrait des difficultés à prédire correctement la catégorie 'Autre', principalement composée des motifs de sortie 'Sans nouvelle' et 'Chômage'. Afin d'améliorer les performances de prédiction, nous envisageons donc d'exclure cette catégorie du modèle. L'objectif est de concentrer l'apprentissage sur des catégories présentant des motifs de sortie plus clairs et distincts.

De plus, nous allons appliquer les choix numéro 2 et 4 (définis en introduction de partie) concernant le traitement des labels.
Nous considérons désormais deux catégories distinctes : 'Positif' et 'A retirer' et nous renommons les 'A retirer' par Négatif pour plus de clarté.

Étudions la proportion de nos deux nouvelles classes.

Statut de sortie	**Proportion**
Positive	$65,52\%$
Négative	$34,48\%$

Table 4.3 – Proportion des statuts de sortie avec 2 classes.

Une fois de plus, nous constatons un déséquilibre dans nos classes, ce qui nous amènera à étudier l'impact de SMOTE dans un second temps.

Nous anticipons une nette amélioration des performances de notre modèle par rapport à la version précédente pour plusieurs raisons. Tout d'abord, les classes 'Positive', 'Transition' et 'Durable' ont été prédites avec une grande précision auparavant. En les regroupant en une seule catégorie, nous nous attendons à maintenir, voire à améliorer, cette performance de prédiction. De plus, en éliminant la catégorie 'Autre', qui était difficile à prédire, et en réduisant le nombre de labels à seulement deux, nous simplifions considérablement le problème de classification.

Les performances du modèle, avec une moyenne de $77,79\%$ et un intervalle de confiance à 95% situé entre 73.18% et $83,45\%$, illustrent donc cette amélioration par rapport aux approches précédentes. Comparé à la proportion de 65.52% de la classe majoritaire, ces performances démontrent clairement que ce modèle surpasse largement un algorithme naïf, qui se limiterait à assigner systématiquement toutes les prédictions à la classe majoritaire.

Cela indique que notre modèle ne se contente pas de suivre une stratégie de prédiction basique, mais qu'il capte effectivement les nuances dans les données pour faire des prédictions justifiées.

Intéressons nous à la matrice de confusion pour ce modèle :

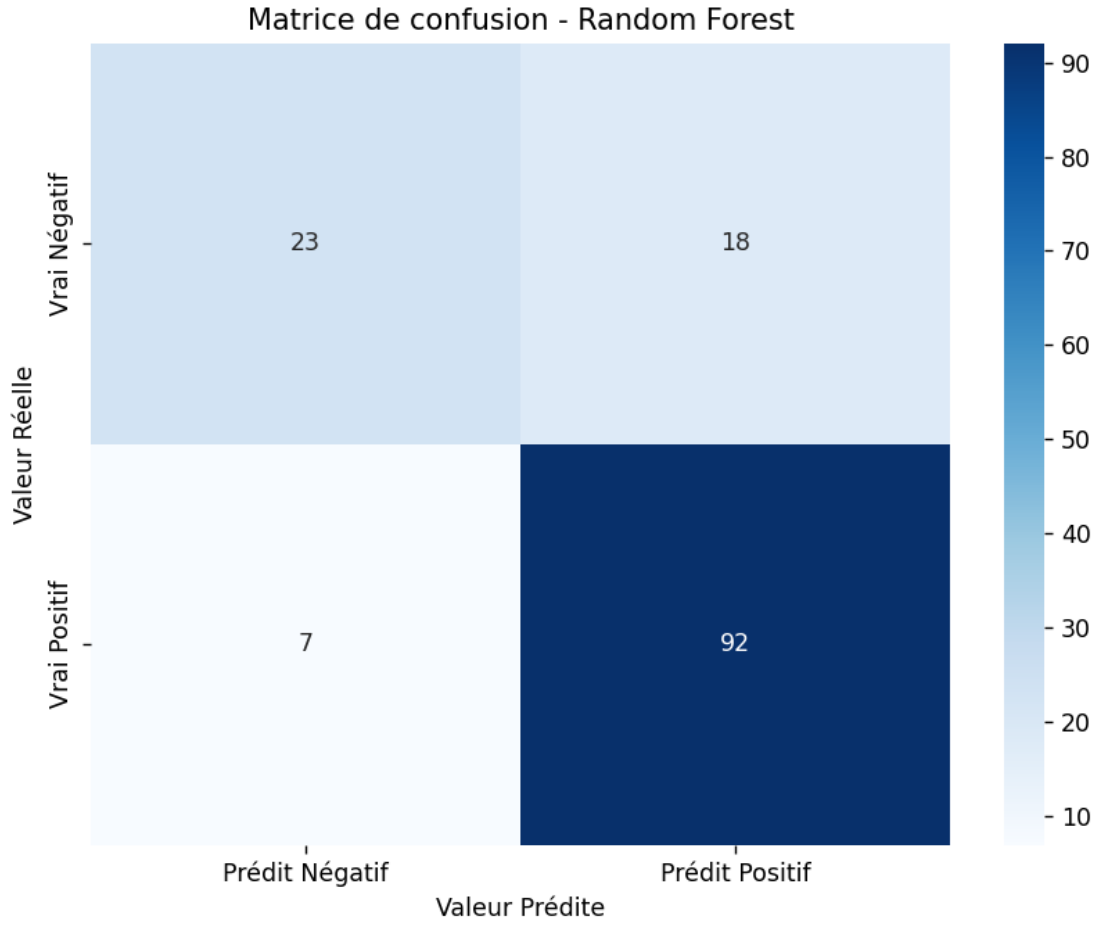


FIGURE 4.4 – Matrice de confusion après le regroupement des catégories positives et la suppression du label 'Autre'

Nous remarquons que les erreurs de prédiction présentent un déséquilibre, avec un taux de faux positifs supérieur au taux de faux négatifs. Malheureusement, l'erreur la plus critique, celle des faux positifs, est donc la plus fréquente. En effet, prédire à tort qu'un statut de sortie est positif quand le label réel est négatif a des implications plus graves, car cela signifie que l'on passe à côté de l'opportunité d'identifier les personnes à risque de sortie négative. Si ces cas étaient correctement identifiés, nous pourrions envisager pour eux un accompagnement supplémentaire, potentiellement améliorant leurs chances de réussite.

Regardons les variables les plus influentes dans notre modèle :

1. **Durée** : Le temps consacré aux actions.
2. **Nombre d'actions** : La quantité d'action.
3. **Libellé de l'action - Entretien et bilan régulier** : Les actions portant le libellé Entretien et bilan régulier.
4. **Polarité de l'action** : La nature des actions (positive ou négative).
5. **Âge**
6. **Nombre de problèmes**
7. **Avancement - Identifié** : L'état d'avancement des difficultés.
8. **Adaptation / Évolution des compétences** : L'une des catégories possibles pour les difficultés
9. **Libellé de l'action - Suivi social** : Les actions portant le suivi social.
10. **Durée formation**

Nous constatons donc une cohérence avec le modèle précédent, car nous retrouvons principalement les mêmes variables importantes. De plus, de nouvelles variables émergent comme significatives, telles que la difficulté liée à 'l'Adaptation / Évolution des compétences'. Cette comparaison met en lumière la cohérence dans la sélection

des variables influentes par le modèle.

Étant donné que nous disposons uniquement de deux labels, le rôle de ces variables est plus facilement interprétable. De plus, la plupart de ces variables ont été analysées dans l'étude descriptive des différentes tables. Nous avons constaté, par exemple, que l'augmentation du nombre d'actions est associée à un taux de sortie plus négatif, tandis que les actions de suivi social sont liées à un taux de réussite élevé. Les entretiens et bilans réguliers, quant à eux, présentent un taux de réussite moyen d'environ 50%, ce qui suggère que la prédiction est également influencée par la fréquence de ces activités. En ce qui concerne l'adaptation/évolution des compétences, il a été observé que c'est le type de difficulté avec le pire taux de réussite, indiquant ainsi que l'absence de ces problèmes est préférable pour favoriser une sortie positive.

Nous allons appliquer SMOTE à nos données dans l'espoir d'améliorer la précision du modèle. Cependant, nous obtenons des résultats comparables au modèle sans ré-échantillonnage, avec une précision moyenne de 78.4%, et un intervalle de confiance situé entre 73.57% et 84.16%, construit après validation croisée. Malheureusement, l'amélioration n'est pas significative. Regardons plus en détails l'impact sur la matrice de confusion :

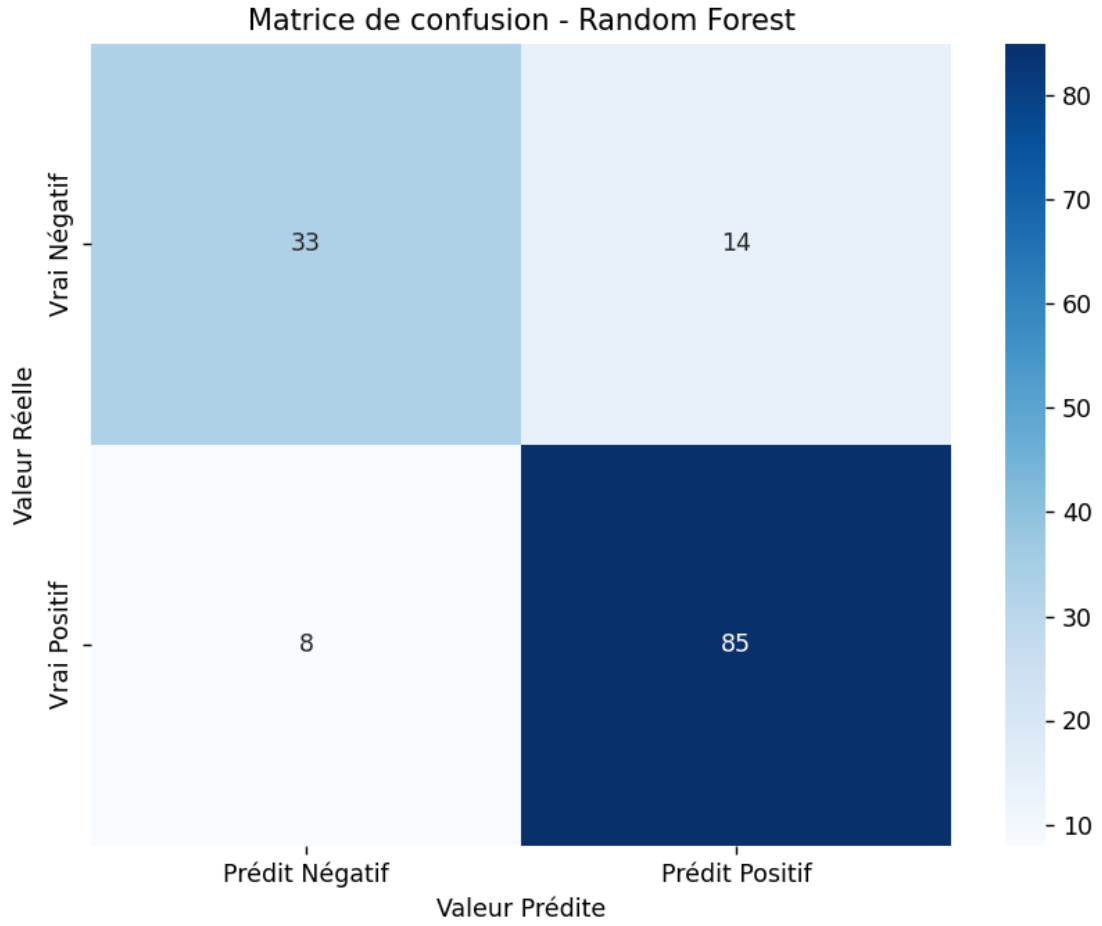


FIGURE 4.5 – Matrice de confusion avec l'utilisation de SMOTE

Nous constatons une diminution des faux positifs dans la matrice de confusion. Cependant, étant donné que cette observation est basée sur une seule itération de l'algorithme et compte tenu de la variabilité inhérente à nos résultats, nous ne pouvons pas affirmer avec certitude que cette réduction est statistiquement significative. Malgré cela, l'application de la technique SMOTE semble contribuer à équilibrer les erreurs de notre modèle. Globalement, les performances demeurent satisfaisantes.

Ainsi, ce modèle peut constituer un indicateur utile pour évaluer si un individu, en considérant un ensemble d'actions à un moment donné, a de bonnes chances d'obtenir une issue positive.

4.1.5 Résumé de nos résultats

Nous avons également exploré d'autres options pour le traitement des labels, dont les résultats sont récapitulés dans le tableau ci-dessous. Il convient de rappeler que pour l'identification initiale des individus à risque, nous avons utilisé uniquement le bilan d'entrée, tandis que pour la prédiction du statut de sortie, nous avons exploité les données provenant des quatre tables disponibles en plus de la sortie.

TABLE 4.4 – Résultats des modèles Random Forest et Random Forest avec $SMOTE$ pour différents choix de catégorisation

Catégorisation	Random Forest		Random Forest avec $SMOTE$		Proportion du label majoritaire
Prédiction du statut de sortie avec uniquement le bilan d'entrée.					
Sans toucher aux labels	46.35%	[41.7, 49.55]	43.9%	[41.26, 46.6]	0.44
En regroupant positif et négatif	62.2%	[58.2, 66]	62.0%	[58.2, 65]	0.59
En regroupant positif et négatif et supprimant la catégorie Autre	76.4%	[72.5, 79.9]	75.45%	[72.06, 79.8]	0.72
Avec des suppressions ciblées de motifs de sortie imprévisibles par nature (Décès, Congés de longue durée (maternité, maladie)) et on regroupe les positifs et négatifs	63.5%	[59.86, 67.32]	61.85%	[58.36, 66.8]	0.60
Prédiction du statut de sortie avec les 5 tables					
Sans toucher aux labels	53.0%	[47.75, 58.65]	50.59%	[46.0, 54.9]	0.34
En regroupant positif et négatif	70.2%	[65.9, 75.2]	69.4%	[64.19, 73.62]	0.56
En regroupant positif et négatif et supprimant la catégorie Autre	78.7%	[73.89, 83.64]	78.33%	[71.23, 84.28]	0.65
Avec des suppressions ciblées de motifs de sortie imprévisibles par nature (Décès, Congés de longue durée (maternité, maladie))	70.3%	[65.5, 76.3]	68%	[62.2, 75.3]	0.57

La modification des labels, notamment le regroupement des catégories positives et négatives et en supprimant la catégorie 'Autre', a significativement amélioré la précision des modèles. Cela suggère que simplifier les catégories de sortie peut contribuer à réduire la complexité du problème et à améliorer l'efficacité de la classification. De plus, les informations contenues dans les tables actions sont importantes pour déterminer le label, car les performances sont nettement meilleures lorsque ces données sont prises en compte, démontrant ainsi leur impact sur le statut de sortie. Étant donné que la prédiction du label 'Autre' était particulièrement complexe, il est donc logique d'observer une nette amélioration des résultats lors de son exclusion.

4.1.6 Limite du modèle et perspectives d'améliorations

Nous avons ainsi développé un modèle capable de prédire le statut de sortie avec une précision acceptable (en moyenne 78.7%), en utilisant les informations présentes dans l'ensemble des tables. En revanche, le modèle basé uniquement sur le bilan d'entrée semble inexploitable en raison des performances insuffisantes. Cependant, plusieurs pistes pourraient être explorées pour améliorer leur précision.

L'une des pistes à explorer serait la redéfinition des labels. Nous pourrions envisager la création d'une nouvelle classe neutre ou d'une classe regroupant les événements sur lesquels les actions ne peuvent pas avoir d'impact, tels que les décès, les congés de longue durée pour maternité ou maladie.
De plus, nous avons observé que certaines personnes, par exemple, ont pour projet de passer leur permis de conduire, mais si elles ne trouvent pas d'emploi, leur sortie n'est pas considérée comme positive. Il serait donc intéressant de pouvoir prendre en compte ces réalisations pour nuancer les sorties.
En résumé, il serait donc pertinent d'avoir une meilleure définition des labels, robuste aux sorties imprévisibles et peut-être plus adaptée aux projets des individus.

Ensuite, une autre perspective d'amélioration concerne le traitement du texte libre. L'algorithme actuel pour déterminer le score de réussite d'une action est assez sommaire. Nous pouvons imaginer que si nous parvenons à améliorer la précision à ce niveau, nous gagnerions également en précision sur le modèle dans son ensemble.

Enfin, pour améliorer davantage notre modèle, il serait préférable de disposer d'un plus grand nombre d'individus, comme nous l'avions souligné dans la partie 2.4, puisque cet échantillon ne compte que 895 individus.

En outre, il est important de noter que bien que notre modèle soit utile, il est loin d'être parfait et ne remplace pas le jugement humain et le suivi apporté, surtout lorsqu'il s'applique à des individus. Cela est d'autant plus vrai que les entretiens et les bilans réguliers, ressortent parmi les variables les plus importantes.

4.2 Recherche de profils type et des actions à associer

Pour notre seconde phase de modélisation, notre objectif était de créer un algorithme de clustering pour identifier les différents profils à partir des données disponibles. L'idée sous-jacente était d'utiliser ces profils comme base pour prédire les actions à entreprendre en vue d'une réinsertion réussie. Dans les prochains paragraphes, nous détaillerons les différentes étapes de cette approche ainsi que les méthodes que nous avons employées.

Nous avons conçu un algorithme qui prend en entrée les difficultés identifiées lors du diagnostic initial ainsi que les autres caractéristiques spécifiques à l'individu. Nous ne considérerons donc ici que celles que nous avons jugées pertinentes dans la table `difficultes_export` suite à notre analyse descriptive. Nous écartons les variables issues des autres tables, car outre bien sûr les actions, leur nombre et durée, nous partons du principe que des informations liées à la formulation d'un projet professionnel et à la participation à une formation ne sont pas connus lors du diagnostic initial et ne correspondent pas aux informations que nous souhaitons prédire.
Cet algorithme renverrait alors une liste d'actions à entreprendre, classée de la plus prometteuse à la moins prometteuse. Nous expliquerons en détail les critères utilisés dans ce processus.

4.2.1 Traitement des variables et des données

Comme dans le modèle précédent, le pré-traitement des données implique des manipulations simples sur la table `difficulte_export`. Nous avons commencé par éliminer les colonnes correspondant aux caractéristiques que nous avons choisi de ne pas conserver, telles que 'situation', 'societe', 'code_postal', 'num_dep', 'date_reso_diff', 'avancement', 'commentaire', 'resolution', 'date_difficulte' et 'situation'. Cette démarche nous permet de travailler avec un dataframe plus léger et concentré sur les informations pertinentes. Dans la même optique, nous avons supprimé toutes les lignes relatives aux individus pour lesquels aucune difficulté n'a été identifiée jusqu'à présent.

Par ailleurs, comme lors du traitement précédent, l'aspect central de cette étape réside dans la transformation des données afin qu'une seule ligne de la table représente un individu dans son intégralité. Pour ce faire, plutôt que de modéliser les différentes difficultés par des indicatrices, nous avons choisi une approche différente. Nous avons regroupé toutes les difficultés rencontrées par chaque individu dans une liste unique de chaînes de caractères, organisées dans le même ordre déterminé par le numéro de difficulté, afin d'éviter les permutations. Cette méthode nous a permis d'obtenir directement une caractéristique distinctive représentée par une liste, sans avoir à ajouter de nombreuses colonnes. Nous avons adopté le même procédé pour les libellés associés.

4.2.2 Modélisation

Le raisonnement derrière cette modélisation est très simple et est directement inspiré de la demande. Nous avons décomposé l'objectif global en trois sous-objectifs distincts : l'identification des profils type, la recherche

des actions correspondantes, et enfin, la prédiction des actions recommandées pour un profil donné.
Dans la suite, nous allons détailler chacune de ces étapes ainsi que les méthodes que nous avons employées.

4.2.2.1 Recherche des profils types

Dès le départ, nous avons pensé aux méthodes de segmentation mises en place en marketing par exemple pour étudier les profils des clients et décider des actions à mettre en place conformément à cela.

Nous nous sommes aperçus qu'en termes de modélisation, ce terme de "segmentation" s'apparentait tout simplement aux méthodes de clusterings. Ces méthodes sont des techniques d'analyse de données, basées sur de l'apprentissage non supervisé, qui permettent de regrouper des éléments similaires en fonction de critères de distance. Ce qui rend cette approche particulièrement intéressante dans notre contexte, c'est qu'elle peut être appliquée à des données non étiquetées. Ainsi, nous sommes en mesure d'identifier et de classifier différents profils sans avoir préalablement défini de catégories spécifiques.

Parmi ces méthodes, nous pouvons citer les algorithmes des Kmeans ou la classification ascendante hiérarchique très connus. Le problème étant qu'ils ne pouvaient pas s'appliquer à nos données, toutes catégorielles, sans encodage et donc un pré-traitement beaucoup plus lourd. Nous avons alors découvert le modèle des Kmodes.

Il s'agit d'une variante de l'algorithme des Kmeans adaptée aux données catégorielles. Les détails concernant le principe de fonctionnement des Kmeans sont d'ailleurs exposés en *Annexe* 2.
Ce nouvel algorithme utilise la distance de Hamming pour évaluer la dissimilarité entre les points de données et les centres des clusters. Il procède par deux étapes : une étape d'affectation, où chaque point est assigné au cluster dont le centre est le plus proche, et une étape de mise à jour, où chaque centre de cluster est remplacé par le mode (la valeur la plus fréquente) de chaque attribut catégoriel dans le cluster.

Globalement les hyperparamètres sont les mêmes pour les deux algorithmes. Nous avons toutefois effectué plusieurs essais sur un petit échantillon de données (sur un département) pour sélectionner les plus optimaux pour notre étude. L'idée était d'abord de pouvoir vérifier rapidement la pertinence et l'efficacité de ce nouvel algorithme, mais aussi de comprendre comment l'appliquer.

Le premier paramètre à sélectionner est le nombre de clusters. Pour le déterminer, nous avons tracé le coût associé à la classification en fonction du nombre de clusters. Ce coût représente la somme des distances entre les points et leurs centres de cluster les plus proches et permet donc de juger la qualité des clusters obtenus. Nous y cherchons la valeur pour laquelle nous y voyons un coude.

Voici un exemple pour le département de l'Aisne :

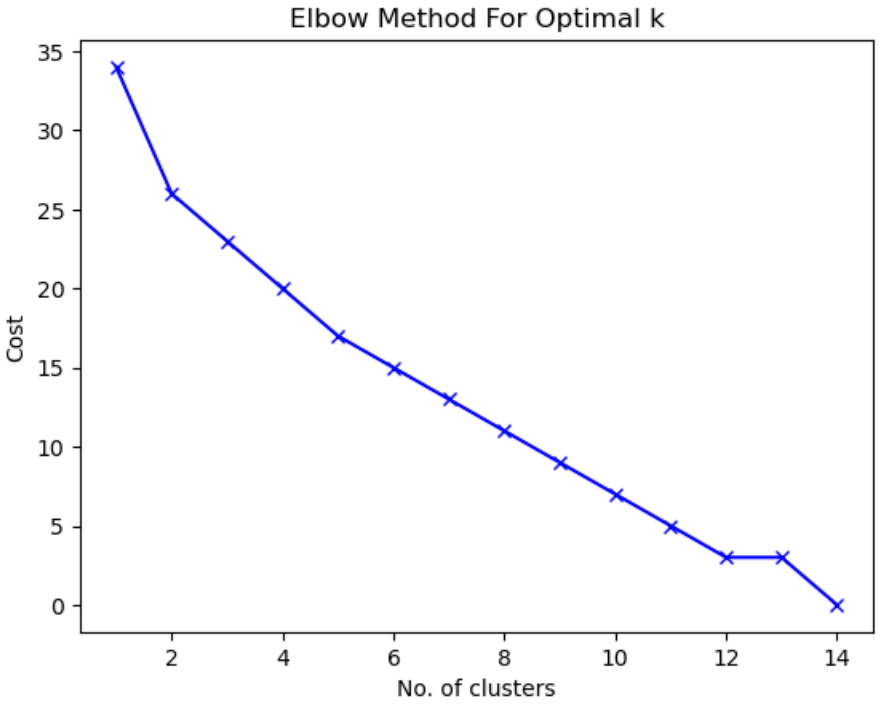


FIGURE 4.6 – Recherche du nombre de clusters optimaux pour les données concernant le département de l'Aisne (15 individus)

Sur cet exemple, le point d'inflexion n'est pas très net. La courbe en présente un pour les valeurs 2 et 12. Deux nous semblent trop peu, nous pouvons choisir donc 12 clusters pour classifier les individus étudiés.

Nous nous sommes cependant aperçus que les résultats associés à ce clustering pouvaient être très bons, comme mauvais, selon l'itération. Afin de pallier à cela, nous avons donc augmenté le nombre d'itérations (n_init) correspondant au nombre de fois que l'algorithme est exécuté avec des centres initiaux différents ; mais nous avons également testé et comparé les différentes méthodes d'initiation possibles, à savoir 'random', 'Huang' et 'Cao'.

Il est à noter que le numéro de cluster attribué peut varier d'une itération à l'autre. L'objectif est de ne pas tenir compte de ces variations, mais plutôt d'examiner la répartition des individus dans ces clusters.
Voici un aperçu de nos résultats :

cle	Cluster	Cluster	Cluster	Cluster	Cluster	civilite	nom_dep	nom_region	cat_difficulte	libelle	tranche_age
100080269140238	8	1	6	0	2	M	Aisne	Hauts-de-France	1. Logement,2. Mobilité	Difficultés de maintien dans le logement,Absen...	18-25
161019938301862	5	3	4	3	4	M	Aisne	Hauts-de-France	5. Administratifs / Judiciaires/ Financiers	Accès aux Droits	>45
166060236115334	4	10	9	6	10	M	Aisne	Hauts-de-France	4. Santé	Problématique physique	>45
167030269111368	10	9	7	8	3	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Difficulté avec les techniques de Recherche d'...	>45
17106940690980901/05/2023	6	9	7	8	3	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Méconnaissance du marché du travail	>45
173109920603957	0	4	10	2	0	M	Aisne	Hauts-de-France	2. Mobilité	S Mobilité : absence de permis (non obtenu, su...	>45
17812024080418722/03/2023	5	3	4	3	7	M	Aisne	Hauts-de-France	3. Problématiques sociales	Prise en charge d'une personne à charge	>45
18805026911589323/09/2022	9	8	5	1	9	M	Aisne	Hauts-de-France	1. Logement	Sans domicile fixe	35-45
190099505201679	0	2	3	9	6	M	Aisne	Hauts-de-France	3. Problématiques sociales	Pas de mode de garde adapté	25-35
19102023612164015/05/2023	0	6	8	2	0	M	Aisne	Hauts-de-France	2. Mobilité	Absence de permis (non obtenu / suspendu / perdu)	25-35
19406514546150412/10/2022	1	5	0	4	0	M	Aisne	Hauts-de-France	1. Logement,2. Mobilité	Logement inadapté/vétuste/insalubre,Absence de...	25-35
19408273752314312/10/2022	0	6	8	2	0	M	Aisne	Hauts-de-France	2. Mobilité	Absence de permis (non obtenu / suspendu / perdu)	25-35
198030269133078	7	7	0	7	5	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Difficulté avec les techniques de Recherche d'...	25-35
198040269150021	3	2	2	5	8	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences,2. M...	Analphabétisation / illettrisme / non maîtrise...	25-35
266100269103464	2	0	1	10	1	Mme	Aisne	Hauts-de-France	5. Administratifs / Judiciaires/ Financiers,6....	Accès aux Droits,Difficulté avec les technique...	>45

FIGURE 4.7 – Résultats du clustering pour les données concernant le département de l'Aisne (15 individus) avec 'Random'

L'analyse des cinq itérations indépendantes révèle que les individus présentant exactement le même profil appartiennent systématiquement aux mêmes clusters, même si leur étiquette de cluster peut changer (comme illustré par les lignes 10 et 12). Pourtant, il est également observé (par exemple, les lignes 2 et 7) que certains individus peuvent appartenir au même cluster lors de certaines itérations, même s'ils présentent des problèmes totalement différents.

cle	Cluster	Cluster	Cluster	Cluster	Cluster	civilite	nom_dep	nom_region	cat_difficulte	libelle	tranche_age
100080269140238	1	0	7	1	3	M	Aisne	Hauts-de-France	1. Logement,2. Mobilité	Difficultés de maintien dans le logement	18-25
161019938301862	5	6	1	6	2	M	Aisne	Hauts-de-France	5. Administratifs / Judiciaires/ Financiers	Accès aux Droits	>45
166060236115334	8	1	2	0	6	M	Aisne	Hauts-de-France	4. Santé	Problématique physique	>45
167030269111368	3	3	0	10	0	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Difficulté avec les techniques de Recherche d'...	>45
171069406909809 01/05/2023	3	4	3	10	8	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Méconnaissance du marché du travail	>45
173109920603957	9	2	4	7	5	M	Aisne	Hauts-de-France	2. Mobilité	S Mobilité : absence de permis (non obtenu, su...	>45
178120240804187 22/03/2023	10	10	5	2	4	M	Aisne	Hauts-de-France	3. Problématiques sociales	Prise en charge d'une personne à charge	>45
188050269115893 23/09/2022	2	9	9	5	7	M	Aisne	Hauts-de-France	1. Logement	Sans domicile fixe	35-45
190099505201679	7	5	8	8	10	M	Aisne	Hauts-de-France	3. Problématiques sociales	Pas de mode de garde adapté	25-35
191020236121640 15/05/2023	4	8	6	4	1	M	Aisne	Hauts-de-France	2. Mobilité	Absence de permis (non obtenu / suspendu / perdu)	25-35
194065145461504 12/10/2022	6	7	10	3	9	M	Aisne	Hauts-de-France	1. Logement,2. Mobilité	Logement inadapté/vétuste/insalubre	25-35
194082737523143 12/10/2022	4	8	6	4	1	M	Aisne	Hauts-de-France	2. Mobilité	Absence de permis (non obtenu / suspendu / perdu)	25-35
198030269133078	3	3	0	9	0	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Difficulté avec les techniques de Recherche d'...	25-35
198040269150021	4	8	6	4	1	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences,2. M...	Absence de permis (non obtenu / suspendu / perdu)	25-35
266100269103464	0	6	1	6	2	Mme	Aisne	Hauts-de-France	5. Administratifs / Judiciaires/ Financiers,6....	Accès aux Droits	>45

FIGURE 4.8 – Résultats du clustering pour les données concernant le département de l'Aisne (15 individus) avec 'Huang'

Avec cette seconde méthode d'initialisation des centres de clusters, nous constatons que les regroupements d'individus restent globalement similaires, bien que parfois plus cohérents.
Cependant, la variabilité de la répartition des clusters d'une itération à l'autre pose problème, car elle remet en question la fiabilité des conseils donnés pour l'accompagnement d'un individu, ces conseils pouvant varier d'une itération à l'autre si l'on se base sur cela.

Enfin voici nos résultats pour l'initialisation avec 'Cao' :

cle	Cluster	Cluster	Cluster	Cluster	Cluster	civilite	nom_dep	nom_region	cat_difficulte	libelle	tranche_age
100080269140238	2	2	2	2	2	M	Aisne	Hauts-de-France	1. Logement,2. Mobilité	Difficultés de maintien dans le logement	18-25
161019938301862	6	6	6	6	6	M	Aisne	Hauts-de-France	5. Administratifs / Judiciaires/ Financiers	Accès aux Droits	>45
166060236115334	8	8	8	8	8	M	Aisne	Hauts-de-France	4. Santé	Problématique physique	>45
167030269111368	0	0	0	0	0	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Difficulté avec les techniques de Recherche d'...	>45
171069406909809 01/05/2023	0	0	0	0	0	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Méconnaissance du marché du travail	>45
173109920603957	5	5	5	5	5	M	Aisne	Hauts-de-France	2. Mobilité	S Mobilité : absence de permis (non obtenu, su...	>45
178120240804187 22/03/2023	7	7	7	7	7	M	Aisne	Hauts-de-France	3. Problématiques sociales	Prise en charge d'une personne à charge	>45
188050269115893 23/09/2022	3	3	3	3	3	M	Aisne	Hauts-de-France	1. Logement	Sans domicile fixe	35-45
190099505201679	9	9	9	9	9	M	Aisne	Hauts-de-France	3. Problématiques sociales	Pas de mode de garde adapté	25-35
191020236121640 15/05/2023	1	1	1	1	1	M	Aisne	Hauts-de-France	2. Mobilité	Absence de permis (non obtenu / suspendu / perdu)	25-35
194065145461504 12/10/2022	10	10	10	10	10	M	Aisne	Hauts-de-France	1. Logement,2. Mobilité	Logement inadapté/vétuste/insalubre	25-35
194082737523143 12/10/2022	1	1	1	1	1	M	Aisne	Hauts-de-France	2. Mobilité	Absence de permis (non obtenu / suspendu / perdu)	25-35
198030269133078	0	0	0	0	0	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences	Difficulté avec les techniques de Recherche d'...	25-35
198040269150021	1	1	1	1	1	M	Aisne	Hauts-de-France	6. Adaptation / Evolution des compétences,2. M...	Absence de permis (non obtenu / suspendu / perdu)	25-35
266100269103464	4	4	4	4	4	Mme	Aisne	Hauts-de-France	5. Administratifs / Judiciaires/ Financiers,6....	Accès aux Droits	>45

FIGURE 4.9 – Résultats du clustering pour les données concernant le département de l'Aisne (15 individus) avec 'Cao'

Sur des itérations indépendantes, l'algorithme a, cette fois, fourni des prédictions constantes (même au niveau des labels même, si ceci n'est pas important) et cohérentes. C'est donc la méthode d'initialisation que nous avons sélectionné pour la suite.

En ce qui concerne la fiabilité de notre modélisation, notons que nous ne pouvons évaluer les résultats de notre algorithme que visuellement, en regardant la cohérence des clusters associés. Pour cette raison, mais également dans l'optique que la méthode du coude ne s'exécute pas en un temps raisonnable et que nous n'avons pas réussi à mettre en place le critère de la silhouette (car les données sont non numériques). Cela nous a amené à penser que nous devions faire un compromis sur le choix du nombre de clusters afin que ce nombre puisse toujours être choisi de manière optimale et pertinente indépendamment soit le nombre d'individus à différencier.

Nous avons donc choisi de sélectionner comme nombre de clusters quelque chose de fixe et de facilement calculable : le nombre de lignes identiques (sans tenir compte de la clé).

Pour reprendre notre exemple de l'Aisne, cela nous ferait considérer 14 clusters différents au lieu de 11 ou 12 selon les initialisations choisies précédemment par la méthode du coude. L'écart n'est pas très grand, nous pouvons toutefois nous imaginer qu'il serait d'autant plus grand selon le nombre de données considérées, mais cela nous assure ainsi de regrouper ensemble des personnes au même profil avec les mêmes difficultés et d'avoir des clusters fiables et invariants. Nous verrons d'ailleurs par la suite comment notre algorithme final nous permet de pallier ceci.

De cette façon, nous avons appliqué ce modèle ainsi paramétré à nos 3979 individus avec difficultés diagnostiquées et créés 1915 profils de personnes à accompagner différents.
En pratique, nous préférons travailler avec de plus petits nombres de clusters, car cela garantit leur interprétabilité. Or, ce nombre de 1915 profils est toutefois très restreint par rapport aux 16080 que nous pouvions établir. De plus, en ce qui concerne l'interprétabilité, nous savons ici que chaque cluster correspond à un profil, c'est-à-dire à des caractéristiques et difficultés précises. Nous sommes assurés de ne pas regrouper des personnes, dont les difficultés diffèrent et de prendre en compte toutes les caractéristiques qui leur sont propres en plus de cela.

4.2.2.2 Recherche des actions correspondantes

Une fois ces profils établis, nous avons cherché à identifier les actions associées à chacun des clusters.
Nous avions envisagé, pour commencer, de considérer des séquences d'actions. C'est-à-dire de considérer que l'enchaînement des actions avait une importance dans l'accompagnement. Dans cette optique, nous avions extrait de la table `action_export` ces différentes actions pour les considérer dans l'ordre chronologique, avec la clé de l'individu correspondant et sa sortie. Cependant, nous nous sommes aperçus que ces séquences revenaient rarement entre les différents individus d'un même cluster et qu'il serait donc compliqué, à partir de cela, de prédire les actions à mettre en place pour un nouvel individu.

Nous avons alors choisi de considérer les différentes actions de manière indépendante, en espérant ensuite pouvoir repérer celles qui reviennent le plus souvent dans les cas où la sortie est positive. Pour cela, nous avons extrait de la table `action_export` ces différentes actions en conservant une ligne par action cette fois, avec la clé de l'individu correspondant et sa sortie.

Pour faire le lien entre les clusters et les actions, nous avons fusionné une table contenant les différents clusters et leurs caractéristiques propres, ainsi que la table d'actions créée précédemment. Ainsi, il est facilement possible de la filtrer par rapport à la sortie et de sommer les actions mises en place pour chaque cluster. De cette façon, nous savons maintenant sur quoi nous baser pour nos prédictions.

4.2.2.3 Prédiction

L'idée est donc de pouvoir prédire pour de nouveaux individus, les actions à mettre en place à partir de ce qui a déjà été mise en place pour les autres.

Étant donné que nos individus d'intérêt sont étiquetés vis-à-vis des clusters, nous avons pensé prédire le cluster d'appartenance des nouveaux individus à partir de leurs caractéristiques renseignées lors du diagnostic d'entrée en utilisant à nouveau l'algorithme *Random Forest*.
Afin de garantir des résultats pertinents, il est nécessaire de considérer que des clusters contenant plusieurs individus pour que le modèle puisse les observer à la fois dans une démarche d'apprentissage et de test. Le modèle atteint alors une précision parfaite.
Cependant, le problème avec ce modèle va être sa façon de traiter un nouvel individu dont aucun profil similaire n'a été observé en phase d'apprentissage, il sera alors classé par défaut dans un autre cluster même s'il n'en a pas forcément les caractéristiques.

Nous avons pensé à ajouter un cluster 'Autre', mais cela aurait été délicat, en plus du fait que si ceci était réalisable, ce serait rapidement devenue une classe majoritaire qui aurait biaisé le modèle.

Nous avons réalisé que le problème serait le même pour un autre modèle statistique tel que la régression logistique, nous sommes donc revenus à une idée de modélisation plus rudimentaire mais efficace : rechercher directement les individus ayant le même profil pour y associer les actions relatives en utilisant le système de la partie précédente.

Pour ce faire, nous avons créé une fonction qui prend en paramètre les caractéristiques de l'individu relatives à sa civilité, sa tranche d'âge, sa région d'origine, sa liste de difficultés et de libellés de difficultés et qui, à partir de cela, va implicitement chercher le cluster correspondant. À noter que cette façon d'identifier les clusters d'appartenance des individus est ici valable et efficace (avec 100% de précision) car nous sommes sûrs que les caractéristiques fournies ne peuvent correspondre qu'à un seul et unique cluster. Par leur définition, il n'y a pas de part d'aléatoire dans cette association. L'avantage est que si un individu n'y appartient pas, il n'y sera pas associé à un autre qui ne correspond pas par défaut.

Le cluster bien identifié, nous retournons alors le nombre d'individus le composant, le nombre d'individus le composant ayant réussi leur réinsertion, le nombre d'actions nécessaires en moyenne, ainsi que les actions conseillées, de la plus fréquente à celle qui revient le moins souvent.
Nous avons ajouté le taux de réussite associé à chaque action, c'est-à-dire le nombre de fois où une action donnée revient pour les individus du cluster qui ont une sortie positive divisée par le nombre de fois où cette action revient pour tous les individus du cluster. Par ailleurs, cette démarche aurait été la même en identifiant les clusters par un modèle statistique.

Afin que la façon de rentrer les informations ne soit pas trop rigide pour les utilisateurs, nous autorisons les permutations dans les listes de difficultés et de libellés associés. Nous vérifions toutefois que les différentes caractéristiques rentrées sont bien valides.

Nous avions évoqué plus haut que le peu d'informations que nous avions à disposition pour certains clusters pouvaient être assez limitant pour décider des actions à mettre en place pour un nouvel individu. Afin de contourner ceci, nous avons créé une variante de notre première fonction qui permet de choisir les caractéristiques à passer en paramètre en conservant au minimum une liste de difficultés à fournir. Les informations retournées sont les mêmes, mais cela permet de prendre plus de recul sur l'accompagnement possible en regardant des profils qui diffèrent.

Enfin, ce modèle mis en place ne se basait pas sur le traitement des informations non structurées mis en place dans le *Chapitre* 3. Nous avons donc cherché à intégrer la polarité des actions issues du traitement des colonnes $objectif$, *resultat* et *commentaire* de la table *action* pour affiner les actions proposées en s'assurant de ne garder que les plus pertinentes. Nous pouvons donc envisager d'écarter celles dont le résultat est considéré comme négatif. En effet, nous pouvons supposer qu'une action dont le résultat est jugé ainsi ne sera pas explicative d'une sortie positive même. Pour cela, il nous a suffi de développer une troisième version en ajoutant un filtre lors de la comptabilisation des actions et du calcul du taux d'efficacité.

4.2.3 Résultats

La modélisation mise en place, nous allons maintenant illustrer le fonctionnement et les résultats de notre algorithme de prédiction à travers quelques exemples concrets.

Pour commencer, imaginons le cas d'une femme de plus de 45 ans, résidant dans les Hauts-de-France et rencontrant des problèmes de mobilité, notamment du fait qu'elle n'a pas de permis de conduire (non obtenu/suspendu/perdu). Avant de commencer, nous avons tenté de prédire le statut de sortie pour cette personne accompagnée en utilisant un modèle de prédiction similaire à celui décrit dans la section 4.1.4.2. Ce modèle a été entraîné sur l'ensemble des données à l'exception de l'individu en question, dans le but de prédire son statut de sortie. Nous avons regroupé les différents statuts en deux catégories positives et négatives. L'algorithme a anticipé un résultat négatif pour cet individu, une prédiction logique, car nous allons voir que les profils similaires ont majoritairement abouti à des issues négatives.

L'algorithme développé dans cette partie nous retourne, quant à lui, que
"3 individus présentent les mêmes caractéristiques dans notre base de données. Parmi eux, 1 ont une sortie positive. À partir de ceci, voici les actions conseillées pour accompagner les personnes ayant ce profil :

— **Entretien et bilan régulier** : est revenu 8 fois (taux de succès : 0.73)

— **Définition d'un projet professionnel** : est revenu 1 fois (taux de succès : 1.00)

En moyenne, 9.0 actions fructueuses par individu ont été mises en place pour mener à une sortie positive.".

Face à cela, il nous semble clair que nous avons tout intérêt à proposer à la personne à accompagner de définir un projet professionnel ainsi que des entretiens et bilans réguliers.

Nous pourrions alors nous faire la remarque, que baser notre décision seulement sur le retour obtenu à propos du parcours fructueux d'un seul individu n'était pas forcément très pertinent. Nous pouvons alors choisir de ne pas tenir compte de la civilité pour avoir plus de recul par exemple, nous obtenons alors
"11 individus présentent les mêmes caractéristiques dans notre base de données. Parmi eux, 3 ont une sortie positive.
À partir de ceci, voici les actions conseillées pour accompagner les personnes ayant ce profil :

— **Entretien et bilan régulier** : est revenu 12 fois (taux de succès : 0.31)
— **Passerelle vers les entreprises classique** : est revenu 6 fois (taux de succès : 0.50)
— **Diagnostic socio-professionnel** : est revenu 1 fois (taux de succès : 0.20)
— **Solution aux problématiques sociales** : est revenu 1 fois (taux de succès : 0.14)
— **Suivi parcours : Entretien de renouvellement** : est revenu 1 fois (taux de succès : 1.00)
— **Suivi social** : est revenu 1 fois (taux de succès : 0.11)
— **Technique de recherche d'emploi** : est revenu 1 fois (taux de succès : 0.50)
— **Définition d'un projet professionnel** : est revenu 1 fois (taux de succès : 0.50)

En moyenne, 8.0 actions fructueuses par individu ont été mises en place pour mener à une sortie positive.". Cette fois, plus d'actions sont proposées. Celles retenues précédemment ne semblent pas forcément les plus indiquées ici. Nous pourrions donc également proposer un suivi de parcours à notre cliente ainsi que de quoi l'aider à rechercher un emploi ou encore à effectuer une passerelle vers une entreprise classique.

Supposons maintenant que l'on diagnostique également à cette personne un problème lié au logement, l'algorithme renvoie "Aucun profil correspondant aux critères de recherche n'a été trouvé.". Notre base de données ne semble pas assez fournie pour avoir de quoi se baser pour l'accompagnement. Il faudrait procéder comme l'entreprise le fait habituellement de manière classique.

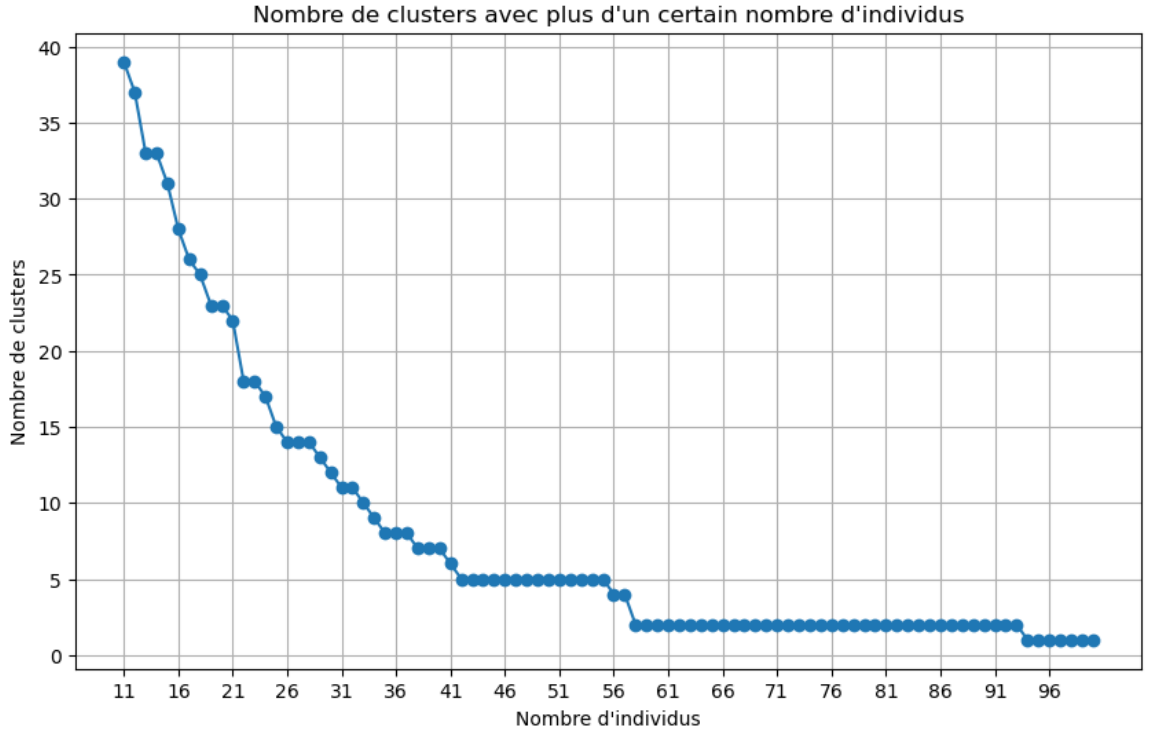
Enfin, nous pouvons imaginer que peut-être d'autres personnes ont également un problème de logement, mais pas le même libellé. Si nous retirons les libellés de difficultés dans notre recherche, nous obtenons "Il n'y a pas d'individus avec ce profil qui ont eu une sortie positive à ce jour.". Ainsi, dans ce cas de figure, la conclusion est la même que celle faite ci-dessus. Cela montre d'ailleurs que de tels algorithmes ne remplaceront jamais le travail des accompagnateurs.

4.2.4 Limites et améliorations

Nous disposons donc actuellement d'un modèle capable de suggérer des actions à partir des caractéristiques propres aux individus et à leurs difficultés en se basant sur les clusters identifiés à partir de la table `difficultes_export` et les actions de la table `action_export`. Toutefois, plusieurs pistes pourraient être explorées pour augmenter la pertinence et la maniabilité de notre modèle.

Premièrement, comme nous avons pu l'évoquer, les clusters que nous avons contiennent peu d'individus en général. Il y a donc assez peu de recul sur les actions proposées au final.

Nous avions pensé ne sélectionner que les clusters pour lesquels nous disposions d'un certain nombre d'individus au minimum afin de n'avoir que de véritables "profils type" sur lequel nous avons suffisamment de matière pour en tirer un apprentissage. Nous avions donc tracé la courbe du nombre de clusters selon le nombre de personnes minimums regroupées pour choisir un seuil :



Nous hésitions alors entre prendre des valeurs telles que 10, 15, 20 ou même des valeurs beaucoup plus élevées comme 50 ou 100. Idéalement, il aurait été intéressant d'avoir un retour métier pour éclaircir ce point. Dans le doute, nous avions mis de côté ce point pour construire notre modèle. Nous nous sommes toutefois rendu compte en le testant que nous disposions déjà de peu d'informations sur les sorties positives, même lorsque le cluster était grand.

Ce serait donc un point essentiel pour renforcer la pertinence de notre modèle et donner plus de sens aux comptages d'actions ainsi qu'aux taux retournés. Cependant, ce n'est pas quelque chose dont on peut se permettre ici, car cela nécessiterait d'avoir une base de données plus fournie surtout en ce qui concerne des retours positifs. Une façon éventuelle d'avoir plus d'étiquettes de sortie serait de faire une analyse de survie sur les données censurées. Nous pourrions peut-être prendre en compte les données temporelles et regarder si l'individu est rentré et a eu des actions mises en place il y a moins de deux ans, temps de suivi par l'entreprise. Le cas échéant, nous pourrions supposer que cet individu n'a plus donné de nouvelles et que sa sortie est donc négative. Ceci serait cependant laborieux, de plus, nous n'aurions pas de certitude que la sortie d'un tel individu soit réellement négative et d'un autre côté, cela ne nous donnerait pas plus de contenu sur les sorties positives comme nous le souhaiterions.

En parlant de précision, nous devons rester critique sur l'incorporation des données concernant la polarité des actions à notre modèle. Nous n'avons pas trouvé d'autres idées pour l'incorporer que d'écarter les actions jugées négatives. Toutefois, il est important de prendre en considération que la négativité est souvent basée sur des mots tels que "retard", or un retard n'est pas forcément significatif d'une action qui n'a pas été utile.

Toujours concernant la précision de notre modèle, une autre amélioration envisageable concerne les libellés des difficultés. Nous nous sommes aperçus, assez tard, que deux libellés exprimés différemment pouvaient exprimer le même problème. Par exemple : "S Administratif/ judiciaire/ financier : Problèmes financiers/" et "S Administratif/ judiciaire/ financier : Problèmes financiers" à un caractère près, mais cela pourrait être également associé à "Problèmes financiers/ Dettes et gestion difficile des ressources".
Une idée serait donc d'ajouter une étape de pré-traitement pour regrouper les champs proches. À la main, ceci serait faisable, mais serait assez long.

Peut-être que ceci permettrait d'avoir des clusters globalement un peu plus gros en regroupant des profils identiques qui n'étaient pas identifiés comme similaires jusqu'à présent, mais aussi, de simplifier l'utilisation de notre algorithme de recherche d'action.
Malgré la non-obligation de rentrer tous les champs, celui-ci est totalement rigide en ce qui concerne la rentrée d'information. Nous pouvons imaginer que cela ne le rend pas forcément très pratique pour un utilisateur. Autant nous avons fait en sorte de signaler une erreur pour un champ mal rentré et de rappeler ceux autorisés, autant lister tous les libellés possibles ne donne pas quelque chose de facilement lisible et qui permette à l'utilisateur de corriger facilement sa requête.

En outre, il pourrait être pertinent d'affiner la définition de positivité et de négativité du statut de sortie, comme abordé précédemment dans la section 3.2.

Pour finir, comme évoqué dans l'introduction de la partie, nous n'avons pas tenu compte ici des informations relatives à la formulation d'un projet professionnel ou du suivi d'une formation, car nous nous étions concentrés sur le fait d'établir une relation entre un diagnostic d'entrée et les actions à mettre en place pour une sortie réussie. Toutefois, nous avions vu que ces informations pouvaient être clés pour caractériser une sortie positive. Nous pourrions donc envisager d'ajouter, à la sortie de notre fonction, les formations éventuellement suivies par les personnes ayant le même profil si cela contribue à favoriser une sortie positive ainsi que les projets professionnels/les métiers de sortie qui reviennent le plus pour ces individus. Il nous suffirait d'aller chercher ces caractéristiques comme nous sommes allés chercher les actions.

Une fois de plus, il est important de souligner qu'un tel modèle ne remplacera jamais l'expertise d'un expert et qu'il n'a pour objectif que l'aide à la prise de décision.

En conclusion, ce chapitre de modélisation a mis en lumière une approche pour comprendre et anticiper les tendances et comportements des personnes accompagnées, à travers la création de modèles prédictifs basés sur la forêt aléatoire et l'exploitation de techniques de clustering adaptées aux données catégorielles, comme l'algorithme des K-modes.

L'identification de profils type et la prédiction des actions à entreprendre selon ces profils représentent des avancées qui pourraient permettre l'optimisation de l'accompagnement proposé. Cependant, la précision et la pertinence des modèles, bien que prometteuses, soulignent la nécessité du jugement d'expert pour l'accompagnement. Les limites identifiées ouvrent la voie à des améliorations futures, notamment en ce qui concerne le regroupement des libellés de difficultés pour affiner les clusters et les prédictions, l'incorporation plus nuancée de la polarité des actions, et l'enrichissement des recommandations par l'ajout d'informations sur les formations suivies et les projets professionnels envisagés.

Ce travail a donc mis en évidence l'importance d'une collaboration étroite entre les données disponibles et l'expertise humaine pour développer des outils d'aide à la décision qui, tout en étant basés sur des analyses de données approfondies, restent sensibles aux réalités individuelles de chaque personne accompagnée.

Conclusion

Dans le cadre de notre projet, nous avons entrepris une analyse approfondie des liens existants entre les différentes tables de données fournies. Cette démarche nous a permis de mettre en lumière l'importance de chaque étape de l'accompagnement dans la résolution de la problématique posée par l'entreprise. En identifiant les variables influentes pour notre modèle, nous avons constaté qu'elles étaient cohérentes avec les attentes concernant les facteurs déterminants du statut de sortie. Notamment, les actions régulières telles que les entretiens et les bilans réguliers se sont avérées être parmi les variables les plus significatives, soulignant ainsi l'importance et l'efficacité d'un suivi continu.

Le traitement du texte libre a été particulièrement enrichissant, car nous n'avions pas l'habitude de travailler avec ce type de données. Cette technique, qui pourrait d'ailleurs être étendue à l'ensemble des tables, a renforcé nos algorithmes en extrayant des informations précieuses jusqu'alors inexploitables.

Nos différentes approches de modélisation, basées sur la création de modèles prédictifs via les forêts aléatoires et l'utilisation d'algorithmes de clustering adaptés aux données catégorielles (l'algorithme K-modes), nous ont permis d'anticiper avec une certaine précision les comportements et orientations des participants. Ce travail de modélisation, loin d'être une simple tâche académique évidente, s'est révélé être une mission passionnante, offrant la perspective d'appliquer nos découvertes à des situations réelles et d'avoir un impact significatif en aidant les individus à retrouver des opportunités d'emploi.
L'intérêt que nous avons ressenti tout au long de ce projet témoigne de la satisfaction que nous avons tirée de notre travail.

Ainsi, ce projet n'a pas seulement été une occasion d'appliquer des connaissances théoriques à des données réelles pour un travail d'entreprise. Il a également été une opportunité d'apprentissage constant.
D'un côté, il est d'ailleurs un peu frustrant d'arrêter un tel projet ainsi, sans avoir pu l'améliorer davantage ou surmonter les obstacles identifiés. Nous pourrions, par exemple, envisager son extension par l'ajout de plus de données ou encore l'utilisation d'avancées en intelligence artificielle et traitement du langage pour obtenir des résultats encore plus performants. De plus, nous avions pensé à la possibilité de transformer cette étude en une application concrète ou en un tableau de bord interactif, comme une page Power BI afin de faciliter son utilisation aux professionnels.

Ce travail a été une excellente occasion d'appliquer nos connaissances et de conclure notre semestre, nous préparant ainsi à aborder plus sereinement les projets futurs en entreprise.

Annexe 1 : Explication de l'algorithme de *Random Forest*

Nous allons tenter d'expliquer de manière simple les algorithmes utilisés dans ce rapport. Afin de faciliter la compréhension intuitive des principes. Pour comprendre le principe de la *Random Forest*, nous devons d'abord comprendre les arbres de décision, car cet algorithme utilise une collection d'arbres de décision. La construction d'un arbre de décision repose sur le partitionnement successif de l'espace des données. À chaque nœud (division de l'arbre), l'algorithme de classification sélectionne la caractéristique et le seuil qui divisent au mieux les données selon un critère (par exemple, l'indice de Gini ou l'entropie, que nous ne détaillerons pas ici) pour séparer les différentes classes. Étudions un arbre de décision sur un exemple fictif :

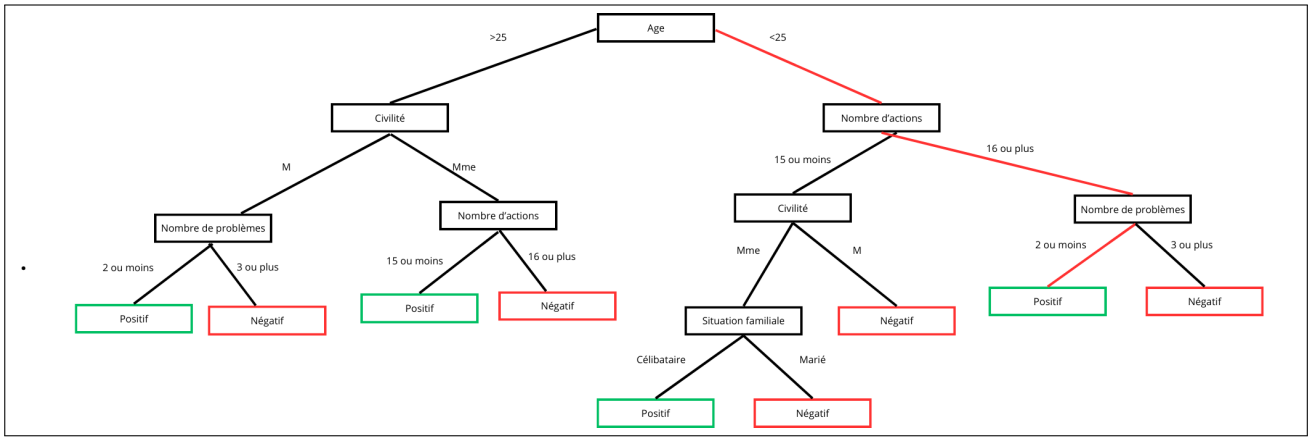


FIGURE 4.10 – Exemple d'arbre de décision

Nous avons une première division basée sur la variable âge, qui vient scinder nos données en deux groupes : les individus de moins de 25 ans et ceux de plus de 25 ans. Ensuite, nous avons d'autres divisions jusqu'à obtenir les nœuds terminaux (feuilles de l'arbre), qui représentent la prédiction finale. La construction du nœud se fait par l'algorithme qui évalue toutes les caractéristiques disponibles et leurs valeurs possibles pour déterminer le meilleur point de division (selon un critère comme l'indice de Gini ou l'entropie). Une fois l'arbre construit, il est facile de prédire le label d'un nouvel individu. Prenons, par exemple, un homme célibataire de moins de 25 ans ayant réalisé plus de 20 actions et rencontré 1 problème. Il suffit donc de suivre le bon chemin de l'arbre pour prédire le label, cela correspond au chemin indiqué en rouge sur le schéma.

Après avoir compris le fonctionnement des arbres de décision, nous pouvons comprendre la *Random Forest*. Cette méthode utilise plusieurs arbres de décision, chacun construit à partir d'un sous-ensemble aléatoire des données et des caractéristiques. C'est l'ensemble de ces arbres qui constitue la "forêt". Lors de la prédiction pour un nouvel individu, chaque arbre de la forêt vote, et la prédiction finale est déterminée par un vote majoritaire. Cette approche d'ensemble améliore la robustesse et la précision du modèle, rendant l'algorithme de *Random Forest* particulièrement efficace pour traiter des ensembles de données complexes avec un mélange de caractéristiques numériques et catégorielles.

Annexe 2 : Explication de l'algorithme des K-means

L'algorithme des K-means est une méthode de clustering non supervisée qui partitionne un ensemble de données en K clusters distincts. Chaque cluster regroupe des individus présentant des caractéristiques communes ou étant proches selon une certaine mesure de distance, cette dernière nécessitant une définition adaptée pour les variables catégorielles (d'où l'utilisation des Kmodes).

En pratique, l'algorithme attribue chaque donnée au centroïde le plus proche, formant ainsi des clusters. Le centroïde, à définir, représente le centre d'un cluster et est déterminé en calculant la moyenne des points qui lui sont associés.

Après avoir initialisé les centroïdes de manière aléatoire dans le jeu de données, K-means procède par une alternance de deux étapes pour affiner la position des centroïdes et la composition des clusters :

— L'affectation de chaque objet au centroïde le plus proche.

— L'actualisation de la position de chaque centroïde pour qu'il corresponde à la moyenne des points de son cluster.

Après plusieurs itérations, l'algorithme parvient à un partitionnement stable des données, signifiant sa convergence.

Regardons un exemple en 2 dimensions :

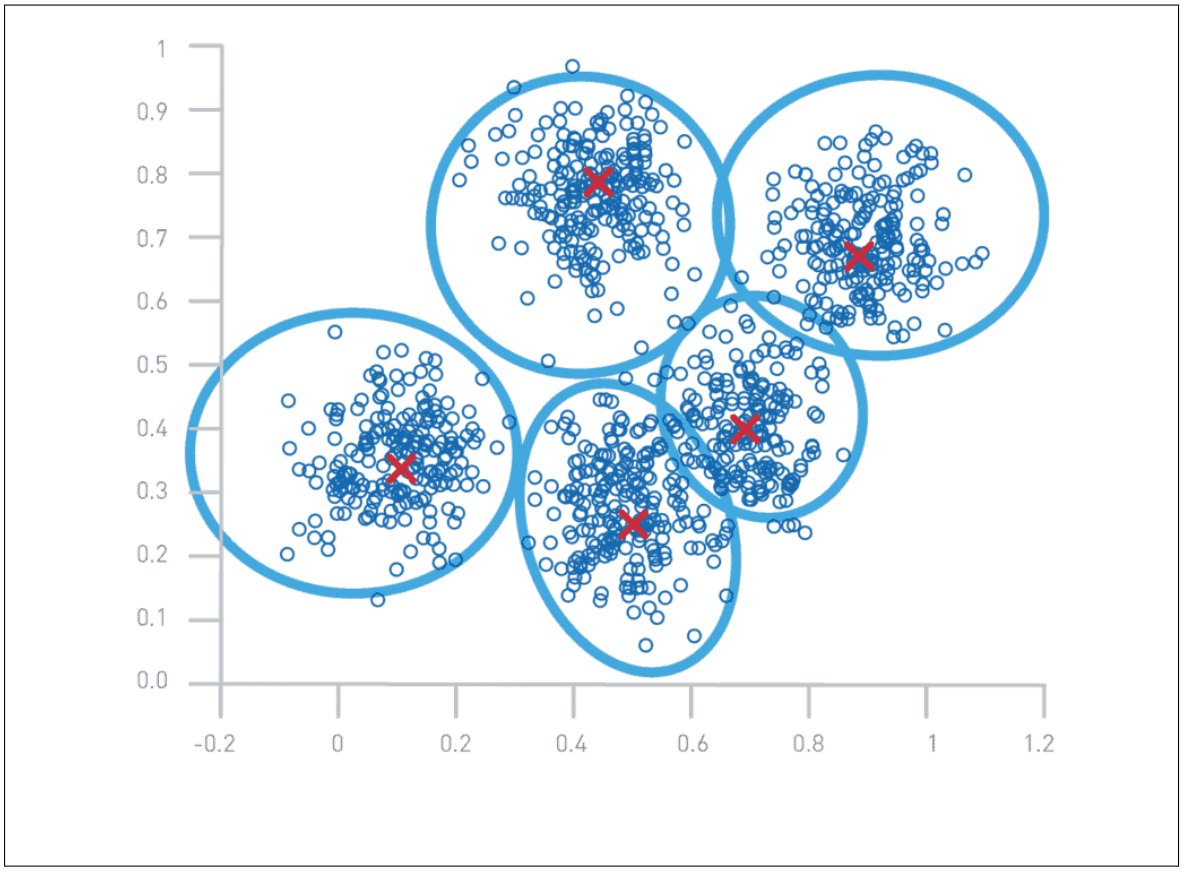


FIGURE 4.11 – Exemple de cluster après application des K means.
Source : https ://www.data-transitionnumerique.com/k-means/

Cette illustration montre les résultats de l'algorithme avec 5 clusters identifiés, où les centroïdes de chaque cluster sont marqués en rouge.

Dans notre cas, bien que l'algorithme fonctionne de la même manière, nous ne pouvons pas visualiser ces clusters de façon aussi directe, car le nombre de dimensions (qui correspond au nombre de variables) est beaucoup plus élevé.