UNIVERSITÉ D'ARTOIS

Université d'Artois

FACULTE DE SCIENCES JEAN PERRIN

FOUILLE DE DONNÉES POUR LA PRÉDICTION DE CLASSES D'HEURE D'ABSENTÉISME AU TRAVAIL

Présenté par

Mohamed YOULA

Supervisé par

Mr. Karim Tabia

Le 30/11/2020

SOMMAIRE

INTRODUCTION :

Dans ce projet, nous disposons d'un jeu de données sur l'absentéisme au travail. Il contient **20 variables explicatives** (numériques et nominales) ainsi qu'une **variable à expliquer**, correspondant au nombre d'heures d'absence de certains individus. Certains individus apparaissent sur plusieurs lignes dans cette base de données. Le jeu de données comprend **740 lignes** et **ne comporte aucune donnée manquante**.

Ce jeu de données, nommé **Absenteeism_at_work.csv**, est disponible à l'adresse suivante :

https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Avant toute tâche de fouille de données, les prétraitements suivants ont été effectués :

- **Conversion de certaines variables** : certaines variables saisies comme numériques ont été converties en variables nominales. Cette opération a été appliquée aux variables suivantes : *ID, Reason_for_absence, Month_of_absence, Day_of_the_week, Seasons, Education, Social_drinker* et *Social_smoker*.

- **Normalisation des variables numériques** : les attributs numériques ont été normalisés à l'aide de la technique de normalisation *Min-Max*, car certaines variables n'étaient pas sur la même échelle (par exemple, l'âge et le poids).

Durant ce projet, le logiciel **R** a été utilisé pour le prétraitement, la préparation et la visualisation des données, tandis que l'application **Weka** a été employée pour l'évaluation des classifieurs.

L'objectif principal de ce projet est d'extraire le maximum d'informations à partir des données du fichier *Absenteeism_at_work.csv*. Pour ce faire, les étapes suivantes ont été réalisées :

1. **Analyse de la variable cible** (*Absenteeism_time_in_hours*) en estimant sa distribution à l'aide de la technique du *bagging*, puis détection et suppression des valeurs aberrantes.

2. **Problème de classification** : création de deux classes de la variable cible afin d'identifier les facteurs influençant un nombre d'heures d'absence élevé ou faible au travail.

3. **Extraction de règles d'association** pour enrichir l'interprétation et mieux comprendre les résultats de la fouille de données.

Analyse de la variable cible via des estimateurs de densité

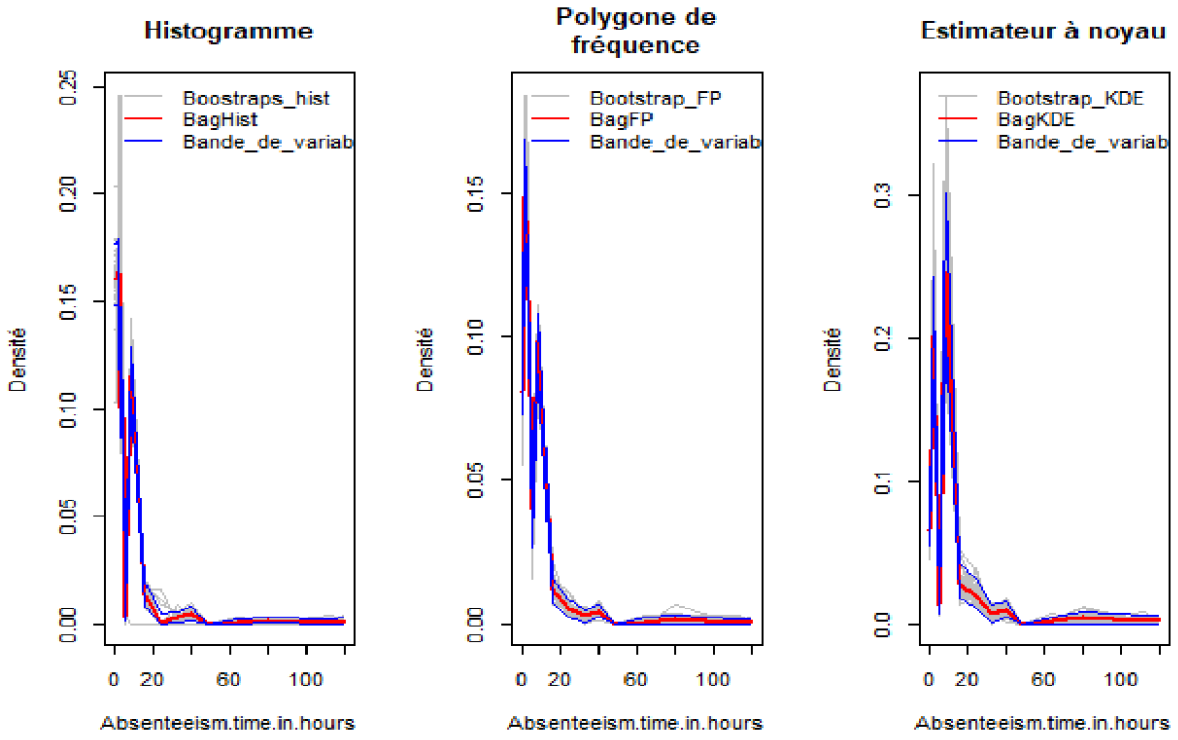
Dans cette partie, nous utilisons trois types d'**estimateurs de densité** :

- l'**histogramme**,
- le **polygone de fréquence**,
- et l'**estimateur à noyau gaussien**.

Afin de rendre ces estimateurs plus robustes, nous appliquons la technique du **bagging**, qui permet :

- de générer et d'agréger plusieurs estimateurs de densité du même type,
- et de construire une bande de variabilité à partir de laquelle une **métrique** est calculée pour sélectionner le meilleur estimateur.

Après application de cette méthode aux valeurs de la variable `Absenteeism_time_in_hours`, nous obtenons les résultats suivants :

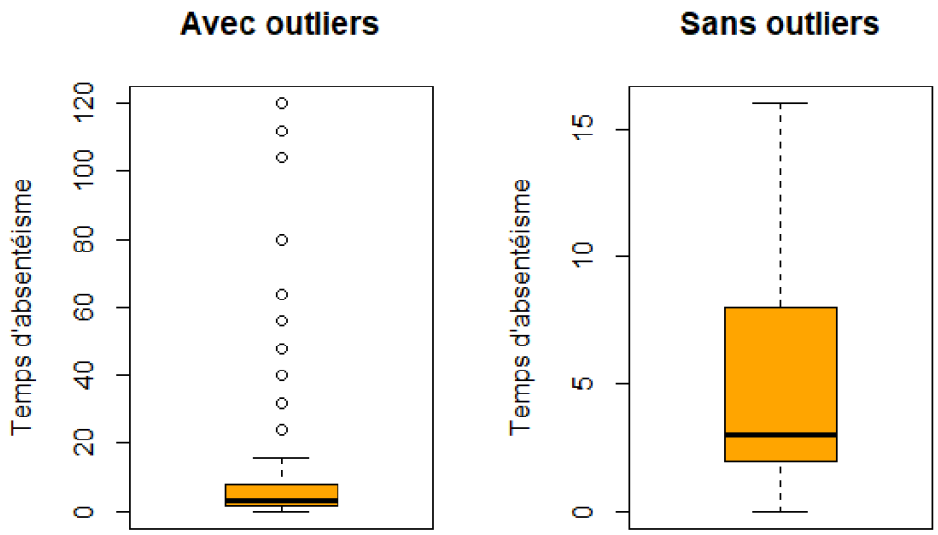


Erreur-Histogramme	Erreur-Polygone-de-frequence	Eurreur-Estimateur-à-noyau
0.026	**0.023**	0.072

D’après ces résultats, nous constatons que le **polygone de fréquence** est le meilleur estimateur, car il fournit l’erreur la plus faible.

L’allure de la courbe de densité associée à cet estimateur montre que les **valeurs aberrantes** apparaissent à partir d’un peu moins de 20 heures jusqu’à au-delà de 20 heures, et qu’elles sont peu fréquentes. Ainsi, les valeurs inférieures à 20 heures sont les plus denses dans la variable *Absenteeism_time_in_hours* de notre base de données.

Les **boîtes à moustaches** ci-dessous permettent d’obtenir des informations numériques supplémentaires. La première met clairement en évidence la présence des outliers et les valeurs qu’ils prennent.



	Min	1st Qu	Mediane	Moyenne	3rd Qu	Max
Avec outliers	0.000	2.000	3.000	6.924	8.000	120.000
Sans outliers	0.000	2.000	3.000	**4.293**	8.000	16.000

La **moyenne**, étant très sensible aux valeurs aberrantes, est de **6,924** lorsqu'elles sont prises en compte, et de seulement **4,293** lorsqu'elles sont retirées de l'ensemble des valeurs de notre variable.

Pour la suite de l'analyse, nous supprimons donc les lignes correspondant à ces valeurs aberrantes, afin de **réduire le bruit** dans nos données.

Classification pour prédire les classes de la variable cible

Pour notre problème de classification, nous travaillons sur le jeu de données **sans outliers**. La variable à prédire est transformée en **variable de classe** en utilisant la moyenne (**4,293 heures**) comme seuil.

- **Classe A** : individus dont l'absence est **inférieure à 4,293 h**
- **Classe B** : individus dont l'absence est **supérieure ou égale à 4,293 h**

Nous cherchons donc à résoudre ce problème en appliquant différents classifieurs.

Le jeu de données utilisé est **normalisé**, mais les classes y sont **déséquilibrées** :

- 461 instances pour la classe **A**
- 235 instances pour la classe **B**

Après évaluation des classifieurs, nous obtenons les résultats suivants :

	Accuracy (%)	**Précision**	**Rappel**	**ROC Area**
J48-10.cv	83.477	0,833	0,835	0,862
JRip-10.cv	77.5862	0,772	0,776	0,761
Baging-J48-100	82.9023	0,827	0,829	**0,883**

Baging-JRip-100	80.1724	0,798	0,802	0,850
AdaBoostM1 J48.100	79.454	0,794	0,795	0,864
AdaBoostM1 JRip.100	79.7414	0,795	0,797	0,847
RandomForest	81.0345	0,807	0,810	0,879

Tab1

Comme les classes sont **déséquilibrées**, le choix du meilleur classifieur repose sur la valeur de l'**aire sous la courbe ROC (ROC Area)**, car cette métrique est insensible à la répartition des instances entre les classes.

Le meilleur classifieur est donc **Bagging-J48 (100 itérations)**, qui fournit la valeur la plus élevée (**0,883**).

Ensuite, nous travaillons sur un **jeu de données équilibré** en affectant des poids aux instances, de manière à ce que chaque classe ait le même poids total.
Nous obtenons alors les valeurs des différentes métriques pour chaque classifieur, présentées dans le tableau ci-dessous.

	Accuracy (%)	**Précision**	**Rappel**	**ROC Area**
J48-10.cv	83.1463	0,832	0,831	0,864
JRip-10.cv	77.3143	0,774	0,773	0,805
Baging-J48-100	82.3912	0,824	0,824	0,885
Baging-JRip-100	**83.6636**	0,837	0,837	0,894
AdaBoostM1 J48.100	77.6185	0,782	0,776	0,865

AdaBoostM1 JRip.100	80.468	0,805	0,805	0,872
RandomForest	80.714	0,810	0,807	0,882

Tab2

Cette fois-ci, en observant les valeurs de l'**accuracy**, mais aussi celles de la **précision** et du **rappel**, on constate que le meilleur classifieur est le **Bagging-JRip (100 itérations)**. Pour ces trois métriques, ce modèle obtient les meilleures performances.

Évalué sur un **jeu de données équilibré**, ce classifieur s'avère supérieur au **Bagging-J48 (100 itérations)**, que nous avions sélectionné comme meilleur modèle sur le jeu de données déséquilibré. En effet, ses valeurs d'accuracy, de précision, de rappel et même de **ROC Area** sont toutes plus élevées.

Règles d'association

Ci-dessous figurent les premières règles d'association extraites à l'aide de l'algorithme **JRip** avec une validation croisée à 10 folds. Ces règles permettent de mieux caractériser les absences au travail.
Rappel : `Absenteeism_time_in_hours = B` correspond à une absence supérieure à **4 heures**.

1. **(transportation.expense ≥ 235) AND (Age ≤ 33) AND (Disciplinary.failure = 0) → Classe B**
   → Les individus de moins de 33 ans, sans défaillance disciplinaire, et avec des frais de transport supérieurs à 235, ont tendance à s'absenter plus de 4 heures.

2. **(Reason.for.absence = 26) AND (Age ≤ 43) → Classe B**
   → Les individus de moins de 43 ans dont l'absence est injustifiée (code 26) présentent une probabilité plus élevée d'absence supérieure à 4 heures.

3. **(Reason.for.absence = 22) → Classe B**
   → Les individus dont l'absence est liée à des raisons de santé (code 22) ont tendance à s'absenter plus de 4 heures.

4. **(Weight ≤ 67.99) AND (Work.load.Average.day ≥ 264.60) AND (Age ≥ 41) → Classe B**
   → Les individus de plus de 41 ans, avec un poids inférieur à 68 kg et une charge de travail quotidienne moyenne supérieure à 264,60, s'absentent généralement plus de 4 heures.

5. **(Reason.for.absence = 13) AND (Service.time ≥ 11) → Classe B**
   → Les individus dont la raison d'absence est le code 13 et ayant plus de 11 années de service tendent à dépasser 4 heures d'absence.

Conclusion

Les investigations menées dans ce projet ont porté sur un **problème de décision** lié à l'absentéisme au travail.

- Dans un premier temps, les **valeurs aberrantes** de la variable cible ont été identifiées et supprimées, afin de réduire le bruit dans les données.

- Ensuite, il a été montré que, sans ces outliers, le meilleur classifieur est obtenu via le **bagging** associé aux règles d'association.

- Enfin, l'analyse des règles extraites avec **JRip** a confirmé que la présence d'outliers aurait perturbé la qualité de la prédiction des deux classes construites