

Projet Python & Statistiques ML

Introduction

De plus en plus, les entreprises cherchent à mieux connaître et à mieux comprendre leurs employés pour optimiser leurs performances. Pour parler de performance, il faut déjà un travail effectif, une présence effective de travailleurs. Les absences temporaires constatées alors qu'imprévues sont désignées par « absentéisme ».

L'absentéisme peut être indicateur d'une inadéquation du niveau de la qualité de l'environnement de travail avec le bien-être des employés dans l'entreprise. Par conséquent sa caractérisation a une importance notable pour le management. L'absentéisme peut aussi être indicateur de la structure sociodémographique avec des conséquences sur le fonctionnement du service et les recrutements futurs.

L'objectif de ce travail est de contribuer à mieux comprendre le problème de l'absentéisme à travers une démarche de catégorisation et de classification avec les absences constatées. La réalisation de cet objectif dépend de la conception qu'on a de l'absentéisme.

Nous proposons une approche basée sur la conception de l'absentéisme comme « toute absence qui aurait pu être évitée par une prévention suffisamment précoce des facteurs[...] »¹. De cette conception, et au regard des enjeux économiques des absences, se dégagent quelques questions intéressantes :

- **Quel est la durée des absences au sein de l'entreprise ?**
- **Quelles sont les variables qui influencent l'absentéisme² ?**

L'objectif principal de notre travail est de répondre à ces questions en utilisant les données d'une entreprise du Brésil. De plus, notre travail a pour but d'apporter une description globale des caractéristiques des employés afin d'aider à mieux connaître les employés.

Pour réaliser ces objectifs, nous procédons à une analyse exploratoire des données pour répondre au besoin de la description. Ensuite, nous procédons à l'identification des facteurs influençant l'absentéisme selon son niveau d'importance. Enfin, au moyen de ces facteurs, nous proposons une classification des employés.

Plusieurs modèles ont été mobilisées pour identifier les facteurs déterminants : le SVM, l'arbre de décision, le Boosting, les KNN et la forêt aléatoire. A chaque méthode, nous utilisons la cross validation pour avoir des résultats plus stables. Globalement, tous les modèles permettent une prédiction avec plus de 65% de précision.

¹ L'ABSENTÉISME OUTILS ET MÉTHODES POUR AGIR, ANACT

² L'ANALYSE LONGITUDINALE DE L'ABSENTEISME : CONSIDERATIONS METHODOLOGIQUES Jean-Paul DUMOND, 2006

I. Méthodologie

Sturman (1996)³ indique l'importance du choix de la méthode d'analyse de l'absentéisme. Selon cet auteur, le choix doit être guidée par la nature des données et l'objectif. De plus, l'auteur souligne que pour un objectif de recherche de facteurs importants relatifs à une variable comme les absences, la méthode classique de l'analyse par la corrélation réduit les facteurs importants aux facteurs ayant un lien direct avec les absences. Cette réduction des champs des facteurs importants n'est pas toujours pertinente.

Pour trouver des facteurs déterminant dans quelle catégorie d'absentéisme appartient une absence, nous utilisons une méthode de classification. La classification consiste à classer des données d'entrée à des catégories connues (classification supervisée) ou non connues (classification non supervisée). Dans le cadre de la classification supervisée, le but est d'identifier la méthode et les facteurs qui permettent de classer les valeurs dans les bonnes classes avec le moins d'erreurs possibles. Dans le cas de la classification non supervisée, les classes d'appartenance n'étant pas connues d'avance, le but est d'avoir des classes les plus homogènes possibles.

Dans notre approche d'analyse, nous avons des classes d'absence que nous avons prédéfinies suivant des objectifs stratégiques différents. Cette situation nous oriente vers une classification supervisée. L'hypothèse de la classification non supervisée offre la possibilité de définir 'automatiquement' les catégories d'appartenance selon la proximité de certaines caractéristiques. Par conséquent, nous utilisons plusieurs algorithmes d'optimisation afin d'obtenir le meilleur résultat : les KNN, l'arbre de décision, la forêt aléatoire, le SVM, le Boosting.

Les KNN permettent de prédire la classe en regardant la classe la plus répandue chez les K points les plus proches. L'arbre de décision nous permet de classer les individus selon des critères de test appliqués à chacune des variables caractéristiques. Il résume ainsi les variables permettant de classer les individus et les critères de choix utilisés à chaque niveau. A chaque partition est associée un indice d'homogénéité. Les classes retenues sont les plus homogènes possibles. Toutefois, les résultats de l'arbre de décision sont sensibles aux changements des données. Pour pallier cet éventuel problème d'instabilité, nous utilisons la démarche de classification de la forêt aléatoire. Cette démarche tient compte de la différence de résultats issus d'ensembles aléatoires d'échantillons, puis fournit un résultat agrégé. Elle réduit donc l'instabilité éventuelle. Cependant, une variance réduite ne signifie pas des erreurs peu importantes. Pour réduire les erreurs de prédiction et se rapprocher de la réalité, nous utilisons la méthode du gradient Boosting. Le SVM permet de classer des données même celles apparemment non dissociables.

En plus de la recherche de l'algorithme d'optimisation le plus adapté, pour chaque algorithme retenu, nous itérons les estimations de sorte à retenir le modèle avec le moins de variables avec le niveau de précision élevé. Notre démarche de sélection de variables (features) consiste à considérer toutes les variables, puis de retenir progressivement une à une les variables qui permettent d'améliorer la qualité

³ Multiple Approaches to Absenteeism Analysis Michael C. Sturman, 1996

prédictive déjà atteinte du modèle : « forward features selection ». En plus de cette stratégie, nous avons une liste de variables pré-sélectionnées suivant différents filtres tels que la corrélation, la variance, le chi-2 etc. L'idée avec ces filtres est simple : proposer une liste de variables vraisemblablement pertinentes au regard de la quantité d'information ou au regard du lien avec la variable cible.

Nous estimons les modèles généralement sur 3 listes⁴ : celle pré-sectionnée par le filtrage avant entraînement, celle proposée par la sélection avec la méthode forward et celle contenant l'ensemble des variables sans filtrage. Ensuite, nous comparons les résultats en matière de performance. Certaines variables sont transformées (regroupement ou éclatement) pour des raisons techniques ou de synthèse d'informations avant les estimations.

Pour apprécier la qualité des modèles, nous utilisons les métriques telles que le taux de précision, la sensibilité (recall) et le f1-Score (matrice de confusion et rapport de performance). Le principe de ces métriques est de comparer les prédictions à la réalité (vérité terrain). Elle permet de dénombrer les prédictions correctes et incorrectes. Pour chaque mesure, on cherche à avoir des valeurs les plus proches possible de 100%. De plus, nous prêtons attention à la qualité de prédiction de chaque classe. Pour obtenir des précisions plus stables, nous appliquons la cross validation.

La section suivante présente les résultats.

II. Principaux résultats

La base de données est composée de 740 lignes et 21 colonnes sans valeurs manquantes. Avec 13 variables quantitatives et 8 variables qualitatives. Les différentes caractéristiques constituant les colonnes sont l'ID de l'employé, l'âge (en années), le nombre d'heure d'absence, la distance domicile travail, les frais de transport, les raisons de l'absence, le poids, la taille, l'indice de masse corporelle, le nombre d'enfants, le nombre d'animaux de compagnie, la charge de travail, la réalisation des objectifs, le niveau de diplôme, les saisons de l'année, les jours de la semaine, le mois d'absence, temps de travail, discipline au travail, le fait d'être fumeur ou non et le fait d'être un buveur ou non (voir Annexe pour plus d'informations).

Notre variable cible est le nombre d'heures d'absence. Nous l'avons catégorisée en classe suivant trois objectifs stratégiques.

Objectif 1: distinguer 3 classes d'absentéisme avec des enjeux différents selon la durée:

- les absences assimilables à un retard: courte durée (<3h)
- les absences de moyenne durée (>=3h et <8h)
- les absences assimilables à des arrêts: longue durée (>= 1 jour)

⁴ Dans certaines situations (KNN), nous estimons seulement avec la liste sélection par la méthode « forward » car nous avons constaté que généralement celle méthode propose la meilleure performance

Objectif 2: distinguer deux classes d'absentéisme relatif à l'organisation des remplacements

- les absences de moins d'une journée: difficile à remplacer
- les absences d'au moins une journée

Objectif 3: objectif statistique tiré de la distribution de la variable

- les absences d'au plus 7 heures (moins d'une journée)
- les absences de 8 à 16 heures
- les absences exceptionnelles (>20 h)

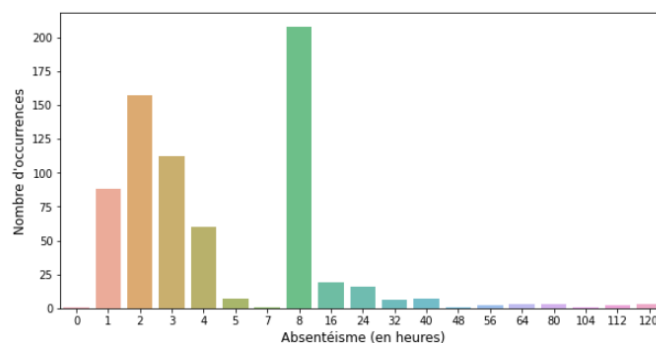
Dans ce premier rapport nous nous concentrons sur le premier objectif.

1. Résultats issus de l'exploration

Le fichier technique contient tous les détails importants. Nous présentons ici quelques éléments de résultats.


a) Les heures et les périodes d'absence

L'observation du nombre d'heures d'absence indique une fréquence élevée des absences d'une journée (8h). L'absence la plus longue est de 15 jours (120 heures). On peut donc distinguer les absences d'une journée et plus, les absences de moins de 3 heures pouvant être assimilées à des retards et les absences d'au moins une demie journée mais inférieures à une journée. En s'appuyant sur ce constat, nous avons regroupé les absences en 3 catégories correspondant aux limites



susmentionnées.

Le nombre d'absences le plus élevé est constaté les lundis. Les jeudis sont les jours avec moins d'absences. Les mois avec le plus d'absence sont respectivement le mois de mars et le mois de février et le mois de juillet. Le mois avec le moins d'absence est le mois de septembre.



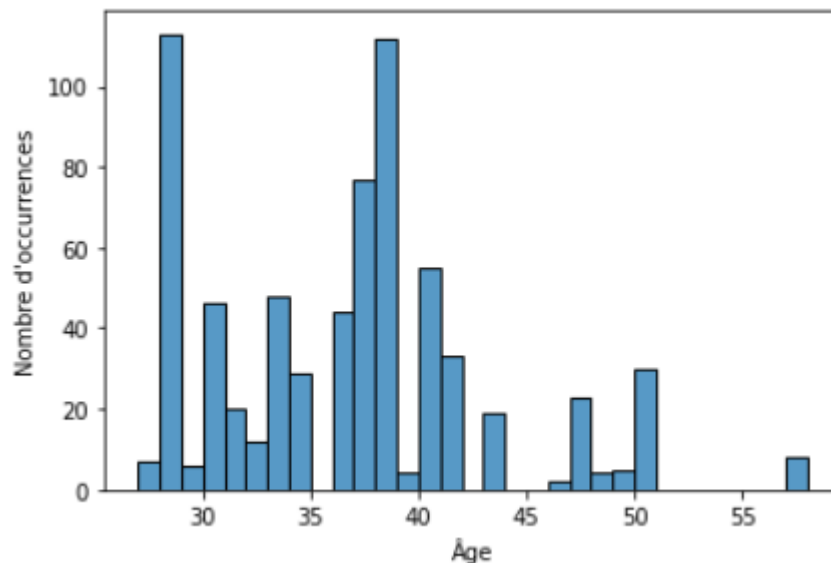
Type d'absence	Pourcentage
Absence justifiée	96%
Absence injustifiée	4%

Toutes les raisons d'absence sont essentiellement des raisons médicales. On constate que les raisons d'absences les plus fréquentes sont les consultations médicales (n°23) et la consultation dentaires (n°28)



c) La caractéristique âge

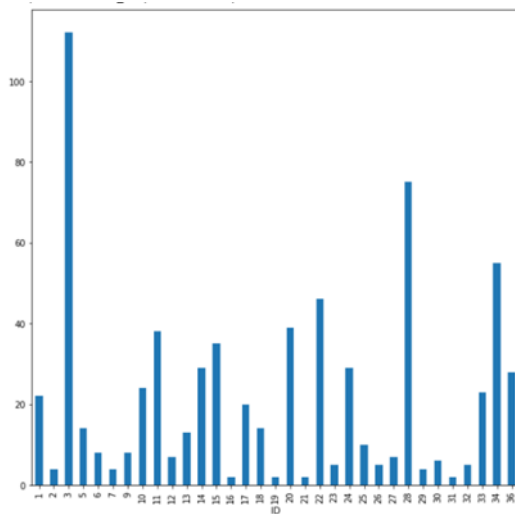
L'âge des employés dont les absences ont été enregistrées au cours de la période d'étude, est compris entre 27 et 58 ans avec 22 valeurs différentes. L'âge la plus fréquente dans le registre des absences constatées est 28 ans suivi de 38 ans. La figure ci-dessous suggère la possibilité de regrouper les âges en différentes catégories.



d) La caractéristique ID

L'analyse de ID montre que l'entreprise a **36 employés** (au cours de la période concernée) concernés par l'enregistrement des absences. Les employés ont une fréquence d'absences différente. L'individu 3 (voir figure ci-dessous) est celui qui a la plus grande fréquence d'absence (112 fois). Si on s'intéresse de plus près à cette personne, on constate que c'est une personne de 38 ans. Son domicile est à 51 kilomètres du travail. Son indice de masse corporelle (IMC) indique que c'est une personne obèse.

Les personnes avec le nombre d'absences le moins élevé sont les individus avec les ID 16,19, 21 et 31.

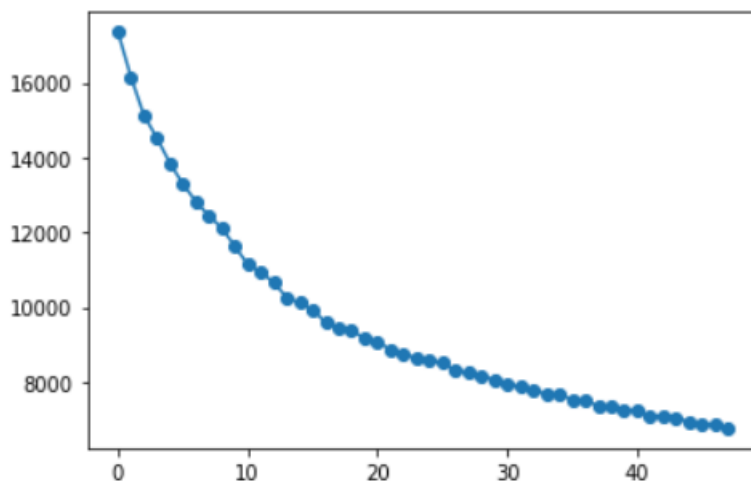


Par exemple, la personne avec l’ID a été absente 2 fois pour avoir fait une consultation médicale et pour des douleurs oculaires.

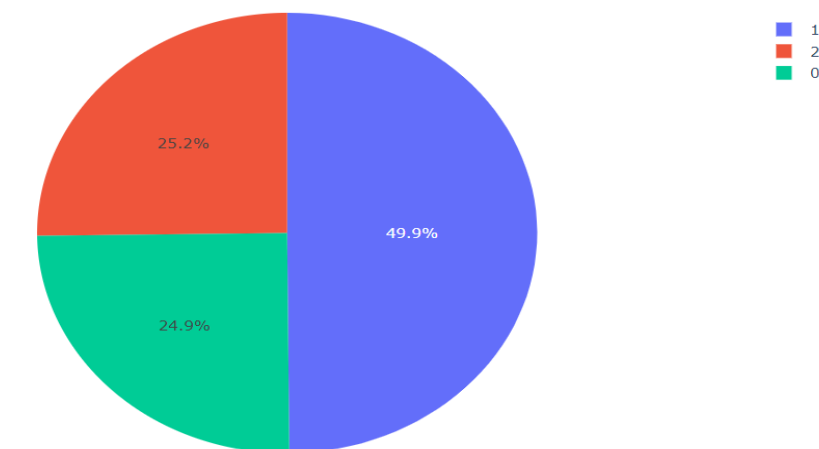
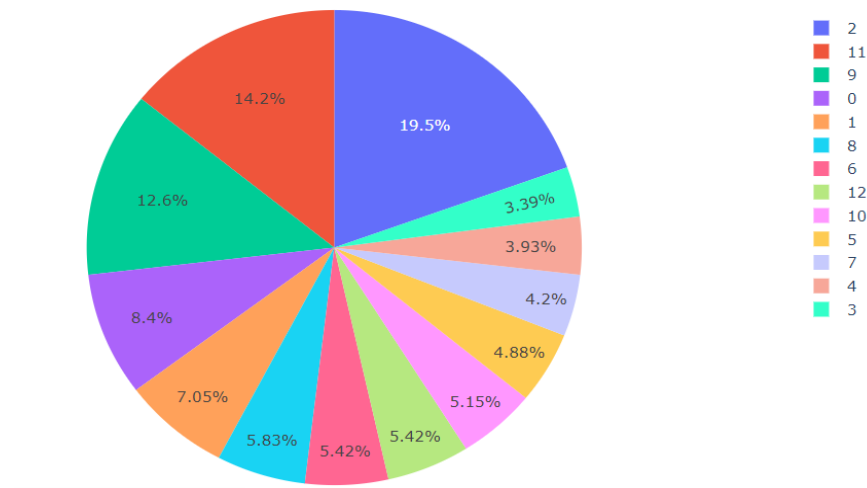
2. Résultats clustering

Une des attentes sur notre travail est la proposition d’un clustering des informations. Ce regroupement est une première approche de synthèse des informations.

Nous avons proposé donc un regroupement basé sur la moyenne des distances vectorielles entre les variables. Sur une plage de 2 à 50 groupes essayés, la courbe d’inertie indique que le meilleur niveau de regroupement possible est celui avec 13 groupes ou à 3 groupes. Ainsi, 13 groupes de situations peuvent être distingués.



Ce nombre élevé est aussi un signal de la difficulté à résumer les informations en un nombre plus réduit. La classe 2 à 19,5% de l'effectif total : c'est la plus grande classe. La classe avec la plus petite la classe 3 avec 3,39% des informations. On constate que 3 grandes classes se distinguent par la proportion d'information. Cela suggère la possibilité d'avoir 3 classes.



Avec 3 classes nous avons la répartition suivante : la première classe à 50% des informations. Les classes 2 et se partagent l'autre 50%.

Avec ces résultats, pour chaque variable, par exemple, il est possible de voir les différents groupes d'appartenance et apprendre de leur point commun ou détecter les comportements assez singuliers.

3. Résultats classifications

L'approche du clustering est une approche non supervisée. Dans le point suivant nous avons des classes d'absence et nous cherchons à y classer les différentes situations enregistrées : nous sommes à mesure de vérifier la qualité de la classification en comparant la prédiction à la réalité.

Pour notre objectif (1), un arbre de décision sur la liste des variables permet d'avoir des prédictions précises à près de 75%

L'autre modèle qui permet d'avoir un résultat proche est le SVM: il semble proposer le meilleur modèle de prédiction avec une forte stabilité.

a) L'arbre de décision

Avec cet arbre, en fonction des valeurs de certaines variables, il est possible de classer.

Par exemple si la raison de l'absence n'est pas 0 (raison non définie) et est 23 (XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified). Et Si en plus la personne a 4 enfants alors l'absence sera entre 4 et 8 heures.

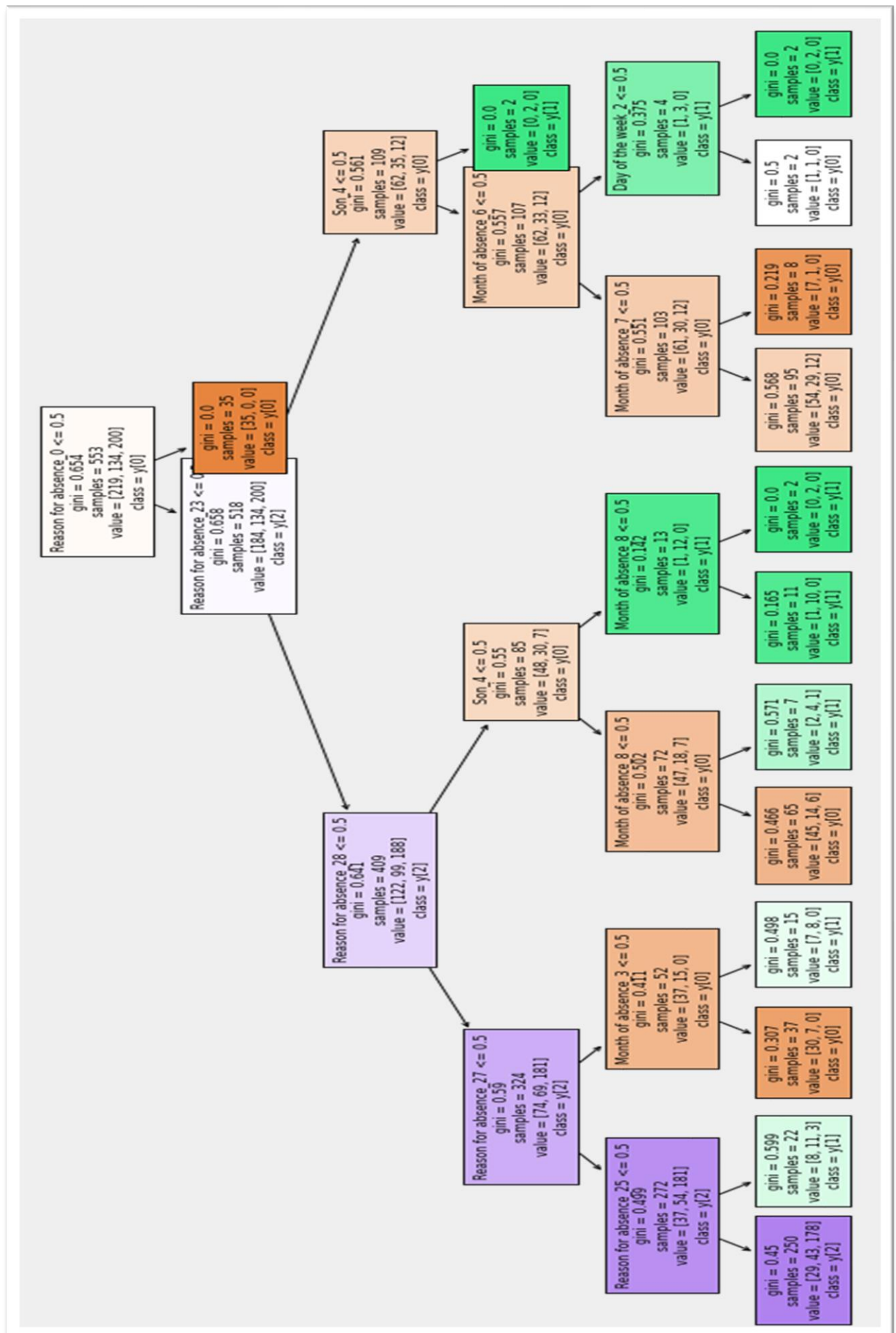
Selon cet arbre, les informations les plus pertinentes pour la classification sont les raisons de l'absence, le nombre d'enfants et la période de l'absence (mois, jours).

Le rapport de performance associé indique une précision de près de 70% : environ 7 absences sur 10 sont globalement bien classées.

La classe 1 (deuxième classe) est la plus difficile à prévoir. La sensibilité (recall) de cette classe est faible : seul un quart des absences appartenant réellement à cette classe y sont effectivement bien classées. Cependant, de manière générale, plus de 60% des absences classées dans cette classe sont bien classées (précision de la classe).

CLASSE	PRECISION	RECALL	F1-SCORE
0	0.69	0.81	0.75
1	0.67	0.26	0.38
2	0.79	0.96	0.86
ACCURACY			0.73
MACRO AVG	0.71	0.68	0.66
WEIGHTED AVG	0.72	0.73	0.70

Malgré cette performance globale, cet arbre de décision est sujette de possibles instabilités. Ce qui nous amène à rechercher une méthode permettant d'avoir un résultat plus stable.



b) Vers un résultat plus stable

Le tableau de performance suivant donne les performances obtenues avec différents modèles.

De manière générale, le SVM permet d'avoir la performance la plus élevée devant le Random Forest et le Boosting. Cependant en fonction du métrique d'intérêt et de la classe d'intérêt, le choix varie. La classe la plus difficile à prévoir est la classe 1 pour tous les modèles. Ainsi, le modèle à retenir lorsque l'entreprise prête attention à la sensibilité (recall pour s'assurer une meilleure classification des absences) de la classe 1 (deuxième classe) est le Boosting. De même, si l'entreprise cherche particulièrement à avoir le meilleur niveau de performance en matière de précision pour la classe 1 alors elle se penchera vers le Random Forest.

CLASSE	SUPPORT VECTOR MACHINE			GRADIENT BOOSTING		
	precision	recall	f1-score	precision	recall	f1-score
0	0.71	0.90	0.79	0.75	0.63	0.68
1	0.75	0.20	0.32	0.49	0.39	0.43
2	0.75	0.91	0.82	0.68	0.91	0.78
ACCURACY	0.74			0.66		
MACRO AVG	0.74	0.67	0.65	0.64	0.64	0.63
WEIGHTED AVG	0.74	0.74	0.69	0.66	0.66	0.65
CLASSE	RANDOM FOREST			K-NEAREST NEIGHBOR		
	precision	recall	f1-score	precision	recall	f1-score
0	0.67	0.82	0.73	0.63	0.55	0.59
1	0.85	0.24	0.38	0.48	0.32	0.38
2	0.72	0.88	0.79	0.65	0.87	0.74
ACCURACY	0.70			0.62		
MACRO AVG	0.74	0.65	0.64	0.59	0.58	0.57
WEIGHTED AVG	0.73	0.70	0.67	0.60	0.62	0.60

Par rapport aux variables importantes, globalement tous les modèles évoquent les raisons de l'absence, le nombre d'enfants, la distance domicile travail. A celles-ci viennent s'ajouter, selon le modèle, parfois la tranche d'âge, le temps de service ou le moment des absences.

Conclusion

Après une exploration des données et un clustering, ce travail propose différentes solutions alternatives pour répondre au besoin de classification des absences en 3 catégories relatives à une stratégie de gestion des absences bien précisée. Il ressort qu'une catégorie des absences est plus difficile à prévoir même avec des modèles globalement performant pour 70% des cas. En termes de recommandations, cette difficulté suggère un besoin de flexibilité dans la gestion du personnel afin de mieux gérer ces incertitudes par l'ajustement des tâches. Il ressort aussi que l'un des facteurs déterminant les absences sont les raisons médicales : cela questionne aussi la qualité de l'environnement de travail par rapport à la situation des employés.

Une perspective possible de prolongement ce travail est de proposer des solutions répondant aux autres objectifs de gestion des absences. La réduction à deux classes pourrait améliorer les performances de prédiction.

Annexes: Attribute Information:

1. Individual identification (ID)
2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons

6. Transportation expense
7. Distance from Residence to Work (kilometers)
8. Service time
9. Age
10. Work load Average/day
11. Hit target
12. Disciplinary failure (yes=1; no=0)
13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son (number of children)
15. Social drinker (yes=1; no=0)
16. Social smoker (yes=1; no=0)
17. Pet (number of pet)
18. Weight
19. Height
20. Body mass index
21. Absenteeism time in hours (target)