UNIVERSITÉ
CÔTE D'AZUR

BUT SD - $3^{ème}$ année

Rapport projet Data Mining

Analyse et classification de l'absenteisme au travail.

Algassimou Diallo

Décembre 2023

SAE: Mise en oeuvre d'un processus de Datamining
Enseignant : Celia Da Costa Pereira

Contents

1 Introduction

L'absentéisme au travail se profile comme un défi majeur dans le contexte professionnel contemporain, engendrant des répercussions significatives sur la productivité, la satisfaction des employés, et leur bien-être. La compréhension approfondie des motifs sous-jacents à ces absences ainsi que le développement de stratégies préventives sont devenus cruciaux pour les responsables des ressources humaines. Dans cette optique, notre projet de data mining s'engage à explorer et à analyser le jeu de données "Absenteeism at work", recensant des informations sur l'absentéisme au travail au sein d'une entreprise de messagerie au Brésil entre juillet 2007 et juillet 2010.

L'objectif central de cette étude est d'appliquer des techniques avancées de data mining pour élaborer des modèles prédictifs capables d'identifier les facteurs déterminants de l'absentéisme professionnel. À travers l'utilisation d'algorithmes de classification, notre démarche cherche à anticiper et à classifier les absences potentielles des individus, dotant ainsi les gestionnaires d'outils pertinents pour prendre des mesures proactives et formuler des politiques adaptées.

Ce rapport détaille de manière exhaustive le déroulement de notre projet, depuis le prétraitement des données et l'analyse exploratoire des données jusqu'à l'implémentation des modèles de machine learning et à l'analyse approfondie des résultats obtenus.

2 Prétraitement des données

Dans cette première étape cruciale de notre projet, nous avons entrepris le prétraitement des données afin de garantir la qualité et la cohérence des informations manipulées. Le jeu de données initial s'est avéré relativement propre, sans présence de valeurs manquantes (NA), et les types de données étaient corrects. Cependant, une observation particulière a attiré notre attention concernant les variables "Reason for absence" et "Month of absence".

Une exploration plus approfondie a révélé des valeurs minimales égales à zéro pour ces deux variables, ce qui semblait incohérent par rapport à la documentation du dataset. Pour la variable "Reason for absence", nous avons identifié que ces valeurs nulles correspondent en réalité à une situation de "non absence" ou que tout simplement l'absence a duré moins d'une heure(juste un petit retard) car la variable "Absenteeism time" est en heure. Cette constatation est importante, car ces cas ne constituent pas des erreurs, et nous avons donc décidé de conserver ces lignes, car elles fournissent des informations pertinentes sur l'absence effective.

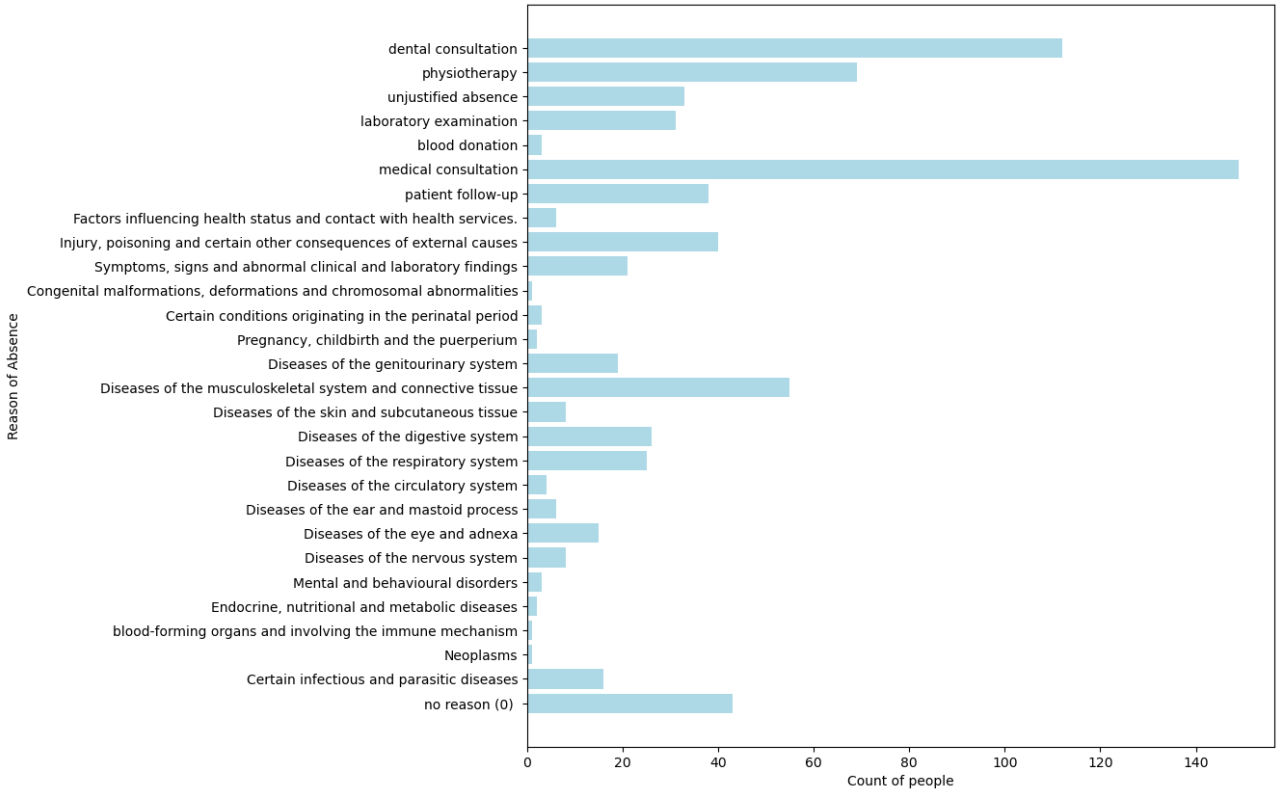
En ce qui concerne la variable "Month of absence", nous avons noté que ces occurrences correspondaient également à des individus avec un temps d'absence nul, et leur nombre était juste de trois. Afin de maintenir la cohérence des données, nous avons opté pour le remplacement de ces valeurs par la moyenne des mois des individus précédents c'est à dire les individus où la raison d'absence est égale à zéro et le temps d'absence est également nul. Cette approche vise à garantir la précision des données tout en préservant la pertinence des informations fournies par le dataset.

Cette phase de prétraitement est essentielle pour assurer la robustesse de nos analyses ultérieures et la fiabilité de nos modèles de machine learning. Ces modifications ne représentent qu'une partie des ajustements que nous avons effectués. En poursuivant nos analyses, en particulier lors de l'exploration des données, d'autres problèmes ont émergé, et nous avons entrepris des démarches supplémentaires pour les résoudre. Ces démarches seront discutées en détail dans les parties à venir.

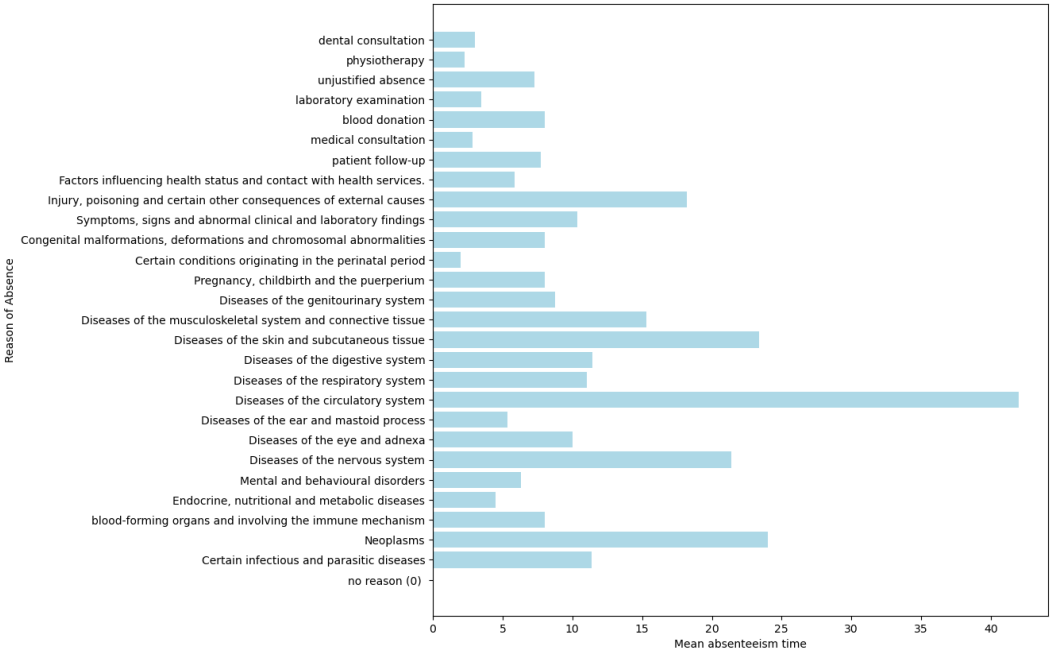
3 Analyse exploratoire des données(EDA)

Dans cette phase de notre étude, nous avons conduit une analyse exploratoire des données pour mieux appréhender les relations et les tendances inhérentes à la variable que nous cherchons à prédire, à savoir "Absenteeism time". Voici un aperçu des principales observations tirées de nos analyses :

"Reason of Absence" et "Absenteeism time: : En examinant la relation entre la raison d'absence et la durée d'absence, nous avons identifié que la raison la plus fréquente est la raison 23 correspondant à des "consultations médicales", suivie de près par les raisons 28 ("Consultation dentaire") et 13 ("Maladies du système musculo-squeletique et du tissu conjonctif").



Cependant, lorsque nous nous intéressons à la moyenne du temps d'absence, une observation notable émerge: la raison d'absence qui affiche la durée moyenne la plus longue est la raison 9(Diseas of the respiratory system).Toutefois, après une inspection approfondie, il s'est avéré qu'une donnée aberrante(outlier) biaisait ces résultats. En prenant en compte cet outlier, nous constatons que la vraie raison d'absence qui présente la durée moyenne la plus longue est la raison 1("néoplasmes"). Les néoplasmes, ou tumeurs, englobent une variété de maladies, dont certaines nécessitent des traitements intensifs et prolongés, impactant ainsi significativement la durée d'absence au travail. Par exemple, le traitement d'un cancer peut impliquer des sessions de chimiothérapie et de radiothérapie étalées sur plusieurs mois, nécessitant des périodes d'absence prolongées pour les employés. Cette rectification souligne l'importance de l'examen minutieux des données pour identifier et corriger tout effet indésirable d'outliers, garantissant ainsi la précision des analyses.

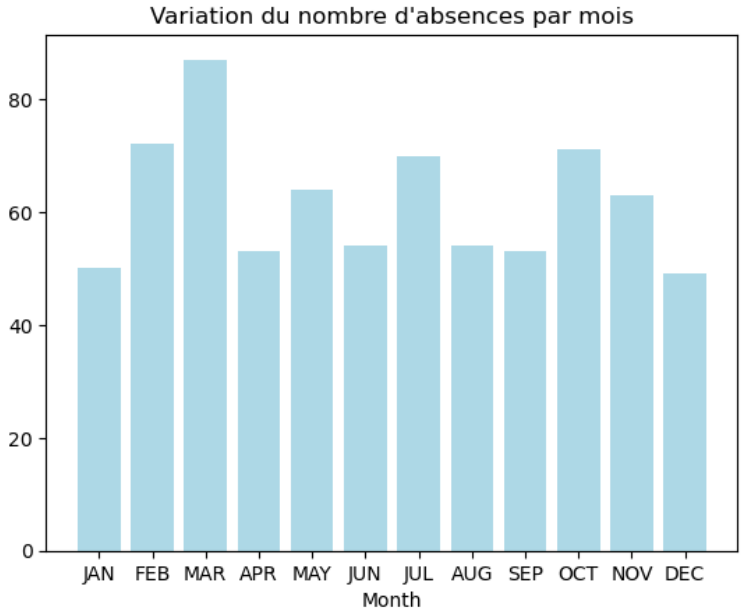


Moyenne d'absences par raisons

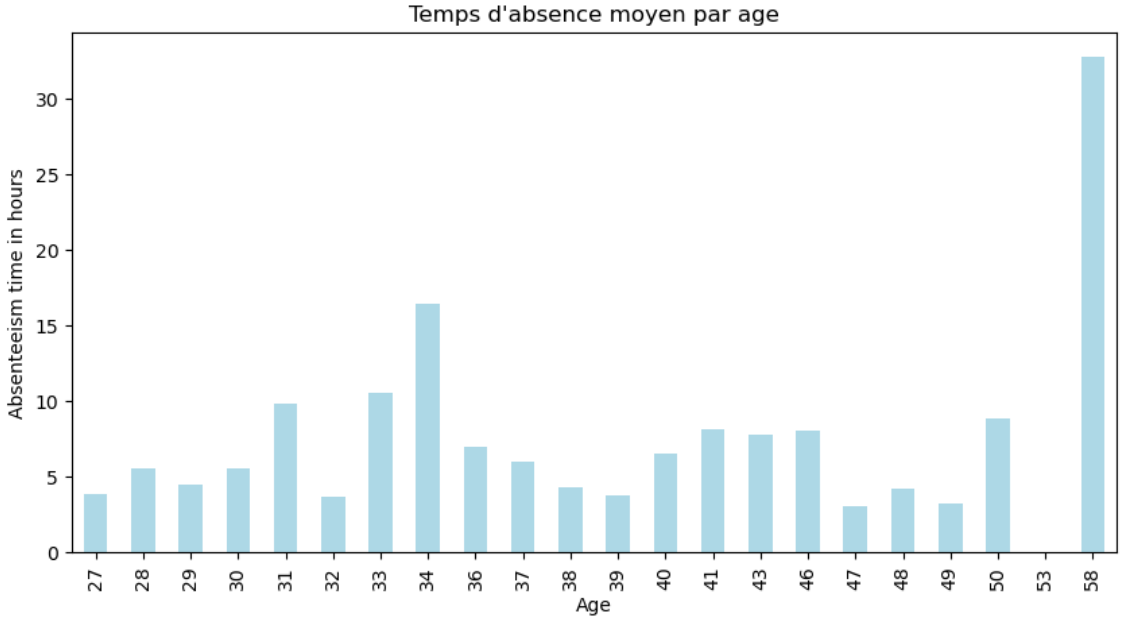
"Month of absence" et "Absenteeisme time": En analysant le mois d'absence, il est remarquable que le mois de Mars enregistre le plus grand nombre d'absences, suivi de Février et Octobre. Une explication plausible réside dans le contexte culturel brésilien, où le mois de Mars est souvent associé à divers festivals et célébrations nationales.

En particulier, le célèbre "Carnaval de Rio de Janeiro" se déroule généralement en février ou mars. Cette célébration annuelle attire des milliers de personnes du monde entier, créant une ambiance festive et entraînant parfois des absences temporaires au travail. Outre le Carnaval, d'autres festivités telles que la "Fête de l'Indépendance du Brésil" le 7 septembre et le "Jour de la République" le 15 novembre sont également des événements nationaux qui peuvent influencer les schémas d'absentéisme au cours de ces mois.

Ainsi, la présence marquée d'absences au mois de Mars peut être expliquée en partie par l'impact des célébrations culturelles et des festivités qui caractérisent cette période au Brésil.



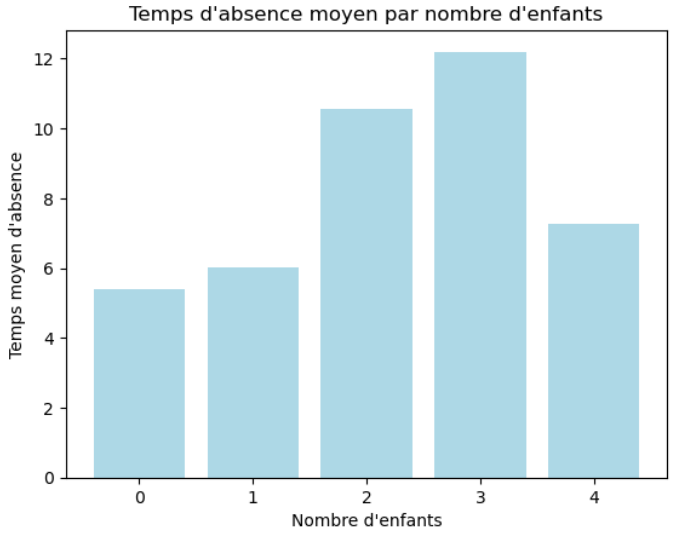
"Age" et "Absenteeisme time": L'âge des individus ne semble pas être un facteur déterminant dans le temps d'absence, à l'exception d'un outlier notable. Les individus dans la trentaine affichent cependant la durée moyenne d'absence la plus élevée.



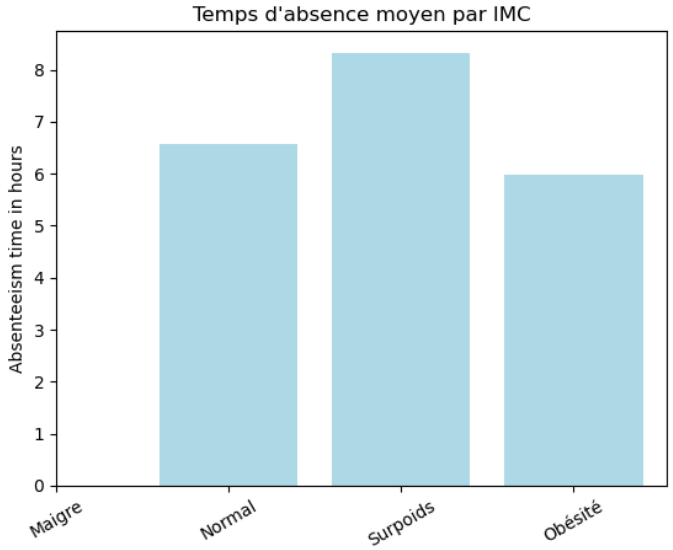
"Nombre d'enfants" et "Absenteeisme time": les individus ayant deux ou trois enfants présentent une durée moyenne d'absence plus longue. Cette constatation suggère que la charge familiale peut jouer un rôle significatif dans les périodes d'absence au travail.

Par exemple, un parent avec plusieurs enfants peut être confronté à des situations imprévues telles que des maladies familiales, des rendez-vous médicaux, ou d'autres responsabilités parentales, ce qui peut influencer la fréquence et la durée de leurs absences. Une autre observation intéressante ressort quand on regarde combien d'enfants les gens ont et combien de temps ils s'absentent. Contrairement à ce à quoi on pourrait s'attendre, ceux qui ont quatre enfants semblent s'absenter moins longtemps en moyenne. Cela pourrait s'expliquer par le fait que dans les familles nombreuses, les enfants peuvent se soutenir mutuellement.

Par exemple, dans une famille avec beaucoup d'enfants, ils peuvent partager les responsabilités et s'entraider. Cela pourrait signifier que les parents ont moins besoin de prendre du temps libre pour s'occuper des enfants, car les enfants peuvent jouer un rôle actif les uns envers les autres. Ainsi on peut dire que la taille de la famille peut avoir des effets différents sur le temps d'absence.

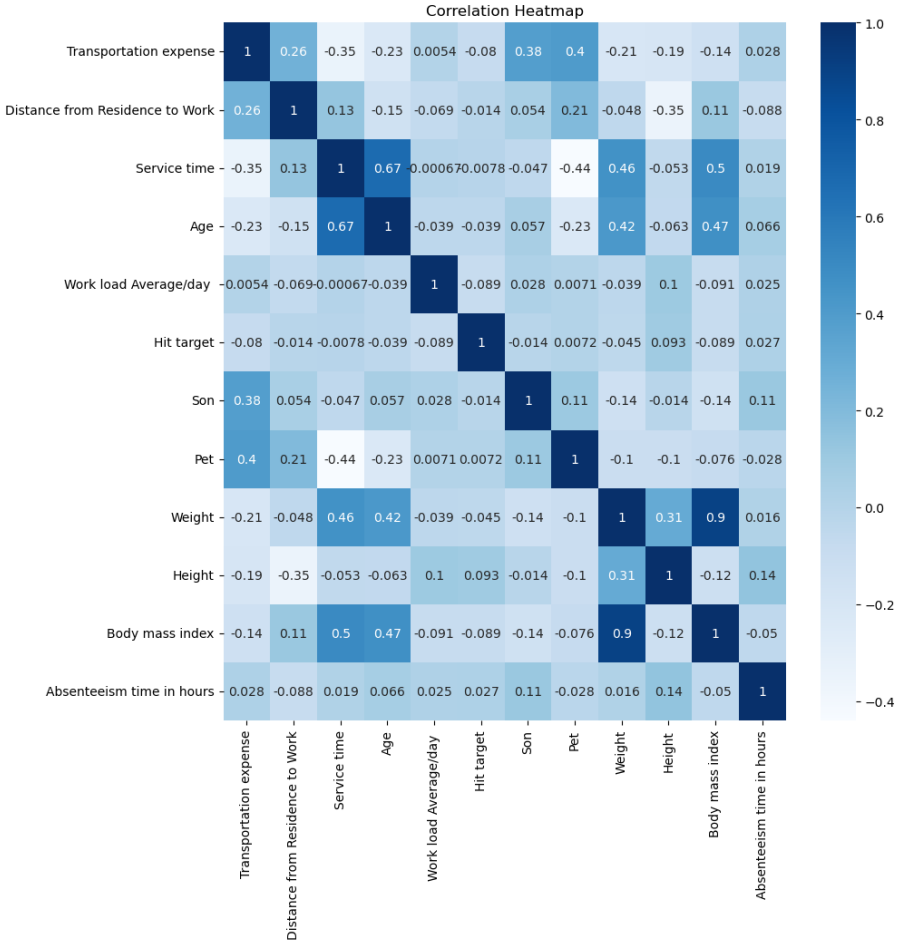


Les analyses des autres variables telles que le jour de la semaine, les habitudes de consommation de tabac et d'alcool et l'indice de masse corporelle (IMC) montrent que ces facteurs semblent avoir un impact relativement limité sur la durée d'absence. Par exemple, le jour de la semaine ne semble pas être un indicateur significatif, les habitudes de consommation de tabac et d'alcool montrent des variations mineures et l'IMC n'affiche pas de variations majeures malgré que les gens en "surpoids" aient le temps moyen d'absence le plus élevé.



Ces constatations suggèrent que, dans le contexte spécifique de notre étude, ces variables ne sont peut-être pas des facteurs prépondérants dans la détermination du temps d'absence au travail.

Pour renforcer cette idée, nous avons utilisé une matrice de corrélation. Les résultats de cette analyse confirment ce que nous avons observé auparavant. On constate aussi que a plupart des variables du dataset ont respectivement une corrélation très faible avec la durée d'absence au travail.



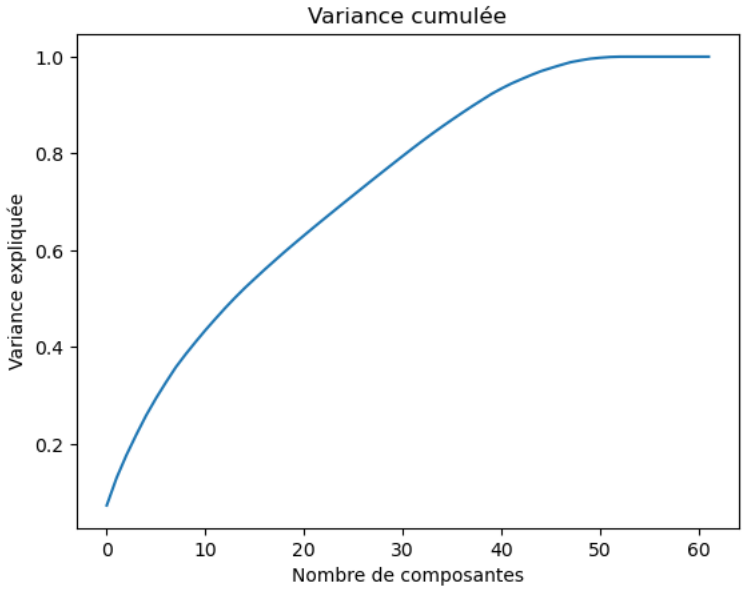
De plus, la matrice de corrélation met en évidence des relations entre certaines variables, telles que l'indice de masse corporelle (IMC) et la taille, ce qui est normal. Elle révèle également des corrélations entre des paires de variables telles que "service time" et "age", ainsi que "service time" et "IMC".Nous avons donc décider d'éliminer la variable "Height" et de garder tout simplement l'IMC et le poids des individus. De plus dans le jeu de données, la variable "reason of absence" est codifiée sous forme de chiffres de 1 à 28 pour représenter différentes raisons d'absence. Cependant, cette représentation numérique peut induire en erreur les modèles de machine learning en leur suggérant que certaines raisons sont plus impactantes que d'autres, simplement en raison de leur valeur numérique,ce qui fausserait l'analyse. Pour remédier à ce problème, nous avons choisi de créer de nouvelles colonnes en effectuant un "one-hot encoding" pour chaque raison, créant ainsi des colonnes binaires distinctes pour chaque raison d'absence. On fait de même pour les variables "Seasons", "Day of the week" et "Education"

Enfin, en raison de la faible corrélation observée entre les variables existantes et notre variable cible "absenteeism time", nous avons pris la décision de recourir à une Analyse en Composantes Principales (ACP). Cette approche se révèle pertinente dans notre contexte, car elle permet de réduire la dimensionnalité du jeu de données tout en préservant l'essentiel de l'information.

L'ACP est particulièrement adaptée lorsque les variables initiales ne montrent pas une corrélation significative avec la variable cible, car elle vise à identifier les combinaisons linéaires de variables qui maximisent la variance des données. En réduisant le nombre de dimensions, l'ACP facilite la détection de tendances et de motifs cachés qui pourraient ne pas être apparents dans l'ensemble initial de variables.

Ainsi, notre choix de l'ACP ne se limite pas seulement à surmonter le défi de la faible corrélation, mais aussi à améliorer la compréhension des structures sous-jacentes du jeu de données. Cela nous permettra d'obtenir des composantes plus significatives pour la modélisation ultérieure, renforçant ainsi la qualité et la pertinence de nos

analyses. Notre Analyse en Composantes Principales (ACP) a démontré une capacité impressionnante à expliquer plus de 90% de la variance présente dans le jeu de données avec un nombre assez important de composantes générées.



Nous avons opté pour une approche plus pragmatique en ne retenant que les 31 premières composantes principales, celles-ci expliquant 80 % de la variance totale.

Cette décision est motivée par le besoin de maintenir un équilibre entre la réduction de la dimensionnalité et la conservation d'une proportion significative de l'information originale. En limitant le nombre de composantes à 31, nous cherchons à simplifier la représentation du jeu de données tout en préservant une bonne explication de la variance des données.

En définitive, il est clair que la plupart des variables examinées présentent des corrélations très faibles, voire négligeables, avec la variable cible "Absenteeism time in hours". Ainsi, il apparaît que la création d'un modèle visant à prédire le **temps d'absence** en utilisant ces variables spécifiques ne serait pas très efficace même avec l'ACP, comme illustré par nos tests.En effet, en utilisant plusieurs modèles de classification**(Regression,Naive Bayes,Arbre de décision et random forest)**, le résultat le plus satisfaisant que nous avons obtenu est un model de random forest avec une accuracy de seulement 53%. Nous avons donc opter pour une approche alternative.

Plutôt que de chercher à prédire directement la durée d'absence, nous choisissons d'introduire une nouvelle variable avec 3 modalités qui classent de manière arbritaire les individus en fonction de leur durée d'absence. Ces modalités sont définies comme "retardataire", "absence normale", et "absentéisme problématique". Cette classification devrait permettre une analyse plus qualitative et nuancée de l'absentéisme, en mettant l'accent sur la gravité des absences plutôt que sur leur durée exacte.

4 Transformation des données

Mais comment parvenons-nous à classer les individus dans ces différentes catégories de niveaux d'absence ? Cette question constitue le point de départ essentiel de notre démarche, mettant en lumière la nécessité de définir des critères de classification pertinents pour évaluer l'impact de l'absentéisme. Dans cette section, nous explorerons la logique de notre approche de classification, en examinant comment nous attribuons les labels "retard", "normal" et "absenteiste" en fonction de la durée d'absence.

Tout d'abord nous faisons l'hypothèse que chaque journée de travail est de 8 heures qui est la durée normale d'une journée de travail au Brésil formant ainsi notre base pour la catégorisations des individus.

Ensuite nous avons définis une focntion qui attribue une catégorie en fonction de la valeur de la variable "Absenteeisme time in hours". Voici comment la catégorisation est effectuée :

- Si la valeur est inférieure à 2 ($y < 2$), cet asbence de l'individu est classée comme "retard".

- Si la valeur est supérieure à 2 mais inférieure à 24 ($2 \leqslant y \leqslant 24$: moins de 3 journées de travail), cet asbence de l'individu est classée comme "normal".
- Si la valeur est supérieure 24 ($y > 24$: plus de 3 journées de travail), cet asbence de l'individu est classée comme "absence problematique".(Par soucis de simplicité, on note "absenteiste" dans les données)

Ainsi nous obtenons des données avec une nouvelle variable cible à 3 modalités:

```
normal        564
retard        132
absenteiste    44
```

.

Nous identifions maintenant un déséquilibre dans les données, ce qui peut affecter la performance du modèle. Pour remédier à cette situation, nous avons mis en place une technique de suréchantillonnage aléatoire (random oversampling). Cette méthode consiste à augmenter le nombre d'échantillons de la classe minoritaire (dans ce cas, les classes "absenteiste" et "retard") en ajoutant des exemples supplémentaires de manière aléatoire. Cette approche vise à équilibrer la distribution des classes, renforçant ainsi la capacité du modèle à généraliser de manière équitable sur l'ensemble des catégories d'absence.

Cependant une personne prenant en compte les raisons d'absences pourrait se dire que cette classication n'est pas correcte car parfois une personne peut être absent plusieurs jours à cause d'une raison médicale de manière involontaire. Mais même dans ce cas là cette absence reste problématique. Par exemple, considérons le cas d'une absence due à une maladie grave(ou une grossesse)nécessitant plusieurs jours d'absence. Bien que cela puisse être justifié du point de vue individuel et de la santé de l'employé, du point de vue de l'entreprise, cette absence prolongée peut entraîner des perturbations opérationnelles et nécessiter une gestion appropriée des ressources humaines(retard sur un projet, embauche CDD).

Ainsi, en dépit de la simplification induite par l'hypothèse de la journée de travail standard, cette approche demeure pratique pour évaluer les implications générales de l'absentéisme, en permettant une classification qualitative qui tient compte de la réalité opérationnelle de l'entreprise.

5 Modèles et Résultats

Dans cette partie consacrée à la présentation des modèles et de leurs résultats, notre approche vise à évaluer les performances de diverses méthodes de classification automatiques des individus en tant que "retardataires", "absences normales" ou "absences problématiques". Nous débutons par la mise en œuvre d'un modèle Naive Bayes, choisi délibérément comme référence de base. Par la suite, nous avons classifier avec un modèle d'arbre de decision.

5.1 Naïve Bayes

Avec notre modèle baseline Naïve Bayes, nous avons obtenu une précision de 46%, une performance notablement proche du meilleur modèle que nous avions construit précédemment avant l'intégration de la variable de classe définissant les catégories "retard", "normale" et "absenteiste". Cette proximité de résultats démontre clairement que l'ajout de ces catégories a considérablement facilité et amélioré la conception des modèles.

```
Rapport de classification:
                precision    recall  f1-score   support

 absenteiste       0.21      0.67      0.31        12
      normal       0.95      0.34      0.50       169
      retard       0.30      0.90      0.45        41

    accuracy                           0.46       222
   macro avg       0.49      0.64      0.42       222
weighted avg       0.79      0.46      0.48       222
```

Le rapport de classification révèle des performances variées pour chaque classe du modèle Naive Bayes. En ce qui concerne la catégorie "absenteiste", le modèle affiche une précision de 21%, suggérant que parmi les cas prédits comme "absenteiste", seulement 21% le sont réellement. Cependant, le recall, qui mesure la capacité du modèle à détecter tous les cas réels de "absenteiste", est relativement élevé à 67%. Cela indique que le modèle a une

propension à identifier un nombre significatif de véritables cas d'"absenteiste", bien que ses prédictions puissent inclure des faux positifs.

Pour la classe "normal", le modèle présente une précision remarquable de 95%, indiquant que la grande majorité des cas prédits comme "normal" le sont effectivement. Cependant, le recall est plus bas à 34%, suggérant que le modèle peut manquer de détecter certains vrais cas d'"normal".

Quant à la classe "retard", le modèle affiche une précision de 30%, ce qui signifie que parmi les prédictions positives de "retard", 30% sont correctes. Le recall est élevé à 90%, suggérant que le modèle est efficace pour identifier la plupart des vrais cas de "retard".

5.2 Decision Tree

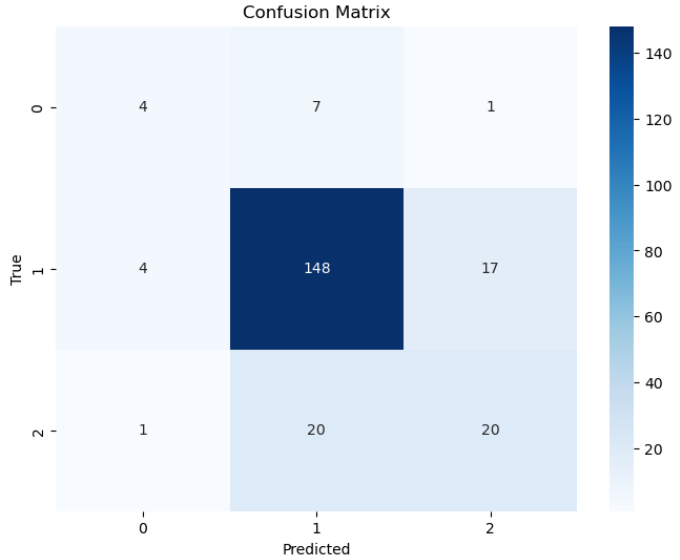
Comparé au modèle précédent basé sur le classificateur Naive Bayes, qui présentait une précision de seulement 46%, le modèle de décision Tree démontre une amélioration significative avec une précision globale de 78.%. Cette hausse substantielle suggère que le modèle de décision Tree offre une meilleure classification des données.

Rapport de classification:

	precision	recall	f1-score	support
absenteiste	0.44	0.33	0.38	12
normal	0.85	0.88	0.86	169
retard	0.53	0.49	0.51	41
accuracy			0.77	222
macro avg	0.61	0.57	0.58	222
weighted avg	0.77	0.77	0.77	222

En examinant le rapport de classification, on constate que le modèle obtient des résultats équilibrés pour chaque classe. La classe "normal" affiche une précision et un recall élevés, atteignant respectivement 85% et 88%, indiquant que le modèle est capable de bien classifier les individus avec des absences considérées comme normales. La classe "retard" présente également des scores de précision (53%) et de recall (49%) raisonnables.

La classe "absenteiste" montre une précision de 44%, ce qui signifie que parmi les cas prédits comme "absenteiste", 44% sont corrects. Le recall est de 33%, indiquant que le modèle a du mal à détecter tous les vrais cas d'"absenteiste". Cela suggère qu'il pourrait être nécessaire d'ajuster le modèle pour améliorer sa capacité à identifier cette catégorie spécifique d'absence.



La matrice de confusion met en évidence le nombre de prédictions correctes et incorrectes pour chaque classe. Elle révèle que le modèle a quelques difficultés à distinguer entre les classes "retard" et "absenteiste", comme en témoignent les éléments hors diagonale correspondant à ces deux classes.

Bien que le modèle Decision Tree ait montré de bonnes performances, il est important de noter qu'aucun modèle n'est parfait et que des ajustements pourraient encore être nécessaires pour améliorer la discrimination entre les différentes catégories. Cependant, avec une précision globale de 78%, il est actuellement considéré comme le meilleur modèle parmi ceux testés, offrant un équilibre entre la capacité de prédiction pour chaque classe. Des itérations supplémentaires et des ajustements fins pourraient potentiellement améliorer davantage les performances du modèle.

6 Conclusion

En conclusion, l'analyse des deux modèles, Naive Bayes et Decision Tree, révèle des performances variables dans la classification des niveaux d'absence. Le modèle Naive Bayes présente une précision relativement faible, particulièrement pour la catégorie "absenteiste", ce qui suggère des difficultés dans la prédiction précise de cette classe spécifique. En revanche, le Decision Tree offre des performances plus équilibrées avec une précision globale de 78

Le Decision Tree se distingue par ses scores de précision et de recall robustes pour la classe "normal", démontrant une capacité notable à identifier les absences considérées comme normales. Cependant, il montre des difficultés à différencier entre les classes "retard" et "absenteiste", comme indiqué par des scores de précision et de recall relativement modestes pour ces deux catégories. Mais ceci est du e partie au fait que les données ne soient pas équilibrées entre les différentes classesc car nous disposons que de très peu de temps d'absence très élevés dans les données.